# 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 in 2005, and has since become enormously popular as a measure of $\beta$-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 proxies like ASVs/OTUs) 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, [3] introduced the generalized UniFrac distance ($GU^R$), in which the impact of abundant lineages can be mitigated by decreasing the parameter $\alpha$:

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

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.

We were curious as to the implications of using this "Absolute Unifrac" ($U^A$) compared to $U^R$, and the impact of absolute abundances between phylogenetic and non-phylogenetic distance measures like the Bray-Curtis Dissimilarity.

## References

40 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

41 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

42 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