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 we 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

List of Figures

List of Tables

List of Appendices

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 [93]. Trillions of these bacteria live in the gut [38], and have a massive metabolic potential. For example, the gut microbiome has been shown to produce changes in hormone levels [63], short chain fatty acid levels [104], and ethanol levels [49], to name a few. The human gut microbiome can even digest polysaccharides otherwise unusable by humans [29].

This massive metabolic potential produces measurable symptomatic 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 [103]. 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 [71].

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 [21]. 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 [61], people who are susceptible and resistant to a condition [100], or patients before and after a medical intervention [37]. 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 metagenomic experiments that can be done with microbiome next generation DNA sequencing data: gene tag abundance (Fig. 1.4.3) and metagenomic sequencing (Fig. 1.5.1) [87]. The tag used for gene tag abundance here is the conserved 16S rRNA gene that is presumed to track taxonomic identity [35].

1.3 Illumina next generation sequencing platform

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.

[TO DO: add advantages/limitations section of illumina tech, compared to Solid, 454]

Unlike a technology such as qPCR, Illumina and other next generation DNA sequencing technologies deliver data as parts per machine limit, not a count of molecules in the sample.

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 and non-pathogenic in other diseases.

3. If the microbe is isolated in pure culture, it can be used to induce the disease [47].

One group has created a modified set of postulates that takes DNA sequencing into account [30] which can be applied to differentially abundant taxa detected by gene tag sequencing. However, Koch's postulates do not account for when the same bacteria can have a very different expression profile in health and disease, such as Lactobacillus iners in bacterial vaginosis [61]. Gene tag abundance takes ua 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 [113]. 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, where the

microbiome is much more diverse [42]. 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.

[TODO: figure out what Greg's comment "or AU" means] [TODO: find citation for above]

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.

In atopic dermatitis, a flare-up is defined as an acute exacerbation of disease despite standard treatment. Flare-ups are associated with an increase in the proportion of *Staphylococcus aureus* on the skin [48].

[TODO: make the above a paragraph, consider adding in something about how s. aureus affects inflammation]

Bacteria have also been used for therapy in the treatment of *Clostridium difficile*. 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 [80].

[TODO: add bit about "stool transplant (NIEJM)"]

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 DBM, the UniFrac distance, and the Bray Curtis dissimilarity.

[TODO: figure out what dbm is]

Sometimes when the data is plotted, there appears to be separation between groups, even if specific taxa are not differentially abundant. One example of this is a study on discordant gut microbiomes between twins in Malawi where one twin has kwashiorkor and the other is healthy [96]. In this case the microbiomes 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 [84], the Ribosomal Database Project [17], and Greengenes [20].

Specifically, this work uses the 16S rRNA gene primers from the Earth Microbiome Project protocol [32], 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 [109].

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 are 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.

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 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. Closed refernece 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.

[TODO: "REF HERIZ, EDGAR SOMEWHERE"]

1.4.3 General protocol and rationale

The 16S rRNA gene sequencing experiment (Fig. 1.4.3) 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.

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

1. Take a biological sample and extract the DNA
   The sample can be collected swabbing the target body site or by collecting samples in some other way. DNA extraction is usually done with common commercial kits.

2. Run a PCR amplification
   As discussed previously, the gene tag experiments in this thesis amplify the V4 region of the 16S rRNA gene, following the Earth Microbiome Project protocol [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 [35].

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 variable region.

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

1. 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 to cluster the reads into groups of 97% identity [24].

4. Annotate the OTUs with bacterial taxonomy
   Annotation was done by matching our OTUs to the SILVA database [84].

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 [23].

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].

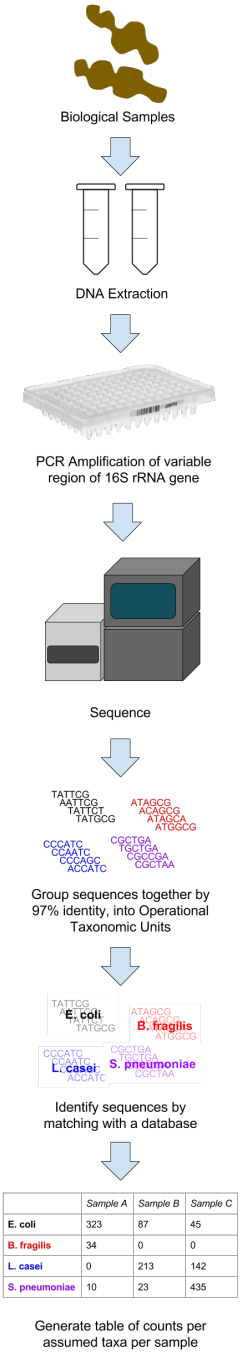

Figure 1.1: **16S rRNA gene tag experiment workflow.** This shows the workflow from sample collection to data generation. The end result is a count table of reads per operational taxonomic unit per sample.

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 Co-ordinate Analysis (PCA) is necessary for multivariate statistics. However, the OTU abundances derived from the 16S rRNA gene sequencing experiment are proportional, so the Principal Co-ordinate Analysis (which assumes a linear differences) ia 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 Principal Component Analysis (PCoA) 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 [59]. 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 [58]. Since then, papers reporting these metrics have garnered over a thousand citations, and enabled insights about everything from how kwashiorkor causes malnutrition [96] to how people can have a similar microbiome to their pet dog [97]. Except for Generalized UniFrac, used to make hybrid unweighted and weighted UniFrac comparisons [15], few advances in the metric have occurred since 2007.

Unweighted UniFrac

Unweighted UniFrac uses an inferred evolutionary distance to measure similarity between samples (Fig. 1.4.4). 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

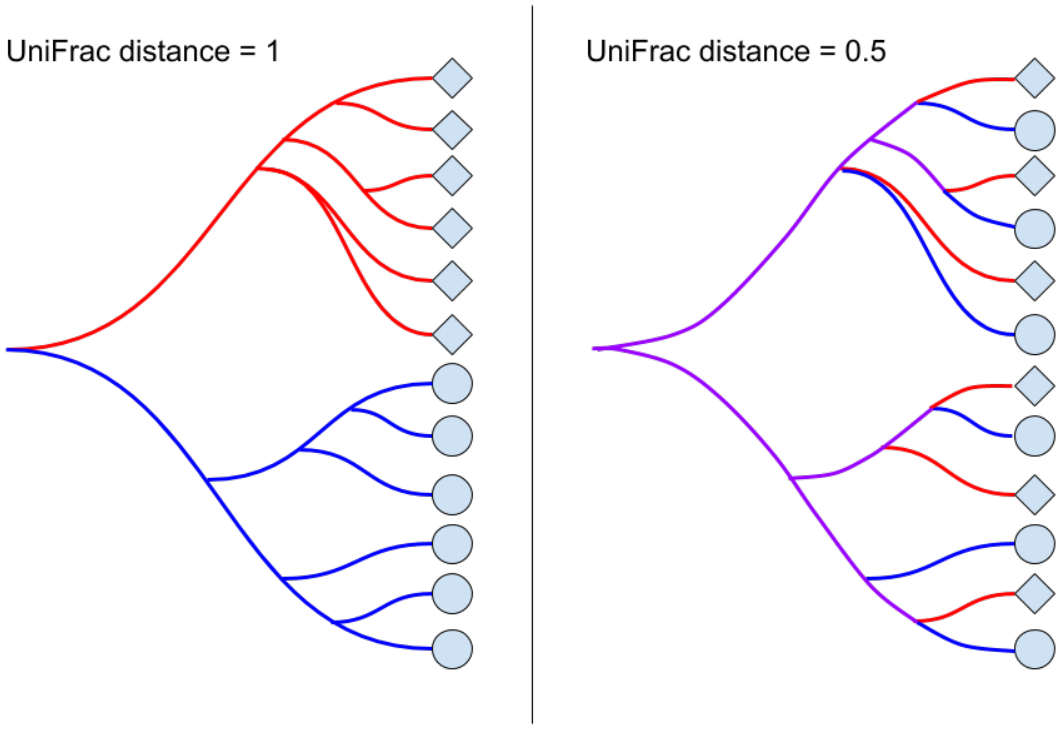

Figure 1.2: **Unweighted UniFrac.** When two samples do not share any branches of the phylogenetic tree, the unweighted UniFrac distance is maximized at 1. When two samples share half of their branch lengths on the phylogenetic tree, the unweighted UniFrac distance is 0.5. If the two samples contain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

rarefaction to normalize the sequencing depth across samples by random sampling without replacement [14], although this is controversial.

[TODO: add waste not want not citation]

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.

[TODO: figure out which UnIFrac paper to cite for above]

This is disregarded even by the original authors.

[TODO: find some citations of unifrac usage by lozupone/knight]

[TODO: generate the above figure (just use left half of sample migration figure from unifrac chapter)]

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. 1.3). Any published finding done with an unweighted UniFrac analysis is suspect,

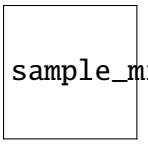

Figure 1.3: **Sample migration in different rarefactions, plotted on principal components, measured with unweighted UniFrac.** The plot is of the tongue data set while the right plot is the tongue dorsum vs. buccal mucosa data set. Red samples have moved from the left cluster to the right cluster between rarefactions. Blue samples have moved from the right cluster to the left. Samples are taken from the tongue dorsum body site from the Human Microbiome Project database. If the experiment were run once, one might mistakenly assume that there are two clusters of data, however, the inconsistent sample membership of the two groups between rarefactions proves the clustering irreproducible. Note that the variance explained 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.

especially 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*.

Weighted UniFrac

Weighted UniFrac is an implementation of the Kantorovich-Rubinstein distance in mathematics, also known as the earth mover's distance [26]. 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 reduced impact on the total distance reported by the metric.

[TODO: add figure with unweighted and weighted UniFrac (no information), cite in text]

UniFrac is constituted as either a presence/absence (unweighted UniFrac) [59], a linear proportion in the form of weighted UniFrac [58], or some combination of the two in the form of Generalized UniFrac [15]. However, the data are not linear, because the sum of the total number of reads is constrained by the sequencing machinery [31]. Alternative weightings and nonlinear transformations of data need to be explored, and this is the focus of Chapter 2.

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 [22], shortened as PCoA. PCoA is a single value decomposition of the distane 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.

[TO DO: add tongue/tongue and tongue/cheek comparison to show PCoA and variation explained, cite in text]

After multidimensional scaling the data can be analyzed in several ways. The data can be examined by k-means analysis clustering [101] or by unsupervised clustering.

[TODO: add citation for 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 coordinate 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 PCA 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 has more genes related to methanogenesis, compared to the average sequenced microbe [33].

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 [103], sparking more research into the gut microbiome and obesity related ailments such as diabetes [53] and nonalcoholic fatty liver disease [114]. 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. 1.5.1) 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 [74] or KEGG [46] 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 [52] 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 sampled multiple times before the sequencing data is retrived: The biological sample collected from the patient is only part of the full bacterial community. The amount of DNA extracted is a sample of the 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 varation present in the microbiomes being sampled, there is an additional layer of technical 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.

[TODO: find a citation for above]

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.

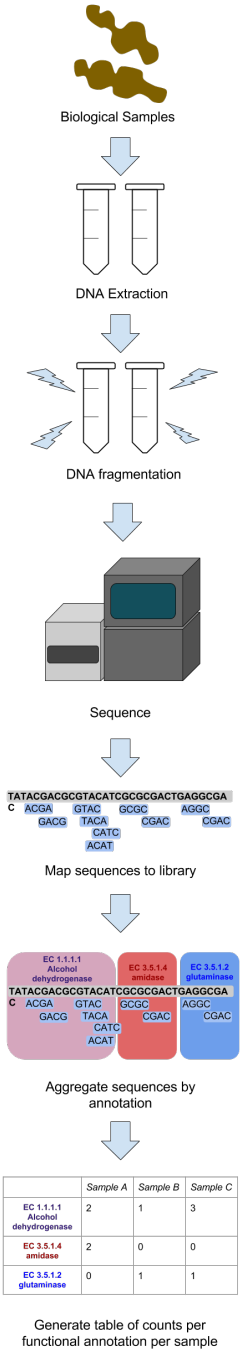

Figure 1.4: **Metagenomic experiment workflow.** This shows the workflow from sample collection to data generation. The end result is a table of number of sequencing reads per functionally annotated gene per sample.

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 [51]. PiCrust uses the Greengenes database [20] to identify the bacterial taxa in the sample, and pulls their genomes from the Integrated Microbial Genomes database [64]. With the genomes, the program tries to predict what would be seen if the samples underwent deep metagenomic sequencing. For taxa without a fully sequenced genome, PiCrust infers the genetic content based on ancestors in the phylogenetic tree. PiCrust produces metagenome predictions with Spearman r= 0.7 [51], 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

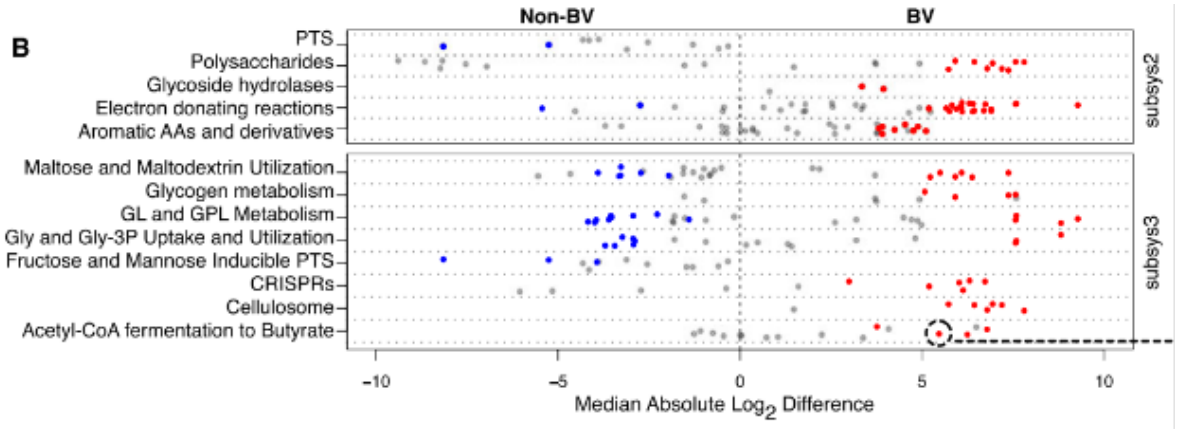

Figure 1.5: **Example stripcharts for subsystem 2 and 3 functional categorizations.** Dots on the left side are SEED subsystem 4 annotations found to be more abundant in the healthy condition while dots on the right side are subsystem 4 annotations found to be more abundant in the bacterial vaginosis condition. Colored dots were found to be significantly differentially abundant. Figure taken from [61].

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. 1.5.3). 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 [46]. 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. Next generation sequencing is also a sensitive technology, and the data can be confounded by contaminants 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.

[TODO: cite brady bunch paper above]

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 [105]. 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.

[TODO: reference fat data and combinatorics]

Generally experiments of this nature typically have low sample sizes due to budget constraints, sample collection difficulties, patient compliance, and other issues. To increase cost effectiveness and reduce batch effects, we run all the samples in an experiment on the same sequencing run, by means of a combinatorial barcode primer design [35].

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 [91] and Metastats [77] for microbiome analysis. Most of these were originally designed for RNA-seq experiments on single organisms [75]. Currently the most popular tools for analyzing differential abundance are EdgeR [89], DESeq2 [56], MetagenomeSeq [78], and LEfSe [91]. EdgeR was cited by 1,130 papers in 2015 according to Google Scholar. DESeq2 and MetagenomeSeq are part of the QIIME pipeline, which was cited by 1,620 papers in 2015. LEfSe has been cited by 393 papers.

[TODO: limit citations to microbiome stuff only (not rna-seq)]

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 [27].

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

[TODO: add section for lefse]

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 [28]. After a number of samples, the mean value and mean variance are used to determine if OTUs are differentially abundant between groups, an

approach that is believed to result in greater sensitivity and equivalent specificity compared to the point estimate approach [28].

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 is to be reliable [73].

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 ofen 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.

[TODO: add NAFLD venn diagram and cite in text]

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.

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 (1×100^5) 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. This is analogous to microbiome data where raw counts are normalized by dividing by the total number of counts [79].

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 analysed as multivariate data. All non-parametric tests (Principal Coordinates Analysis, correlations, etc.) assume that the abundance measured for each feature is independant. 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.

[TODO: include greg's window and animals figure, and cite in above text]

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. A correlation between two OTUs is suspect if it is dependant 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]. 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 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 [28] and the Analysis of Composition of Microbiomes (ANCOM) framework [62] 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 non overlapping results.

[TODO: include figure refernece above]

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 [34] [27] [28]. 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 [76].

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 [81]. 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 in the liver, and decreasing the 5 year survival rate to 67% [83]. If we can shed some light on the process by which people progress from NAFLD to NASH, we might be able to find treatments to prevent NASH.

Several genetic [99] [88], epigenetic [70], hormonal [112], and metabolite [85] factors are known to affect the risk progression to NASH. The relationship between the gut microbiome and non alcoholic fatty liver disease is less clear.

A 2001 paper performed C-D-xylose-lactulose breath tests and measured tumor necrosis factor alpha levels to determine presence of bacterial overgrowth, and found increased bacterial overgrowth in 22 patients with NASH compared to 23 healthy controls [110]. Some papers claim a link between ethanol-producing gut bacteria and NAFLD [114] [44], however, no multiple test correction was performed in these studies. Five published studies claiming to have found differentially abundant bacteria in the gut microbiome between healthy controls and patients with non alcoholic fatty liver disease have nearly non-overlapping results [114] [111] [85] [44] [8].

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

[TODO: EXPAND] [TODO: discuss prior results]

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 microbiomes that incorporated phylogenetic distance [59]. 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 [58]. Since then, papers reporting these metrics have garnered over a thousand citations, and enabled research about everything from how kwashiorkor causes malnutrition [96] to how people can have similar microbiomes to their pet dogs [97]. Except for generalized UniFrac, used to make hybrid unweighted and weighted UniFrac comparisons [15], 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 variable region are considered to be the same taxa [16]. 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 [24]) that partitions the groups and then later assign taxonomic identity by matching the seed or central sequences with public databases, such as SILVA [84], the Ribosomal Database Project [17], or Greengenes [20]. 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: 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 [102]. 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 [23], or using a guide tree, such as through Greengenes [20] 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. 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. 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. *Prevotella* was found to increase through non-BV to intermediate to BV, while *Lactobacillus iners* stayed relatively the same. 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 *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. 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 [59] 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 calcualtions 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 [60], and 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 [90].

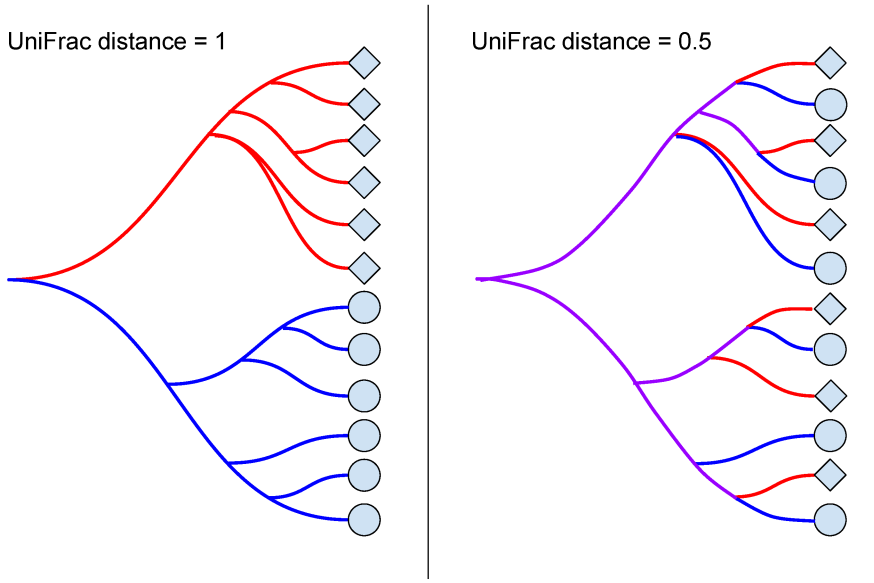

Figure 2.1: **Unweighted UniFrac.** When two samples do not share any branches of the phylogenetic tree, the unweighted UniFrac distance is maximized at 1. When two samples share half of their branch lengths on the phylogenetic tree, the unweighted UniFrac distance is 0.5. If the two samples contain 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 [58] is an implementation of the Kantorovich–Rubinstein distance in mathematics, also known as the earth mover's distance [26]. 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) [59], a linear proportion (weighted UniFrac) [58], or some combination of the two (generalized UniFrac) [15]. 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 [31] [27] [28] [57]. 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 different read count totals. Vaginal samples from patients with bacterial vaginosis compared to patients without can have total counts that differ one magnitude [115]. Alternative weightings

and non-linear transformations of data need to be explored. Furthermore, unweighted UniFrac is known to be unreliable, but it is not generally known or understood how this can impact results.

Materials and Methods

2.0.5 Analytical techniques

Rarefaction

Rarefaction normalizes the samples OTU counts to a standard sequencing depth [95]. This resulting table can be thought of as a random point estimate of the dataset, as the output is a sub-sample 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 [72].

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.

Weighted UniFrac

Weighted UniFrac [58] also incorporates each branch length of the phylogenetic tree, and weights them according to proportional abundance of the two samples. The formula for weighed UniFrac is as follows, where A and B are the two samples, b is the set of branch lengths, and $\frac{A_i}{A_T}$ and $\frac{B_i}{B_T}$ are the proportional abundances associated with branch length b_i:

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

Information UniFrac

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

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

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, because a given sequence read is more likely to be taxa A. When the amount of uncertainty that a taxa has in one sample corresponds with the amount of uncertainty the same taxa has in a different sample, the abundance of that taxa is mutually informative between samples. Weighting UniFrac by uncertainty combines the the concept of uncertainty with phylogenetic relationships to identify taxa that are differentially informative between groups.

The formula for Information UniFrac is as follows:

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

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 [25].

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 [27]. 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 outliers. The formula for ratio UniFrac is as follows, where gm is the geometric mean:

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

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

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

One challenge when it comes to the analysis of read count data is 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 [27]. A Bayesian approach can be used to give a posterior estimate of the likelihood for zero: this is implemented by the cmultRepl command in the zCompositions package in R [76].

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: `https://github.com/ruthgrace/ruth_unifrac_workshop`

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:

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

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

$$S_i = \text{Specimen counts at site i}$$

$$S_j = \text{Specimen counts at site j}$$

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/JRWu/R_Scripts`.

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 [106] 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 component analysis is drawn from each distance matrix per metric, and for the first principal component of each metric, the resultant value (V_{res}) is computed per each first principal component 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 [106] 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 [72] 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 component 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 [107]. 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`. 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 components or coordinate analysis 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 [60] 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 component 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 component 1. Principal component 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 hundred separate rarefactions were conducted on the data to a minimum sampling depth of 378. For each individual rarefied OTU table, a distance matrix was computed using 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 rarefation 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

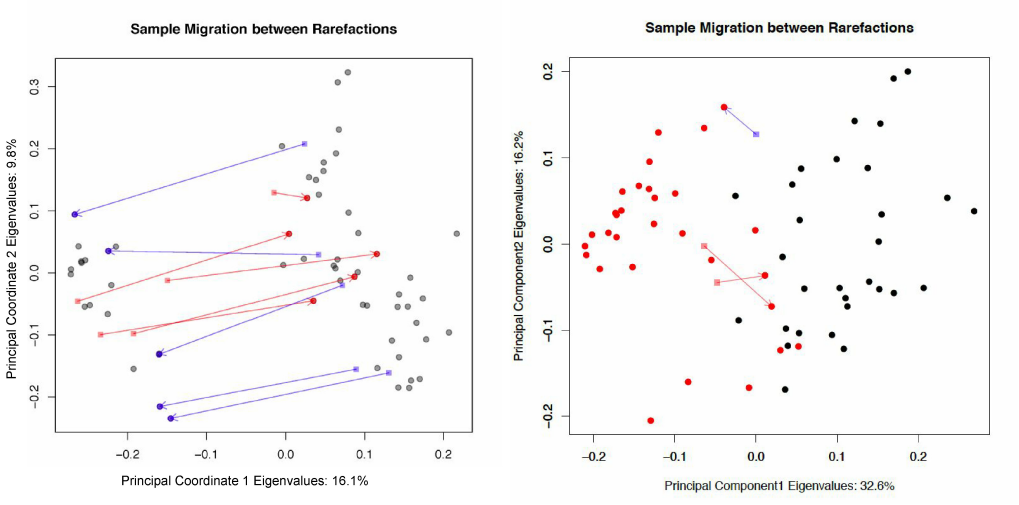

Figure 2.2: **Sample migration in different rarefactions, plotted on principal components, measured with unweighted UniFrac.** The left plot is of the tongue data set while the right plot is the tongue dorsum vs. buccal mucosa data set. On the left panel red samples have moved from the left cluster to 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 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 (compare with 32.6% and 16.2% in the tongue dorsum vs. buccal mucosa data set). The variance explained in the first and second component in the 2011 UniFrac commentary [60] was even smaller, at 8.6% and 5.6%.

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 component 1 and 2 effects, we suggest caution in their interpretation. For example, see the use of unweighted UniFrac in these papers about the human microbiome published in Cell[41], where the first and second principal components axis explain 14% and 9.5% of the variation in Figure 2A, as well as in Nature [98], where the first principal component explains 14% of the variation in Figure 1. In both of these examples, less variance is explained by the first principal component than in our uniform tongue data set.

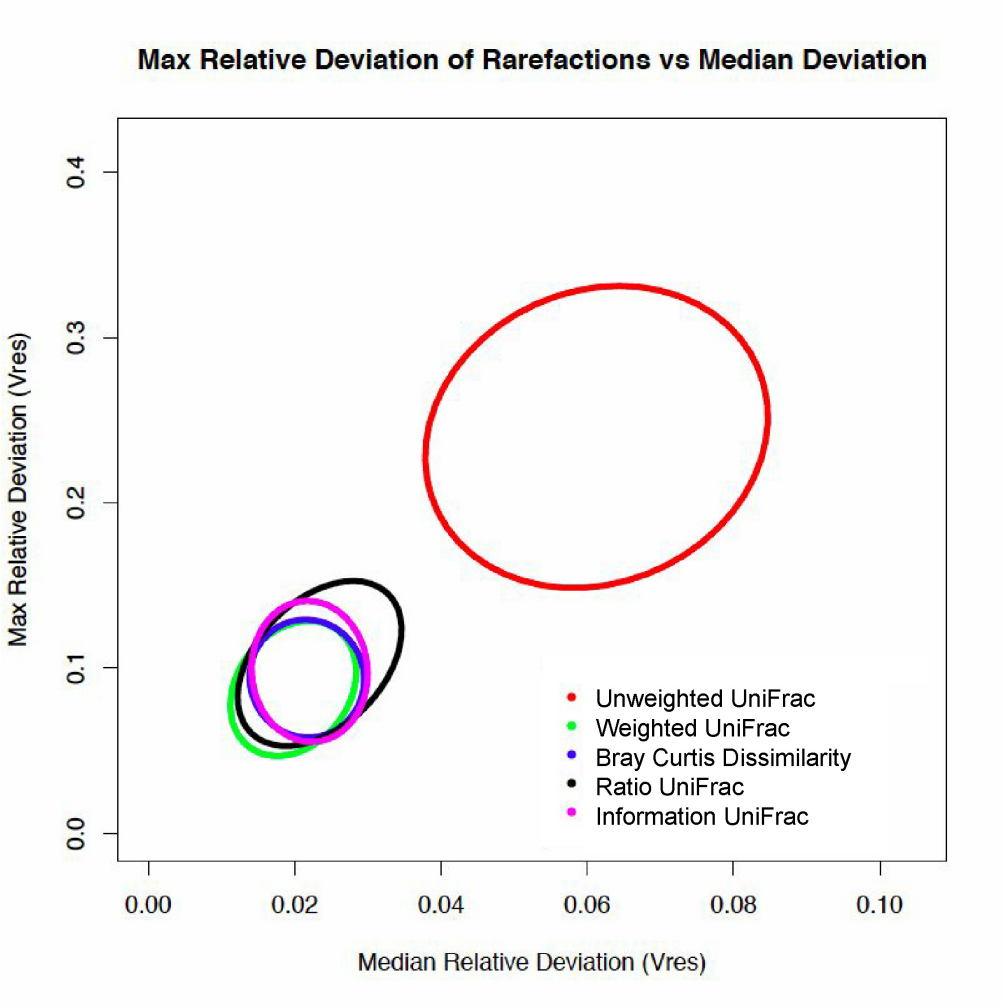

Figure 2.3: **Maxiumum relative deviation of rarefactions versus median deviation for traditional and non-traditional microbiome dissimilarity metrics.** Sixty samples from the tongue dorsum were taken from the Human Microbiome Project [106], and rarefied 100 times. The maximum relative deviation was 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 distance, ratio UniFrac, and information UniFrac.

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 [36]. 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.

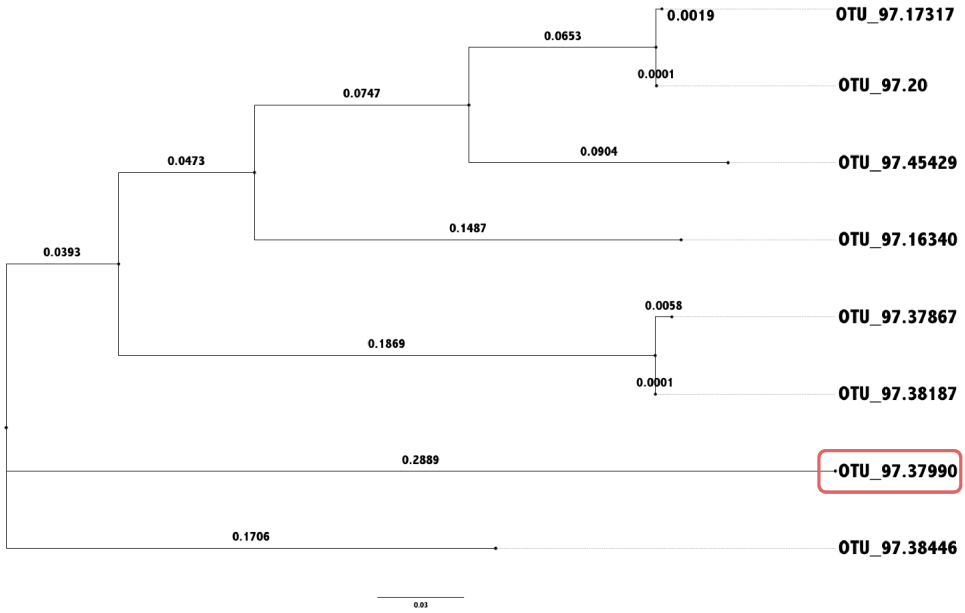

Figure 2.4: **Phylogenetic tree with long isolated branches.** Variation in different rarefactions of data in unweighted UniFrac analysis is exacerbated by the presence of long isolated branches in the phylogenetic tree, such as the circled OTU in this example.

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

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. 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 unweighetd 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

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.

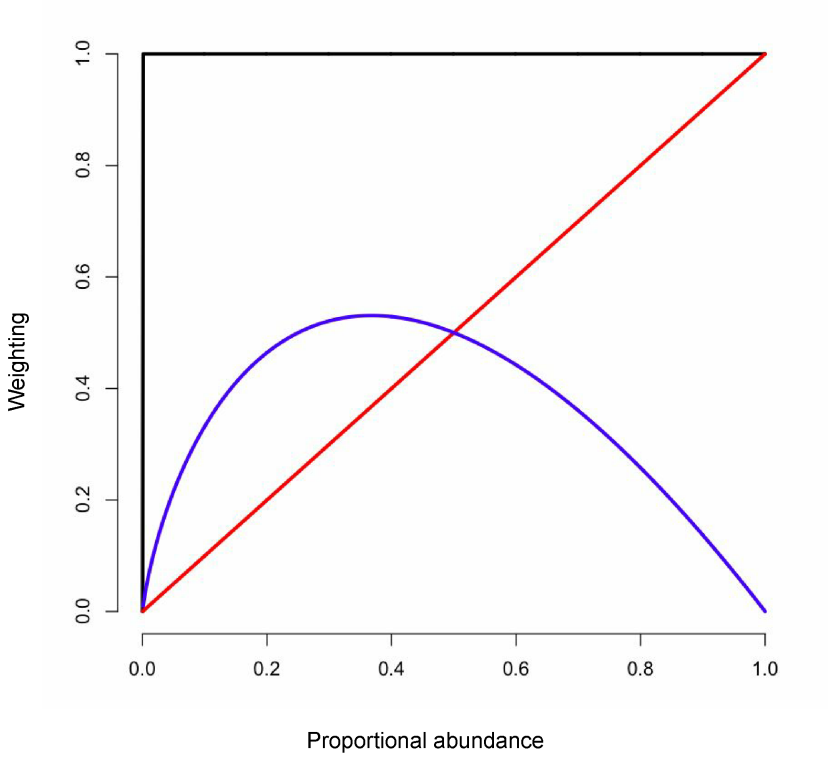

Figure 2.5: **UniFrac weights.** Each UniFrac weighting is plotted with the corresponding proportional abundance. The black line is unweighted UniFrac, the red line is weighted UniFrac,and the blue line is information UniFrac. From 0 to 0.2 on the x-axis information UniFrac has a higher slope, and therefore more discovery power with 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 component 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.

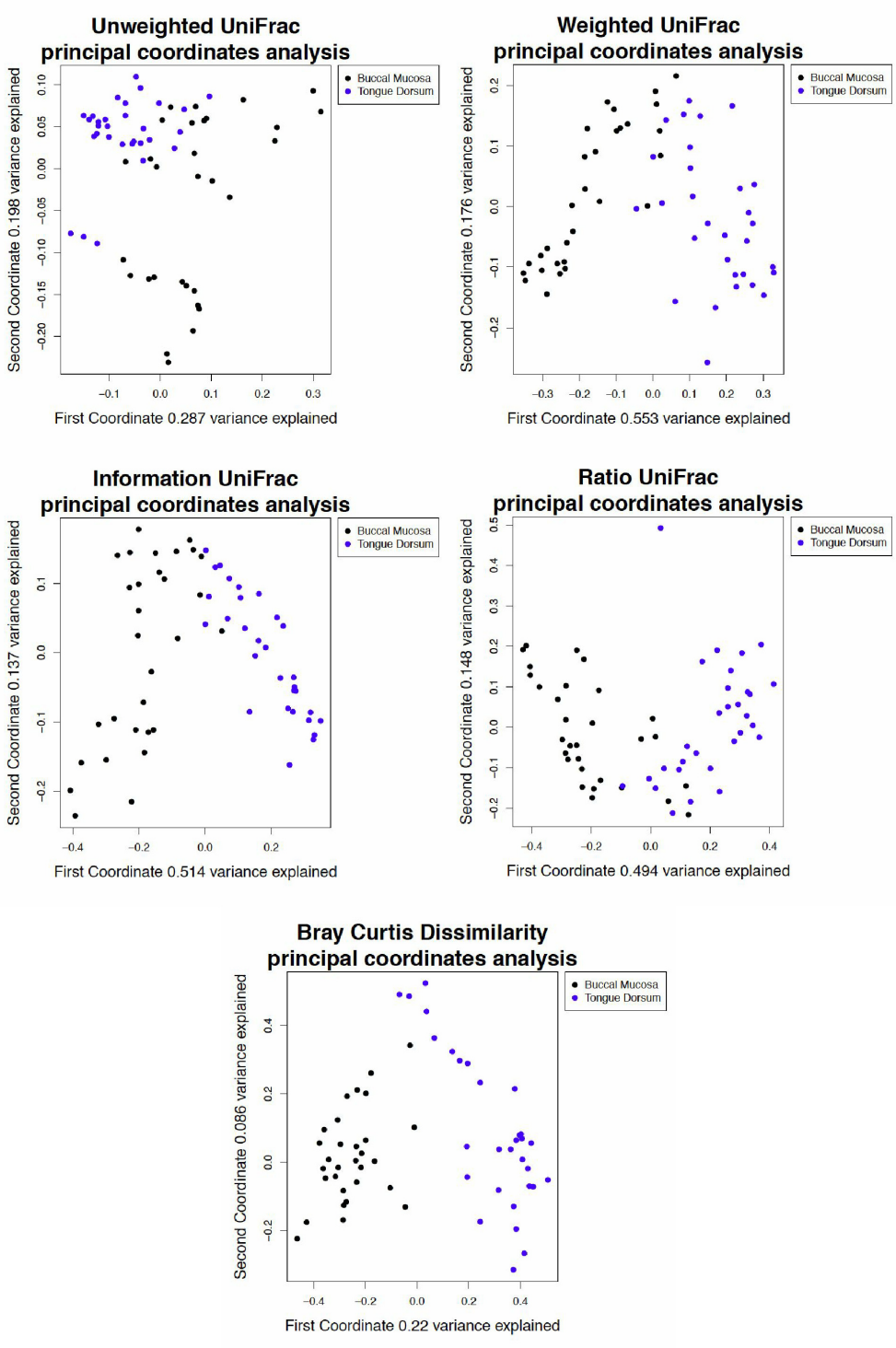

Figure 2.6: **Analysis of tongue and buccal mucosa data using different UniFrac weightings.** A principal component analysis of a 16S rRNA gene tag experiment done on samples from the tongue and buccal mucosa, selected from the Human Microbiome Project [106]. All weightings and the Bray-Curtis dissimilarity show separation between the samples by body site. Note that the variance explained by the first and second principal component 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 component analysis of a 16S rRNA gene sequencing experiment done on microbiome samples from breast milk [107]. 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.

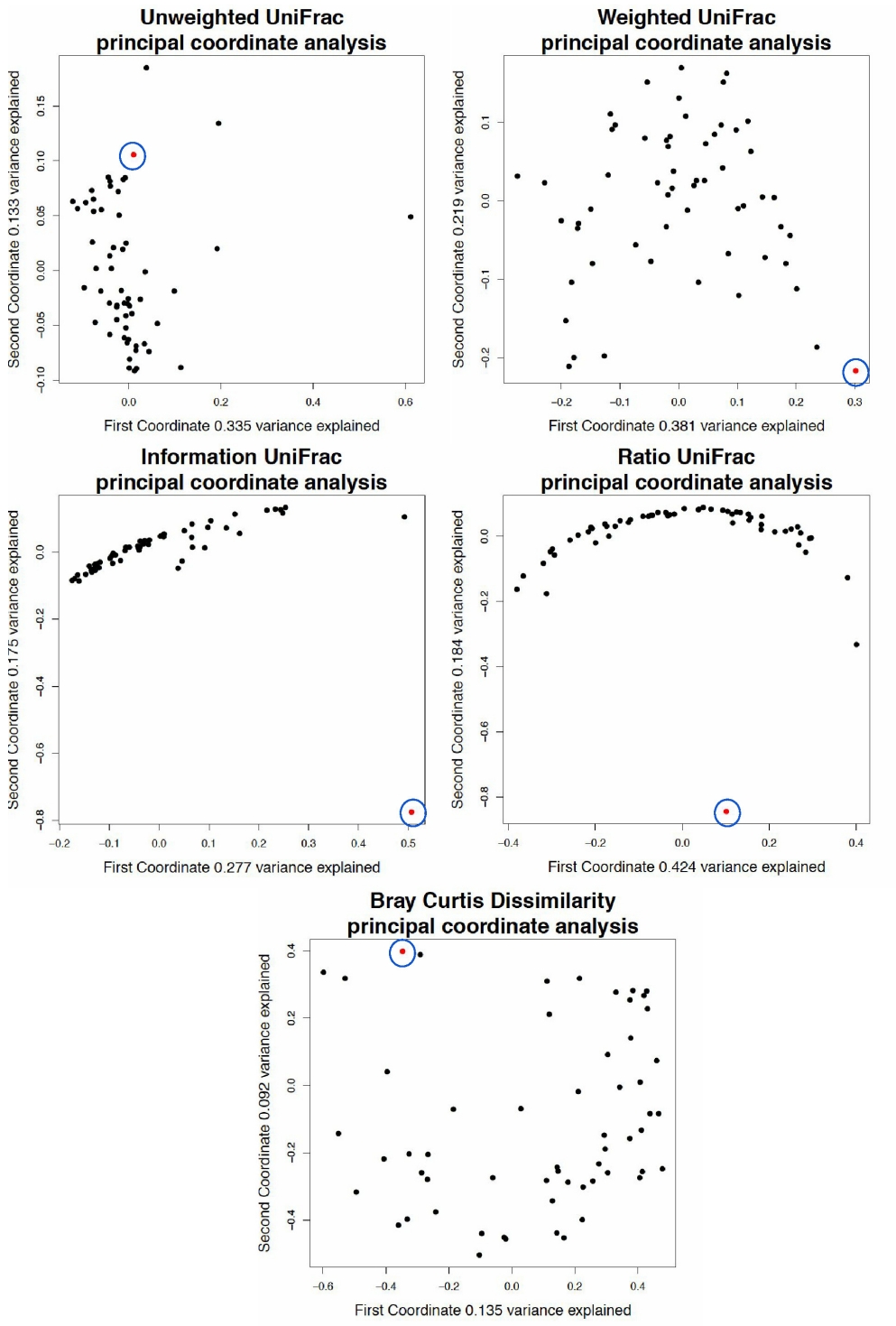

Figure 2.7: **Analysis of breast milk data using different UniFrac weightings.** A principal component analysis of a simulated 16S rRNA gene tag experiment based on the breast milk data. Red samples are dominated at 07% by *Pasteurella*, black samples are dominated by *Staphylococcus*, and cyan samples 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 components 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 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 component (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.

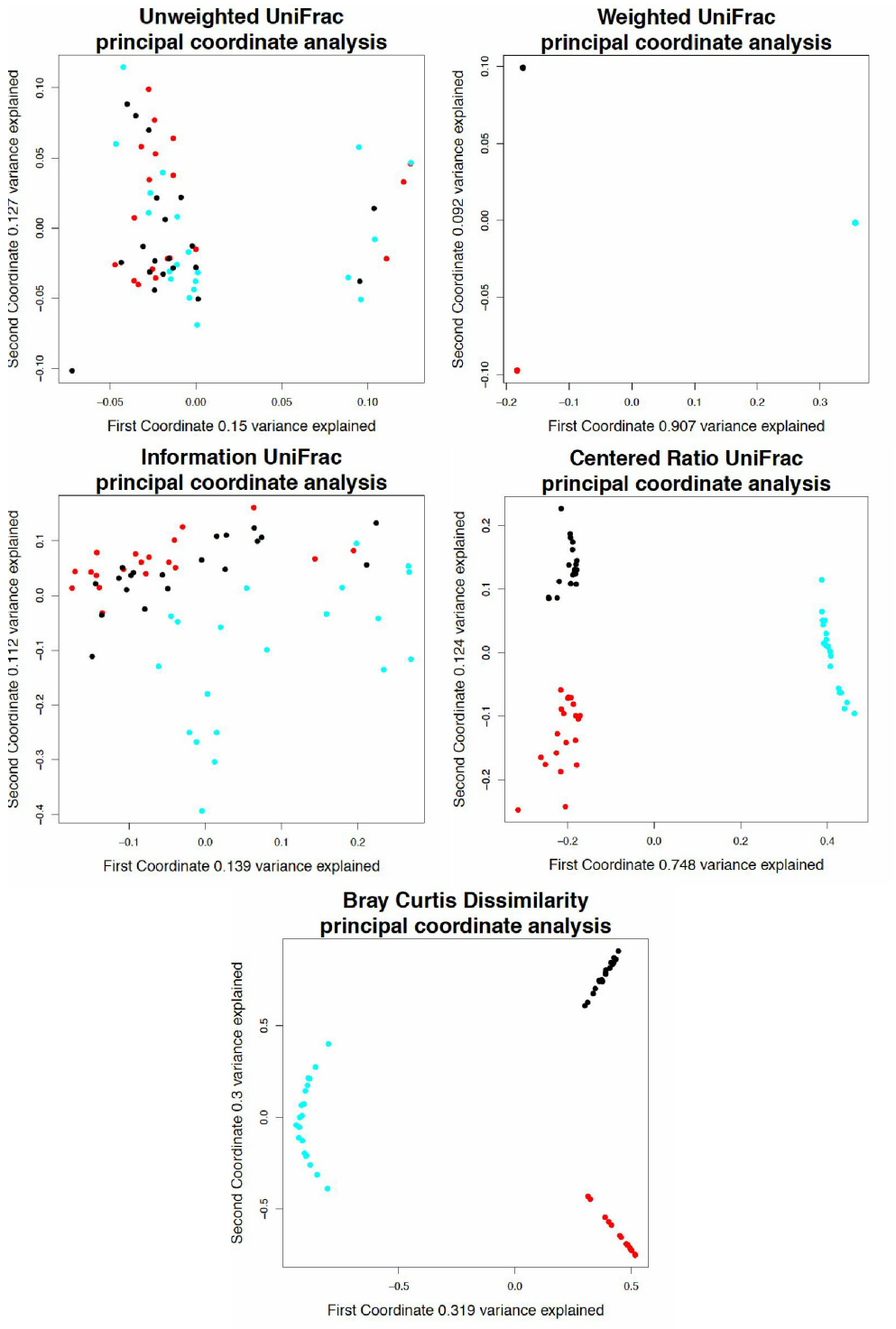

Figure 2.8: **Analysis of simulated monocultures using different UniFrac weightings.** A principal component analysis of a simulated 16S rRNA gene tag experiment based on the breast milk data. Red samples are dominated at 97% by *Pasteurella*, black samples are dominated by *Pseudomonas*, and cyan samples 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 components 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 our 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 the similar sets of taxa between groups. Ratio UniFrac is especially good for examining data sets that have more subtle variations, due to its 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.

Acknowledgments

Thanks to Camilla Urbaniak for providing the data from her breast milk study [107].

unifrac

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 a fifth to a third of the North American population [81]. Most people with NAFLD remain asymptomatic, however, in up to a third of patients NAFLD can progress to nonalcoholic steatohepatitis (NASH), causing inflammation and scarring (fibrosis) in the liver, and decreasing the 5 year survival rate to 67% [83]. 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.

Mouse
In mice non alcoholic fatty liver disease is often modelled with a methionine/choline-deficient diet (MCD), which induces steatohepatitis in wildtype mice. Mice with a toll-like receptor 4 knockout had lower lipid and injury accumulation markers when fed a MCD diet [88].

Rat
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 [112]. From this, hormones are also a factor in nonalcoholic fatty liver disease progression.

Human
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 [99]. The heterozygous gene was found to be associated with fatty liver disease in genome wide association studies, but some additional studies have faileid to replicate the relationship with NASH [99].

On the epigenetic level, many genes are differentially methylated in advanced NAFLD compared to mild NAFLD. 11% of genes are differentially hypomethylated in advanced NAFLD (compared to 3% hypermethylated), leading to increased expression [70]. In advanced NASH specifically, some tissue repair genes were hypomethylted 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 [70].

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 [85]. 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 [33]. Fermenters produce short chain fatty acids, which make up 10% of the calories in a Western diet [67] Some groups claim a link between ethanol-producing gut bacteria and NAFLD [114] [44], however the evidence was inconclusive since no multiple test correction was performed.

3.1.2 Data

Applying next generation 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

Generally 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 taxa or gene functions 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 [73].

Researchers should include multiple test corrections to ensure that the results they are reporting are true, at the expense of having p-values less than 0.05. Unfortunately many studies have been published in high impact journals without multiple test corrections, including a famous paper linking the gut microbiome to autism published in Cell [41].

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. 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.

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:

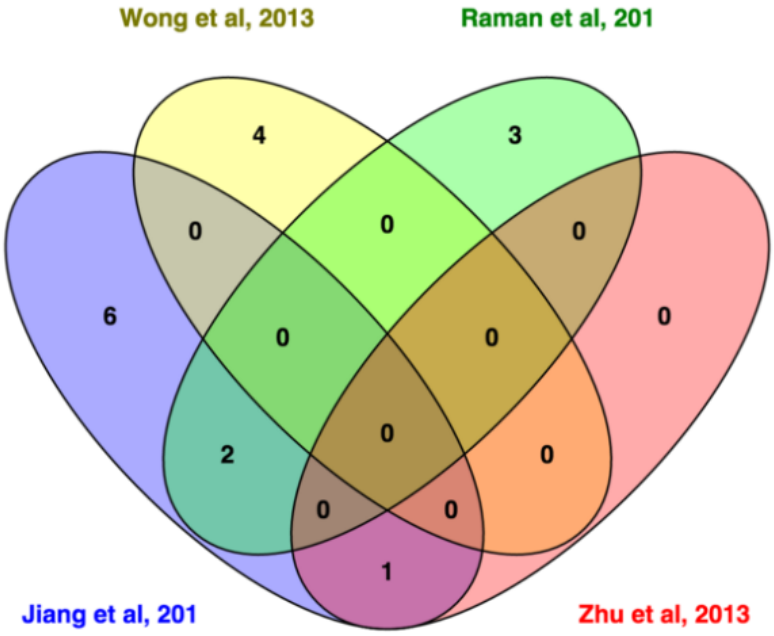

Figure 3.1: **Venn diagram of genera found to be differentially abundant by different studies between NASH/NAFLD and healthy controls.** Only 3 out of the 16 genera claimed to be differentially abundant were found in two studies: members of the *Escherichia* genus were found in the Zhu [114] and Jiang [44] studies, and members of the *Lactobacillus* and *Oscillibacter* genus were found in the Jiang [44] and Raman [85] studies.

Jiang et al, 2015 [44]
Study: This group compared 53 NAFLD patients with 32 healthy controls. The NAFLD patients had a significantly higher BMI ($P < 0.01$).

Sequencing: 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.
Analysis: The reads were annotated with the Ribosomal Database Project [17] and differential abundance was determined using Projection on Latent Structures - Discriminant Analysis (PLS-DA) methods.
Results: 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 [114]
Study: This group 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.
Sequencing: A 16S rRNA gene tag sequencing experiment was performed and reads were sequenced in a 454 pyrosequencer.
Analysis: This group used MG-RAST [68] and QIIME [12].
Results: Note that in the PCoA plot, there is only 11% variance explained by the first component, and they had to plot the first component with the 3rd component (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 [85]
Study: This group 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.
Sequencing: The 16S rRNA gene was amplified and sequenced with 454 pyrosequencing, yielding 2000 reads per sample.
Analysis: Reads were annotated with the Ribosomal Database Project [17]. UniFrac analysis was performed with QIIME [12], and differential abundance was tested with Metastats [77].
Results: 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 [111]
Study: This group compared 16 NASH patients with 22 healthy controls.
Sequencing: They amplified the V1-V2 variable region of the 16S rRNA gene with pyrosequencing, yielding 4-11 thousand reads per sample.
Analysis: Reads were clustered with UCLUST [24] and annotated with the Ribosomal Database Project [17].
Results: 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]
Study: This group compared 30 patients with F0 or F1 fibrosis to 27 patients with F2 or greater fibrosis, 35 of which had NASH
Sequencing: A gene tag experiment was performed on the V4 region of the 16S rRNA gene, and sequenced on an Illumina platform, yeilding an average of 0.2 million reads per sample.
Analysis: Reads were annotated with the Greengenes database [20], and differential abundance was measured by Mann-Whitney's test. A metagenomic imputation was performed with PiCrust [51], annotated with KEGG [46], and analysed with LEfSE [91].
Results: 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 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. 3.1.3 shows a Venn diagram illustrating the inconsistency of the literature on the gut microbiome and NAFLD. Of these, only Raman et al [85] 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 rigorously, such that our results are replicable. 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 **??**, 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 Celcius, 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 [35]. 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 [65], clustered into Operational Taxonomic Units (OTUs) using UCLUST [24], and annotated with the SILVA database [84] using mothur [90], producing a table of counts per operational taxonomic unit per sample. Twelve milion (48%) of

the reads were succesfully overlapped and annotated into 232 OTUs. Differential abundance was analyzed using ALDEx2 [28].

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

The MetaPhlAn tutorial (`https://bitbucket.org/nsegata/metaphlan/wiki/MetaPhlAn_Pipelines_Tutorial`) was followed, using an additoinal marker$_{c}ountsoptioninthemerge_{m}etaphlan_{t}ables.pystep$.

3.2.3 Metagenomic experiment

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. 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 to separate the reads for each 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 [55] to decrease the number of ORFs in the library from 3,495,887 to 2,256,844. Annotation was performed with the SEED database [74], 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. 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 [39], producing 8,847,816 sequences, and removed sequences that matched our reference library with 90% identity as

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.

determined by BLAST [2], leaving 5,876,423 sequences. 3,571,905 of these assembled sequences were successfully annotated with the SEED database [74], and sequenced reads were mapped onto this. [FILL IN THIS] additional reads were annotated by this method, over [FILL IN THIS] unique SEED hierarchy annotations. The code for the custom assembly pipeline is on GitHub. The data from both prongs was amalgamated into a single table of counts per annotation per sample.

Differential abudance was analyzed using ALDEx2 [28]. A full description of the workflow for this process is included in Appendix A.

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.

3.3 Results

3.3.1 16S rRNA gene tag experiment

The top four genus detected by 16S rRNA gene sequencing (excluding unclassified bacteria) were: *Bacteroides*, *Faecalibacterium*, *Blautia*, and *Pseudobutyrivibrio* (Fig. 3.3.1).

No obvious structure or separation is evident from the principal components analysis in Fig. 3.3.1 or the principal coordinate analysis in Fig. 3.3.1. Furthermore the variance explained by each principal component axis is not notably high, indicating a rather uniform data set. Additionally, no OTUs are significantly differentially abundant between groups (Fig. 3.3.1)

When comparing all the healthy samples with all the NASH samples, the genus 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 genus 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.

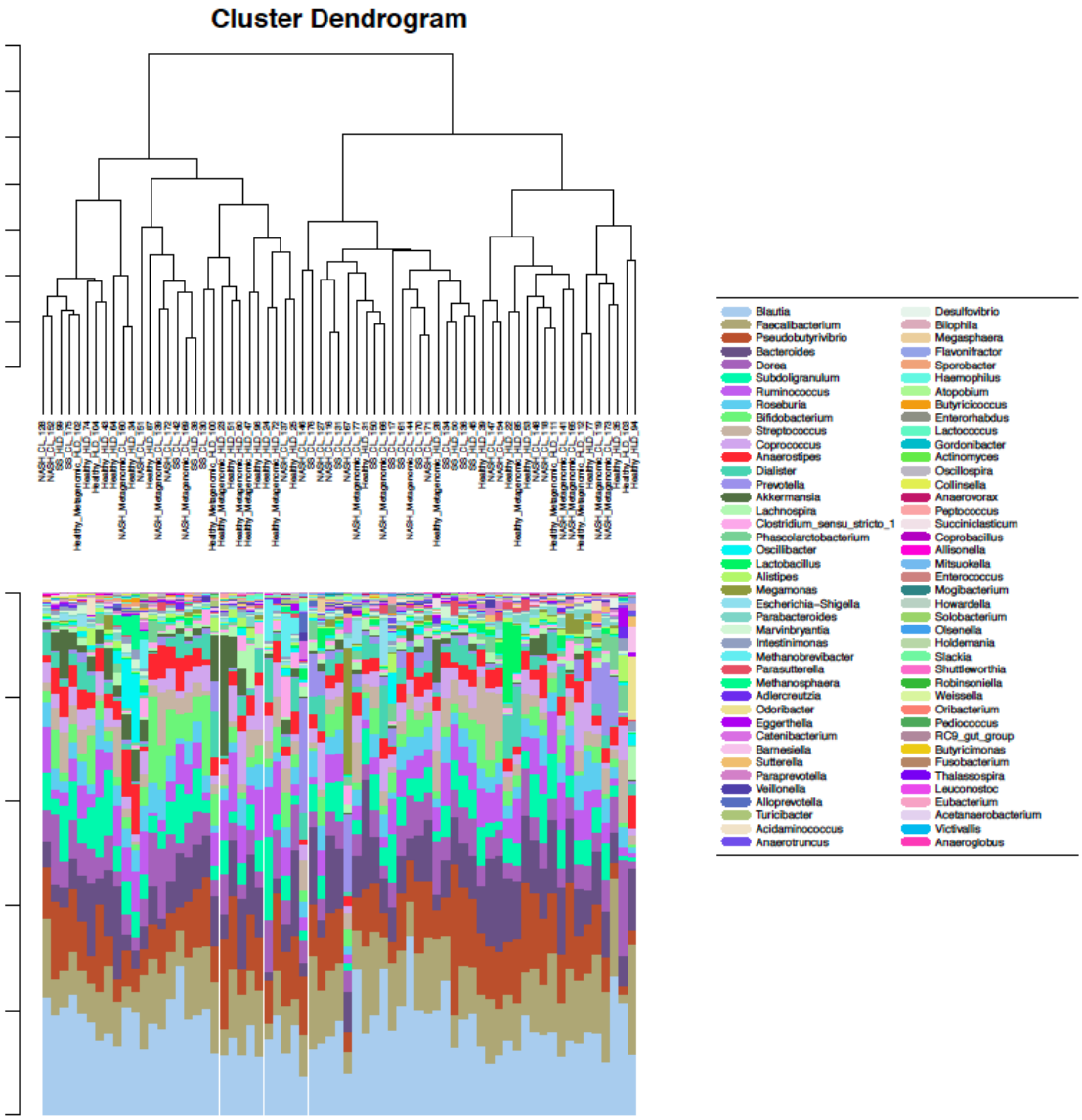

Figure 3.2: **Bar plot of 16S rRNA gene tag sequencing experiment.** Each column of this bar plot represents one sample, and each color represents one bacterial genus. Genus are listed in the legend in order of decreasing total abundance across all samples. Samples do not cluster according to their condition (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.

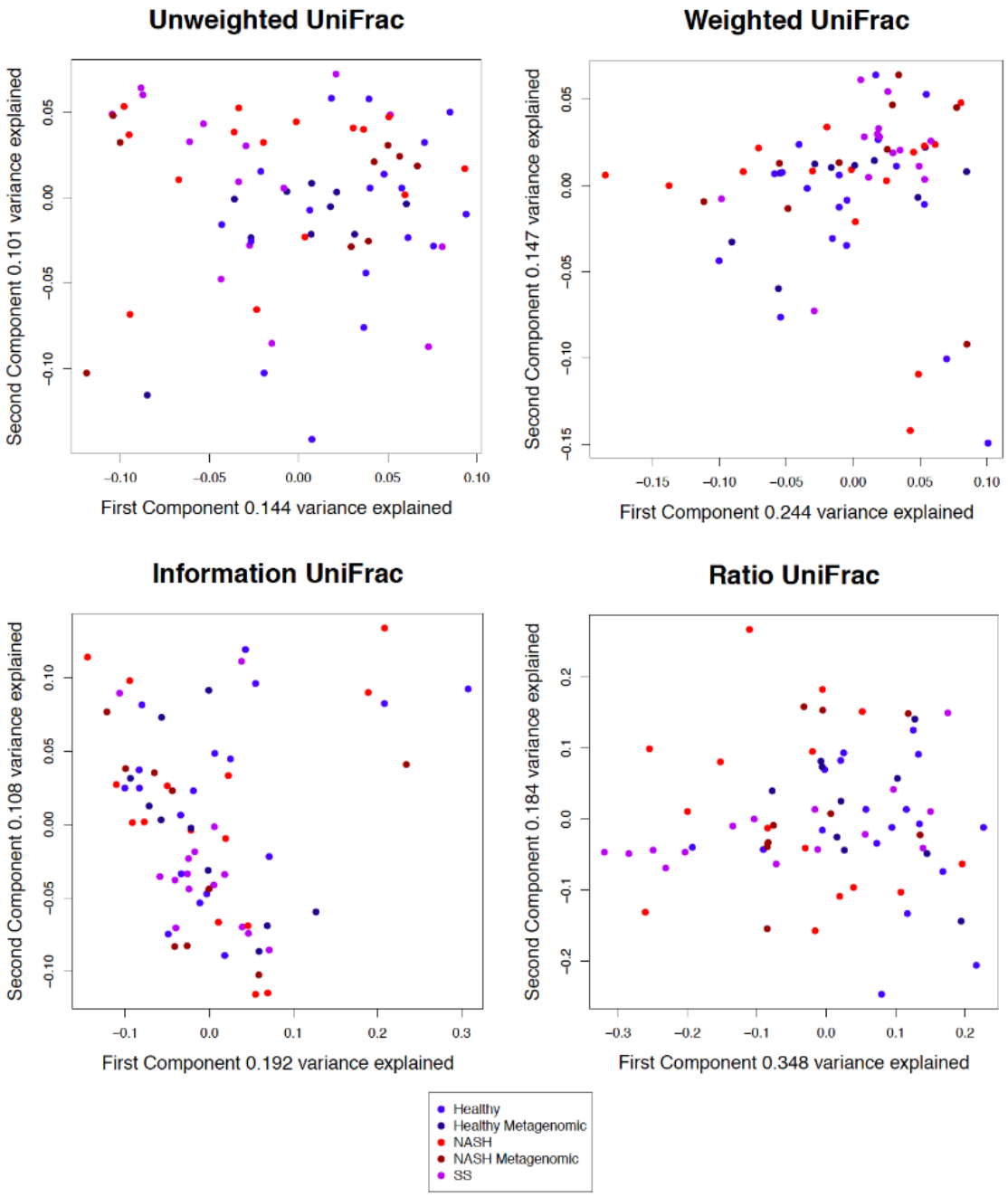

Figure 3.3: **Principal Components Analysis of 16S rRNA gene tag sequencing data with different UniFrac weightings.** Each point represents one sample, and the distances between the samples have been calculated using different UniFrac metrics, taking into account phylogenetic as well as abundance information.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

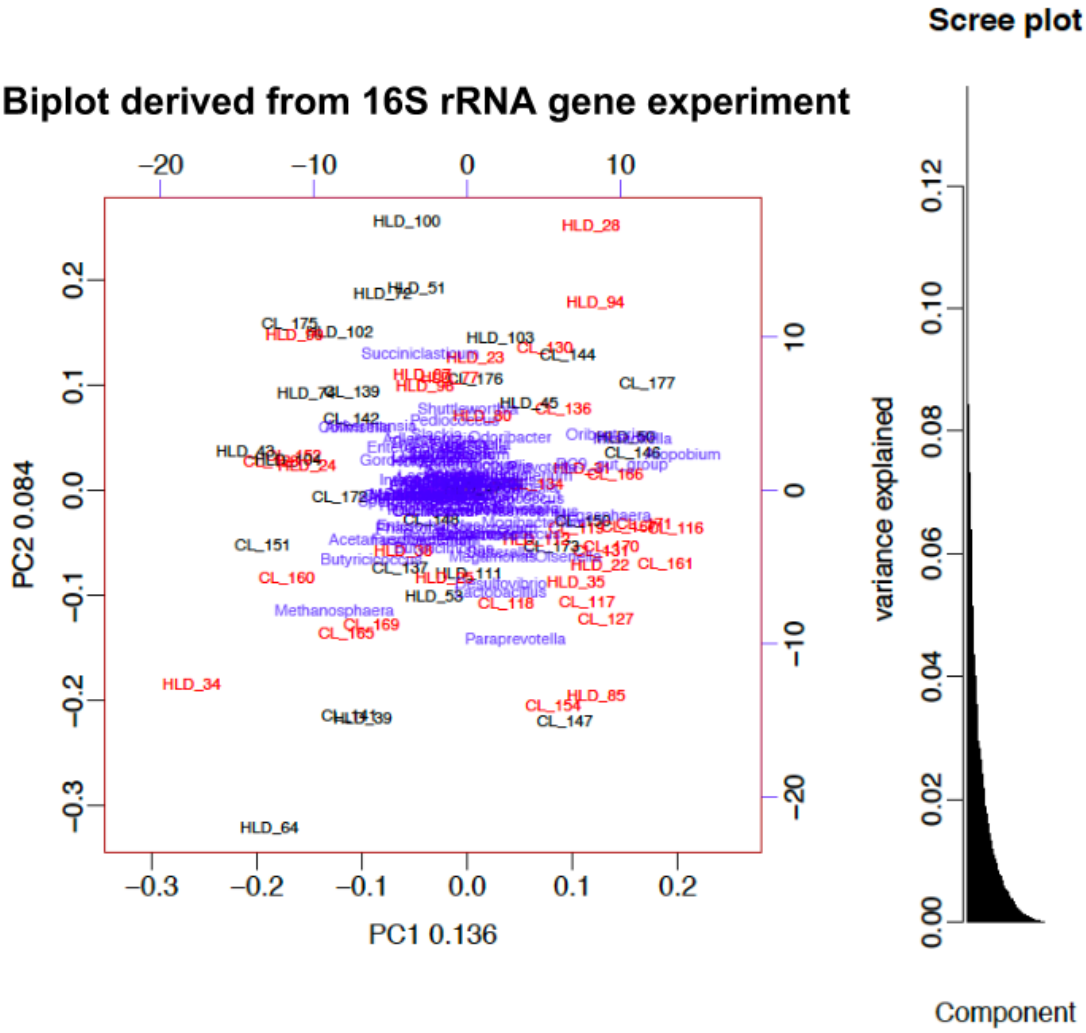

Figure 3.4: **16S rRNA gene tag sequencing experiment biplot.** Compositional data analysis is done by transforming the counts with a centered log ratio transform, and then performing a principal coordinate analysis. The variance explained by each genus is overlayed on the same principal coordinate analysis plot. Note that the variance explained by the first and the second coordinate is 9% and 8% 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.

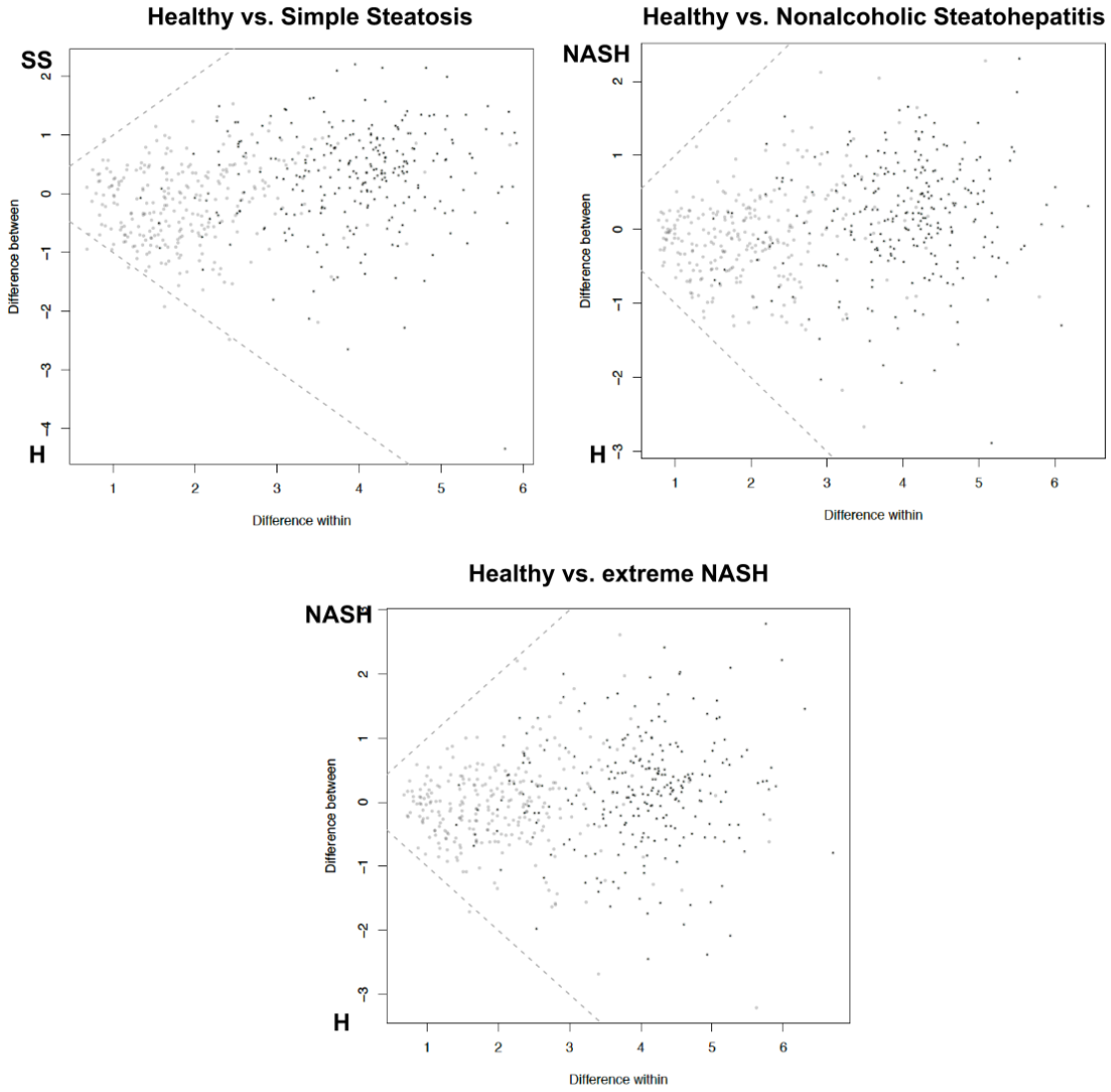

Figure 3.5: **Difference within vs. difference between groups.** Each point represents one OTU, and the differential abundance of that OTU within groups is plotted against the differential abundance between groups. None of the OTUs are more different between groups than within groups. The healthy samples used for these comparisons are the 10 healthy samples used for the metagenomic study. The extreme NASH samples used for these comparisons are the subset of the NASH patients selected for the metegenomic study.

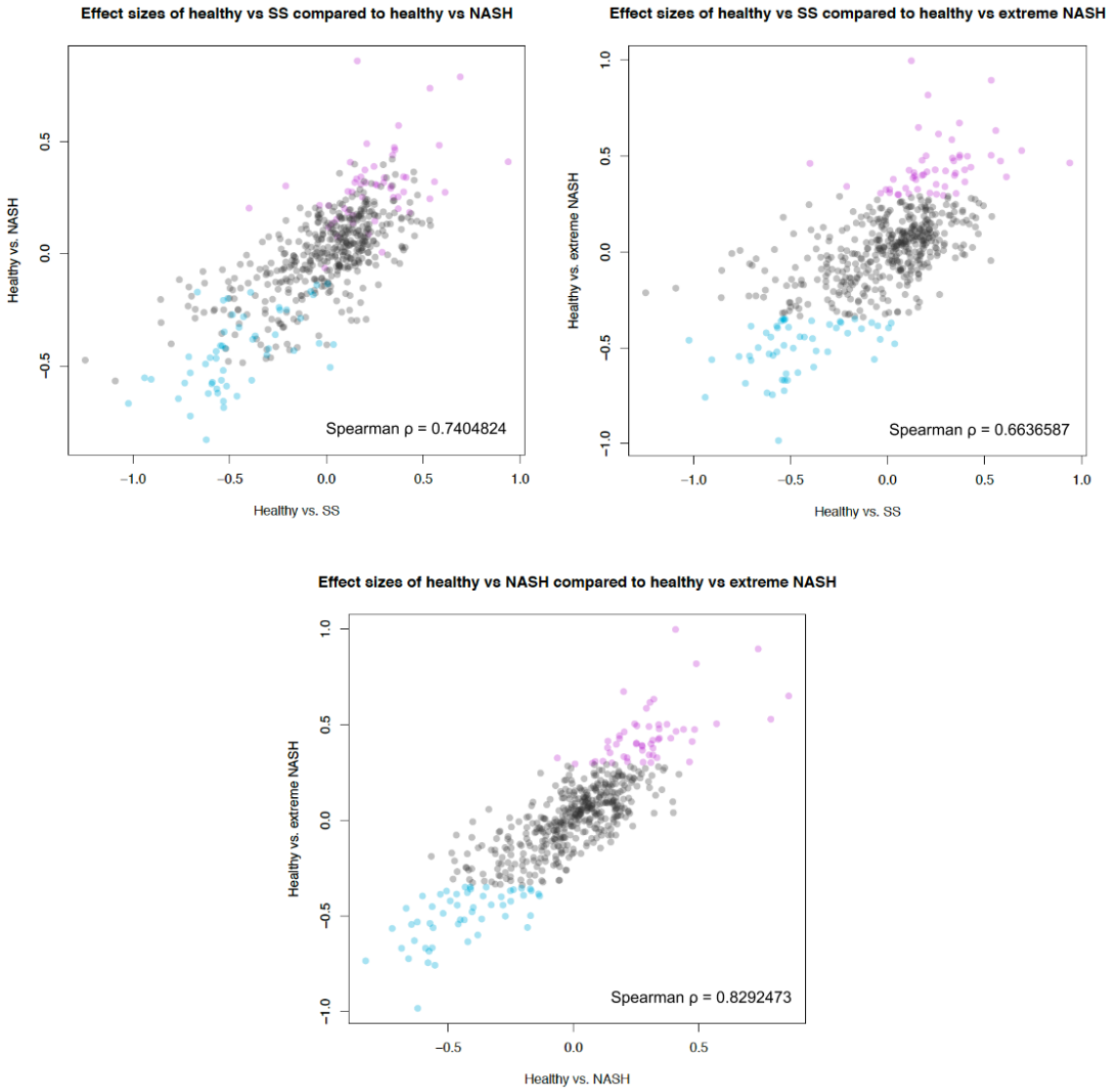

Figure 3.6: **Correlation in effect sizes of different group experiments.** Each point represents one OTU, and the effect size of that OTU in one comparison (for example, comparing the gut microbiome of healthy patients with patients who have simple steatosis) is plotted against the effect size of that 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 median difference in the absolute effect sizes is -0.005547 for Healthy vs. NASH - Healthy vs. SS, 0.01094 for Healthy vs. extreme NASH - Healthy vs. SS, and 0.03105 for Healthy vs. extreme NASH - Healthy vs. NASH. 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.

A differential expression analysis performed with ALDEx2 between healthy vs. SS, healthy vs. NASH, and the 10 healthy samples selected for the metagenomic study vs. the 10 NASH samples for the metagenomic study yielded no significantly differentially abundant OTUs (Fig. 3.3.1). However, the effect size (difference between groups divided by the difference within groups) of each OTU in each comparison is correlated (Fig. 3.3.1). The effect sizes are higher in the Healthy vs. extreme NASH compared to the Healthy vs. SS or Healthy vs. NASH comparison.

OTU family	OTU genus	SILVA bootstrap value	H Vs. NASH metagenomic study effect sizes	H Vs. SS effect sizes	H vs. NASH effect sizes	16S genus effect sizes	MetaPhlAn effect sizes
Acidaminococcaceae	Phascolarctobacterium	100	0.998	0.122	0.407	0.827	0.081
Lactobacillaceae	Lactobacillus	97	0.896	0.534	0.736	0.587	−0.899
Prevotellaceae	Paraprevotella	100	0.819	0.208	0.489	0.843	0.064
Lachnospiraceae	Incertae Sedis	98	0.673	0.37	0.2	NA	NA
Lachnospiraceae	Marvinbryantia	77	0.65	0.159	0.858	0.077	0.269
Lachnospiraceae	Incertae Sedis	73	0.634	0.557	0.32	NA	NA
Bifidobacteriaceae	Bifidobacterium	100	0.616	0.262	0.304	0.188	0.032
Ruminococcaceae	Incertae Sedis	72	0.586	0.331	0.291	NA	NA
Prevotellaceae	Paraprevotella	100	0.529	0.691	0.787	0.843	0.064
Lachnospiraceae	unclassified	100	0.505	0.371	0.571	NA	NA
unclassified	unclassified	72	0.505	0.533	0.244	NA	NA
Ruminococcaceae	Butyricicoccus	71	0.502	0.198	0.372	0.449	0.28
Lachnospiraceae	Incertae Sedis	91	0.5	0.411	0.339	NA	NA
Ruminococcaceae	Ruminococcus	93	0.494	0.369	0.253	−0.866	0.023
Coriobacteriaceae	unclassified	97	0.491	0.334	0.301	NA	NA
Lachnospiraceae	unclassified	98	0.478	0.178	0.341	NA	NA
Lactobacillaceae	Lactobacillus	98	0.476	0.341	0.439	0.587	−0.899
Ruminococcaceae	Subdoligranulum	87	0.475	0.582	0.483	−0.087	−0.177
Ruminococcaceae	Incertae Sedis	98	0.465	0.939	0.409	NA	NA
Ruminococcaceae	unclassified	100	0.462	−0.4	0.202	NA	NA
Coriobacteriaceae	Olsenella	91	0.443	0.429	0.183	0.318	0.141
Ruminococcaceae	Subdoligranulum	98	0.429	0.245	0.388	−0.087	−0.177
Lachnospiraceae	unclassified	100	0.429	0.397	0.342	NA	NA
Prevotellaceae	Prevotella	99	0.427	0.111	0.183	0.212	0.188
Lachnospiraceae	Anaerostipes	100	0.423	0.298	0.338	0.244	0.344
Prevotellaceae	unclassified	70	0.419	0.203	0.316	NA	NA
unclassified	unclassified	92	0.414	0.136	0.137	NA	NA
Ruminococcaceae	Incertae Sedis	99	0.412	0.35	0.473	NA	NA
Ruminococcaceae	Faecalibacterium	100	0.404	0.185	0.251	−0.226	−0.173
Alcaligenaceae	Sutterella	100	0.4	0.177	0.309	0.162	0.051
unclassified	unclassified	73	0.4	0.345	0.25	NA	NA
Rikenellaceae	Alistipes	100	0.397	0.139	0.17	−0.687	0.053
Lachnospiraceae	Roseburia	98	0.392	0.612	0.273	0.18	0.168
Prevotellaceae	unclassified	75	0.387	0.131	0.274	NA	NA
Coriobacteriaceae	Enterorhabdus	72	0.38	0.029	0.135	0.429	0.224
Ruminococcaceae	Ruminococcus	100	0.378	0.146	0.317	−0.866	0.023
unclassified	unclassified	98	0.366	0.398	0.275	NA	NA
Veillonellaceae	Dialister	100	0.353	0.25	0.145	−0.297	−0.038
Lachnospiraceae	unclassified	100	0.342	−0.211	0.301	NA	NA
Family XIII	Incertae Sedis	100	0.341	0.295	0.317	NA	NA
Bacteroidaceae	Bacteroides	100	0.333	0.093	0.201	−0.356	−0.124
Lachnospiraceae	unclassified	100	0.327	0.155	0.333	NA	NA
Lachnospiraceae	unclassified	100	0.327	0.01	0.214	NA	NA
Desulfovibrionaceae	Desulfovibrio	100	0.326	−0.009	−0.065	0.283	0.046
Lachnospiraceae	unclassified	100	0.309	0.011	0.117	NA	NA
Lachnospiraceae	Blautia	96	0.308	0.218	0.087	−0.031	0.192
Lachnospiraceae	Blautia	97	0.306	−0.036	0.215	−0.031	0.192
Ruminococcaceae	unclassified	100	0.305	0.353	0.462	NA	NA
Christensenellaceae	unclassified	99	0.303	0.11	0.277	NA	NA
Alcaligenaceae	Parasutterella	100	0.302	0.252	0.308	0.389	0.521
Lachnospiraceae	Incertae Sedis	100	0.3	0.052	0.153	NA	NA
Ruminococcaceae	Subdoligranulum	92	0.299	0.056	0.076	−0.087	−0.177
Acidaminococcaceae	Acidaminococcus	100	0.295	0.287	0.006	0.345	−0.737

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, their effect sizes in all the comparisons, as well as the corresponding genus-level effect sizes in the 16S rRNA gene tag experiment and MetaPhlAn comparisons between the 10 healthy and 10 NASH samples selected for the metagenomic study. The OTUs were picked by open reference, by clustering and comparison with the SILVA database [84]. Positive effect sizes indicate that the feature was found to be relatively increased in NASH while negative effect sizes indicate that the featuer 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. *Incertae Sedis* and unclassified genera were not analyzed at the genus level.

OTU family	OTU genus	SILVA bootstrap value	H Vs. NASH metagenomic study effect sizes	H Vs. SS effect sizes	H vs. NASH effect sizes	16S genus effect sizes	MetaPhlAn effect sizes
Ruminococcaceae	Incertae Sedis	100	−0.349	−0.539	−0.411	NA	NA
Verrucomicrobiaceae	Akkermansia	100	−0.35	−0.169	−0.433	−0.496 0.343	
Porphyromonadaceae	Parabacteroides	100	−0.35	−0.529	−0.349	−0.313 0.016	
Rikenellaceae	Alistipes	100	−0.356	−0.534	−0.208	−0.687 0.053	
Lachnospiraceae	Incertae Sedis	73	−0.36	−0.392	−0.173	NA	NA
Lachnospiraceae	unclassified	100	−0.361	−0.549	−0.411	NA	NA
Streptococcaceae	Streptococcus	100	−0.363	−0.245	−0.24	−0.18 0.233	
Lachnospiraceae	unclassified	100	−0.369	−0.082	−0.169	NA	NA
Lachnospiraceae	Dorea	100	−0.369	−0.241	−0.252	−0.267 −0.154	
Lachnospiraceae	Roseburia	83	−0.371	0.019	−0.507	0.18 0.168	
Lachnospiraceae	Incertae Sedis	82	−0.379	−0.3	−0.424	NA	NA
Christensenellaceae	unclassified	98	−0.386	−0.051	−0.14	NA	NA
Christensenellaceae	unclassified	100	−0.387	−0.571	−0.467	NA	NA
Lachnospiraceae	Blautia	93	−0.387	−0.704	−0.532	−0.031 0.192	
Ruminococcaceae	unclassified	100	−0.393	−0.511	−0.2	NA	NA
Lachnospiraceae	Roseburia	91	−0.397	−0.568	−0.603	0.18 0.168	
Porphyromonadaceae	Odoribacter	100	−0.397	0.005	−0.135	−0.541 −0.333	
Lachnospiraceae	Incertae Sedis	85	−0.397	−0.265	−0.362	NA	NA
Lachnospiraceae	unclassified	98	−0.401	−0.135	−0.289	NA	NA
Porphyromonadaceae	Odoribacter	100	−0.422	−0.625	−0.492	−0.541 −0.333	
Erysipelotrichaceae	Turicibacter	100	−0.424	−0.206	−0.251	−0.52 −0.717	
Christensenellaceae	unclassified	100	−0.443	−0.452	−0.329	NA	NA
Ruminococcaceae	Ruminococcus	100	−0.444	−0.429	−0.282	−0.866 0.023	
Christensenellaceae	unclassified	99	−0.445	−0.602	−0.464	NA	NA
Lachnospiraceae	unclassified	100	−0.452	−0.387	−0.564	NA	NA
Bacteroidaceae	Bacteroides	100	−0.456	−0.038	−0.4	−0.356 −0.124	
Ruminococcaceae	Incertae Sedis	84	−0.461	−1.024	−0.668	NA	NA
Veillonellaceae	Dialister	100	−0.479	0.036	−0.405	−0.297 −0.038	
Bacteroidaceae	Bacteroides	100	−0.488	−0.534	−0.521	−0.356 −0.124	
Prevotellaceae	Alloprevotella	100	−0.5	−0.667	−0.172	−0.001 0.863	
Bacteroidaceae	Bacteroides	100	−0.502	−0.487	−0.272	−0.356 −0.124	
Rikenellaceae	Alistipes	100	−0.517	−0.369	−0.367	−0.687 0.053	
Coriobacteriaceae	Adlercreutzia	100	−0.521	−0.309	−0.452	−0.372 −0.298	
Ruminococcaceae	unclassified	100	−0.522	−0.571	−0.436	NA	NA
Family XIII	Anaerovorax	91	−0.533	−0.611	−0.624	−0.476 −0.152	
Ruminococcaceae	unclassified	76	−0.54	−0.59	−0.573	NA	NA
Lachnospiraceae	Pseudobutyrivibrio	98	−0.543	−0.712	−0.46	−0.423 −0.091	
Bacteroidaceae	Bacteroides	100	−0.546	−0.766	−0.647	−0.356 −0.124	
Bacteroidaceae	Bacteroides	100	−0.561	−0.069	−0.184	−0.356 −0.124	
Lachnospiraceae	Blautia	85	−0.562	−0.906	−0.561	−0.031 0.192	
Ruminococcaceae	Faecalibacterium	100	−0.566	−0.704	−0.724	−0.226 −0.173	
unclassified	unclassified	85	−0.601	−0.381	−0.382	NA	NA
Ruminococcaceae	Incertae Sedis	100	−0.63	−0.463	−0.635	NA	NA
Ruminococcaceae	Subdoligranulum	99	−0.636	−0.523	−0.422	−0.087 −0.177	
Ruminococcaceae	Incertae Sedis	97	−0.668	−0.544	−0.565	NA	NA
Ruminococcaceae	Ruminococcus	100	−0.669	−0.516	−0.591	−0.866 0.023	
Lachnospiraceae	Incertae Sedis	85	−0.67	−0.532	−0.686	NA	NA
Lachnospiraceae	Coprococcus	91	−0.686	−0.733	−0.577	−0.647 0.469	
Ruminococcaceae	unclassified	74	−0.725	−0.533	−0.658	NA	NA
Erysipelotrichaceae	unclassified	100	−0.736	−0.621	−0.83	NA	NA
Ruminococcaceae	Subdoligranulum	85	−0.746	−0.594	−0.581	−0.087 −0.177	
Lachnospiraceae	Dorea	100	−0.759	−0.94	−0.554	−0.267 −0.154	
Family XIII	Incertae Sedis	91	−0.985	−0.563	−0.622	NA	NA

Table 3.4: **Bottom decile of OTUs relatively increased in healthy based on effect size from healthy vs. NASH comparison.** This table lists the OTUs, their effect sizes in all the comparisons, as well as the corresponding genus-level effect sizes in the 16S rRNA gene tag experiment and MetaPhlAn comparisons between the 10 healthy and 10 NASH samples selected for the metagenomic study. The OTUs were picked by open reference, by clustering and comparison with the SILVA database [84]. Positive effect sizes indicate that the feature was found to be relatively increased in NASH while negative effect sizes indicate that the featuer 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. *Incertae Sedis* and unclassified genera were not analyzed at the genus level.

3.3.2 Metagenomic experiment

Functional analysis

MetaPhlAn

We ran the metagenomic sequences through MetaPhlAn to infer what the results would be with 16S rRNA gene tag sequencing, so that we could compare with our empirical 16S rRNA gene tag sequencing results. The effect size Spearman coefficient (Fig. 3.3.2) is smaller than the effect size coefficient between the healthy vs. SS and healthy vs. NASH comparison, even though in this case the same samples are being compared.

The operational taxonomic units in the MetaPhlAn and 16S rRNA gene analysis were derived from different databases, and could not be compared directly, so the effect size comparison in Fig. 3.3.2 was done at the genus level. Note that OTUs can reside in between genuses, such that the genus classificaiton 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.

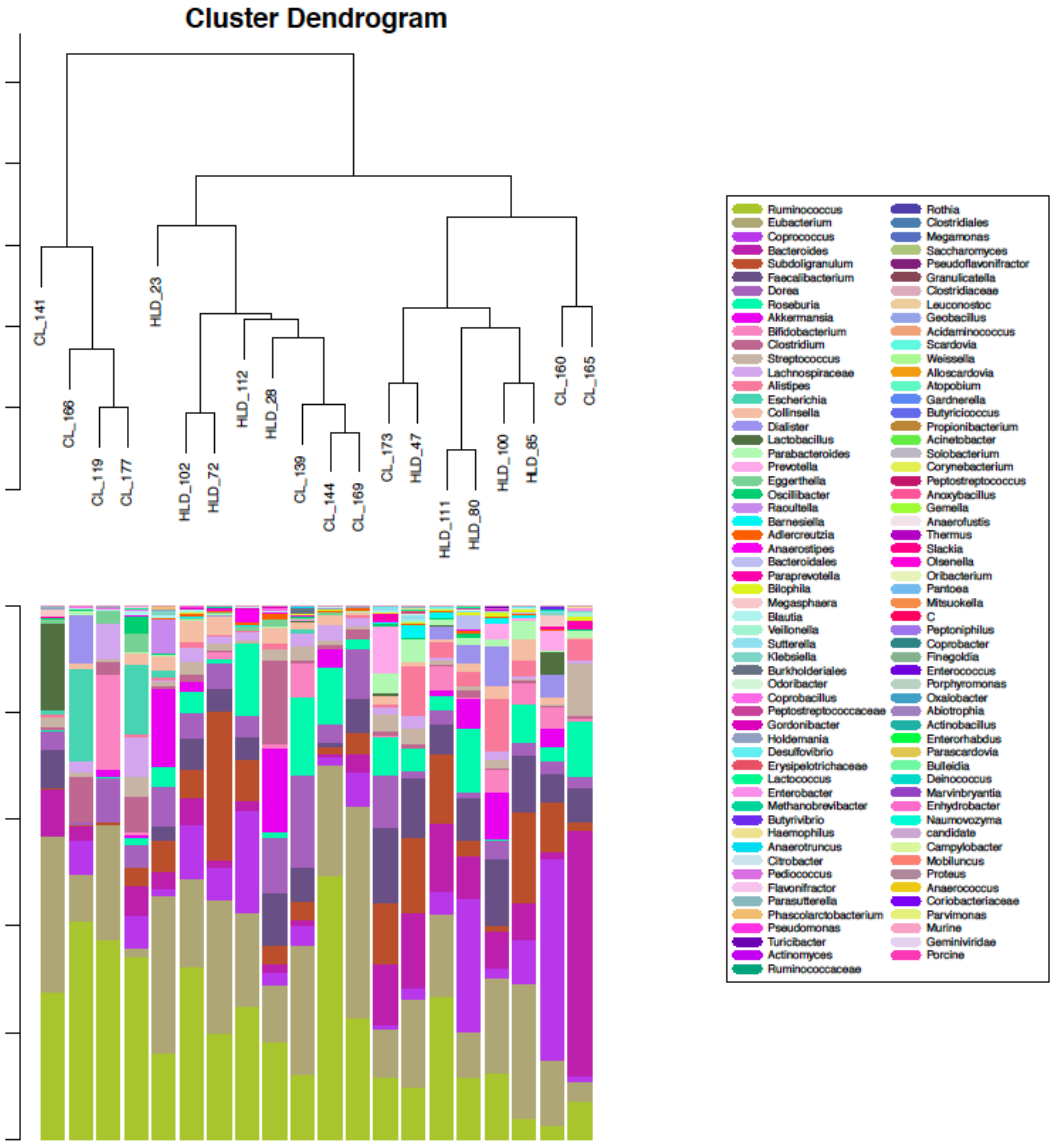

Figure 3.7: **Taxa barplot dendogram derived from MetaPhlAn.** The metagenomic reads were input into MetaPhlAn to generate a count table. The taxa in the count table were filtered such that only taxa with at least 1% abundance in any sample was kept. In this barplot, each bar represents one sample, and each color represents one genus, with the size of the colored segments corresponding with the relative abundance of the genus.

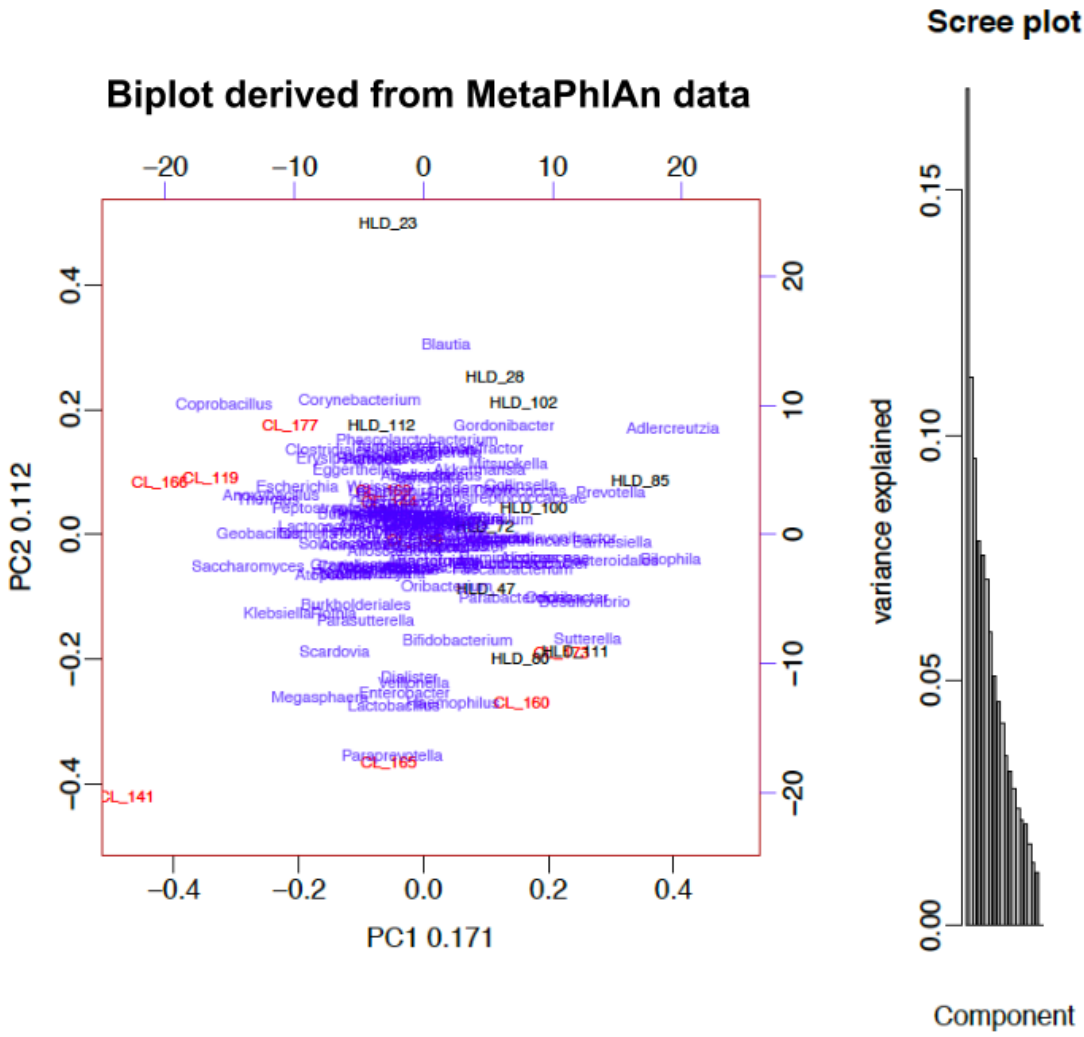

Figure 3.8: **Biplot derived from MetaPhlAn.** Compositional data analysis is done by transforming the counts with a centered log ratio transform, and then performing a principal coordinate analysis. The variance explained by each genera is overlayed on the same principal coordinate analysis plot. This biplot was generated from the count table inferred by MetaPhlAn, with taxa filtered such that only taxa with at least 1% abundance in any sample was kept. Note that the variance explained by the first and the second coordinate is 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.

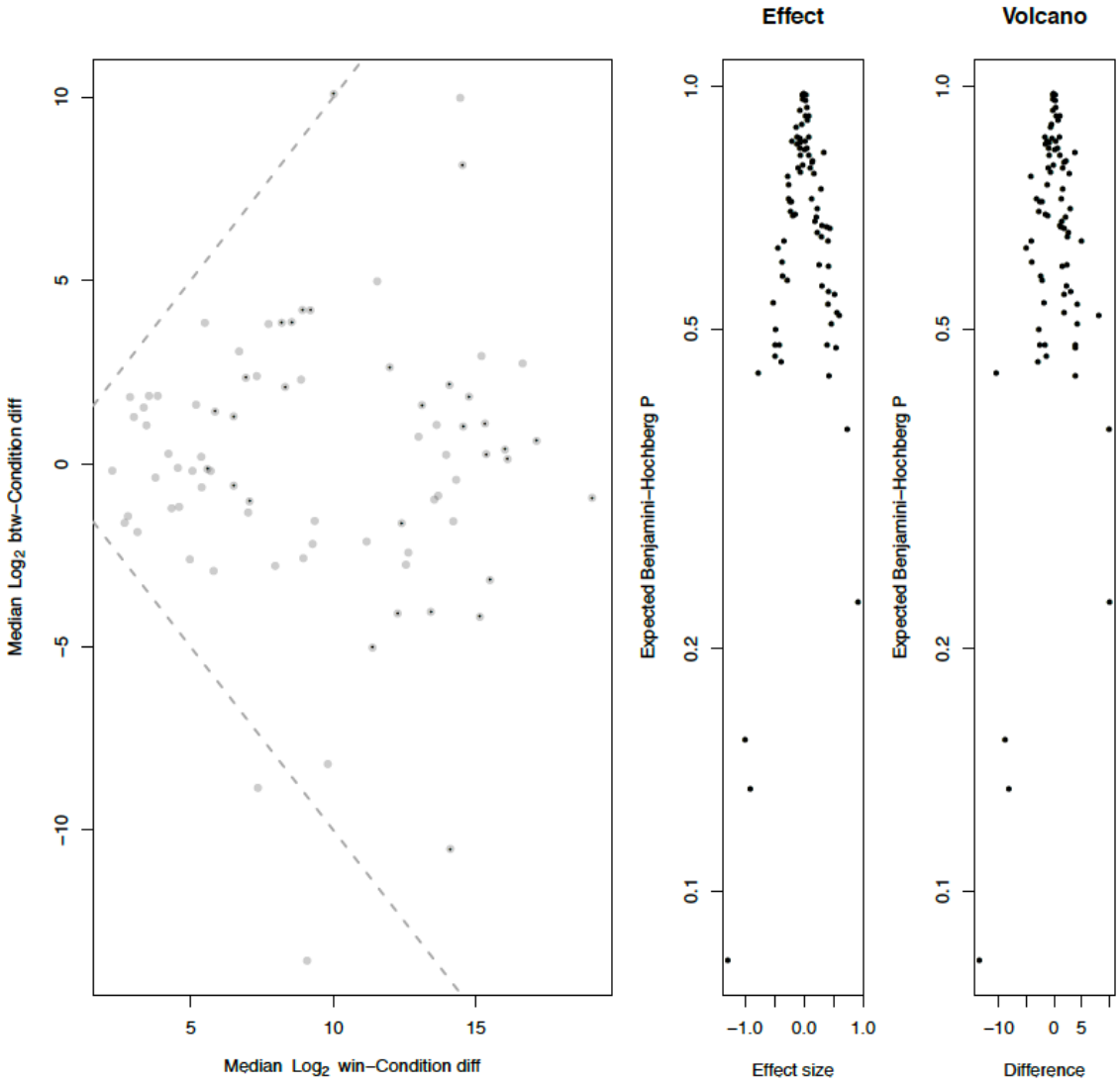

Figure 3.9: **Difference within groups vs. difference between groups per taxa, derived from MetaPhlAn.** This plot was generated from the count table inferred by MetaPhlAn, with taxa filtered such that only taxa with at least 1% abundance in any sample was kept. No taxa are more differential between 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 OTU level.

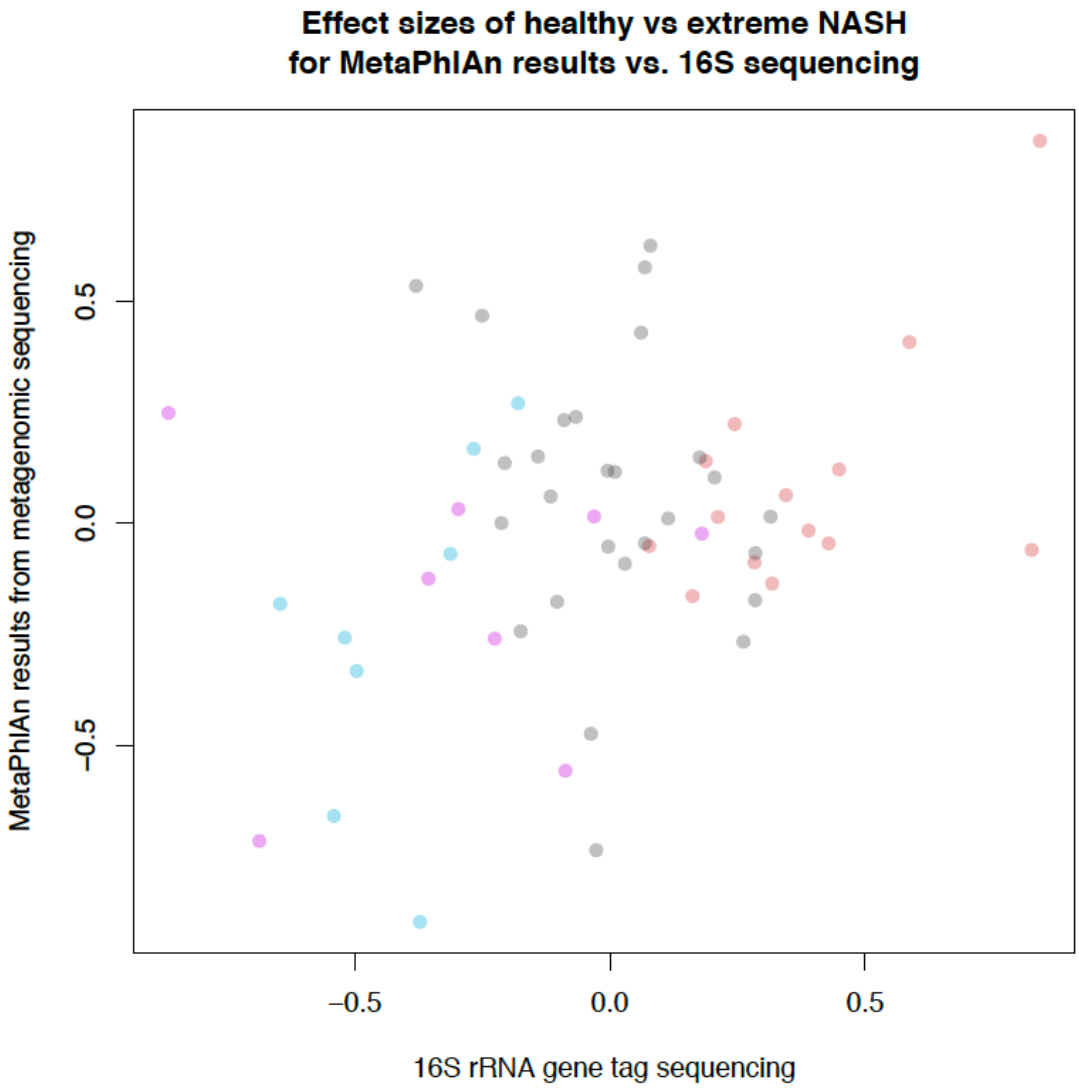

Figure 3.10: **Effect size correlation between MetaPhlAn and 16S rRNA gene tag sequencing.** For this plot, taxa were amalgamated at the genus level. The effect sizes for the comparison of the healthy and NASH samples selected for the metagenomic study are shown. The Spearman coefficient is 0.2021001. Genera corresponding with OTUs identified to have the top deciles of effect sizes in the 16S rRNA gene tag experiment comparison are colored - red for OTUs relatively abundant in NASH, and blue for OTUs relatively abundant in the healthy condition. Genera corresponding to both OTUs in the NASH and healthy deciles are colored purple.

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. 3.3.1) or at the level of the individual OTUs (Fig. 3.3.1).

There are several factors that would make such a study underpowered. First, the gut microbiome is highly diverse between individuals. This is compounded by the fact that the samples were taken from a diverse Toronto population, including people who immigrated from other countries who likely have different diets. The literature shows that differences in the gut microbiome are often driven by diet [18]. 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. 3.3.1, the correlation shows that even though there is not enough power to detect a significant difference, the difference from the healthy baseline are moving in the same direction through simple steatosis to nonalcoholic steatohepatitis to extreme NASH.

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

Chapter 4

Discussion

4.1 Lack of reproducibility

It is clear that more robust statistical analysis is necessary in this field. The chapter on expanding the UnIFrac toolbox highlighted examples of misuse of unweighted UniFrac in papers published in Cell and Nature. One claimed to find differences in the gut microbiome of mice modelling autism spectrum disorder, compared to healthy controls. 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. 3.1.3). 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 [43]. The biomedical sciences suffer similar issues, prompting Nature to publish a collection of statistics for biologists (`http://www.nature.com/collections/qghhqm/`).

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 for principal components analysis or principal coordinates analysis, along with software such as metagenomeSeq [78], DESeq2 [56], and metastats [77] for differential expression analysis. These are commonly accessed through pipelines such as QIIME and mothur. Many of these have roots in ecology, for example, the Shannon diversity index. Shannon diversity is commonly measured in microbiome papers, but does not make sense for complex biological samples where diversity can be increased by performing deeper sequencing to uncover more bacterial taxa.

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* [108], 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 correations include SPARCC [31], SpiecEasi [50] or the phi metric [57]. Recommended software and tools for differential expression analysis include the analysis of composition of microbiomes (ANCOM) [62] and ALDEx2 for differential expression analysis [28].

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 (weak evidence). An estimated 17-25% of such reported results are expected to be wrong, even without p-hacking [45]. 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 analyse their data without looking at the data raw. We recommend visualizing the data in bar graphs (as in Fig. 3.3.1), principal components (as in Fig. 3.3.1), 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 [54], ancestry differences [82], and variations in genotype calling methods [69], and recommendations were made to avoid pooling the sequences together [66], and for using sample sizes in the thousands [9], stratification detection [82], and technical replicates [40], and experimental validation [66]. Patient genomes must be sequenced at 30 times coverage or higher to validate the presence of SNPs [86]

Interestingly, the field of microbiome research seems to have standardized too early. Efforts such as the Human Microbiome Project, set a precident 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 [90] make it comparatively difficult for researchers to explore other analysis options, both in terms of analysing 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 [103], and leading to clinical interventions for diseases such as *C. diff* [80]. 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] Jun Chen et al. “Associating microbiome composition with environmental covariates using generalized UniFrac distances”. In: *Bioinformatics* 28.16 (2012), pp. 2106–2113.

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

[17] 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.

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

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

[20] 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.

[21] 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.

[22] 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.

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

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

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

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

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

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

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

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

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

[36] 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.

[37] 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.

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

[39] 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.

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

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

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

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

[44] 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).

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

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

[47] 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.

[48] 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.

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

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

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

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

[58] 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.

[59] 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.

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

[61] 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.

[62] 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).

[63] 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.

[64] 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.

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

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

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

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

[73] 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.

[74] 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.

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

[76] 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.

[77] 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.

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

[79] Karl Pearson. “Mathematical Contributions to the Theory of Evolution.–On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs”. In: *Proceedings of the royal society of london* 60.359-367 (1896), pp. 489–498.

[80] Elaine O Petrof et al. “Stool substitute transplant therapy for the eradication of Clostridium difficile infection:‘RePOOPulating’the gut”. In: *Microbiome* 1.1 (2013), p. 1.

[81] David Preiss and Naveed Sattar. “Non-alcoholic fatty liver disease: an overview of prevalence, diagnosis, pathogenesis and treatment considerations”. In: *Clinical science* 115.5 (2008), pp. 141–150.

[82] 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.

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

[84] 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.

[85] 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.

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

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

[88] 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.

[89] 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.

[90] 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.

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

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

[93] Ron Sender, Shai Fuchs, and Ron Milo. “Revised estimates for the number of human and bacteria cells in the body”. In: *bioRxiv* (2016), p. 036103.

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

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

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

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

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

[99] Silvia Sookoian and Carlos J Pirola. "Meta-analysis of the influence of I148M variant of patatin-like phospholipase domain containing 3 gene (PNPLA3) on the susceptibility and histological severity of nonalcoholic fatty liver disease". In: *Hepatology* 53.6 (2011), pp. 1883–1894.

[100] 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).

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

[102] 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.

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

[104] 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.

[105] 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.

[106] 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.

[107] 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.

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

[109] 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.

[110] AJ Wigg et al. "The role of small intestinal bacterial overgrowth, intestinal permeability, endotoxaemia, and tumour necrosis factor α in the pathogenesis of non-alcoholic steatohepatitis". In: *Gut* 48.2 (2001), pp. 206–211.

[111] 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.

[112] 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.

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

[114] 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.

[115] 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

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 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$_TYPE$ = $BlastSearchBLAST_SPEC = MicrobialGenomes$).

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.

A.1.3 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 [19] to predict open reading frames.

A.1.4 Annotate reference library coding sequences

Annotation was performed by querying the SEED database [74] 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.

A.1.5 Map seqeunced reads to reference library

Once the sequenced reads are available, they can be mapped to the reference library using Bowtie2 [52]. 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 [28].

All of the custom scripts used to perform the above for the metagenomic non-alcoholic fatty liver disease experiment can be found at `https://github.com/ruthgrace/make_functional_mapping_library`.

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.