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.

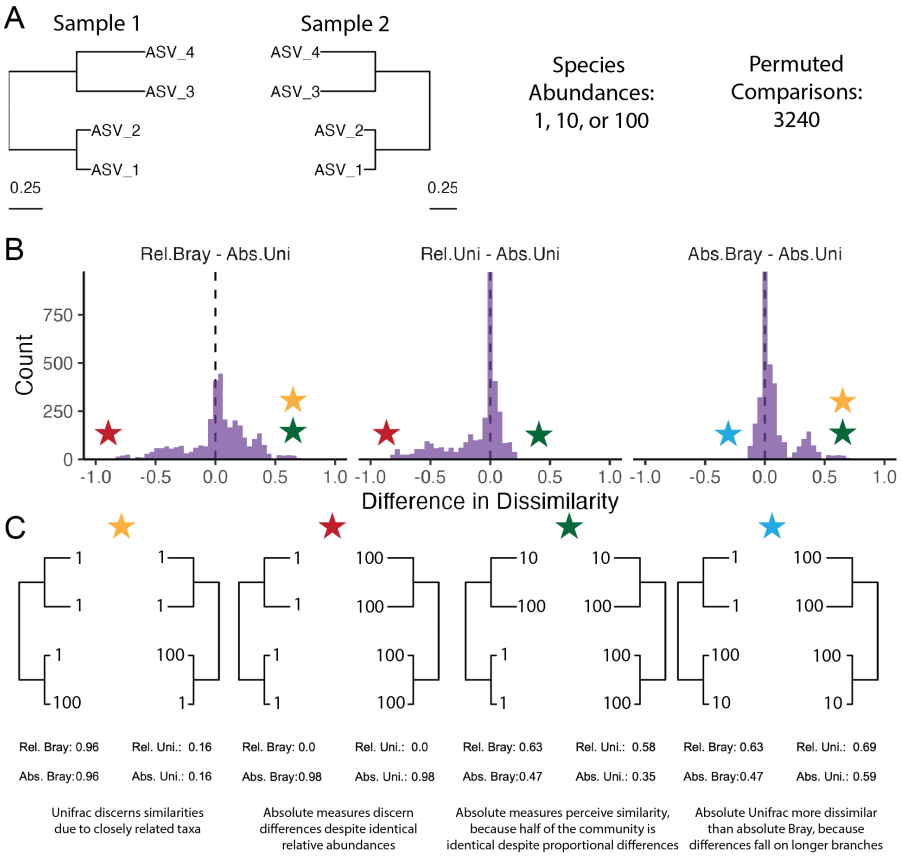

Figure 1. Simple simulations to compare the impact of absolute abundance on phylogenetic and non-phylogenetic distance measures. (A) We created a small community represented by 4 ASVs. We explored all permutations of each ASV holding an absolute count of 1, 10, or 100 to create 81 communities, producing 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.

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
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
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