PhyloDivNet

Trinh
Willis

June 2019

1 Introduction

The human microbiome is comprised of a vast number of microorganisms that interact in complex ways to influence human health. A common question of interest for microbiome researchers is whether microbial diversity is related to characteristics such as disease states or trait expression. The terms alpha, beta, and gamma diversity were introduced in 1972 by R.H. Whittaker as frameworks to characterize the spatial component of biodiversity. The two most commonly used terms to describe microbial diversity are alpha-diversity which refers to the diversity of a site and beta diversity which characterizes the differences in species composition between sites. Summarizing microbial communities

(What do I want to say about alpha and beta diversity here that might transition into focusing on beta diversity?) Alpha diversity may be of particular interest when understanding whether species diversity is related to some characteristic of the site whereas beta diversity is of particular interest when understanding whether species composition differs between sites.

Beta diversity metrics provide a measure of the differences in composition between two sites which can reveal aspects of microbial ecology that are not apparent when examining microbial composition within individual sites (Goodrich 2014). Beta diversity metrics can be categorized as either phylogenetic vs. non-phylogenetic and quantitative vs. qualitative. Qualitative measures such as Jaccard or unweighted UniFrac consider the presence-absence of species in their calculations whereas quantitative measures such as weighted UniFrac and Bray-Curtis incorporate information on sequence abundance (Goodrich 2014). Phylogenetic metrics such as weighted and unweighted UniFrac are based on phylogeny whereas non-phylogenetic metrics such as Bray-Curtis and Jaccard do not.

Here we focus on the commonly used phylogeny-based UniFrac metrics first introduced by Lozupone et al. in 2008. UniFrac stands for the unique fraction

- Talk about beta-diversity and why we analyze microbiome data in this way. Microbiome studies are interested in diversity because of the
- Talk about UniFrac and its differenct flavors of weighted/unweighted.

Here PhyloDivNet is appropriate for weighted UniFrac because we are testing group differences and as such there's a pooling of information that occurs for inferring some mean group relative abundances for each taxon. Because unweighted UniFrac is a phylogenetic quantitative measure that uses presence/absence information, this approach is not well suited for that metric.

- UniFrac has not traditionally been discussed as an estimand.
- Bacteria are networked and as such DivNet and by extension PhyloDivNet provides a way to incorporate the network structure in our phylogenetic beta-diversity estimation
- Segue into the use of PhyloDivNet and its advantages, particularly when evaluating its performance in hypothesis testing

The way this paper distinguishes itself is through a robust look at the pairing of the log-ratio model for estimating relative abundances (which we know DivNet is a good estimator of non-phylogenetic diversity metrics) with the phylogenetic tree in estimating UniFrac. Basically we really need to find situations in which this proposed method might be limited so that would require me to really look at different tree structures more carefully.

2 Estimating Weighted UniFrac

Weighted UniFrac is a quantitative phylogenetic diversity index and as such utilizes abundance information and a phylogenetic tree in its calculation. We start with samples from $i = 1, ..., n$ ecosystems that have a known vector of covariates $X_i \in R^p$. Q represents the number of species present in one or more ecosystems and we use $q = 1, ..., Q$ to index the Q taxa across all ecosystems. $Z_{iq}[0, 1]$ refers to the latent relative abundance of taxon q in ecosystem i with the relative abundances of all taxon q in ecosystem i summing to 1. For each ecosystem i there are M_i total sequence counts observed that have been classified into q taxa. W_{iq} will represent the number of observed counts of taxon q in ecosystem i.

For the phylogenetic tree, let $m = 1, ..., M$ index the branches of the phylogenetic tree where M represents the total number of branches of the phylogenetic tree. b_m denotes the branch length of the nth branch. d is the distance from each taxon q to the root. And let S_m represent the set of taxa that are descendents from branch m.

The weighted UniFrac (Lozupone et. al 2007) diversity metric is defined as

$$\hat{\beta}_{ij,WU,plug-in} = \sum_{i=1}^{n} b_n \left| \sum_{q \in S_{in}} \frac{W_{iq}}{M_i} - \sum_{q \in S_{jn}} \frac{W_{jq}}{M_j} \right| \tag{1}$$

Based on Eq. 1, the target estimand for weighted UniFrac is

$$\beta_{ij,WU} = \sum_{i=1}^{n} b_n | \sum_{q \in S_{in}} Z_{iq} - \sum_{q \in S_{jn}} Z_{jq} | \tag{2}$$

To our knowledge, there has been no discussion of the target estimand for weighted UniFrac in the literature to date. There also exists a normalized weighted UniFrac metric where $\beta_{ij,WU} \in [0, 1]$. The normalized UniFrac plug-in estimate divides the weighted UniFrac by the average distance of each taxon q from the root.

$$\hat{\beta}_{ij,NWU,plug-in} = \frac{\sum_{i=1}^{n} b_n | \sum_{q \in S_{in}} \frac{W_{iq}}{M_i} - \sum_{q \in S_{jn}} \frac{W_{jq}}{M_j} |}{\sum_{q=1}^{Q} d_q (\frac{W_{iq}}{M_i} + \frac{W_{jq}}{M_j})} \tag{3}$$

The target estimand for normalized weighted UniFrac is therefore

$$\beta_{ij,NWU} = \frac{\sum_{i=1}^{n} b_n | \sum_{q \in S_{in}} Z_{iq} - \sum_{q \in S_{jn}} Z_{jq} |}{\sum_{q=1}^{Q} d_q (Z_{iq} + Z_{jq})} \tag{4}$$

Compositional data can be modeled using a multinomial distribution where the covariances between components are negative constrained. If the sum of all components is equal to some known fixed n, an increase in one component will result in a decrease in another component and therefore be negatively correlated. To address this negative constrained covariance issue we follow the `DivNet`(Willis and Martin[under review]) approach to estimating the latent composition matrix $Z \in R^{nxQ}$ and beta-diversity. `DivNet` uses Aitchison's log-ratio methodology to estimate the latent composition matrix by first modeling W_{iq} from a multinomial distribution,

$$p(W|Z) \propto \prod_{i=1}^{n} \prod_{q=1}^{Q} Z_{iq}^{W_{iq}} \tag{5}$$

and then performing the log-ratio transformation by fixing some baseline taxon D as a comparison group:

$$Y_{iq} = \phi Z_{iq} = \{ log(\frac{Z_{iq}}{Z_{iD}}) \}_{q=1,...,D-1,D+1,...Q} \tag{6}$$

Note that the log-ratio transformation is invertible $Z_{iq} = \phi^{-1} Y_{iq}$. The log-ratios Y_i are then modeled using a multivariate normal distribution. We link the mean of Y_i to covariates using $\mu_i = X_i^T \beta$. Our expected value of Y_i can therefore be expressed as

$$\hat{Y}_i = X_i^T \hat{\beta} \tag{7}$$

where $\hat{\beta}$ is the maximum likelihood estimate of β. The expected value of our random variable Y_i can then be used to derive the fitted values of the latent composition since $\hat{Z}_i = \phi^{-1}(Y_i)$.

Using `DivNet`'s estimation approach of β-diversity we propose the following estimate of weighted UniFrac and normalized weighted UniFrac

$$\hat{\beta}_{ij,WU,proposed} = \sum_{i=1}^{n} b_n | \sum_{q \in S_{in}} \hat{Z}_{iq} - \sum_{q \in S_{jn}} \hat{Z}_{jq} | \tag{8}$$

$$\hat{\beta}_{ij,NWU,propose} = \frac{\sum_{i=1}^{n} b_n | \sum_{q \in S_{in}} \hat{Z}_{iq} - \sum_{q \in S_{jn}} \hat{Z}_{jq} |}{\sum_{q=1}^{Q} d_q (\hat{Z}_{iq} + \hat{Z}_{jq})} \tag{9}$$

For the phylogenetic tree, we assume that the phylogenetic tree that is constructed is correctly specified. Here we define an estimand for weighted UniFrac and a proposed method which we call `PhyloDivNet` for estimating the weighted UniFrac estimand.

3 Variance Estimation

Accurate estimation of the variance of UniFrac estimates has important implications towards valid hypothesis testing. Variance estimation of our UniFrac estimates is carried out using parametric and nonparametric bootstrap approaches and evaluated under simulation. Parametric bootstrapping to estimate the $Var(\hat{\beta}_{ij,NWU})$ is conducted as follows: Let $\hat{\beta}$ and $\hat{\Sigma}$ be the estimates of β and Σ of our log-ratio model and B represent the number of bootstrap iterations. We simulate B number of datasets from the log-ratio model using our $\hat{\beta}$ and $\hat{\Sigma}$ estimates. For each simulated dataset $b = 1, ..., B$ we use `DivNet` to estimate the $\hat{\beta}^b$ and $\hat{\Sigma}^b$. From there, we estimate the weighted UniFrac ($\hat{\beta}^b_{ij}$) for $b = 1, ..., B$. The $\widehat{Var}(\hat{\beta}^b_{ij,NWU})$ is therefore the parametric bootstrap estimate of the $Var(\hat{\beta}_{ij,NWU})$.

Nonparametric bootstrapping to estimate the $Var(\hat{\beta}_{ij,NWU})$ begins with some dataset (W, X). We define n_{sub} as the number of samples we want to subsample from the dataset (W,X) and let $i = 1, ..., I$ index the samples. We then use a uniform random selection of n_{sub} elements from $1, ..., n_{sub}$ that will correspond to the sample indices subsampled from some dataset (W, X). The subsampled set of samples will be referred to as P. With each set P we estimate $\hat{\beta}^P$ and $\hat{\Sigma}^P$ to obtain $\hat{\beta}^P_{ij,NWU}$. We repeat this subsampling from (W,X) and estimation of $\hat{\beta}^P_{ij,NWU}$ for B iterations until we obtain a set of UniFrac estimates $\hat{\beta}^{P_b}_{ij,NWU}$ for $b = 1, ..., B$ bootstrap iterations. From there we obtain the nonparametric bootstrap estimate of $Var(\hat{\beta}_{ij,NWU})$ by calculating $\widehat{Var}(\hat{\beta}^{P_b}_{ij,NWU})$.

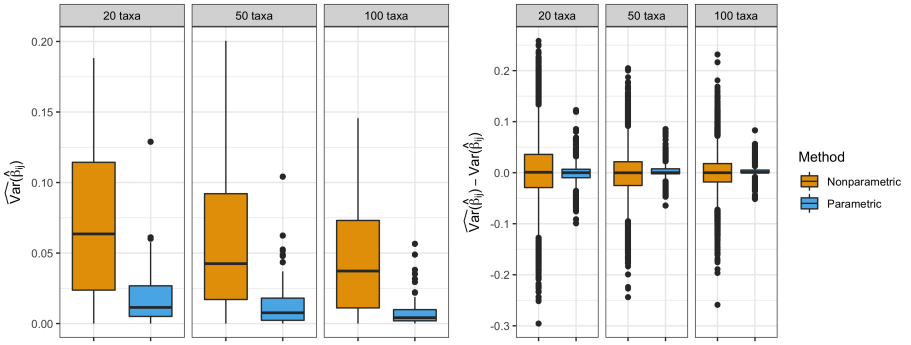
3.1 Variance Estimation Simulation Study

As mentioned by Willis and Martin [Under Review], $q >> n$ in microbiome studies and as such the generalized inverse approach may inappropriate for estimating the inverse covariance. Willis and Martin therefore compared several

approaches to estimate the inverse covariance such as using a regularize estimate from the graphical lasso, a maximum likelihood estimate restricted to the diagonal covariance matrices, and the generalized inverse of the sample covariances. They found that there was no substantial advantage in estimating the inverse covariance between any of the three approaches and as such we choose the generalized inverse of the sample covariances approach as our method for estimating the inverse covariance in our variance estimation simulations.

We compare both nonparametric and parametric bootstrapping approaches to estimating $Var(\hat{\beta_{ij,NWU}})$ under simulation. We simulate our data from a log-ratio model by specifying β, Σ, X, and M. β is generated from a random normal distribution with $\mu = 0$ and $SD = 1$. Σ is constructed by first creating a matrix $A \in \mathbb{R}^{(Q-1)(Q-1)}$ whose elements are drawn from a uniform distribution that range from (-1,1). From there a diagonal matrix D is created where the diagonal elements are an arithmetic sequence of length Q starting from σ_{max} and ending at σ_{min}. Σ is then calculated as $\Sigma = A^T DA$. $X = (1_n^T, (0_{n/2}, 1_{n/2})^T, (0,1)_{n/2}^T)$ and M is the sequencing depth for each ecosystem i set to 10,000. We fix $n = 30$, $\sigma_{min} = 0.01$, $\sigma_{max} = 5$, and evaluate varying $Q = 20, 50, 100$ for 50 simulations under a log-ratio model. To simulate from the log-ratio model with specified parameters β, Σ, X, M, we simulate a matrix Y_i from a $N(X_i^T \beta, \Sigma)$ then calculate $Z_i = \phi^- 1(Y_i)$ for the ith row to obtain the Z matrix. With the Z matrix we can simulate our count data W_i from a $Multinomial(M_i, Z_i)$. We select $B = 3$ iterations for the parametric bootstrap and $B = 3$ subsamples of $n_{sub} = 26$ samples for the nonparametric bootstrap.

Figure 1 Variance Estimation Results



Results from our simulations are displayed in Figure 1. We observe that the parametric bootstrap variances have a lower median estimated variance for all Q with increasing number of taxa resulting in lowering median estimated variances for both the nonparametric and parametric bootstrap approaches (left panel). We present the difference between the estimated variance and the true variance for both approaches in the right panel. The true variance was estimated by simulating data according to (β, Σ, and M), calculating the weighted UniFrac for each dataset, then calculating the variance of the weighted UniFrac

estimates. We see that the median difference between our estimated variance and the true variance is closer to zero for the parametric bootstrap approach across all Q than the nonparametric bootstrap approach. For our evaluation of `PhyloDivNet` on an empirical dataset (Section: Data Analysis) we will therefore be choosing a parametric bootstrap approach to estimating our variances.

4 Weighted UniFrac Estimation Simulation Study

To examine the performance of `PhyloDivNet` in estimating weighted UniFrac we compare our proposed method against the *empirical plug-in* estimator (Vu et al. 2007) and the *zero-replace* estimator (Martín-Fernández et al. 2003). The *empirical plug-in* estimator estimates the weighted UniFrac between ecosystems i and j by replacing Z_{iq} and Z_{jq} with the observed or empirical relative abundances $\frac{W_{iq}}{Mi}$ and $\frac{W_{jq}}{Mj}$. The *empirical plug-in* estimator can also be referred to interchangeably as the maximum likelihood estimate (MLE) as the observed relative abundances are considered the maximum likelihood estimates of the unknown abundances (Vu et al. 2007).

$$\hat{\beta}_{ij,NWU,plug-in} = \frac{\sum_{i=1}^{n} b_n |\sum_{q\in S_{in}} \frac{W_{iq}}{M_i} - \sum_{q\in S_{jn}} \frac{W_{jq}}{M_j}|}{\sum_{q=1}^{Q} d_q(\frac{W_{iq}}{M_i} + \frac{W_{jq}}{M_j})} \tag{3}$$

The *zero-replace* method was proposed by Martín-Fernández et al. 2003 to estimate the unknown composition by replacing zero values with 0.5. The resulting *zero-replace* estimator for weighted UniFrac would therefore be considered

$$\hat{\beta}_{ij,NWU,ZR} = \frac{\sum_{i=1}^{n} b_n |\sum_{q\in S_{in}} \frac{W_{iq}\vee 0.5}{\sum_{r\in C} M_{ir}\vee 0.5} - \sum_{q\in S_{jn}} \frac{W_{jq}\vee 0.5}{\sum_{r\in C} M_{jr}\vee 0.5}|}{\sum_{q=1}^{Q} d_q(\frac{W_{iq}\vee 0.5}{\sum_{r\in C} M_{ir}\vee 0.5} + \frac{W_{jq}\vee 0.5}{\sum_{r\in C} M_{jr}\vee 0.5})} \tag{10}$$

We evaluate `PhyloDivNet` against the *empirical plug-in* estimator and the *zero-replace* estimator under simulation where W is simulated from the log ratio model by specifying $\beta \in \mathbb{R}^{pxQ}$, $X \in \mathbb{R}^{nxp}$, $\Sigma \in \mathbb{R}^{QxQ}$, and $M \in \mathbb{R}^{n}$. We set $p = 3$ and $X = (1_n^T, (0_{n/2}, 1_{n/2})^T, (0,1)_{n/2}^T)$ for all K simulations. Additionally, a matrix $\beta \in \mathbb{R}^{pxQ}$ is generated using a random normal distribution with a mean of 0 and a standard deviation of 1 for all K. It is of note that the latent relative abundance vector can be obtained for a sample i through $Z_i = \phi^{-1}(X_i^T\beta)$. Finally, Σ is constructed by first creating a matrix $A \in \mathbb{R}^{(Q-1)(Q-1)}$ whose elements are drawn from a uniform distribution that range from (-1,1). From there we create a diagonal matrix D where the diagonal elements are an arithmetic sequence of length Q starting from σ_{max} and ending at σ_{min}. Σ is then calculated as $\Sigma = A^T DA$ for all K. We simulate data according to the log-ratio model as outlined in previously in section 2.3.1. This simulation study is managed using the `simulator` (Bien 2016). We evaluate the weighted UniFrac estimates using the mean squared error (MSE) across all simulated samples. The MSE of the estimated weighted UniFrac for the kth simulation where i

indexes each pairwise weighted UniFrac estimate is therefore

$$MSE(\widehat{UniFrac}^{(k)}) = \frac{1}{n(n-1)/2} \sum_{i<j} (\widehat{UniFrac}_{ij}^{(k)} - UniFrac_{ij})^2 \qquad (11)$$

4.1 Increasing Sample Sizes Reduces Estimation Error

To evaluate the relationship between increasing sample size and estimation error, we simulate data according to the approach oultined in section 2.4 and set $Q = 50$, $\sigma_{min} = 0.01$, $\sigma_{max} = 5$, $M_i = 10^5$ for all i, $K = 100$ simulations, and evaluate four samples sizes $n = 10, 20, 40, 80$. An input phylogenetic tree is generated for each set of simulated sample sizes using the `rtree` function from `ape` (Paradis 2004) where Q taxa are specified and the edges of the tree are randomly split until Q tips have been reached. The branch lengths of the tree are taken from a Uniform(0,1).

Figure 2 Normalized Weighted UniFrac: n-vary

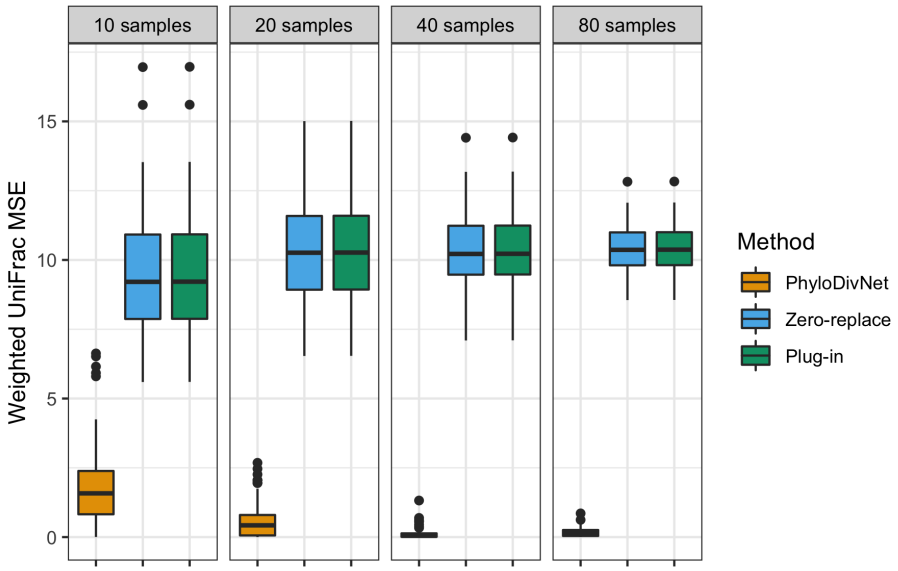


Figure 1 illustrates the performance of `PhyloDivNet` against the *zero-replace* and *plug-in* estimators using the MSE. Increasing sample size results in lower estimation error when using `PhyloDivNet`, but the same pattern does not hold for the *zero-replace* and *plug-in* estimators. This is expected as these estimators do not utilize information from the covariate matrix in their estimation approaches and thereby do not benefit from the increased information afforded with larger sample sizes. Additionally, for all values of n we see that the estimation error of `PhyloDivNet` is uniformly lower than the estimation error of the *zero-replace*

and *plug-in* estimators. These results are consistent with `DivNet`'s estimation of non-phylogenetic α and β diversity metrics (Willis and Martin [Under Review]).

4.2 Estimation Error is Small and Stable With Increasing Number of Taxa

We next investigate the performance of `PhyloDivNet` against the *zero-replace* and *plug-in* estimators with varying number of taxa. For these simulations we set $n = 30$, $\sigma_{min} = 0.01$, $\sigma_{max} = 5$, $M_i = 10^5$ for all i, $K = 50$ simulations, and evaluate four different community sizes $Q = 20, 50, 150, 300$. An input phylogenetic tree for each simulated set of Q was constructed using the `rtree` function from `ape` (Paradis 2004) where Q taxa are specified and the edges of the tree are randomly split until Q tips have been reached. The branch lengths of the tree are drawn from a Uniform(0,1).

Figure 3 Normalized Weighted UniFrac: Q-vary

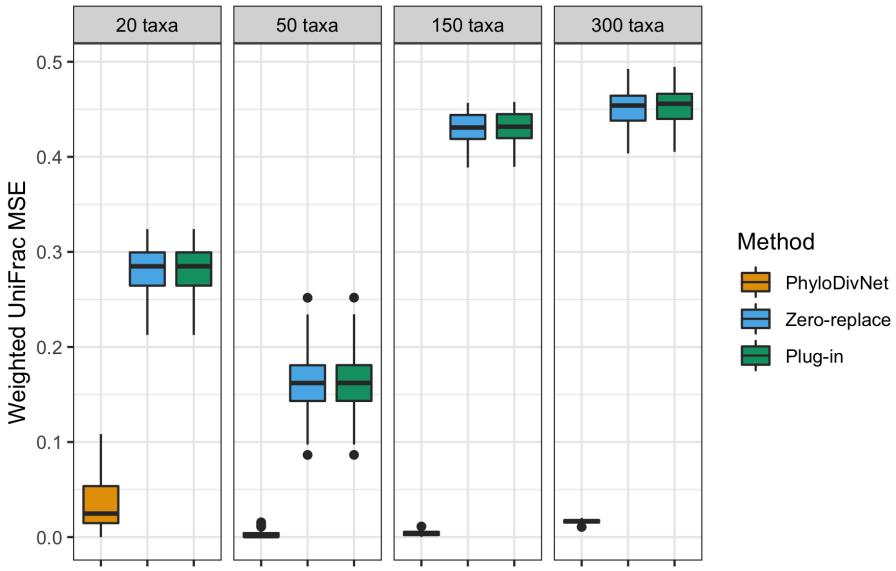
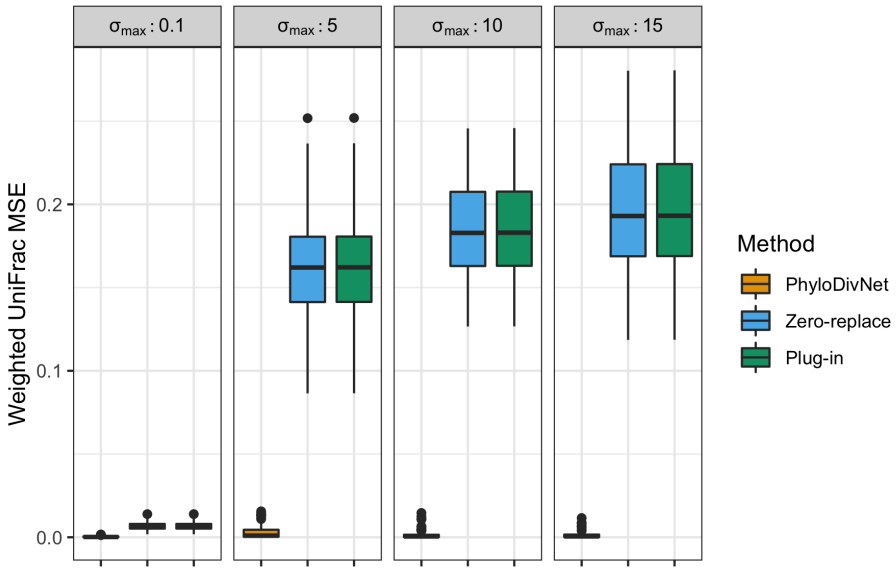


Figure 2 shows the simulation results. We see that with increasing number of taxa `PhyloDivNet` maintains smaller estimation error than both the *plug-in* and *zero-replace* estimators. In fact, in all community sizes the 75th quantiles of $MSE(\widehat{UniFrac}^{(k)})_k$ for `PhyloDivNet` are uniformly lower than the *zero-replace* and *plug-in* methods. It common in microbiome studies for $q >> n$ and we see that estimation error is typically larger with larger numbers of taxa for the *plug-in* and *zero-replace* methods. Here we demonstrate that `PhyloDivNet` outperforms the *zero-replace* and *plug-in* methods in estimating the weighted UniFrac in terms of the MSE even in communities with large numbers of taxa.

4.3 Estimation Error is Small and Stable Across Increasing Strengths of Co-occurrences

We know that bacteria live in networked communities and that co-occurrence patterns exist between microorganisms that may be due to factors such as shared physiologies, habitat affinities, or functional ecological roles (Barberán et al. 2012). As such, a method for estimating diversity that can take into consideration co-occurrences would be appropriate for microbial community analysis. We investigate the estimation performance of `PhyloDivNet` with differing networks of microbial co-occurrences by examining different σ_{max}. σ_{max} refers to the maximum eigenvalue of a Σ matrix where larger eigenvalues correspond to stronger co-occurrences. For our simulations we set $Q = 50$, $n = 30$, $\sigma_{min} = 0.01$, $M_i = 10^5$ for all i, $K = 100$ simulations, and evaluate four strengths of co-occurrences $\sigma_{max} = 0.1$, 5, 10, 15. An input phylogenetic tree is constructed for each simulated set of σ_{max} was constructed using the `rtree` function from `ape` (Paradis 2004) where Q taxa are specified and the edges of the tree are randomly split until Q tips have been reached. The branch lengths of the tree are drawn from a Uniform(0,1). Figure 3 displays the results of our simulation.

Figure 4 Normalized Weighted UniFrac: Sigma-vary



We observe that with stronger co-occurrences the estimation error increases for the *zero-replace* and *plug-in* estimators whereas `PhyloDivNet`'s estimation error remains consistently smaller. We thus conclude that `PhyloDivNet` is well suited for estimating weighted UniFrac in the presence of strong and weak mi-

crobial co-occurrence networks.

4.4 Estimation Error is Small and Stable for Balanced and Unbalanced Trees

Weighted UniFrac is a phylogenetic quantitative β-diversity measure that uses relative abundance information to weight the branch lengths of the tree in its calculation of distances between ecosystem i and j. There are two distinct features of rooted phylogenetic trees. First the topology of the tree and secondly the branch lengths which correspond to periods of time separating evolutionary events. The shape of a tree carries useful information about diversification rates among species. One of the most widely used indices to measure tree balance is Colless' index which looks at each internal node indexed by $j = 1, ..., n-1$ for a phylogenetic tree with n tips and partitions the leaves that descend from each j node into L_j and R_j sized groups. The absolute difference between L_j and R_j is calculated at each node and summed across all nodes to compute the Colless metric.

$$I_c = \sum_{j=1}^{n-1} |L_j - R_j| \tag{12}$$

A standardization of the Colless metric under the Yule model allows for comparison across different tree sizes. Under the Yule model the standardized Colless index is

$$I_{yule} = \frac{I_c - n * log(n) - n(\gamma - 1 - log(2))}{n} \tag{13}$$

where γ is the Euler constant. A value of I_c equal to 0 represents a balanced tree and as follows the larger the positive value the greater the imbalance. Therefore more negative values of I_{yule} represent more balanced trees and more positive values reflect more imbalanced trees. To investigate the performance of `PhyloDivNet` on balanced versus balanced phylogenetic trees we simulate a set of 10,000 Yule trees using `phytools` (CITE) fixing $q = 60$ tips. Because the Yule process assumes that every lineage is equally likely to speciate at any given time with no extinction and only birth until q tips are reached, the phylogenetic trees generated are ultrametric.

To evaluate our method on balanced and unbalanced tree topologies we calculated the Yule standardized Colless metric for all 10,000 trees and defined balanced trees as $I_c < -1$ and unbalanced trees as $I_c > 1$. This yielded 1000 trees of our 10,000 simulated trees that were considered unbalanced and 1000 trees that were considered balanced. We designated the tree with the lowest Colless statistic as the true balanced tree and similarly the tree with the highest Colless statistic as the true unbalanced tree. Of the remaining trees we randomly drew without replacement from our pool of unbalanced and balanced trees for use in our simulations. We set $n = 40$, $p = 3$, $\sigma_{max} = 5$, $\sigma_{min} = 0.01$, $M_i = 10^5$ for all i, K = 50 simulations. Figure 5 shows the simulation results.

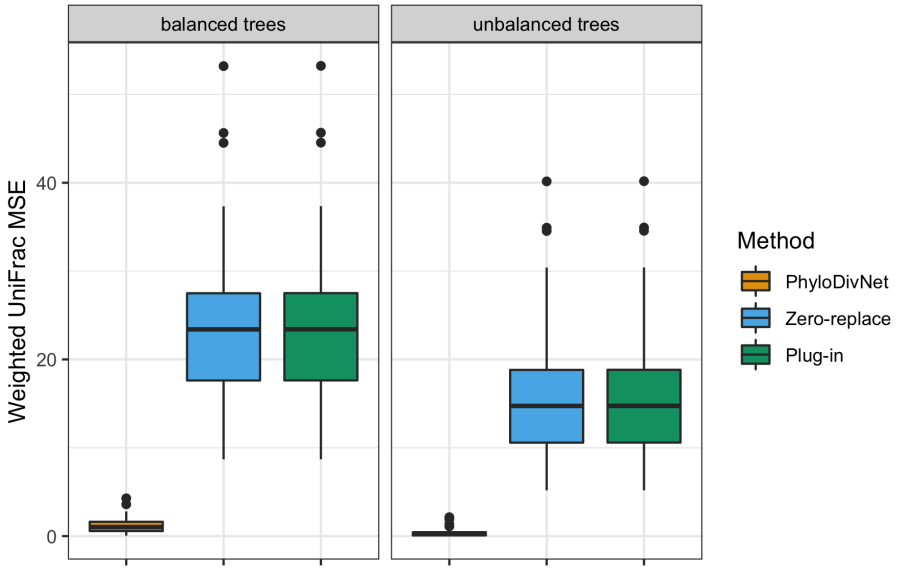


Figure 5 Weighted UniFrac: Unbalanced vs. Balanced Trees

In Figure 5 we see that estimation error is small and stable for our proposed method compared to the *plug-in* and *zero-replace* methods for both balanced and unbalanced tree shapes.

Note: Amy, as I as going through this section I am dissatisfied with how I did these phylogenetic tree simulations and I believe I can do better. I had several thoughts (1) You had suggested that I bin the Colless metric so that I didn't choose an arbitrary cut-off point for balanced/unbalanced trees but I wasn't actually sure what would be considered the "true tree" for each Colless interval so I left it the way it is currently until I talked about it with you a bit more in April. (2) building on that, I was delving into the phylogenetic tree simulation package TreeSim and thinking about how phylogenetic trees are constructed and how that might affect weighted UniFrac, I think I could do a better job with testing different phylogenetic trees not based on the balanced vs unbalanced design but maybe based on evolutionary rate and extinction rate. I'd like to think about simulating a birth-death tree (which will get me non ultrametric trees) so I wanted to make this my first task coming back that first week to improve on this section.

5 Data Analysis: Healthy Dairy Worker Study

The Healthy Dairy Worker (HDW) study is a on-going cohort study that explores the health effects of microbial exposures from a dairy farm on dairy workers in the Pacific Northwest. One of the primary aims of the study is to understand the microbial sharing that occurs between dairy workers, dairy cows, and the dairy farm environment. The study thus provides an especially relevant setting to understand differences in community composition between cows, workers, and the environment through the use of weighted UniFrac. By January 2019, fecal samples for 38 dairy workers and 14 dairy cows as well as 22 environmental swabs were collected at baseline. DNA was extracted using MoBio DNeasy PowerLyzer PowerSoil Kits. Library prep was performed according to the Earth Microbiome Project (EMP) protocol (CITE EMP) and samples were sequenced using an Illumina MiSeq targeting the V4 region of the 16s rRNA subunit.

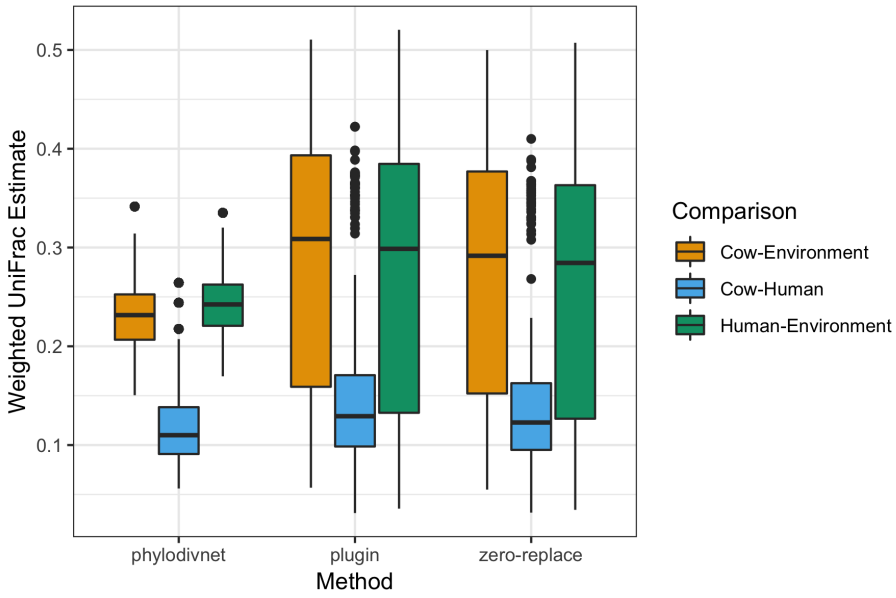
Paired end sequences were processed using DADA2 (Callahan et al. 2015). First, primer sequences were filtered and trimmed from forward and reverse reads. Based on quality scores of 20 or greater, forward sequences were trimmed left at nucleotide position 20 and truncated right at position 200. Reverse reads were trimmed left at position 50 and truncated right at position 200. Paired end sequences were then merged and chimeras removed to create a table of amplicon sequence variants (ASVs) and their counts. A phylogenetic tree was then constructed using SATé-enabled phylogenetic placement (SEPP) (Mirarab et al. 2012). SEPP allows for phylogenetic tree construction by inserting sequences into a reference phylogeny and thereby creating trees with more accurate branch lengths than those generated using de novo phylogeny reconstruction (Janssen et al. 2018). The reference phylogeny and matching alignment used in this analysis was Greengenes 13_8 at 99% (CITE). The ASV table was then filtered to include only the fragments that were present in the insertion tree. A final filtering of the data was done using a minimum ASV abundance cutoff of 0.005% of total reads as is recommended when a mock community is not incorporated in sequencing to minimize the generation of spurious ASVs (Bokulich et al. 2013). A total of 3,441 amplicon sequence variants (ASV) remained after quality filtering and processing for analysis. Data processing was done using R 3.5.0 and q2-fragment-insertion (Janssen et al. 2018).

[INSERT HERE: Investigating the different choices of D (reference taxon) and Amy's plot idea of the continuous UniFracs along a phylogenetic tree for each taxon selected as the reference]

We compare our proposed method `PhyloDivNet` with the *zero-replace* and *plug-in* methods for estimating weighted normalized UniFrac. We use the most abundant taxon as the reference taxon D (Eq.6). A parametric bootstrap approach as described in Section 2.3 is used to estimate the variance with $B = 100$. The results are displayed in Figure 6 with the 25% and 75% quantiles shown.

Comparing the weighted UniFrac estimates in Figure 6 generated by each method, we see that the intervals for `PhyloDivNet` are smaller than the intervals for the *plug-in* and *zero-replace* methods. Additionally our method pro-

Figure 6 Data Analysis Results



duces interval estimates that are more symmetric around the median than the comparison methods. Another noted advantage of our method compared to the *plug-in* and *zero-replace* methods is the use of multiple covariates in estimating the weighted UniFrac.

6 Hypothesis Testing

We evaluate the performance of PhyloDivNet under several proposed hypothesis testing procedures. In particular we were interested in evaluating the control of Type I and Type II error rates with four different hypothesis testing procedures: (1) Constructing a Wald test statistic and comparing this against the chi-square distribution, (2) constructing a Wald test statistic and comparing this against a nonparametric bootstrapped distribution, (3) constructing a Wald test statistic and comparing this against a permutation-derived distribution, and (4) **I forgot what the fourth approach was and took very sparse notes on this suggestion you had for a procedure that wouldn't require estimating the variance**. We are also interested in comparing our proposed method and hypothesis testing procedures for PhyloDivNet with the commonly used plug-in + Permutational Multivariate Analysis of Variance (PERMANOVA) approach (Anderson 2001, Anderson 2017) to testing for differences in community composition between ecosystems.

6.1 Control of Type I Error Rate

We evaluate control of Type I error rate by first simulating data under the null hypothesis that the weighted UniFrac distances between two ecosystems is equal to 0. Note: Ecosystems can be considered as samples or groups of samples.

$$H_0 : \beta_{ij,NWU} = 0 \tag{14}$$

$$H_A : \beta_{ij,NWU} \neq 0 \tag{15}$$

A weighted UniFrac distance of 0 suggests that the two ecosystems are no different in their community composition. A weighted UniFrac of 0 can occur in two scenarios: (1) when the relative abundances of taxa are the same and present in both samples/communities or (2) if the branch lengths of the tree are all equal to 0.

6.1.1 Step 1: Simulate data under the null hypothesis that the UniFrac distances between ecosystems are equal to 0.

To simulate data under the null hypothesis that the weighted UniFrac distance between two ecosystems is 0, we first construct a $p \times (Q + 1)$ matrix of β's generated from a random normal distribution with $\mu = 0$ and $SD = 1$. We set $p = 1$ and choose $X = (1_n^T)$. Using our vector of β's and the design matrix $X = (1_n^T)$ we create a matrix of abundances on the log ratio scale by taking the product of β and X to calculate $Y_i = X_i^T \beta$. From there we set the last taxon as D and calculate $\mathrm{Z}_i = \phi^{-1}(Y_i)$ for the ith row to obtain the Z matrix of relative abundances. We then simulate W_i from a $Multinomial(M_i, Z_i)$.

6.1.2 Step 2: Generate PhyloDivNet estimates for weighted UniFrac between ecosystems along with the parametric bootstrapped variance estimate.

With W_i generated so that the relative abundances are approximately equivalent between all ecosystems we choose $p = 2$ and define our design matrix as $X = (1_n^T, (0_{n/2}, 1_{n/2})^T)$. We follow a log-ratio approach to estimating the latent composition matrix Z as outlined in detail in section 2 and obtain our PhyloDivNet estimates of the weighted UniFrac distance between ecosystem i and ecosystem j (Eq. 9). We use a parametric bootstrapping approach to estimate the $Var(\hat{\beta}_{ij,NWU})$ as defined in section 3.

6.1.3 Step 3: Define the different hypothesis testing procedures:

We evaluate PhyloDivNet under three different hypothesis testing procedures and compare the type I error rate control to the plug-in + PERMANOVA approach to testing for differences in community composition between ecosystems. The first hypothesis testing procedure consists of constructing a Wald

statistic using PhyloDivNet's estimates of the weighted UniFrac $\hat{\beta}_{ij,NWU}$ and the $Var(\hat{\beta}_{ij,NWU})$.

$$W^2 = \frac{(\hat{\beta}_{ij,NWU} - \beta_{ij,NWU})^2}{Var(\hat{\beta}_{ij,NWU})} \tag{16}$$

We compare this Wald test statistic which we will refer to as W^2_{obs} to a chi-square distribution with n-p degrees of freedom and specify $\alpha = 0.05$. An $\alpha = 0.05$ is the rejection of the null hypothesis 5% of the time when the null hypothesis is true. **What is an appropriate degrees of freedom in this situation? I thought that since I chose p = 2 to estimate the latent composition matrix Z it might be reasonable to say n-2 degrees of freedom?**

The second hypothesis testing procedure also constructs the Wald test statistic W^2_{obs} as in Eq. 16 but instead of comparing to a chi-square distribution we compare it to a nonparametric bootstrapped distribution of the Wald test statistics. Specifically we can generate a distribution of our Wald test statistics by first defining n_{sub} as the number of samples we want to subsample from a dataset (W,X) with i = 1,..., I indexing the samples. We use a uniform random selection with replacement of n_{sub} elements from 1,...,n_{sub} that corresponds to the sample indices selected for subsampling from some dataset (W,X). For each subsampled set P we estimate $\hat{\beta}^P$ and $\hat{\Sigma}^P$ using the log ratio approach described in section 2 and obtain PhyloDivNet estimates of the weighted UniFrac $\hat{\beta}^P_{ij,NWU}$. Using $\hat{\beta}^P$ and $\hat{\Sigma}^P$ we take a parametric bootstrapping approach to estimating the $Var(\hat{\beta}_{ij,NWU})$ as described in Section 3 for 10 times. We then calculate the Wald test statistic for this subsample P as defined in Eq. 16 and refer to this as W^2_B. We repeat this entire process B times to generate a distribution of B number of Wald test statistics for a given (W,X). The desired p-value can be expressed as

$$p - value = \frac{\#oftimes W^2_B > W^2_{obs}}{B} \tag{17}$$

We reject the H_0 if the $p - value < \alpha = 0.05$.

The third hypothesis testing procedure similarly constructs the Wald test statistic W^2_{obs} as in Eq. 16 but compares the test statistic to a permutation derived distribution of Wald test statistics. The permutational approach posits that if the null hypothesis is true then the groupings and the individuals within the groups should be interchangeable. We can compute the Wald test statistic distribution by first shuffling the group labels onto different individuals for a given (W,X). For each reshuffling of labels (for R times) we calculate the PhyloDivNet estimates for weighted UniFrac $\hat{\beta}_{ij,NWU}$ and $Var(\hat{\beta}_{ij,NWU})$ to then calculate the W^2_R. We repeat this permutational reshuffling for R times for a given (W,X) to get R number of Wald test statistics that serve as our distribution of Wald test statistics. The desired p-value can be expressed as

$$p - value = \frac{\#oftimes W^2_R > W^2_{obs}}{R} \tag{18}$$

We reject the H_0 if the $p - value < \alpha = 0.05$.

6.2 Different Sample Sizes

For all simulations we are interested in varying sample sizes and seeing how this might have an affect on our hypothesis testing procedures ability to control type 1 error rates.

What do I expect to happen with differing sample sizes to the type 1 error control?

6.3 Different Bacterial Community Sizes

What do I expect to happen with differing q size to the type 1 error control?

6.4 Different Tree Structures

What do I expect to happen under the following 3 tree-relative abundance distribution situations? 3 different tree situations: - Lower abundance organisms are related to longer branches/ Larger abundance on shorter branches - Lower abundance organisms are on shorter branches/ Larger abundance on longer branches - Relatively equal abundances on equal branch lengths

7 Ignore Me! Extra meandering thoughts maybe useful later. somewhat maniacal.

Alternative proposed way to generate the data:

To simulate data under the null hypothesis we first construct a vector of q relative abundance proportions that sum to 1. We do that by randomly drawing q elements from a uniform distribution between (0,1). The relative proportions of each element q are then calculated by dividing each element by the sum of the elements. With a vector of relative abundances for each q taxa we construct our count table of abundances by drawing from a multinomial distribution with the specified proportions and a sequencing depth $M = 10,000$ for each sample i.

So we want to be sure we're generating count data that is equivalent. Why wouldn't I just use an design matrix that is intercept only and everyone should be identical?

So once we're assured that the data that we're generating is arising from identical populations with the same mean relative abundances between both populations then PhyloDivNet should be able to estimate that these mean relative abundances are similar and thereby their UniFrac distances should be equal to 0 regardless of how I set up my design matrix for inferring relative abundances. Is this one degree of freedom?

So what are we trying to evaluate here? With PhyloDivNet and DivNet in general we are using a regression approach with covariates as predictors of the expected value of the log ratio abundances. So we interpret Y_i as the difference in the log ratio between those who are sick versus those who aren't (Or the

expected mean value). We can plug in our covariate values for each sample and get the expected mean log ratio abundance for a given taxon in each sample. We transform the log ratios back into their compositional scale so what we're left with is the expected mean transformed difference in abundance between those who are sick versus those who aren't. So we then have the predicted value for the relative abundance of taxon q in ecosystem i. With these predicted values we plug them into UniFrac to get the predicted weighted UniFrac between ecosystems. So let's test the null hypothesis that the predicted weighted UniFrac between ecosystems = 0. To test this we need to simulate data under the null hypothesis that the weighted UniFrac between ecosystems is 0. The weighted UniFrac between ecosystems is 0 when the relative abundances are the same between ecosystems.

We want to evaluate different sample sizes with equal groups. Different sample sizes with unequal group sizes. We want to see differing q's. We also want to look at different phylogenetic tree topologies.

We fix $n = 10$, $\sigma_{min} = 0.01$, $\sigma_{max} = 5$, and evaluate varying $Q = 20, 50, 100$ for 50 simulations under a log-ratio model.

- hypothesis testing evaluation section here

8 Discussion

- Discuss the idea that groups of people based on some design matrix/covariates can be considered replicates in the same way that we do with other regression methods when testing for differences between groups
- Why we didn't estimate unweighted UniFrac
- Emphasize the importance of type 1 error control and hypothesis testing
- If people want to make inferential estimates of individual-level pairwise UniFracs then they need to incorporate biological replicates.

9 Conclusion

itemize)

`github.com/statdivlab/PhyloDivNet`

Martín-Fernández 2003 https://link.springer.com/content/pdf/10.1023

Vu 2007 https://statistics.berkeley.edu/sites/default/files/tech-reports/727.pdf

Paradis 2004 https://cran.r-project.org/web/packages/ape/index.html

Barberán https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2728295/

Goodrich 2014 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5074386/S31title