**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Jessica Lawanna Chen Date

# The Effect of Unclassified Taxa on the UniFrac Distance Measurement

By

Jessica Chen

Master of Science in Public Health

Department of Biostatistics

---

Vicki Hertzberg
Committee Chair

---

Hao Wu
Committee Member

# The Effect of Unclassified Taxa on the UniFrac Distance Measurement

By

Jessica Chen

Bachelor of Arts
University of California, Berkeley
2014

Thesis Commitee Chair: Vicki Hertzberg, Ph.D.

An Abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2016

# Abstract

The Effect of Unclassified Taxa
on the UniFrac Distance Measurement

By Jessica Chen

The microbes in our microbiome can both benefit and harm the human host. Researchers are still figuring out what combinations of microbes are beneficial to humans in order to prevent or fight against diseases. The UniFrac measure is a distance measurement between two samples that shows how similar the two samples are biologically. Many Operational Taxonomic Units (OTUs) have unclassified taxa. Researchers usually either keep the OTUs with unclassified taxa or delete those OTUs from analyses. This study analyzes the weighted UniFrac distance measurement when the counts for the OTUs with unclassified taxa are imputed onto a known OTU. These UniFrac distance measurements calculated from imputation are compared to the UniFrac distance measurements when the original OTUs are used and when the OTUs with unknown taxa are removed from analyses. We find that the UniFrac distances created from deletion of OTUs with unknown taxa are on average smaller than the UniFrac distances created from the original OTUs. We find that the UniFrac distances created from the imputation method are on average greater than the UniFrac distances created from the original OTUs.

# The Effect of Unclassified Taxa on the UniFrac Distance Measurement

By

Jessica Chen

Bachelor of Arts
University of California, Berkeley
2014

Thesis Commitee Chair: Vicki Hertzberg, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2016

## Acknowledgements

*I would like to thank Dr. Vicki Hertzberg for her guidance, support, and encouragement throughout this journey.*
*I would also like to thank Dr. Hao Wu for taking the time to read my thesis.*

# Contents

# List of Figures



# List of Tables

# 1 Introduction

## 1.1 Human Microbiome

The human body can consist of about 100 trillion cells that are not all human. For every human cell there are about 10 cells that consist of bacteria, viruses, or other microorganisms [1]. Until very recently, these microorganisms have gone unstudied. These trillion of microorganisms that live in our mouth, gut, skin, and other parts of our body are collectively called the human microbiome. The microbiome is imperative in helping to maintain human health [2]. The microbes in our microbiome provide nutrients and vitamins for us that we do not have the genes to produce, break down our food in order to extract essential nutrients, as well as prevent and defend against disease causing microbes. On the flip side, certain communities in our microbiome can also foster disease [1].

Scientists use DNA sequencing in order to identify specific microbes in samples from nasal, oral, skin, and gastro-intestinal sites. A common technique to do so is to sequence a marker, which is a short, unique DNA sequence that can be used to identify the genome that contains it. This process helps researchers to identify all the species in a large amount of samples quickly. A common DNA marker that is utilized is the gene that codes for the 16S subunit of ribosomal RNA [3].

The microbiome of the same area for two different people will look more similar than different sites on one person [2]. For example, the microbiome of the

gut looks vastly different than the microbiome in the same persons's mouth. The variability of the microbiome by site and by person makes it difficult to describe a typical microbiome. The reason for diversity of different sites is generally unknown, though the environment, one's diet, genetics, and early exposure to microbiota have been attributed [2]. Learning more about the microbiome can help us better understand what microbes are located at certain sites in our body as well as the interaction between microbes that are either beneficial or detrimental to our health. There is a growing amount of literature that suggests changes in our microbiome correlate to disease states, suggesting that taking control of our microbiome can possibly help prevent and treat disease.

## 1.2 Measuring Biological Diversity

There are two different classifications for measuring biological diversity. $\alpha$-diversity measures within sample diversity, i.e. which bacterium are in the sample and how many of each bacterium. $\beta$-diversity measures between sample diversity, i.e. how similar are two different samples in terms of bacterium. Non-phylogenetic metrics are utilized when all operational taxonomic units (OTU) are treated as being equally related while phylogenetic metrics include evolutionary relationships between OTUs [4].

Phylogenetic measures offer more information and advantages over non-phylogenetic measures. Dan Faith originally developed phylogenetic diversity in the

early 1990s. This measure originally was meant to measure communities of macro-organisms instead of microorganisms. However, phylogenetic diversity has translated well to measuring microorganisms. Phylogenetic diversity can be measured simply by the sum of the branch length in a phylogenetic tree that is represented in a specific sample [4].

### 1.2.1 $\alpha$-diversity

For $\alpha$-diversity, one could focus on either richness or evenness. Richness refers to how many different types of organisms are present in one sample. Researchers could count the number of different OTUs observed in each sample and that would be a basic measure of richness for each sample. Evenness refers to how even or uneven the distribution of species are in a sample. If the most abundant organism in one sample is as common as the least abundant organism in the sample, then the evenness of the sample is said to be high. If the most abundant organism is much more common than the least abundant organism, then the evenness of the sample is said to be low [4].

### 1.2.2 $\beta$-diversity

There are many different metrics for computing pairwise dissimilarity of samples. One such metric is the Bray-Curtis dissimilarity measure, which is a quantitative non-phylogenetic $\beta$-diversity metric. The Bray-Curtis dissimilarity metric

between a pair of samples, j and k, is defined as:

$$BC_{jk} = \frac{\sum_i \mid X_{ij} - X_{ik} \mid}{\sum_i (X_{ij} + X_{ik})} \tag{1}$$

where i is the observation such as OTU, $X_{ij}$ is the count of observation i in sample j, and $X_{ik}$ is the count of observation i in sample k [5].

For this study, we will be focusing on the weighted UniFrac measurement. UniFrac stands for "unique fraction metric". The weighted UniFrac measurement is a type of phylogenetic metric that estimates β-diversity. Weighted UniFrac takes into account the relative abundance of species and other taxa shared between samples, whereas the unweighted UniFrac, a type of qualitative measurement, only considers the presence or absence of species and other taxa. The unweighted UniFrac measure is calculated by the following

$$u = \frac{\sum_{i=1}^{N} l_i \mid A_i - B_i \mid}{\sum_{i=1}^{N} l_i max(A_i, B_i)} \tag{2}$$

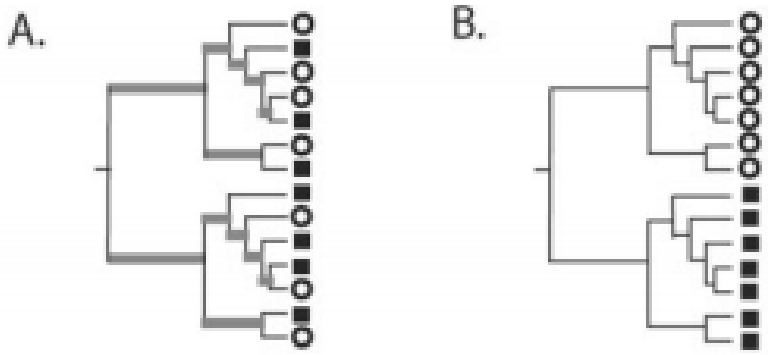
where N is the number of nodes in the tree, $l_i$ is the branch length between node i and its parent, and $A_i$ and $B_i$ are indicator values of 0 or 1 as descendants of node i are absent or present in communities A and B respectively [6]. The weighted UniFrac measure is normalized so that each sequence contributes equally to the distance calculated in order to account for varying rates of evolution between sequences. The weighted UniFrac measure will be examined more in depth in the Methods section [7].

The UniFrac distance measures the phylogenetic distance between two sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads down to either set of taxa, but not both. The UniFrac measurement captures the total amount of evolution that is unique to each set of taxa [4].

If two environments are quite similar, few adaptations would be required to transfer one environment to the other. A phylogenetic tree would have descendants from both environments, with much of the branch length in the tree shared between the two similar environments. Alternatively, if two environments are starkly different, then the lineages in the phylogenetic tree would be distinct for each environment, with most of the branch length leading to descendants of one of the two environments. When comparing how similar several different environments are simultaneously, a distance matrix describing the pairwise phylogenetic distances between sets of environments is created to comprehensively show all possible pairwise dissimilarities [4].

Figure 1 [8] on page 6 exhibits two examples of the UniFrac distance of two very similar communities and two very dissimilar communities. The squares and circles represent sequences from different communities. Branches are dark grey and thicker if they are unique to a particular environment and thin lines if they are shared. (A) is a tree representing communities that are phylogenetically similar, a significant amount of branches in the tree are shared. (B) is a tree representing communities that are 100% dissimilar with no branches shared [8].

Figure 1: Example showing the UniFrac distance measurement [8]




## 1.3 Human Microbiome Project

The NIH Common Fund Human Microbiome Project (HMP) was established in 2008. The main mission of the HMP was to generate resources that would allow researchers to comprehensively characterize the human microbiome and the role microbes have in human health. The HMP utilizes metagenomics, which allows for a thorough investigation of genomes taken from microbial communities in the natural environments without the need for cultivation in the laboratory [9].

Samples were collected from human subjects at 1-3 visits. The subjects had samples taken from 15-18 body sites over the course of the series of visits. There were 1,642 subjects who participated in the first visit. Of these 1,642 subjects for the first visit, 108 had their stool samples collected for analyses. There were 1,244 subjects who participated in the second visit. Of these 1,244 subjects for the second visit, 78 had their stool samples collected. Lastly, there were 12 subjects who participated in the third visit. Of these 12 subjects in the last visit, only 1 subject had their stool sample collected for analysis.

Each individual sample was assigned a primary sample number (PSN) id. DNA was extracted from the primary samples and were assigned a Nucleic Acid Preparation (NAP) id. Extractions were divided into aliquots as necessary, where each aliquot was assigned a sample number (SN) id. Sequencing centers performed sequencing on each SN sample. The centers sequenced 16S variable regions 3-5 (V35) from every sample in the study, and variable regions 1-3 (V13) and variable regions 6-9 (V69) from subsets of samples. Data was submitted by the center to the National Center of Biotechnology Information (NCBI) where the data were organized [9].

The mapping file contains the sample ID, subject identifier, the visit number, sex of the subject, the body site, and a description. The Final OTU table consists of the count of sequences for each OTU by each sample ID. A representative sequence file contains a representative sequence for each OTU for creating a phylogenetic tree. For this study, sequences sets, mapping files, and OTU tables from the HMP 16s rRNA for the (V13) variable regions will be used.

## 1.4 Unclassified Taxa

Little research has been done on analyzing how unclassified taxa affects the β-diversity measurement and subsequently the conclusions and interpretations drawn from the measure. Some researchers eliminate any sequences with unknown taxonomy from their analyses while others keep all original OTU sequence counts regardless of unclassified taxa [10]. Missing taxa can lead to misleading results.

## 1.5 Purpose

Due to the lack of research in examining the effect of unclassified taxonomy on the UniFrac measurement, this study will measure this effect. We will be comparing the average pairwise UniFrac distance in the HMP V13 regions with unclassified taxa to two different measurements. One is the average pairwise UniFrac distance of 30 matrices where OTUs with missing taxa are randomly distributed to a taxa that is known. The second measurement for comparison is the average pairwise UniFrac distance of samples where any OTUs with unknown taxa are eliminated. We are trying to determine if there exists a bias in the distance measure when different methods are utilized to deal with unclassified taxa.

## 1.6 Definitions

*OTU Table*: a table of the sequence counts for each OTU (rows) per sample (columns)

*Biological Observation Matrix* (BIOM): same information as the OTU Table in a different format that is required for QIIME analyses

*FASTA format*: a file that contains the sequencing data

$\alpha$-*diversity*: measures what species and how many of them are in a sample

$\beta$-*diversity*: measures how similar pairs of samples are

# 2 Methods

QIIME (Quantitative Insights Into Microbial Ecology) is an open-source bioinformatics pipeline for performing microbiome analysis [11]. QIIME was used to subset the mapping files, to create the phylogenetic tree, and to produce the pairwise UniFrac distance matrices. R was used to impute the OTUs with unclassified taxa onto those with known taxa and to calculate the average UniFrac distance of a matrix.

## 2.1 Cleaning Data

We are interested in only the first visit for the subjects and the stool samples in order to control for any bias that might incur from observing all site samples at all visits. We end up with a total of 108 samples that are from the first visit and stool samples. We believe samples from the same subject and site over multiple visits will be closer in UniFrac distance as opposed to samples between different subjects. Also, as mentioned in the Introduction, the same site between two different subjects is thought to be closer biologically than two different sites on the same subject.

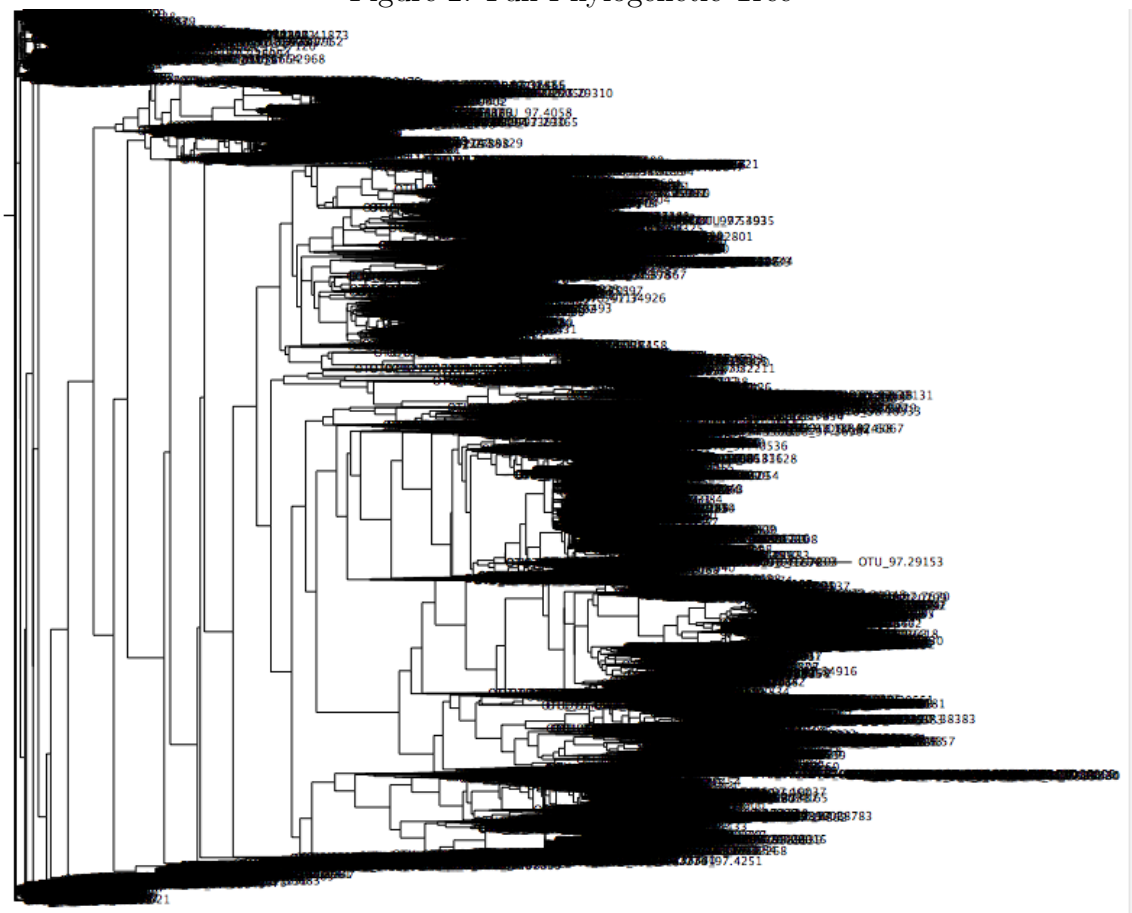
The OTU table is first converted to a Biological Observation Matrix (BIOM) file, which is the required file format in QIIME. The filter_samples_from_otu_table script was used to keep only the samples that had a stool sample (HMPbodysubsite:Stool) on the first visit (visitno:1). A total of 108 samples matched our preset criteria. The OTU table was converted to a text file for subsequent analyses in R.

## 2.2 Phylogenetic Tree

Since the UniFrac distance measures the phylogenetic distance between two samples, a phylogenetic tree must be created. In order to create a phylogenetic tree, QIIME requires a multiple sequence alignment. Trees are created with a set of sequences representative of OTUs. The HMP provides a representative sequence set for the V13 variable regions. The representative sequence set is a single FASTA file that contains one sequence for each OTU.

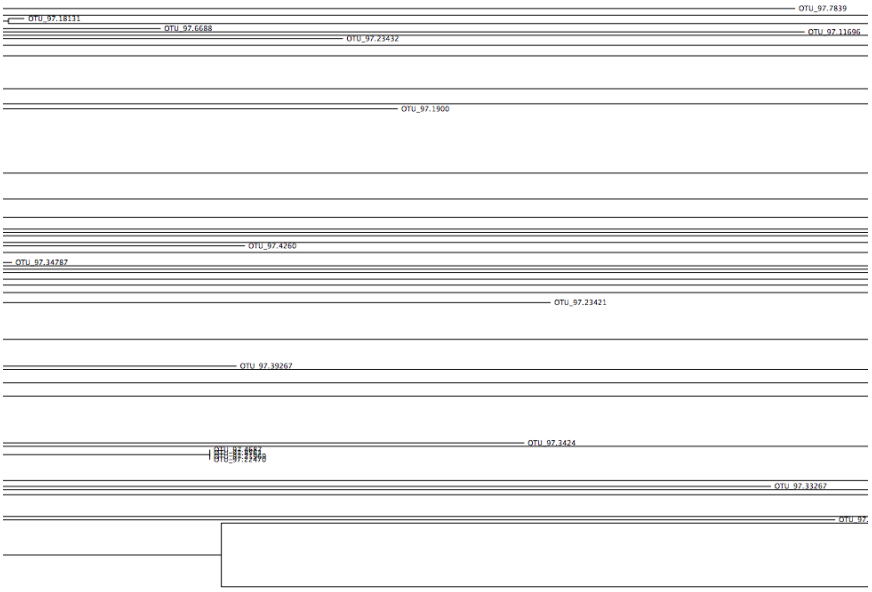
In order to prepare the representative sequence set for creating the phylogenetic tree, the align_seqs script is used to align the sequences in a FASTA file to each other or to a template sequence alignment. The alignment method used is PyNAST, the default alignment method in QIIME. PyNAST is the python implementation of the NAST alignment algorithm. The NAST algorithm aligns each provided sequence (the "candidate" sequence) to the best-matching sequence in a pre-aligned database of sequences (the "template" sequence). The candidate sequence is not allowed to produce new gap characters into the template, therefore the algorithm produces local mis-alignments to preserve the existing template sequence. The default matching criteria for matching between the template sequences and a candidate sequence is a minimum sequence length of 150 and the minimum percent id is 75%.

Figure 2: Full Phylogenetic Tree




This dataset contains 43,140 OTUs. The phylogenetic tree that represents the relationship between these OTUs is quite extensive and much more complex than the phylogenetic tree represented in Figure 1 on page 6, as we can see in Figure 2 on page 11. Figure 3 on page 12 is zoomed in for better understanding of the phylogenetic tree created. Though there are lots of branches and nodes, we are able to distinguish between shared and unshared branches between OTUs. The phylogenetic tree is necessary for calculating the pairwise distance between two samples by examining the shared and unshared branches in a tree.

Figure 3: Phylogenetic Tree Zoomed In




## 2.3 Imputation

The imputation of the OTU table was done in R. The original OTU table lists out what each OTU is by phylum, class, order, family, and genus with each taxum separated by a semicolon. The species level was not specified for the OTUs. The separated taxa entries were added to the subsetted OTU table. All blank entries for any taxa were coded as “NOS”, standing for ”Not Otherwise Specified”. If the class of an OTU was classified as NOS, then all subsequent entries down the taxonomic rank (order, family, and genus) were coded as “NA”. If the order of an OTU was classified as NOS, then all subsequent entries down the taxonomic rank (family, and genus) were coded as “NA”. Similar coding was done for NOS for the family taxa. Table 1 on page 13 shows the number of unique taxa per taxonomic level (phylum, class, order, family, genus). It also shows the number of OTUs that are “NOS” per

taxonomic level (highest level of an OTU that is "NOS"). There are a total of 43,140 OTUs to begin with, including fullly classified OTUs as well as unclassified OTUs at any taxa level).

| Taxonomic Level | Number of unique taxa observed | Number of OTUs with level NOS |
|---|---|---|
| Phylum | 29 | 521 |
| Class | 51 | 751 |
| Order | 93 | 1042 |
| Family | 159 | 4204 |
| Genus | 353 | 8124 |

Table 1: Number of unique taxa and number of OTUs labeled NOS per taxonomic level

If an OTU had a class that was NOS then a second OTU that had the same phylum and a specified class was weighted randomly sampled. The OTU with the NOS class was then redistributed into the second OTU; the count of the number of sequences observed in the first OTU for each sample was added to the second OTU count of sequences observed. The OTU with the NOS class was deleted thereafter. The same concept was done for the order, family, and genus levels in the respective order. For example, OTU_97.100 is an OTU with its phylum as Firmicutes, its class as Clostridia, its order as Clostridiales, its family as Veillonellaceae, and its genus as NOS. A second OTU will be weighted randomly sampled that has the same phylum, class, order, and family, with a classified genus. The counts of number of sequences observed for OTU_97.100 will be redistributed to the second OTU. This process was done 30 times for all OTUs with NOS taxa, creating 30 different OTU tables that are redistributing the OTU counts of unclassified taxa to those with known taxa. Doing more than 30 iterations proved to be computationally intensive.

## 2.4 Measuring Weighted UniFrac

The weighted UniFrac is measured by the quantity W defined as

$$W = \frac{\sum_{i=1}^{N} I_l \mid \frac{A_i}{A_T} - \frac{B_i}{B_T} \mid}{\sum_{j=1}^{S} L_j} \tag{3}$$

where N is the number of nodes in the tree, S is the number of sequences represented by the tree, $I_l$ is the branch length between node i and its "parent", $L_j$ is the total branch length from the root to the tip for sequence j. $A_i$ and $B_i$ represent the number of sequences from communities A and B that descend from the node. $A_T$ and $B_T$ represent the total number of sequences from communities A and B [7].

QIIME cannot read an OTU table with characters so the imputed OTU tables are exported as text files with the OTU names, sample ID, and counts. The newly exported OTU table is converted to a BIOM file. The beta_diversity script was used in QIIME to create a 108 by 108 matrix of pairwise sample dissimilarity for each of the 30 matrices. There are 108 columns ad rows since the analyses were run on 108 samples (first visit and stool samples). There are 30 matrices since 30 iterations were run.

The beta_diversity script for a phylogenetic measure like the UniFrac distance requires the OTU table in biom format, the beta-diversity metric to use, and the phylogenetic tree filepath. The OTU table contains the number of sequences observed in each OTU (rows) for each sample (columns).

Thirty-two different distance matrices (1 with all OTUs, 1 with OTUs with unclassified taxa deleted, and 30 for 30 iterations done with the imputation method) containing a dissimilarity value for each pairwise comparison were created. One distance matrix was created using an OTU table with all the original OTUs, including the ones with unclassified taxa. The matrix below is a subset of the first five samples of the distance matrix created using an OTU table with all the original OTUs. The matrix below is showing the upper left 5x5 matrix out of the 108x108 matrix, where element ij is the UniFrac distance between sample i and sample j. The distance matrices produced are all symmetric matrices with 0's down the diagonal. This is the case since the distance between community A and community B is the same as the distance between community B and community A and there is no dissimilarity between community A and community A.

$$\begin{bmatrix} 0 & 0.3037 & 0.3163 & 0.4190 & 0.3923 & \dots \\ 0.3037 & 0 & 0.3577 & 0.4160 & 0.4031 & \vdots \\ 0.3162 & 0.3577 & 0 & 0.3451 & 0.3284 & \vdots \\ 0.4190 & 0.4160 & 0.3451 & 0 & 0.4932 & \vdots \\ 0.3923 & 0.4031 & 0.3284 & 0.4932 & 0 & \vdots \\ \vdots & \dots & \dots & \dots & \dots & \ddots \end{bmatrix}$$

One distance matrix was created using an OTU table with any OTUs with unclassified taxa eliminated from the table. The matrix below is the same subset as above, the first five samples, of the distance matrix created using an OTU table with the OTUs

with unclassified taxa deleted. The matrix below shows the upper left 5x5 matrix out of the 108x108 matrix, where element ij is the UniFrac distance between sample i and sample j, similar as above. We can attempt to eyeball the differences between this matrix and the first matrix for the first five samples. It is hard to compare every single pairwise distance to its same pair in the original OTU matrix.

$$\begin{bmatrix} 0 & 0.3207 & 0.2803 & 0.3895 & 0.3044 & \dots \\ 0.3207 & 0 & 0.3510 & 0.4115 & 0.3209 & \vdots \\ 0.2803 & 0.3510 & 0 & 0.2365 & 0.2536 & \vdots \\ 0.3895 & 0.4115 & 0.2365 & 0 & 0.3335 & \vdots \\ 0.3044 & 0.3209 & 0.2536 & 0.3335 & 0 & \vdots \\ \vdots & \dots & \dots & \dots & \dots & \ddots \end{bmatrix}$$

As mentioned above, thirty distance matrices were created using thirty OTUs using the imputation method. The matrix below is a subset of the first five samples created using an OTU table based on the imputation method for the first iteration. The matrix below is just one realization out of the 30 iterations produced by imputation. The matrix below shows the upper left 5x5 matrix out of the 108x108 matrix, where element ij is the UniFrac distance between sample i and sample j. Again, we can try to eyeball the first five samples of this matrix versus the first matrix, but this does not accurately represent how the method affects the UniFrac distance over all the samples.

$$\begin{bmatrix} 0 & 0.2368 & 0.2910 & 0.3975 & 0.4376 & \dots \\ 0.2368 & 0 & 0.3314 & 0.4069 & 0.4383 & \vdots \\ 0.2910 & 0.3314 & 0 & 0.3440 & 0.3975 & \vdots \\ 0.3975 & 0.4069 & 0.3440 & 0 & 0.5539 & \vdots \\ 0.4376 & 0.4383 & 0.3975 & 0.5539 & 0 & \vdots \\ \vdots & \dots & \dots & \dots & \dots & \ddots \end{bmatrix}$$

In R, the average pairwise UniFrac distance was calculated after disregarding the "0" distances in the matrix between the same sample. The average and standard error of the mean of the thirty matrices created by the imputation method were calculated. The standard error of the mean was calculated by

$$SE = \frac{SD(A)}{\sqrt{N}} \tag{4}$$

where SD(A) is the standard deviation of the 30 iteration averages and N is the number of iterations. The average of the 30 averages, the average of the matrix created from OTUs with unclassified taxa, and the average of the matrix created from any OTUs with unclassified taxa cut out were compared to see if there exists any differences in measurement between these three different types of distance matrices (1 with all OTUs, 1 with OTUs with unclassified taxa deleted, and 1 done with imputation) created in QIIME.

# 3 Results

We find that the OTU table where the OTUs were left untouched yields an average UniFrac distance measure of 0.3538 with a standard deviation of 0.1300. We find that the OTU table where any OTUs with unclassified taxa were promptly eliminated yields an average UniFrac distance measure of 0.3212 with a standard deviation of 0.1306, which is smaller than the average UniFrac distance where the OTUs were left untouched by 9.21%. The standard deviations about the average UniFrac distance are fairly equivalent. Table 2 on page 19 lists out the average UniFrac distance between any two samples and standard deviation about the average for each pairwise matrix, including the 30 matrices produced by imputation.

We find that the mean of the 30 averages of the pairwise matrices computed by imputation is 0.3779. The standard error of the mean is 0.0006. The mean of the averages of the 30 iterations is a 6.81% increase in UniFrac distance from the mean of the matrix with all original OTUs included. These values can be found in Table 3 on page 20. The average standard deviation of the 30 iterations is 0.1316. The standard error of mean of the standard deviations is 0.00042. The standard deviations of the 30 iterations are very similar to the standard deviation of the original pairwise matrix as well as the standard deviation of the pairwise matrix with unclassified OTUs deleted.

We can see in figure 4 on page 20 that the interquartile range for all the matrices are similar. The upper-tail outliers are at a lower range for the matrix with deleted unclassified OTUs (yellow) compared to the matrix with all the original OTUs

| Matrix | Average UniFrac distance between 2 samples | Standard Deviation |
|---|---|---|
| NOS Taxa Included | 0.3538 | 0.1300 |
| OTUs with NOS Taxa Deleted | 0.3212 | 0.1306 |
| Impute 1 | 0.3771 | 0.1305 |
| Impute 2 | 0.3711 | 0.1301 |
| Impute 3 | 0.3737 | 0.1289 |
| Impute 4 | 0.3755 | 0.1302 |
| Impute 5 | 0.3716 | 0.1300 |
| Impute 6 | 0.3776 | 0.1316 |
| Impute 7 | 0.3725 | 0.1321 |
| Impute 8 | 0.3726 | 0.1311 |
| Impute 9 | 0.3736 | 0.1295 |
| Impute 10 | 0.3711 | 0.1281 |
| Impute 11 | 0.3780 | 0.1346 |
| Impute 12 | 0.3790 | 0.1357 |
| Impute 13 | 0.3761 | 0.1302 |
| Impute 14 | 0.3779 | 0.1336 |
| Impute 15 | 0.3779 | 0.1334 |
| Impute 16 | 0.3744 | 0.1288 |
| Impute 17 | 0.3732 | 0.1289 |
| Impute 18 | 0.3811 | 0.1353 |
| Impute 19 | 0.3768 | 0.1318 |
| Impute 20 | 0.3764 | 0.1314 |
| Impute 21 | 0.3804 | 0.1332 |
| Impute 22 | 0.3709 | 0.1308 |
| Impute 23 | 0.3834 | 0.1355 |
| Impute 24 | 0.3722 | 0.1264 |
| Impute 25 | 0.3798 | 0.1330 |
| Impute 26 | 0.3741 | 0.1304 |
| Impute 27 | 0.3775 | 0.1319 |
| Impute 28 | 0.3743 | 0.1337 |
| Impute 29 | 0.3756 | 0.1326 |
| Impute 30 | 0.3756 | 0.1336 |

Table 2: Average UniFrac and SD per matrix

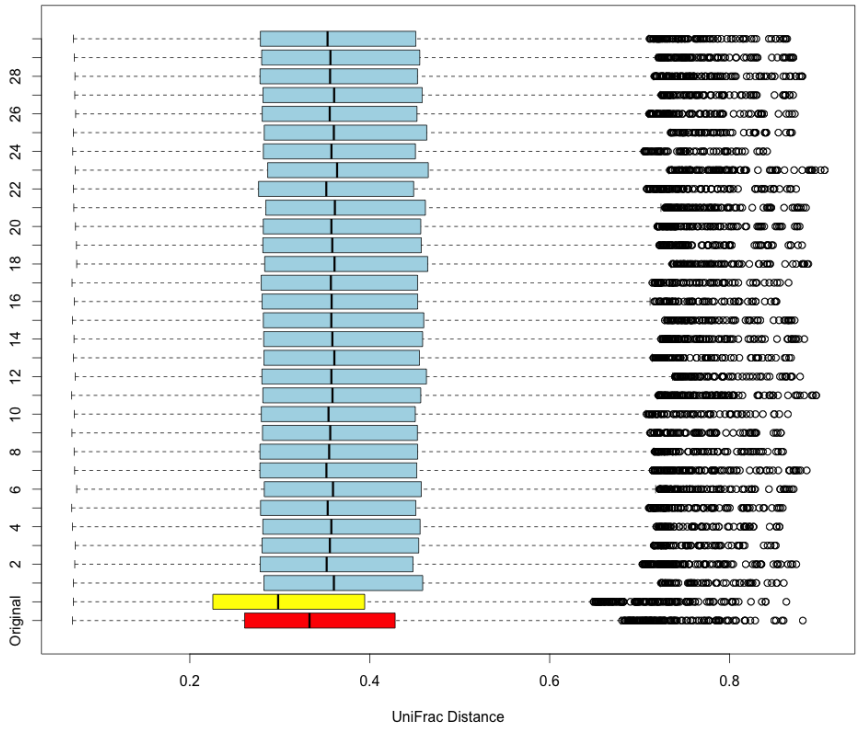
(red). On the contrary, The upper-tail outliers are at a higher range for the matrices created by the imputation method (light blue) compared to the matrix with all the original OTUs. By manually examining and comparing the matrices, the outlier

points for higher or lower UniFrac distance measurements are generally between the same pairwise samples for each matrix.

| Mean of Averages of 30 Iterations | Standard Error |
|---|---|
| 0.3779 | 0.000602 |

Table 3: Statistics of 30 Iterations of Imputed Matrices

Figure 4: Boxplot of UniFrac measure for all matrices



# 4 Discussion

We have found that by excluding all OTUs with unclassified taxa in the HMP data, the average pairwise distance between two samples is smaller than when all OTUs were included in the measurements. This means researchers are interpreting samples to be more similar when deciding to eliminate all OTUs with unclassified taxa from their analyses.

On the other hand, our results yielded an average pairwise distance that is larger than the average pairwise distance with all OTUs included for pairwise matrices produced by our imputation method. The standard error of mean is relatively small. Our method of imputing shows that a bias exists when researchers decide to leave unclassified taxa as unknown in their analyses. Researchers are interpreting samples to be more similar when deciding to keep unclassified OTUs as unknown rather than use the imputation method to redistribute the sequence counts.

In order to determine whether these results are valid, the imputation method should be done to other datasets for future studies with more careful simulations in order to determine if the similar biases are found. Another interesting study would be to use the same sequence, mapping, and OTU files over different programs such as QIIME, R, and Mothur to determine if different microbiome programs yield similar statistical findings.

# 5 Bibliography

[1] Second Genome Solutions. Microbiome to medicine. *Second Genome The Microbiome Company*, 2016.

[2] Ed Yong. An introduction to the microbiome. *National Georgraphic*, 2010.

[3] Genetic Science Learning Center. How we study the microbiome. *http://learn.genetics.utah.edu/content/microbiome/study/*, 2016.

[4] An introduction to applied bioinformatics. *http://readiab.org/book/latest/*.

[5] COLOSS. Exploratory techniques: beta diversity. *http://www.coloss.org/beebook/I/gut-symbionts/2/2/5*.

[6] Unweighted unifrac algorithm. *http://www.mothur.org/wiki/Unweighted_UniFrac_algorithm*.

[7] Weighted unifrac algorithm. *http://www.mothur.org/wiki/Weighted_UniFrac_algorithm*.

[8] Catherine Lozupone , Rob Knight. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 2005.

[9] The human microbiome project. *http://hmpdacc.org/overview/about.php*.

[10] Peter J Turnbaugh , Ruth E Ley, Micah Hamady, Claire Fraser-Liggett, Rob Knight, Jeffrey I Gordon. The human microbiome project: exploring the

microbial part of ourselves in a changing world. *Nature*, 449(7164):804–810, 2007.

[11] What is qiime? *http://qiime.org/index.html*.

# 6 Code

Listing 1: QIIME Code

```
macQIIME
biom convert -i/Users/gchen/Documents/Thesis/otu_table_psn_v13.txt
o/Users/gchen/Documents/Thesis/table.otu.biom --table-type="OTU
_table" --to-json
filter_samples_from_otu_table.py
-i /Users/gchen/Documents/Thesis/table.otu.biom
o/Users/gchen/Documents/Thesis/otu_table_visit1_v13.biom -m
/Users/gchen/Documents/Thesis/v13_map_uniquebyPSN.txt -s
'visitno:1'
biom convert -i
/Users/gchen/Documents/Thesis/otu_table_visit1_v13.biom -o
/Users/gchen/Documents/Thesis/otu_table_visit1.txt --to-tsv
filter_samples_from_otu_table.py -i
/Users/gchen/Documents/Thesis/otu_table_visit1_v13.biom -o
/Users/gchen/Documents/Thesis/otu_table_stoolvisit1_v13.biom -m /Users/gchen/Do
-s 'HMPbodysubsite:Stool'
biom convert -i
/Users/gchen/Documents/Thesis/otu_table_stoolvisit1_v13.biom -o
/Users/gchen/Documents/Thesis/otu_table_stoolvisit1.txt --to-tsv

biom convert -i/Users/gchen/Documents/Thesis/otu_table_v13.txt
o/Users/gchen/Documents/Thesis/seq.tableotu.biom --table-
type="OTU_table" --to-json
filter_samples_from_otu_table.py -i
/Users/gchen/Documents/Thesis/seq.tableotu.biom
o/Users/gchen/Documents/Thesis/seqotu_tablevisit1.biom -m
/Users/gchen/Documents/Thesis/otu2reads_v13.txt -s 'visitno:1'
biom convert -i
/Users/gchen/Documents/Thesis/seqotu_tablevisit1.biom -o
/Users/gchen/Documents/Thesis/seqotu_tablevisit1.txt --to-tsv
filter_samples_from_otu_table.py -i
/Users/gchen/Documents/Thesis/seqotu_tablevisit1.biom -o
/Users/gchen/Documents/Thesis/seqotu_tablestoolvisit1.biom -m
/Users/gchen/Documents/Thesis/v13_map_uniquebyPSN.txt -s
'HMPbodysubsite:Stool'
biom convert -i
/Users/gchen/Documents/Thesis/seqotu_tablestoolvisit1.biom -o
/Users/gchen/Documents/Thesis/seqotu_tablestoolvisit1.txt --to-
tsv

extract_seqs_by_sample_id.py -i
/Users/gchen/Documents/Thesis/seqs_v13.fna -o
/Users/gchen/Documents/Thesis/seqs_subset1.fna -s
```

```
700015181.V13.241719,SRS012191.SRX020675,SRS013543.SRX020677,SRS0
13762.SRX020677,SRS014345.SRX020546,SRS014369.SRX020554,SRS014885
.SRX020554,SRS014999.SRX020577,SRS015190.SRX020554,SRS015247.SRX0
20577,SRS015281.SRX020554,SRS015332.SRX020577,SRS015518.SRX020539
,SRS015578.SRX020577,SRS015599.SRX020539,SRS015663.SRX020539,SRS0
15724.SRX020539,SRS015782.SRX020539,SRS015960.SRX020577,SRS016018
.SRX020539,SRS016056.SRX020536,SRS016095.SRX020516,SRS016203.SRX0
20516,SRS016267.SRX020577,SRS016335.SRX020516,SRS018559.SRX020516
,SRS018655.SRX020572,SRS018712.SRX020572,SRS018733.SRX020516,SRS0
18872.SRX020572,SRS018920.SRX020572,SRS018968.SRX020572,SRS019267
.SRX020516,SRS019381.SRX020572,SRS019534.SRX020572,SRS020470.SRX0
22097,SRS020584.SRX022097,SRS020641.SRX022097,SRS020811.SRX022097
,SRS021065.SRX022097,SRS021109.SRX022097,SRS021853.SRX022232,SRS0
42415.SRX020523,SRS043299.SRX020523,SRS044415.SRX020523,SRS045414
.SRX020563,SRS045607.SRX020523,SRS045613.SRX020523,SRS046313.SRX0
20550,SRS046382.SRX020540,SRS047642.SRX022224,SRS048722.SRX020523
,SRS048838.SRX020523,SRS049823.SRX020523,SRS050733.SRX020550,SRS0
52196.SRX020523,SRS052697.SRX020550,SRS054461.SRX020550,SRS054590
.SRX020550,SRS054608.SRX020523,SRS055482.SRX020550,SRS055563.SRX0
20523,SRS056620.SRX020523,SRS057447.SRX020523,SRS058416.SRX020523
,SRS063307.SRX022232,SRS063324.SRX020519,SRS064321.SRX020519,SRS0
65466.SRX020532,SRS065665.SRX020519,SRS065725.SRX020519
align_seqs.py -i /Users/gchen/Documents/Thesis/seqs_subset1.fna
o /Users/gchen/Documents/Thesis/align/
filter_alignment.py -i
/Users/gchen/Documents/Thesis/align/seqs_subset1_aligned.fasta -o
/Users/gchen/Documents/Thesis/filtered_alignment/
make_phylogeny.py -i
/Users/gchen/Documents/Thesis/filtered_alignment/seqs_subset1_ali
gned_pfiltered.fasta -o /Users/gchen/Documents/Thesis/phylo.tre
beta_diversity.py -i /Users/gchen/Documents/Thesis/seqotu.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i/Users/gchen/Documents/Thesis/OTU_v13_subset.txt
o/Users/gchen/Documents/Thesis/seqotu_subset.biom --table-
type="OTU_table" --to-json

Using representative sequences
biom convert -i/Users/gchen/Documents/Thesis/otu_table_v13.txt
o/Users/gchen/Documents/Thesis/seqotu.biom --table-type="OTU
table" --to-json
align_seqs.py -i /Users/gchen/Documents/Thesis/rep_set_v13.fna
  o   /Users/gchen/Documents/Thesis/align/
filter_alignment.py -i /Users/gchen/Documents/Thesis/seqs_subset1.fna -o
/Users/gchen/Documents/Thesis/filtered_alignment/
filter_alignment.py -i
/Users/gchen/Documents/Thesis/align/rep_set_v13_aligned.fasta -o
```

```
/Users/gchen/Documents/Thesis/filtered_alignment/
make_phylogeny.py -i
/Users/gchen/Documents/Thesis/filtered_alignment/rep_set_v13_alig
ned_pfiltered.fasta -o /Users/gchen/Documents/Thesis/phylo.tre
beta_diversity.py -i
/Users/gchen/Documents/Thesis/seq.tableotu.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre

Making own representative sequence file
pick_rep_set.py -i /Users/gchen/Documents/Thesis/seqmap_v13.txt
f /Users/gchen/Documents/Thesis/seqs_subset1.fna -o
/Users/gchen/Documents/Thesis/create_rep_set1.fna

batch set of beta diversity
beta_diversity.py-i /Users/gchen/Documents/Thesis/OTU_Impute/ -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i /Users/gchen/Documents/Thesis/OTU_Impute/OTU_6.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_6.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_6.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_7.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_7.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_7.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_8.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_8.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_8.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_9.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_9.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_9.biom -m
```

```
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_10.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_10.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_10.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_11.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_11.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_11.biom  m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_12.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_12.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_12.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_13.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_13.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_13.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_14.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_14.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_14.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_15.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_15.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
```

```
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_15.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_16.txt
-o /Users/gchen/Documents/Thesis/OTU_Impute/OTU_16.biom --table-
type="OTU table" --to-json
beta_diversity.py -i /Users/gchen/Documents/Thesis/OTU_Impute/OTU_16.biom
-m weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_17.txt -o
 /Users/gchen/Documents/Thesis/OTU_Impute/OTU_17.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_17.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_18.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_18.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_18.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_19.txt -o
 /Users/gchen/Documents/Thesis/OTU_Impute/OTU_19.biom --table-
type="OTU table" --to-json
beta_diversity.py -i /Users/gchen/Documents/Thesis/OTU_Impute/OTU_19.biom
-m weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_20.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_20.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_20.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_21.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_21.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_21.biom -m
```

```
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_22.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_22.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_22.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_23.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_23.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_23.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_24.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_24.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_24.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_25.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_25.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_25.biom  m
 weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_26.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_26.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_26.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_27.txt  o
 /Users/gchen/Documents/Thesis/OTU_Impute/OTU_27.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
```

```
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_27.biom -m
 weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_28.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_28.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_28.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_29.txt -o
 /Users/gchen/Documents/Thesis/OTU_Impute/OTU_29.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_29.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_30.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_30.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_30.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_31.txt -o
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_31.biom --table-
type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_31.biom -m
weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
 /Users/gchen/Documents/Thesis/phylo.tre
biom convert -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_nounknown.txt -o
 /Users/gchen/Documents/Thesis/OTU_Impute/OTU_nounknown.biom
table-type="OTU table" --to-json
beta_diversity.py -i
/Users/gchen/Documents/Thesis/OTU_Impute/OTU_nounknown.biom -m
 weighted_unifrac -o /Users/gchen/Documents/Thesis/beta_div/ -t
/Users/gchen/Documents/Thesis/phylo.tre
```

Listing 2: R Code

```
small=read.table("/Users/gchen/Documents/Thesis/
otu_table_stoolvisit1.txt",fill=T, header=T)
```

```
otu_root1=data.frame(small,all_roots)
otu_root1[otu_root1==""] <- NA
otu_root1$OTU_ID=as.character(otu_root1$OTU_ID)

otu_root1[is.na(otu_root1)]="NOS"
otu_root1$class[otu_root1$phylum=="NOS" | is.na(otu_root1$phylum)]=NA
otu_root1$order[otu_root1$class=="NOS" | is.na(otu_root1$class)]=NA
otu_root1$family[otu_root1$order=="NOS" | is.na(otu_root1$order)]=NA
otu_root1$genus[otu_root1$family=="NOS" | is.na(otu_root1$family)]=NA

otu_root1=read.table("/Users/gchen/Documents/Thesis/OTU_Impute/
OTU_orig.txt",fill=T, header=T)
otu_root1$OTU_ID=as.character(otu_root1$OTU_ID)
otu_root1$phylum=as.character(otu_root1$phylum)
otu_root1$class=as.character(otu_root1$class)
otu_root1$order=as.character(otu_root1$order)
otu_root1$family=as.character(otu_root1$family)
otu_root1$genus=as.character(otu_root1$genus)
testing=otu_root1

### number of unique taxa and number NOS per level ###
length(unique(testing$phylum))
length(testing[testing$phylum=="NOS",1])
length(unique(testing$class))
length(testing[testing$class=="NOS",1])
length(unique(testing$order))
length(testing[testing$order=="NOS",1])
length(unique(testing$family))
length(testing[testing$family=="NOS",1])
length(unique(testing$genus))
length(testing[testing$genus=="NOS",1])

### start ###
classna=testing[testing$class=="NOS",110]
classnaotu=testing[testing$class=="NOS",1]
classna=classna[!is.na(classna)]
classnaotu=classnaotu[!is.na(classnaotu)]
classnaadd=classna
class.samplefrom=data.frame(classna,classnaadd)

for (i in 1:length(classnaotu))
{
  classnaadd[i]=sample(subset(otu_root1$OTU_ID,
  otu_root1$phylum==classna[i]),1)
  class.samplefrom=data.frame(classnaotu,classnaadd)
  testing[testing$OTU_ID==class.samplefrom[i,2],2:109]=
          testing[testing$OTU_ID==class.samplefrom[i,1],2:109]+
```

```
            testing[testing$OTU_ID==class.samplefrom[i,2],2:109]
  testing1=testing[-c(class.samplefrom[,1]),]
}

orderna=testing1[testing1$order=="NOS",111]
ordernaotu=testing1[testing1$order=="NOS",1]
orderna=orderna[!is.na(orderna)]
ordernaotu=ordernaotu[!is.na(ordernaotu)]
ordernaadd=orderna
order.samplefrom=data.frame(ordernaotu,ordernaadd)

for (i in 1:length(ordernaotu))
{
  ordernaadd[i]=sample(subset(testing1$OTU_ID,testing1$class==
                              orderna[i]),1)
  order.samplefrom=data.frame(ordernaotu,ordernaadd)
   [otu_root1$OTU_ID==order.samplefrom[i,2],2:109]
  testing1[testing1$OTU_ID==order.samplefrom[i,2],2:109]=
    testing1[testing1$OTU_ID==order.samplefrom[i,1],2:109]+testing1
    [testing1$OTU_ID==order.samplefrom[i,2],2:109]
  testing2=testing1[-c(order.samplefrom[,1]),]
}

familyna=testing2[testing2$family=="NOS",112]
familynaotu=testing2[testing2$family=="NOS",1]
familyna=familyna[!is.na(familyna)]
familynaotu=familynaotu[!is.na(familynaotu)]
familynaadd=familyna
family.samplefrom=data.frame(familynaotu,familynaadd)

#for (i in 1:43140)
for (i in 1:length(familynaotu))
{

  familynaadd[i]=sample(subset(testing2$OTU_ID,testing2$order==
                               familyna[i]),1)
  family.samplefrom=data.frame(familynaotu,familynaadd)
  testing2[testing2$OTU_ID==family.samplefrom[i,2],2:109]=
           testing2[testing2$OTU_ID==family.samplefrom[i,1],2:109]+
           testing2[testing2$OTU_ID==family.samplefrom[i,2],2:109]
  testing3=testing2[-c(family.samplefrom[,1]),]
}

genusna=testing3[testing3$genus=="NOS",113]
genusnaotu=testing3[testing3$genus=="NOS",1]
genusna=genusna[!is.na(genusna)]
genusnaotu=genusnaotu[!is.na(genusnaotu)]
```

```
genusnaadd=genusna
genus.samplefrom=data.frame(genusnaotu,genusnaadd)


for (i in 1:length(genusnaotu))
{
  genusnaadd[i]=sample(subset(testing3$OTU_ID,testing3$family==
                              genusna[i]),1)
  genus.samplefrom=data.frame(genusnaotu,genusnaadd)
  testing3[testing3$OTU_ID==genus.samplefrom[i,2],2:109]=
            testing3[testing3$OTU_ID==genus.samplefrom[i,1],2:109]+
            testing3[testing3$OTU_ID==genus.samplefrom[i,2],2:109]
  testing4=testing3[-c(genus.samplefrom[,1]),]
}

testing5=testing4[,1:109]
write.table(testing5, "/Users/gchen/Documents/Thesis/OTU_Impute/
OTU_31.txt", sep="\t", row.names=FALSE,quote=FALSE)

testing_no1= subset(otu_root1,otu_root1$phylum!="NOS")
testing_no2=subset(testing_no1,testing_no1$class!="NOS")
testing_no3=subset(testing_no2,testing_no2$order!="NOS")
testing_no4=subset(testing_no3,testing_no3$family!="NOS")
testing_no5=subset(testing_no4,testing_no4$genus!="NOS")
testing_no5=testing_no5[,1:109]
write.table(testing_no5, "/Users/gchen/Documents/Thesis/OTU_Impute/
OTU_nounknown.txt",sep="\t",row.names=FALSE,quote=FALSE)
### matrices ###
originalmatrix=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_otu_table_stoolvisit1_v13.txt",fill=T, header=T)
originalmatrix[originalmatrix == 0] = NA
orig_avg = mean(unlist(originalmatrix), na.rm=TRUE)

matrix2=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_2.txt",fill=T, header=T)
matrix2[matrix2== 0] = NA
avg2 = mean(unlist(matrix2), na.rm=TRUE)

matrix3=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_3.txt",fill=T, header=T)
matrix3[matrix3== 0] = NA
avg3 = mean(unlist(matrix3), na.rm=TRUE)

matrix4=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_4.txt",fill=T, header=T)
matrix4[matrix4== 0] = NA
avg4 = mean(unlist(matrix4), na.rm=TRUE)
```

```
matrix5=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_5.txt",fill=T, header=T)
matrix5[matrix5== 0] = NA
avg5 = mean(unlist(matrix5), na.rm=TRUE)

matrix6=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_6.txt",fill=T, header=T)
matrix6[matrix6== 0] = NA
avg6 = mean(unlist(matrix6), na.rm=TRUE)

matrix7=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_7.txt",fill=T, header=T)
matrix7[matrix7== 0] = NA
avg7 = mean(unlist(matrix7), na.rm=TRUE)

matrix8=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_8.txt",fill=T, header=T)
matrix8[matrix8== 0] = NA
avg8 = mean(unlist(matrix8), na.rm=TRUE)

matrix9=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_9.txt",fill=T, header=T)
matrix9[matrix9== 0] = NA
avg9 = mean(unlist(matrix9), na.rm=TRUE)

matrix10=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_10.txt",fill=T, header=T)
matrix10[matrix10== 0] = NA
avg10 = mean(unlist(matrix10), na.rm=TRUE)

matrix11=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_11.txt",fill=T, header=T)
matrix11[matrix11== 0] = NA
avg11 = mean(unlist(matrix11), na.rm=TRUE)

matrix12=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_12.txt",fill=T, header=T)
matrix12[matrix12== 0] = NA
avg12 = mean(unlist(matrix12), na.rm=TRUE)

matrix13=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_13.txt",fill=T, header=T)
matrix13[matrix13== 0] = NA
avg13 = mean(unlist(matrix13), na.rm=TRUE)

matrix14=read.table("/Users/gchen/Documents/Thesis/beta_div/
```

```
weighted_unifrac_OTU_14.txt",fill=T, header=T)
matrix14[matrix14== 0] = NA
avg14 = mean(unlist(matrix14), na.rm=TRUE)

matrix15=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_15.txt",fill=T, header=T)
matrix15[matrix15== 0] = NA
avg15 = mean(unlist(matrix15), na.rm=TRUE)

matrix16=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_16.txt",fill=T, header=T)
matrix16[matrix16== 0] = NA
avg16 = mean(unlist(matrix16), na.rm=TRUE)

matrix17=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_17.txt",fill=T, header=T)
matrix17[matrix17== 0] = NA
avg17 = mean(unlist(matrix17), na.rm=TRUE)

matrix18=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_18.txt",fill=T, header=T)
matrix18[matrix18== 0] = NA
avg18 = mean(unlist(matrix18), na.rm=TRUE)

matrix19=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_19.txt",fill=T, header=T)
matrix19[matrix19== 0] = NA
avg19 = mean(unlist(matrix19), na.rm=TRUE)

matrix20=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_20.txt",fill=T, header=T)
matrix20[matrix20== 0] = NA
avg20 = mean(unlist(matrix20), na.rm=TRUE)

matrix21=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_21.txt",fill=T, header=T)
matrix21[matrix21== 0] = NA
avg21 = mean(unlist(matrix21), na.rm=TRUE)

matrix22=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_22.txt",fill=T, header=T)
matrix22[matrix22== 0] = NA
avg22 = mean(unlist(matrix22), na.rm=TRUE)

matrix23=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_23.txt",fill=T, header=T)
matrix23[matrix23== 0] = NA
```

```
avg23 = mean(unlist(matrix23), na.rm=TRUE)

matrix24=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_24.txt",fill=T, header=T)
matrix24[matrix24== 0] = NA
avg24 = mean(unlist(matrix24), na.rm=TRUE)

matrix25=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_25.txt",fill=T, header=T)
matrix25[matrix25== 0] = NA
avg25 = mean(unlist(matrix25), na.rm=TRUE)

matrix26=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_26.txt",fill=T, header=T)
matrix26[matrix26== 0] = NA
avg26 = mean(unlist(matrix26), na.rm=TRUE)

matrix27=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_27.txt",fill=T, header=T)
matrix27[matrix27== 0] = NA
avg27 = mean(unlist(matrix27), na.rm=TRUE)

matrix28=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_28.txt",fill=T, header=T)
matrix28[matrix28== 0] = NA
avg28 = mean(unlist(matrix28), na.rm=TRUE)

matrix29=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_29.txt",fill=T, header=T)
matrix29[matrix29== 0] = NA
avg29 = mean(unlist(matrix29), na.rm=TRUE)

matrix30=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_30.txt",fill=T, header=T)
matrix30[matrix30== 0] = NA
avg30 = mean(unlist(matrix30), na.rm=TRUE)

matrix31=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_31.txt",fill=T, header=T)
matrix31[matrix31== 0] = NA
avg31 = mean(unlist(matrix31), na.rm=TRUE)

unifrac_all=c(avg2,avg3,avg4,avg5,avg6,avg7,avg8,avg9,avg10,avg11,
avg12,avg13,avg14,avg15,avg16,avg17,avg18,avg19,avg20,avg21,avg22,
avg23,avg24,avg25,avg26,avg27,avg28,avg29,avg30,avg31)
avg_30=mean(unifrac_all)
sd_30=sd(unifrac_all)
```

```
matrixunknown=read.table("/Users/gchen/Documents/Thesis/beta_div/
weighted_unifrac_OTU_nounknown.txt", fill=T, header=T)
matrixunknown[matrixunknown==0]=NA
avg_no=mean(unlist(matrixunknown), na.rm=TRUE)

boop2=unlist(matrix2)
twoboop=subset(boop2, boop2 >= 0.6)
boop3=unlist(matrix3)
threeboop=subset(boop3, boop3 >= 0.6)
boop4=unlist(matrix4)
fourboop=subset(boop4, boop4 >= 0.6)
boop5=unlist(matrix5)
fiveboop=subset(boop5, boop5 >= 0.6)
boop6=unlist(matrix6)
sixboop=subset(boop6, boop6 >= 0.6)
boop7=unlist(matrix7)
sevenboop=subset(boop7, boop7 >= 0.6)
boop8=unlist(matrix8)
eightboop=subset(boop8, boop8 >= 0.6)
boop9=unlist(matrix9)
nineboop=subset(boop9, boop9 >= 0.6)
boop10=unlist(matrix10)
tenboop=subset(boop10, boop10 >= 0.6)
boop11=unlist(matrix11)
elevenboop=subset(boop11, boop11 >= 0.6)

boopog=unlist(originalmatrix)
ogboop=subset(boopog, boopog>=0.6)

boopunknown=unlist(matrixunknown)
unknownboop=subset(boopunknown, boopunknown>=0.6)
### boxplot ###
boxplot(unlist(originalmatrix), unlist(matrixunknown), unlist(matrix2),
        unlist(matrix3), unlist(matrix4), unlist(matrix5),
        unlist(matrix6), unlist(matrix7), unlist(matrix8),
        unlist(matrix9), unlist(matrix10), unlist(matrix11),
        unlist(matrix12), unlist(matrix13), unlist(matrix14),
        unlist(matrix15), unlist(matrix16), unlist(matrix17),
        unlist(matrix18), unlist(matrix19), unlist(matrix20),
        unlist(matrix21), unlist(matrix22), unlist(matrix23),
        unlist(matrix24), unlist(matrix25), unlist(matrix26),
        unlist(matrix27), unlist(matrix28), unlist(matrix29),
        unlist(matrix30), unlist(matrix31), horizontal=TRUE,
        names=c("Original","Unknown Deleted","Iteration 1","2","3",
        "4","5","6","7","8","9","10","11","12","13","14","15","16",
        "17","18","19","20","21","22","23","24","25","26","27","28",
```

```
"29","Iteration 30"),col=c("red","yellow","light blue",
"light blue","light blue","light blue","light blue",
"light blue","light blue","light blue","light blue",
"light blue","light blue","light blue","light blue",
"light blue","light blue","light blue","light blue",
"light blue","light blue","light blue","light blue",
"light blue","light blue","light blue","light blue",
"light blue","light blue","light blue","light blue",
"light blue"),xlab="UniFrac Distance")
```