COMPUTATIONAL MICROBIOME ANALYSIS: METHODS AND APPLICATIONS

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

by

Ruth Wong

Graduate Program in Biochemistry

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

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

THE UNIVERSITY OF WESTERN ONTARIO
School of Graduate and Postdoctoral Studies

CERTIFICATE OF EXAMINATION

Supervisor:

.....
Dr. Gregory B. Gloor

Supervisory Committee:

.....
Dr. Lindi M. Wahl

.....
Dr. David R. Edgell

Examiners:

.....
Dr. ExaminerA

.....
Dr. ExaminerB

.....
Dr. ExaminerC

The thesis by

Ruth Grace Wong

entitled:

Computational microbiome analysis: methods and applications

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

.....
Date

.....
Chair of the Thesis Examination Board

Abstract

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

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

Contents

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

1.1 The human microbiome

Approximately half of the cells that make up the human body are bacterial (Sender, 2016). Trillions of these bacteria live in the gut (Guarner, 2003), and have a massive metabolic potential. For example, the gut microbiome has been shown to produce changes in hormone levels (Markle, 2013), short chain fatty acid levels (Turnbaugh, 2008), and ethanol levels (Krebs, 1970), to name a few. The human gut microbiome can even digest polysaccharides otherwise unusable by humans (Flint, 2008).

This massive metabolic potential produces measurable symptomatic effects. Transplanting gut bacteria from obese mice to lean mice have been shown to convert lean mice to absorb more calories from the same food (Turnbaugh, 2006). The microbiome can also affect behavior: Completely germ free mice exhibit more anxiety-like behaviors than specific pathogen free mice (Neufeld, 2011).

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

With the advent of next generation sequencing, scientists can obtain a more comprehensive snapshot of the bacterial composition of the microbiome, what genes they have, and what proteins they produce (Di Bella, 2013). We are in a phase of developing the experiments and accompanying statistical techniques to elucidate the exact mechanisms by which the human microbiome affects health and disease. Armed with a deeper understanding of how the microbiome works, we may be able to develop probiotic techniques to improve quality of life.

1.2 Exploring the human microbiome

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

Usually a study involves an experimental group and a control group. These can be patients with disease and healthy controls (Macklaim, 2013), people who are susceptible and resistant to a condition (Theriot, 2014), or patients before and after a medical intervention (Graessler, 2013). The questions that scientists in this field generally want to answer are: Is the human microbiome driving or associated with the difference between the two groups? If so, what is the mechanism of action? There are also exploratory studies which try to determine what the core microbiome for a body site in a single condition is by examining what people who fit the condition have in common.

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

The two metagenomic experiments that can be done with microbiome next generation sequencing data used in this thesis are gene tag abundance and deep metagenomic sequencing (Riesenfeld, 2004). The tag used for gene tag abundance here is the 16S rRNA gene (Gloor, 2010). The process and resulting data of each experiment is described in the next section, followed by a piece about data analysis and points of failure.

1.3 Illumina next generation sequencing platform

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

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

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

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

The Illumina technology has been used for years (Bentley, 2008), and standard protocols exist for library preparation, with kits available commercially.

1.4 Gene tag abundance

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

What bacterial taxa make up the microbial community? Scientists often want to characterize microbiomes for certain conditions. For example, the core gut microbiome was described by one group to have three enterotypes (Arumugam, 2011), however, when another group studied a diverse population including non-Western people, the enterotypes did not hold (Yatsunenko, 2012). The vaginal microbiome is known to be Lactobacillus dominated, except in bacterial vaginosis, where the microbiome is much more diverse (Hummelen, 2010). The idea is that characterizing the core microbiome is can lead to insight on core functions and how they can be altered when the core microbiome is disrupted.

Are there any differentially abundant taxa between conditions? Some theories of disease progression include the involvement of bacteria as pathogens. Others involve bacteria as probiotics, preventing disease progression. Salient examples include atopic dermatitis where flare-ups are associated with an increase in the proportion of Staphylococcus aureus on the skin (Kong, 2012), and RePOOPulate, a probiotic therapy where 33 microbes cultured from a healthy donor were used to successfully treat symptoms of C. difficile (Petrof, 2013).

Historically, Kochs postulates have been used to determine if a microbe is a disease-causing pathogen: First, the microbe must be present in all cases of the disease. Second, the microbe must not be present and non-pathogenic in other diseases. Third, if the microbe is isolated in pure culture, it can be used to induce the disease (Koch, 1890). One group has created a modified set of postulates that takes DNA sequencing into account (Fredericks, 1996), which can be applied to differentially abundant taxa detected by gene tag sequencing. However, Kochs 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 (Macklaim, 2013).

Do samples from different conditions cluster together? Sometimes when the data is plotted, there appears to be separation between groups, even if specific taxa are not differentially abundant. One example of this is a study on discordant gut microbiomes between twins in Malawi where one twin has kwashiorkor and the other is healthy (Smith, 2013). In this case the microbiomes diverge the most during treatment with ready-to-use therapeutic food.

1.4.1 16S rRNA gene sequencing experiment

The gene tag chosen throughout this thesis is the gene for the 16S subunit of ribosomal RNA. The 16S rRNA gene is present in all known bacteria and has regions of variability interspersed with regions of high conservation. This allows primers to be made to match the conserved regions, such that the variable regions can be amplified, sequenced, and used to infer taxa. Entire databases exist specifically to match the 16S rRNA gene with taxonomy, such as SILVA (Quast, 2013), the Ribosomal Database Project (Cole, 2009), and Greengenes (DeSantis, 2006).

Specifically, we have been using the 16S rRNA primers from the Earth Microbiome Project protocol (Gilbert, 2014), 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 (Walters, 2011).

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

However, it is difficult to determine how a batch of sequences should be partitioned into groups of 97

1.4.3 General protocol and rationale

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

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

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

2. Run a PCR amplification As discussed previously, the gene tag experiments in this thesis amplify the V4 region of the 16S rRNA gene, following the Earth Microbiome Project protocol (Caporaso, 2012). The set of primers that we use are barcoded, so that we can sequence all the samples in the same sequencing run and differentiate them afterwards.

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

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

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

2. Assemble the paired ends of sequenced DNA The paired sequences are overlapped in the middle, resulting in the full variable region amplified by the primers.

3. Group the reads into operational taxonomic units (OTUs) We used the mothur software suite to cluster the reads into groups of 97

4. Annotate the OTUs with bacterial taxonomy Annotation was done by matching our OTUs to the SILVA database (Quast, 2013).

Alternatively, an Individual Sequence Unit (ISU) 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 (Callahan, 2015).

1.4.4 Data analysis

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

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

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

UniFrac Principal Component Analysis is necessary for multivariate statistics, and It is well known that the Principal Component Analysis cannot be performed on proportions, such as the OTU abundances derived from gene tag sequencing. Instead, a Euclidean distance is required (Anderson, 2003).

In 2005, Lozupone et al introduced the UniFrac distance metric, a measure to calculate the difference between microbiomes that incorporated phylogenetic distance (Lozupone, 2005). 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 (Lozupone, 2007). Since then, papers reporting these metrics have garnered over a thousand citations, and enabled research about everything from how kwashiorkor causes malnutrition (Smith, 2013) to how people can have similar microbiomes to their pet dogs (Song, 2013). Except for Generalized UniFrac, used to make hybrid unweighted and weighted UniFrac comparisons (Chen, 2012), few advances in the metric have occurred since 2007.

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

The qualitative rather than quantitative nature of unweighted UniFrac makes the metric very sensitive to sequencing depth. A greater sequencing depth generally results in the detection of a greater number of taxa. To account for this problem, ecologists use a technique called rarefaction to normalize the sequencing depth across samples by random sampling without replacement (de Crcer, 2011). However, in unweighted UniFrac samples move relative to the other samples in different rarefaction instances, to the point where they can switch from being a member of one cluster of data to another, as demonstrated in the chapter Expanding the UniFrac Toolbox.

Weighted UniFrac Weighted UniFrac is an implementation of the Kantorovich-Rubinstein distance in mathematics, also known as the earth mover's distance (Evans, 2012). Rather than looking only at the presence or absence of taxa, each branch length of the phylogenetic tree is weighted by the difference in proportional abundance of the taxa between the two samples. This technique reduces the problem of low abundance taxa being represented as a 0 or by a low count depending on sampling depth. In unweighted UniFrac, such taxa would flip from absent to present, and could skew the measurement: this would be especially problematic if the taxa are on a long branch. In weighted UniFrac, low abundance taxa have a much lower weight and so will have a lower impact on the total distance reported by the metric.

UniFrac is constituted as either a presence/absence (unweighted UniFrac) (Lozupone, 2005), a linear proportion in the form of weighted UniFrac (Lozupone, 2007), or some combination of the two in the form of Generalized UniFrac (Chen, 2012). However, the data are not linear, because the sum of the total number of reads is constrained by the sequencing machinery (Friedman, 2012). Alternative weightings and non-linear transformations of data need to be explored.

Principal Components Analysis Once the distances between each pair of samples has been calculated, they can be visualized on a plot, with each sample represented as one point. For visualization, the data should be placed so distances are preserved as much as possible, so that clustering and separation of samples can be clearly seen. This is done using the Principal Coordinate Analysis method of multidimensional scaling (Dollhopf, 2001), shortened as PCoA.

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

of separation or clustering.

After multidimensional scaling the data can be analyzed in several ways. The data can be examined for clustering by k-means analysis (Tibshirani, 2005). The points can also be measured for separation by looking only at their position on the first principal component axis, especially if the first axis covers the majority of the variation in the data set. With each sample associated with a number on the first principal component axis, one can examine the effect size of two different groups by taking the mean positions and dividing by the standard deviation.

1.5 The metagenomic experiment

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

What is the metabolic potential of the microbial community? The metabolic potential is made up of all the protein functions that are coded by the genetic material present in the sample. Biologically speaking, these protein functions represent the enzymatic reactions that the microbiome could produce if all the genes were expressed. For example, the human gut microbiome has more genes related to methanogenesis, compared to the average sequenced microbe (Gill, 2006).

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 (Turnbaugh, 2006), sparking more research into the gut microbiome and obesity related ailments such as diabetes (Larsen, 2010) and non-alcoholic fatty liver disease (Zhu, 2013). The ability to check if genes, functional categories of genes, or pathways are differentially abundant between groups allows scientists to find clues about the mechanisms by which the microbiome affects certain diseases.

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

1.5.1 Sequencing

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

1. Take a biological sample and perform DNA extraction The sample can be collected by swabbing the target body site or collecting excretions.

2. Prepare the DNA for sequencing Fragment the DNA, and filter for the desired size. These steps are all part of the standard Illumina library prep protocol for the HiSeq. There are

two options for fragment size, either 50 or 100 nucleotides in length, and we chose the longer one for easier assembly and mapping.

3. Sequence the DNA. We performed single end sequencing on the Illumina HiSeq platform, with our samples barcoded so that they could be pooled into the same sequencing run.

4. Create an annotated library of reference sequences The annotated library contains annotations about what kind of protein each sequence codes for. The first step to creating the annotated library is to gather a database of sequences. The database of sequences can be created before the sequencing is complete by gathering all the genomes of all the bacterial strains predicted to be present in the sample, or it can be created after sequencing by assembling the sequenced reads into parts of genomes. The second step is to annotate the sequences with predicted protein functions. Some publically available genomes already have protein annotations. For genomes or partial genomes without annotations, the placement of genes can be predicted by looking for open reading frames, and these predicted genes can be aligned with databases such as SEED (Overbeek, 2005) or KEGG (Kanehisa, 2000) to match them with functional annotations, using the BLAST algorithm (Altschul, 1990).

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 (Langmead, 2012) to map our sequenced reads to the annotated library created in the previous step. Bowtie2 aligns similar sequences together.

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

Issues with sequencing and the analysis of sequencing data arise from sampling and the fat nature of the data. The sequences that are read by the sequencer are only a small fraction of the DNA from the sample. Additionally, primers used for sequencing may be biased for certain sequences more than others. Lastly, the data is very fat, which is to say that there are magnitudes more variables (in the form of functional annotations of genes) than there are samples. This makes it difficult to have enough power to detect small differences in the data, a concept expanded upon in the Points of Failure section below.

1.5.2 Imputation

Deep metagenomic sequencing can be imputed using a tool called PiCrust from a gene tag experiment (Langille, 2013). PiCrust uses the Greengenes database (DeSantis, 2006) to identify the bacterial taxa in the sample, and pulls their genomes from the Integrated Microbial Genomes database (Markowitz, 2012). 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 (Langille, 2013), compared to a full metagenomic sequencing experiment.

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

1.5.3 Data analysis

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

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

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

Pathway enrichment Biological pathways can be thought of as made up of a series of chemical reactions, each catalyzed by a protein enzyme, which is encoded by a gene. KEGG (Kyoto Encyclopaedia of Genes and Genomes) is a manually curated annotation database that matches genes to pathways (Kanehisa, 2000). This database allows researchers to see if there is differential abundance of pathways encoded by functionally annotated genes, even when the genes may not be differentially abundant by themselves.

1.6 Points of failure

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

1.6.1 Collection methods differ

These experiments are very sensitive to batch effects because microbiome composition can be very variable within groups such that the effect size of a difference between groups can be

small. Wherever possible, all samples should be processed in the same batch. Analysis should also be done to check if samples extracted on different dates or sequenced with different primers separate into clusters, to make sure that there is no systematic bias in the data.

1.6.2 Microbiome data is highly variable between individuals

One highly studied body site is the gut, and the gut microbiome can be affected very strongly by diet (Turnbaugh, 2009). This among other factors lead to a highly diverse gut microbiome between subjects for reasons unrelated to the disease being studied, creating a lot of noise, potentially obscuring real effects or even creating the appearance of false effects.

Generally experiments of this nature typically have low sample sizes due to budget constraints, sample collection difficulties, patient compliance, and other issues. To increase cost effectiveness and reduce batch effects, we run all the samples in an experiment on the same sequencing run, by means of a primer design (Gloor 2010).

There are several models for computationally analyzing the variance within conditions in order to determine if operational taxonomic units are significantly differentially abundant, most of which were originally desigend for RNA-seq experiments on single organisms (Pachter, 2011). Currently the most popular tools for analyzing differential abundance are EdgeR (Robinson, 2010), DESeq2 (Love, 2014), and MetagenomeSeq (Paulson, 2014). EdgeR was cited by 1,130 papers in 2015 according to Google Scholar. DESeq2 and MetagenomeSeq are part of the QIIME pipeline, which was cited by 1,620 papers in 2015.

EdgeR and DESeq2 use the negative binomial distribution. The negative binomial distribution allows the variance of data to be estimated given the mean, through a function. The function is determined by collecting the mean and variance for all the counts for each OTU in each experimental condition, and fitting the variances according to the negative binomial distribution. This vastly underestimates the variance at low counts, which represent the sampling of low abundance OTUs, and can be very different between replicates. Underestimating the variance at low counts produces spurious low p-values for low count OTUs (Fernandes, 2013).

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 doesnt 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 (Paulson, 2013).

For my differential abundance analysis, Ive used ALDEx2, which samples from the Dirichlet distribution to model variation in the data (Fernandes, 2014). After a number of samples, the mean value and mean variance are used to determine if OTUs are differentially abundant between groups, an approach that is believed to result in greater sensitivity and equivalent specificity compared to the DESeq2 approach (Fernandes, 2014).

1.6.3 Microbiome data involves the comparison of many features

Oftentimes, the number of taxa or gene functions comparisons is a magnitude larger than the sample size. This is known in statistics as having more variables than observations, or having

fat data. The higher the ratio of variables to observations are, the less likely the principal components analysis is to be reliable (Osborne, 2004).

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 (Hsiao, 2013).

1.6.4 Microbiome data is compositional

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

First, the total number of reads per sample is irrelevant to the biological implications of the data, as it is limited by how the samples were processed and the sequencing platform. Based on spurious correlations discovered in organ size research, it is known that given compositional data (such as bone lengths as a proportion of height, or OTU abundances that add up to the total number of counts per sample), analysis with the assumption that the variables (bone lengths or OTU counts) are independent lead to spurious positive correlations (Pearson, 1896). The variables thought to be independent are related by the sum they are divided by. Additionally, the constrained sum causes the abundance of different taxa to appear to be negatively correlated with each other when analyzed by conventional statistics. When one taxa increases in abundance, the counts detected in other taxa decrease in abundance, even if the taxa are not decreasing in abundance biologically.

Second, removing an entire variable (an OTU in gene tag sequencing, or a functional annotation in deep metagenomic sequencing) from the analysis should not change correlations between OTUs. Removing variables occur routinely in microbiome research, such as when rare OTUs are discarded. Without a data transformation, removing variables will change the correlation between variables (Aitchison, 1986).

To ensure that these conditions are met, data should be analyzed in a compositional way. Several types of log ratio data transformations are recommended to allow the data to be analyzed by standard Euclidean methods (Aitchison, 1986). The type that makes the most sense for microbiome data is the centered log ratio transform. The centered log ratio transform is performed by dividing each proportional abundance by the geometric mean of all the proportional abundances, and taking the logarithm. The geometric mean acts as a low level baseline abundance in microbiome data. Taking the logarithm of the ratio allows for a consistent measurement whether the large number is in the numerator or denominator of the ratio.

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

Compositional techniques such as those espoused in the Analysis of Composition of Microbiomes (ANCOM) framework (Mandal, 2015) and the ANOVA-Like Differential Expression 2 (ALDEx2) software (Fernandes, 2014) should be used to prevent spurious correlations and promote consistent data analysis. However, these techniques are not yet mainstream in the field.

1.6.5 Microbiome data is sparse

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

Two methods have been suggested in the literature to account for zeros. The first is simply to add a small arbitrary value to each zero, as suggested in the original literature about the statistical analysis of compositional data (Aitchison, 1986). This is used in ALDEx2, and the arbitrary value is chosen to be 0.5, representing complete uncertainty in whether or not a zero count in one sample (where the OTU or gene has non zero counts in other samples) would be a 0 or a 1 in a technical replicate (Fernandes, 2013).

The second method is to take a Bayesian approach where the likelihood that a zero could be changed to a positive count if the sample were resequenced is estimated, based on . This is implemented by the cmultRepl command in the zCompositions package in R (Palarea-Albaladejo, 2015). Based on the shape of the rest of the data for the same sample, the average value of the count detected if a zero were resequenced is determined, and the zeros are all replaced by this fraction.

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

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

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

Here is a picture of a long memory time series.

Here's a table.

n	α	$n\alpha$	β
1	0.2	0.2	5
2	0.3	0.6	4
3	0.7	2.1	3

Table 1.1: A random table

Figure 1.1: A long memory time series

$$y = mx + b \tag{1.1}$$
$$= ax + c \tag{1.2}$$

This is an un-numbered equation, along with a numbered one.

$$u = px$$
$$p = P(X = x) \tag{1.3}$$

Look at Table 1.1 and Figure 1.1 and equations 1.1, 1.2, and 1.3.
Let's do some matrix algebra now.

$$det\left(\begin{vmatrix} 2 & 3 & 5 \\ 4 & 4 & 6 \\ 9 & 8 & 1 \end{vmatrix}\right) = 42 \tag{1.4}$$

In the equation and eqnarray environments, you don't need to have the dollar sign to enter math mode.

$$\alpha = \beta_1 \Gamma^{-1} \tag{1.5}$$

This is citing a reference [**?**]. This is citing another [**?**]. Nobody said something [**?**].

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

Microbiome analysis is frequently performed using the UniFrac distance metric to separate groups. Here we demonstrate that unweighted UniFrac is highly sensitive to rarefaction instance and to sequencing depth in uniform data sets. We show that this arises because of subcompositional effects. We introduce information UniFrac and centered ratio UniFrac, two new weightings that are not sensitive to rarefaction and allow greater separation of outliers than classic unweighted and weighted UniFrac. With this expansion of the UniFrac toolbox, we hope to empower researchers to extract more varied information from their data.

Introduction

In 2005, Lozupone et al introduced the UniFrac distance metric, a measure to calculate the difference between microbiomes that incorporated phylogenetic distance [**?**]. 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 [**?**]. Since then, papers reporting these metrics have garnered over a thousand citations, and enabled research about everything from how kwashiorkor causes malnutrition [**?**] to how people can have similar microbiomes to their pet dogs [**?**]. Except for generalized UniFrac, used to make hybrid unweighted and weighted UniFrac comparisons [**?**], few advances in the metric have occurred since 2007. In this paper we examine a data set where unweighted UniFrac gives misleading results, and discuss some alternative weightings for UniFrac.

2.0.1 Unweighted UniFrac

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

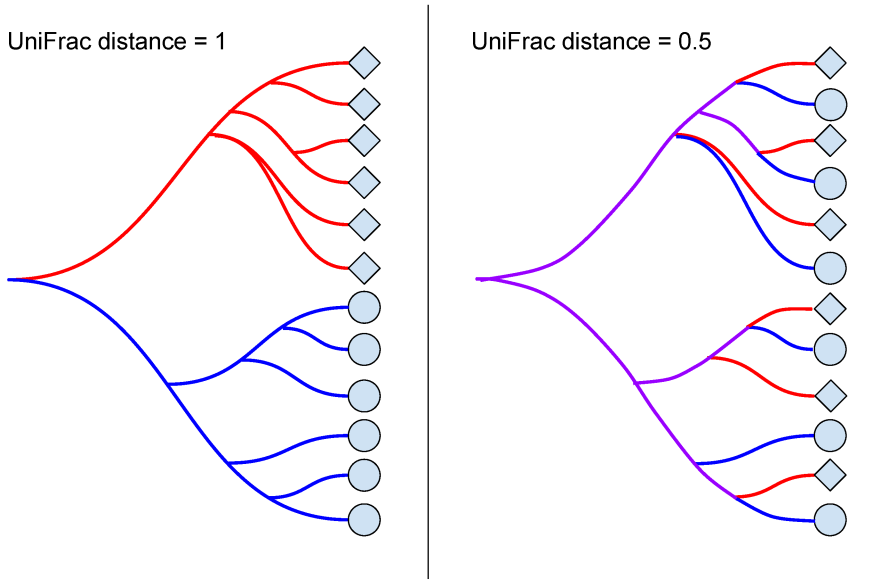

Figure 2.1: **Unweighted UniFrac.** When two samples do not share any branches of the phylogenetic tree, the unweighted UniFrac distance is maximized at 1. When two samples share half of their branch lengths on the phylogenetic tree, the unweighted UniFrac distance is 0.5. If the two samples contain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

2.0.2 Weighted UniFrac

Weighted UniFrac [**?**] is an implementation of the KantorovichRubinstein distance in mathematics, also known as the earth movers distance [**?**]. Rather than looking only at the presence or absence of taxa, each branch length of the phylogenetic tree is weighted by the difference in proportional abundance of the taxa between the two samples.

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

UniFrac is constituted as either a presence/absence (unweighted UniFrac) [**?**], a linear proportion (weighted UniFrac) [**?**], or some combination of the two (generalized UniFrac) [**?**]. However, the data are not linear, because the sum of the total number of reads is constrained by the sequencing machinery [**?**] [**?**] [**?**] [**?**]. 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

2.0.3 Analytical techniques

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

Rarefactions can be performed using the Qiime software [**?**] or using the vegan package in R [**?**].

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 is unshared when one sample has a non-zero abundance but not the other, and a branch is shared when both samples 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:

$$\frac{\sum b_{unshared}}{\sum b_{unshared} + \sum b_{shared}}$$

Weighted UniFrac
Weighted UniFrac [**?**] 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:

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

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

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

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

combines the the concept of uncertainty with phylogenetic relationships to identify taxa that are differentially informative between groups.

The formula for Information UniFrac is as follows:

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

Centered Ratio UniFrac

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

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

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:

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

One challenge when it comes to the analysis of read count data is the presence of zero counts. Whether a low-abundance taxa appears in the data as a zero or a low positive count is up to chance, and assuming that a zero count represents the absence of a taxa can be very misleading [**?**]. A Bayesian approach can be used to estimate the likelihood that a zero could be changed to a positive count if the sample were resequenced, implemented by the cmultRepl command in the zCompositions package in R [**?**].

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

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

Bray-Curtis dissimilarity metric

The Bray Curtis dissimilarity metric [**?**] quantifies how dissimilar two sites are based on counts.

A index of 0 means that two samples are identical, while a index of 1 means samples do not share any species. It is computed as a proportion through the formula:

$$C_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$
$$\text{where } C_{ij} = \text{dissimilarity index bound by [0,1]}$$
$$S_i = \text{Specimen counts at site i}$$
$$S_j = \text{Specimen counts at site j}$$

2.0.4 Data preparation

Tongue dorsum data set
Samples from this experiment were sourced from the Human Microbiome Project [**?**] Qiime Community profiling v35 otu tables (http://hmpdacc.org/HMQCP/). Analysis of the data was conducted on a Late 2011 15 inch MacBook Pro 2.4 GHz i7 with 16GB of 1333 MHz DDR3 RAM. Rarefaction was conducted through Qiime version 1.8.0-20140103, and generation of the ellipse figures was done in R version 3.2.3 (2015-12-10) "Wooden Christmas-Tree" x86_64-apple-darwin13.4.0 (64 bit).

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

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

Tongue dorsum and buccal mucosa data set

Thirty random samples were selected from the tongue site of the Human Microbiome Project [**?**] and 30 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 [**?**] before the unweighted UniFrac distance was calculated. Weighted, information, and centered log UniFrac were calculated on the data set without rarefaction. The resulting distances were plotted for principal coordinate analysis.

The script used to run this analysis can be referenced at https://github.com/ruthgrace/exponentUnifrac/blob/

Breast milk data set

The breast milk data set used here has also been published in the Microbiome Journal [**?**]. The count table was analyzed using our custom UniFrac script. Data was rarefied to the sample with the smallest number of read counts (3072) before the unweighted UniFrac distance matrix was calculated. Non-rarefied data was used for weighted, information, and centered ratio UniFrac. Data was plotted using a principal components or coordinate analysis as appropriate.

The script used to run this analysis can be referenced at https://github.com/ruthgrace/exponentUnifrac/blob/mas

Results

2.0.5 Unweighted Unifrac is highly sensitive to rarefaction variants

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

Sixty tongue dorsum subsamples were drawn from the Human Microbiome Project data without replacement and filtered such that each gene had a minimum sum of 100 counts across samples. 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. For Fig. 2.2, two independent rarefactions of the data were conducted in order to observe the effect of rarefaction variants on the metrics. The unweighted UniFrac distance was then computed for each rarefaction, and Procrustes adjustment was applied in order to overlay the second rarefaction onto the first. A PCA of rarefaction 1 was plotted, and any samples that changed between rarefactions one and two were visualized with red and blue on the plot. If the sample moved from one cluster to another between the rarefactions, it was indicated with either a blue or a red arrow.

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

For the ellipse plot in Fig. 2.3, 60 tongue dorsum subsamples were randomly drawn without replacement and the gene compositions per sample were also filtered to a minimum of 100. A hundred separate rarefactions were conducted on the data to a minimum sampling depth of 378. For each individual rarefied OTU table, a distance matrix was computed using the unweighted Unifrac, weighted UniFrac, Bray-Curtis Dissimilarity, information UniFrac, and centered ratio UniFrac as a weighting method. By generating 100 separate datasets for each metric, it is

possible to assess the magnitude of difference each metric has by analyzing what is essentially the same data. In other words, what does the effect of random sampling (rarefaction) have on the output of each metric? Each distance matrix generated per metric was adjusted with a Procrustes adjustment to overlay the subsequent rarefactions onto the first.

Thus, 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[**?**] and Nature [**?**].

The maximum value of Vres for each rarefaction is plotted against the median value per rarefaction in Fig. 2.3. This plotting serves to highlight the maximum potential change for an analysis given that there is no difference in the data. Unweighted UniFrac shows by far the highest maximum potential change between rarefactions, compared to weighted, information, and centered ratio UniFrac, as well as Bray-Curtis.

2.0.6 Why does Unweighted Unifrac have discrepancies when analyzing rarefied data?

The UniFrac distance is defined as the sum of unshared branches divided by the sum of all branch lengths [**?**]. Samples that are dissimilar will have values closer to 1 as they should have more unshared branches relative to one another. Similar samples have a value closer to 0 since they will have fewer unshared branches. As defined previously, rarefaction serves the purpose of standardizing sample counts to a common denominator, which is usually defined as the lowest sequencing depth(cite rarefaction paper here). 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 [**?**]. A combination of unevenly sampled genes and distantly related genes will contribute to UniFrac's variability when genes are ultimately rarefied. Distance matrices between samples will be affected when rare genes 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 genes disappear.

$$Distance_{A:B} for Rarefaction 1$$

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

Table 2.1: **Original abundance of taxa and rarefied abundance of taxa.**

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

$$Distance_{A:B} for Rarefaction 2$$

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

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

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

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

2.0.7 Information UniFrac

The difference in information content between low proportional abundances (which make up the bulk of microbiome data) is generally higher than the difference between the proportional

abundances themselves, potentially allowing scientists to differentiate groups with subtle differences.

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

2.0.8 Tongue and cheek comparison

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

Fig. 2.6 shows a principal coordinate analysis plot with four different metrics: unweighted UniFrac, weighted UniFrac, information UniFrac, and centered ratio UniFrac. We observe that the difference in the microbiome between the human tongue and cheek 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 that weighted UniFrac, information UniFrac, and centered ratio UniFrac do not tend to show spurious separation in uniform data sets to the degree that unweighted UniFrac does (Fig. 2.3), while reliably separating samples in data with a defined difference between groups.

2.0.9 Breast milk Data

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

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

Discussion

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

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

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

Acknowledgments

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

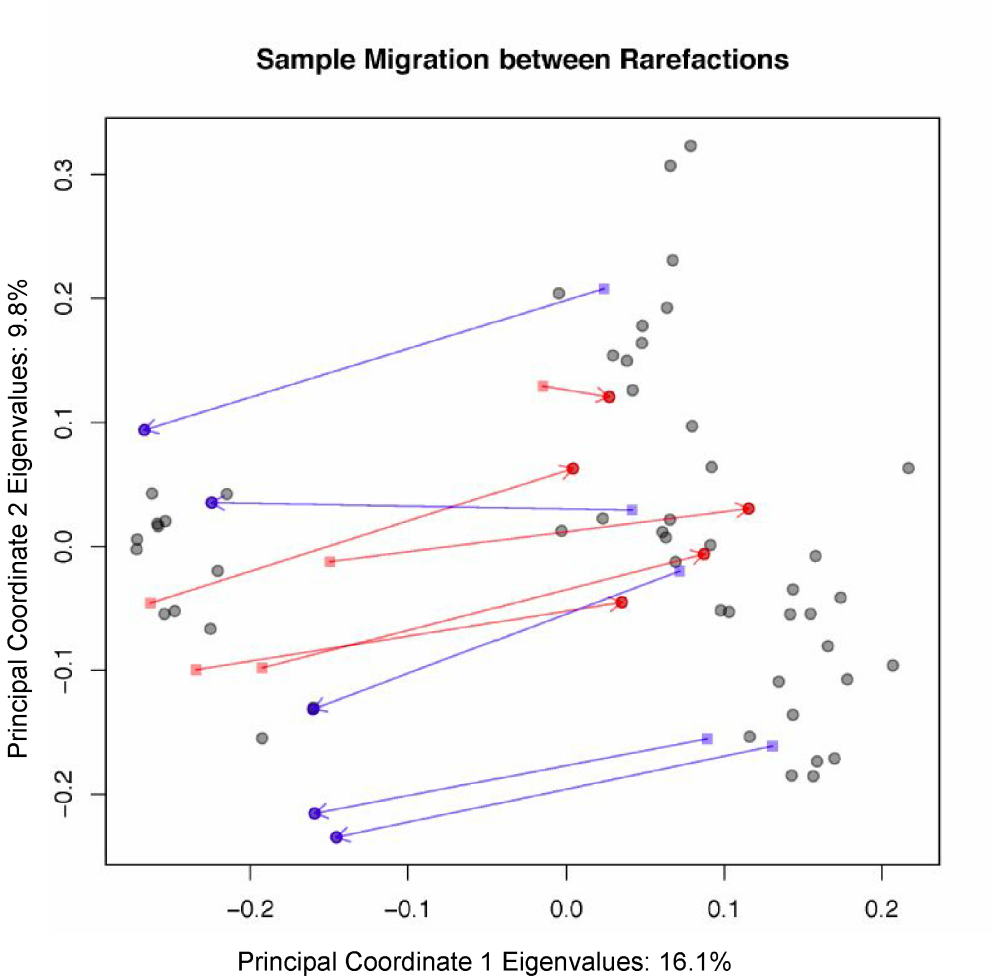

Figure 2.2: **Sample migration in different rarefactions, plotted on principal coordinates, measured with unweighted UniFrac.** Red samples have moved from the left cluster to the right cluster between rarefactions. Blue samples have moved from the right cluster to the left. Samples are taken from the tongue dorsum body site from the Human Microbiome Project database.

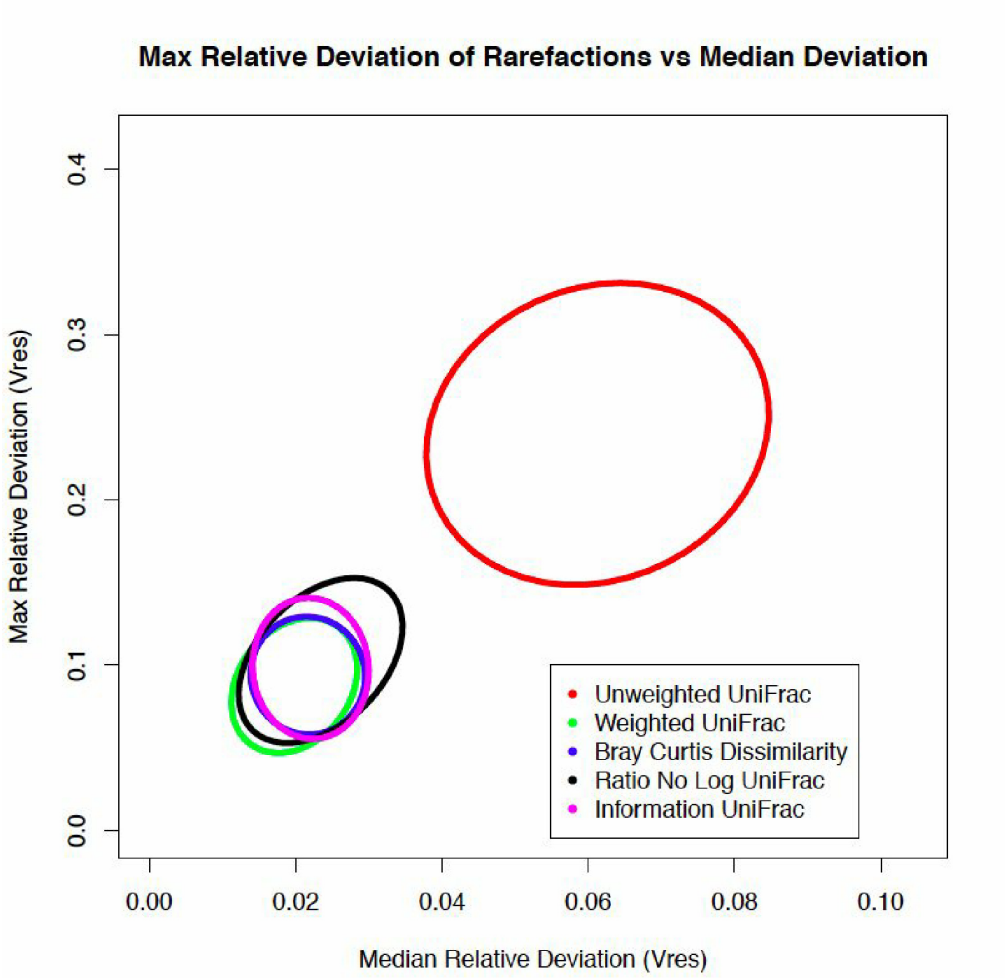

Figure 2.3: **Maxiumum relative deviation of rarefactions versus median deviation for traditional and non-traditional microbiome dissimilarity metrics.** Sixty samples from the tongue dorsum were taken from the Human Microbiome Project [**?**], and rarefied 100 times. The maximum relative deviation was 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. Both the maximum relative deviation of rarefied data and the median relative deviation of rarefied data are greater in unweighted UniFrac than in weighted UniFrac, Bray Curtis distance, centered ratio UniFrac, and information UniFrac.

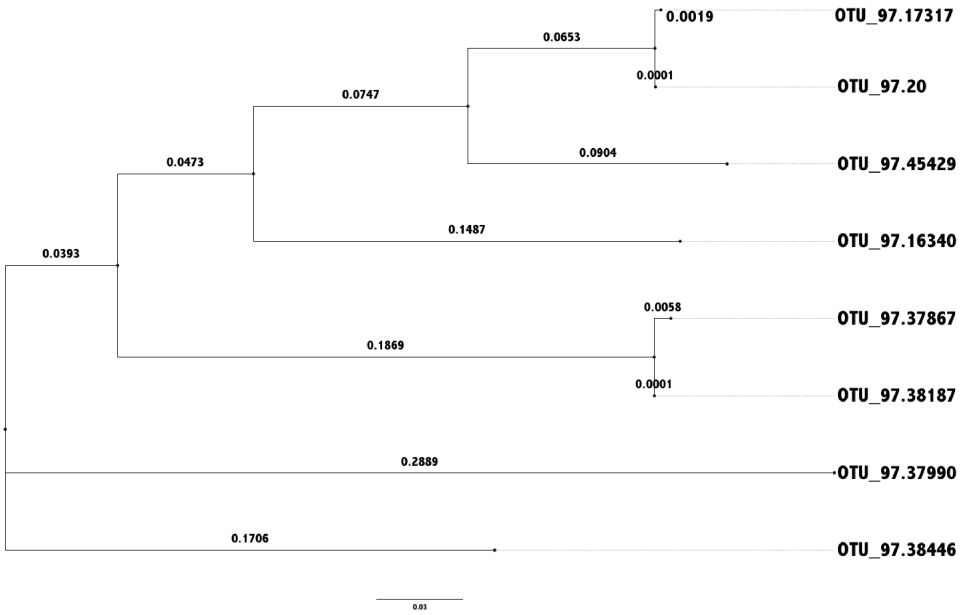

Figure 2.4: **Phylogenetic tree with long isolated branches.** Variation in different rarefactions of data in unweighted UniFrac analysis is exacerbated by the presence of long isolated branches in the phylogenetic tree.

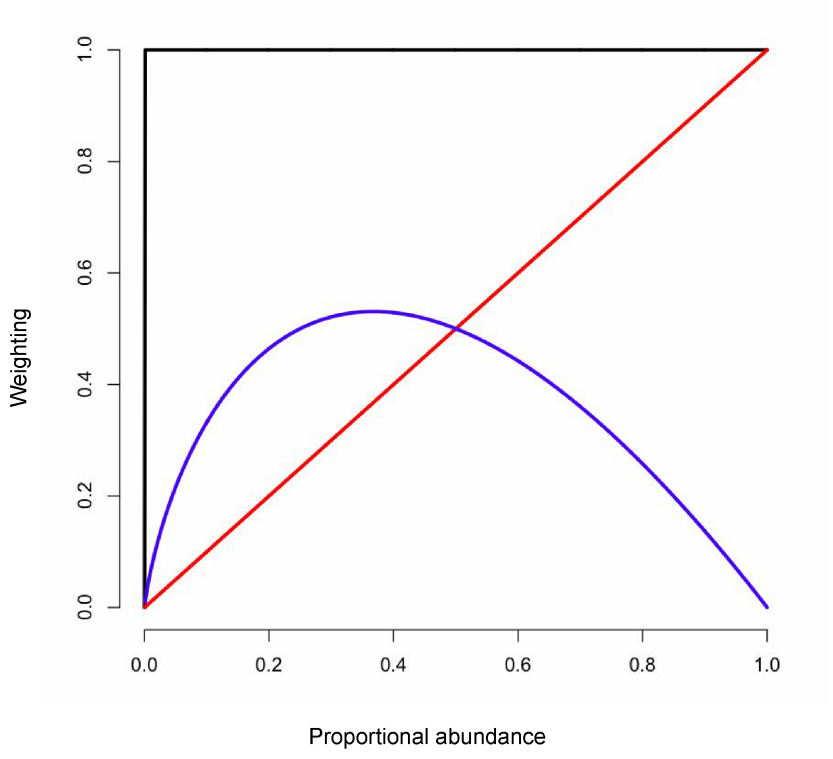

Figure 2.5: **Unifrac weights.** Each UniFrac weighting is plotted with the corresponding proportional abundance. The black line is unweighted UniFrac, the red line is weighted UniFrac,and the blue line is information UniFrac.

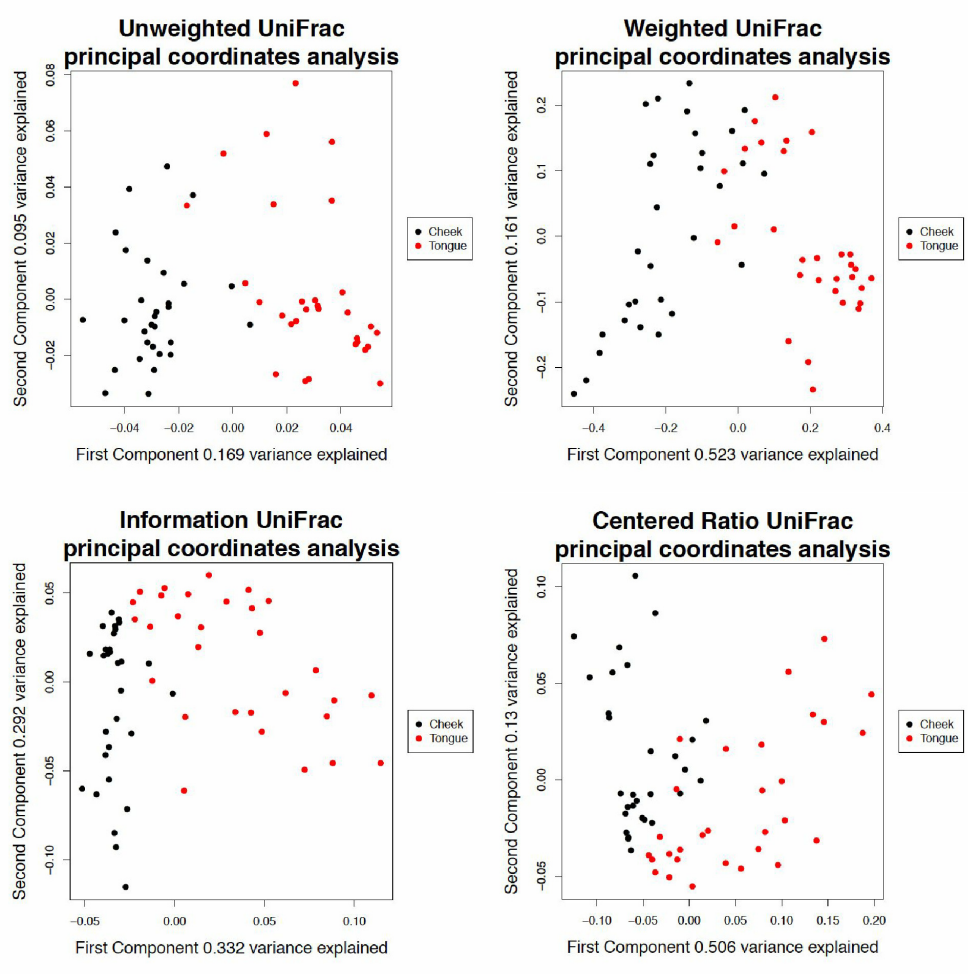

Figure 2.6: **Analysis of tongue and cheek data using different UniFrac weightings.** A principal coordinate analysis of a 16S rRNA experiment done on samples from the tongue and cheek, selected from the Human Microbiome Project [**?**]. All weightings show separation between the samples by body site.

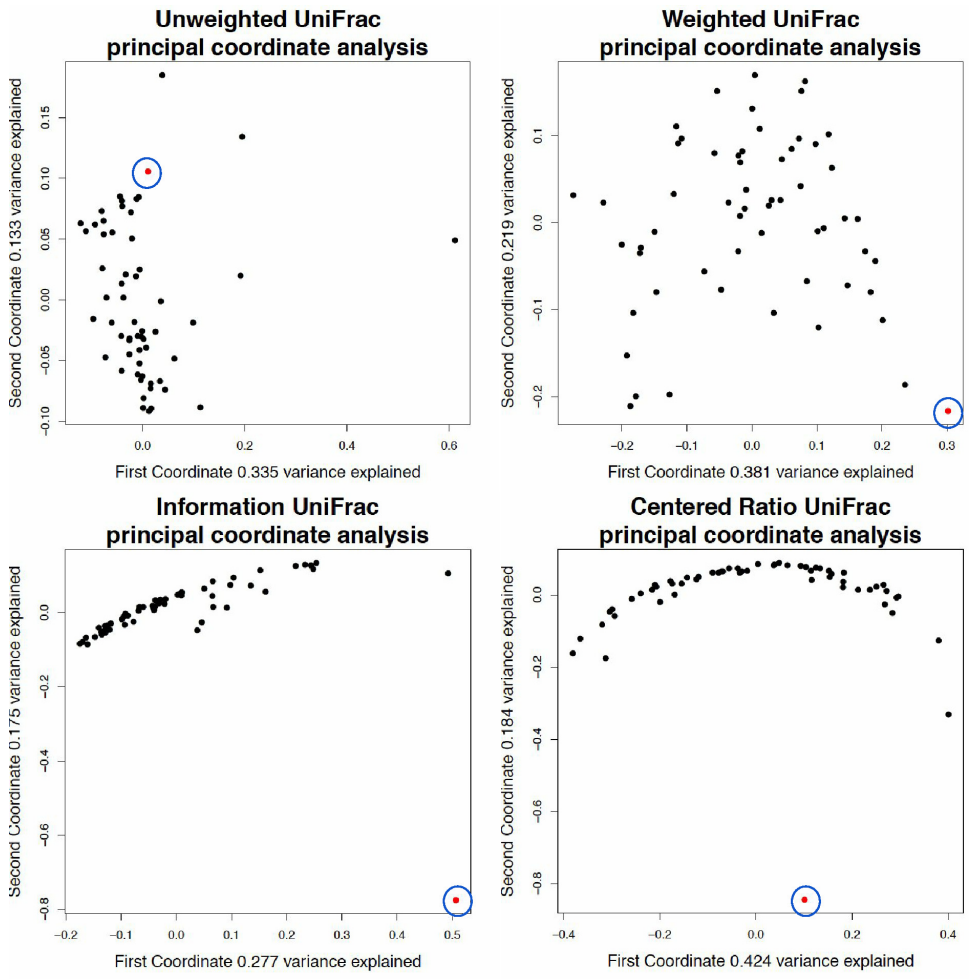

Figure 2.7: **Analysis of breast milk data using different UniFrac weightings.** A principal coordinate analysis of a 16S rRNA experiment done on samples from a 16S rRNA experiment on breast milk. The circled sample is infected with 97% *Pseudomonas*, compared to 15-20% in the other samples.

Bibliography

Bibliography

Appendix A

Proofs of Theorems

Proof of Theorem ??

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

$$= -1 \quad \text{(A.2)}$$

∎

Curriculum Vitae

Name:	Ruth Wong
Post-Secondary Education and Degrees:	The University of Western Ontario London, ON 2010-2014 B.M.Sc. University of Western Ontario London, ON 2014-2016 M.Sc.
Honours and Awards:	Western Gold Medal 2014 Leland Ritcey Prize 2011
Related Work Experience:	Summer Intern, Persistent Disk Team Google Inc., New York office Summer 2015 Google Summer of Code Participant Bader Lab, University of Toronto Summer 2014

Publications:

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