RESEARCH ARTICLE

Understanding microbial community diversity metrics derived from metagenomes: performance evaluation using simulated data sets

Germán Bonilla-Rosso[1], Luis E. Eguiarte[1], David Romero[2], Michael Travisano[3] & Valeria Souza[1]

[1]Department of Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, México D.F., México; [2]Programa de Ingeniería Genómica, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, México D.F., México; and [3]Departement of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, MN, USA

Correspondence: Valeria Souza, Department Ecología Evolutiva, Instituto de Ecología, Apartado Postal 70-275, Universidad Nacional Autónoma de México, Ciudad Universitaria, México D.F., México.
Tel.: 52 555 622 9006;
fax: 52 555 616 1976;
e-mail: souza@servidor.unam.mx

Received 21 November 2011; revised 25 April 2012; accepted 27 April 2012.
Final version published online 25 May 2012.

DOI: 10.1111/j.1574-6941.2012.01405.x

Editor: Julian Marchesi

Keywords
community structure; diversity; dominance; metagenomics; microbial communities; simulations.

Abstract

Metagenomics holds the promise of greatly advancing the study of diversity in natural communities, but novel theoretical and methodological approaches must first be developed and adjusted for these data sets. We evaluated widely used macroecological metrics of taxonomic diversity on a simulated set of metagenomic samples, using phylogenetically meaningful protein-coding genes as ecological proxies. To our knowledge, this is the first approach of this kind to evaluate taxonomic diversity metrics derived from metagenomic data sets. We demonstrate that abundance matrices derived from protein-coding marker genes reproduce more faithfully the structure of the original community than those derived from SSU-rRNA gene. We also found that the most commonly used diversity metrics are biased estimators of community structure and differ significantly from their corresponding real parameters and that these biases are most likely caused by insufficient sampling and differences in community phylogenetic composition. Our results suggest that the ranking of samples using multidimensional metrics makes a good qualitative alternative for contrasting community structure and that these comparisons can be greatly improved with the incorporation of metrics for both community structure and phylogenetic diversity. These findings will help to achieve a standardized framework for community diversity comparisons derived from metagenomic data sets.

Introduction

In recent years, advances in the metagenomic analysis of microbial communities have been fuelled not only by decreasing sequencing costs, but also by the promise for the identification of general patterns in microbial community ecology. Metagenomics can significantly advance the study of community ecology by a simultaneous access to both functional and taxonomic diversity. It has already been applied to a wide range of environments (Rusch *et al.*, 2007), providing an unprecedented opportunity to identify ecological patterns in the structure and distribution of natural microbial communities (Kemp & Aller, 2004; Lozupone & Knight, 2007; Smith, 2007). Nonetheless, the estimation of taxonomic diversity has long proved to be a difficult task (Hurlbert, 1971; Hill, 1973; Venter *et al.*, 2004; Roesch *et al.*, 2007; Rusch *et al.*, 2007; Bent & Forney, 2008; Quince *et al.*, 2008; Shaw *et al.*, 2008; Sharpton *et al.*, 2011).

Historically, microbial community ecology has relied on SSU-rRNA genotyping as the standard approach, and many studies have estimated species richness directly from SSU-rRNA clone libraries (Roesch *et al.*, 2007; Fulthorpe *et al.*, 2008; Biers *et al.*, 2009). Although SSU-rRNA are powerful phylogenetic markers, the scattered distribution of hypervariable regions across its full length (~1500 bp) makes it very hard to recover comparable, phylogenetically informative fragments that are mutually overlapping (Mills *et al.*, 2006; Kembel

FEMS MICROBIOLOGY ECOLOGY

et al., 2011), and efforts have focused on circumventing this problem through the use of reference alignments and phylogenetic trees (Huson *et al.*, 2007; Rusch *et al.*, 2007; Berger *et al.*, 2011; Sharpton *et al.*, 2011). In addition, concerns have recently been raised against its use to study community structure because variability in gene copy number per genome can lead to biased estimations (Venter *et al.*, 2004; Biers *et al.*, 2009; Kembel *et al.*, 2011; Roux *et al.*, 2011). This is why attention has turned to the use of multiple single-copy, universally conserved protein-coding phylogenetic markers (Ciccarelli *et al.*, 2006; Wu & Eisen, 2008) as ecological proxies of community structure in metagenomic studies (Venter *et al.*, 2004; von Mering *et al.*, 2007; Rusch *et al.*, 2007; Biers *et al.*, 2009; Kembel *et al.*, 2011; Roux *et al.*, 2011).

The large number of microbial sequencing projects is stressing the need to develop new theoretical and methodological approaches to measure diversity across data sets (Rodriguez-Brito *et al.*, 2006; Huson *et al.*, 2009). While a wide range of diversity metrics have been used to compare microbial community richness (Roesch *et al.*, 2007; Schloss & Handelsman, 2008) and ranking (Hughes *et al.*, 2001; Shaw *et al.*, 2008; Youssef & Elshahed, 2009), testing their suitability to be used with microbial communities has received less consideration (Hughes *et al.*, 2001; Curtis *et al.*, 2002; Hill *et al.*, 2003; Quince *et al.*, 2008; Kuczynski *et al.*, 2010). To our knowledge, the applicability of macroecological diversity metrics has been evaluated mostly for SSU-rRNA clone libraries (Hughes *et al.*, 2001; Mills *et al.*, 2006; Bent & Forney, 2008; Shaw *et al.*, 2008; Youssef & Elshahed, 2009; Kuczynski *et al.*, 2010), and only the choice of ecological distances has been explored for metagenomic data sets (Mitra *et al.*, 2010). Furthermore, the use of mathematical models and computer simulated data sets for accurate evaluation of diversity metrics has been scarce (Curtis & Sloan, 2006; Sloan *et al.*, 2006; Green & Plotkin, 2007; Bent & Forney, 2008; Kuczynski *et al.*, 2010), even though it is not possible to test the efficiency of these metrics without knowing the real diversity in natural communities (Shaw *et al.*, 2008). To address this problem, we evaluated the applicability of widely used diversity metrics on a simulated set of metagenomic samples from nine source communities with contrasting structure and proposed a set of considerations for the qualitative comparison of the diversity in metagenomic data sets.

Materials and methods

To evaluate the applicability of macroecological diversity measures to metagenomic data sets, we chose to simulate the sequencing of nine theoretical microbial communities, estimate relative abundances from protein-coding phylogenetic markers and calculate diversity with canonical macroecological metrics. We generated other similar data sets to compare against and evaluate the effect of marker choice, taxonomic composition bias and sampling bias. A summary of the generation of matrices is presented as a flux chart in Fig. 1.

Design of source communities from completely sequenced genomes

As microbial ecology heavily relies on genomic molecular markers, the first step was to design *in silico* a set of theoretical, artificial microbial communities with contrasting diversity that will serve as the known template and starting point for the sequencing simulation. We took advantage of the availability of complete genome sequences from several microbial organisms deposited in public databases and randomly sampled them to construct these source communities (Supporting Information, Table S1). We assume that their relative abundance in the community is equal to the relative abundance of the genome in the community metagenome, so that all species included have only one genomic copy per genome and there is no polyploidy.

To better represent the multidimensional nature of diversity, each source community belonged to one of the three species richness levels (low: 10 species, medium: 100 spp., high: 500 spp.) and one of the three dominance levels. In the low-dominance level, all species had exactly the same number of individuals (a total-evenness scenario). The medium-dominance level was constructed in a way that four species equally contained half of the individuals in the community (one-eighth of the community each), for a scenario of four equally dominant species and a long tail of rare species. The high-dominance level was constructed so that only three species contained half of the individuals of the community, with one species containing one quarter of the community, and the other quarter shared by the other two species. This represents a scenario with one dominant species, two half-dominant and a long tail of rare species. A total of nine source communities were constructed as the result of the cross-product of all three richness levels and all three dominance levels (Fig. S1). The total number of individuals was kept to 1000 for all communities to standardize dominance comparisons, and the abundances were calculated as proportions of the total community, so that the dominance level was conserved across different richness levels. To avoid taxonomic biases, dominance was modified over the same taxa, in a way that community composition at the lower richness levels are a subset of the higher richness levels.

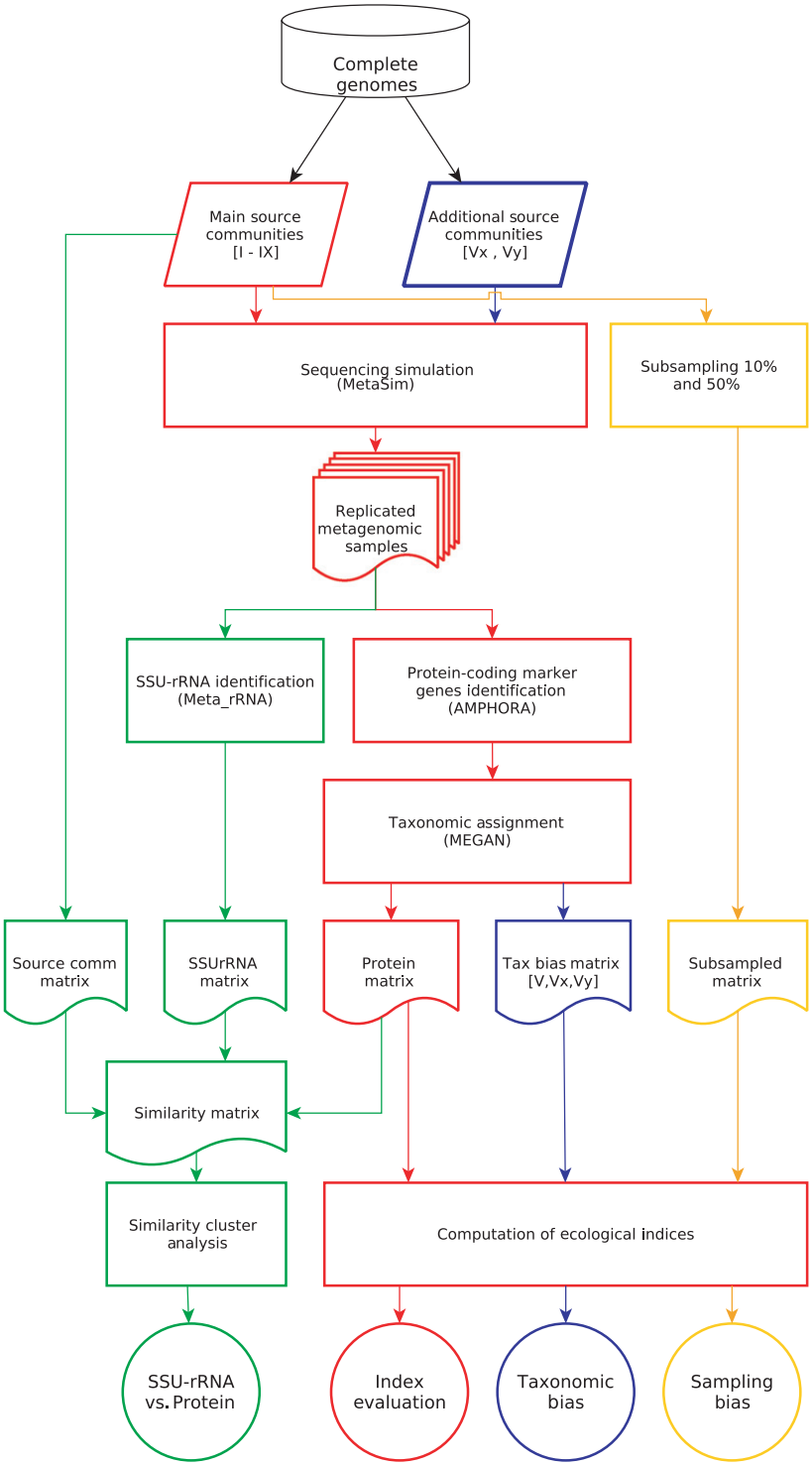

Fig. 1. A flowchart illustrating the main steps in the methodology towards the comparison of diversity metrics calculated from protein-marker matrices derived from simulated metagenomes (red), its contrast against SSU-rRNA derived matrices (-green) and the evaluation of sampling (yellow) and taxonomic (blue) biases.

Metagenomic data sets sequencing simulation

We next simulated the pyrosequencing of each source community with the sequencing simulator software MetaSim (Richter *et al.*, 2008). Briefly, MetaSim generates a set of synthetic sequencing reads (a metagenomic data set) from a species-abundance matrix and a database of the complete genomes, according to the characteristics and error models produced by different sequencing technologies (Richter *et al.*, 2008). We used our source communities as the species-abundance matrices input and simulated five pyrosequencing replicated runs for each source community (450 000 reads each, error model = 454, read length = ~250 bp, distribution mean = 0.23, distribution SD = 0.15, proportionality constant = 0.15, scale SD with square root of mean = true, error clone distribution = normal, error clone mean = 2000, 2nd parameter = 200). Each resulting simulated metagenome replica was corrected for pyrosequencing noise with CDHIT-454 (Li & Godzik, 2006), and ORFs were predicted and translated into proteins with GeneMark (Lukashin & Borodovsky, 1998).

Construction of community matrices

Each translated protein sample replica was scanned for 31 universally conserved, single-copy, protein-coding genes with AMPHORA (*dnaG*, *frr*, *infC*, *nusA*, *pgk*, *pyrG*, *rplA*, *rplB*, *rplC*, *rplD*, *rplE*, *rplF*, *rplK*, *rplL*, *rplM*, *rplN*, *rplP*, *rplS*, *rplT*, *rpmA*, *rpoB*, *rpsB*, *rpsC*, *rpsE*, *rpsI*, *rpsJ*, *rpsK*, *rpsM*, *rpsS*, *smpB*, and *tsf*, Table S2; Wu & Eisen, 2008). These genes are commonly used as taxonomic molecular markers because they are phylogenetically informative, and because they are single-copy in the genomes, we can use them as ecological proxies as an indirect measure of the relative abundance of their species of origin. Each of the identified protein-marker fragment was assigned to a taxonomic category using the last common ancestor (LCA) algorithm implemented in the metagenomic analysis software MEGAN (Huson *et al.*, 2007; Min Support = 1, Min Score = 35, Top Per cent = 3). This software employs the phylogenetic information within the top best BLAST hits of each fragment against the nonredundant protein database and the NCBI taxonomy tree to assign each fragment to a taxonomic category. Once all reads were classified, we used a parsing script to summarize the total number of reads within each taxonomic category in each metagenomic sample in the form of a community or species-abundance matrix.

Diversity metrics calculation

All the canonical macroecological diversity metrics in this work are estimated from ecological distance matrices. We used the protein-marker matrices obtained in the previous section to calculate the Hellinger transformation of ecological distances (Eqn 1), both because the Hellinger distances are more representative of real ecological distance (Legendre & Gallagher, 2001) and because their use with metagenomic data sets has already been evaluated with positive results (Mitra *et al.*, 2010).

Equation (1) Hellinger's Distance

$$D = \sqrt{\sum_{i=1}^{S}\left(\sqrt{\frac{x_i}{\hat{x}}} - \sqrt{\frac{y_i}{\hat{y}}}\right)^2}, \hat{x} = \sum_{i=1}^{S} x_i$$

where x_i = abundante of ith species at site x; y_i = abundante of ith species at site y.

These ecological distance matrices were in turn used to calculate diversity metrics commonly used in macroecology. The richness estimators used were observed richness (S), the nonparametric richness estimator *Chao1* (Chao, 1984), and abundance-based coverage estimator *ACE* (Chao & Lee, 1992). Dominance-based diversity metrics used were Simpson's probability that two randomly sampled individuals belong to the same species (D; Simpson, 1949), and Berger–Parker's proportion of the most abundant species (*BP*; Berger & Parker, 1970). Metrics that incorporate both richness and dominance used in this work are Shannon's diversity index (H; Shannon, 1948), and its derived evenness metrics J and E (Kindt & Kindt, 2008) and Fisher's alpha (α) parameter for a log-series fitted species-abundance curve (Fisher *et al.*, 1943).

All the previous metrics are only point descriptions of diversity (Hurlbert, 1971; Hill, 1973), while parametric diversity families provide a more complete, multidimensional summary of community diversity (Hill, 1973; Patil & Taillie, 1982; Ricotta, 2003). Rényi's entropy profiles (Rényi, 1961) are a generalization of Shannon's informational measure extrapolated to particular moments of the same function with a scale parameter (alpha = 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, infinity) that reflects the partition of abundance between species, constituting the best representation of a continuum of possible diversity measurements (Ricotta, 2003). In consequence, Rényi's metrics span from richness to dominance, across approximations to most of the individual metrics previously mentioned (Fig. S4). All Rényi's profiles were calculated as in Eqn (2) after Tóthmérész (1995). All ecological and statistical analyses were performed in R with packages VEGAN (Oksanen *et al.*, 2007) and BiodiversityR (Kindt & Kindt, 2008).

Equation (2) Rényi's Entropy

$$H\alpha = \frac{\ln\left(\sum_{i=1}^{S} p_i^{\alpha}\right)}{1-\alpha}$$

Where p_i = relative abundante of ith species; α = scale parameter.

Evaluation of individual diversity indices

To assess the performance of each diversity metric relative to the true diversity parameter values from their community of origin, each index was tested against their corresponding value from the source communities for significant differences. The 'real' ecological distance matrices were constructed from the raw source-community species-abundance matrices, and the 'real' diversity metrics of the original communities were calculated from these as described in the previous section. We then tested for statistically significant differences between the estimated diversity metrics (calculated from the replicated metagenomic data sets) and the real diversity values (calculated from the source communities) with a *T*-test for single samples, using the values from the replicated metagenomic data sets as observations and the values from the source communities as the population parameter (Sokal & Rohlf, 1995). Samples were then ranked according to the values of each diversity metric and compared against the ranking obtained from their respective source community (Table 1).

Choice of molecular marker as ecological proxy

To evaluate whether protein-coding genes are superior to SSU-rRNA as ecological proxies of the original community, we scanned each untranslated sample from the replicated metagenomes for SSU-rRNA gene fragments using Meta_RNA, a high-sensitivity algorithm for the detection of ribosomal metagenomic fragments using hidden Markov Models (Huang *et al.*, 2009). To date, there is no consensus on the best methodology and reference database to taxonomically classify complete SSU-rRNA genes, let alone fragmented sequences (McDonald *et al.*, 2011; Sharpton *et al.*, 2011), and their choice can profoundly affect the resulting community matrices. One of the advantages of using simulated metagenomes is that we can track each of the SSU-rRNA fragments back to their genome of origin, allowing us to reconstruct a species-abundance matrix without incorporating the selection of a classifying method as an additional confusion factor. As this results in a highly confident classification of the identified fragments, it gives the SSU-rRNA matrix in this work an advantage over the protein matrix, but we chose this comparison for the sake of simplicity. A matrix of ecological distances was constructed between all the metagenomic SSU-rRNA matrices, protein-marker matrices and the original source-communities matrices (derived from the raw, 'real' data without simulation). The similarities between samples were analysed with a hierarchical cluster analysis by complete linkage as implemented in the CLUSTER package in R (Maechler *et al.*, 2002), and the distances between the SSU-rRNA and protein-marker matrices to their corresponding source communities were tested for statistical differences with a completely randomized block design for ANOVA in R (R Development Core Team, 2006).

Taxonomic composition biases

To analyse the effect of taxonomic composition bias, two additional communities were built with the same community structure as sample 'V' but the dominant species were randomly shifted from the pool of available complete genomes to modify community composition (Table S1). This allowed us to compare three communities with exactly the same diversity but different taxonomic composition. The two resulting source communities (Vx and Vy) were subjected to the exact same procedures as the others as described above, and their diversity metrics compared against sample V.

The mean pairwise phylogenetic distance (MPD; Webb *et al.*, 2002) is a diversity measure that explicitly incorporates differences in community structure, and it was determined between all members in each community following the procedure presented in Kembel *et al.* (2011) using the R package PICANTE (Kembel *et al.*, 2010). Briefly,

Table 1. Rank ordering according to diversity indices values of source communities (S) and metagenomic samples (M)

Rank	Shannon		Simpson		Logalpha		Berger-Parker		Jevenness		Eevenness		Chao1	
1st	**S1**	**M1**	**S1**	**M1**	**S1**	**S1**	S6	M3	S7	M1	S7	M1	S1	M3
2nd	**S4**	**M4**	**S4**	**M4**	S3	S4	S3	M6	S8	M4	S8	M4	**S2**	**M2**
3rd	**S2**	**M2**	**S2**	**M2**	S2	S3	**S9**	**M9**	S4	M7	S4	M7	S3	M1
4th	S3	M5	**S5**	**M5**	S4	S2	**S5**	**M5**	S9	M8	S9	M8	**S4**	**M4**
5th	S5	M3	**S7**	**M7**	S5	S5	S2	M8	S1	M9	S1	M9	**S5**	**M5**
6th	**S6**	**M6**	**S8**	**M8**	S6	S6	S8	M2	**S5**	**M5**	**S5**	**M5**	**S6**	**M6**
7th	**S7**	**M7**	S3	M6	**S7**	**S7**	**S7**	**M7**	S6	M2	S6	M2	S7	M8
8th	**S8**	**M8**	S6	M3	**S9**	**S8**	**S4**	**M4**	S2	M6	S2	M6	S8	M7
9th	**S9**	**M9**	**S9**	**M9**	**S8**	**S9**	**S1**	**M1**	**S3**	**M3**	**S3**	**M3**	**S9**	**M9**

Ranks conserved in both cases are shown in bold.

each protein-marker gene fragment was aligned to a concatenated reference alignment and then placed onto a reference phylogeny using the evolutionary placement of short sequences implemented in RAxML v.7.2.8 (Berger *et al.*, 2011). MPD was then calculated from this phylogenetic tree. The reference phylogeny was calculated via maximum likelihood with a WAG+G model partitioned by gene families from the reference alignment provided in Kembel *et al.* (2011). Statistical differences in MPD were calculated with ANOVA. In addition and because average genome size is deeply affected by the taxonomic community composition, the effective genome size (EGS) was calculated from the protein-marker abundance matrices following the methodology in Raes *et al.* (2007).

Incomplete sampling bias

An important source of bias that is unrelated to the methodology evaluated here is incomplete sampling of the natural community. Because no complex natural community has been sampled to exhaustion (to our knowledge), the effects of these kinds of bias are of the greatest importance. To separate the bias observed because of incomplete sampling from methodological bias, two additional species-abundance matrices were constructed by randomly sampling 10% and 50% of the individuals directly from the source communities, without a sequencing simulation. These samples were processed to obtain Rényi diversity profiles in exactly the same way that has been previously described.

Results and discussion

The comparison of microbial community structure by means of metagenomic data sets relies on the estimation of diversity from abundance matrices. While this comparison is promising for testing ecological hypotheses, in practice, the construction of accurate abundance matrices from metagenomic data sets is challenging and far from being standardized. To address this problem, we designed nine source communities with contrasting structure and simulated the sequencing of five replicas from each. Next, we took the advantage of the fact that we knew the real values of the diversity metrics parameters from the source communities and evaluated the performance of its estimators by contrasting them against the estimated values from the metagenomic samples.

On the type of molecular markers as ecological proxies

The first step towards contrasting communities is the construction of the abundance matrix, and so the choice of molecular markers as ecological proxies for species abundances is fundamental. Previous studies have used SSU-rRNA gene clone libraries and metagenomic fragments (Kemp & Aller, 2004; Edwards *et al.*, 2006; Mills *et al.*, 2006; Roux *et al.*, 2011), conserved protein-marker genes (Kembel *et al.*, 2011; Roux *et al.*, 2011) and even all metagenomic reads (Edwards *et al.*, 2006) as ecological proxies to address community structure and composition. Here, we compared the performance of abundance matrices built from SSU-rRNA fragments or from protein markers recovered from the metagenome sample data sets to reflect the real structure of the source communities.

Overall, the protein-marker matrices were consistently more similar and showed smaller ecological distances (mean = 0.45) to the source communities than the SSU-rRNA matrices (mean = 0.50; Fig. 2). SSU-rRNA were directly classified by their genome of origin, and are free of other common sources of error such as misalignment, misclassification and a lower resolution for detecting taxonomic groups (Roux *et al.*, 2011). This means that even if we had error-free classification methods for SSU-rRNAs, the protein-marker gene matrices would still be more similar to the real source community structure. The exception is sample I, where the SSU-rRNA matrices were more similar to the source communities than the protein matrices. Sample I has the higher richness and evenness, and although this could indicate

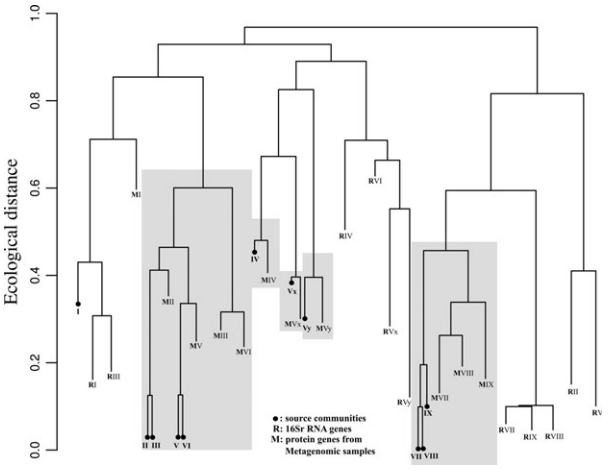

Fig. 2. Dendogram resulting from the complete cluster analysis based on Hellinger distances between source communities (black dots numbered 1–9), rRNA gene abundance matrices (R1–R9) and matrices derived from the 31 protein genes (M1–M9). Agglomerative coefficient = 0.82. The mean distance from protein gene matrices to source communities is 0.45. The mean distance from rRNA gene matrices to source communities is 0.50. This difference in distance is statistically significant (d.f. = 1, $F = 14.055$, $P < 0.01$; $\text{d.f.}_{\text{blocks}} = 10$, $F_{\text{blocks}} = 78.35$, $P_{\text{blocks}} < 0.01$). Source communities are marked with circular tips.

that SSU-rRNA matrices perform better with very complex communities, a most plausible interpretation is that the effect of classification bias on protein matrices is more strongly revealed in complex communities. Misclassification and low resolution of reference databases are prone to modify precisely the relative dominance of closely related clades, affecting samples with large numbers of species and high evenness. We would expect then that sample I would be more strongly affected if our SSU-rRNA matrices were subjected to a classification algorithm. This finding supports the choice of universally conserved, single-copy protein-coding marker genes over SSU-rRNA genes for the estimation of diversity metrics.

Evaluation of diversity metrics

With the 31 protein-marker matrices, we calculated the most commonly reported diversity metrics and Rényi's entropy profiles for each of the metagenomic samples. Because five metagenomic samples were produced by the sequencing simulation replications from each source community, we were able to directly compare the estimated values of each diversity metric (from the sample replicas) against their corresponding known community parameter (from the source community). None of the metrics estimated were statistically similar to their corresponding parameter from the source communities ($P > 0.05$; Table S3), and their results were inconsistent across samples. This means that the particular values for individual diversity metrics from metagenomic data sets differ quantitatively from the ones derived from the real, known community structure. It has been shown that some ecological problems can be approached by qualitative relative measures of diversity like ordering a set of samples according to their diversity rankings relative to each another (Shaw *et al.*, 2008). The ordering and ranking of communities according to individual diversity metrics has already been applied in microbial ecology studies using clone libraries (Hughes *et al.*, 2001; Shaw *et al.*, 2008; Youssef & Elshahed, 2009) and also metagenomes (Biers *et al.*, 2009). However, no individual diversity index recovered the same ranking from their corresponding source community, and the inconsistency of the ranking across different indices prevented us from achieving a consensus ranking (Table 1). This can be attributed to the fact that individual metrics are only point descriptors of particular aspects of diversity, and so a bias in their estimation will result in an erroneous ranking of the samples. Hence, the use of metrics that explores the multidimensional aspects of diversity (Preston, 1948; Hill, 1973) appears as a better option to compare communities. We chose Rényi's entropy profiles (Rényi, 1961) because it clearly depicts diversity graphically (Tóthmérész, 1995), but other possible alternatives are Hill's numbers (Hill, 1973), Patil and Taille's parameter families (Patil & Taillie, 1982), and even a combination of individual metrics that measure richness and different degrees of weight to dominance and richness like the *Chao1* and *BP* indices. Although also biased, the relationship between each pair of source communities Rényi's profiles (Fig. 3a) is faithfully reflected by the relationships of the metagenomic samples (Fig. 3b). Samples are difficult to rank using Rényi's profiles because one sample can be more diverse in one scale and less diverse in other (Tóthmérész, 1995) as the case of samples II and IV in Fig. 3a, but their strength relies on their ability of depicting exactly that complex relationship between the two samples, where sample II has a larger richness than IV, but it has a larger dominance than the even sample IV. An analysis of the Rényi's profiles from our samples reveals that the inconsistencies observed at the ranking with individual indices are caused by real differences in the community structure. Moreover, our results indicate that the relative positions between samples are more faithfully reflected when replicated data sets are pooled together as shown in Fig. 3d. In summary, ranking by single-diversity metrics might not be sufficient to accurately compare the diversity in two communities, and we suggest the use of multidimensional metrics to describe the rankings at different scales of diversity that might be differentially affected during manipulative studies.

Possible sources of estimation bias

There are three factors expected to cause the majority of the estimation bias observed in metagenomic data sets: DNA extraction and sequencing, choice of molecular marker selected as ecological proxy and the effect of an insufficiently sampled community. Biases in DNA extraction are beyond the scope of this work and have been addressed elsewhere (Morgan *et al.*, 2010; Lombard *et al.*, 2011), and because our metagenomic data sets were simulated *in silico,* they are free from this bias. To differentiate biases introduced by the methodology and the effect of subsampling, we constructed abundance matrices by sampling 10% and 50% of the individuals in the source communities directly, without sequencing simulation or taxonomic classification (Fig. 3c). The effect of subsampling is similar to the patterns of general reduction in diversity, and sample aggregation observed in the metagenomic data sets (Figs. 3b and 3c). A much clearer separation among samples is observed when 50% of the source community is sampled (Fig. 3e). Unfortunately, the fraction of the community present in any given sample is very hard to estimate for natural communities, and the definition of the number of sequences required

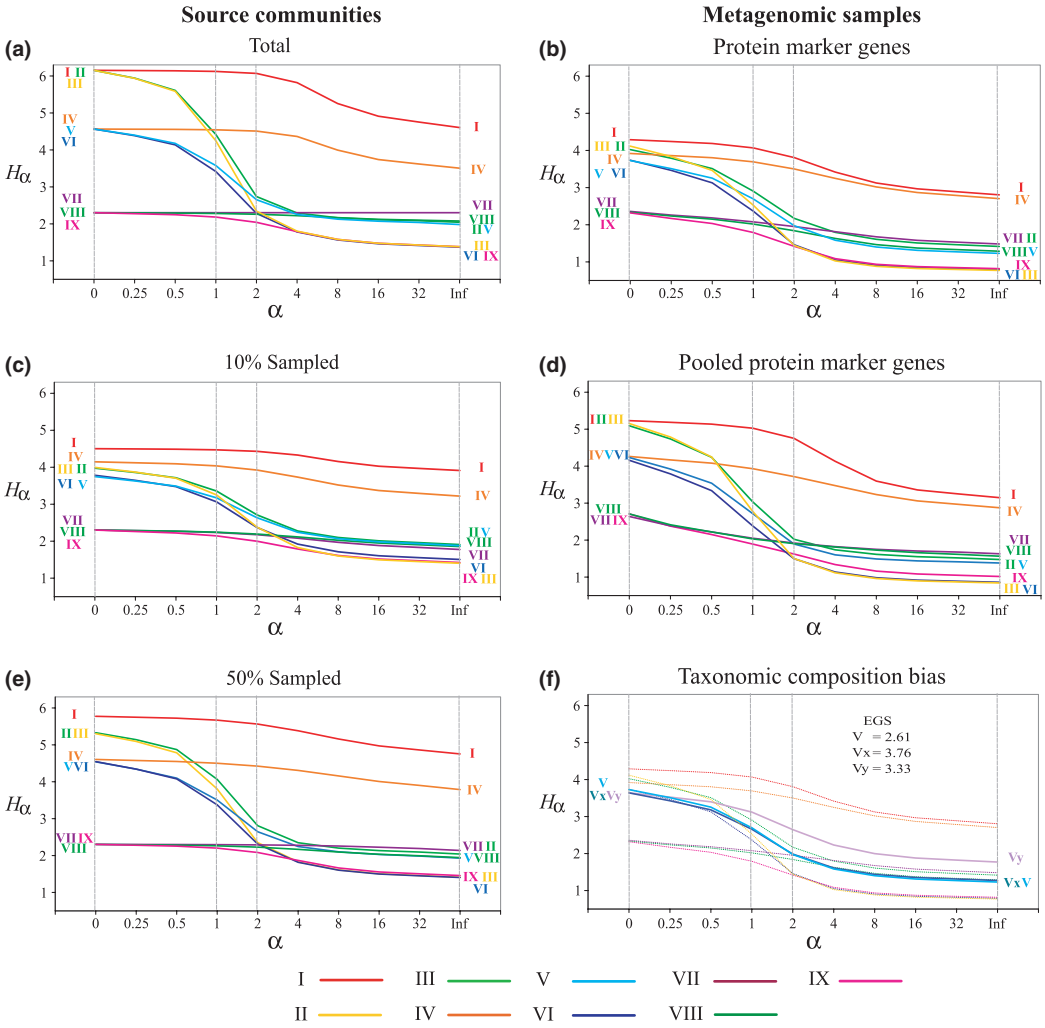

Fig. 3. Rényi's entropy profiles for (a) the source communities; (b) the matrices derived from the averaged 31 protein-marker genes; subgroups of the source communities where only (c) 10% or (e) 50% of the community was sampled; (e) the pooled replicas from the 31 protein-marker gene matrices; and (f) the protein-marker abundances of samples with different taxonomic composition. Samples Vx and Vy in (d) have exactly the same structure as V, but the species showing larger abundances are different. The rest of the samples in (d) are attenuated with dotted lines as they are given only as reference (EGS). The richness, Shannon, Simpson and Berger–Parker indices can be conceived as single moments of the entropy function, and are marked with vertical dashed lines over (α = 0, 1, 2, infinity), respectively.

to obtain a representative data set is one of the major challenges in metagenomic research. Quince *et al.* (2008) suggested that a slight increase in sequencing effort would produce a significant increase in coverage in moderately complex communities. We observed that richness categories can already be differentiated with the pooling of only two samples, each sample being roughly equivalent to the sequencing of one plate in the 454-FLX platform (Fig. S3). Moreover, samples are readily separated by their diversity profiles when the five simulated replicas are pooled together (Fig. 3d). This suggests that most of the confusing factors observed are due to subsampling, which is promising because this is expected to be less of a problem in the future with the decreasing costs of sequencing technologies.

Another potential source of bias for comparing community structure with metagenomics comes from phylogenetic community composition. This arises from the fact that the probability of sequencing any given molecular marker is a factor of the density of that marker in its genome of origin, which in turn depends of the genome size of each particular organism (Raes *et al.*, 2007). This problem is exclusive of metagenomic data sets because other approaches are usually based on the direct observations of species from individual counts. Beszteri *et al.* (2010) proposed that single-copy protein genes suf-

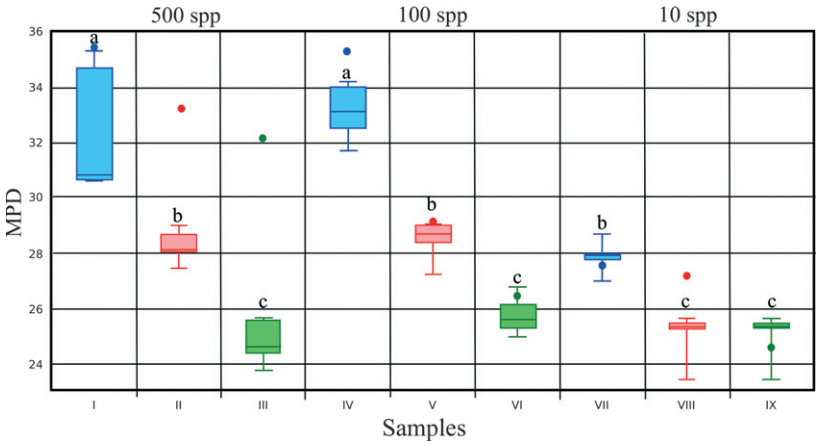

Fig. 4. Box-and-whisker plot comparing the mean pairwise phylogenetic distance values observed for the protein-marker matrices in each sample. Midlines represent the median and box limits represent the first and third quantiles, while whiskers are the maximum and minimum values and bold circles mark the corresponding source-community value. Samples are grouped by their source community richness category in bold black boxes and by their evenness by colours as follows: total evenness, blue; medium evenness, red; high dominance, green. Letters above the boxes denote membership to the statistically significant groups obtained by a *post hoc* Tukey multiple comparison with 95% of confidence.

fer from a reduced sampling probability in metagenomic data sets (as a 'dilution effect'), as they are directly affected by the mean genome size of all individuals in the community (measured as the EGS, Raes *et al.*, 2007). To address this issue, two additional source communities with identical community structure to sample V, but with different taxonomic composition (samples Vx and Vy) were built. Samples V and Vx are more similar than sample Vy, and significant differences were observed between the three samples, with evenness being more profoundly affected than richness (Fig. 3f). The observed pattern is precisely what we would expect from the dilution effect, but the EGS sample ordination does not follow the observed diversity pattern, Vy being the middle value between V and Vx (Fig. 3f). As all the dominant species in these samples belong to different bacterial phyla, this suggests that there are phylogenetic factors other than EGS affecting the estimation of community structure metrics. Our approach is not suited to address these factors, but the variable phylum representation in the reference protein databases is most likely to affect the resolution for classification and relative abundance estimation.

Because we used phylogenetically informative molecular markers as ecological proxies, it seems natural to incorporate that very same phylogenetic information into diversity metrics. Again, we measured the MDP, but other alternatives are available (Cadotte *et al.*, 2010). Differences in evenness were corroborated by variations in MPD; for instance, group *a* (I–IV) had a large MPD that was explained by a large evenness in the Rényi profile (Fig. 4), but these two samples differed in their richness. Samples VII and VIII were undifferentiated by the Rényi profile, but could be separated by the MPD and showed that although the structure was very similar in both samples, a greater clustering was observed in sample VIII. The metagenomic samples are separated by diversity metrics when they are first grouped according to their richness category and then by their MPD category, effectively reflecting the ranking of source communities by their structure (Fig. 4). Although the estimated values were statistically different from that of the source communities, the grouping of samples by MPD reflected the evenness categories from the source communities (Fig. 4). The MPD values from samples in the low richness category (i.e. samples VIII and IX) are equivalent to samples in the medium and high-dominance categories (i.e. samples III and VI), most likely because MPD is also affected by richness (Kembel *et al.*, 2010). These results suggest that measures of phylogenetic diversity can further differentiate communities by their composition and that these values naturally reflect the structure of the community and so can help differentiate samples that have not been differentiated by other multidimensional metrics that do not consider community composition.

It should also be noted that because these simulated metagenomes were constructed using known genomes, these comparisons are a best-case scenario. The diversity metrics resulted in biased estimations even under these optimal conditions, so it is reasonable to expect greater biases with real data sets where the majority of the species are only distantly related to known organisms with sequenced genomes (a case study with real metagenomic data can be found in Fig. S2). Furthermore, misclassifica-

tion errors are expected to be reduced with the advancement of classification algorithms, the availability of sequencing technologies that deliver longer sequencing reads and the phylogenetic expansion of the reference genomes, and these fields have shown significant improvements in recent days (Wu *et al.*, 2009; Ghosh *et al.*, 2010; Meinicke *et al.*, 2011; Parks *et al.*, 2011; Pati *et al.*, 2011). In the meantime, the LCA algorithm allows for reads from organisms that are phylogenetically distant from reference genomes to be only be assigned to high taxonomic ranks, so that a more accurate community structure estimation can be achieved sacrificing phylogenetic resolution. In practice, this means that more representative abundance matrices can be built from metagenomes with phylogenetically uncharacterized members if they are built at genus or family level instead of species level.

A last source of biases that was not addressed here is precisely the combined effect of uncharacterized species in a highly complex community, and we recognize that the behaviour of both diversity metrics and choice of ecological proxy might change at higher complexity. Nevertheless, these differences are difficult to address because no complex community metagenomes have been sequenced to exhaustion. Until then, these problems can be only addressed with the comparison of natural communities where contrasting levels of diversity can be presumed with confidence (Fig. S4).

Conclusion

Modern microbial ecology needs new tools to quantify microbial diversity in a statistically realistic fashion, if we expect to identify general patterns of community structure, composition and assemblage. Moreover, we need to distinguish true patterns from possible artefacts caused by the massive amounts of fragmentary data whose statistical properties are poorly understood and are possibly biased because of genetic, biological and sampling factors. Although it is natural to borrow ecological methods directly from macroecology, microbial ecologists should adjust or develop and evaluate tools and methodological practices, in a way that properly fits the biological and ecological properties of natural microbial communities.

Diversity is a complex community property, and this study illustrates the need to carefully evaluate the behaviour of the metrics used to estimate it using simulated data sets where the real community structure and composition are known. Our results showed that abundance matrices derived from protein-coding marker genes reproduce more faithfully the structure from the original community than those derived from SSU-rRNA genes, even without taking into account the alignment and misclassification biases. We found that, when calculated from metagenomic samples, the most commonly used diversity metrics are biased estimators and differ significantly from their real community parameter counterpart. Our analyses further suggest that these biases are most likely the consequence of insufficient sampling and that, as expected, this problem could be overcome by increasing sequencing coverage depth. We also found that the differences in taxonomic community composition can affect community structure estimation so phylogenetic diversity measures should be incorporated to account for this source of bias. Nevertheless, we show that correct qualitative comparisons can be achieved by the ordering and ranking of samples using a metric that contemplates the multidimensional nature of community structure diversity.

Finally, the incorporation of metrics for both community structure and phylogenetic diversity provides additional understanding of diversity in metagenomic data sets. Although the causes and alternatives to diversity metric bias are to be addressed by mathematical theory, our findings are a first attempt to achieve a standardized framework for community diversity comparisons derived from metagenomic data sets. This will support ongoing work towards the identification of general diversity patterns across geographic space and along environmental gradients.

Acknowledgements

Many thanks to L. Segovia, C. Rooks, F. Reverchon, E. Lopez-Lozano, L. Espinosa-Asuar and E. Aguirre for their suggestions on this project, as well as the comments of two anonymous reviewers that greatly improved this manuscript. Dr Blackburn and his team walked with us through the final versions of the manuscript. Financial support was received from Consejo Nacional de Ciencia y Tecnología – Secretaría de Educación Pública (grant 57507) and Secretaría de Medio Ambiente y Recursos Naturales (grant 2006-C01-23459). All research was carried out at Instituto de Ecología (UNAM), as part of GBR's PhD program at Programa de Doctorado en Ciencias Biomédicas UNAM. G.B.R. was supported with a PhD scholarship from Consejo Nacional de Ciencia y Tecnología (196814). V.S. and L.E.E. worked on this manuscript while in sabbatical at UCI (US) supported by UC-Mexus and DGPA-UNAM, respectively.

References

Bent SJ & Forney LJ (2008) The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. *ISME J* **2**: 689–695.

Berger WH & Parker FL (1970) Diversity of planktonic foraminifera in deep-sea sediments. *Science* **168**: 1345–1347.

Berger SA, Krompass D & Stamatakis A (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol* **60**: 291.

Beszteri B, Temperton B, Frickenhaus S & Giovannoni SJ (2010) Average genome size: a potential source of bias in comparative metagenomics. *ISME J* **4**: 1075–1077.

Biers EJ, Sun S & Howard EC (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol* **75**: 2221–2229.

Cadotte MW, Jonathan Davies T, Regetz J, Kembel SW, Cleland E & Oakley TH (2010) Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecol Lett* **13**: 96–105.

Chao A (1984) Nonparametric estimation of the number of classes in a population. *Scand J Stat* **11**: 265–270.

Chao A & Lee S-M (1992) Estimating the number of classes via sample coverage. *J Am Stat Assoc* **87**: 210–217.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B & Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**: 1283–1287.

Curtis TP & Sloan WT (2006) Towards the design of diversity: stochastic models for community assembly in wastewater treatment plants. *Water Sci Technol* **54**: 227.

Curtis TP, Sloan WT & Scannell JW (2002) Estimating prokaryotic diversity and its limits. *Proc Nat Acad Sci* **99**: 10494–10499.

Edwards R, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC & Rohwer F (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**: 57.

Fisher R, Corbet S & Williams C (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *J Anim Ecol* **12**: 42–58.

Fulthorpe RR, Roesch LFW, Riva A & Triplett EW (2008) Distantly sampled soils carry few species in common. *ISME J* **2**: 901–910.

Ghosh TS, M MH & Mande SS (2010) DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics* **11**: S14.

Green JL & Plotkin JB (2007) A statistical theory for sampling species abundances. *Ecol Lett* **10**: 1037–1045.

Hill M (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**: 427–432.

Hill TCJ, Walsh KA, Harris JA & Moffett BF (2003) Using ecological diversity measures with bacterial communities. *FEMS Microbiol Ecol* **43**: 1–11.

Huang Y, Gilna P & Li W (2009) Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* **25**: 1338.

Hughes JB, Hellmann JJ, Ricketts TH & Bohannan BJM (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67**: 4399–4406.

Hurlbert SH (1971) The nonconcept of species diversity: a critique and alternative parameters. *Ecology* **52**: 577–586.

Huson DH, Auch AF, Qi J & Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* **17**: 377.

Huson D, Richter D, Mitra S, Auch A & Schuster S (2009) Methods for comparative metagenomics. *BMC Bioinformatics* **10**: S12–S22.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP & Webb CO (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**: 1463.

Kembel SW, Eisen JA, Pollard KS & Green JL (2011) The phylogenetic diversity of metagenomes. *PLoS ONE* **6**: e23214.

Kemp PF & Aller JY (2004) Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. *FEMS Microbiol Ecol* **47**: 161–177.

Kindt R & Coe R (2005) *Tree diversity analysis. A manual and software for common statistical methods for ecological and biodiversity studies*. World of Agroforestry Centre (ICRAF), Nairobi.

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.

Legendre P & Gallagher E (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia* **129**: 271–280.

Li W & Godzik A (2006) Cd-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658.

Lombard N, Prestat E, van Elsas JD & Simonet P (2011) Soil-specific limitations for access and analysis of soil microbial communities by metagenomics. *FEMS Microbiol Ecol* **78**: 31–49.

Lozupone CA & Knight R (2007) Global patterns in bacterial diversity. *P Natl Acad Sci USA* **104**: 11436–11440.

Lukashin AV & Borodovsky M (1998) GeneMark. hmm: new solutions for gene finding. *Nucleic Acids Res* **26**: 1107.

Maechler M, Rousseeuw P, Struyf A, Hubert M & Hornik K (2002) *cluster: Cluster Analysis Basics and Extensions*. R package version 1.14.2.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R & Hugenholtz P (2011) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610–618.

Meinicke P, Aßhauer KP & Lingner T (2011) Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics* **27**: 1618–1624.

Mills DEK, Entry JA, Voss JD, Gillevet PM & Mathee K (2006) An assessment of the hypervariable domains of the 16S rRNA genes for their value in determining microbial community diversity: the paradox of traditional ecological indices. *FEMS Microbiol Ecol* **57**: 496–503.

Mitra S, Gilbert JA, Field D & Huson Daniel H (2010) Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. *ISME J* **4**: 1236–1242.

Morgan JL, Darling AE & Eisen JA (2010) Metagenomic sequencing of an *in vitro*-simulated microbial community. *PLoS ONE* **5**: e10209.

Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Stevens MHH & Wagner H (2007) *Vegan: community ecology package*. R package version 1.8-8.

Parks DH, MacDonald NJ & Beiko RG (2011) Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics* **12**: 328.

Pati A, Heath LS, Kyrpides NC & Ivanova N (2011) ClaMS: a classifier for metagenomic sequences. *Stand Genomic Sci* **5**: 248–253.

Patil GP & Taillie C (1982) Diversity as a concept and its measurement. *J Am Stat Assoc* **77**: 548–561.

Preston F (1948) The commonness, and rarity, of species. *Ecology* **29**: 254–283.

Quince C, Curtis TP & Sloan WT (2008) The rational exploration of microbial diversity. *ISME J* **2**: 997–1006.

R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.

Raes J, Korbel JO, Lercher MJ, von Mering C & Bork P (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol* **8**: R10.

Rényi A (1961) On measures of entropy and information. Fourth Berkeley Symposium on Mathematical Statistics and Probability, pp. 547–561.

Richter DC, Ott F, Auch AF, Schmid R & Huson DH (2008) Metasim—a sequencing simulator for genomics and metagenomics. *PLoS ONE* **3**: e3373.

Ricotta C (2003) On parametric evenness measures. *J Theor Biol* **222**: 189–197.

Rodriguez-Brito B, Rohwer F & Edwards RA (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**: 162.

Roesch LFW, Fulthorpe RR, Riva A, Casella G & Hadwin A (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283–290.

Roux S, Enault F, Bronner G & Debroas D (2011) Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems. *FEMS Microbiol Ecol* **78**: 617–628.

Rusch DB, Halpern A, Sutton G *et al.* (2007) The sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

Schloss P & Handelsman J (2008) A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics* **9**: 34.

Shannon CE (1948) A mathematical theory of communication. *AT&T TECH J* **27**: 379–423 and 623–653.

Sharpton TJ, Riesenfeld SJ, Kembel SW, Ladau J, O'Dwyer JP, Green JL, Eisen JA & Pollard KS (2011) PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Comput Biol* **7**: e1001061.

Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC & Martiny JBH (2008) It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10**: 2200–2210.

Simpson EH (1949) Measurement of diversity. *Nature* **163**: 688.

Sloan WT, Woodcock S, Lunn M, Head IM & Curtis TP (2006) Modeling taxa-abundance distributions in microbial communities using environmental sequence data. *Microb Ecol* **53**: 443–455.

Smith VH (2007) Microbial diversity–productivity relationships in aquatic ecosystems. *FEMS Microbiol Ecol* **62**: 181–186.

Sokal RR & Rohlf FJ (1995) *Biometry: The Principles of Statistics in Biological Research*. WH Freeman and Company, New York, NY.

Tóthmérész B (1995) Comparison of different methods for diversity ordering. *J Veg Sci* **6**: 283–290.

Venter JC, Remington K, Heidelberg JF *et al.* (2004) Environmental genome shotgun sequencing of the sargasso sea. *Science* **304**: 66–74.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N & Bork P (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**: 1126–1130.

Webb CO, Ackerly DD, McPeek MA & Donoghue MJ (2002) Phylogenies and community ecology. *Annu Rev Ecol Syst* **33**: 475–505.

Wu M & Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**: R151.

Wu D, Hugenholtz P, Mavromatis K *et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**: 1056–1060.

Youssef NH & Elshahed MS (2009) Diversity rankings among bacterial lineages in soil. *ISME J* **3**: 305–313.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Schematic representation of the nine source communities designed as a product of three richness categories (with 10, 100 and 500 species on the *x*-axis) and three dominance/richness categories (total evenness on the extreme right over the *y*-axis, highest dominance at the extreme left).

Fig. S2. Renyi's entropy profiles for the SSU-rRNA derived (a) and the protein-markers derived (b) matrices of three solar saltern ponds with low (squares), medium (triangles) and high (circles) salinity.
Fig. S3. Rarefaction analysis showing the increase in the expected number of species with increasing sequencing depth.
Fig. S4. Quick tutorial for the interpretation of Rényi's entropy profiles.
Table S1. Community profiles describing the structure and genomes of the source communities.
Table S2. Summary of the number of reads found for each one of the 31 protein-coding marker genes.
Table S3. Summary of the resulting *P*-values for the statistical comparisons between the observed diversity values and the values calculated from the original source communities.
Table S4. Summary of the values for all diversity metrics calculated.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.