Interpretation and application of absolute abundance in Weighted UniFrac distance

Augustus Pendleton[1]* & Marian L. Schmidt[1]*

[1]Department of Microbiology, Cornell University, 123 Wing Dr, Ithaca, NY 14850, USA

Corresponding Authors: Augustus Pendleton: arp277@cornell.edu; Marian L. Schmidt: marschmi@cornell.edu

Author Contribution Statement: Both authors contributed equally to the manuscript.

Keywords: Microbial Ecology - Beta Diversity - Absolute Abundance - Bioinformatics

The UniFrac distance was first introduced by Lozupone & Knight (2005), and has since become enormously popular as a measure of β-diversity within the field of microbial ecology [1]. A major draw of the UniFrac distance is that is considers phylogenetic information when estimating the distance between two communities. After first generating a phylogenetic tree representing species (or amplicon sequence variants, ASVs) from all samples, the UniFrac distance computes the fraction of branch-lengths which is *shared* between communities, relative to the total branch length represented in the phylogenetic tree. UniFrac can be both *unweighted*, in which only the incidence of species is considered, or *weighted*, wherein the contribution a branch makes to the overall distance is weighted by the proportional abundance of taxa descended from that branch [2]. The weighted UniFrac is derived as:

$$U^R = \frac{\sum_{i=1}^{n} b_i |p_i^a - p_i^b|}{\sum_{i=1}^{n} b_i (p_i^a + p_i^b)}$$

Where we weight the length of each branch, b_i, by the difference in the relative abundance of all species (p_i) descended from that branch in sample a or sample a. As such, we denote this distance as U^R, for "Relative Unifrac". Popular packages which calculate weighted Unifrac, including QIIME and the R packages `phyloseq` and `GUniFrac` run this normalization by default.

Because U^R is most sensitive to changes in abundant lineages, it can sometimes obscure compositional changes occurring in rare to moderately-abundant taxa [3]. To address this weakness, Chen et al. (2012) introduced the generalized UniFrac distance (GU^R), in which the impact of abundant lineages can be mitigated by decreasing the parameter α:

$$GU^R = \frac{\sum\limits_{i=1}^{n} b_i (p_i^a + p_i^b)^\alpha \left| \frac{p_i^a - p_i^b}{p_i^a + p_i^b} \right|}{\sum\limits_{i=1}^{n} b_i (p_i^a + p_i^b)^\alpha}$$

However, if one wishes to use absolute abundances, both U^R and GU^R can be derived without normalizing by total counts:

$$U^A = \frac{\sum\limits_{i=1}^{n} b_i |c_i^a - c_i^b|}{\sum\limits_{i=1}^{n} b_i (c_i^a + c_i^b)} \qquad GU^A = \frac{\sum\limits_{i=1}^{n} b_i (c_i^a + c_i^b)^\alpha \left| \frac{c_i^a - c_i^b}{c_i^a + c_i^b} \right|}{\sum\limits_{i=1}^{n} b_i (c_i^a + c_i^b)^\alpha}$$

Where c_i^a and c_i^b stands for the absolute counts of species descended from branch b_i in community a and b, respectively, yielding "Absolute Unifrac" (U^A/GU^A) c. There is growing recognition that microbial load matters, and methods for discerning absolute abundance like flow cytometry and qPCR are growing in popularity [4, 5]. However, it can be conceptually difficult to predict and interpret the impact of absolute abundance on complex metrics like UniFrac.

As such, we began with a simulated community, represented by four ASVs with simple phylogenetic relationships (Fig. 1A). We assigned each ASV an absolute count of 1, 10, or 100, yielding 81 samples and 3240 cross-sample comparisons for which we calculated the Bray-Curtis dissimilarity and weighted Unifrac distance using both relative

and absolute abundances (BC^R, BC^A, U^R, and U^A). We then calculated the difference in distance calculated by each metric compared to Absolute Unifrac; the distributions of these differences are shown in Fig. 1B.

For all metrics, there are scenarios where U^A estimates either greater or less distance (in the improbable scenario that all branch lengths are equal, U^A will always be equal to or less than BC^A, Fig. S1). We then isolated specific comparisons to illustrate where and when U^A differs from other metrics (Fig. 1C). Scenario one (gold star) indicates the canonical benefit of UniFrac distances: the phylogenetic similarity of ASV_1 and ASV_2 allow U^R and U^A to discern greater similarity between samples than the phylogenetically-oblivious BC^R/BC^A. Scenario two (red star) illustrates the importance of absolute abundances; BC^R and U^R perceive two communities as identical despite large differences in absolute abundance.

At times, the incorporation of absolute abundance can also *decrease* the perceived dissimilarity (green star). Here, BC^A and U^A are both smaller than their relative counterparts, as over half of the community is perceived as identical, despite differences in their relative abundances due to proportionality. Finally, U^A is highly correlated with BC^A (Pearsons $r = 0.82$, $p < 0.0001$) compared to BC^R ($r = 0.41$) and U^R ($r = 0.55$). However, there are scenarios in which U^A is greater than BC^A (blue star). Because the absolute differences are observed on longer branches relative to the rest of the tree, U^A discerns greater distance than BC^A.

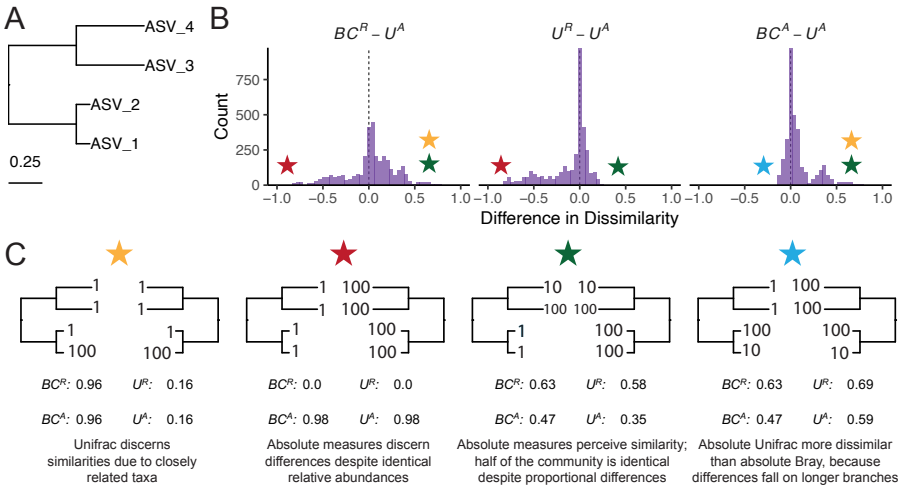

Figure 1. Simple simulations to compare the impact of absolute abundance on phylogenetic and non-phylogenetic distance measures. (A) We created a small community represented by 4 ASVs. We explored all permutations of each ASV holding an absolute count of 1, 10, or 100 to create 81 communities, producing 3240 comparisons. We calculated the Bray-Curtis dissimilarity and weighted Unifrac distance using both relative and absolute abundances (BC^R, BC^A, U^R, and U^A). (B) Distributions of the difference between weighted Unifrac using absolute abundance (U^A), Bray-Curtis using relative abundance (BC^R), weighted Unifrac using relative abundance (U^R), and Bray-Curtis using absolute abundance (BC^A). (C) Scenarios (comparisons) highlighting occasions when U^A is larger or lesser compared to other metrics. The stars indicate what region of the distributions in (B) this scenario would fall. The actual values for each metric are displayed beneath the scenario.

We next explored the practical effect U^A has on separating groups across a real dataset. We used a previously published 16S dataset from Lake Ontario, where samples clustered into three main groups defined by depth and month due to changes in both composition and abundance (Fig. S2, [6]). To illustrate the impact of absolute abundance,

we calculated the generalized absolute UniFrac (GU^A) across three levels of α: 0.0 (close to unweighted UniFrac), 0.5, and 1.0 (identical to U^A). Fig. 2A shows PCoA ordinations of the same samples across this α-range. As changes in abundance become more weighted (α increases), we see a greater recognition of similarity between Shallow September and Shallow May, reflecting their relatively higher cell counts compared to the Deep samples. Note that an increasing amount of variation is assigned to the first axis, going from 18.3% up to 76.7%. This trend was also true for U^R across multiple α, but to a much weaker degree (Fig. S3).

We used PERMANOVAs to quantify the variance (R^2) and statistical power (F-statistic) our three depth-month groups can explain across different distance metrics (GU^R, GU^A, BC^R and BC^A) and different alpha values (Fig. 2B and 2C). We see that absolute abundance metrics consistently allow for greater R^2 values and F-statistics than their relative counterparts. For UniFrac distances, this separation increases as alpha increases. As such, the incorporation of absolute abundances can be a powerful tool to emphasize differences in microbial load between groups.

That said, we identify a major caveat; U^A can become strongly correlated to the cell-counts alone compared to other distance metrics (Fig. 2D). We used a Mantel test to assess the correlation between each distance metric to the absolute difference in cell counts between each sample. Absolute abundance measures are more sensitive to differences in cell counts compared to their relative counterparts. This is intuitive, and to a degree, intentional; the utility of these metrics is to identify changes in absolute abundance even in the case of compositional similarity. However, at $\alpha = 1$, the value of U^A becomes almost entirely based on the overall abundance of a sample. We see this reflected in the ordinations as well; as α increases, the first Axis increasingly represents (and correlates to) the absolute abundance, and its possible that any 2D structure we observed in the third panel of Fig. 2A is simply the result of the horseshore effect [7].

With this in mind, we encourage others to think critically about how heavily they want their distance metric to correspond to absolute abundance, and modulate this effect using GU^A rather than U^A. As a simple start from these data, we recommended using an α value of 0.5, consistent with the original recommendation of Chen et al. (2012) but of heightened importance when using absolute abundances.

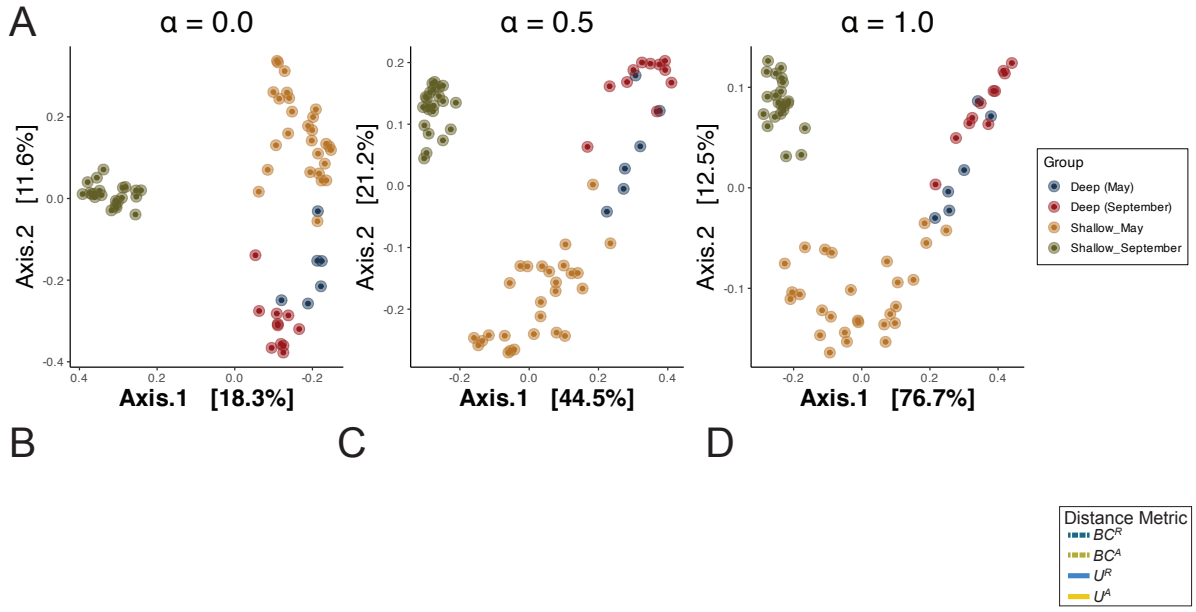

Figure 2. Impact of absolute abundance on separating freshwater microbial communities (A) Samples represent microbial communities from Lake Ontario taken at different times (May and September) and at different depths (Shallow and Deep), as originally described in Pendleton et al. (2025). We used Principal Coordinates Analysis (PCoA) to ordinate samples based on the GU^A at α values of 0.0, 0.5, and 1.0. The proportion of variance assigned to each axis is provided in brackets. Note that the x-axis is reversed in the first panel, to provide visual symmetry across ordinations. We then used PERMANOVAs to quantify the variance (R^2) (B) and statistical power (F-statistic) (C) our three depth-month groups can explain across different distance metrics (GU^R, GU^A, BC^R and BC^A) and different alpha values. (D) Correlations between different distance metrics and differences in over-all cell abundances as assessed by a Mantel test.

The incorporation of absolute abundance allows microbial ecologists to assess more realistic, ecologically-relevant differences in microbial communities. As β-diversity metrics are an essential tool for every microbial ecologist, and UniFrac distances are highly popular within microbiome research specifically, we encourage other researchers to adopt the usage of GU^A in comparing their samples. This index could be especially relevant in instances where microbial load varies widely and is ecologically relevant, including the development of the infant gut, within bioreactors, and after antibiotic treatment [8–10].

As our data show, however, the interpretation of these metrics can be complex, as they integrate differences in composition, phylogeny, *and* abundance in ways which can be unpredictable. As ever, interpreting the results of these metrics critically, and exploring the sensitivity of your conclusions across multiple metrics, is encouraged [11].

References

1. Lozupone C, Knight R. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology* 2005;**71**:8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005
2. Lozupone CA et al. Quantitative and Qualitative Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Applied and Environmental Microbiology* 2007;**73**:1576–1585. https://doi.org/10.1128/AEM.01996-06
3. Chen J et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 2012;**28**:2106–2113. https://doi.org/10.1093/bioinformatics/bts342

137 4. Props R et al. Absolute quantification of microbial taxon abundances. *ISME J* 2017;**11**:584–587. https://doi.org/10.1038/ismej.2016.117

138 5. Wang X et al. Current Applications of Absolute Bacterial Quantification in Microbiome Studies and Decision-Making Regarding Different Biological Questions. *Microorganisms* 2021;**9**:1797. https://doi.org/10.3390/microorganisms9091797

139 6. Pendleton A, Wells M, Schmidt ML. Upwelling periodically disturbs the ecological assembly of microbial communities in the Laurentian Great Lakes.

140 7. Morton JT et al. Uncovering the Horseshoe Effect in Microbial Analyses. *mSystems* 2017;**2**:10.1128/msystems.00166–16. https://doi.org/10.1128/msystems.00166-16

141 8. Rao C et al. Multi-kingdom ecological drivers of microbiota assembly in preterm infants. *Nature* 2021;**591**:633–638. https://doi.org/10.1038/s41586-021-03241-8

142 9. Shapiro OH, Kushmaro A, Brenner A. Bacteriophage predation regulates microbial abundance and diversity in a full-scale bioreactor treating industrial wastewater. *The ISME Journal* 2010;**4**:327–336. https://doi.org/10.1038/ismej.2009.118

143 10. Wagner S et al. Absolute abundance calculation enhances the significance of microbiome data in antibiotic treatment studies. *Front Microbiol* 2025;**16**. https://doi.org/10.3389/fmicb.2025.1481197

144 11. Kers JG, Saccenti E. The Power of Microbiome Studies: Some Considerations on Which Alpha and Beta Metrics to Use and How to Report Results. *Front Microbiol* 2022;**12**. https://doi.org/10.3389/fmicb.2021.796025