Bioinformatics, 31(21), 2015, 3451–3459
doi: 10.1093/bioinformatics/btv394
Advance Access Publication Date: 30 June 2015
Original Paper

Sequence analysis

Estimating beta diversity for under-sampled communities using the variably weighted Odum dissimilarity index and OTUshuff

Daniel K. Manter[1,*] and Matthew G. Bakker[2]

[1]USDA-ARS, Soil-Plant-Nutrient Research, Fort Collins, CO 80526, USA and [2]USDA-ARS, National Laboratory for Agriculture and the Environment, Ames, IA 50011 USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on February 24, 2015; revised on May 27, 2015; accepted on June 25, 2015

Abstract

Motivation: In profiling the composition and structure of complex microbial communities via high throughput amplicon sequencing, a very low proportion of community members are typically sampled. As a result of this incomplete sampling, estimates of dissimilarity between communities are often inflated, an issue we term pseudo β-diversity.

Results: We present a set of tools to identify and correct for the presence of pseudo β-diversity in contrasts between microbial communities. The variably weighted Odum dissimilarity (D_{wOdum}) allows for down-weighting the influence of either abundant or rare taxa in calculating a measure of similarity between two communities. We show that down-weighting the influence of rare taxa can be used to minimize pseudo β-diversity arising from incomplete sampling. Down-weighting the influence of abundant taxa can increase the sensitivity of hypothesis testing. OTUshuff is an associated test for identifying the presence of pseudo β-diversity in pairwise community contrasts.

Availability and implementation: A Perl script for calculating the D_{wOdum} score from a taxon abundance table and performing pairwise contrasts with OTUshuff can be obtained at http://www.ars.usda.gov/services/software/software.htm?modecode=30-12-10-00.

Contact: daniel.manter@ars.usda.gov

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

1 Introduction

The size and diversity of microbial populations pose challenges to characterizing and contrasting communities. For instance, sequence-based analyses of bacterial populations in soil frequently suggest the presence of more than 10^6–10^9 individuals distributed across thousands of distinct taxa per gram of soil (Deng *et al.*, 2012; Kuffner *et al.*, 2012). Although the advent of high-throughput sequencing technologies has increased the depth at which we are able to survey microbial community composition, in many cases sampling still detects only a fraction of the diversity present at a given site. For instance, working with rhizosphere soil samples from a grass host plant species, Lagos *et al.* (2014) obtained 30 000–60 000 sequence reads per sample, yet estimated that they were able to detect only 90% of the OTUs that could have been observed at greater sampling depth. Taxa that are present at low abundance may nevertheless be important to understanding community dynamics. For instance, Shade *et al.* (2014) have demonstrated from analyses of time series in several different environments that many taxa are only conditionally rare, and may be present at dramatically higher abundances at other time points.

While there is debate over the relative benefits of increasing sampling depth versus increasing the number of samples analyzed (Kuczynski *et al.*, 2010) it is becoming increasingly common to trade sample depth for more extensive replication. For instance, in a recent characterization of the Arabidopsis-associated microbiome, researchers analyzed 1248 samples at 1000 sequence reads per

sample (Lundberg *et al.*, 2012). These developments highlight the need for analytical approaches that are specifically designed to account for effects attributable to shallow sampling. It should be noted that incomplete sampling will be a more pressing issue for environments that harbor the most diverse microbial communities, such as soil.

Beyond simply creating a census of which bacterial taxa are present in a given environment, researchers frequently desire to compare two or more species assemblages for differences in species composition and structure. A large number of indices are available for such purposes [summarized in (Magurran, 2004)], but many of these were developed in systems for which it is possible to be nearly comprehensive in detecting species that are present in a given location (i.e. plant and animal communities). In contrast, the majority of species are likely to go undetected in shallow sequence-based surveys of complex microbial communities.

Indices for describing the resemblance between communities are often based either on species composition (i.e. presence-absence measures) or on community structure (i.e. taking into consideration the relative abundance of species within each community). Both index types are sensitive to incomplete sampling (Chao *et al.*, 2005).

When dealing with complex microbial communities, the inflation of dissimilarity due to incomplete sampling is a concern; differential detection of low abundance taxa may be simply due to insufficient sampling, and not to true absence from the community (Chao *et al.*, 2005). Furthermore, rare sequence variants may not always reflect true taxon presence or abundance; rather, they may be introduced through errors in DNA replication or sequencing (Quince *et al.*, 2011). Thus, the ability to adjust the influence of low abundance taxa on a measure of community resemblance is desirable.

Indices that give less weight to low abundance community members exist within the current suite of community resemblance metrics (Magurran, 2004). For instance, Beck *et al.* (2013) compared 14 different measures of β-diversity and concluded that measures with high dependency on the distribution of abundant taxa were the least sensitive to under-sampling (e.g. Morisita-Horn). However, differences among microbial communities, or microbial community response to experimental manipulation, may lie in changes in taxon abundance anywhere across the distribution from dominant to rare community members. Thus, a measure of dissimilarity should ideally be tunable in its sensitivity toward rare or dominant members.

Among non-phylogenetic measures of community resemblance, the only method that we are aware of that is tunable in sensitivity to rare taxa is the Normalized Expected Species Shared method of Grassle and Smith (1976). However, this method adjusts the influence of rare taxa by re-sampling an observed taxon abundance table at varying depths. Thus, this method relies upon the very problem we are trying to address (incomplete sampling), and leaves room for improvement. A variably weighted version of the phylogenetic distance measure UniFrac has been proposed (Chen *et al.*, 2012), and will be discussed below. Recent parallel developments in the field of community diversity assessment have come to similar conclusions regarding the utility of calculating diversity indices across a range of sensitivities toward low abundance community members (Leinster and Cobbold, 2012).

A number of statistical tests are available to assess the significance of differences in microbial community structure, including UniFrac, Parsimony, AMOVA and HOMOVA. Schloss (2008) has provided a comparison of these methods, and suggests that each carries distinct benefits. However, the effects of incomplete sampling have not been well explored for statistical tests that contrast microbial community structure. Here, we present a new analytical method to contrast two samples, which indicates the likelihood that the observed samples were drawn from a common community. We illustrate the method using simulated data, as well as bacterial sequence data (partial 16S rRNA gene fragments) derived from 51 soil samples, collected across a range of geographic locations, plant communities and land uses.

2 Methods

2.1 The variably weighted Odum dissimilarity index

The particular characteristics of sequence-based profiling of microbial communities demand an appropriate index of community resemblance. We use as our starting point the traditional Odum index (Odum, 1950):

$$D_{\mathrm{Odum}} = \frac{\sum_{i=1}^{S} |A_i - B_i|}{\sum_{i=1}^{S} (A_i + B_i)} \quad (1)$$

where A and B are the samples being compared, and S is the total number of observed taxa. When the original taxon counts have been converted into proportions (as we will assume throughout), the Odum index is mathematically equivalent (Somerfield, 2008) to the commonly used Bray-Curtis dissimilarity index (Bray and Curtis, 1957):

$$D_{\mathrm{BC}} = 1 - \frac{\sum_{i=1}^{S} 2\ \min(A_i,\ B_i)}{\sum_{i=1}^{S} (A_i + B_i)} \quad (2)$$

The proper naming and origins of these formulas are a topic of some debate (Somerfield, 2008; Yoshioka, 2008); however, we use the terms Odum index (Eq. 1) and Bray-Curtis dissimilarity (Eq. 2) to differentiate the two notations. Regardless of the notation, with this metric, differences in the abundance of common or rare taxa are given equal weighting. As a modification that allows adjustable influence based on taxon rarity, we propose a variably weighted version of the Odum index, which we call wOdum:

$$D_{\mathrm{wOdum}} = \frac{\sum_{i=1}^{S} \left(\frac{|A_i - B_i|}{(A_i + B_i)} * (A_i + B_i)^{\alpha} \right)}{\sum_{i=1}^{S} \left((A_i + B_i)^{\alpha} \right)} \quad (3)$$

where A and B are the samples being compared, S is the total number of observed taxa and α is a weighting parameter, such that $\alpha \geq 0$. D_{wOdum} is bounded between 0 and 1, where a score of 0 would reflect the contrast of a community to itself (all taxa shared, and at the same relative abundances in both communities) and a score of 1 would reflect completely dissimilar communities (no taxa shared between communities). When $\alpha = 1$, D_{wOdum} is reduced to Equation (1) above and D_{wOdum} is equal to D_{Odum} or D_{BC}; when $\alpha < 1$, the influence of abundant taxa will be down-weighted; and when $\alpha > 1$, the influence of low abundance taxa will be down-weighted. Figure 1 depicts a simple simulation in order to illustrate the effect of the weighting parameter on D_{wOdum} scores.

In this example, we started with two samples having the following taxon abundance distributions (Community A: [10, 9, 8, 7, 6, 5, 4, 3, 2, 1]; Community B: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]). In order to

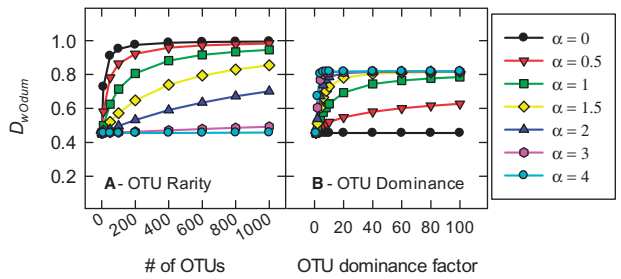

Fig. 1. Illustration of the effects of the weighting parameter, α, on the variably weighted Odum dissimilarity score (D_{wOdum}), on datasets simulated to accentuate (**A**) rare OTUs or (**B**) dominant OTUs (i.e. low community evenness)

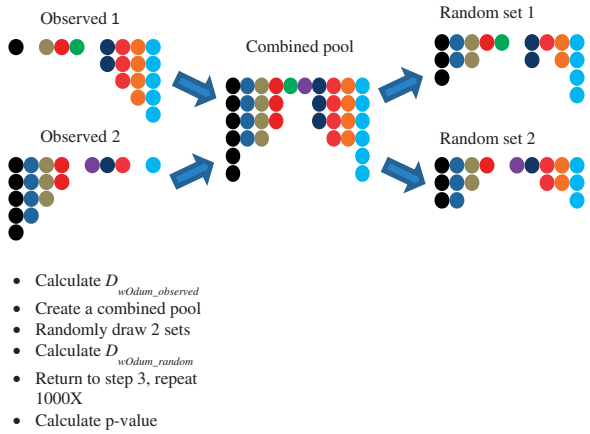

Fig. 2. Schematic representation of the OTUshuff procedure for testing whether two observations of species abundances may have been drawn from the same community. Each circle represents an observation of an individual; different colors indicate different taxa

simulate rarity, additional singleton taxa (A_i: 0, B_i: 1 or A_i: 1, B_i: 0] were added up to a total of 1000 taxa per sample. As expected, as α approaches 0 the D_{wOdum} score shows increasing responsiveness to rare taxa; whereas, as α increases above 1, a reduced sensitivity to rare taxa is evident (Fig. 1A). In order to simulate taxon dominance (i.e. low community evenness), the abundance of Taxon1 [A_1: 10, B_1: 1] was increased x-fold for both populations (i.e. for $x = 100$, A_1: 1000, B_1: 100) with all other OTUs unchanged. In this case, sensitivity toward dominant taxa decreased when $\alpha < 1$ and increased when $\alpha > 1$ (Fig. 1B).

2.2 OTUshuff

We also introduce a new method, termed OTUshuff, which utilizes a Monte Carlo simulation to test whether two samples are likely to have been drawn from a common starting population (see schematic in Fig. 2). From the observed abundances of operational taxonomic units (OTUs), OTUshuff calculates the D_{wOdum} score for a pair of samples ($D_{wOdum\_observed}$). Next, the two samples are combined to create a single pool, which is re-divided into two sets via random draws without replacement, using a weighted probability based on the combined abundance distribution. For example, if the relative abundance of OTU_i was 0% in Community A and 2% in Community B, the probability of any sequence read being assigned to OTU_i during drawing for either sample is 1%. This process is repeated until the two samples are repopulated with the number of reads in the smaller of the two originally observed communities and D_{wOdum} is re-computed (D_{wOdum_random}). This process is repeated

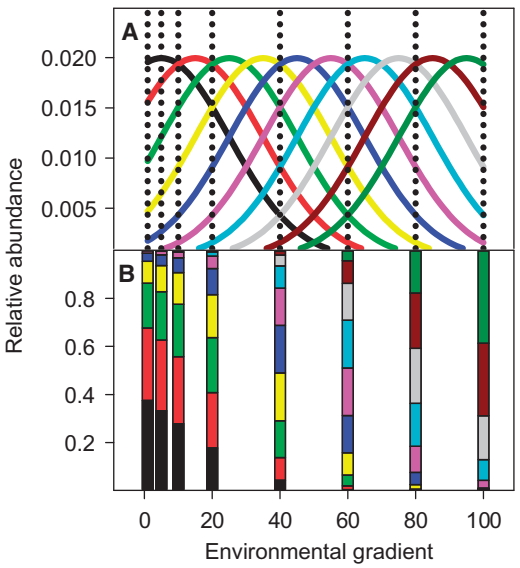

Fig. 3. Illustration of the method used to simulate a range of microbial community structures. Different colored lines or bars represent different microbial community types. (**A**) Ten microbial community types were assumed to have abundances that followed normal distributions about varying optima along a hypothetical environmental gradient. Within each community type, there are 1000 taxa whose abundances follow a logarithmic distribution. (**B**) Sampling various points along this hypothetical environmental gradient yields a range of observed microbial community structures

multiple times (user-defined) and the reported *P*-value is determined as follows:

$$p = 1 - \sum_{1}^{n} \delta$$

where $\delta = 1/\text{number of iterations}$, and $n =$ the count of iterations in which $D_{wOdum\_random} < D_{wOdum\_observed}$. D_{wOdum} and OTUshuff can be obtained from: http://www.ars.usda.gov/services/software/software.htm?modecode=30-12-10-00. OTUshuff.pl allows the user to calculate D_{wOdum} at defined values of α and to conduct OTUshuff, with results output in both tabular and lower-triangular matrix forms. The program requires as input a taxon abundance table formatted as a Mothur 'shared' file (http://www.mothur.org). Users can specify the threshold for taxon delimitation (i.e. sequence similarity threshold for binning sequences into operational taxonomic units, as defined by Mothur), and the number of iterations to conduct. For datasets containing more than two samples, OTUshuff will automatically conduct each pairwise comparison for all samples in the dataset. Because not all comparisons may be of biological interest, OTUshuff does not correct *P*-values for multiple comparisons (e.g. Bonferoni's adjustment). We leave multiple test correction to the user's discretion.

2.3 Simulated data

We simulated samples varying in microbial community structures by assuming a hypothetical environmental gradient, across which 10 microbial community types could be found, with the abundance of each community type following a normal distribution about a unique optimum ($\mu = 5$, 15, 25, 35, 45, 55, 65, 75, 85, or 95), but with a common variance ($\sigma = 20$). Each community type consisted of 1000 taxa (i.e. OTUs), with relative abundances following a logarithmic distribution ($\Theta = 0.9999$). Thus, sampling any given location along this hypothetical gradient (Fig. 3A, dotted black lines) could yield up to 10 000 possible OTUs. Individual observations (i.e.

sequence reads) at a given location along the gradient were assigned to taxa probabilistically, leading to differences in observed communities among locations (Fig. 3B). With this approach, diversity and evenness will be lower at the extremes, compared to the center of the gradient. The phylogenetic distances required for $D_{GUniFrac}$ were modeled in a manner similar to that of Schloss (2008). Briefly, distances between OTUs within a community were randomly generated from a two-dimensional circle ($r = 0.1$) and each community centroid was offset by 0.1 (e.g. community 1: 0; community 2: 0.1; community 3: 0.2; etc.).

From these defined communities, we simulated true β-diversity (that is, measured distance or dissimilarity between two fully-sampled communities). It should be noted that true β-diversity can only be known in the context of simulated communities for which composition is determined a priori and for which the taxon abundance distributions are defined. We selected 15 locations along the environmental gradient (1–10, 20, 40, 60, 80 and 100), and at each location allowed the abundance of each OTU to vary randomly by±0–100%. This process was repeated five times to generate a set of replicate samples at each location, and then the full simulation was repeated 100 times. Complete sampling was assumed to occur at 10^9 individuals per sample, or the point where each OTU present along the hypothetical environmental gradient was detected at least once.

To simulate pseudo β-diversity (that is, the portion of a pairwise distance or dissimilarity that is attributable purely to having an incomplete census of the two communities in question), a set number of observations could be drawn probabilistically to generate community surveys with incomplete sampling.

2.4 Observed data

Soils were collected from 11 different sites representing a wide range of geographic locations, plant communities and land uses (Table 1) within the United States. Within each site, a single soil sample (0–7.5 cm depth) was collected from each plot ($n = 3$–4) unless specified differently. For each of the three Alaska sites (AK.1, AK.2, AK.3), plots were centered on a single white spruce tree with each tree located a minimum of 1 km apart. The CA.1 site is a commercial potato farm located in Tulelake, California and was sampled immediately prior to harvest. The CO.1 site was part of cover crop study located in the San Luis Valley, Colorado (Essah *et al.*, 2012). Plots were sampled at the end of a cover crop rotation (i.e. mowing and incorporation). The remaining three Colorado sites (CO.2, CO.3, CO.4) are all part of the Dryland Agroecosystems Project (Sherrod *et al.*, 2014) with each plot ($n = 4$) consisting of a single soil sample from a randomly selected plot under wheat management. The sites MA.1 and MA.2 are golf courses located in Massachusetts, where each plot consists of a fairway and rough sample from three different holes per course. The NE site consists of a switchgrass cultivar trial at the University of Nebraska's Agricultural Research and Development Center with each plot consisting of twelve switchgrass plants of the same cultivar. From each plot ($n = 3$) rhizosphere soil (0 to 15 cm) was collected from a single plant.

Table 1. Locations and characteristics of sites from which soil microbial communities were characterized by sequencing

Site code	Over-story	Location	Plots	Sub-samples[a]
AK.1	White spruce	Alaska	3	3, 3, 2
AK.2	White spruce	Alaska	3	3, 3, 3
AK.3	White spruce	Alaska	3	3, 3, 3
CA.1	Potato	Tulelake, CA	3	2, 3, 2
CO.1	Canola	Center, CO	3	3, 3, 3
	Fallow		3	3, 2, 3
	Mustard		3	3, 3, 3
CO.2	Wheat	Sterling, CO	4	2, 3, 2, 3
CO.3	Wheat	Stratton, CO	4	3, 3, 3, 3
CO.4	Wheat	Walsh, CO	4	3, 3, 3, 2
MA.1	Golf course—fairway	Vineyard Haven, MA	3	3, 3, 3
	Golf course—rough		3	3, 3, 3
MA.2	Golf course—fairway	Edgartown, MA	3	3, 2, 3
	Golf course—rough		3	2, 3, 2
NE.1	Switchgrass 'Kanlow'	Ithaca, NE	3	3, 3, 2
	Switchgrass 'Summer'		3	3, 3, 2

[a]Subsamples of 1000 sequence reads were randomly drawn without replacement from each plot-specific 16 S library. Samples with a sufficient number of reads were subsampled multiple times.

2.5 DNA extraction and PCR amplification

DNA was extracted from 0.5 g subsamples of each soil collection using the MoBio UltraClean-HTP™ soil DNA isolation kit (Carlsbad, CA, USA) following the manufacturer's recommendations plus an additional purification step with AMPure beads (Agencourt, MA, USA) to further remove humic acids and other PCR inhibitors. Extracted DNA was quantified by spectrophotometry and diluted to a final concentration of 10 ng μl^{-1}. Amplification of bacterial 16S rRNA gene fragments was performed as described by Manter *et al.* (2010), using primers 27F and 388R to amplify across the V1–V3 hypervariable regions of the gene. Primers were modified to add a unique barcode (Hamady *et al.*, 2008) to amplicons from each of 51 distinct samples (Table 1). Unidirectional (Lib-L) pyrosequencing was performed under contract with Duke University's IGSP Sequencing Core Facility using a 454 Life Sciences GS FLX System with standard chemistry.

2.6 Sequence processing

All sequence read editing and processing was performed with Mothur Ver. 1.32 (Schloss *et al.*, 2009) using the default settings unless otherwise noted. Briefly, sequence reads were (i) trimmed (bdiff = 0, pdiff = 0, qaverage = 25, minlength = 100, maxambig = 0, maxhomop = 10); (ii) aligned to the bacterial-subset SILVA alignment available at the Mothur website (http://www.mothur.org); (iii) filtered to remove vertical gaps; (iv) screened for chimeras with UCHIME (Edgar *et al.*, 2011); (v) classified using the RDP training set Vers. 9 (http://www.mothur.org) and the naïve Baysian classifier (Wang *et al.*, 2007) embedded in Mothur, after which all sequences identified as chloroplast or mitochondria were removed; (vi) sequences were screened (optimize = minlength-end, criteria = 95) and filtered (vertical = T, trump=.) so that all sequences covered the same genetic space; and (vii) all sequences were pre-clustered (diff=2) to remove potential pyrosequencing noise and clustered (calc = onegap, countends = F, method = average) into operational taxonomic units or OTUs (Huse *et al.*, 2010). After processing, each site- or plot-specific library was randomly split into subsamples of 1000 sequence reads in order to standardize sampling effort for each sample, and to generate replicate subsamples from within a given library for comparison. Differences between replicate subsamples from the same dataset represent sensitivity to low sample coverage, and should be minimized. If a method is truly suitable for severely under-sampled communities, a comparative analysis should not detect a statistical difference between subsamples from the same library.

2.7 Analyses

Traditional measures of community resemblance (i.e. Bray-Curtis, D_{BC}; Canberra, $D_{Canberra}$; Gower, D_{Gower}; Manhattan, $D_{Manhattan}$;

Morisita-Horn, D_{MH}; Odum, D_{Odum}; Soergel, $D_{Soergel}$; and Yue-Clayton, $D_{\theta YC}$) and principal coordinate analyses were calculated in Mothur Ver. 1.32. Generalized UniFrac scores ($D_{GUniFrac}$) were calculated using the R package GUniFrac (Chen *et al.*, 2012) because this score is not presently implemented in Mothur.

3 Results

3.1 Simulated data

3.1.1 Comparing indices of community resemblance where true β-diversity is known

When assessing true β-diversity with D_{wOdum}, the most notable effect of adjusting the influence of low abundance taxa via the α parameter was the effect on estimates of β-diversity among very similar communities. For example, as α increased, especially when $\alpha > 1$, the calculated scores exhibited greater variability, which increased as sample similarity increased (Fig. 4, Supplementary Fig. S1). A smaller, but still noticeable effect was that as α increased, scores tended to spread out. For example, at $\alpha = 0$, mean D_{wOdum} scores ranged from 0.394 to 0.890 for the 1 versus 1 and 1 versus 100 location comparisons, respectively; whereas, at $\alpha = 4$, scores ranged from 0.267 to 1.000 (Fig. 4, Supplementary Fig. S1). $D_{GUniFrac}$ shows a similar behavior to D_{wOdum} with distance scores expanding as α increases; although variation remained small regardless of α-level (Supplementary Fig. S2).

Interestingly, some other existing indices (e.g. $D_{Manhattan}$, $D_{Soergel}$ and $D_{\Theta YC}$) were particularly sensitive to small changes in community structure, leading to increased variability and higher estimates of β-diversity than expected. For example, all three of these measures indicated that locations 1 versus 10 were more different than locations 1 versus 20 (Supplementary Fig. S3C, E, F), while we know from the mathematically defined species distributions that communities become increasingly different with separation along the hypothetical environmental gradient.

The variability inherent in an index of community resemblance has a major influence on the ability to visually detect clustering and differentiate among treatments. For instance, principal coordinates ordination plots revealed a wide variation in the amount of overlap between communities from our simulated environmental gradient (Supplementary Figs S4–S6). For instance, communities drawn between locations 1 and 20 were virtually indistinguishable when D_{MH} and $D_{\Theta YC}$ were used, while $D_{Canberra}$ and D_{Gower} provided better separation among these communities (Supplementary Fig. S4).

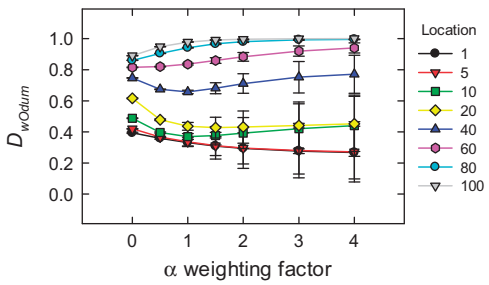

Fig. 4. For simulated communities drawn from a hypothetical environmental gradient, calculated dissimilarities (D_{wOdum}) for the contrast of each location versus location 1, with variations to the influence of low abundance taxa achieved via the weighting parameter, α. The lines connect adjacent means and are not fitted curves. The error bars represent 1 SD, based on 5 instances of perturbing the same original community abundance distribution. The contrast of location 1 versus location 1 represents dissimilarities among perturbed samples, and is not the contrast of a single community to itself

These differences among the indices become more easily interpretable within the context of a family of dissimilarity scores which vary in a coherent way in sensitivity to low abundance community members. For instance, contrasting principal coordinate plots derived from D_{wOdum} scores in which α is varied between 0 and 3 reveals that the true relationships among the locations, in this simulated dataset, are best revealed by down-weighting the impact of dominant taxa (Supplementary Fig. S5). The $D_{GUniFrac}$ scores reveal a similar insight, providing better cluster separation among communities located proximately on the simulated gradient when less weight is given to the high abundance branches (Supplementary Fig. S6).

Variability naturally also has a strong effect on hypothesis testing. For example, the closest location on the simulated gradient for which community structure was found to be significantly different (AMOVA, $P < 0.05$) from location 1 occurred at location 2, 4, 7, 8 and 10 for D_{wOdum} with α at 0, 0.5, 1, 2 and 3, respectively. Statistical power similarly varied among the suite of existing indices, ranging from high to low: $D_{Canberra}$ and $D_{Gower} > D_{Manhattan} > D_{Soergel} > D_{MH}$ and $D_{\Theta YC}$. However, we reiterate that there is no clear framework for evaluating contrasts among these disparate indices, and that this promotes an arbitrary selection of the index that happens to provide a clear pattern, often with no understanding of why this might be the case.

3.1.2 Comparing indices of community resemblance where pseudo β-diversity is present

Incomplete sampling of simulated communities was used to demonstrate the effects of pseudo β-diversity on measures of community resemblance. For each location of interest along the simulated environmental gradient, five independent subsamples were generated at various sampling depths. Pseudo β-diversity resulting from incomplete sampling would be expected to inflate estimates of distance or dissimilarity, particularly where differences between communities are derived from low abundance taxa (e.g. contrasts between communities proximally located along the simulated environmental gradient). This is illustrated nicely by varying α for the D_{wOdum} score: increasing α gives more weight to dominant taxa and reduces pseudo β-diversity associated with incomplete sample (Fig. 5).

In contrast, some existing indices responded erratically to incomplete sampling. For instance, D_{Gower} and $D_{Manhattan}$, which are influenced by sampling effort (i.e. unbounded) and $D_{Canberra}$, which is normalized by the total number of taxa, gave erroneously low estimates of β-diversity when sampling was incomplete (Supplementary Fig. S7; Table 2). The $D_{GUniFrac}$ score also responded less consistently to incomplete sampling, compared to the D_{wOdum} score. For instance, at low sampling depth, $D_{GUniFrac}$ actually underestimated the distance between very distinct communities (Supplementary Fig. S8C–F, location 1 versus location 100). There was also a spike in $D_{GUniFrac}$ score for our simulated communities (except at the lowest α) when the number of sequence reads obtained was equal to the number of possible taxa (Supplementary Fig. S8). As might be expected, accurate $D_{GUniFrac}$ scores are only obtained once the majority of the OTUs (i.e. branches) have been successfully sampled; in our simulation, this occurs at ca. 10^4 individuals or the expected number of OTUs at each location.

3.1.3 Combining true and pseudo β-diversity

The final simulation conducted included both true and pseudo β-diversity by allowing each OTU to randomly vary by ±0–100% and sub-sampling to 1000 reads. As expected, estimates of

β-diversity increased due to incomplete sampling. Because incomplete sampling primarily results in the failure to detect low abundance taxa, this inflation in estimates of β-diversity represents the influence of low abundance community members that may be universally present, but are differentially detected by chance. Accordingly, increasing α to give more weight to dominant taxa successfully corrects for this pseudo β-diversity (Fig. 6). Once $\alpha \geq 2$, there was no longer any significant effect of pseudo β-diversity (due to incomplete sampling) on measures of within-location community distance or dissimilarity; in other words, pseudo β-diversity was removed by adjusting α.

In an effort to determine the appropriate sampling depth required to achieve accurate estimates of β-diversity, we calculated a bias for each index of community resemblance (Table 2), where bias is defined as the difference in score with incomplete sampling versus the score at complete sampling. At 1000 reads per sample, a common sampling depth in many studies, the bias in calculated scores differed dramatically depending upon the index chosen. For both the D_{wOdum} and D_{GUniFrac} scores, increasing α substantially reduced bias. For example, the mean bias for D_{wOdum} and D_{GUniFrac} when $\alpha = 4$ was 0.025 and 0.051, respectively. In comparison, the bias for D_{MH} was 0.075 (Table 2).

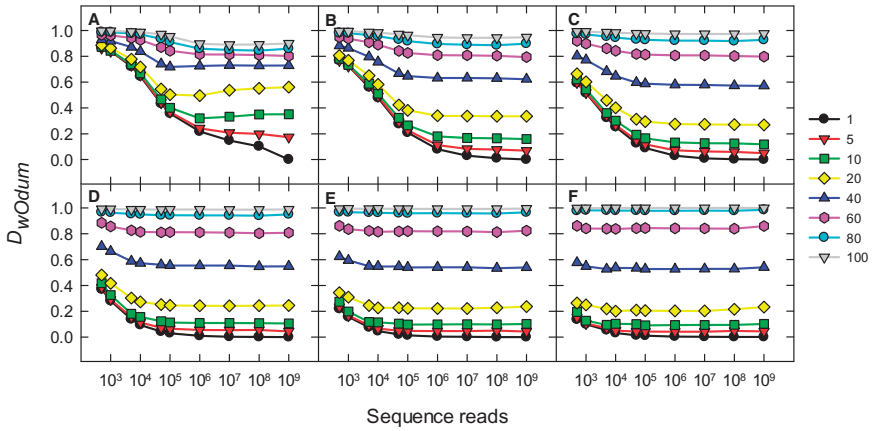

Fig. 5. Sensitivity of a dissimilarity score to incomplete sampling can be reduced by down-weighting the influence of low-abundance community members. Shown are D_{wOdum} scores for pairwise contrasts between location 1 and other locations along the hypothetical environmental gradient, at a range of sampling depths. The D_{wOdum} weighting parameter varies between panels: (A) $\alpha = 0$, (B) $\alpha = 0.5$, (C) $\alpha = 1$, (D) $\alpha = 1.5$, (E) $\alpha = 2$ and (F) $\alpha = 3$

Table 2. Estimates of bias in measuring β-diversity across a simulated environmental gradient, using a number of distance and dissimilarity measures, at a range of sampling depths

		Bias[a]											
		500 reads			1000 reads			5000 reads			10 000 reads		
Score[b]	Alpha	Max. Negative	Max. Positive	MAV[c]	Max. Negative	Max. Positive	MAV	Max. Negative	Max. Positive	MAV	Max. Negative	Max. Positive	MAV
D_{Canberra}	–	−0.788	0.084	0.47	−0.759	0.101	0.45	−0.588	0.174	0.35	−0.447	0.217	0.28
$D_{\text{MorisitaHorn}}$	–	−0.001	0.230	0.13	–	0.136	0.08	−0.007	0.032	0.016	−0.013	0.015	0.008
D_{Soergel}	–	–	0.744	0.32	–	0.679	0.29	–	0.491	0.20	–	0.404	0.15
$D_{\Theta\text{YC}}$	–	–	0.355	0.18	–	0.227	0.11	−0.004	0.062	0.028	−0.008	0.029	0.013
D_{wOdum}	0	–	0.864	0.37	–	0.840	0.36	–	0.720	0.30	–	0.645	0.25
	0.5	–	0.767	0.39	–	0.724	0.36	–	0.561	0.28	–	0.479	0.23
	1	–	0.592	0.31	–	0.515	0.26	–	0.325	0.16	–	0.253	0.12
	1.5	–	0.370	0.19	–	0.281	0.14	–	0.139	0.056	−0.001	0.093	0.034
	2	−0.002	0.218	0.10	–	0.156	0.065	−0.001	0.076	0.020	−0.009	0.046	0.014
	3	−0.007	0.135	0.052	−0.018	0.098	0.028	−0.020	0.053	0.017	−0.032	0.030	0.013
	4	−0.014	0.124	0.046	−0.033	0.087	0.025	−0.035	0.047	0.023	−0.045	0.027	0.017
D_{GUniFrac}	0	0.052	0.349	0.16	–	0.323	0.15	–	0.267	0.13	–	0.209	0.087
	0.5	−0.016	0.259	0.16	−0.021	0.215	0.14	–	0.203	0.12	–	0.052	0.017
	1	−0.047	0.267	0.14	−0.050	0.252	0.12	–	0.267	0.13	−0.017	0.007	0.006
	1.5	−0.058	0.296	0.14	−0.063	0.257	0.11	–	0.382	0.14	−0.020	0.003	0.006
	2	−0.056	0.300	0.13	−0.066	0.231	0.10	–	0.381	0.14	−0.019	0.002	0.006
	3	−0.023	0.286	0.13	−0.052	0.169	0.085	–	0.279	0.11	−0.014	0.006	0.005
	4	–	0.277	0.13	−0.011	0.127	0.051	−0.002	0.224	0.072	−0.009	0.011	0.005

[a]Bias is the difference in score at the defined sampling level, compared to the score for complete sampling (10^9 individuals/reads per sample). Values are limited to the location comparisons shown in Supplementary Figure S1–S3 and represent the average for 100 simulation runs.

[b]For clarity, scores not bound between 0 and 1 (D_{Gower} and $D_{\text{Manhattan}}$) are not shown.

[c]MAV is the mean absolute value.

3.2 Using D_{wOdum} and OTUshuff with biological data

We also present real biological data derived from natural soil communities to demonstrate the effects of D_{wOdum} and OTUshuff on observed dissimilarities between communities from different sites, plots within a site and unique subsamples of sequence reads from the same sample. Dissimilarity between subsamples from the same sample indicates an inflated estimate of β-diversity due to incomplete sampling. Such pseudo β-diversity should be minimized in a way that does not hinder the ability to detect biologically important differences in plot- or site-level comparisons.

When low abundance taxa were more heavily weighted, subsamples of sequence reads drawn from a common pool displayed substantial β-diversity (Fig. 7C, $\alpha \leq 1$). However, reducing the influence of low-abundance taxa on the dissimilarity score reduced this pseudo β-diversity dramatically (Fig. 7C, $\alpha > 1$). Importantly, this down-weighting of low abundance taxa did not substantially impact the scores for cross-site contrasts (Fig. 7A). Varying the influence of low abundance taxa on the score index had a moderate impact on contrasts between plots within a site; calculated dissimilarity between plots within a site decreased as low abundance taxa were given less influence on the score. However, over 75% of pairwise contrasts between communities observed in different plots at the same site remained significant (OTUshuff, $P < 0.05$) even at $\alpha = 4$ (Fig. 7B).

The influence of low abundance taxa on site-specific estimates of β-diversity (i.e. distance or dissimilarity among plots within sites) varied by greatly by site (Fig. 8). For example, when $\alpha = 1$ the range of site-specific β-diversity was 0.678 (AK.2) to 0.852 (AK.1); whereas, when $\alpha = 4$ the range was 0.162 (MA.2) to 0.702 (CO.1). In addition, the patterns observed when rare taxa were down-weighted ($\alpha = 4$) appear to be more consistent with known site attributes, compared to when rare taxa influenced the score to a greater degree ($\alpha = 1$). For example, we would expect to observe the highest β-diversity at the CO.1 site because a variety of different plant species were sampled at this site. In contrast, all other sites tended to be dominated by a single over-story plant species and might therefore be expected to have lower β-diversity among plots. It is only when the influence of rare taxa is down-weighted (Fig. 8B) that is expected pattern becomes evident.

Regardless, it is clear from our simulated and real biological data that pseudo β-diversity can drastically increase β-diversity estimates particularly as samples become more similar (e.g. plot and/or subsample comparisons). Ideally, one could increase sampling depth as sample similarity increases; however, in lieu of this solution the down-weighting of rare OTUs with D_{wOdum} appears to be a viable option to reduce the effects of pseudo β-diversity on estimates of community resemblance. By comparison, $D_{GUniFrac}$ is less effective

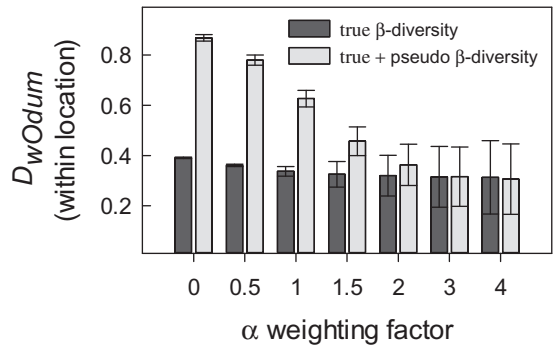

Fig. 6. Effect of varying the influence of low abundance community members (via the α weighting parameter) on a dissimilarity score (D_{wOdum}), using simulated communities. Contrasts are between communities drawn from the same location on our hypothetical environmental gradient, but then perturbed to simulate true β-diversity. Pseudo β-diversity was introduced by shallow sampling (1000 reads). Bars are the mean ± 1 SD for all replicate comparisons (e.g. location 1 versus 1, 5 versus 5, 10 versus 10, etc.)

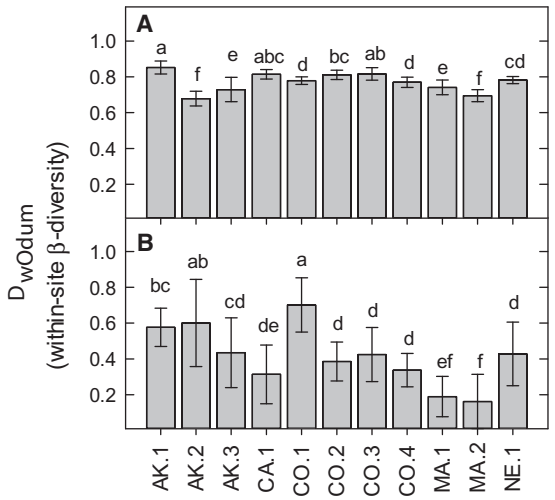

Fig. 8. Within-site β-diversity (i.e. mean dissimilarity among plots within a site) with (A) low abundance community members given moderate influence ($\alpha = 1$) and (B) the influence of low abundance community members reduced ($\alpha = 4$). Within each panel, bars with different letters are significantly different (one-way ANOVA, $P < 0.05$)

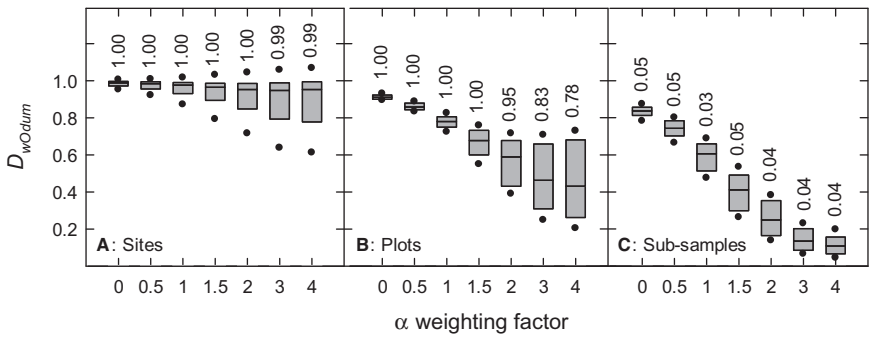

Fig. 7. Effect of D_{wOdum} α weighting parameter on β-diversity for samples from **A:** different sites, **B:** different plots within a site, or **C:** subsamples from the same pool of sequences. All samples contained 1000 sequence reads. Bars represent the first, second and third quartiles, points are ± 1 SD of the mean. Numbers above bars are the proportion of pairwise contrasts that were considered significant (OTUshuff, $P < 0.05$)

at reducing pseudo β-diversity because of limitations to the range of α. Extending α above 1 in an effort to reduce the potential effects of pseudo β-diversity on $D_{GUniFrac}$ scores resulted in distances for all comparisons (sites, plots within sites and subsamples) approaching zero (Supplementary Fig. S9). In other words, site differences were not preserved with $D_{GUniFrac}$ as α was increased.

Existing indices varied widely in their ability to differentiate soil bacterial communities from different sites and to avoid pseudo β-diversity associated with incomplete sampling. For instance, $D_{Canberra}$ between subsamples drawn from a common pool was low, but this index was also unable to detect appreciable difference between sites (Supplementary Fig. S10). In contrast, $D_{Soergel}$ between communities from different sites was high, but this index was also prone to over-estimating β-diversity between subsamples drawn from a common pool (Supplementary Fig. S10).

4 Discussion

Ecologists have a well-developed set of techniques for comparing the composition and structure of communities. However, extension of these approaches into new systems and the implementation of new techniques for surveying community membership may require new approaches and on-going refinement of traditional methods. In particular, the widespread adoption of DNA sequence-based methods of characterizing microbial communities in complex environments such as soil has created unique analytical challenges.

Comprehensive surveys of microbial inhabitants are often unfeasible, and limited resources require balancing a trade-off between depth of characterization and breadth of sampling. It has become quite common for microbial ecologists to sample communities to a very shallow depth, where the number of observations is lower than the number of taxa that are present (Roesch *et al.*, 2007; Eiler *et al.*, 2012; Lundberg *et al.*, 2012); as a result, estimates of distance or dissimilarity between samples tend to be very large, and may arise from either true or pseudo β-diversity depending upon the true underlying similarity (e.g. Fig. 7).

Currently, there are no statistical tools available to evaluate the potential for pseudo β-diversity to inflate estimates of community distance or dissimilarity. Our procedure for pairwise contrasts, OTUshuff, is insensitive to pseudo β-diversity, correctly identifying all subsamples drawn from the same community as not significantly different. Therefore, OTUshuff is a useful tool to identify if scores are likely to be associated with pseudo β-diversity and whether reliable estimates of β-diversity may require additional sampling, or an alternative index. For instance, in lieu of additional sampling, we show that D_{wOdum} with $\alpha \geq 1$ is a suitable approach to remove pseudo β-diversity and provide more accurate estimates of community dissimilarity (Table 2; Fig. 7).

When samples are drawn from genuinely different communities, OTUshuff consistently shows significant differences. In these meaningful contrasts, as we discuss above, varying the choice of community resemblance index, or D_{wOdum} α level, may increase the sensitivity to detect differences. Although a wide array of indices and tests are available, each has its own inherent limitations (Magurran, 2004; Schloss, 2008). While users could pick and choose among existing indices, we have shown that these measures can be sensitive to incomplete sampling, often in unpredictable ways. We also suggest that it is preferable to be able to adjust weightings within the context of a consistent index, rather than switching definitions of community resemblance in order to achieve the desired weightings. The tunable nature of the D_{wOdum} and $D_{GUniFrac}$ scores offers maximum selectivity and sensitivity. For example, D_{wOdum} and $D_{GUniFrac}$ were the only two scores that could consistently detect a significant difference (AMOVA, $P < 0.05$) between location 1 and 2 in our simulated environmental gradient, and this was achieved by increasing the influence of rare taxa ($\alpha = 0$).

We have highlighted at several points that the variable weighting of D_{wOdum} mirrors the capabilities of the phylogenetic distance measure $D_{GUniFrac}$. However, there are substantial differences between these methods. Primarily, $D_{GUniFrac}$ is a phylogenetic distance, which requires the ability to place observed taxa onto a phylogenetic tree so that branch lengths can be used in calculating distances. However, there are applications for which it may not be possible to determine a reliable phylogenetic tree, particularly when using highly variable DNA sequences such as internal transcribed spacer (ITS) data. The weighting parameter for $D_{GUniFrac}$, as originally described, was limited to $\alpha \in [0,1]$. We report $D_{GUniFrac}$ scores at $\alpha > 1$ for purposes of comparison with D_{wOdum}, but with our biological data $D_{GUniFrac}$ became insensitive to between plot and between site β-diversity at these higher levels of α (Supplementary Fig. S9). Thus, D_{wOdum} may be preferable for more severe down-weighting of low abundance taxa. When comparing D_{wOdum} and $D_{GUniFrac}$ scores, two patterns become apparent (i) $D_{GUniFrac}$ variability is not affected as strongly as D_{wOdum} by α level, and (ii) $D_{GUniFrac}$ is prone to large deviations in resultant scores depending upon the number of sequence reads employed (e.g. Supplementary Fig. S8—5000 reads). We suggest that this is due to the additional branch-length weighting in $D_{GUniFrac}$, and depending upon the sampling intensity and the user's interests, such weighting may not be desired.

In setting α for contrasting observed microbial communities, users should consider both the coverage of their sampling (i.e. the proportion of the estimated total number of taxa present that were actually observed) and the expected degree of similarity among the communities that are being contrasted. Lower coverage indicates that many low abundance taxa may have been differentially detected simply due to incomplete sampling, and therefore that the influence of low abundance taxa should be down-weighted in calculating community dissimilarity. If the communities being contrasted are expected to be highly similar (for instance, samples drawn from the same individual host, samples collected under stringently controlled environmental conditions, or samples collected at frequent intervals over a time series), than the low abundance taxa may be critical to measuring community resemblance, and their influence on the dissimilarity score may be increased. As with any framework for simulating datasets, our approach to simulating microbial communities has limitations. For instance, simulated communities found along our hypothetical environmental gradient followed a normal distribution in the decline of abundance with distance from their optimum. This is a simplifying assumption. We did not attempt to contrast communities influenced simultaneously by more than one ecological gradient, or differing in underlying population structures (e.g. abundance distributions, evenness, or richness). As for any new metric, additional testing under various simulation strategies is warranted to better understand its controlling factors.

In conclusion, the new analytical tools introduced here will facilitate progress in understanding the forces that structure complex microbial communities. D_{wOdum} is a dissimilarity index for which the influence of low abundance taxa is tunable. OTUshuff is a new statistical test to determine whether two samples are drawn from genuinely different communities.

Acknowledgements

The authors thank Steve Swenson for soil samples collected in Alaska and Elisha Allen for sampling the golf courses. Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. This article was the work of U.S. Government employees engaged in their official duties and is exempt from copyright.

Funding

A portion of this work was supported by a USDA NIFA AFRI Postdoctoral Fellowship (grant number 2011-67012-30938 to M.B.). Funding for supplies was provided in part from the Joint Venture Agreement between the Forest Service, Pacific Northwest Research Station and Oregon State University (09-JV-112261957-021).

Conflict of Interest: none declared.

References

Bray,R.J. and Curtis,J.T. (1957) An ordination of upland forest communities of southern Wisconsin. *Ecol. Monogr.*, **27**, 325–349.

Beck,J. *et al.* (2013) Undersampling and the measurement of beta diversity. *Method Ecol. Evol.*, **4**, 370–382.

Chao,A. *et al.* (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol. Lett.*, **8**, 148–159.

Chen,J. *et al.* (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distance. *Bioinformatics*, **28**, 2106–2113.

Deng,Y. *et al.* (2012) Elevated carbon dioxide alters the structure of soil microbial communities. *Appl. Environ. Microbiol.*, **78**, 2991–2995.

Edgar,R.C. *et al.* (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Eiler,K.G. *et al.* (2012) Digging deeper to find unique microbial communities: The strong effect of depth on the structure of bacterial and archaeal communities in soil. *Soil Biol. Biochem.*, **50**, 58–65.

Essah,S.Y.C. *et al.* (2012) Cover crops can improve potato tuber yield and quality. *HortTechnology*, **22**, 185–190.

Grassle,J.F. and Smith,W. (1976) A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities. *Oecologia*, **25**, 13–22.

Hamady,M. *et al.* (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods*, **5**, 235–237.

Huse,S.M. *et al.* (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.*, **12**, 1889–1898.

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

Kuffner,M. *et al.* (2012) Effects of season and experimental warming on the bacterial community in a temperate mountain forest soil assessed by 16 S rRNA gene pyrosequencing. *FEMS Microbiol. Ecol.*, **82**, 551–562.

Lagos,L.M. *et al.* (2014) Bacterial community structures in rhizosphere microsites of ryegrass (*Lolium perenne* var. Nui) as revealed by pyrosequencing. *Biol. Fertil. Soils*, **50**, 1253–1266.

Leinster,T. and Cobbold,A. (2012) Measuring diversity: the importance of species similarity. *Ecology*, **93**, 477–489.

Lundberg,D.S. *et al.* (2012) Defining the core *Arabidopsis thaliana* root microbiome. *Nature*, **488**, 86–90.

Magurran,A.E. (2004) *Measuring Biological Diversity*. Blackwell Publishing, Oxford.

Manter,D.K. *et al.* (2010) Pyrosequencing reveals a highly diverse and cultivar-specific bacterial endophyte community in potato roots. *Microb. Ecol.*, **60**, 157–166.

Odum,E.P. (1950) Bird populations of the Highlands (North Carolina) plateau in relation to plant succession and avian invasion. *Ecology*, **31**, 587–605.

Quince,C. *et al.* (2011) Removing noise from pyrosquenced amplicons. *BMC Bioinf.*, **12**, 38.

Roesch,L.F. *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*, **1**, 283–290.

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

Schloss,P.D. *et al.* (2009) Introducing mothur: open source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

Shade,A. *et al.* (2014) Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio*, **5**, e01371–e01314.

Sherrod,L.A. *et al.* (2014) Soil and rainfall factors influencing yields of a dryland cropping system in Colorado. *Agron. J.*, **106**, 1179–1192.

Somerfield,P.J. (2008) Identification of the Bray-Curtis similarity index: Comment on Yoshioka (2008). *Mar. Ecol. Prog. Ser.*, **372**, 303–306.

Wang,Q. *et al.* (2007) Naive Bayesian classifier for rapid assessment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.

Yoshioka,P.M. (2008) Misidentification of the Bray-Curtis similarity index. *Mar. Ecol. Prog. Ser.*, **368**, 309–310.