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 [1]. 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 [2]. Since then, papers reporting these metrics have garnered over a thousand citations, and enabled research about everything from how kwashiorkor causes malnutrition [3] to how people can have similar microbiomes to their pet dogs [4]. Except for generalized UniFrac, used to make hybrid unweighted and weighted UniFrac comparisons [5], 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.

0.1 Data

UniFrac requires two pieces of data: A phylogenetic tree and a table of counts per inferred taxa per sample. This data is derived from a gene tag experiment, the most common of which is the 16S rRNA gene tag experiment for microbiome research [CITE]. A gene tag experiment is run by amplifying the gene tag through PCR, and then sequencing the resulting amplicon.

Operational Taxonomic Units

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

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

However, it is difficult to determine how a batch of sequences should be partitioned into groups of 97% identity. One way is to perform a clustering algorithm (using software such as UCLUST [6]) that optimally partitions the groups and then later assign taxonomic identity by matching the sequences with public databases, such as SILVA [7], the Ribosomal Database Project [8], and Greengenes [9]. Another way is to start off with seed sequences from known bacteria and perform the clustering such that the 97% identity groups are centered on the seed sequences. In any case, the resulting taxonomic groupings are known as Operational Taxonomic Units (OTUs), and are used consistently within the same experiment. While OTUs can be annotated with standard taxonomic names such that results can be compared between experiments, technically the taxonomic groupings used by different experiments are not the same.

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

Phylogenetic tree

To generate the phylogenetic tree [10], a representative sequence is taken from each of the cluster of sequences that belong to each OTU. If seed sequences were used to generate the OTU, then these can be used for the phylogenetic tree, otherwise a sequence in the center of the cluster should be selected.

The phylogenetic tree is created through a multiple sequence alignment with the representative OTU sequences, using software such as MUSCLE [11]. Each leaf of the tree represents one of the OTUs, and each of the branches of the tree has a length. Additionally, the tree needs to be rooted for the UniFrac calculation to be performed. This can be done by rooting the tree by its midpoint, which can be performed by software such as the phangorn R package [12].

0.2 Unweighted UniFrac

Unweighted UniFrac [1] uses an inferred evolutionary distance to measure similarity between samples. It requires a reference phylogenetic tree containing all the taxa present in the samples to be examined, plus information about which taxa were detected in each sample. The calculation is performed by dividing the branch lengths 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.

0.3 Weighted UniFrac

Weighted UniFrac [2] is an implementation of the Kantorovich–Rubinstein distance in mathematics, also known as the earth mover's distance [13]. 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
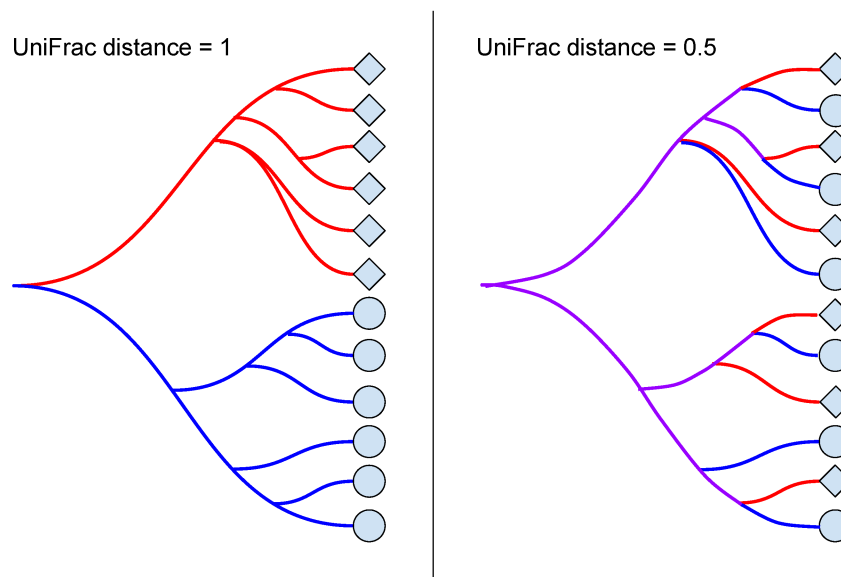
Figure 1. Unweighted UniFrac. When two samples do not share any branches of the phylogenetic tree, the unweighted UniFrac distance is maximized at 1. When two samples share half of their branch lengths on the phylogenetic tree, the unweighted UniFrac distance is 0.5. If the two samples contain exactly the same taxa, the unweighted UniFrac distance is minimized at 0, since the samples share all branches.

especially problematic if the taxa are on a long branch. In weighted UniFrac, low   71
abundance taxa have a much lower weight and so will have a lower impact on the total   72
distance reported by the metric.   73
   UniFrac is constituted as either a presence/absence (unweighted UniFrac) [1], a   74
linear proportion (weighted UniFrac) [2], or some combination of the two (generalized   75
UniFrac) [5]. However, the data are not linear, because the sum of the total number of   76
reads is constrained by the sequencing machinery [14] [15] [16] [17]. Alternative   77
weightings and non-linear transformations of data need to be explored. Furthermore,   78
unweighted UniFrac is known to be unreliable, but it is not generally known or   79
understood how this can impact results.   80

Materials and Methods   81

0.4   Analytical techniques   82

Rarefaction   83
   Rarefaction normalizes the samples OTU counts to a standard sequencing depth [18].   84
This resulting table can be thought of as a random point estimate of the dataset, as the   85
output is a sub-sample of the original table. This standardization process is   86
recommended by the authors of UniFrac [19] in order to account for the sensitivity of   87
UniFrac to sequencing depth.   88
   Rarefactions can be performed using the Qiime software [20] or using the vegan   89
package in R [21].   90

Unweighted UniFrac

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

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

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

Weighted UniFrac

Weighted UniFrac [2] also incorporates each branch length of the phylogenetic tree, and weights them according to proportional abundance of the two samples. The formula for weighed UniFrac is as follows, where A and B are the two samples, b is the set of branch lengths, and $\frac{A_i}{A_T}$ and $\frac{B_i}{B_T}$ are the proportional abundances associated with branch length b_i:

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

Information UniFrac

Information UniFrac is calculated by weighing each branch length by the difference in the uncertainty of the taxa abundance between the two samples. Uncertainty is calculated as follows, where p is the proportional abundance [22]:

$$uncertainty = -p \times log_2(p) \tag{1}$$

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

The formula for Information UniFrac is as follows:

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

Centered Ratio UniFrac

In complex microbiome communities, there are very many bacterial taxa with a low level of counts. Taking the geometric mean of the proportional abundances of taxa in a microbiome sample represents an unbiased baseline [23]. Experiments generally do not have power to detect differences at abundances below the mean [15]. 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. 125
The formula for centered ratio UniFrac is as follows, where gm is the geometric mean: 126

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

Note that the geometric mean is calculated by combining all children in the subtree 127
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 128
proportional abundances separately. The one combined proportional abundance and the 129
remaining single taxa proportional abundances are input into the geometric mean 130
formula, as set a: 131

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

One challenge when it comes to the analysis of read count data is the presence of zero 132
counts. Whether a low-abundance taxa appears in the data as a zero or a low positive 133
count is up to chance, and assuming that a zero count represents the absence of a taxa 134
can be very misleading [15]. A Bayesian approach can be used to estimate the likelihood 135
that a zero could be changed to a positive count if the sample were resequenced, 136
implemented by the cmultRepl command in the zCompositions package in R [24]. 137
The use of this weighting for UniFrac produces measurements that violate the 138
triangle inequality, such that Euclidean statistics are technically invalid. Thus this, like 139
the Bray-Curtis metric, is a dissimilarity, not a distance. 140
For this paper, we calculate UniFrac metrics using a custom R script, which includes 141
unweighted UniFrac, weighted UniFrac, information UniFrac, and centered ratio 142
UniFrac: https://github.com/ruthgrace/exponentUnifrac/blob/master/UniFrac.r 143

Bray-Curtis dissimilarity metric 144
The Bray Curtis dissimilarity metric [25] quantifies how dissimilar two sites are based 145
on counts. A index of 0 means that two samples are identical, while a index of 1 means 146
samples do not share any species. It is computed as a proportion through the formula: 147

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

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

0.5   Data preparation 148

Tongue dorsum data set 149
Samples from this experiment were sourced from the Human Microbiome 150
Project [26] Qiime Community profiling v35 otu tables (`http://hmpdacc.org/HMQCP/`). 151
Analysis of the data was conducted on a Late 2011 15 inch MacBook Pro 2.4 GHz i7 152
with 16GB of 1333 MHz DDR3 RAM. Rarefaction was conducted through Qiime 153
version 1.8.0-20140103, and generation of the ellipse figures was done in R version 3.2.3 154
(2015-12-10) "Wooden Christmas-Tree" x86_64-apple-darwin13.4.0 (64 bit). 155

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

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

where V_{res} = Set of computed PC1s,

V_1 = Reference PC1 (the first),

V_i = Each subsequent PC1,

Tongue dorsum and buccal mucosa data set

Thirty random samples were selected from the tongue site of the Human Microbiome Project [26] 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 [21] before the unweighted UniFrac distance was calculated. Weighted, information, and centered log UniFrac were calculated on the data set without rarefaction. The resulting distances were plotted for principal coordinate analysis.

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

Breast milk data set

The breast milk data set used here has also been published in the Microbiome Journal [27]. The count table was analyzed using our custom UniFrac script. Data was rarefied to the sample with the smallest number of read counts (3072) before the unweighted UniFrac distance matrix was calculated. Non-rarefied data was used for weighted, information, and centered ratio UniFrac. Data was plotted using a principal components or coordinate analysis as appropriate.

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

Results

0.6 Unweighted Unifrac is highly sensitive to rarefaction variants

A commentary by Lozupone et al. 2011 [28] 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
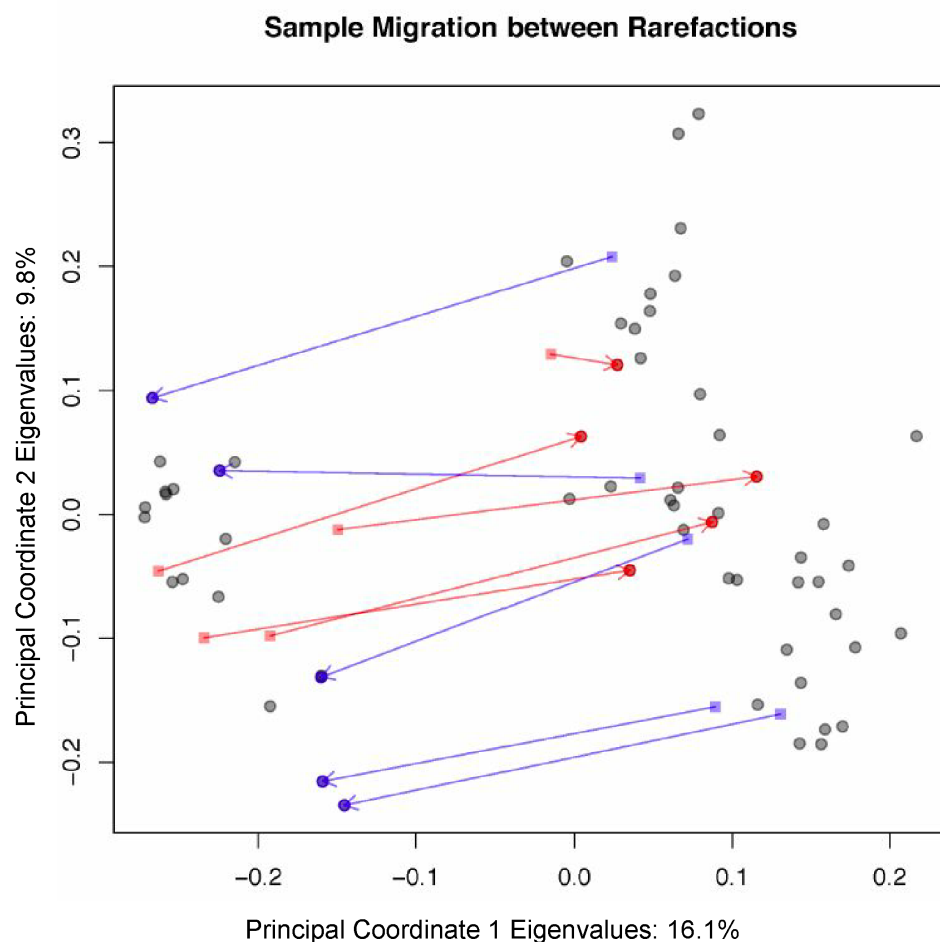
Figure 2. Sample migration in different rarefactions, plotted on principal coordinates, measured with unweighted UniFrac. Red samples have moved from the left cluster to the right cluster between rarefactions. Blue samples have moved from the right cluster to the left. Samples are taken from the tongue dorsum body site from the Human Microbiome Project database.

samples. For Fig. 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, 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

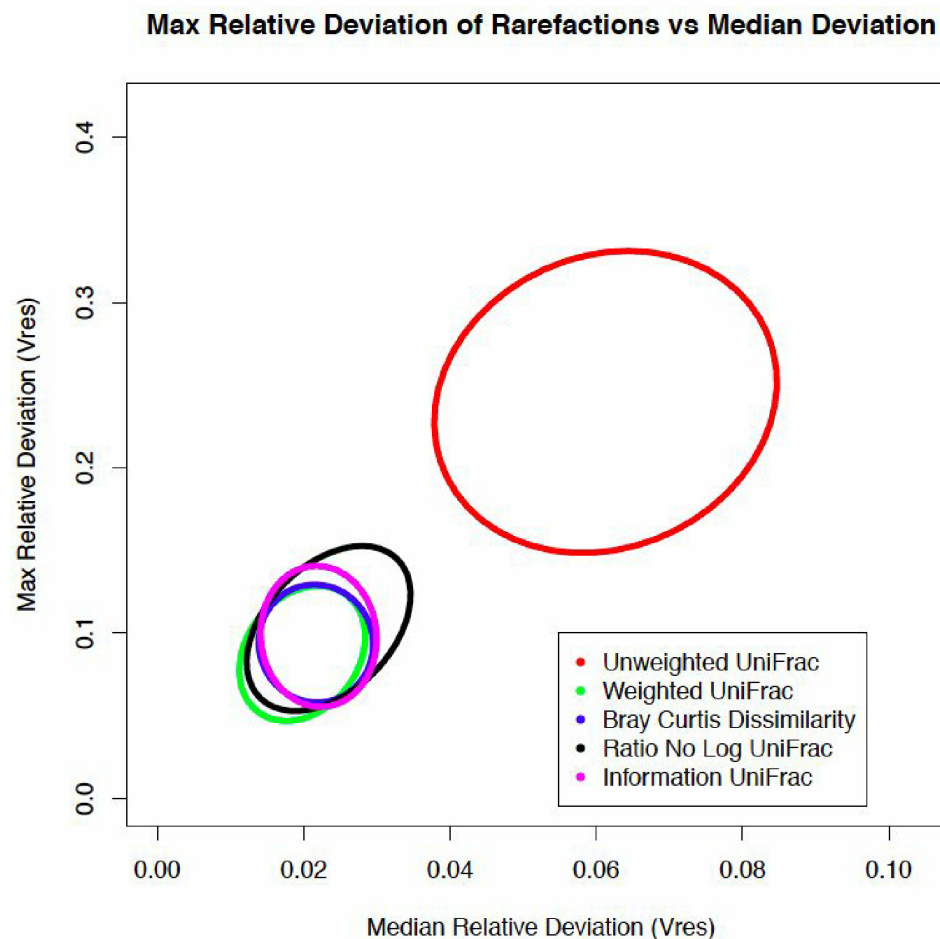appear to be different through the unweighted UniFrac distance metric.   208

Max Relative Deviation of Rarefactions vs Median Deviation

Figure 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 [26], 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.

For the ellipse plot in Fig. 3, 60 tongue dorsum subsamples were randomly drawn   209
without replacement and the gene compositions per sample were also filtered to a   210
minimum of 100. A hundred separate rarefactions were conducted on the data to a   211
minimum sampling depth of 378. For each individual rarefied OTU table, a distance   212
matrix was computed using the unweighted Unifrac, weighted UniFrac, Bray-Curtis   213
Dissimilarity, information UniFrac, and centered ratio UniFrac as a weighting method.   214
By generating 100 separate datasets for each metric, it is possible to assess the   215
magnitude of difference each metric has by analyzing what is essentially the same data.   216
In other words, what does the effect of random sampling (rarefaction) have on the   217

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 [29] and Nature [30].

The maximum value of Vres for each rarefaction is plotted against the median value per rarefaction in Fig. 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.

0.7 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 [1]. 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 [31]. 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.

Table 1. Original abundance of taxa and rarefied abundance of taxa.

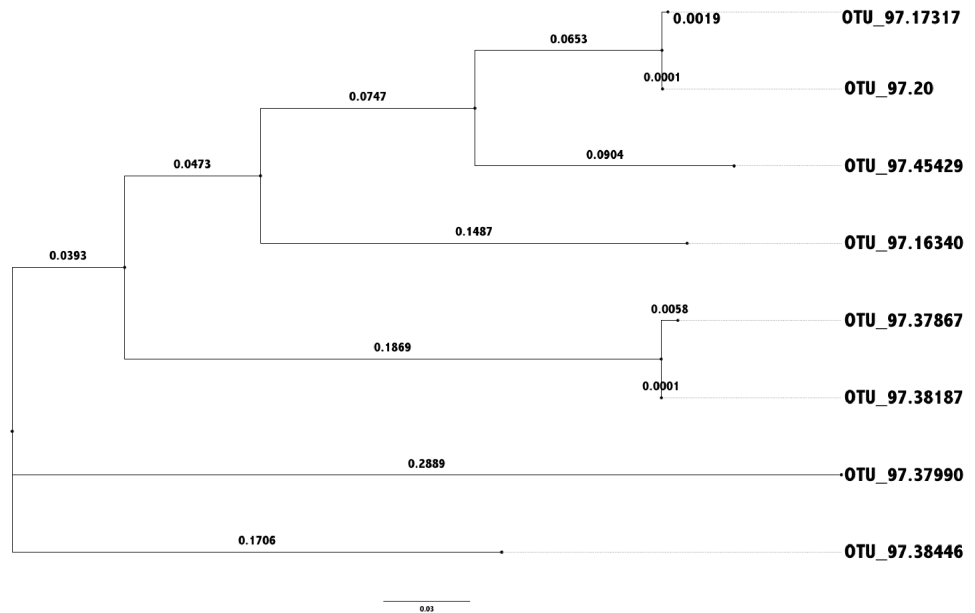| OTU.ID	A	B	A R1	B R1	A R2	B R2
OTU.16340	52	1	8	1	12	1
OTU.17317	17	4	3	4	5	4
OTU.20	70	18	14	18	20	18
OTU.37867	59	10	9	10	11	10
OTU.37990	7	59	0	59	1	59
OTU.38187	646	115	132	115	122	115
OTU.38446	6	8	0	8	1	8
OTU.45429	218	6	55	6	49	6

Figure 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.

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

$$Distance_{A:B} = \frac{\sum UnsharedBranches}{\sum TotalBranches}$$

$$= \frac{(0.2889 + 0.1706)}{1.12}$$

$$= \frac{0.5281}{1.12}$$

$$= 0.4175$$

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

$$Distance_{A:B} = \frac{\sum UnsharedBranches}{\sum TotalBranches}$$

$$= \frac{0}{1.12}$$

$$= 0$$

With rare genes and long branch lengths in the phylogenetic tree (Fig. 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 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 ⁣⁣254
doubled in proportion between samples, and this is much more biologically significant ⁣⁣255
than the change in proportional abundance in the second taxa. Additionally, it does not ⁣⁣256
account for how the counts add up to a constrained sum determined by the sequencing ⁣⁣257
machine model. Because the sum is constrained, as an example, an increase in growth ⁣⁣258
of one taxa can make the data look like there is a decrease in abundance in other taxa, ⁣⁣259
even if in reality the population of the other taxa stayed the same. ⁣⁣260

Here we explore some alternatives to unweighted and weighted UniFrac, and discuss ⁣⁣261
their merits and shortfalls. ⁣⁣262

0.8   Information UniFrac ⁣⁣263

The difference in information content between low proportional abundances (which ⁣⁣264
make up the bulk of microbiome data) is generally higher than the difference between ⁣⁣265
the proportional abundances themselves, potentially allowing scientists to differentiate ⁣⁣266
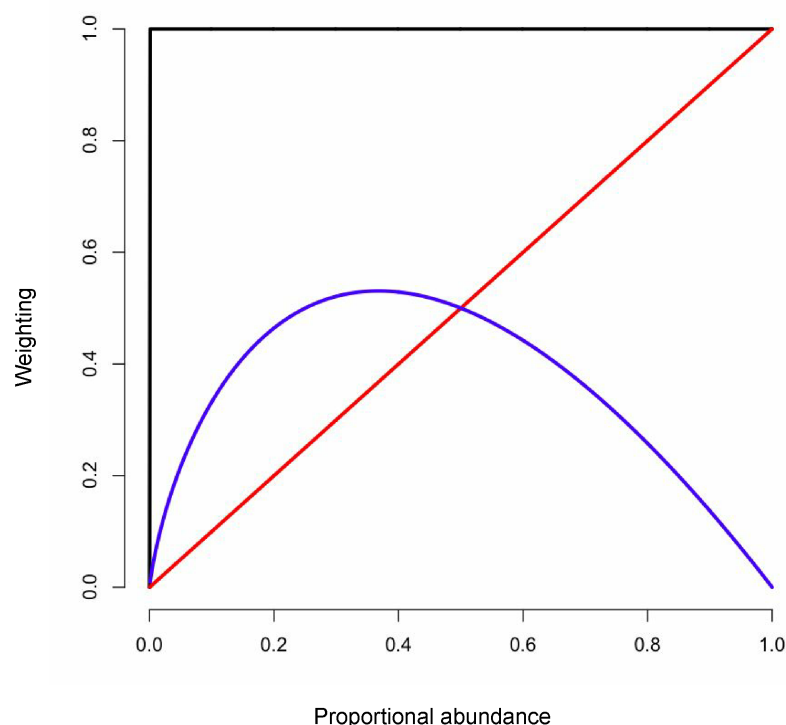groups with subtle differences. ⁣⁣267

Figure 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.

Near the 0, 0 point in Fig. 5, the proportional abundances are low. Here there is ⁣⁣268
higher differentiation between weights of different pairs of low proportional abundances ⁣⁣269
for information UniFrac, as shown by the higher slope of the curved graph. The ⁣⁣270
centered ratio UniFrac (not depicted) depends on the geometric mean of the taxonomic ⁣⁣271
abundances, and would have a different slope in the weight graph depending on how ⁣⁣272
evenly the abundances were distributed. ⁣⁣273

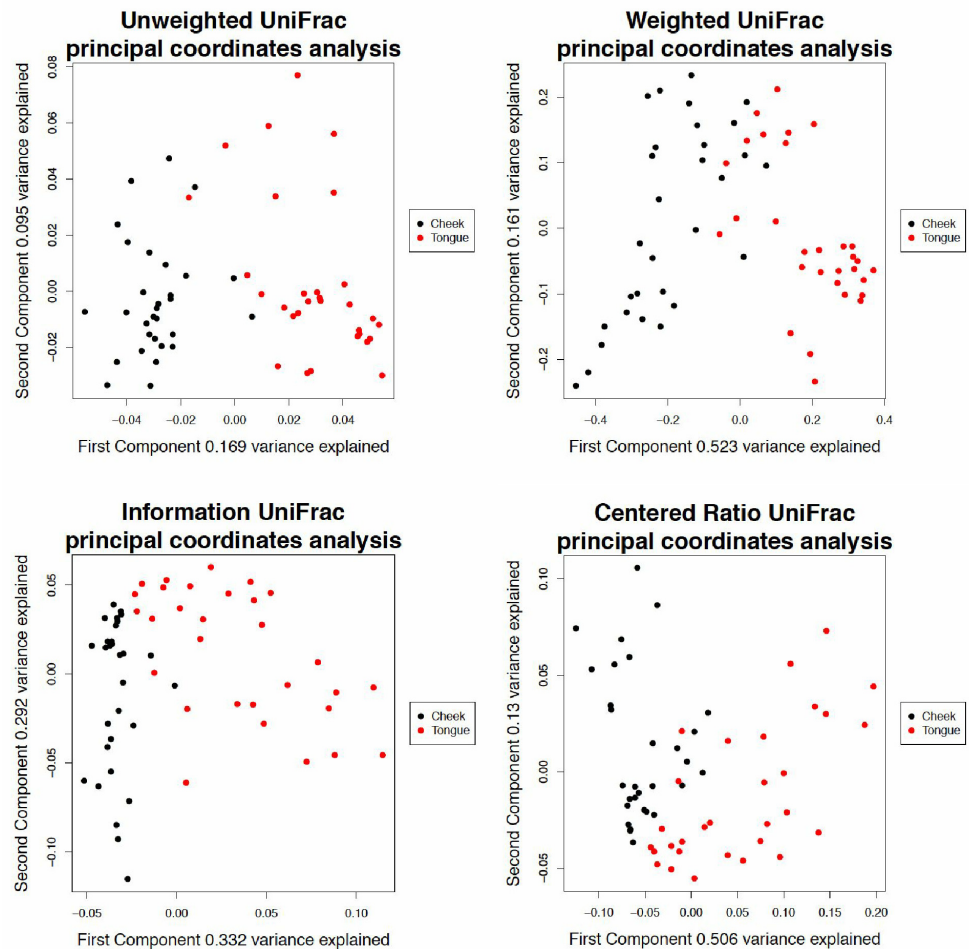## 0.9 Tongue and cheek comparison

Figure 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 [26]. All weightings show separation between the samples by body site.

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. 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. 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. 3), while reliably separating samples in data with a defined difference between groups.
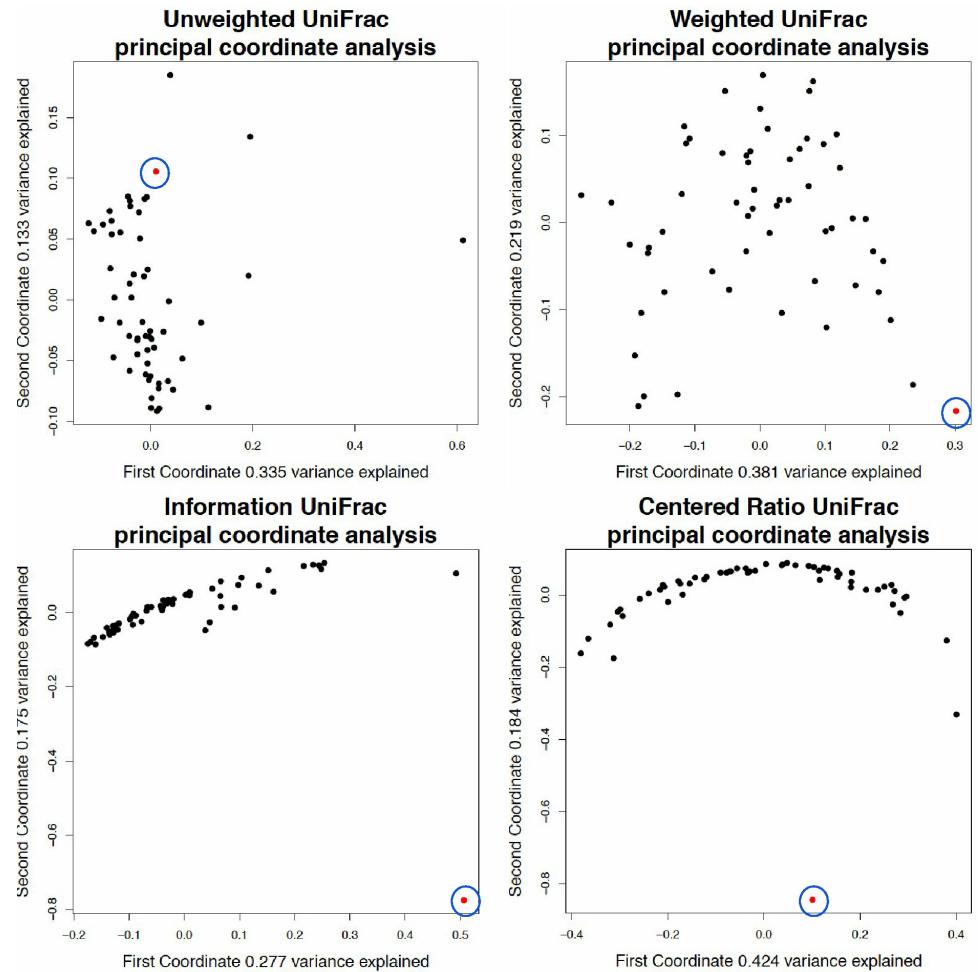
Figure 7. Analysis of breast milk data using different UniFrac weightings.
A principal coordinate analysis of a 16S rRNA experiment done on samples from a 16S rRNA experiment on breast milk. The circled sample is infected with 97% *Pseudomonas*, compared to 15-20% in the other samples.

0.10   Breast milk Data

Fig. 7 is a principal coordinate analysis of a 16S rRNA gene sequencing experiment done on microbiome samples from breast milk [27]. 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

References

1. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology. 2005;71(12):8228–8235.

2. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Applied and environmental microbiology. 2007;73(5):1576–1585.

3. Smith MI, Yatsunenko T, Manary MJ, Trehan I, Mkakosya R, Cheng J, et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. Science. 2013;339(6119):548–554.

4. Song SJ, Lauber C, Costello EK, Lozupone CA, Humphrey G, Berg-Lyons D, et al. Cohabiting family members share microbiota with one another and with their dogs. Elife. 2013;2:e00458.

5. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics. 2012;28(16):2106–2113.

6. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–2461.

7. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research. 2013;41(D1):D590–D596.

8. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic acids research. 2009;37(suppl 1):D141–D145.

9. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology. 2006;72(7):5069–5072.

10. Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. science. 2006;311(5765):1283–1287.

11. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792–1797.

12. Schliep KP. phangorn: Phylogenetic analysis in R. Bioinformatics. 2011;27(4):592–593.

13. Evans SN, Matsen FA. The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2012;74(3):569–592.

14. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8(9):e1002687.

15. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One. 2013;8(7):e67019.

16. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014;2(1):1.

17. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol. 2015;11(3):e1004075.

18. Simberloff D. Use of rarefaction and related methods in ecology. In: Biological data in water pollution assessment: quantitative and statistical analyses. ASTM International; 1978. .

19. de Cárcer DA, Denman SE, McSweeney C, Morrison M. Evaluation of subsampling-based normalization strategies for tagged high-throughput sequencing data sets from gut microbiomes. Applied and environmental microbiology. 2011;77(24):8795–8798.

20. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010;7(5):335–336.

21. Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ, et al. The vegan package. Community ecology package. 2007;p. 631–637.

22. Shannon CE. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review. 2001;5(1):3–55.

23. Aitchison J. The statistical analysis of compositional data. Journal of the Royal Statistical Society Series B (Methodological). 1982;p. 139–177.

24. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems. 2015;143:85–96.

25. Beals EW. Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. Advances in Ecological Research. 1984;14(1):55.

26. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett C, Knight R, Gordon JI. The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature. 2007;449(7164):804.

27. Urbaniak C, Angelini M, Gloor GB, Reid G. Human milk microbiota profiles in relation to birthing method, gestation and infant gender. Microbiome. 2016;4(1):1–9.

28. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. The ISME journal. 2011;5(2):169.

29. Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell. 2013;155(7):1451–1463.

30. Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. Diet-induced extinctions in the gut microbiota compound over generations. Nature. 2016;529(7585):212–215.

31. Gorzelak MA, Gill SK, Tasnim N, Ahmadi-Vand Z, Jay M, Gibson DL. Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool. PloS one. 2015;10(8):e0134802.