

ORIGINAL ARTICLE

Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities

Donovan H Parks and Robert G Beiko

Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

High-throughput sequencing techniques have made large-scale spatial and temporal surveys of microbial communities routine. Gaining insight into microbial diversity requires methods for effectively analyzing and visualizing these extensive data sets. Phylogenetic β -diversity measures address this challenge by allowing the relationship between large numbers of environmental samples to be explored using standard multivariate analysis techniques. Despite the success and widespread use of phylogenetic β -diversity measures, an extensive comparative analysis of these measures has not been performed. Here, we compare 39 measures of phylogenetic β diversity in order to establish the relative similarity of these measures along with key properties and performance characteristics. While many measures are highly correlated, those commonly used within microbial ecology were found to be distinct from those popular within classical ecology, and from the recently recommended Gower and Canberra measures. Many of the measures are surprisingly robust to different rootings of the gene tree, the choice of similarity threshold used to define operational taxonomic units, and the presence of outlying basal lineages. Measures differ considerably in their sensitivity to rare organisms, and the effectiveness of measures can vary substantially under alternative models of differentiation. Consequently, the depth of sequencing required to reveal underlying patterns of relationships between environmental samples depends on the selected measure. Our results demonstrate that using complementary measures of phylogenetic β diversity can further our understanding of how communities are phylogenetically differentiated. Open-source software implementing the phylogenetic β -diversity measures evaluated in this manuscript is available at <http://kiwi.cs.dal.ca/Software/ExpressBetaDiversity>.

The ISME Journal (2013) 7, 173–183; doi:10.1038/ismej.2012.88; published online 2 August 2012

Subject Category: microbial ecology and functional diversity of natural habitats

Keywords: β diversity; community ecology; metagenomics; phylogenetic β -diversity; similarity indices

Introduction

Advances in DNA sequencing technology allow high-throughput recovery of genetic material directly from environmental samples. By using the small ribosomal subunit RNA (16S rRNA) gene to establish the members of naturally occurring microbial communities, large-scale surveys have shed light on spatio-temporal patterns of microbial diversity (Martiny *et al.*, 2006; Caporaso *et al.*, 2011). Recent studies have revealed the relative influences of environmental factors on global patterns of diversity (Lozupone and Knight, 2007; Lauber *et al.*, 2009; Rousk *et al.*, 2010), the impact of antibiotics on the gut microbiota of mice and humans (Dethlefsen *et al.*, 2008; Ubeda *et al.*, 2010), and established

that human-associated communities differ between individuals and body habitats (Costello *et al.*, 2009; Turnbaugh *et al.*, 2009; Fierer *et al.*, 2010). With surveys now encompassing hundreds of environmental samples, a primary challenge is to identify the environmental factors that engender differences in microbial community structure. β -diversity measures address this challenge by providing a univariate statistic establishing the relative similarity of any pair of samples. Exploratory multivariate statistical techniques, such as hierarchical clustering and ordination, can then be used to identify trends across large numbers of samples.

While β -diversity measures have traditionally been determined on the basis of overlap between discretely defined sets of entities (species or operational taxonomic units, that is, OTUs), recent methods have incorporated phylogenetic information in order to establish the relative similarity of OTUs (Clarke and Warwick, 1998; Martin, 2002; Lozupone and Knight, 2005; Graham and Fine, 2008). By exploiting the hierarchical relatedness of organisms,

Correspondence: RG Beiko, Faculty of Computer Science, Dalhousie University, 6050 University Avenue, Halifax, Nova Scotia, B3H 1W5 Canada.

E-mail: beiko@cs.dal.ca

Received 6 February 2012; revised 18 June 2012; accepted 22 June 2012; published online 2 August 2012

Table 1 Details of empirical data sets

<i>Data set</i>	<i>Samples</i>	<i>Seqs/sample</i> (mean \pm s.d.)	<i>Study design</i>	<i>Main results</i>	<i>Reference</i>
Keyboard	89	1183 \pm 250	Samples taken from the fingertips and keyboards of three individuals	Samples from fingertips and keyboards clustered by individuals	Fierer <i>et al.</i> (2010)
Mouse	40	901 \pm 394	Samples taken from the ileum and cecum of four groups of mice before and after treatment with antibiotics	Antibiotic treatment nearly completely displaced the normal microbiota of the small and large intestine	Ubeda <i>et al.</i> (2010)
Human	76	1534 \pm 635	Samples taken from 27 body sites in 7–9 individuals on four occasions	Community composition was determined primarily by body habitat	Costello <i>et al.</i> (2009)
Soil	22	1662 \pm 459	Samples collected from soil across a pH gradient ranging from 4.0 to 8.3	Relative abundance and diversity of bacteria were positively related to pH	Rousk <i>et al.</i> (2010)

phylogenetic β -diversity measures are often more effective at revealing underlying ecological patterns (Hamady *et al.*, 2010; Nipperess *et al.*, 2010). However, as a univariate statistic, a single β -diversity measure cannot address all manners in which the similarity between samples may be usefully defined. Consequently, many different measures of β diversity have been proposed that vary in their treatment of community properties such as the presence of rare OTUs or the relative abundance of OTUs (Legendre and Legendre, 1998). This latter factor is commonly used to classify β -diversity measures as either *quantitative* or *qualitative*. These two classes of measures provide complementary information as quantitative measures consider the relative abundance of each OTU to indicate if ecological differences between habitats have caused the abundance of taxonomic groups to change, while qualitative measures consider only the presence or absence of an OTU and suggest whether or not ecological factors prohibit a taxonomic group from occupying certain habitats.

Due to the complexity of naturally occurring communities and the wide range of mechanisms that can cause communities to differentiate, it can be beneficial to apply several phylogenetic β -diversity measures. Our aim is to establish a set of properties and a methodology for determining a practical subset of measures that will provide complementary information on the similarity of microbial samples. We build upon initial surveys that have considered a limited number of phylogenetic β -diversity measures under a restrictive set of conditions (Schloss, 2008; Nipperess *et al.*, 2010; Root and Nelson, 2011; Swenson, 2011) by establishing key properties and performance measures for 24 quantitative and 15 qualitative measures (Supplementary Table S1). We consider phylogenetic β -diversity measures popular within microbial and classical ecology along with newly established phylogenetic extensions of commonly used taxon-based (non-phylogenetic) measures. Performance measures are established over two distinct models of differentiation that highlights critical aspects of the considered measures, the need to apply multiple measures, and shortcomings of the taxon-based measures previously recommended for

assessing differences between microbial communities (Kuczynski *et al.*, 2010). We contrast the performance of measures on complete and random subsets of four pyrosequencing data sets (Table 1): (1) fingertip and keyboard samples from three individuals used for forensic identification (Fierer *et al.*, 2010); (2) small and large intestinal samples taken from four groups of mice before and after antibiotic treatment (Ubeda *et al.*, 2010); (3) samples collected from the navel, mouth, hair and stool of several individuals (Costello *et al.*, 2009); and (4) soil samples taken across a substantial pH gradient (Rousk *et al.*, 2010). By considering random subsets of samples from these four distinct data sets we are able to evaluate measures over a range of tree topologies spanning samples with varying levels of inter- and intra-sample diversity.

Materials and methods

Empirical data sets

The properties and effectiveness of measures were assessed using the four data sets mentioned above. Data sets were processed using a common pipeline, but with the data set-specific filtering criteria specified in the original publications. Sequences were removed from the analysis if they were <200 bp or greater than a specific length (keyboard, soil: 300 bp; human, mouse gut: 400 bp), had a quality score <25 , contained ambiguous characters, contained an unrecognized barcode or did not contain the primer sequence. Sequences were aligned using the mothur v1.22.1 (<http://www.mothur.org/>; Schloss *et al.*, 2009) implementation of the NAST algorithm with the Greengenes reference alignment (DeSantis *et al.*, 2006). We removed sequences with an alignment length <150 or an identity with the reference alignment of $<75\%$ along with any samples containing an insufficient number of sequences, as defined by the original study authors (keyboard, human: 800; soil: 600; mouse gut: no filtering). Hypervariable columns of the alignment were removed using the PH Lane mask. Phylogenetic trees were inferred using FastTree v2.1.4 (<http://www.microbesonline.org/fasttree/>; Price *et al.*, 2009) with a generalized time-reversible model.

Trees were rooted with a monophyletic outgroup of three archaeal sequences from distinct phyla.

Evaluating properties of phylogenetic β -diversity measures

We evaluated 39 phylogenetic β -diversity measures used within microbial and classical ecology along with newly established phylogenetic extensions of commonly used taxon-based measures (Supplementary Methods S1; Supplementary Table S1). The properties of these measures were evaluated using 100 randomly selected subsets of 10 samples from each of the four empirical data sets. All sequences within a selected sample were used in all cases. Subsets of samples were considered in order to gauge the robustness of results in light of varying tree topologies, amounts of diversity spanned by a data set and patterns of phylogenetic similarity between samples. The hierarchical similarity of measures was determined by applying the unweighted paired group method with arithmetic mean (UPGMA) clustering algorithm to a matrix indicating the mean Pearson dissimilarity, $d = 1 - r$, between each pair of measures. Correlation, r , was determined using the Pearson correlation coefficient as implemented in SciPy v0.9.0 (www.scipy.org). To evaluate the influence of sequence clustering, we clustered sequences using the furthest neighbor algorithm in mothur v1.22.1. Trees at different OTU thresholds were obtained by randomly selecting a representative sequence from each cluster and pruning the tree to the set of representative sequences. To evaluate the robustness of measures to root placement, phylogenies were randomly rerooted 100 times for each of the 100 subsets. Trees were rooted by randomly selecting a new node to be the root and using BioPython v1.58 (<http://biopython.org>; Cock *et al.*, 2009) to reroot the tree. The addition of an outlying basal lineage was evaluated by appending a new lineage to the root of each data set's phylogeny. This lineage consisted of a single branch whose length was set to the average distance from a leaf node to the root. Additional sequences were added to a sample and placed at the leaf node of the outlying lineage. Ordination plots indicating the similarity of samples were obtained using principal coordinate analysis. UPGMA and principal coordinate analysis results were obtained using custom C++ code.

Simulated cluster data

We simulated samples belonging to distinct clusters under two different models of differentiation, which we term the equal-perturbation and dominant-pair models. The equal-perturbation model extends the methodology proposed by Kuczynski *et al.* (2010) to the evaluation of phylogenetic β -diversity measures. This model simulates microbial communities where a process stochastically influences the abundance of each OTU by an amount dependent on the initial

abundance of that OTU. For each of the empirical data sets, we randomly selected a seed sample. We then perturbed this seed sample by multiplying the relative abundance of each OTU by a random number drawn from a normal distribution with unit mean and standard deviation $\sigma_1 = 1.0$. This was repeated three times in order to create starting distributions for three distinct clusters. These starting distributions were then renormalized to sum to 1.0. We generated 30 samples within each cluster by perturbing these three starting distributions, using a random number drawn from a normal distribution with unit mean and standard deviation $\sigma_2 = 0.5$. The simulated samples were then renormalized and sequence counts obtained by drawing, with replacement, a specified number of sequences from each of these sample distributions. The values of σ_1 and σ_2 were set to 1.0 and 0.5, respectively, in order to approximate the clustering pattern of the keyboard data set (Kuczynski *et al.*, 2010).

The dominant-pair model simulates microbial communities where a process primarily influences the abundance of the two most abundant OTUs and only has a small stochastic effect on the remaining OTUs. This model is a simplified version of the shift seen in many communities where one predominant OTU is replaced with another. For example, in enhanced biological phosphorus removal communities, the primary strains of phosphate-accumulating *Candidatus* 'Accumulibacter phosphatis' can decrease dramatically, with a concomitant increase in the frequency of other organisms such as *Candidatus* 'Competibacter', possibly due to viral predation (Barr *et al.*, 2010; Slater *et al.*, 2010). To simulate this scenario, we initially perturb only the two most abundant OTUs in the seed sample in order to create the starting distributions for each cluster. The two most abundant OTUs were perturbed by an amount $\delta = d \cdot x - x$, where x is the abundance of the most abundant OTU in the seed sample and d is the central absolute moment of a normal distribution with $\sigma = 1.0$ (that is, 1.797). The three starting distributions were created by modifying the two most abundant OTUs in the seed sample by $(+\delta, -\delta)$, $(-\delta, +\delta)$ and $(0, 0)$. Samples within each cluster were then obtained as before. Since δ is the average expected change of an OTU under the equal-perturbation model, differences in the performance of a measure are expected to be the result of the models themselves and not an artifact of the relative distinctiveness of clusters. Randomizations under both models were repeated using 100 different seed samples from each empirical data set.

Evaluation of measures on simulated cluster data

The ability of measures to recover simulated patterns of clustering was evaluated using two statistics. For the first statistic, we clustered samples with the k -medoids algorithm implemented in BioPython

v1.58 and calculated the fraction of samples whose k -medoids clustering matched the known clustering of the samples. For the second statistic, we calculated the UPGMA clustering of the simulated samples and determined the consistency index (Kluge and Farris, 1969) of this hierarchical cluster tree (that is, minimum number of state changes divided by the number of state changes required to explain the observed distribution of cluster assignments within the tree). When sorting and summarizing the performance of measures we favored results obtained with moderate sequence depth (1000 sequences/sample) and the k -medoids statistic as it operates more directly on the dissimilarity matrix compared with the consistency index which is calculated on the inferred UPGMA tree. Despite their differences, the two statistics were found to be highly correlated. Note that these measures differ from those previously proposed by Kuczynski *et al.* (2010) as we apply them directly to a measure's dissimilarity matrix as opposed to the distance between samples within an ordination plot, which represents only one of many possible visualizations of a dissimilarity matrix.

Classifying measures by the branches they consider
Phylogenetic β -diversity measures can be classified according to the set of branches that influence the calculation of community dissimilarity (Supplementary Figure S1). A measure only influenced by branches within the most recent common ancestor (MRCA) subtree spanned by a pair of samples is classified as an MRCA measure. In contrast, a measure also influenced by the 'deep branches' that extend from the root of the MRCA subtree to the root of the subtree spanning all samples is termed a complete lineage (CL) measure, and a measure influenced by all branches in the tree inferred from all data under consideration is termed a complete tree (CT) measure.

Results

Identifying complementary measures

To explore the similarity of measures, we calculated statistics over randomly selected subsets of the four empirical data sets and visualized the mean correlation between measures using hierarchical cluster trees (Figure 1; see Supplementary Figure S2 for

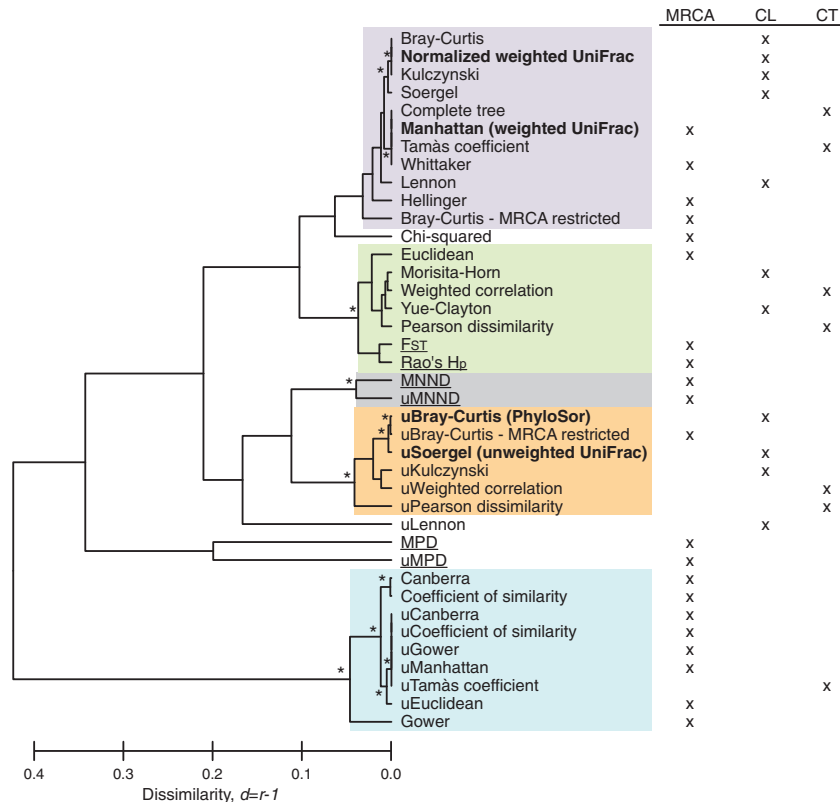


Figure 1 Similarity of phylogenetic β -diversity measures. Branch lengths are transformed Pearson's r values, $d = r - 1$, averaged over 100 random subsets of 10 samples drawn from each of the four empirical data sets. The hierarchical relationship between measures was obtained using the UPGMA clustering algorithm. Branches supported by at least 70% of the trials are indicated with asterisks. The five most highly correlated and consistently clustered groups of measures are highlighted in different colors. These clusterings are nearly perfectly recovered on all four data sets (Supplementary Figure S2). Phylogenetic measures commonly used within microbial ecology are shown in bold while measures popular in classical ecology are underlined. Measures are specified by their common quantitative name and qualitative counterparts indicated by prefixing a 'u' (for 'unweighted'). Each measure is classified as a MRCA, CL, or CT measure.

data set-specific results). This revealed a number of highly and perfectly correlated measures (Supplementary Table S2). Notably, unweighted UniFrac (uSoergel) and PhyloSor (uBray-Curtis) are highly correlated (Pearson's $r = 1.00 \pm 1.1 \cdot 10^{-3}$ s.d.) while normalized weighted UniFrac is identical to the Bray-Curtis measure (Supplementary Equations S1). Corresponding quantitative and qualitative measures were found to be only moderately correlated (Pearson's $r = 0.66 \pm 0.16$ s.d.) with the notable exception of the MNND, Canberra, Gower and coefficient of similarity measures. Measures commonly used within microbial ecology (for example, UniFrac variants) were found to be distinct from those popular for studying macroorganisms (that is, MNND, MPD, Rao's H_p , F_{ST}), and from the Gower and Canberra measures recently recommended for assessing microbial community relationships from species profiles (Kuczynski *et al.*, 2010).

Robustness to sequence clustering

In order to reduce computational requirements, similar sequences can be clustered together and a single representative sequence from each cluster used during sequence alignment and phylogenetic inference. A common threshold for clustering full-length 16S rDNA sequences is 97% as this roughly corresponds to the working definition of a microbial

species (Stackebrandt and Goebel, 1994), but OTUs may be usefully defined over a wide range of similarity thresholds, for example, 80–99% (Schloss and Handelsman, 2004; Bryant *et al.*, 2008). To evaluate the influence of sequence clustering on phylogenetic β -diversity measures, we assessed the degree of correlation between measures on random subsets of samples before and after clustering at various levels of sequence similarity.

Clustering sequences results in a substantial reduction in branch length and number of leaf nodes within a phylogeny (Supplementary Figure S3). At 97% sequence similarity, branch length was reduced by between 8% (soil study) and 44% (mouse gut study). Nonetheless, the majority of quantitative phylogenetic β -diversity measures exhibited only a slight decrease in correlation with the dissimilarity results obtained before clustering (Figure 2a). Even at 85% sequence similarity, all measures remained highly correlated (Pearson's $r > 0.92$ for all trials) except for the Canberra, coefficient of similarity, Gower, and Pearson dissimilarity measures (Figure 2b; Supplementary Table S3). Qualitative measures were more sensitive to sequence clustering and varied more substantially between data sets (Figure 2c). While all qualitative measures were more sensitive to sequence clustering than their quantitative counterparts, the uMPD measure exhibited extreme sensitivity to the extent of being

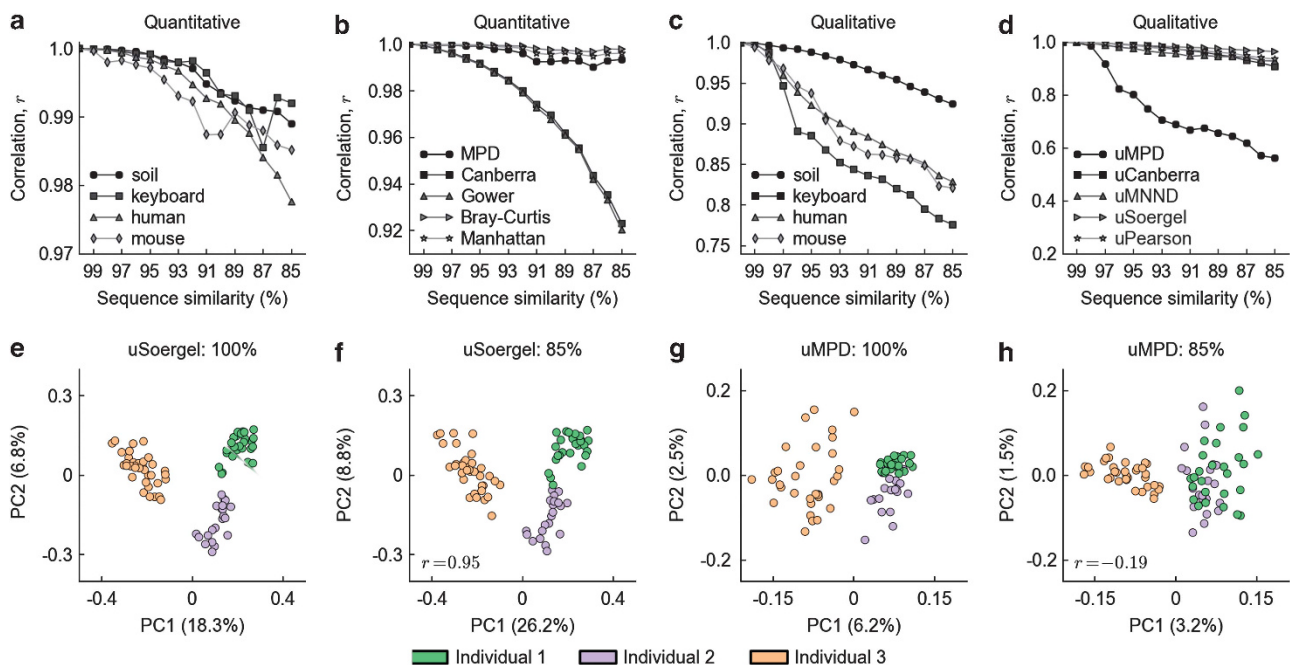


Figure 2 Influence of sequence clustering on phylogenetic β -diversity measures. (a, c) Mean correlation across all quantitative (a) and qualitative (c) measures on 100 randomly selected subsets of 10 samples from each empirical data set. (b, d) Correlation of select quantitative (b) and qualitative (d) measures averaged over all four empirical data sets. (e, f) Ordination plots obtained by applying the qualitative Soergel measure to the keyboard data set with sequences clustered at 100% (e) and 85% (f) sequence similarity. (g, h) Ordination plots for the qualitative MPD measure with sequences clustered at 100% (g) and 85% (h) sequence similarity. Principal coordinate analysis was used to generate the ordination plots. The percentage of total variance explained by each axis is shown in parentheses. Each data point represents a sample taken from one of three individuals. Pearson's correlation coefficient, r , between dissimilarity values measured before and after clustering is given in the bottom-left corner of each plot.

negatively correlated with the unclustered results for certain trials (Figure 2d; Supplementary Table S4). Although the majority of measures revealed the same biological patterns between microbial samples even at 85% sequence similarity (Figures 2e and f), measures sensitive to clustering can fail to recover the same patterns (Figures 2g and h). The observed robustness to the choice of OTU clustering threshold may be perceived as a positive attribute of these measures, but also demonstrates a lack of sensitivity to fine-scale differences in community structure. In the remainder of this paper we consider OTUs formed at 97% sequence similarity as all measures (with the exception of uMPD) were found to be highly correlated (Pearson's $r > 0.93$ for all trials) with their original dissimilarity values at this clustering threshold.

Robustness to outlying lineages

Outlying lineages may occur due to errors in sequence alignment or phylogenetic inference, or simply due to the stochastic detection of rare outlying taxonomic groups. To evaluate the robustness of measures to the addition of an outlying basal lineage, we added a single branch to the root node of each phylogenetic tree and set the length of this branch to the average distance from each leaf node to the root. This simulates an outlying lineage such as a deeply branching phylum or superkingdom not always observed in the community. We then generated random subsets of samples from each of the empirical data sets and randomly selected half of these samples to contain sequences within the outlying lineage. Even with 5% of sequences assigned to the outlying lineage, nearly all quantitative measures were well correlated (Pearson's $r > 0.81$ for every trial except MNND, $r > 0.57$, and Pearson dissimilarity, $r > 0.67$) with their original dissimilarity values (Figure 3; Supplementary Table S5). Despite the all-or-nothing nature of qualitative measures, they were robust (Pearson's $r > 0.88$ for every trial) to the addition of an outlying basal lineage (Figure 3; Supplementary Table S6). The sole exception is the uPearson dissimilarity measure, which was highly sensitive (mean Pearson's $r = 0.42$) to the inclusion of an outlying lineage and failed to recover the expected relationship between samples on both the human (Figure 3) and soil (Supplementary Figure S4) data sets. The general robustness of measures can be contributed to the outlying branch representing only a small portion of the total branch length and consequence having only a moderate influence on the dissimilarity calculated between samples. Most measures will be robust to any perturbation of the phylogeny influencing only a small portion of the total branch length.

Robustness to root placement

In the absence of a credible outgroup for rooting a phylogeny, it is beneficial to have measures that can

be applied to unrooted trees. Several measures are invariant to root placement including the MNND, MPD, Rao's H_p and F_{ST} measures popular within classical ecology, and several other quantitative measures including the Gower, Manhattan and Euclidean measures (Supplementary Equations S2). We evaluated the robustness of the remaining measures to root placement by considering dissimilarity values obtained on random subsets of samples before and after randomly rerooting their corresponding phylogeny (Supplementary Tables S7 and S8). The quantitative Canberra and coefficient of similarity measures were found to be highly robust to root placement (Pearson's $r > 0.99$ for all trials). The remaining quantitative measures showed sensitivity to at least some random root placements, including the Bray-Curtis (normalized weighted UniFrac, minimum Pearson's $r = 0.30$) measure. All qualitative measures were found to be robust to root placement (Pearson's $r > 0.92$ for all trials) with the exception of the uPearson dissimilarity measure (minimum Pearson's $r = 0.73$). In general, measures are robust to root placement as only branches along the path to the original root to the new root differ in their contribution to the dissimilarity measured between a pair of samples.

Robustness to rare OTUs

Measures vary in their treatment of rare OTUs. For randomly selected subsets of samples we determined the correlation between dissimilarity values obtained before and after filtering OTUs containing only a single sequence, $< 0.1\%$ of sequences, or $< 1\%$ of sequences. All quantitative measures were relatively insensitive to the removal of lineages containing only rare OTUs (Pearson's $r > 0.85$ for all trials at 0.1% filtering) with the exception of the Canberra (minimum Pearson's $r = 0.17$), coefficient of similarity (minimum Pearson's $r = 0.08$) and Gower (minimum Pearson's $r = 0.07$) measures that were highly sensitive (Supplementary Table S9). In contrast, all qualitative measures can be substantially affected by the removal of rare OTUs as they are sensitive to the removal of lineages irrespective of the number of sequences assigned to a lineage (Supplementary Table S10).

Revealing clusters of samples

We assessed the ability of measures to identify discrete clusters of samples under two models of phylogenetic differentiation. Under the equal-perturbation model, an initial seed sample was selected from one of the four data sets and the relative abundance of *each* OTU perturbed by a random percentage in order to create three starting distributions. We then applied a relatively small perturbation to these starting distributions in order to generate three clusters each containing 30 distinct samples. Model parameters were selected to mimic

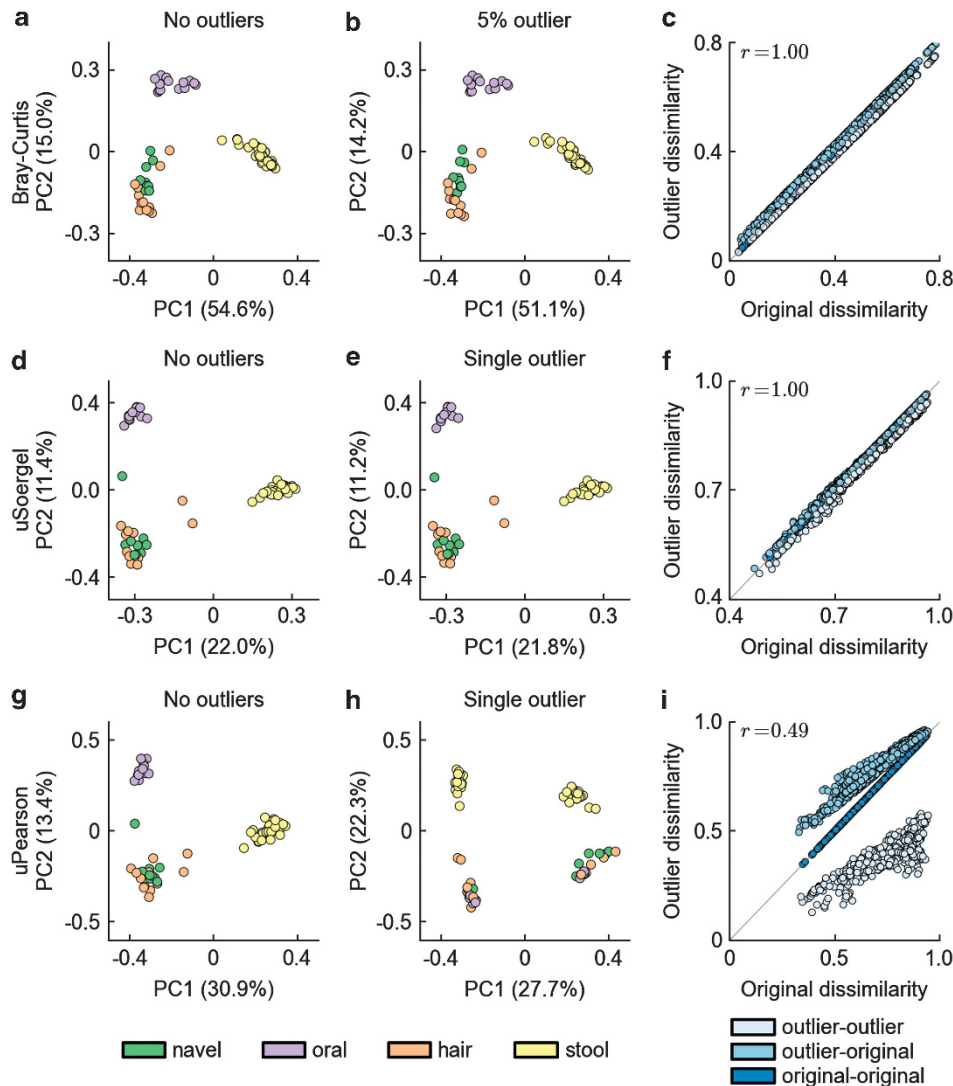


Figure 3 Recovery of clusters is influenced by a measure's robustness to outlying basal lineages. The quantitative Bray-Curtis (a–c), qualitative Soergel (d–f), and qualitative Pearson dissimilarity (g–i) measures were applied to the human data set. (a, d, g) All three methods revealed three clusters: a stool cluster, an oral cluster and a mixed navel and hair cluster. The addition of an outlying basal lineage to half the samples did not substantially affect the Bray-Curtis (b: 5% of sequences assigned to the outlying lineage) or uSoergel (e) measures, but obscured the underlying biological clusters for the uPearson dissimilarity (h) measure. For qualitative measures, a single sequence is sufficient to include the outlying lineage and the addition of further sequences does not influence these measures. Each data point in the scatter plots (c, f, i) indicates the dissimilarity measured between a pair of samples before (x-axis) and after (y-axis) adding sequences to the outlying lineage. For all measures, the addition of the outlying lineage caused pairs of samples where both contained the outlying lineage to become more similar (outlier–outlier) and pairs of samples where only one sample contained the outlying lineage to become less similar (outlier–original). Pairs of samples where neither contained the outlying lineage were unaffected (original–original). However, the degree to which the outlier–outlier and outlier–original pairs were affected depended on the measure used. The Pearson's correlation coefficient, r , between dissimilarity values measured before and after addition of the outlying lineage is given in the upper-left corner of each scatter plot.

the clustering pattern of the keyboard data set. Under the dominant-pair model, the initial perturbation of the seed sample was restricted to the two most abundant OTUs. This was followed by a more subtle stochastic process applied to all OTUs in order to again generate three clusters of 30 distinct samples. Clusters were simulated under both models for 100 randomly selected seed samples from each data set and at varying sequencing depths.

The relative effectiveness of measures was dependent on the simulated model (Supplementary Table S11). Measures sensitive to lineages containing rare OTUs such as Canberra and Gower performed strongly under the equal-perturbation model, but failed to identify clustering patterns under the dominant-pair model (Figure 4; Supplementary Tables S12–S19). The most effective measures under the dominant-pair model were those highly sensitive to the most abundant OTUs such as

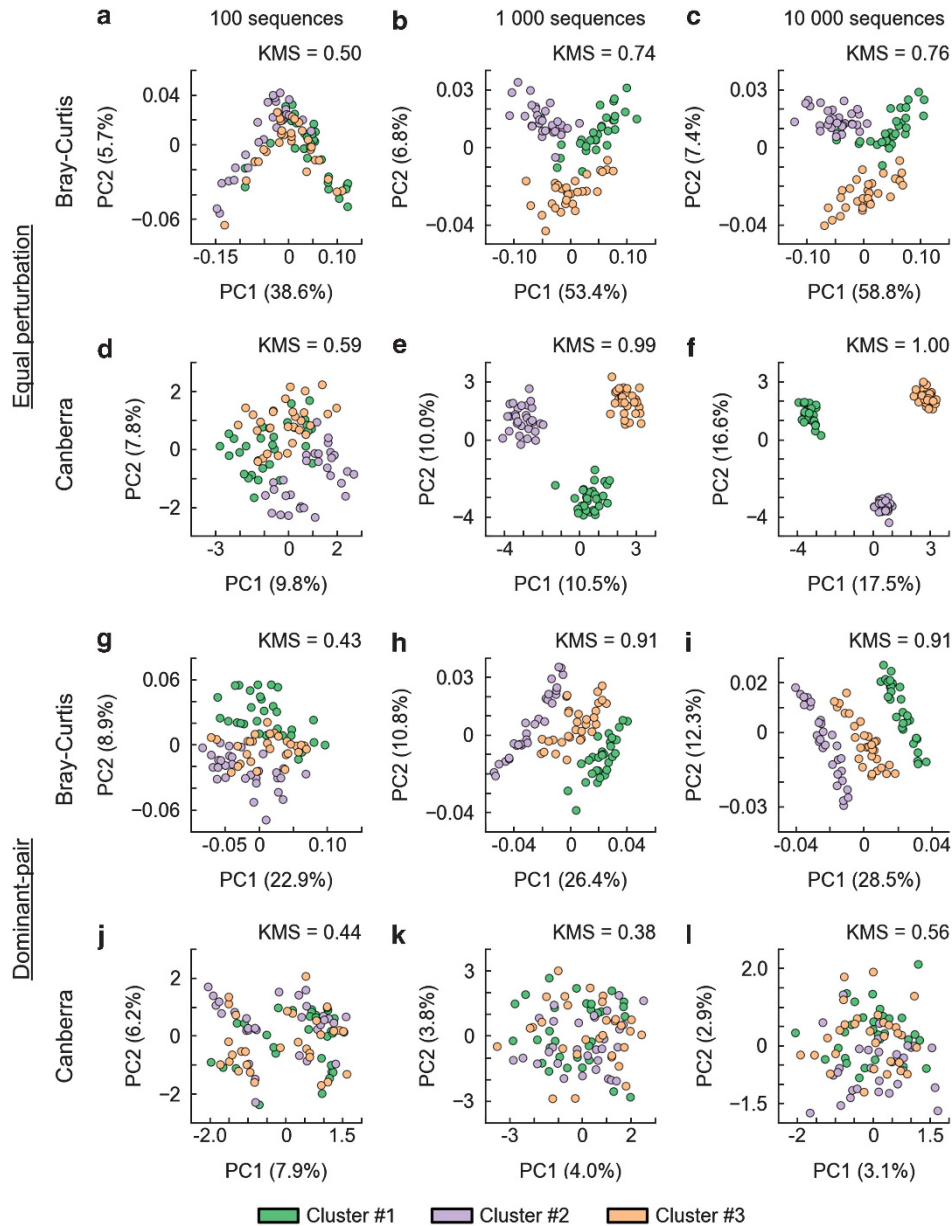


Figure 4 Effectiveness of measures depends on the mechanism of phylogenetic differentiation and sequencing depth. The Bray-Curtis (a–c) and Canberra (d–f) measures were applied to clusters obtained under the equal-perturbation model at sequencing depths of 100, 1000 or 10000 sequences per sample. These measures were also applied to clusters generated under the dominant-pair model (g–l). The *k*-medoids score (KMS) is given in the upper-right corner of each ordination plot.

Morisita-Horn and Euclidean (Magurran, 2004). Measures that are relatively insensitive to rare OTUs while not being overly sensitive to the most abundant OTUs, such as the Bray-Curtis, Soergel and Manhattan measures, performed moderately well under both models (Supplementary Table S20). Although the performance of each measure depended on the empirical data set from which the initial seed sample was drawn, the relative performance of the measures was remarkably stable under both models (Supplementary Tables S21 and S22). Consequently, we are confident that these results are not an artifact of the phylogenetic structure or diversity of a particular data set. We also found the

hierarchical clustering of measures under these two models of differentiation to largely resemble those obtained on the empirical data sets (Supplementary Figure S5).

Discussion

Phylogenetic measures of β diversity can be classified as MRCA, CL or CT based on the set of branches that influence the dissimilarity calculated between a pair of communities. Measures operating over different sets of branches can be highly correlated (Figure 1), while measures operating over the same

Table 2 Properties and performance of phylogenetic β -diversity measures

Contributing branches	Root invariant	Sensitive to rare OTUs	Robust to outlying lineages	Highly effective on equal-perturbation model	Highly effective on dominant-pair model	Measure(s)
<i>Quantitative measures</i>						
MRCA	Yes	Yes	Yes	Yes	No	Gower
MRCA	Yes	No	Yes	No	No	Manhattan, MPD, Whittaker
MRCA	Yes	No	Yes	No	Yes	Euclidean, Rao's H_p , F_{ST}
MRCA	Yes	No	Yes	Yes	No	MNND
CT	Yes	No	Yes	No	No	CT, Tamàs coefficient
MRCA	No	Yes	Yes	Yes	No	Canberra, Coefficient of similarity
MRCA	No	No	Yes	Yes	No	Hellinger
MRCA	No	No	Yes	No	No	Bray-Curtis (MRCA restricted)
CT	No	No	Yes	No	No	Pearson dissimilarity
CT	No	No	Yes	No	Yes	Weighted correlation
CL	No	No	Yes	No	No	Bray-Curtis, Kulczynski, Lennon, Soergel
MRCA	No	No	Yes	Yes	No	Chi-squared
CL	No	No	Yes	No	Yes	Yue-Clayton, Morisita-Horn
<i>Qualitative measures</i>						
MRCA	Yes	Yes	Yes	No	No	uMNND, uMPD
MRCA	No	Yes	Yes	No	No	uCanberra, uEuclidean, uGower, uManhattan, uCoefficient of similarity, uBray-Curtis (MRCA restricted)
CT	No	Yes	Yes	No	No	uTamàs coefficient, uWeighted correlation
CT	No	Yes	No	No	No	uPearson dissimilarity
CL	No	Yes	Yes	No	No	uBray-Curtis, uKulczynski, uLennon, uSoergel

Abbreviations: CL, complete lineage; CT, complete tree; MRCA, most recent common ancestor. A measure was deemed sensitive to rare OTUs if the minimum Pearson's correlation coefficient was <0.8 on any subset at 0.1% filtering. A measure was considered robust to outlying lineages if the minimum Pearson's correlation coefficient on any subset of samples was >0.8 when 5% of sequences were assigned to the outlying lineage. A measure was considered highly effective at identifying the underlying clustering pattern for a given model of differentiation only if it was within 10% of the top performing measure on all four empirical data sets.

set of branches can differ substantially in their properties and effectiveness at revealing patterns of clustering (Table 2). While using the shared absence of species has been criticized in classical ecology as being uninformative to the ecological similarity of sites (Legendre and Legendre, 1998), we found CT measures to perform relatively well under both models of differentiation considered (Supplementary Table S20) and we contend that with sufficiently deep sampling shared lineage absence is informative. Recently, the use of deep branches for conservation assessment has been debated (Crozier *et al.*, 2005; Faith and Baker, 2007). Our results indicate that MRCA, CL and CT measures can all efficiently recover biologically informative patterns. Interestingly, implementations of the normalized weighted UniFrac (Bray-Curtis) measure have differed in their inclusion of deep branches, for example, the Fast UniFrac web services (Lozupone *et al.*, 2006; Hamady *et al.*, 2010) calculate diversity over the CL subtree whereas mothur (Schloss *et al.*, 2009) considers the MRCA subtree by default. Restricting the Bray-Curtis measures to the MRCA subtree can have a notable influence on the dissimilarity measured between communities, highlighting the importance of specifying the set of branches a measure is calculated over (Supplementary Table 2).

The evaluated phylogenetic β -diversity measures differed in their properties and ability to reveal

clustering patterns under alternative models of differentiation (Table 2). For example, the Canberra and Gower measures easily identified clusters under the equal-perturbation model with only 1000 sequences per sample while the Morisita-Horn measure generally failed to reveal clusters even with 10000 sequences per sample. In contrast, Morisita-Horn readily identifies clusters under the dominant-pair model while the Canberra and Gower measures proved ineffective. The performance of a measure on samples that have differentiated according to a particular model can often be inferred from its properties. Five of the six most effective measures on the equal-perturbation model are either sensitive to rare OTUs (Canberra, Coefficient of similarity, Gower) or downweight the contribution of abundant OTUs (Hellinger, Chi-squared). Since these are properties of the measures themselves, it is unsurprising that our results on phylogenetic-based measures are in general agreement with those obtained for taxon-based measures under this model (Kulczynski *et al.*, 2010). These results illustrate the need to consider the performance of a measure under multiple models of differentiation, and indicate that the Canberra and Gower measures can perform poorly under some models of community variation and must be interpreted with regards to their high sensitivity to rare OTUs. These results likely apply to the taxon-based variants of the Canberra and Gower measures recently recommended

by Kuczynski *et al.* (2010) as high sensitivity to rare OTUs is an inherent property of these measures.

The variation in the performance of measures under alternative models of differentiation is the direct result of measures focusing on different aspects of phylogenetic relatedness. Measures may produce contrasting biological patterns indicating the relative importance of factors such as rare OTUs, root placement or abundance information. For example, when applied to cecal microbiota from lean and obese mice, the unweighted UniFrac (uSoergel) measure identified high similarity between the microbiota of mothers and their offspring, while the weighted UniFrac (Manhattan) measure indicated that community composition was associated with obesity genotype (Lozupone *et al.*, 2007). Our observed mean correlation between these measures was 0.83, suggesting that even a relatively high correlation between two measures does not necessarily preclude the recovery of contrasting results in parallel ordination analyses.

A number of measures were found to be highly correlated under both random sampling of empirical data sets (Figure 1; Supplementary Figure S2) and the evaluated models of differentiation (Supplementary Figure S5). This suggests that these measures will be highly correlated for many data sets. Here, we recommend specific measures based on this clustering. The blue, purple and green clusters appear to be driven by the sensitivity of measures to rare or abundant OTUs. We recommend the Gower, Soergel and Morisita-Horn measures as representative measures as their taxon-based variants are well studied and widely used (Legendre and Legendre, 1998; Magurran, 2004). The Gower measure is sensitive to rare OTUs, the Soergel measure takes a more balanced approach, and the Morisita-Horn measure places additional emphasis on highly abundant OTUs. If root-invariant measures are required, the Manhattan and Euclidean measures may be preferred to Soergel and Morisita-Horn, respectively. All measures within the orange cluster are qualitative and we recommend the uSoergel measure (that is, Jaccard index) as its taxon-based variant is well studied and widely used (Koleff *et al.* 2003; Magurran, 2004), as is its phylogenetic variant under the guise of unweighted UniFrac (Lozupone and Knight, 2005; Lozupone *et al.*, 2007). Notably, the Soergel measure is equivalent to the Jaccard index when applied to qualitative data (Pielou, 1984), motivating its use over other measures in the purple cluster. Both the root-invariant MNND and uMNND (gray cluster) often produce only weakly correlated results compared with other measures and are of interest due to their wide use in classical ecology (Webb *et al.*, 2008). The Chi-squared, MPD, uMPD and uLennon measures tended to produce rather distinct results from all other measures so may warrant consideration by studies conducting a thorough analysis of β diversity. While the above clustering

of measures was highly similar across all four empirical data sets and under the two models of variation considered it may differ for specific data sets, especially those considering alternative genes or specific lineages within a 16S phylogeny. As such, our open-source software provides functionality for identifying data set-specific subsets of measures within a given correlation threshold and inferring *de novo* hierarchical cluster trees based on the data set-specific correlation between measures.

We have explored a number of important properties of phylogenetic β -diversity measures and their performance under two models of differentiation, with a focus on the correlation between measures. Additional work is required to assess how the magnitude of dissimilarity values change under different conditions. For example, while the dissimilarity values of most measures remain highly correlated with the addition of an outlying lineage, we observed that the magnitude of dissimilarity values changed more substantially for measures sensitive to rare OTUs. Further efforts to relate the performance of measures to different mechanisms of differentiation would also be of substantial benefit. Different patterns of phylogenetic β -diversity have been associated with different types of ecological processes: for example, Horner-Devine and Bohannan (2006) found considerable evidence for phylogenetic clustering of samples, suggesting the presence of adaptive traits in restricted sets of lineages. We would also welcome efforts to model communities along environmental gradients as well as microbial-based sampling models that consider the spatial structuring of microbes (Plotkin and Muller-Landau, 2002; Green and Plotkin, 2007) in relation to sampling probability in a metagenomic study.

Acknowledgements

DHP was supported by the Killam Trusts; RGB acknowledges the support of Genome Atlantic, the Canada Foundation for Innovation, and the Canada Research Chairs program.

References

- Barr JJ, Slater FR, Fukushima T, Bond PL. (2010). Evidence for bacteriophage activity causing community and performance changes in a phosphorus-removal activated sludge. *FEMS Microbiol Ecol* **74**: 631–642.
- Bryant JA, Lamanna C, Morlon H, Kerkhoff AJ, Enquist BJ, Green JL. (2008). Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proc Natl Acad Sci USA* **105**: 11505–11511.
- Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J *et al.* (2011). Moving pictures of the human microbiome. *Genome Biol* **12**: R50.
- Clarke KR, Warwick RM. (1998). A taxonomic distinctness index and its statistical properties. *J Appl Ecol* **35**: 523–531.
- Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A *et al.* (2009). Biopython: freely available Python tools

- for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.
- Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JL, Knight R. (2009). Bacterial community variation in human body habitats across space and time. *Science* **329**: 1694–1697.
- Crozier RH, Dunnett LJ, Agapow PM. (2005). Phylogenetic biodiversity assessment based on systematic nomenclature. *Evol Bioinform Online* **1**: 11–36.
- DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM *et al.* (2006). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res* **34**: W394–W399.
- Dethlefsen L, Huse S, Sogin ML, Relman DA. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* **6**: e280.
- Faith DP, Baker AM. (2007). Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evol Bioinform Online* **17**: 121–128.
- Fierer N, Lauber CL, Zhou N, Daniel M, Costello EK, Knight R. (2010). Forensic identification using skin bacterial communities. *Proc Natl Acad Sci USA* **107**: 6477–6481.
- Graham CH, Fine PVA. (2008). Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecol Lett* **11**: 1265–1277.
- Green JL, Plotkin JB. (2007). A statistical theory for sampling species abundances. *Ecol Lett* **10**: 1037–1045.
- Hamady M, Lozupone C, Knight R. (2010). Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* **4**: 17–27.
- Horner-Devine MC, Bohannan BJM. (2006). Phylogenetic clustering and overdispersion in bacterial communities. *Ecology* **87**: S100–S108.
- Kluge AG, Farris JS. (1969). Quantitative phyletics and the evolution of anurans. *Syst Zool* **18**: 1–32.
- Koleff P, Gaston KJ, Lennon JJ. (2003). Measuring beta diversity for presence absence data. *J Anim Ecol* **72**: 367–382.
- Kuczynski J, Liu Z, Lozupone C, McDonald D, Fierer N, Knight R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods* **7**: 813–819.
- Lauber CL, Hamady M, Knight R, Fierer N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community composition at the continental scale. *Appl Environ Microbiol* **75**: 5111–5120.
- Legendre P, Legendre L. (1998). *Numerical Ecology*, 2nd English edn. Elsevier: Amsterdam, The Netherlands.
- Lozupone C, Knight R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228–8235.
- Lozupone CA, Hamady M, Kelley ST, Knight R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* **73**: 1576–1585.
- Lozupone CA, Hamady M, Knight R. (2006). UniFrac—An online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7**: 371.
- Lozupone CA, Knight R. (2007). Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104**: 11436–11440.
- Magurran AE. (2004). *Measuring Biological Diversity*. Blackwell Publishing: Oxford, UK.
- Martin AP. (2002). Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol* **68**: 3673–3682.
- Martiny JB, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL *et al.* (2006). Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**: 102–112.
- Nipperess DA, Faith DP, Barton K. (2010). Resemblance in phylogenetic diversity among ecological assemblages. *J Veg Sci* **21**: 809–820.
- Pielou EC. (1984). *The interpretation of ecological data: a primer of classification and ordination*. John Wiley and Sons. Plotkin JB, Muller-Landau HC (eds). (2002). Sampling the species composition of a landscape. *Ecology* **83**: 3344–3356.
- Plotkin JB, Muller-Landau H. (2002). Sampling the species composition of a landscape. *Ecology* **83**: 3344–3356.
- Price MN, Dehal PS, Arkin AP. (2009). FastTree: computing large minimum-evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641–1650.
- Root HT, Nelson PR. (2011). Does phylogenetic distance aid in detecting environmental gradients related to species composition? *J Veg Sci* **22**: 1143–1148.
- Rousk J, Bååth E, Brookes PC, Lauber CL, Lozupone C, Caporaso JG *et al.* (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J* **4**: 1340–1351.
- Schloss PD. (2008). Evaluating different approaches that test whether microbial communities have the same structure. *ISME J* **2**: 265–275.
- Schloss PD, Handelsman J. (2004). Status of the microbial census. *Mol Biol Rev* **68**: 686–691.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.
- Slater FR, Johnson CR, Blackall LL, Beiko RG, Bond PL. (2010). Monitoring associations between clade-level variation, overall community structure and ecosystem function in enhanced biological phosphorus removal (EBPR) systems using terminal-restriction fragment length polymorphism (T-RFLP). *Water Res* **44**: 4908–4923.
- Stackebrandt E, Goebel BM. (1994). Taxonomic Note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* **44**: 846–849.
- Swenson NG. (2011). Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. *PLOS One* **6**: e21264.
- Turnbaugh PJ, Hamady M, Yatsunencko T, Cantarel BL, Duncan A, Ley RE *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.
- Ubeda C, Taur Y, Jeng RR, Equinda MJ, Son T, Samstein M *et al.* (2010). Vancomycin-resistant *Enterococcus* domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. *J Clin Invest* **120**: 4332–4341.
- Webb CO, Ackerly DD, Kembel SW. (2008). Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics* **24**: 2098–2100.

Supplementary Information accompanies the paper on The ISME Journal website (<http://www.nature.com/ismej>)