

Expanding the UniFrac toolbox

Ruth G Wong¹ , Jia R Wu¹ , Gregory B Gloor¹ 

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

 **These authors contributed equally to this work.**

* gloor@uwo.ca

Abstract

~~Microbiome analysis is frequently performed using the UniFrac distance metric is often used to separate groups in microbiome analysis, but requires a constant sequencing depth to work properly.~~ Here we demonstrate that unweighted UniFrac is highly sensitive to rarefaction instance and to sequencing depth in uniform data sets ~~with no clear structure or separation between groups~~. We show that this arises because of subcompositional effects. We introduce information UniFrac and ~~entered~~-ratio UniFrac, two new weightings that are not ~~as~~ sensitive to rarefaction and allow greater separation of outliers than classic unweighted and weighted UniFrac. With this expansion of the UniFrac toolbox, we hope to empower researchers to extract more varied information from their data.

Introduction

In 2005, Lozupone et al introduced the UniFrac distance metric, a measure to calculate the difference between microbiomes that incorporated phylogenetic distance [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 data sets where~~ UniFrac gives misleading results, and ~~present and~~ discuss some alternative weightings for UniFrac.

Operational Taxonomic Units

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

~~Historically bacteria that are have 97% identity in a variable region are considered to be the same taxa [6]. The 97% cutoff was arbitrarily chosen to best map sequence data to bacterial classifications. This threshold is thought to maximizes the grouping of bacteria classified as the same species while minimizing the grouping of bacteria classified as different species [7]. 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 [8]) that partitions the groups and then later assign taxonomic identity by matching the seed or central sequences with public databases, such as SILVA [9], the Ribosomal Database Project [10], or Greengenes [11]. Another method is closed reference OTU picking, which starts off with seed sequences from known bacteria and perform the clustering such that the 97% identity groups are centered on the seed sequences. In any case, the resulting taxonomic groupings are known as Operational Taxonomic Units (OTUs), and are used consistently within the same experiment. While OTUs can be annotated with standard taxonomic names such that results can be compared between experiments, technically the taxonomic groupings used by different experiments are not the same, except with closed reference OTUs, or individual sequence unit methods. Individual sequence unit (ISU) methods which do not use OTUs can be run with software such as DADA2 [12].

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

1 Data

UniFrac requires two pieces of information: phylogenetic tree and a table of counts per inferred taxa per sample. These are derived from a gene tag sequencing experiment, such as the commonly used 16S rRNA gene [13]. The sequenced gene contains a variable region, allowing the sequences to be grouped into OTUs as described in the previous section. A count table can then be generated with the number of reads per OTU per sample. The center sequence of each OTU group can be put into a multiple sequence alignment, from which a phylogenetic tree can be inferred.

The phylogenetic tree is created through a multiple sequence alignment with the representative OTU sequences, using software such as MUSCLE [14], or using a guide tree, such as through Greengenes [11] or the QIIME software [15]. Each leaf of the tree represents one of the OTUs, and each of the branches of the tree has a length. Additionally, the tree needs to be rooted for the UniFrac calculation to be performed. This is often done by rooting the tree at its midpoint.

2 Compositional Data Analysis

Microbiome data is in the form of a list of counts per feature (OTUs in this case), with the features composing an aspect of the microbiome for each sample. This is compositional data. There are several core truths about microbiome data and its compositional nature that should be considered when making an analysis strategy.

First, the total number of reads per sample is influenced by sample collection, extraction, sequencing library preparation, and sequencing platform, and is irrelevant to the biological implications of the data. Additionally, the constraint of the count total causes the abundance of different taxa to appear to be negatively correlated with each other when analyzed by conventional statistics. When one taxa increases in abundance, the counts detected in other taxa decrease in abundance, even if the taxa are not decreasing in abundance biologically. For example, one study compared the microbiome of vaginal swab samples from women with bacterial vaginosis (BV), women without BV, and women with intermediate BV, using qPCR to quantify the taxa. *Prevotella* was found to increase through non-BV to intermediate to BV, while *Lactobacillus iners* stayed relatively the same. If the same samples were put through a

gene tag sequencing experiment where the taxa could not be quantified and the total read counts were constrained, one might incorrectly conclude that the abundance *Lactobacillus iners* was decreasing while *Prevotella* was increasing.

To prevent incorrect conclusions, data should be analyzed in a compositional way. In Euclidean space, data points can increase or decrease freely. Compositional data is under a sum constraint, and exist in a non-Euclidean space known as the Aitchison simplex [16]. A data transformation can be performed to put the data into Euclidean space, so that it can be analyzed with standard statistical methods that depend on Cartesian coordinates and linear relationships. These transformations involve examining the ratios of different OTU abundances to each other, so that the total number of reads do not unduly affect the result. In the example with bacterial vaginosis, using ratios of taxa to each other would elucidate the nature of the biological change in the data.

3 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 that are not shared between the two samples by the branch lengths covered by either sample. Figure 1 shows example calcualtions for UniFrac based on the tree overlap. A distance of 0 means that the samples are identical, and a distance of 1 means that the two samples share no taxa in common.

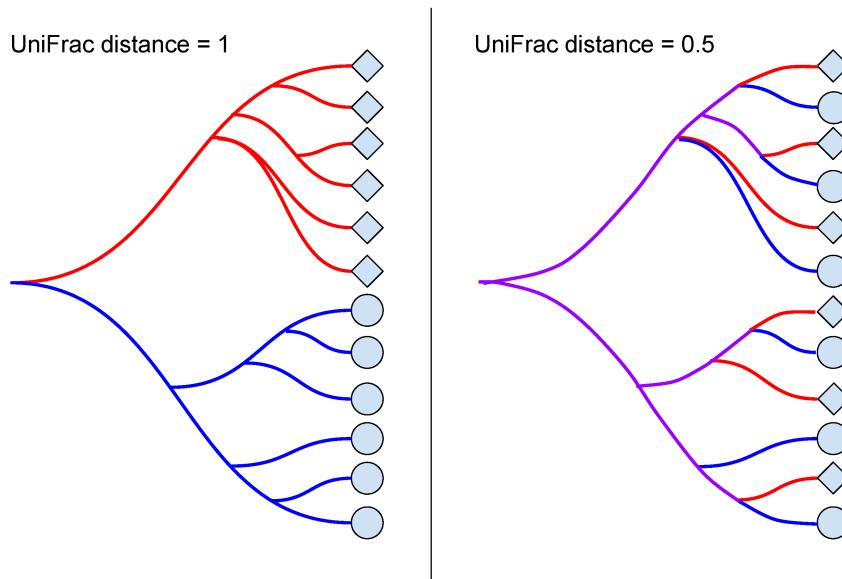


Figure 1. Unweighted UniFrac. When two samples do not share any branches of the phylogenetic tree, the unweighted UniFrac distance is maximized at 1. When two samples share half of their branch lengths on the phylogenetic tree, the unweighted UniFrac distance is 0.5. If the two samples contain exactly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

As UniFrac is a binary test of absence, it is sensitive to sequencing depth, and assumes that the data has been normalized to a common sequencing depth [17], and rarefaction prior to unweighted UniFrac has become a standard part of the microbiome analysis workflow, with built in rarefaction functions in QIIME [15] and mothur [18].

4 Weighted UniFrac

Weighted UniFrac [2] is an implementation of the Kantorovich–Rubinstein distance in mathematics, also known as the earth mover’s distance [19]. Rather than looking only at the presence or absence of taxa, each branch length of the phylogenetic tree is weighted by the difference in proportional abundance of the taxa between the two samples.

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

UniFrac is constituted as either a presence/absence-binary weighting (unweighted UniFrac) [1], a linear proportion (weighted UniFrac) [2], or some combination of the two (generalized UniFrac) [5]. However, it is a misconception that the data are not linear-linear because the sum of the total number of reads is constrained by the sequencing machinery [20] [21] [22] [23]. Microbiome communities can exhibit tremendous variation in their total bacterial count. For example, a stool sample may produce more highly concentrated DNA extract than a skin swab sample, resulting in different read count totals. Vaginal samples from patients with bacterial vaginosis compared to patients without can have total counts that differ one magnitude [24]. Alternative weightings and non-linear transformations of data need to be explored. Furthermore, unweighted UniFrac is known to be unreliable, but it is not generally known or understood how this can impact results.

Materials and Methods

5 Analytical techniques

Rarefaction

Rarefaction normalizes the samples OTU counts to a standard sequencing depth [25]. This resulting table can be thought of as a random point estimate of the dataset, as the output is a sub-sample without replacement of the original table. This standardization process is recommended by the authors of UniFrac [26] in order to account for the sensitivity of UniFrac to sequencing depth.

Rarefactions can be performed using the QIIME-QIIME software [15] or using the vegan package in R [27].

Unweighted UniFrac

Unweighted UniFrac is calculated based on the presence or absence of counts for each branch in the phylogenetic tree, when comparing two samples. A branch is unshared when one sample has a non-zero abundance but not the other, and a branch is shared when both samples belongs to a sample when at least one of the OTUs in the leaves below it have a non-zero abundance. The formula for unweighted UniFrac is as follows, where b is the set of branch lengths in the phylogenetic tree, A and B

represent the two samples being compared, Δ is the symmetric difference between two sets, and \cup is the union between two sets:

$$\frac{\sum b_{unshared}}{\sum b_{unshared} + \sum b_{shared}} Unweighted_{AB} = \frac{\sum b_A \Delta b_B}{\sum b_A \cup b_B}$$

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 information is calculated as follows, where p is the proportional abundance [28]:

$$information = -p \times log_2(p) \quad (1)$$

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

The formula for Information UniFrac is as follows:

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

Information UniFrac approaches a minimum of zero (Fig. 5) when a sample is composed of a monoculture. It also related to the Aitchison distance in compositional data analysis [29].

Centered-Ratio UniFrac

In complex microbiome communities, there are very many may be a large number of bacterial taxa with a low level of counts few counts, such that the data is sparse. Taking the geometric mean of the proportional abundances of taxa in a microbiome sample represents an unbiased baseline of the average abundance of features with

geometric growth characteristics - such as bacteria which divide by fission [16].
 Experiments generally do not have power to detect differences at abundances below the
 mean [21]. Centering the proportional abundances around the geometric mean thus
 allows one to examine the data in this context, muting differences that are close to
 baseline the baseline abundance and accentuating outliers. The formula for eentered
 ratio UniFrac is as follows, where gm is the geometric mean:
 170
 171
 172
 173
 174
 175

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

Note that the geometric mean is calculated by combining all children in the subtree
 of b_i into $\frac{A_i}{A_T}$ for sample A or $\frac{B_i}{B_T}$ for sample B, and including the rest of the single taxa
 proportional abundances separately. The one combined proportional abundance and the
 remaining single taxa proportional abundances are input into the geometric mean
 formula, as set a:
 176
 177
 178
 179
 180

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

One challenge when it comes to the analysis of read count data is the presence of
 zero counts that the data is very sparse. Whether a low-abundance taxa or feature
 appears in the data as a zero or a low positive count is up to chance, and assuming that
 a zero count represents the absence of a taxa can be very misleading [21]. A Bayesian
 approach can be used to estimate the likelihood that a zero could be changed to a
 positive count if the sample were resequenced, give a posterior estimate of the
 likelihood for zero: this is implemented by the cmultRepl command in the
 zCompositions package in R [30].
 181
 182
 183
 184
 185
 186
 187
 188
 189
 190
 191

The use of this ratio weighting for UniFrac produces measurements that violate the
 metric triangle inequality, such that Euclidean statistics are technically invalid. Thus
 this metric, like the Bray-Curtis metric, is a dissimilarity, not a distance.
 192
 193
 194
 195

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

Bray-Curtis dissimilarity metric

The Bray Curtis dissimilarity metric [31] quantifies how dissimilar two sites are
 based on counts. A Bray-Curtis index of 0 means that two samples are identical, while
 a Bray-Curtis index of 1 means samples do not share any species. It is computed as a
 proportion through the formula:
 196
 197
 198
 199
 200

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

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

S_i = Specimen counts at site i

S_j = Specimen counts at site j

6 Data preparation

The data used comes in the form of a table of counts per operational taxonomic unit per sample, plus a phylogenetic tree. All of our data are derived from 16S rRNA gene tag sequencing experiments, and the data and scripts can be accessed at https://github.com/JRWu/R_Scripts.

Tongue dorsum data set

The tongue dorsum data set is a collection of 60 microbiome samples taken from the tongues of healthy participants. There were 0.3 million reads across 554 OTUs, and a minimum and maximum of 659 and 17176 reads per sample.

Samples from this experiment were sourced from the Human Microbiome Project [32] Qiime Community profiling v35 otu tables (<http://hmpdacc.org/HMQCP/>). Analysis of the data was conducted on a Late 2011 15 inch MacBook Pro 2.4 GHz i7 with 16GB of 1333 MHz DDR3 RAM.

Rarefaction was conducted through Qiime version 1.8.0-20140103 to 659 reads (the lowest number of reads for a sample), and generation of the ellipse figures was done in R version 3.2.3 (2015-12-10) "Wooden Christmas-Tree" x86_64-apple-darwin13.4.0 (64 bit).

A principal coordinate analysis is drawn from each distance matrix per metric, and for the first principal coordinate of each metric, V_{res} —the resultant value (V_{res}) is computed per each first principal coordinate as defined by the formula:

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

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

The tongue dorsum and buccal mucosa data set is a collection of 30 microbiome samples taken from the tongues of healthy participants, plus 30 microbiome samples taken from the buccal mucosa (cheek) of a different set of healthy participants. There were 0.4 million reads across 12701 OTUs, and a minimum and maximum of 5028 and 9861 reads per sample. Note that if the OTUs that are less than 1% abundant in all samples are filtered out, only 179 OTUs remain.

Thirty To create this data set, thirty random samples were selected from the tongue site of the Human Microbiome Project [32] and 30-thirty random samples from the buccal mucosa site. Samples were filtered so that only samples with 5000 to 10,000 reads were included.

Read counts from the HMP data set were rarefied to the smallest total read count per sample using the vegan R package [27] before the unweighted UniFrac distance was calculated. Weighted, information, and eentered-log-ratio 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 .

Breast milk data set

The breast milk data set is a collection of 58 microbiome samples taken from lactating Caucasian Canadian women. The breast milk data set used here has also

been published in [the Microbiome Journal](#) [33], a recent study [33]. There were a total of 5.3 million reads across 115 OTUs, and a minimum and maximum of 3072 and 2.8 million reads per sample. Note that the 2.8 million reads came from a sample that was taken from a patient with an infection, and the next largest number of reads per sample was 282485 (ten times less).

The count table was analyzed using our custom UniFrac script, [which can be accessed at](#) https://github.com/ruthgrace/ruth_unifrac_workshop. Data was rarefied to the sample with the smallest number of read counts (3072) before the unweighted UniFrac distance matrix was calculated. Non-rarefied data was used for weighted, information, and [entered-ratio](#) UniFrac. Data was plotted using a principal components coordinates or coordinate analysis as appropriate.

[The script used to run this analysis can be referenced at](#).

Monoculture data set

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

Results

7 Unweighted Unifrac is highly sensitive to rarefaction variantsinstance

A commentary by Lozupone et al. 2011 [17] addressed the sensitivity of Unweighted UniFrac to sampling. [They utilized Lozupone's group used](#) mean UniFrac values to compute a confidence ellipse [between the first and third quartile](#). However, we observed that this approach under-represented the true variability of unweighted UniFrac as a distance metric by highlighting how individual samples vary. In the absence of true differences and in the presence of uneven sampling, unweighted UniFrac can be sensitive to rarefaction variantsinstances. 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](#) [Rare OTUs with less than](#) 100 [counts across samples](#) [total counts across all the samples were removed](#). The minimum sample count for the subset of 60 we analyzed was 659, therefore we rarefied (subsampled) to the minimum of 659 to normalize the samples, prior to performing a principal coordinates analysis (PCoA). For Fig. 2, two independent rarefactions of the data were conducted in order to observe the effect of rarefaction variants on the metricsinstance on the metric. The unweighted UniFrac distance was then computed for each rarefaction, and Procrustes adjustment was applied in order to overlay the PCoA-derived second rarefaction onto the first. A PCA PCoA of rarefaction 1 was plotted, and any samples that changed between rarefactions one and two were visualized with red and blue on the plot. If the sample moved from one cluster to another between the rarefactions side of the first coordinate axis to the other between the rarefaction instances, it was indicated with either a blue or a red arrow.

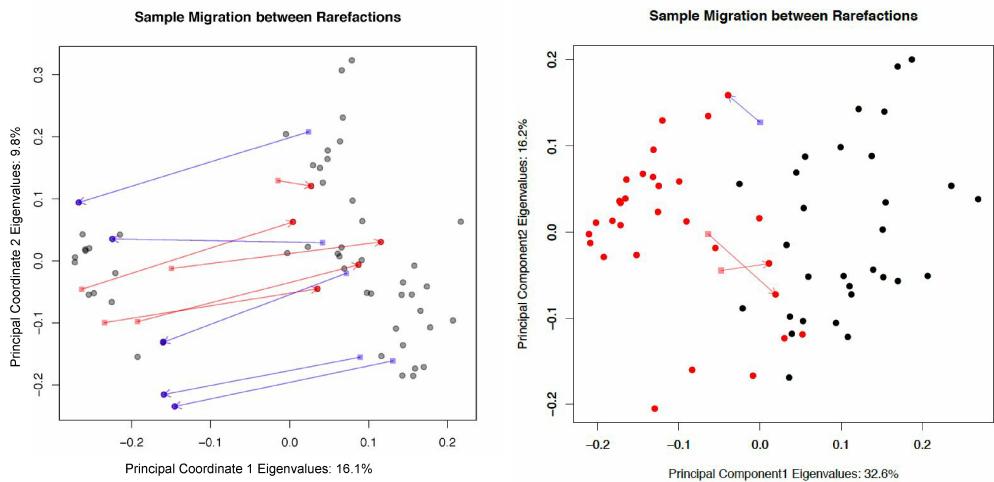


Figure 2. Sample migration in different rarefactions, plotted on principal coordinates, measured with unweighted UniFrac. Red The left plot is of the tongue data set while the right plot is the tongue dorsum vs. buccal mucosa data set. On the left panel red samples have moved from the left cluster to the right cluster between rarefactions. Blue samples have moved from the right cluster to the left. Samples are taken from the tongue dorsum body site from the Human Microbiome Project database. If the experiment were run once, one might mistakenly assume that there are two clusters of data, however, the inconsistent sample membership of the two groups between rarefactions proves the clustering irreproducible. The tongue dorsum and buccal mucosa data set is included for comparison, with the tongue samples colored black and the buccal mucosa samples colored red. Note that the variance explained in the tongue data set by the first and second coordinate is merely 16.1% and 9.8% respectively, indicating that the data is rather spherical, even though the points on the plot appear to show two separated clusters (compare with 32.6% and 16.2% in the tongue dorsum vs. buccal mucosa data set). The variance explained in the first and second coordinate in the 2011 UniFrac commentary [17] was even smaller, at 8.6% and 5.6%.

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 several samples are observed moving between the various clusters to be unstable, exhibiting large differences in location. This example demonstrates that samples with little difference can appear to be different through the unweighted UniFrac distance metric and that rarefaction can lead to misleading and non-reproducible results.

For the ellipse plot in Fig. 3, 60 tongue dorsum subsamples were randomly drawn without replacement and the gene compositions per sample were also filtered to a minimum of 100. Rare OTUs with less than 100 total counts across all samples were removed. A hundred separate rarefactions were conducted on the data to a minimum sampling depth of 378. For each individual rarefied OTU table, a distance matrix was computed using the one of unweighted UniFrac, weighted UniFrac, Bray-Curtis Dissimilarity, information UniFrac, and centered or ratio UniFrac as a the weighting method. By generating 100 separate datasets for each metric, it is possible to assess the

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

~~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 [34] and Nature [35].~~

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 ~~entered~~-ratio UniFrac, as well as Bray-Curtis.

8 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). Given the wide use of unweighted UniFrac in the literature with small principal coordinate 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 [34], where the first and second principal coordinates axis explain 14% and 9.5% of the variation in Figure 2A, as well as in Nature [35], where the first principal coordinate explains 14% of the variation in Figure 1. In both of these examples, less variance is explained by the first principal coordinate than in our uniform tongue data set.~~

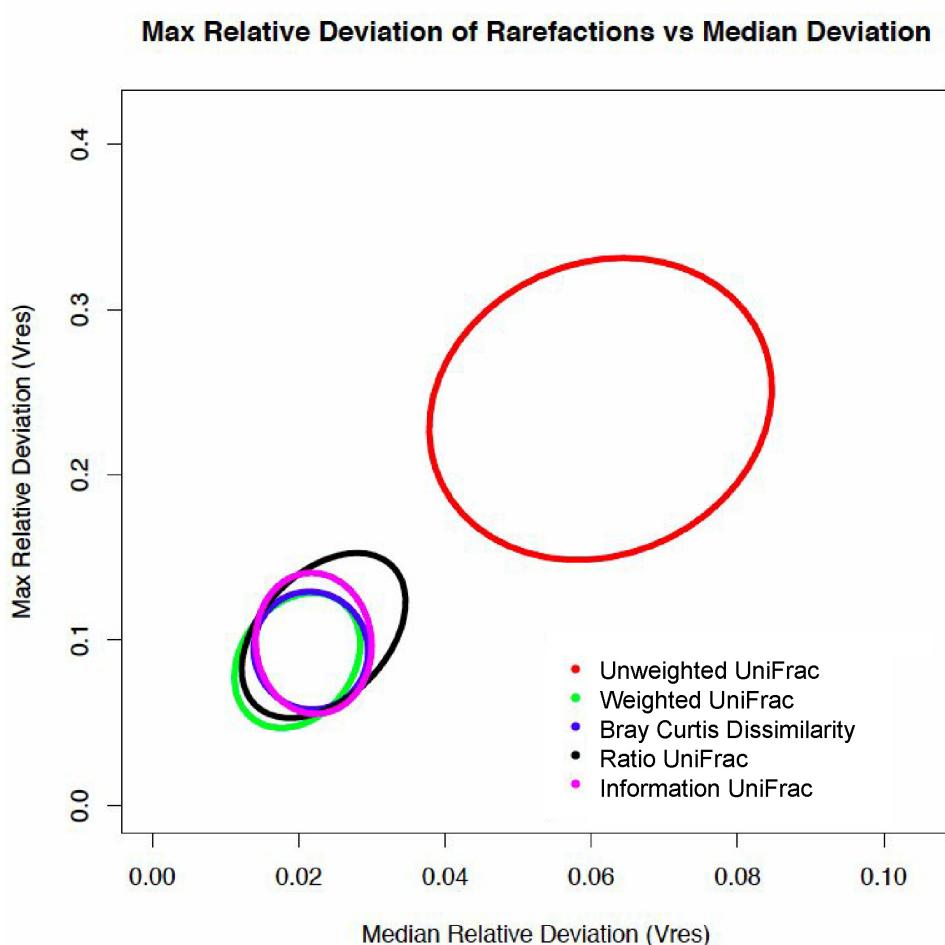


Figure 3. Maximum relative deviation of rarefactions versus median deviation for traditional and non-traditional microbiome dissimilarity metrics. Sixty samples from the tongue dorsum were taken from the Human Microbiome Project [32], and rarefied 100 times. The maximum relative deviation was plotted against the median relative deviation of the rarefied data, and ellipses were drawn at the 95% confidence interval, around the cloud of points for each metric. [A higher maximum and median deviation indicates lower reproducibility of results between rarefaction instances.](#) Both the maximum relative deviation of rarefied data and the median relative deviation of rarefied data are greater in unweighted UniFrac than in weighted UniFrac, Bray Curtis distance, ~~entered~~-ratio UniFrac, and information UniFrac.

8 The cause of rarefaction variation by Unweighted UniFrac

One point to note is that rarefaction carries the assumption that microbiota within samples are homogeneous and randomly distributed. However, this assumption is only valid if proper sampling protocols are observed [36]. A combination of unevenly sampled genes-OTUs and distantly related genes-OTUs will contribute to UniFrac's variability when genes-the variability in unweighted UniFrac when OTUs are ultimately rarefied. Distance matrices between samples will be affected when rare genes-OTUs are left out during the rarefaction processes. It becomes intuitive to see how similar samples may grow dissimilar from each other through unweighted UniFrac on rarefied samples as the number of unshared branches increases as genes-disappear OTUs are removed.

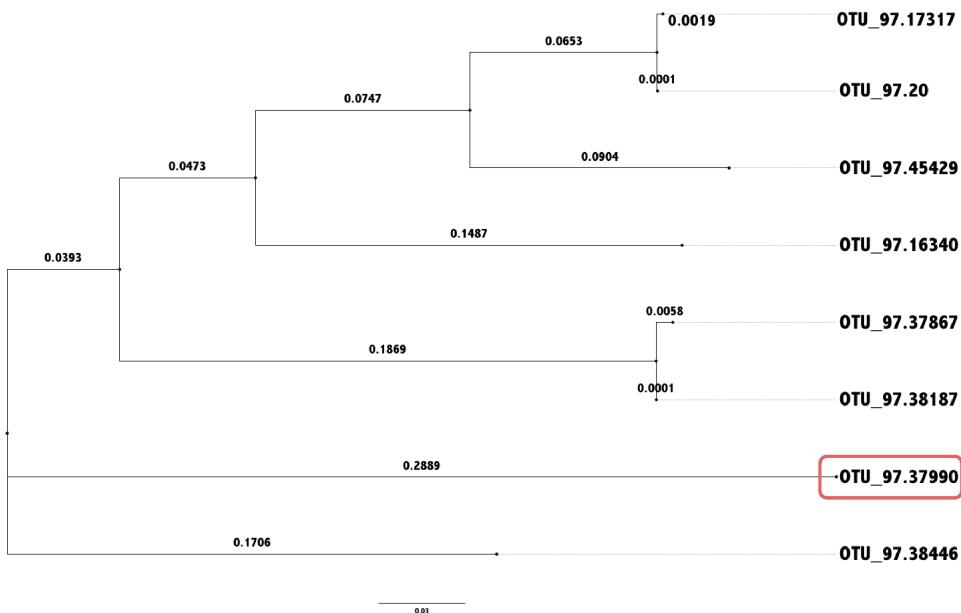


Figure 4. Phylogenetic tree with long isolated branches. Variation in different rarefactions of data in unweighted UniFrac analysis is exacerbated by the presence of long isolated branches in the phylogenetic tree, such as the circled OTU in this example.

337
338
339
340
341
342
343
344
345
346

Distance_{A:B} for Rarefaction1

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

347

Original abundance of taxa and rarefied abundance of taxa.

Table 1. Original abundance of taxa and rarefied abundance of taxa. This data was simulated to demonstrate how rarefaction can change the distances reported by the unweighted UniFrac metric. The OTU in bold has been rarified to a zero count in sample A for one instance and a non zero count in the other instance. In Rarefaction 1, the unweighted UniFrac distance (unshared over total branches) is 0.4175, while in Rarefaction 2 the distance is 1.12.

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

Distance_{A:B} for Rarefaction2

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

With rare genes With rare OTUs 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 doubled in proportion between samples, and this is much more biologically significant than the change in proportional abundance in the second taxa. Additionally, it does not account for how the counts add up to a constrained sum determined by the sequencing machine model. Because the sum is constrained, as an example with the bacterial vaginosis sample earlier, an increase in growth of one taxa can make the data look like there is a decrease in abundance in other taxa, even if in reality the population of the other taxa stayed the same.

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

9 Information UniFrac

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

between the proportional abundances themselves, potentially allowing scientists to differentiate groups samples with subtle differences—such as the infected breastmilk sample in Fig. 7.

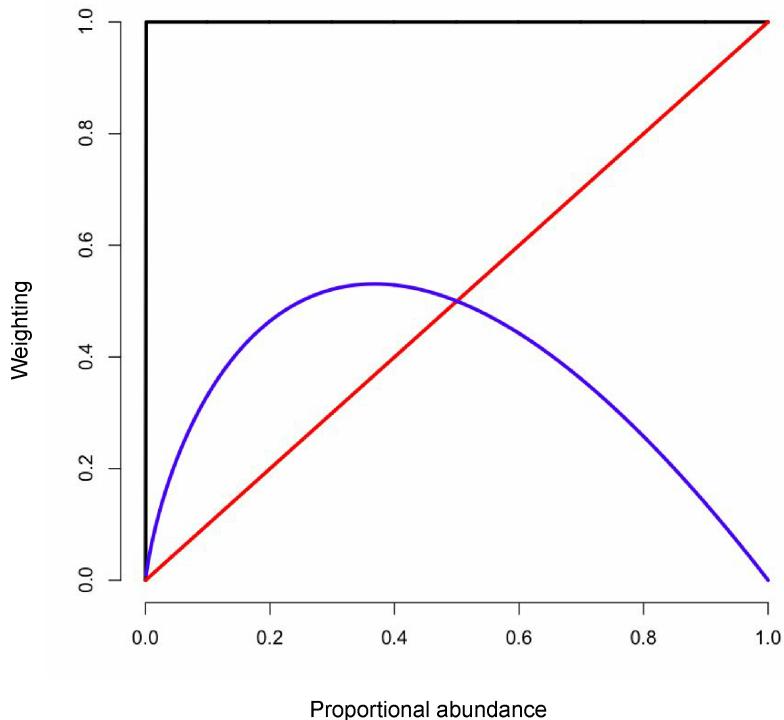


Figure 5. Unifrac–UniFrac weights. Each UniFrac weighting is plotted with the corresponding proportional abundance. The black line is unweighted UniFrac, the red line is weighted UniFrac, and the blue line is information UniFrac. From 0 to 0.2 on the x-axis information UniFrac has a higher slope, and therefore more discovery power with smaller changes in abundance. As the x-axis approaches 1, changes in abundance add little discovery power to information UniFrac.

Near the 0, 0 point in For example, Fig. 5 shows the weighting of a taxon in unweighted, weighted, and information UniFrac as a function of the taxon proportional abundance. Near the 0, the proportional 0 point the proportional abundances are low and information is 0. However, small increases in abundance result in large changes in contribution to UniFrac weighting, as shown by the slope of the curve. Here there is higher differentiation between weights of different pairs of low proportional abundances for information UniFrac, as shown by the higher slope of the curved graph. The centered ratio UniFrac (not depicted) depends on the geometric mean of the taxonomic abundances, and each sample would have a different slope in the weight graph depending on how evenly the abundances were distributed.

10 Tongue-and-cheek comparison

10 Tongue and buccal mucosa comparison

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

Fig. 6 shows a principal coordinate analysis plot with four different metrics: unweighted UniFrac, weighted UniFrac, information UniFrac, and **eentered**-ratio UniFrac. We observe that the difference in the microbiome between the human tongue and cheek-buccal mucosa are well defined by all metrics (Fig. 6), since all of the weightings show separation between the samples according to body site. We conclude from (Fig. 3) that weighted UniFrac, information UniFrac, and **eentered**-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.

388
389
390
391
392
393
394
395
396

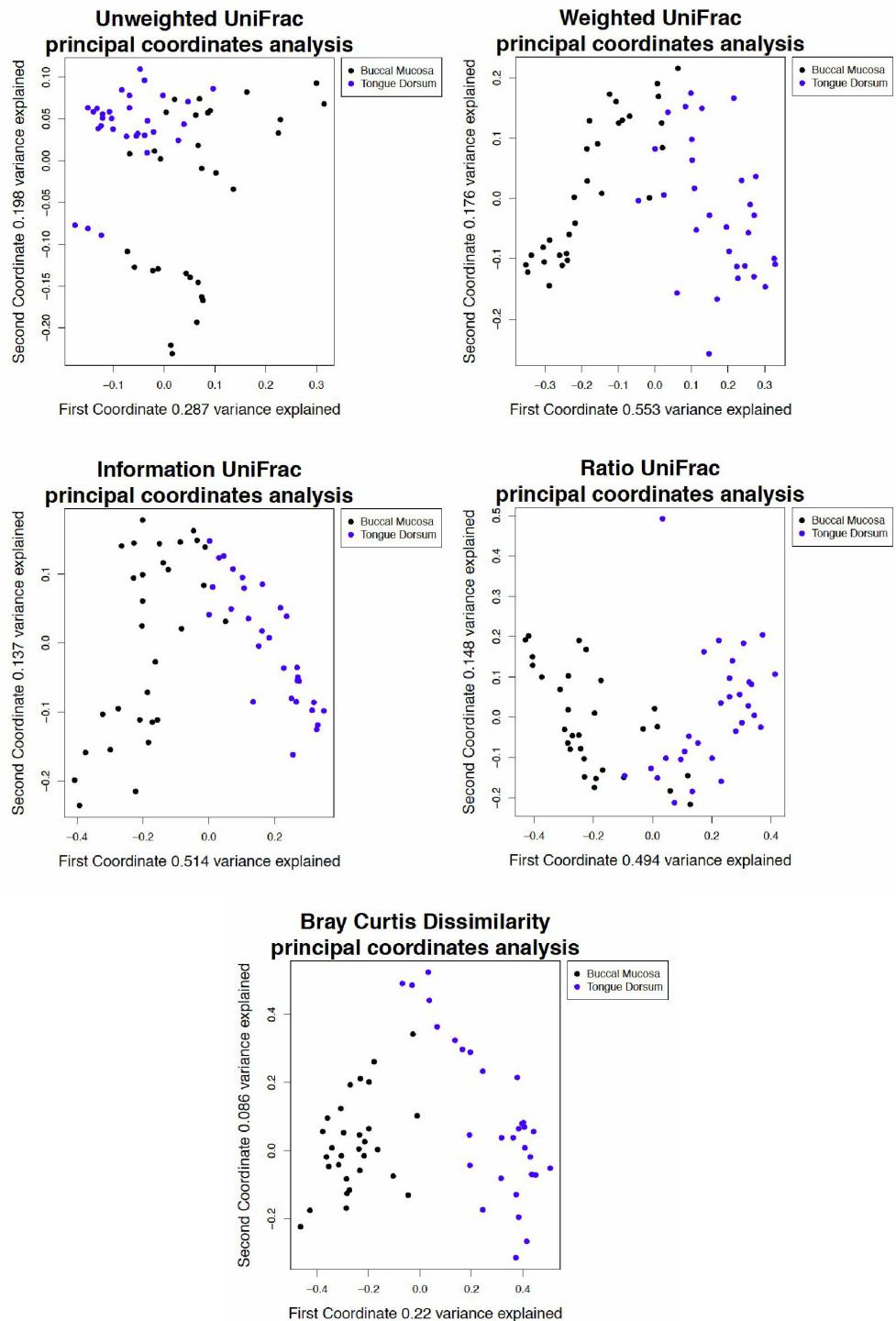


Figure 6. Analysis of tongue and cheek-buccal mucosa data using different UniFrac weightings. A principal coordinate analysis of a 16S rRNA gene tag experiment done on samples from the tongue and cheek-buccal mucosa, selected from the Human Microbiome Project [32]. All weightings and the Bray-Curtis dissimilarity show separation between the samples by body site. Note that the variance explained by the first and second principal coordinate axis is higher than in the tongue-tongue data set from Figure 2, which had 16.1% and 9.8% variance explained, respectively.

11 Breast milk Data

Fig. 7 is a principal coordinate analysis of a 16S rRNA gene sequencing experiment done on microbiome samples from breast milk [33]. Breast milk samples were collected and the V4 region of the 16S rRNA gene was sequenced. One of ~~these samples was infected (circled), consisting the patients who provided a sample had an active infection, producing a sample that consisted~~ of 97% Pasteurella. We noted that this sample was not distinct in unweighted and weighted UniFrac because the distance from the Pasteurella branches of the phylogenetic tree to the root of the tree (rooted by midpoint) were not particularly short or long, measuring at just over the 3rd quartile of all root-to-leaf distances. In addition, the Pasteurella leaves shared a clade with many other taxa.

The reason the infected sample in the breast milk study is so distinct from the rest of the samples in Information UniFrac and ~~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 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.

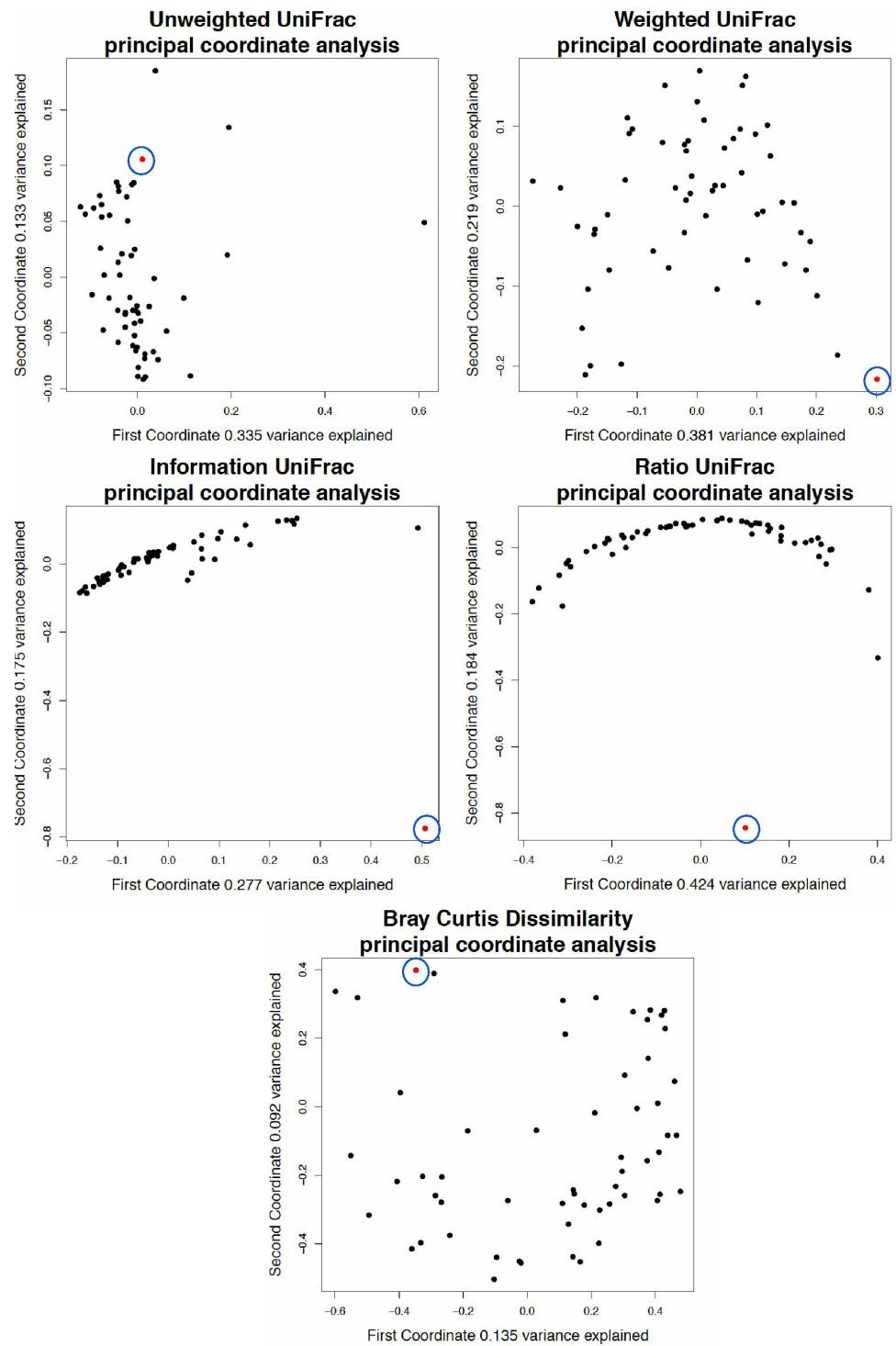


Figure 7. Analysis of breast milk data using different UniFrac weightings. A principal coordinate analysis of a simulated 16S rRNA gene tag experiment done based on samples from a 16S rRNA experiment on the breast milk data. The circled sample is infected with 97% *Pseudomonas* by *Pasteurella*, compared black samples are dominated by *Staphylococcus*, and cyan samples are dominated by *Pseudomonas*. Note that while information UniFrac appears to 15–20% in separate the other samples reasonably well visually, the amount of variance explained by the first two coordinates is much lower than even weighted UniFrac.

12 Monoculture data

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

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

Information UniFrac is known to not perform very well for monocultures, due to taxa with very high and low proportional abundances having uncertainty information values close to zero (Fig. 5). While the samples separate visually with information UniFrac, the variance explained by the separation is low, and the distance matrix does not separate the three groups by hierarchical clustering. Ratio UniFrac and Bray Curtis both separate the samples by monoculture, and also differentiate the samples by their minor variations, showcasing a more representative perspective of this data set.

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

421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442

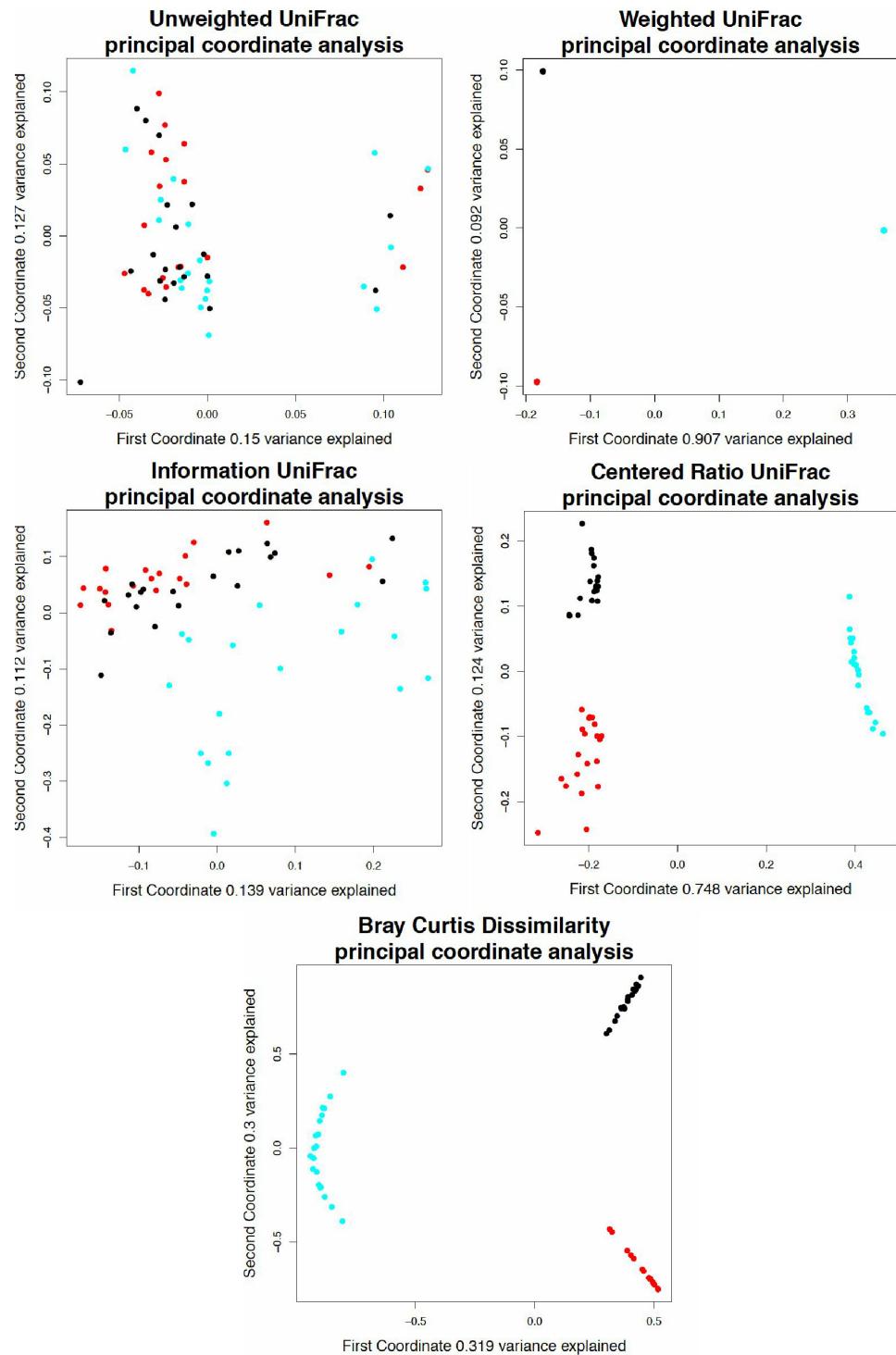


Figure 8. Analysis of simulated monocultures using different UniFrac weightings. A principal coordinate analysis of a simulated 16S rRNA gene tag experiment based on the breast milk data. Red samples are dominated at 97% by *Pasteurella*, black samples are dominated by *Pseudomonas*, and cyan samples are dominated by *Staphylococcus*. Note that while information UniFrac appears to separate the samples reasonably well visually, the amount of variance explained by the first two coordinates is much lower than even weighted UniFrac.

Discussion

As shown in the tongue and ~~cheek~~⁴⁴⁴ buccal mucosa⁴⁴⁵ 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 dorsum⁴⁴⁶ 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 result⁴⁴⁷ in spurious groups, or inclusion of samples in the wrong groups.⁴⁴⁸

We found weighted UniFrac, information UniFrac, ~~entered~~⁴⁴⁹ratio UniFrac, and⁴⁵⁰ Bray-Curtis methods to be more reliable choices. We suggest that investigators use⁴⁵¹ several methods as they can detect outliers in different circumstances. When an outlier⁴⁵² is detected by any metric, an investigation is warranted, as with our example in the⁴⁵³ breast milk data set.⁴⁵⁴

~~We do not believe that any of these weightings are a perfect model for microbiome data. Each tool is prone to its own set of weaknesses. If the difference in groups is driven by presence/absence then UniFrac is a reasonable choice. If the difference is driven by a linear abundance, then weighted UniFrac is a good choice. Information UniFrac and ratio UniFrac are useful for examining data sets that contain the similar sets of taxa between groups. Ratio UniFrac is especially good for examining data sets that have more subtle variations, due to its non linear nature. In any case, inspection should be done to make sure that the tool used accurately represents the data.~~⁴⁵⁵

In summary, with the addition of information UniFrac and ~~entered~~⁴⁵⁶ratio UniFrac,⁴⁵⁷ biologists have more tools at their disposal to prevent spurious interpretations, detect⁴⁵⁸ outliers, and ultimately understand their data better.⁴⁵⁹

Acknowledgments

Thanks to Camilla Urbaniak for providing the data from her breast milk study [33].

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. 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.

7. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*. 2011;108(Supplement 1):4516–4522.
8. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460–2461.
9. 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.
10. 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.
11. 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.
12. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High resolution sample inference from amplicon data. *bioRxiv*. 2015;p. 024034.
13. Tringe SG, Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. *Current opinion in microbiology*. 2008;11(5):442–446.
14. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*. 2004;32(5):1792–1797.
15. 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.
16. Aitchison J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society Series B (Methodological)*. 1982;p. 139–177.
17. 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.
18. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*. 2009;75(23):7537–7541.
19. 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.
20. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol*. 2012;8(9):e1002687.
21. 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.

22. 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.
23. 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.
24. Zozaya-Hinchliffe M, Lillis R, Martin DH, Ferris MJ. Quantitative PCR assessments of bacterial species in women with and without bacterial vaginosis. *Journal of clinical microbiology.* 2010;48(5):1812–1819.
25. Simberloff D. Use of rarefaction and related methods in ecology. In: *Biological data in water pollution assessment: quantitative and statistical analyses.* ASTM International; 1978. .
26. de Cácer 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.
27. 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.
28. Shannon CE. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review.* 2001;5(1):3–55.
29. Egozcue J, Pawlowsky-Glahn V. Evidence information in bayesian updating. In: *Proceedings of the 4th International Workshop on Compositional Data Analysis;* 2011. .
30. 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.
31. Beals EW. Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. *Advances in Ecological Research.* 1984;14(1):55.
32. 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.
33. Urbaniaik 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.
34. 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.
35. 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.
36. 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.