BIOINFORMATICS ORIGINAL PAPER
Vol. 28 no. 16 2012, pages 2106–2113
doi:10.1093/bioinformatics/bts342

Genome analysis
Advance Access publication June 17, 2012

Associating microbiome composition with environmental covariates using generalized UniFrac distances

Jun Chen[1], Kyle Bittinger[2], Emily S. Charlson[2,3], Christian Hoffmann[2], James Lewis[1,4], Gary D. Wu[4], Ronald G. Collman[2,3], Frederic D. Bushman[2] and Hongzhe Li[1,*]

[1]Department of Biostatistics and Epidemiology, [2]Department of Microbiology, [3]Pulmonary, Division of Allergy and Critical Care and [4]Division of Gastroenterology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: The human microbiome plays an important role in human disease and health. Identification of factors that affect the microbiome composition can provide insights into disease mechanism as well as suggest ways to modulate the microbiome composition for therapeutical purposes. Distance-based statistical tests have been applied to test the association of microbiome composition with environmental or biological covariates. The unweighted and weighted UniFrac distances are the most widely used distance measures. However, these two measures assign too much weight either to rare lineages or to most abundant lineages, which can lead to loss of power when the important composition change occurs in moderately abundant lineages.

Results: We develop generalized UniFrac distances that extend the weighted and unweighted UniFrac distances for detecting a much wider range of biologically relevant changes. We evaluate the use of generalized UniFrac distances in associating microbiome composition with environmental covariates using extensive Monte Carlo simulations. Our results show that tests using the unweighted and weighted UniFrac distances are less powerful in detecting abundance change in moderately abundant lineages. In contrast, the generalized UniFrac distance is most powerful in detecting such changes, yet it retains nearly all its power for detecting rare and highly abundant lineages. The generalized UniFrac distance also has an overall better power than the joint use of unweighted/weighted UniFrac distances. Application to two real microbiome datasets has demonstrated gains in power in testing the associations between human microbiome and diet intakes and habitual smoking.

Availability: http://cran.r-project.org/web/packages/GUniFrac

Contact: hongzhe@upenn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 15, 2012; revised on June 6, 2012; accepted on June 8, 2012

1 INTRODUCTION

Understanding the compositional differences of microbial communities (microbiomes) is essential in microbial ecology. With the development of next-generation sequencing technologies, microbiome composition can now be determined by direct DNA sequencing without the need for laborious cultivation (Dinsdale *et al.*, 2008; Gill *et al.*, 2006; Grice *et al.*, 2009; Qin *et al.*, 2010; Tringe *et al.*, 2005; Turnbaugh *et al.*, 2009; von Mering *et al.*, 2007). There has been a great interest in human microbiome studies in different body sites, ranging from skin (Grice *et al.*, 2009) to gut (Arumugam *et al.*, 2011; Muegge *et al.*, 2011; Qin *et al.*, 2010; Wu *et al.*, 2011) and respiratory tract (Charlson *et al.*, 2010). Important insights have been gained from analysis of large-scale human microbiome data, including the discovery of enterotypes (Arumugam *et al.*, 2011) and discovery of the link between diet and these enterotypes (Wu *et al.*, 2011). Although the metagenomic shotgun approach is potentially more powerful and unbiased, 16S rRNA gene targeted sequencing is routinely performed to determine the taxonomic composition. The generated 16S rRNA sequence tags are usually clustered into operational taxonomic units (OTUs) with a specified amount of variability allowed within each OTU (Caporaso *et al.*, 2010; Schloss *et al.*, 2009). At 97% similarity, these OTUs represent ‘species’. Downstream analysis is then performed on the OTU abundance data.

Two central themes in human microbiome studies are to identify potential biological and environmental factors that are associated with microbiome composition, and to define the relationship between microbiome features and biological or clinical outcomes. The goal is to provide a better understanding of the factors that shape our microbiome and, potentially, contribute to the development of new therapeutic strategies to modulate the microbiome composition and affect human health (Spor *et al.*, 2011; Virgin and Todd, 2011). Testing the association of microbiome composition with potential environmental factors using OTU abundances directly is difficult due to high dimensionality, non-normality and phylogenetic structure of the OTU data. Instead, distance-based non-parametric testing, in which a distance measure is defined-between any two microbiome samples, is usually used to achieve this goal (Charlson *et al.*, 2010; Fukuyama *et al.*, 2012; Kuczynski *et al.*, 2010a; Wu *et al.*, 2010, 2011). The power of the distance-based test depends on a proper choice of the distance measure. Numerous distance measures have been proposed to compare microbial communities (Kuczynski *et al.*, 2010b; Swenson, 2011). Phylogenetic distance measures, which account for the phylogenetic relationship among the species, provide far more power because they exploit the degree of divergence between different sequences. Among these, the UniFrac distances are the most popular ones (Lozupone and

[*]To whom correspondence should be addressed.

Knight, 2005; Lozupone *et al.*, 2007). There are two versions of UniFrac distances: an unweighted UniFrac distance that considers only species presence and absence information and counts the fraction of branch length unique to either community, and a weighted UniFrac distance that uses species abundance information and weights the branch length with abundance difference. Unweighted UniFrac distance is most efficient in detecting abundance change in rare lineages. When the abundance of a rare lineage falls below a certain threshold, the sequencing machine may not be able to pick it up and it will appear absent in the final dataset. On the other hand, weighted UniFrac distance is most sensitive to detect change in abundant lineages since it uses absolute abundance difference in its definition. However, unweighted/weighted UniFrac distances may not be very powerful in detecting change in moderately abundant lineages. Recently, a variance-adjusted weighted UniFrac distance (VAW-UniFrac), which moderates the branch proportion difference by its variance, was developed to account for the fact that weighted UniFrac distance does not consider the variation of the weights under random sampling (Chang *et al.*, 2011). VAW-UniFrac was shown to increase the power over weighted UniFrac distance for detecting the difference between two microbial communities.

In this article, we introduce generalized UniFrac distances that unify weighted UniFrac and unweighted UniFrac distances. The new generalized UniFrac distances cover a series of distances ranging from weighted to unweighted UniFrac by adjusting the weight on the branches. The generalized UniFrac distances are designed to provide a robust and powerful tool for detecting a wider range of biologically relevant changes in microbiome composition. We conduct extensive Monte Carlo simulation studies under various conditions to evaluate their power in detecting environmental influence on microbiome composition using PERMANOVA (McArdle, 2001), a distance-based non-parametric test. Although each distance in the series can perform the best in certain scenarios, none has the optimal performance under all conditions considered. However, analyses based on the generalized UniFrac distances are shown to be more robust and has overall the best performances across a range of possible scenarios. We demonstrate the power gain of using this distance in detecting the microbiome differences by analysis of two real human gut microbiome datasets related to linking human gut microbiome composition to long-term diet (Wu *et al.*, 2011) and testing upper respiratory tract microbiome difference between smokers and non-smokers (Charlson *et al.*, 2010).

2 METHODS

2.1 Generalized UniFrac distances between two microbial communities

Consider two microbiome communities A and B and suppose that we have a rooted phylogenetic tree with n branches. Let b_i be the length of the branch i and p_i^A and p_i^B are the taxa proportions descending from the branch i for community A and B, respectively. The unique fraction metric, or UniFrac, measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both. The original definition refers to unweighted UniFrac (Lozupone and Knight, 2005), which is mathematically defined as

$$d^U = \sum_{i=1}^{n} \frac{b_i \left| I(p_i^A > 0) - I(p_i^B > 0) \right|}{\sum_{i=1}^{n} b_i},$$

where $I(.)$ is the indicator function and only presence/absence of species of branch i, $I(p_i^A > 0)$ and $I(p_i^B > 0)$, are used in the definition. The distance definition d^U completely ignores the taxa abundance information. In contrast, the (normalized) weighted UniFrac distance (Lozupone *et al.*, 2007) weights the branch length with abundance difference and is defined as

$$d^W = \frac{\sum_{i=1}^{n} b_i \left| p_i^A - p_i^B \right|}{\sum_{i=1}^{n} b_i (p_i^A + p_i^B)}.$$

Note that d^W cannot be reduced to d^U even if we convert abundance data into presence/absence data. Also note that d^W uses the absolute proportion difference $\left| p_i^A - p_i^B \right|$ in its formulation. The consequence of using the absolute difference is that the value of d^W is determined mainly by branches with large proportions and is less sensitive to the abundance changes on the branches with small proportions. To attenuate the weight on branches with large proportions, we may instead use the relative difference $\left| p_i^A - p_i^B \right| / (p_i^A + p_i^B)$ ($\in [0,1]$) in the formulation. We denote this distance measure as

$$d^{(0)} = \frac{\sum_{i=1}^{n} b_i \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{n} b_i},$$

where $\sum_{i=1}^{n} b_i$ in the denominator is the normalizing factor so that $d^{(0)} \in [0,1]$. Now if we dichotomize the abundance data using the indication function $I(.)$, $d^{(0)}$ is reduced to d^U. So $d^{(0)}$ can be seen as the 'weighted version' of d^U. Using the relative differences, we place equal emphasis on every branch and the distance is not dominated by the branches with large proportions, since the relative difference does not depend on the magnitude of p_i^A, p_i^B. However, the low-abundance branches may be more noisy and the relative difference may amplify such noises. To strike a balance between relative difference and absolute difference, we weight the branch length both by the relative difference and its importance indicated by the branch proportion. We propose the following generalized UniFrac distances

$$d^{(\alpha)} = \frac{\sum_{i=1}^{n} b_i (p_i^A + p_i^B)^{\alpha} \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{n} b_i (p_i^A + p_i^B)^{\alpha}},$$

where $\alpha \in [0,1]$ controls the contribution from high-abundance branches, and $\sum_{i=1}^{n} b_i (p_i^A + p_i^B)^{\alpha}$ is the normalizing factor so that $d^{(\alpha)} \in [0,1]$. Branches with zero proportions for both communities will not be included in the calculation. As α changes from 0 to 1, more emphasis is placed on high-abundance branches. When $\alpha = 1$, $d^{(\alpha)}$ is reduced to d^W. When $\alpha = 0$, we get $d^{(0)}$ defined above.

Therefore, by varying α from 1 to 0 , we achieve a series of distances ranging from d^W to $d^{(0)}$. Note that d^U is obtained by dichotomizing the abundance in $d^{(0)}$, but is different from $d^{(0)}$. We are particularly interested in $d^{(0.5)}$, the distance in the middle of the distance series

$$d^{(0.5)} = \frac{\sum_{i=1}^{n} b_i \sqrt{p_i^A + p_i^B} \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{n} b_i \sqrt{p_i^A + p_i^B}}.$$

We also compare $d^W, d^{(0.5)}, d^{(0)}$ and d^U with VAW-UniFrac distance d^{VAW} (Chang *et al.*, 2011), which is defined as:

$$d^{VAW} = \frac{\sum_{i=1}^{n} b_i \frac{\left| p_i^A - p_i^B \right|}{m(m - m_i)}}{\sum_{i=1}^{n} b_i \frac{p_i^A + p_i^B}{m(m - m_i)}},$$

where m_i is the total number of individuals/reads from both communities on the ith branch and m is total number of individuals/reads.

2.2 Statistical test based on UniFrac distances

We study the power of generalized UniFrac distances using the distance-based non-parametric test for association of microbiome composition with environmental covariates. Suppose we have a set of m environmental covariates. We assume that we have collected microbiome data and the m-dimensional covariates data $\mathbf{X}$ on n samples. We apply the PERMANOVA procedure (McArdle, 2001) [Permutational Multivariate Analysis of Variance Using Distance Matrices, 'adonis' function from R package 'vegan' (Oksanen *et al.*, 2011)], which partitions the distance matrix among sources of variation, fits linear models to distance matrices and uses a permutation test with pseudo-F ratios to obtain the P-values. The pseudo-F statistic is defined as

$$F=\frac{\mathrm{tr}(\mathbf{HGH})/(m-1)}{\mathrm{tr}[(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})]/(n-m)},$$

where tr(.) is the trace function of a matrix, $\mathbf{H}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat (projection) matrix of the design matrix $\mathbf{X}$, $\mathbf{G}$ is Gower's centered matrix and n and m is the number of samples and the number of predictors, respectively. Let d_{ij} be the generalized UniFrac distance between community i and j and denote $\mathbf{A}=(a_{ij})=(-\frac{1}{2}d_{ij}^2)$. The Gower's matrix is defined as

$$\mathbf{G}=\left(\mathbf{I}-\frac{\mathbf{11}'}{n}\right)\mathbf{A}\left(\mathbf{I}-\frac{\mathbf{11}'}{n}\right),$$

where $\mathbf{1}$ is a vector of 1's.

Since d^U and d^W reflect the abundance change in either rare lineages or abundant lineages, combining d^U and d^W may potentially increase the overall power. Instead of applying Bonferroni correction to the P-values from separate PERMANOVA tests using d^U or d^W to control the family-wise Type I error rate, a more powerful approach is to take the maximum of pseudo-F statistics for d^U and d^W as a new test statistic. The significance of the pseudo-F statistics is assessed based on permutations.

2.3 Simulation strategies

We use two simulation strategies to evaluate the power of the generalized UniFrac distances under various conditions. The first strategy is a modification of the simulation method proposed by Schloss (2008), where we draw points (16S rRNA sequences) from a 2D circle with known densities (Fig. 1A). This strategy facilitates simulations of different community characteristics such as species evenness and species richness. The Euclidean distance between points is analogous of the genetic distance between the sequences. The diameter of the circle represents the maximum genetic divergence between any pair of sequences within a sample. The area of the circle is proportional to the richness and the density distribution of the circle is proportional to the evenness. By varying the centroid positions (o) and their radius (r), it is possible to vary the fraction of shared membership and species richness within each sample (Fig. 1B and D). By varying the point distribution on the circle (density proportional to r^{α}, where α controls the degree of evenness and $\alpha=0.5$ for uniformly distribution), it is possible to change the species evenness (Fig. 1C). We also simulate scenarios where lineages of different abundance levels change by a k fold (Fig. 1E–G). These are achieved by simulating the community with point mass concentrated at the circle center ($r^{1.0}$) and varying the point density in different regions of the 2D circle corresponding to abundant lineages ($0-0.2r$ from the center; Fig. 1E), moderately abundant lineages ($0.4r-0.8r$ from the center; Fig. 1F) and rare lineages ($0.8r-1.0r$ from the center; Fig. 1G). We further bin the sampled points into small hexagons as 'OTU's before calculating the UniFrac distance ['hexbin' function from the R package 'hexbin' (Carr *et al.*, 2011)]. The phylogenetic tree of these 'OTU's is built using NJ algorithm (Neighbor Joining, 'nj' function in R) and rooted by midpoint rooting method. Generalized UniFrac distances are then calculated based on the NJ tree and 'OTU' abundances. Each replication consists of drawing 400 points from each community, a bin size of 0.015 units to form 'OTUs' (~300 OTUs per sample), and the maximum distance between any two points is 0.3 units ($r=0.15$), corresponding to typical phylum level divergence of 30% for 16S rRNA gene. These conditions allow us to simulate the sampling intensity and biodiversity found within a typical 16S rRNA gene targeted sequencing experiment (Schloss, 2008).

The second set of simulations utilize a real upper respiratory tract microbiome dataset consisting of 60 samples and 856 OTUs from Charlson *et al.* (2010) (Fig. 1H). A common way of modeling multivariate count data is to use the multinomial model. However, the multinomial model assumes fixed underlying proportions for each sample, which do not hold for real microbiome data due to high degree of heterogeneity among the samples. The real OTU count distribution (Supplementary Fig. S1A) exhibits more variance than expected from a multinomial model (Supplementary Fig. S1B). To realistically simulate the data, it is important to model extra-variation or overdispersion of the OTU counts. This can be achieved by using the Dirichlet-multinomial (DM) model (Mosimann, 1962), which assumes the underlying proportions of the multinomial model come from a Dirichlet distribution. The density function of a DM random variable N is given as

$$P(N=n)=\binom{n}{n}\frac{\prod_{j=1}^{k}\prod_{r=1}^{n_j}\{\pi_j(1-\theta)+(r-1)\theta\}}{\prod_{r=1}^{n}\{1-\theta+(r-1)\theta\}},$$

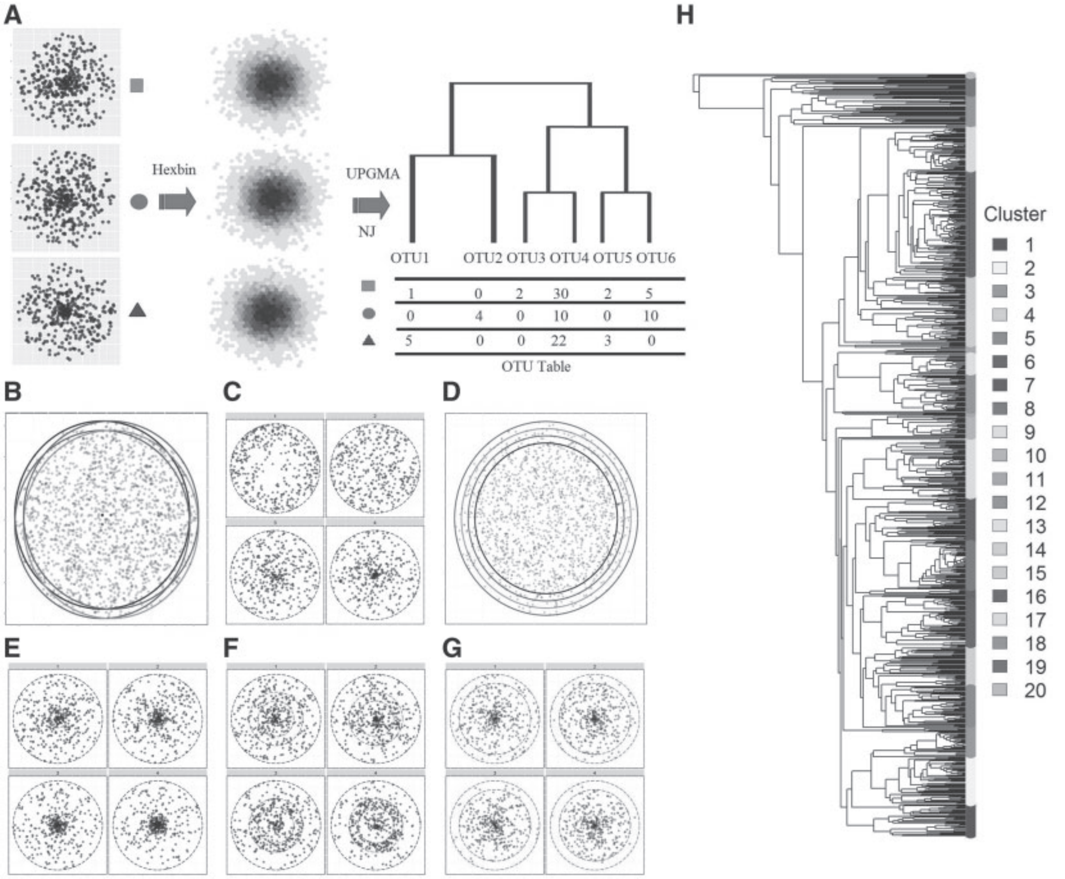
where $n=\sum_j n_j$ is total count, k is the OTU number and proportion mean $\pi=(\pi_1,\pi_2,\cdots,\pi_k)$ and dispersion θ are parameters. When $\theta=0$, it is reduced to multinomial model. We estimate the DM parameters π,θ using maximum likelihood method ('dirmult' function from R package 'dirmult'). We then generate OTU counts using the DM model with the estimated parameters and 1000 counts per sample. Supplementary Figure S1C shows an OTU heatmap generated by the DM model, in which the overdispersion is similar to that of the real data. To study the power of UniFrac variants for identifying potential environmental factors, we let the abundance of a certain OTU cluster change in response to environment. We use the UPGMA tree of the OTUs based on the OTU distance matrix calculated under the K80 nucleotide substitution model (Felsenstein, 2004), QIIME (FastTree algorithm (**?**)) and partition the 856 OTUs into 20 clusters using Partitioning Around Medoids (PAM) ('pam' function from R package 'cluster') based on patristic distances (the length of the shortest path linking two OTUs on the tree). These OTU clusters are highlighted in different colors in Figure 1H.

We call the first strategy 2D circle-based simulation and the second tree-based simulation. For power calculation, we use 2000 replications.

3 RESULTS

3.1 Comparison of the power of different UniFrac variants using 2D circle-based simulations

We use PERMANOVA to test for environmental effect and compare the power of d^W,$d^{(0.5)}$,$d^{(0)}$, d^U and d^{VAW}. Specifically, we simulate two environmental conditions (e.g. smoking versus non-smoking) under which we draw 10 samples each. We then vary the degree of community difference under these two conditions and produce the power curve over a grid of 10 for each UniFrac distance. We investigate six scenarios, where the environmental factor affects the community membership, species evenness, species richness, most abundant lineages moderately abundant lineages and rare lineages (Fig. 1B–G). For each scenario, we vary one community characteristic (Supplementary Table S1). Suppose x_1 and x_2 are the mean values of the community characteristic for Conditions 1 and 2. We simulate 10 communities for each condition with community



	OTU1	OTU2	OTU3	OTU4	OTU5	OTU6
■	1	0	2	30	2	5
●	0	4	0	10	0	10
▲	5	0	0	22	3	0

Fig. 1. Two simulation strategies to evaluate the generalized UniFrac distances. (A–G), 2D circle-based simulation of microbial communities with different characteristics. (**A**) The microbial community is represented by a 2D circle. Points are drawn from the circle to simulate the 16S-based sampling process. These points are further binned into small hexagons as OTUs. UPGMA or NJ method is used to build the OTU phylogenetic tree. Six scenarios are investigated, where the difference occurs in: community membership (**B**), evenness (**C**), richness (**D**), most abundant lineages (**E**), moderately abundant lineages (**F**) and rare lineages (**G**). The affected lineages are indicated by a red circle or ring. H, tree-based simulation of microbial communities based on the phylogenetic tree and DM model. A real OTU phylogenetic tree from a throat microbial community dataset is used. These OTUs are roughly divided into 20 clusters (lineages) by performing PAM method using the OTU patristic distance matrix. Each cluster is subjected to abundance change in response to the environment. Counts are generated from a DM model.

characteristic value $x_{ij} \sim \text{Uniform}(x_j - s, x_j + s)$ for $i = 1 \dots 10$ and $j = 1, 2$, where s controls the variation within each condition and takes different values for different scenarios (see Supplementary Table S1). Each community is sampled once. Initially, we let $x_1 = x_2$ (no difference) and then increase the difference between x_1 and x_2 to simulate stronger environmental effect. PERMANOVA is then performed on the distance matrices and the power curve is created over a grid of 10 using Type I error $\alpha = 0.05$. Figure 2 shows the power curves for different UniFrac distances under the six scenarios considered. When the environmental factor has no effect ($x_1 = x_2$), PERMANOVA controls the Type I error at the nominal level of 0.05 for all five UniFrac distances. As the environmental effect becomes stronger, all the distances have better power. When the environmental factor affects the community membership or richness (Panels 1 and 3), all the distances give a similar power and their power curves are nearly identical. For the evenness change scenario (Panel 2), the power of d^W and $d^{(0.5)}$ is very close and is more powerful than $d^{(0)}$ and d^U. d^W is the most powerful for detecting change in most abundant lineages (Panel 4) but is much less powerful for change in rare lineages (Panel 6). d^U shows an opposite trend: it is the most powerful for detecting change in rare lineages (Panel 6) but has almost no power for change in most abundant lineages (Panel 4). In contrast, $d^{(0.5)}$ is the most powerful for detecting change in moderately abundant lineages (Panel 5). They are also the most robust among the distances investigated: its power is close to the best UniFrac distance under all scenarios. The performance of $d^{(0)}$ lies between $d^{(0.5)}$ and d^U and is also very robust. Finally, under the 2D circle simulations, the performance of d^{VAW} is almost identical to that of $d^{(0.5)}$.

In the above simulations, we use a bin size of 0.015 to form ‘OTU’s (~300 OTUs per sample). To study the effect of bin size, we compare the power curves of the generalized UniFrac distances using a smaller bin size of 0.01 (~700 OTUs per sample) or a larger bin size of 0.03 (~80 OTUs per sample). The bin size does not change the general conclusion (Supplementary Fig. S2). To study the effect of tree construction methods, we also construct the

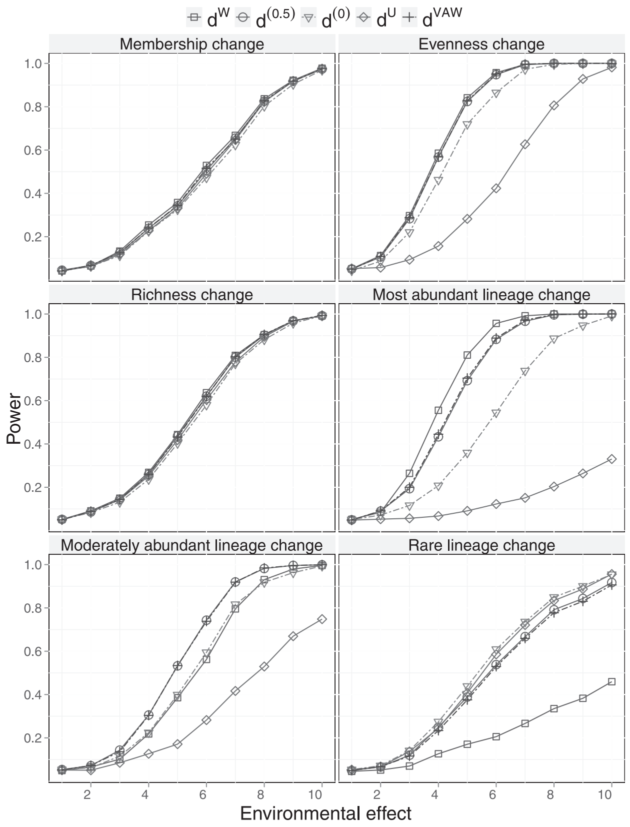


Fig. 2. Power comparison of different UniFrac variants for detecting environmental effect using 2D circle-based simulation. PERMANOVA is used for testing hypotheses. The specific community difference caused by different environmental conditions is indicated in the panel title. The power curves are created by varying the degree of environmental effect. The initial point of the power curve is the power when there is no environmental effect.

phylogenetic tree using UPGMA. The general conclusions still hold (Supplementary Fig. S3).

3.2 Comparison of the power of different UniFrac variants using tree-based simulations

We also compare the power of different UniFrac distances for detecting environmental effect using tree-based simulations that mimic the throat microbiome data (see Section 3.4 for details). A recently proposed variance adjusted UniFrac distance (d^{VAW}) (Chang *et al.*, 2011) is also compared in this setting. d^{VAW} was developed to moderate the branch proportion difference by its variance and was shown to increase the power of detecting the difference between two microbial communities. We use the phylogenetic tree of the 856 OTUs from the throat microbiome dataset and divide them into 20 clusters (Fig. 1H). The mean OTU proportions and the dispersion parameter are estimated from the real data by fitting a DM model. We assume that the environmental factor causes an increase of the abundance of a particular OTU cluster. Specifically, suppose that the proportion of the ith OTU cluster under Condition 1 is p_i. For Condition 2, the proportion of ith OTU cluster is increased by k fold where k varies from 1 (no difference) to $1/\sqrt{p_i}$ (strong effect) on a grid of 10. The proportion vector is re-normalized to sum to 1. Next, 10 samples are simulated for each condition with their OTU counts generated by the DM model with the corresponding proportion vector and the common dispersion parameter. As expected, the five UniFrac distances differ in their power for detecting environmental effect for the 20 OTU clusters tested. Except for $d^{(0)}$, all the UniFrac distances have their best-performance scenarios. d^W, $d^{(0.5)}$, d^U and d^{VAW} achieve the highest power in seven, six, three and one cases, respectively. For the remaining three cases, d^W and $d^{(0.5)}$ are equally the most powerful (Supplementary Fig. S4). The results are consistent with the 2D circle-based simulation: d^W is most powerful in detecting the environmental effect on most abundant lineages, $d^{(0.5)}$ is most powerful for moderately abundant lineages and d^U is most powerful for rare lineages. In contrast, performance of the test with $d^{(0)}$ and d^{VAW} is generally between d^U and $d^{(0.5)}$. The power of d^W and d^U has a reciprocal relationship and neither of them is as robust as $d^{(0.5)}$. Figure 3A shows the power curves of four representative cases. As the proportion of the affected cluster decreases from 19.7% to 0.9%, d^W becomes less powerful and the power of d^U has the opposite trend.

In the simulations presented above, the power is calculated assuming the affected cluster is known. Since the affected cluster can be abundant or rare, we randomly choose an affected OTU cluster in each replication and calculate the power over 2000 replications. We also report the power for the test combining d^W and d^U by taking the maximum of their pseudo-F statistics. We denote this method as d^{MAX}. Figure 3B (left plot) shows that d^U has the lowest overall power and $d^{(0.5)}$ and d^{MAX} have the best power, indicating combining d^U and d^W can lead to power gain. In contrast, the power of $d^{(0)}$, d^{VAW} and d^W is in between. As the environmental effect becomes stronger, $d^{(0)}$ becomes as powerful as $d^{(0.5)}$ and d^{MAX}. Finally, we assume that the environmental factor affects a random set of 40 OTUs instead of a random OTU cluster. At this extreme where the phylogenetic relationship is no longer important, $d^{(0.5)}$ has even higher power than the other distances, followed by $d^{(0)}$, d^{MAX}, d^W, d^U and d^{VAW} (see Fig. 3B, right plot). Overall, $d^{(0.5)}$ has a better power than other UniFrac distances including the one that combines d^W and d^U.

3.3 Results from analysis of a dataset linking long-term diet to gut microbiome composition

Diet strongly affects the human health, partly by modulating gut microbiome composition. Wu *et al.* (2011) studied the habitual diet effect on the human gut microbiome, where the diet information was converted into a vector of micro-nutrient intakes. A total of 98 healthy volunteers were enrolled in this cross-sectional study. Habitual long-term diet information was collected using food frequency questionnaire (FFQ). The questionnaires were converted to intake amounts of 214 micro-nutrients. Nutrient intake was further normalized using the residual method to standardize for caloric intake. Stool samples were collected and DNA samples were analyzed by 454/Roche pyrosequencing of 16S rRNA gene segments of the V1-V2 region. The pyrosequences were denoised (Quince *et al.*, 2009) prior to taxonomic assignment yielding an average of 9265±3864(SD) reads per sample. The denoised sequences were then analyzed by the QIIME pipeline (Caporaso *et al.*, 2010) with the default parameter settings in the QIIME pipeline. The OTU table contains 3068 OTUs after discarding the singleton OTUs. One objective of the study is to identify nutrients that have a significant

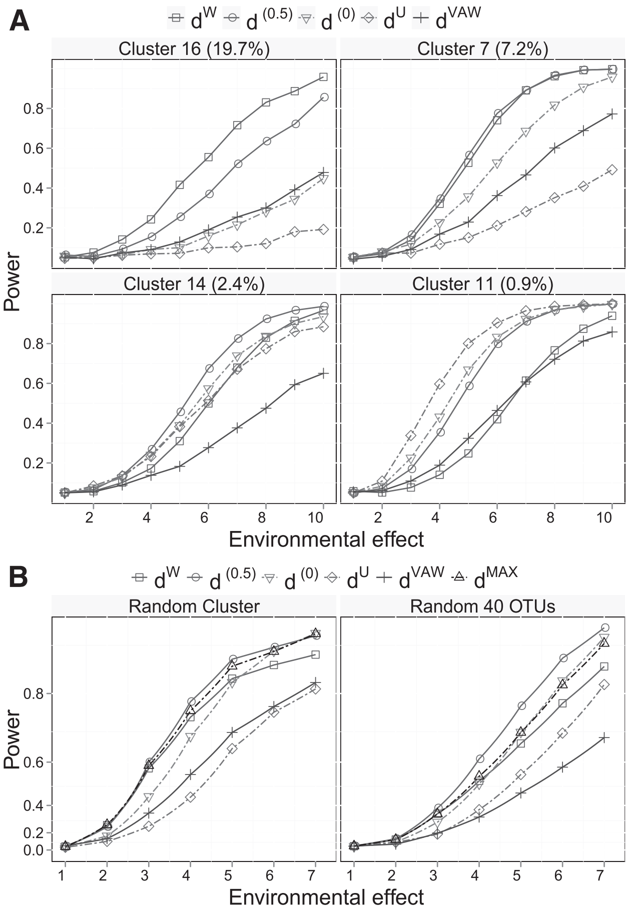


Fig. 3. Power comparison of different UniFrac variants for detecting environmental effect using tree-based simulation. PERMANOVA is used for testing hypotheses. The power curves are created by varying the degree of environmental effect. (**A**) The environmental factor affects a particular lineage (OTU cluster). Four example lineages of different abundance levels that are affected by environment are given. The lineage abundance is given in parentheses in the panel title. (**B**) The environmental factor affects a random lineage (left panel) or a random subset of 40 OTUs (right panel). The initial point of the power curve is the power when there is no environmental effect.

impact on the gut microbiome composition. We use PERMANOVA to test for association of microbiome composition with nutrient intake based on different UniFrac distance matrices. We compare $d^{(0.5)}$ with d^U, d^W, their combination d^{MAX} and d^{VAW}. We plot the number of selected nutrients against different P-value cutoffs to create a ROC-like curve (Fig. 4). For $P < 0.01$, all distances except the d^U identify the same number of nutrients. For $P > 0.03$, the curve for $d^{(0.5)}$ is above all the other four curves. Wilcoxon signed-rank tests show that $d^{(0.5)}$ results in smaller P-values than other distances ($P < 0.05$), indicating that $d^{(0.5)}$ is most powerful in selecting the relevant microbiome-associated nutrients. Using d^W or d^U alone could miss important associations when the nominal P-value of 0.03–0.05 is used. Although using a relatively large nominal P-value can certainly lead to inclusion of possible false-positive nutrient-microbiome associations, there are situations when one might want to include all possible associations for further validation or replications. d^{VAW} performs the second best. Interestingly, d^{MAX}, the joint use of d^W and d^U, does not increase the power over d^W, indicating most associations can be recovered by d^W alone.

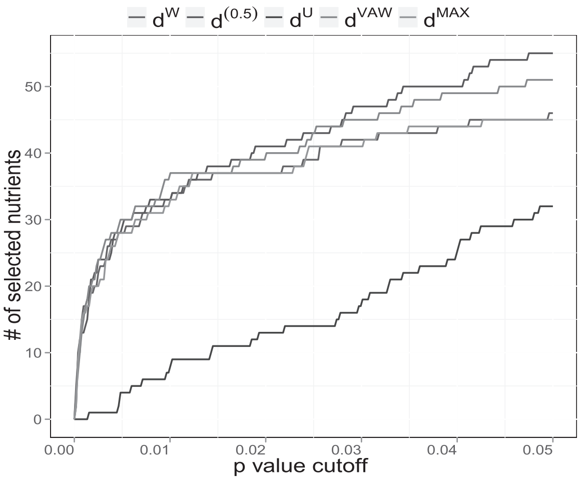


Fig. 4. Comparison of different UniFrac variants for detecting nutrient effects on gut microbiome composition. PERMANOVA is used for testing hypotheses. 214 nutrients are included in the testing. The curves are generated by varying the P-value cutoffs.

3.4 Results from analysis of a throat microbiome dataset of smokers and non-smokers

Cigarette smokers have an increased risk of multiple diseases, including upper respiratory tract infections. Previous studies had linked smoking to specific respiratory tract bacteria, but the consequences of smoking for global airway microbial community composition had not been fully clarified. Charlson *et al.* (2010), investigated the smoking effect on the oropharyngeal and nasopharyngeal bacterial communities using 454 pyrosequencing of 16S sequence tags. Specifically, a total of 291 swab samples from the right and left nasopharynx and oropharynx of 29 smoking and 33 non-smoking healthy asymptomatic adults were collected. The variable region 1-2 (V1-V2) of the bacterial 16S rRNA gene was PCR-amplified using individually barcoded primer and subject to multiplexed pryosequencing. The pyrosequences were denoised (Quince *et al.*, 2009) prior to taxonomic assignment and yielded an average of 1335±603(SD) reads per airway sample. The denoised sequences were then analyzed using the QIIME pipeline (Caporaso *et al.*, 2010) with default parameter setting. We use the left oropharyngeal samples in this study. After removing two samples with read number <500 and discarding singleton OTUs, i.e OTU with only one read, we finally have an OTU table of 60 samples (28 smokers versus 32 non-smokers) and 856 OTUs.

We test the smoking effect on the throat microbial community composition by applying PERMANOVA (10000 permutations). All the five UniFrac distances achieve statistical significance at $\alpha = 0.05$ level, indicating smoking alters the community composition. However, test using $d^{(0.5)}$ produces the smallest P-value of 0.006, followed by 0.008 from $d^{(0)}$. The P-values based on d^W, d^U and d^{VAW} are 0.012, 0.019 and 0.043, respectively. We also perform a principle coordinate analysis using the distance matrices and plot the samples on the first two principle coordinates (Fig. 5). The distance $d^{(0.5)}$ separates the samples better than the other

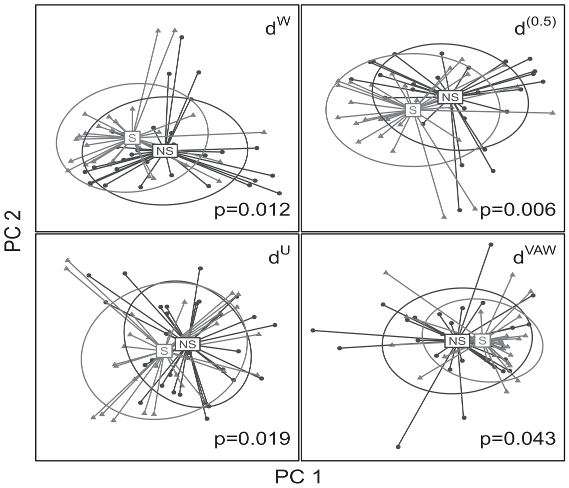


Fig. 5. Comparison of different UniFrac variants for clustering samples from smokers and non-smokers. Principle coordinate analysis is performed on the distance matrices of d^W, $d^{(0.5)}$, d^U and d^{VAW}. The samples are plotted on the first two principle coordinates. The PERMANOVA *P*-values are also indicated in this figure. The ellipse center indicates groups means, its main axis corresponds to the first two principle components from principle component analysis and the height and width are variances on that direction.

three distance measures. This indicates that smoking might affect not only the predominant lineages but also these less abundant lineages in the throat microbial community. We then perform Wilcoxon rank-sum or Fisher's exact test to select the differential OTUs. At $\alpha=0.05$ level, we identify 32 OTUs. These OTUs belong to genera *Prevotella* (8), *Lachnospiraceae* (5), *Veillonella* (3), *Streptococcus* (2), *Fusobacterium* (2), *Treponema* (2), *Neisseria* (1), *Haemophilus* (1), *Megasphaera* (1), *Dialister* (1), *Moryella* (1), *Erysipelotrichaceae* (1) and four genera from *Actinobacteria*. Most of the selected OTUs are moderately abundant or rare, so we expect $d^{(0.5)}$ and $d^{(0)}$ to have better power.

4 DISCUSSION

Microbiome data are multivariate count data in their original form and are statistically challenging to analyze due to their high dimensionality, phylogenetic constraints among species/OTUs, overdispersion and excessive zeros. To circumvent the difficulty, the data are often summarized in the form of distance matrix. Testing association of microbiome composition with environmental covariates is performed using the distance matrix. We have demonstrated in simulations that the weighted and unweighted UniFrac impose large weight either to abundant lineages or to rare lineages; they can be underpowered in detecting change in moderately abundant lineages. Since microbiome composition change could occur in any lineages, our generalized UniFrac distances, which unify the weighted and unweighted UniFrac in a common framework, enable us to detect a much wider range of biologically relevant changes. Our simulation studies have clearly demonstrated that the generalized UniFrac distance $d^{(0.5)}$ is more robust than d^W or d^U, and its performances are in general comparable to the best UniFrac distances among the scenarios we considered. In addition, the generalized UniFrac distances are very robust to tree constructing methods. We suggest the use of $d^{(0.5)}$ for testing association of microbiome composition with environmental covariates to avoid missing important findings.

Both weighted and unweighted UniFrac distances are sensitive to sampling depth (Lozupone *et al.*, 2010). Inflated distances at a low sampling depth are caused by sampling variation especially for the rare lineages. The generalized UniFrac distances are also sensitive to sampling depth (Supplementary Fig. S5). However, as the sampling depth increases, the distance stabilizes. For the gut microbiome dataset, we found a sequencing depth of ~1000 reads is sufficient to stabilize the generalized UniFrac distances. To overcome the potential adverse effects of uneven sampling, rarefaction is usually used to subsample the samples to the same depth. However, when the sampling depth varies greatly across the samples, rarefaction throws away a significant portion of the 16S reads and increases the sampling variation. We found that rarefaction is not necessary, at least, in the context of testing the association of the microbiome composition with covariates (Supplementary Fig. S6).

The power of UniFrac variants can also be compared in the context of testing whether two microbial communities differ significantly as in (Chang *et al.*, 2011; Schloss, 2008). Instead of comparing power for detecting the difference between two communities, we focus our evaluations on the performance of UniFrac distances for associating microbiome composition to environmental covariates by collecting multiple independent samples. The rational is that as the sequence depth increases, two sample comparison will have increased power to detect differences due to sources that we are not interested in (random noises), such as the individual-to-individual variability, day-to-day variability, sampling location variability or even technical variability (e.g. sample preparation). Multiple samples from a population coupled with multivariate statistical methods such as the distance-based PERMANOVA provide powerful design and analysis methods to overcome these potential random noises (Lozupone *et al.*, 2010). As more and more large-scale microbiome datasets are being collected, we expect that our generalized UniFrac distances can help to identify important covariates that are associated with the microbiomes that could be missed using the commonly used UniFrac distances. In addition to identifying environmental covariates that may be determinants of microbiome composition, our approach would be equally suited to identifying microbiome features associated with biological or clinical outcomes, which is needed to begin to understand the impact of the microbiome on health.

ACKNOWLEDGEMENT

We thank the three reviewers for very helpful comments.

Funding: This research was supported by the National Institute of Health grants R01CA127334, U01HL098957 and DK083981.

Conflict of Interest: None declared.

REFERENCES

Arumugam,M. *et al.* (2011) Enterotypes of the human gut microbiome. *Nature*, **473**, 174–180.

Caporaso,J. *et al.* (2010) Qiime allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Carr,D. *et al.* (2011) *hexbin: Hexagonal Binning Routines*.

Chang,Q. *et al.* (2011) Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*, **12**, 118.

Charlson,E. *et al.* (2010) Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS One*, **5**, e15216.

Dinsdale,E. *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629–632.

Felsenstein,J. (2004) *Inferring Phytogenies. Sunderland, MA: Sinauer Associates*.

Fukuyama,J. *et al.* (2012) Comparisons of distance methods for combining covariates and abundances in microbiome studies. *Pac. Symp. Biocomput.*, **12**, 213–224.

Gill,S. *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science*, **312**, 1355–1359.

Grice,E. *et al.* (2009) Topographical and temporal diversity of the human skin microbiome. *Science*, **324**, 1190–1192.

Kuczynski,J. *et al.* (2010a) Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol.*, **11**, 210–218.

Kuczynski,J. *et al.* (2010b) Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods*, **7**, 813–819.

Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8835.

Lozupone, C. *et al.* (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, **73**, 1576–1585.

Lozupone,C. *et al.* (2010) Unifrac: an effective distance metric for microbial community comparison. *ISME J.*, **5**, 169–172.

McArdle,B. (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, **82**, 290–297.

Mosimann,J. (1962) On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, **49**, 65–82.

Muegge, B. *et al.* (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, **332**, 970–974.

Oksanen,J. *et al.* (2011) *vegan: Community Ecology Package*.

Qin,J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

Quince,C. *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods*, **6**, 639–641.

Schloss,P. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537.

Schloss,P. (2008) Evaluating different approaches that test whether microbial communities have the same structure. *ISME J.*, **2**, 265–275.

Spor,A. *et al.* (2011) Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.*, **9**, 279–290.

Swenson,N. (2011) Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. *PloS One*, **6**, e21264.

Tringe,S. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.

Turnbaugh,P. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

Virgin,H. and Todd,J. (2011) Metagenomics and personalized medicine. *Cell*, **147**, 44–56.

von Mering,C. *et al.* (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, **315**, 1126–1130.

Wu,G. *et al.* (2010) Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol.*, **10**, 206.

Wu,G. *et al.* (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science*, **334**, 105–108.