

## ORIGINAL ARTICLE

# Evaluating different approaches that test whether microbial communities have the same structure

Patrick D Schloss

Department of Microbiology, University of Massachusetts—Amherst, Amherst, MA, USA

As microbial ecology investigations have progressed from descriptive characterizations of a community to hypothesis-driven ecological research, a number of different statistical techniques have been developed to describe and compare the structure of microbial communities. Thus far, these methods have only been evaluated using 16S rRNA gene sequence data obtained from incomplete characterizations of microbial communities. In this investigation, simulations were designed to test the statistical power of different methods to differentiate between communities with known memberships and structures. These simulations revealed three important results that affect how the results of the tests are interpreted. First,  $\beta$ -LIBSHUFF, TreeClimber, UniFrac, analysis of molecular variance (AMOVA) and homogeneity of molecular variance (HOMOVA) compare the structure of communities and not just their memberships. Second,  $\beta$ -LIBSHUFF is unable to detect cases when one community structure is a subset of another. Third, AMOVA determines whether the genetic diversity within two or more communities is greater than their pooled genetic diversity, and HOMOVA determines whether the amount of genetic diversity in each community is significantly different.  $\beta$ -LIBSHUFF, TreeClimber and UniFrac lump these and other factors together when performing their analysis making it difficult to discern the nature of the differences that are detected between communities. These findings demonstrate that when correctly employed, the current statistical toolbox has the ability to address specific ecological questions concerning the differences between microbial communities.

*The ISME Journal* (2008) 2, 265–275; doi:10.1038/ismej.2008.5; published online 31 January 2008

**Subject Category:** microbial population and community ecology

**Keywords:** microbial ecology; Monte Carlo; community structure; simulation

## Introduction

The field of microbial ecology has entered an exciting scientific period. Reduced sequencing costs and high-throughput sequencing and analysis are allowing experimental investigations of important ecological hypotheses (Horner-Devine *et al.*, 2004; Eckburg *et al.*, 2005; Ley *et al.*, 2005; Schloss and Handelsman, 2006c). A traditional experiment involves generating a clone library where each clone harbors a PCR amplification product generated using conserved primers (for example, 16S rRNA gene fragments) followed by sequencing a limited number of clones (Pace *et al.*, 1985). Generally, several sequence collections are generated from separate communities and compared. Each clone library must be intensively sampled to obtain adequate coverage as many microbial communities

contain  $10^2$ – $10^4$  species distributed among more than  $10^9$  cells per gram of biomass (Whitman *et al.*, 1998). Considering it is impossible to complete a total census of every cell in these communities, a growing number of statistical approaches have been proposed for describing and comparing microbial communities.

Three general approaches have been pursued. The first approach, which is employed in tools such as DOTUR (Schloss and Handelsman, 2005) and SONS (Schloss and Handelsman, 2006a), assigns sequences to operational taxonomic units (OTUs) based on the genetic distance between sequences. The abundance distribution of sequences among OTUs provides the parameters necessary to estimate the richness, evenness and ecological diversity (that is, the combination of richness and evenness) of individual communities as well as the richness of OTUs shared between communities. A second approach, which is used in LibraryCompare (Cole *et al.*, 2007), compares two communities by using a reference database. A third set of approaches, which is used in LIBSHUFF/ $\beta$ -LIBSHUFF (Singleton *et al.*, 2001; Schloss *et al.*, 2004), TreeClimber (Martin,

Correspondence: PD Schloss, Department of Microbiology, 203 Morrill Science Center IVN, University of Massachusetts—Amherst, 639 North Pleasant Street, Amherst, MA 01003, USA.  
E-mail: pschloss@microbio.umass.edu

Received 26 October 2007; revised 24 December 2007; accepted 26 December 2007; published online 31 January 2008

2002; Schloss and Handelsman, 2006b), UniFrac (Lozupone and Knight, 2005; Lozupone *et al.*, 2006, 2007) and in the analysis of molecular variance (AMOVA, Martin, 2002), use a Monte Carlo testing procedure to evaluate differences between each community. Each of these approaches has been utilized in a complimentary manner to reveal novel insights into the microbial ecology of diverse habitats.

All of these methods are phylogenetic in that they measure differences between communities based on the differences between sequences. The OTU-based approach is popular because it is possible to obtain a quantitative description of a community and its similarity to other communities; it is limited because a large number of sequences are necessary to minimize the underestimation of richness due to inadequate sampling (Schloss and Handelsman, 2006c). The database-based approach is limited because it is based on making comparisons to an incomplete representation of biodiversity within public databases. The Monte Carlo testing procedure-based methods are advantageous because they do not require a large number of sequences to detect significant differences; however, the precise nature of the hypotheses tested by these procedures is not clear. For instance, the generic hypothesis of these methods is that they test whether two or more communities are the same. But this is a relatively generic and uninteresting hypothesis. Is community B a subset of community A? Is the membership or the structure (that is, the relative abundance of members of a community) of communities A and B different? Unfortunately, these tools have not been thoroughly evaluated to determine the nature of the statistical hypotheses they are testing.

Here, I perform a systematic evaluation and comparison of the different Monte Carlo testing procedures using simulated communities with defined structures. The goal of these simulations was to understand how the tests differ and to propose a scheme for comparing communities and interpreting their results. As the field of experimental microbial ecology continues to mature it is essential that the field thoroughly understand these methods to design more robust experiments and make sound ecological inferences from the results of their studies.

## Methods

### *Simulations*

Although the current suite of statistical methods has been extensively applied to published 16S rRNA sequence collections, they have not been applied to sequence collections that were sampled from communities where the membership, structure and overlap of communities was known. This has limited the ability to account for the conflicting results one obtains from analyzing the same data set

with different techniques. To test and evaluate the performance of tools used to compare microbial communities, I instead simulated the analysis of simple microbial communities with specified characteristics. A typical experiment using 16S rRNA sequence collections involves collecting several hundred sequences from multiple communities. When one uses one of the Monte Carlo-based statistical approaches to test for the presence of significant differences between the communities, a distance matrix is generally constructed to represent the genetic diversity (that is, the average genetic distance between all pairs of sequences) contained within and between communities. Ordination methods are often used to graphically represent these distance relationships in a two-dimensional space (Legendre and Legendre, 1998). On the basis of the clustering of points, in which each represent a sequence, one can qualitatively describe the clustering of the sequences according to various treatments.

To simulate these types of experiments, I sampled clusters of sequences from a two-dimensional space by drawing points from circles or ellipses with known shapes and densities. Biologically, the diameter or length represented the maximum genetic diversity between any pair of sequences within a community. The area of the circle was proportional to the richness and described the membership of the community. The density distribution of points within the circle was proportional to the evenness. The centroid of each circle (that is, its center of mass) represented the point that corresponds to a sequence with average genetic diversity. By varying the distance between the centroid of each circle and their radius, it was possible to vary the amount of genetic diversity within each community and the fraction of membership that was shared between the communities. Although the ecological meaning is unclear, it was also possible to simulate ellipses that had the same richness, evenness, length, width, density and centroid, but were pivoted with respect to each other so that they did not share their entire membership.

The specific conditions used to generate each community are described with the results of the simulations. Each simulation consisted of 1000 independent replications. Except where noted, each replication consisted of drawing 200 points from each community and calculating the pairwise Euclidean distance among all points within a single community as well as between all of the points within a separately defined community. Each community was defined so that the maximum distance between any two points within that community was 0.300 units. These conditions allowed us to simulate the sampling intensity and biodiversity commonly found within a generic 16S rRNA gene sequence collection.

For each simulation, the 1000 distance matrices were analyzed using all of the available methods.

I measured the probability of falsely detecting a significant difference (that is,  $\alpha$ ) by the fraction of matrices that yielded a significant  $P$ -value when the communities were identical. I measured the statistical power to correctly detect significant differences (that is,  $1-\beta$ ) by the fraction of matrices that yielded a significant  $P$ -value when the communities were different. By increasing the number of individuals sampled from each simulated community, I was able to measure the relationship between sampling intensity and statistical power.

### $\int$ -LIBSHUFF

LIBSHUFF and  $\int$ -LIBSHUFF implement the Cramer-von Mises statistic to test the generic hypothesis that two communities are the same (Singleton *et al.*, 2001; Schloss *et al.*, 2004). The difference in implementations is primarily cosmetic except that LIBSHUFF uses a discrete summation to calculate the statistic, whereas  $\int$ -LIBSHUFF (below) uses a continuous integration:

$$\Delta C_{AB} = \int_0^{\infty} (C_A(D) - C_{AB}(D))^2 dD$$

where  $C_A$  and  $C_{AB}$  represent the coverage within community A and the coverage of community A onto community B. Both coverage values are dependent on the distance ( $D$ ) considered around each point.  $P$ -values for the observed  $\Delta C_{AB}$  and  $\Delta C_{BA}$  values were determined by determining the fraction of 10 000 matrix permutations that resulted in  $\Delta C_{AB}$  and  $\Delta C_{BA}$  values that were greater than or equal to the observed values. A modified version of  $\int$ -LIBSHUFF was used to facilitate the analysis of a large number of distance matrices (<http://www.plantpath.wisc.edu/fac/joh/s-libshuff.html>).

$\int$ -LIBSHUFF generates two  $P$ -values for each comparison so that the total number of  $P$ -values is equal to  $2(n-1)$ , where  $n$  is the number of treatments under consideration. Because multiple  $P$ -values are generated, it is necessary to correct the experiment-wise false discovery error using the Bonferroni or another type of correction for multiple comparison. As each distance matrix represented a comparison of two libraries, I considered a  $P$ -value to be significant if it was below 0.025. In the past, if both  $P$ -values were significant, then the communities were said to be different and if only one  $P$ -value was significant, then one community represented a subset of the other community. On the basis of this logic, I devised two tests. The first test ('strict') considered the observed differences between two communities to be statistically significant only if both  $P$ -values were significant. The second test ('relaxed') considered the observed differences between two communities to be statistically significant different if either  $P$ -value was significant.

### Parsimony test

In the parsimony test, the external branches of user-supplied phylogenetic trees are labeled using identifiers specific to each treatment and then determines the number of changes along the tree that are necessary to account for the clustering of the identifiers using Fitch's parsimony method (Fitch, 1971; Maddison and Slatkin, 1991; Martin, 2002). An unlimited number of treatments can be compared using this approach without having to use a correction for multiple comparisons. The significance of the parsimony score has been determined using two approaches. In the original approach, which is implemented in TreeClimber (<http://www.plantpath.wisc.edu/fac/joh/treeclimber.html>), trees with random topologies are generated and scored (Maddison and Slatkin, 1991; Martin, 2002; Schloss and Handelsman, 2006b). An alternative approach uses the topology of the user-supplied tree, but randomizes the labels (Lozupone and Knight, 2005). In either case, the parsimony score is calculated for 1000 trees and the fraction of trees with a parsimony score equal to or less than the observed tree score is used as the  $P$ -value. Any  $P$ -values less than 0.05 were considered significant. For this study, dendrograms were generated from each distance matrix using the neighbor-joining tree algorithm implemented in the neighbor program from the PHYLIP package (<http://evolution.genetics.washington.edu/phylip.html>).

### UniFrac

Two methods have been implemented in UniFrac to measure the fraction of the branch length in a phylogenetic tree that is unique to any community (Lozupone and Knight, 2005; Lozupone *et al.*, 2006, 2007). The method has been designed to test the hypothesis that lineages from two or more communities are undergoing equal rates of evolution. The unweighted approach calculates the ratio of branch length unique to any community to the total branch length in the tree ( $U$ ). The weighted approach divides the total branch length of a tree among the different communities using the formula:

$$W = \frac{\sum_{i=1}^N l_i \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|}{\sum_{j=1}^S L_j}$$

where  $N$  is the number of nodes in the tree,  $S$  is the number of sequences represented by the tree,  $l_i$  is the branch length between node  $i$  and its 'parent,'  $L_j$  is the total branch length from the root to the tip of the tree for sequence  $j$ ,  $A_i$  and  $B_i$  are the number of sequences from communities A and B that descend from the node, and  $A_T$  and  $B_T$  are the total number of sequences from communities A and B. Random distributions are obtained using the topology of the user-supplied tree and randomizing the labels 1000

times followed by calculating  $U$  and  $W$ . Each  $P$ -value represents the number of randomizations that generate a  $U$ - or  $W$ -value equal to or greater than the observed values. I validated my implementation of both methods by manually calculating  $U$  and  $W$  from selected trees as well as using the online version of UniFrac (<http://bmf2.colorado.edu/unifrac/index.psp>). Those  $P$ -values less than 0.05 were considered significant.

#### Analysis of molecular variance

Analysis of molecular variance is a nonparametric analog of traditional analysis of variance. This method is widely used in population genetics to test the hypothesis that genetic diversity within two populations is not significantly different from that which would result from pooling the two populations (Excoffier *et al.*, 1992; Anderson, 2001; Martin, 2002). The AMOVA statistic was calculated by

$$SS_W = \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^N d_{ij}^2 \varepsilon_{ij}$$

where  $n$  is the number of sequences per treatment,  $N$  is the number of sequences,  $d_{ij}$  is the distance between sequences  $i$  and  $j$ , and  $\varepsilon_{ij}$  is 1 when  $i$  and  $j$  are from the same treatment and 0 when they are from different treatments. A  $P$ -value is calculated by measuring the fraction of 1000 randomizations of the rows and columns in a distance matrix where the observed  $SS_W$  is less than or equal to the randomized  $SS_W$  values. Those  $P$ -values less than 0.05 were considered significant.

#### Homogeneity of molecular variance

Homogeneity of molecular variance (HOMOVA) is a nonparametric analog of Bartlett's test for homogeneity of variance, which has been used in population genetics to test the hypothesis that the genetic diversity within two or more populations is homogeneous (Stewart and Excoffier, 1996); this test has not been used in the microbial ecology literature. The HOMOVA statistic is calculated by

$$B = \frac{(N - P) \ln\left(\frac{SS_W}{N - P}\right) - \sum_{i=1}^P (N_i - 1) \ln\left(\frac{SS_{W_i}}{N_i - 1}\right)}{1 + \frac{1}{3(P-1)} \left( \sum_{i=1}^P \frac{1}{N_i - 1} - \frac{1}{N - P} \right)}$$

where  $N$  is the total number of sequences in the study,  $P$  is the number of treatments,  $N_i$  is the number of sequences in treatment  $i$ , and  $SS_{W_i}$  is the amount of  $SS_W$  contributed by treatment  $i$ . The  $P$ -value of the observed  $B$  is determined by measuring the fraction of 1000 randomizations of the rows and columns in the distance matrix, where the observed  $B$  is greater than or equal to the randomized  $B$ -values. Those  $P$ -values less than 0.05 were considered significant.

## Results

### Comparing communities with different centroids, but the same genetic diversity

To simulate the sampling of community A, I drew 200 uniformly distributed random points from a circle with a radius of 0.150. This resulted in the maximum distance between any two points being 0.300, which is approximately the distance between sequences from different phyla. To simulate the sampling of community B, I drew 200 uniformly distributed random points from circles that varied in the distance between the centroids of the circles representing the two communities. By changing the distance between centroids from 0.000 to 0.300, it was possible to simulate conditions where the two communities had the same genetic diversity, but where community A shared 0%, 80%, 90%, 95% and 100% of its membership with community B (Table 1; left to right). As expected, when I considered the case where 100% of the membership was shared between both communities, TreeClimber, UniFrac (both variants), AMOVA and HOMOVA each had a false detection rate that was not significantly different from 0.05 (95% confidence interval between 0.036 and 0.064). When using  $f$ -LIBSHUFF, the strict rule yielded a false detection rate of 0.008 and the relaxed rule yielded a false detection rate of 0.047. Therefore, except where noted, the relaxed rule was used in subsequent analysis instead of the strict rule so that it was possible to achieve a false detection rate of 0.05 using  $f$ -LIBSHUFF.

As anticipated, changing the distance between the centroids of the two circles representing communities A and B, while maintaining constant radii for the circles, resulted in an increased frequency of significant  $P$ -values for TreeClimber, UniFrac, AMOVA and  $f$ -LIBSHUFF (Table 1). These fractions represent the statistical power of the test under each condition. When communities with the same genetic diversity, but different memberships were analyzed, the power of the AMOVA test was superior to that of the unweighted UniFrac, which was superior to that of the weighted UniFrac, TreeClimber and  $f$ -LIBSHUFF tests. To obtain a statistical power of 0.80 using AMOVA when the circles had the same radius but were offset so that 95% of community A was shared with community B would require sampling at least 750 points from each community.

### Comparing communities that have the same centroid, but are a subset of each other

By setting the centroids of communities A and B at the same point, but reducing the radius of the circle representing community B, I was able to test the sensitivity of the different methods to the heterogeneity of genetic diversity (Table 1; top to bottom). This may be analogous to comparing a community before and after a perturbation, and the finding that

**Table 1** Performance of statistical tests in detecting differences between simulated communities with known properties

Radius of community B	Offset of community B compared with community A				
	0.000	0.012	0.047	0.096	0.300
0.150					
0.146					
0.134					
0.116					

Abbreviations: AMOVA, analysis of molecular variance; HOMOVA, homogeneity of molecular variance.

The gray box represents communities that are not different from each other. The decimal values represent the fraction of 1000 simulations that had a significant *P*-value.

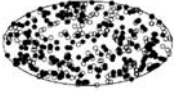




the perturbation selected for a subset of the community. For these simulations, I varied the radius of the circle representing community B between 0.116 and 0.150, so that community A shared 80%, 90%, 95% and 100% of its membership with community B. These simulations indicated that when one community represented a subset of the other, the statistical power of HOMOVA was superior to unweighted UniFrac, which was superior to weighted UniFrac, TreeClimber and  $\beta$ -LIBSHUFF (Table 1). As expected, when using AMOVA the fraction of significant randomizations within each simulation did not vary significantly from 0.05.

Interestingly,  $\beta$ -LIBSHUFF was unable to detect community B as a subset of community A. On the basis of the typical interpretation of *P*-values generated from  $\beta$ -LIBSHUFF, one would expect the comparison between community B and A ( $\Delta C_{BA}$ ) not to be significant while the comparison between A and B ( $\Delta C_{AB}$ ) would be significant. Instead, it was equally likely that either *P*-value would be

significant and the other not significant. For example, in the case where community B contained 90% of the membership found in community A,  $\Delta C_{AB}$  was significant and  $\Delta C_{BA}$  was not significant in 121 randomizations, and  $\Delta C_{BA}$  was significant and  $\Delta C_{AB}$  was not in 113 randomizations out of 1000. Both  $\Delta C_{AB}$  and  $\Delta C_{BA}$  were significant in 94 of the randomizations. These data demonstrate that  $\beta$ -LIBSHUFF can only detect differences in the structures of communities, not whether one community structure is a subset of another.

In this set of simulations, the statistical power of HOMOVA was routinely greater than that of the other tests. To obtain a statistical power of 0.80 using HOMOVA when the circles had the same centroid, but community B shared 95% of its membership with community A, would require sampling at least 2000 points from each community. When I ran simulations that varied both the distance between the centroids of each community and the radius of the circles, the power to detect differences using AMOVA, TreeClimber, UniFrac and

**Table 2** Performance of statistical tests in detecting differences between simulated communities with the same genetic diversity, centroid and abundance distribution, but different membership

Profiles of simulated communities	Simulation results
Pivot=0° 	Overlap: 100% TreeClimber: 0.052 UniFrac: 0.042 WUniFrac: 0.051 $\beta$ -LIBSHUFF: 0.049 AMOVA: 0.050 HOMOVA: 0.052
Pivot=6° 	Overlap: 95% TreeClimber: 0.106 UniFrac: 0.209 WUniFrac: 0.086 $\beta$ -LIBSHUFF: 0.073 AMOVA: 0.046 HOMOVA: 0.050
Pivot=12° 	Overlap: 90% TreeClimber: 0.326 UniFrac: 0.782 WUniFrac: 0.343 $\beta$ -LIBSHUFF: 0.252 AMOVA: 0.055 HOMOVA: 0.051
Pivot=26° 	Overlap: 80% TreeClimber: 0.987 UniFrac: 1.000 WUniFrac: 0.979 $\beta$ -LIBSHUFF: 1.000 AMOVA: 0.054 HOMOVA: 0.053
Pivot=71° 	Overlap: 60% TreeClimber: 1.000 UniFrac: 1.000 WUniFrac: 1.000 $\beta$ -LIBSHUFF: 1.000 AMOVA: 0.059 HOMOVA: 0.054

Abbreviations: AMOVA, analysis of molecular variance; HOMOVA, homogeneity of molecular variance.

The gray box represents communities that are not different from each other. The decimal values represent the fraction of 1000 simulations that had a significant *P*-value.

$\beta$ -LIBSHUFF increased as the level of overlap decreased between the communities. The power of HOMOVA to detect differences between the two communities decreased as the overlap between communities decreased. This is analogous to the negative effect departures from homoscedasticity has on the statistical power of classical analysis of variance (Sokal and Rohlf, 1995).

#### Testing for differences when communities have the same genetic diversity, centroid and abundance distribution but different memberships

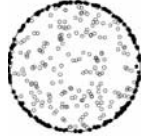
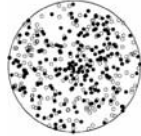
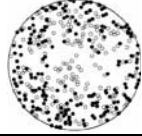
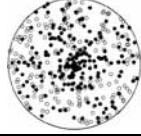
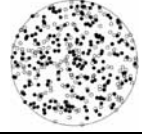
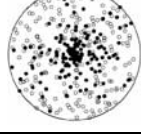
In these simulations, each community was represented as an ellipse with a length of 0.30, width of 0.15 and the same centroid. I altered the membership of the communities by pivoting the ellipse representing community B while fixing community A (Table 2). As expected, for each set of conditions, AMOVA and HOMOVA were unable to detect differences between the communities. Power analysis of the remaining tests showed that unweighted UniFrac had superior power compared with TreeClimber and weighted UniFrac, which had superior power compared with  $\beta$ -LIBSHUFF when communities had the same genetic diversity, centroid and abundance distribution, but different memberships. The ecological relevance of such a scenario is unclear; however, it does indicate that there are situations that specific statistical tests are presently unable to detect differences in community structure.

#### Testing differences in community structure

Because the communities generated for the simulations represented in Table 1, each had a uniform abundance distribution; it is clear that these tests can differentiate between communities with different memberships. Less clear is whether the tests can differentiate between differences in community structure when they have the same membership. This would represent a scenario where an environmental perturbation alters the distribution of a community while not affecting the membership. To test the ability to detect differences in community structure, I constructed two communities, which each had the same membership. The abundance of members within community A was uniformly distributed (for example, Table 3,  $r^{0.50}$ ) and the abundance members of community B were either clumped to the periphery or centroid of the circle (for example, Table 3,  $r^{0.01}$  or  $r^{2.00}$ ). I found that each test was able to detect differences between the communities. This indicates that the methods are sensitive to differences in the abundance distribution of communities with the same membership. When the abundance distribution of community B was skewed so that the centroid was not the geometric center of the circle, AMOVA was able to detect differences between the communities (data not shown). Also, because the genetic diversity was not the same in each community, HOMOVA was able to detect differences between the communities even though their memberships were identical.

In these simulations, the trend in the power of the various tests was different than those I observed when the communities had a uniform abundance distributions. Similar to the earlier simulations, HOMOVA routinely had the best power to detect

**Table 3** Performance of statistical tests in detecting differences between simulated communities with the same membership and genetic diversity, but different structures

Profiles of simulated communities	Simulation results	Profiles of simulated Communities	Simulation results
$r^{0.01}$ 	TreeClimber: 1.000 UniFrac: 1.000 WUniFrac: 1.000 β-LIBSHUFF: 1.000 AMOVA: 0.038 HOMOVA: 1.000	$r^{0.75}$ 	TreeClimber: 0.313 UniFrac: 0.095 WUniFrac: 0.141 β-LIBSHUFF: 0.124 AMOVA: 0.054 HOMOVA: 0.916
$r^{0.25}$ 	TreeClimber: 0.817 UniFrac: 0.413 WUniFrac: 0.335 β-LIBSHUFF: 0.761 AMOVA: 0.035 HOMOVA: 1.000	$r^{1.00}$ 	TreeClimber: 0.876 UniFrac: 0.271 WUniFrac: 0.668 β-LIBSHUFF: 0.610 AMOVA: 0.064 HOMOVA: 1.000
$r^{0.50}$ 	TreeClimber: 0.053 UniFrac: 0.051 WUniFrac: 0.050 β-LIBSHUFF: 0.053 AMOVA: 0.051 HOMOVA: 0.056	$r^{2.00}$ 	TreeClimber: 1.000 UniFrac: 0.908 WUniFrac: 1.000 β-LIBSHUFF: 1.000 AMOVA: 0.048 HOMOVA: 1.000

Abbreviations: AMOVA, analysis of molecular variance; HOMOVA, homogeneity of molecular variance.

The gray box represents communities that are not different from each other. The decimal values represent the fraction of 1000 simulations that had a significant *P*-value. The distribution of A is proportional to  $r^{0.50}$  and the distribution of B is as indicated.

differences of the methods when the community memberships were the same, but the structures were different. However, in this set of simulations TreeClimber had the next best power followed by weighted UniFrac and β-LIBSHUFF, which was superior to unweighted UniFrac. The relative power of weighted UniFrac method over unweighted UniFrac was the opposite of what I observed in the uniformly distributed communities. This may be because the weighted approach puts greater emphasis on closely related individuals from the same community than the unweighted approach.

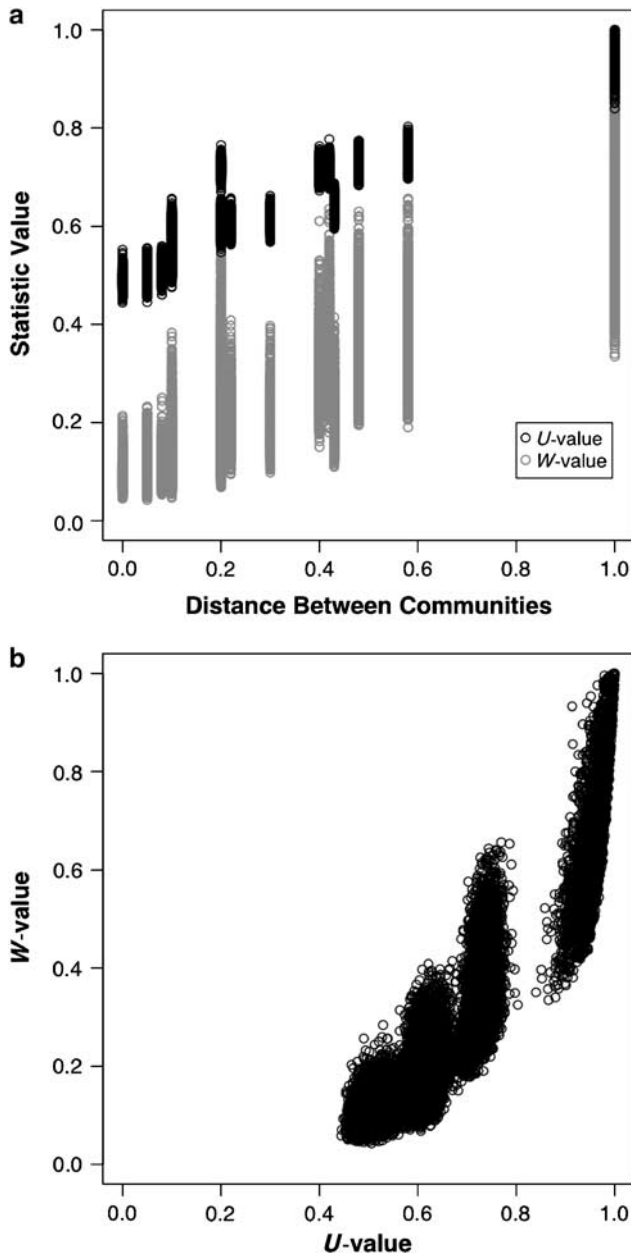
#### Use of test statistics as a measure of similarity between communities

Since its publication in 2005, numerous studies have used the generic UniFrac approach to perform statistical hypothesis tests and to generate dendrograms and ordination plots showing the similarity between multiple communities (Lozupone and Knight, 2005, 2007; Ley *et al.*, 2005; Lozupone *et al.*, 2006, 2007; Rawls *et al.*, 2006; Turnbaugh *et al.*, 2006; Fierer *et al.*, 2007; Frank *et al.*, 2007; Lamarche and Hamelin, 2007; Liu *et al.*, 2007; Walker and Pace, 2007; Wallenstein *et al.*, 2007; Yamada *et al.*, 2007). Because *U* and *W* scale between 0 and 1, they appear to be convenient distance measures to describe the dissimilarity in community memberships; however, their use as a distance metric has not been validated using data

from communities where the actual membership, structure and overlap were known. Ideally, a distance would have a linear correlation with the fraction of overlap between communities and be insensitive to sampling.

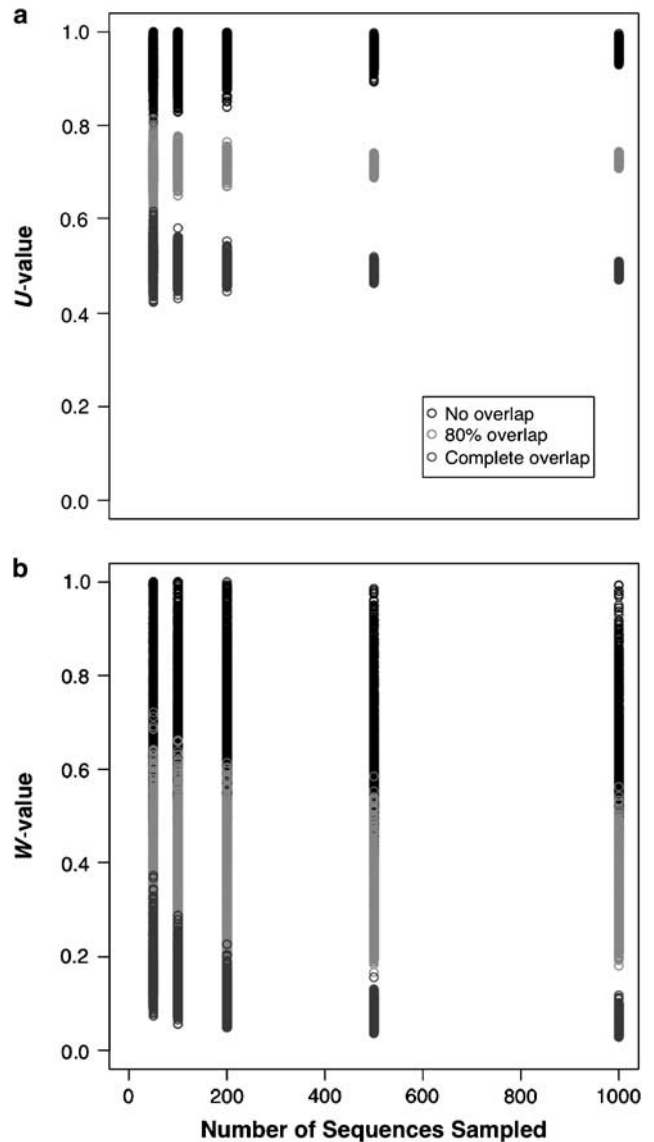
To test the correlation of *U* and *W* with the actual overlap between two communities, I plotted the observed values of *U* and *W* from the simulations performed in Tables 1 and 2 as a function of the membership overlap between them (Figure 1). Although there was considerable variation in the values of *U* and *W* for a specific level of overlap between the communities, there was a strong correlation between the statistic values and the actual distance between the communities ( $R_U = 0.97$  and  $R_W = 0.90$ ; Figure 1a). There was also a strong correlation between *U* and *W*-values ( $R = 0.94$ ; Figure 1b). Interestingly, values of *U* varied between 0.44 and 1.00 and those of *W* varied between 0.04 and 1.00. As indicated in Figure 1a, the variation in the values of *W* was greater than those of *U*.

To test the sensitivity of *U* and *W* to sampling effort, I simulated the comparison of two communities with the same genetic diversity and abundance distribution, but differed in their membership so that the fraction of overlap between them was 0%, 80% and 100%. For each level of overlap, I calculated *U* and *W* when 50, 100, 200, 500 and 1000 individuals were sampled from each community and each set of comparisons was replicated 1000



**Figure 1** The effect of distance between two communities on the unweighted and weighted UniFrac values (a) and the correlation between the UniFrac values (b). Each set of circles represents a separate set of simulations described in Tables 1 and 2. Each circle within a set represents 1 of the 1000 randomizations performed for the simulation.

times. Both  $U$  and  $W$  were sensitive to sampling intensity (Figure 2). The mean value of  $U$  did not change considerably with sampling intensity; however, the standard deviation of the observations decreased with sampling effort. In contrast, the mean value of  $W$  decreased with sampling intensity for the three levels of overlap that were considered. Most alarming was the observation in simulations where the two communities had no overlap. In these simulations, the mean value of  $W$  decreased from 0.716 to 0.626 as sampling increased from 50 to 1000



**Figure 2** The effect of sampling intensity on the unweighted ( $U$ ; a) and weighted ( $W$ ; b) UniFrac values and their variation. Each set of circles represents a separate set of simulations described in Tables 1 and 2. Each circle within a set represents 1 of the 1000 randomizations performed for the simulation.

individuals per community. Finally, as mentioned above, the standard deviation of  $W$  was considerably larger than those observed for the values of  $U$ , but the standard deviations showed little reduction with increased sampling intensity when the communities either had no overlap or 80% overlap. Since  $U$  and  $W$  did not exhibit a consistent linear correlation with the fraction of overlap between communities and they were sensitive to sampling, use of  $U$  and  $W$  as a measure of distance between communities is not recommended.

#### Testing strategy

The simulations in Tables 1–3 have shown that for specific community characteristics the power of the



tests varies. Since it is not possible to ascertain what test will perform the best, the question becomes, ‘which test is most appropriate?’ One strategy that has been employed is to use each of the methods and determine if there are any significant differences between the communities; however, there is a risk that this will be more likely to yield significant *P*-values, although there is not a difference between the communities. Indeed, when I analyzed the two communities in Table 1 that had the same membership and structure (gray box) and required that either TreeClimber, UniFrac, weighted UniFrac, *f*-LIBSHUFF, AMOVA or HOMOVA be significant, 255 of the 1000 randomizations yielded a significant *P*-value. Ideally, less than 50 would have been detected as there was no real difference between the two communities. One approach would be to correct for the multiple tests by forcing each *P*-value to be less than 0.0083 (that is, 0.05/6 tests) to be considered significant. Such a requirement could be overly conservative and limit the power to detect real differences. When applied to the shaded case in Table 1, the probability of falsely detecting a significant difference was 0.054. When applied to the case where 95% of the membership of community B is shared with community A and both communities have the same genetic diversity, the fraction of replicates correctly considered significant decreased from 0.519 to 0.182. When applied to the case in which the centroid of the two communities is the same, but the membership of community B is 95% of the membership found in community A, the fraction of replicates correctly considered significant decreased from 0.342 to 0.067. These results indicate that it is necessary to have a more robust method of implementing the various tests.

To begin to develop a strategy for hypothesis testing, I measured the correlation between the results of the six tests for comparisons where there was no true difference between communities (Table 4). As expected, I found that the *P*-values from the HOMOVA and AMOVA tests did not correlate with one another and neither correlated with the minimum *P*-values from *f*-LIBSHUFF. Interestingly, the *P*-values from the TreeClimber,

UniFrac and weighted UniFrac tests had a marginal correlation with each other. Although these methods all use phylogenetic tree as input, they each emphasize different characteristics of the tree, which perhaps leads to the lack of correlation between test statistics. The *P*-values from the three tree-based methods showed no correlation with the minimum *P*-values from *f*-LIBSHUFF. In the course of the simulations, I observed that *f*-LIBSHUFF generated low *P*-values when two samples were both highly similar and different. In contrast, TreeClimber and UniFrac generated low *P*-values when two communities were different and high *P*-values when they were the same. In light of this result, I corrected the *P*-values for TreeClimber, UniFrac and weighted UniFrac by subtracting those *P*-values larger than 0.5 from 1.0. This was done so that the four statistics could be compared on the same scale. The resulting correlations between the minimum *f*-LIBSHUFF *P*-value and TreeClimber, UniFrac and weighted UniFrac were 0.238, 0.242 and 0.029, respectively.

Assuming that negligible correlations indicated that the methods were testing independent hypotheses, I devised three classes of hypotheses. First, AMOVA determines whether the genetic diversity within each community is significantly different from the genetic diversity of the pooled communities. Second, HOMOVA detects differences in genetic diversity. Finally, *f*-LIBSHUFF, TreeClimber, UniFrac and weighted UniFrac are generic tests that detect these differences as well as differences in the pivot between the communities and possibly other unaccounted for differences between communities (for example, Table 2). Therefore, a more sophisticated testing scheme should involve first conducting parallel tests using AMOVA and HOMOVA and require significant *P*-values to be less than 0.05. For the third test, an investigator should select one test among *f*-LIBSHUFF, TreeClimber, UniFrac and weighted UniFrac and identify those *P*-values less than 0.05 as significant. If AMOVA and HOMOVA are not significant and the third test is significant, then this result would indicate the presence of a pivot between the communities. However, if either AMOVA or HOMOVA is significant and the third test is significant, then it would not be possible to determine whether a significant pivot existed between the communities.

**Table 4** Pearson correlation coefficients of *P*-values generated by different tests for comparison of samples drawn from two communities with the same membership and structure

	<i>Tree Climber</i>	<i>Uni Frac</i>	<i>WUni Frac</i>	<i>f- LIBSHUFF</i>	<i>AMOVA</i>
UniFrac	0.545	—	—	—	—
WUniFrac	0.402	0.248	—	—	—
<i>f</i> -LIBSHUFF	0.061	0.006	0.024	—	—
AMOVA	0.144	0.039	0.489	0.052	—
HOMOVA	0.109	0.030	0.094	-0.008	-0.007

Abbreviations: AMOVA, analysis of molecular variance; HOMOVA, homogeneity of molecular variance.

## Discussion

The recent improvements in sequencing quality and capacity, interesting experimental designs and a desire to test ecological theory developed for macroorganisms at the microbial level have allowed microbial ecology to develop from an observational to an experimental discipline. To match this development, it is necessary to continue to develop and refine the available statistical tools. In this

analysis, I have reconsidered the existing tools using simulated communities with known properties to validate previously held assumptions about the methods and to provide guidance to the field regarding how best to use the different methods.

A necessary limitation of these simulations was the representation of biodiversity in a two-dimensional space. In reality, 16S rRNA gene sequences would need to be represented by hundreds of dimensions. Regardless, the results obtained in these simplified simulations are generalizable to the more complicated reality. The results of the simulations in Tables 1–3 make it clear that AMOVA tests whether two communities have the same centroid. Alternatively stated, AMOVA determines whether the genetic diversity within each community is significantly different from the average genetic diversity of both communities pooled together. HOMOVA tests whether the genetic diversity is the same in multiple communities. The specific hypotheses that the other methods evaluate are less obvious. It has been claimed that UniFrac has the potential to determine whether a community has lineages that are evolving faster than another lineage. This would suggest that UniFrac is a tree-based version of HOMOVA; however, the simulations demonstrate that this is not the case. UniFrac detects any differences in the communities that result in the ability to attribute the total branch length of a tree to one particular community. The weighted UniFrac attempts to perform a similar test with a different weighting scheme. TreeClimber is related to the UniFrac methods and attempts to detect differences in the community that result in the ability to attribute sections of a tree's topology to specific communities. Finally,  $\beta$ -LIBSHUFF evaluates the significance of the probability that the closest relative of any sequence is from the same or different community. As indicated by the correlation values, these tests evaluate similar but seemingly different questions.

The advantage of AMOVA and HOMOVA is their ability to address specific questions that have ecological meaning. Shifts in genetic diversity are ecologically meaningful. Less clear is the ecological meaning of a pivot between two or more communities, except that it indicates that the community structures are different. Another advantage of AMOVA and HOMOVA over the other methods is the ability to incorporate more sophisticated experimental designs including replication, multiple factor analysis and regression. Although the ability to construct a phylogenetic tree improves the flexibility of an analysis, the methods that analyze trees are currently limited by their inability to analyze complicated designs and to isolate tests for specific ecological differences.

A limitation of any significance testing method is that the test provides a probability that the same or more extreme result could be observed by chance. Such a probability does not indicate the similarity of

two or more communities. The simulations conducted in this study have shown that the previous assumptions regarding the ability to detect subsets using  $\beta$ -LIBSHUFF were incorrect. Furthermore, the simulations also indicated that  $U$ - and  $W$ -values are not appropriate distance metrics for constructing dendrograms or ordination plots. Such questions are answered best by using statistical models that predict community parameters as well as the overlap in membership or structure between two communities using OTU-based approaches. Finally, hypothesis-testing methods can only detect statistically significant differences; they do not necessarily predict an ecologically significant difference. As the search continues to identify and quantify interactions between microbes and their environment, parallel use of statistical and biological tools will be essential.

## Acknowledgements

I acknowledge the financial support from the College of Natural Resources at the University of Massachusetts—Amherst.

## References

- Anderson MJ. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecol* **26**: 32–46.
- Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM *et al.* (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res* **35**: D169–D172.
- Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M *et al.* (2005). Diversity of the human intestinal microbial flora. *Science* **308**: 1635–1638.
- Excoffier L, Smouse PE, Quattro JM. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**: 479–491.
- Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R *et al.* (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl Environ Microbiol* **73**: 7059–7066.
- Fitch WM. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst Zool* **20**: 406–416.
- Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* **104**: 13780–13785.
- Horner-Devine MC, Lage M, Hughes JB, Bohannan BJ. (2004). A taxa–area relationship for bacteria. *Nature* **432**: 750–753.
- Lamarche J, Hamelin RC. (2007). No evidence of an impact on the rhizosphere diazotroph community by the expression of *Bacillus thuringiensis* Cry1Ab toxin by

- Bt white spruce. *Appl Environ Microbiol* **73**: 6577–6583.
- Legendre P, Legendre L. (1998). *Numerical Ecology*. Elsevier: New York.
- Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. (2005). Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA* **102**: 11070–11075.
- Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* **35**: e120.
- Lozupone CA, Hamady M, Kelley ST, Knight R. (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, Hamady M, Knight R. (2006). UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7**: 371.
- Lozupone C, Knight R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228–8235.
- Lozupone CA, Knight R. (2007). Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104**: 11436–11440.
- Maddison WP, Slatkin M. (1991). Null models for the number of evolutionary steps in a character on a phylogenetic tree. *Evolution* **45**: 1184–1197.
- Martin AP. (2002). Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol* **68**: 3673–3682.
- Pace NR, Stahl DA, Lane DJ, Olsen GJ. (1985). Analyzing natural microbial populations by rRNA sequences. *ASM News* **51**: 4–12.
- Rawls JF, Mahowald MA, Ley RE, Gordon JI. (2006). Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell* **127**: 423–433.
- Schloss PD, Handelsman J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71**: 1501–1506.
- Schloss PD, Handelsman J. (2006a). Introducing SONS, A tool that compares the membership of microbial communities. *Appl Environ Microbiol* **72**: 6773–6779.
- Schloss PD, Handelsman J. (2006b). Introducing Tree-Climber, a test to compare microbial community structure. *Appl Environ Microbiol* **72**: 2379–2384.
- Schloss PD, Handelsman J. (2006c). Toward a census of bacteria in soil. *PLoS Comput Biol* **2**: e92.
- Schloss PD, Larget BR, Handelsman J. (2004). Integration of microbial ecology and statistics: a test to compare gene libraries. *Appl Environ Microbiol* **70**: 5485–5492.
- Singleton DR, Furlong MA, Rathbun SL, Whitman WB. (2001). Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. *Appl Environ Microbiol* **67**: 4374–4376.
- Sokal RR, Rohlf FJ. (1995). *Biometry: The Principles and Practice of Statistics in Biological Research*, 3rd edn. Freeman: New York, xix, 887pp.
- Stewart CN, Excoffier L. (1996). Assessing population genetic structure and variability with RAPD data: application to *Vaccinium macrocarpon* (American Cranberry). *J Evol Biol* **9**: 153–171.
- Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**: 1027–1031.
- Walker JJ, Pace NR. (2007). Phylogenetic composition of Rocky Mountain endolithic microbial ecosystems. *Appl Environ Microbiol* **73**: 3497–3504.
- Wallenstein MD, McMahon S, Schimel J. (2007). Bacterial and fungal community structure in Arctic tundra tussock and shrub soils. *FEMS Microbiol Ecol* **59**: 428–435.
- Whitman WB, Coleman DC, Wiebe WJ. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95**: 6578–6583.
- Yamada A, Inoue T, Noda S, Hongoh Y, Ohkuma M. (2007). Evolutionary trend of phylogenetic diversity of nitrogen fixation genes in the gut community of wood-feeding termites. *Mol Ecol* **16**: 3768–3777.