Journal of Microbiological Methods 86 (2011) 42–51

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Rethinking microbial diversity analysis in the high throughput sequencing era

Leandro N. Lemos [a], Roberta R. Fulthorpe [b], Eric W. Triplett [c], Luiz F.W. Roesch [a,*]

[a] Universidade Federal do Pampa, Campus São Gabriel, Av. Antônio Trilha, 1847, São Gabriel, RS, Brazil
[b] Department of Physical and Environmental Sciences, University of Toronto at Scarborough, Ontario, Canada
[c] University of Florida, Department of Microbiology and Cell Science, Gainesville, FL, USA

ARTICLE INFO

Article history:
Received 15 February 2011
Received in revised form 21 March 2011
Accepted 23 March 2011
Available online 30 March 2011

Keywords:
16S rRNA genes
Taxon-based approaches
Phylogenetic approaches
Sequencing effort

ABSTRACT

The analysis of amplified and sequenced 16S rRNA genes has become the most important single approach for microbial diversity studies. The new sequencing technologies allow for sequencing thousands of reads in a single run and a cost-effective option is split into a single run across many samples. However for this type of investigation the key question that needs to be answered is how many samples can be sequenced without biasing the results due to lack of sequence representativeness? In this work we demonstrated that the level of sequencing effort used for analyzing soil microbial communities biases the results and determines the most effective type of analysis for small and large datasets. Many simulations were performed with four independent pyrosequencing-generated 16S rRNA gene libraries from different environments. The analysis performed here illustrates the lack of resolution of OTU-based approaches for datasets with low sequence coverage. This analysis should be performed with at least 90% of sequence coverage. Diversity index values increase with sample size making normalization of the number of sequences in all samples crucial. An important finding of this study was the advantage of phylogenetic approaches for examining microbial communities with low sequence coverage. However, if the environments being compared were closely related, a deeper sequencing would be necessary to detect the variation in the microbial composition.

1. Introduction

The analysis of amplified and sequenced 16S rRNA genes has become the most important single approach for microbial diversity studies since it allows hypothesis testing at various taxonomic levels. Many studies have evaluated the structure and the presence/absence of microbial communities in soil (Roesch et al., 2007; Lauber et al., 2009), hydrothermal vents (Sogin et al., 2006; Huber et al., 2007), human microbiome (Grice et al., 2009) marine environments (Schauer et al., 2010) and atmosphere (Bowers et al., 2009).

The analysis of the 16S rRNA genes for microbial ecology studies are conducted based in three ways: a) alpha or beta diversity, relating the taxon organization in a single sample or among different samples; b) tests that take into account the total number of individuals in a taxon or the presence/absence of individuals in different taxons; and c) phylogenetic distance measures (Hamady and Knight, 2009). Generally, several sequences are generated from distinct microbial communities and compared by using one or more approaches cited above. However, to obtain full representativeness of the environment each sequence must be sampled at least twice (Hughes et al., 2001). In an environment like the soil, which presents the most diverse terrestrial microbial communities, sampling intensity is critical to obtain reliable results. Although the next generation sequencing methods allow for obtaining a large number of sequences and the cost of sequencing is dropping, the technology is still expensive. Alternatively, a barcode system can be used for pyrosequencing hundreds of samples in multiplex (Hamady et al., 2008) decreasing the costs per sample. The key question is how many samples can be sequenced without biasing the results due to lack of sequence representativeness?

Analyzing an undersized number of sequences in extremely diverse environments presents potential limitations including the tests that can be performed. Here we attempted to define how much sequence is needed to analyze a microbial community. We hypothesize that if the number of sequences sampled is not representative of the environment, the increment in the sampling effort will produce different results until the number of sequences reaches a minimum necessary to generate reproducible results. This number will vary according to the test performed and the diversity of the sample.

The aims of this work were to: a) demonstrate that the level sequencing effort used for analyzing soil microbial communities biases the results; and b) determine the most effective type of analysis for small and large datasets obtained by next generation sequencing platforms.

* Corresponding author at: Universidade Federal da Pampa, Campus São Gabriel, Av. Antônio Trilha, 1847, São Gabriel, RS, Brazil. Tel./fax: +55 55 3232 6075.
E-mail address: luiz.roesch@unipampa.edu.br (L.F.W. Roesch).

0167-7012/© 2011 Elsevier B.V. Open access under the Elsevier OA license.
doi:10.1016/j.mimet.2011.03.014

Table 1
Source and description of datasets used in this work.

Origin of soil sample	16S rRNA gene region amplified	No. of sequence in the database	Average read length (bases)	Reference
Everglades, Florida	V9	235,366	242	This work
Pu'u Puai bare, Hawaii	V9	53,414	224	Giongo et al. (2010)
Mauna Ulu, Hawaii	V9	28,793	227	Giongo et al. (2010)
King George Island, Antarctica	V4	2,918	221	Teixeira et al. (2010)

2. Material and methods

In order to test the importance of sequencing depth on the interpretation of results, many simulations were performed with four independent pyrosequencing-generated 16S rRNA gene libraries obtained from different environments and two different gene regions (Table 1). The first set of sequences was amplified from DNA isolated from a soil sample collected from a sugar cane field from the Everglades Agricultural Area (EAA) site within the University of Florida's Research and Education Center in Belle Glade, Florida, USA in 2007. Soil sampling and DNA extraction methodology were described previously (Roesch et al., 2007). The second set of sequences was obtained from the 16S rRNA gene amplification products of DNA isolated from two soil samples collected at Hawaii Volcanos National Park in May 2008 (Giongo et al., 2010). The third set of sequences was obtained from a rhizosphere soil from Antarctic vascular plants of Admiralty Bay (Teixeira et al., 2010).

2.1. Processing of the 16S rRNA gene sequences and microbial communities simulations

The dataset from Everglades Agricultural Area, Florida was pre-processed to remove short sequences and trim those sequences that contain bases with low quality scores (Phred quality score smaller than 20). To perform this task we used a script called Trim2 revised and implemented in the Pipeline for Analysis of Next Generation Amplicons called PANGEA (Giongo et al., 2010). This script was also used to remove all unnecessary information from the dataset, such as sequence length and ranking as well as data about the pyrosequencing plate. The sequences that do not present the 18 primer bases (5′-GNTACCTTGTTACGACTT-3′) were also removed from the dataset.

The datasets from Hawaii (Pu'u Puai and Mauna Ulu) and from Antarctic rhizosphere soil were already pre-processed by Giongo et al. (2010) and didn't need any treatment.

Each simulation sequences from the original dataset were randomly selected using a perl script called random_seq_sample.pl downloaded from the Canadian Bioinformatics Help Desk Software Repository (http://gchelpdesk.ualberta.ca/repository/contents.php). This script randomly selects sequences from multiple Fasta formatted file without replacement. To test the effect of different sample sizes from the same environment we randomly sampled six communities with 100, 500, 1000, 5000, 10,000 and 20,000 sequences from three different datasets: (Pu'u Puai and Mauna Ulu form Hawaii and Everglades form Florida). To test the effect of a small sampling intensity from the same environment we randomly generate four microbial communities with 500 sequences each for each of the four datasets obtained.

2.2. Bioinformatics analysis

The 16S rRNA gene collection was used in microbial community simulations for two independent experiments. In the first experiment, the analysis of the microbial communities is simulated within the same environment using different sequencing intensities (100, 500, 1000, 5000, 10,000 and 20,000 sequences). In the second experiment, the analysis of four microbial communities is simulated from the same environment with the same number of sequences but with small sequencing intensity (500 sequences each). The tests performed were divided in taxon-based approaches and hypothesis testing approaches (Table 2).

For the taxon-based approaches, the sequences obtained by each simulation were clustered into Operational Taxonomic Units (OTU) based on the relatedness of the sequences using the tools implemented in Mothur (Schloss et al., 2009). To perform this task, each file was first aligned using the Needleman–Wunsch algorithm. The sequences were aligned against a template-aligned database from greengenes (greengenes.lbl.gov) with 7682 positions and 4938 bacterial and archaeal sequences in it. The program was run with the default parameters with the exception to the kmers parameter, which was set to use 9mers. Based on the alignment, a pairwise distance matrix in phylip format was generated. Based on the genetic distance, the sequences were assigned into OTUs using the nearest neighbor

Table 2
Taxon based approaches and hypothesis testing approach commonly used in microbial ecology studies.

Approaches	Description	Reference
Taxon based approaches		
Shannon index (H')	Measures the average degree of uncertainty in predicting as to what species an individual chosen at random from a collection of S species and N individuals will belong. The value increases as the number of species increases and as the distribution of individuals among the species becomes even.	Ludwig and Reynolds (1988)
Simpson's index (D)	Indicates species dominance and reflects the probability of two individuals that belong to the same species being randomly chosen. It varies from 0 to 1 and the index increases as the diversity decreases.	Simpson (1949)
Chao1 richness estimator	Non-parametric estimator that calculates the minimal number of OTUs present in a sample.	Chao(1984)
Coverage (C)	Measures how well an environment was sampled and indicates the percentage of individuals sampled in a microbial community.	Good (1953)
Shared OTUs	Identify sequences that are either unique or shared by specific groups at a specific OTU designation.	Schloss and Handelsman (2006)
Hypothesis testing approach		
Principal Coordinates Analysis (PCoA)	Multivariate statistical technique that can be used for the overall comparison for significant differences among the bacterial communities by finding clusters of samples that will reflect the similarity of the biological communities.	Krzanowski (2000)

algorithm. Since there are no accepted dissimilarity cutoffs for the different microbial taxonomic levels we used the clustering threshold proposed by Kunin et al. (2010) of 0.03 and 0.20 dissimilarity. According to the authors, diversity estimates is grossly overestimated when clustering threshold are higher than 97% identity. The OTU profiles were then used for the calculations of diversity indexes and richness (Shannon, Simpson and Chao1), sampling intensity (coverage) and shared OTUs.

For the hypothesis testing approach, the sequences were first grouped from each library into operational taxonomic units (OTUs) (with a cutoff value for assigning a sequence to the same group equal to or greater than 0.03 or 0.20 dissimilarity) using the program CD-HIT — Cluster Database at High Identity with Tolerance (Li and Godzik, 2006). Representative sequences (the longest sequence of the cluster) were chosen and merged in a single file. This file was used as input for MUSCLE (Edgar, 2004), which built a guide tree using UPGMA (unweighted pair group method with arithmetic mean) agglomerative clustering method. Quantitative and qualitative PCoA were done using UniFrac (Lozupone et al., 2006). UniFrac analysis required a phylogenetic tree prepared using MUSCLE and the number of sequences found on each OTU in each sample. By using UniFrac, PCoA was performed to find clusters of similar groups of samples. A matrix using the UniFrac metric for each pair of environments is calculated. The distances are converted to points in space with the number of dimensions one less than the number of samples. The dimensions were used to plot bi-dimensional graphs also called cluster diagrams. The cluster diagrams are useful for showing which environments are most closely related to one another. In order to support the clusters observed in the PCoA analysis we applied Jeckknife environmental clusters that sample a smaller number of sequences from each environment and tell whether the clusters are well supported. For each analysis 1000 permutations were performed. The results are presented in the Supplementary Figs. 1 to 10.

3. Results

3.1. Taxon based approaches: how much sampling is necessary to work with OTUs?

The first step to decide the better approach for analyzing a dataset of sequences is to measure the representativeness of this dataset. To analyze how well each simulation was representative of the bacterial community in the environment, sequence coverage was calculated (Tables 3 and 4). The analyzed sequence coverage representing the percentage of individuals sampled in the microbial community is presented in Tables 3 and 4. As expected, the low stringency in the grouping criteria was reflected in greater coverage. Even with a low number of sequences it was possible to achieve more than 90% coverage when the grouping criteria was 20% dissimilarity. However, irrespective of the soil tested, more than 10,000 sequences were necessary to obtain at least 80% coverage when the grouping criteria among sequences was 3% dissimilarity.

Table 3
Sequence coverage calculated for the re-sampling simulations with increasing sequencing intensity.

Number of sequences	Coverage (%)		
	Everglades, Florida	Mauna Ulu, Hawaii	Pu'u Puai, Hawaii
0.03 dissimilarity			
100	52	57	53
500	54	76	70
1000	61	79	76
5000	77	88	86
10,000	82	90	88
20,000	87	92	91
0.20 dissimilarity			
100	91	96	98
500	96	99	97
1000	97	99	98
5000	98	99	99
10,000	99	99	99
20,000	99	99	99

Table 4
Sequence coverage calculated for the re-sampling simulations with the same sequencing intensity.

Number of sequences	Coverage (%)			
	Everglades, Florida	Mauna Ulu, Hawaii	Pu'u Puai, Hawaii	King George Island, Antarctic
0.03 dissimilarity				
500 (A)	51	81	73	47
500 (B)	55	82	73	47
500 (C)	59	78	71	42
500 (D)	56	81	76	45
0.20 dissimilarity				
500 (A)	95	98	97	91
500 (B)	95	99	98	89
500 (C)	96	99	97	90
500 (D)	96	98	98	90

A largely used approach to compare two or more communities at a particular taxonomic unit is the definition of shared OTUs. This technique is useful to identify the members of the communities that are either unique or shared. In order to test the power of this approach, the overall overlap of taxonomic units between four microbial communities was obtained by re-sampling simulations with the same sequencing intensity (500 sequences and 20,000 sequences) in each dataset tested (Fig. 1). In the experiment with 500 sequences with 3% as the cutoff distance for grouping the sequences, a very small number of sequences shared within the same dataset was observed. For the Everglades dataset, the total richness of all the groups was 889 but only 45 OTUs (5%) were found to be in common with the four simulations comprising 500 sequences sampled from the dataset. For the dataset from Antarctic soil only 15% of the sequences were shared among the four sub-sets of 500 sequences (the total richness of all groups was 390). Moreover, only 9.3% and 12.4% of the sequences from Mauna Ulu and Pu'u Puai bare, Hawaii, respectively, were shared among the four sub-sets. The total richness for Mauna Ulu and Pu'u Puai bare was 214 and 549, respectively. Considering that the four communities were represented by 500 sequences belonging to the same dataset, a larger number of sequences in common was expected.

In order to test the same approach with datasets with high sequence coverage, the analysis was performed by re-sampling subsets of 20,000 sequences (Fig. 1B). The total richness of all the groups increased in all datasets (total richness of Everglades, Mauna and Pu'u Puai was 48%, 59% and 47% respectively) but the percentage of shared species was still below the expected for subsets of the same dataset. The results illustrated that evaluating a dataset with poor coverage by using an approach based on shared OTUs may not be biologically meaningful since randomly sampling sequences from the same dataset can generate subsets with different community membership. In order to apply such an approach, a large sampling intensity (coverage ≥ 90%) is needed to get reliable results. For most of our dataset tested this coverage was achieved with 20,000 sequences.

Diversity indexes are also used by many researchers to compare and further characterize the differences between the communities. Here we performed many simulations to test the resolution of the diversity indexes (Shannon and Simpson) and the richness estimation (Chao1) with low and high number of sequences. For the experiment with increasing sequencing intensity (Fig. 2), the diversity indices

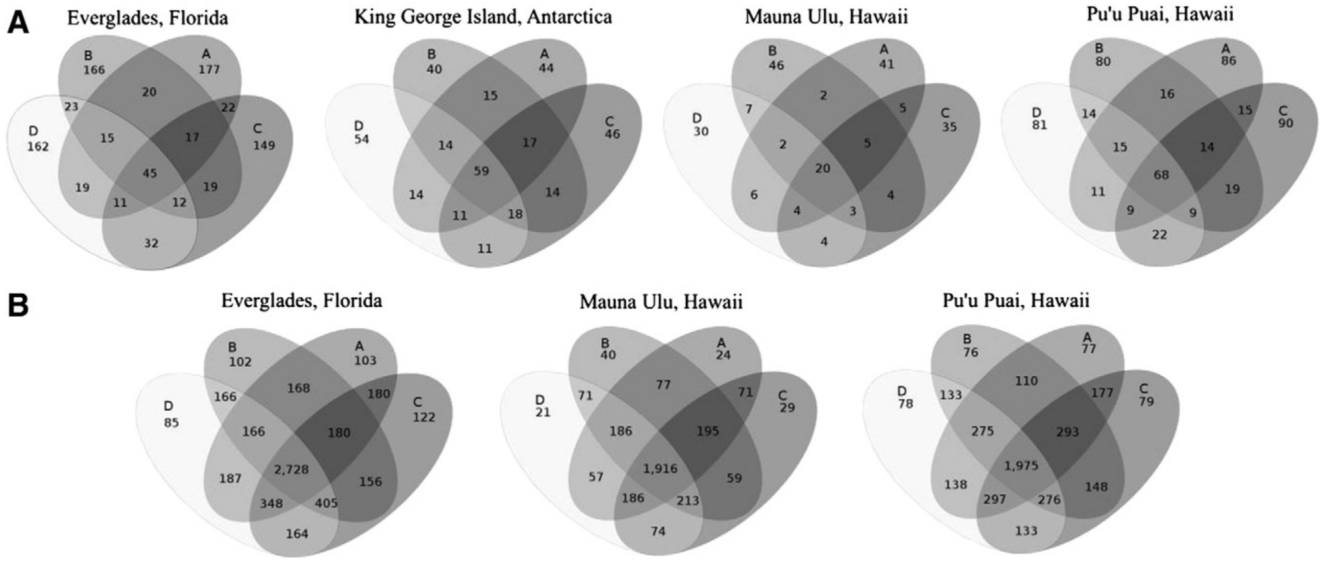

Fig. 1. *Venn diagrams* showing overall overlap of taxonomic units between four microbial communities obtained by re-sampling simulations with the same sequencing intensity. A) 500 sequences in each of the four different datasets tested. B) 20,000 sequences in each of the four different datasets tested. The criterion used for grouping the sequences was 3% dissimilarity among them.

increased as the number of sequences increased for each analysis. When applying these indices in the experiment with the same sequencing intensity, no significant change in diversity was observed at any level of dissimilarity (Fig. 3, Table 4). To confirm these results, another two re-sampling experiments were performed with the dataset from the Everglades site, generating four files with 10,000 sequences each and four files with 20,000 each. Shannon's and Simpson's diversity indices were determined (Supplementary Table S1 and S2). The experiment confirms the results observed with the data presented in Fig. 3. Even with a low coverage, the results obtained in this experiment show that Shannon's diversity indexes are suitable for comparing environments when the number of sequences is normalized. However, the Chao1 richness estimator presented a large variation when applied for the same number of sequences. For the level of 3% dissimilarity in the experiment with increasing sequencing intensity, the estimator varied from 1215 to 882 OTUs representing a difference of about 28% (Fig. 2). Even when the same number of sequences was sampled many times from the same environment (Fig. 3), the Chao1 richness estimator does not provide constant values meaning that it is not a suitable estimator for poor sequencing effort experiments.

3.2. Phylogeny approaches: when should it be used?

In contrast to OTU based approaches, which are based on estimating the number and proportion of individuals according to a predefined cutoff value for grouping sequences, the hypothesis testing approach used here was based on phylogenetic distance measures. To detect broad trends of similarities and differences in the simulations Principal Coordinates Analysis (PCoA) and Jackknife clusters analysis were used for the following simulations: a) randomly sampled microbial communities with increasing sequencing intensity using three different datasets; b) randomly sampled microbial communities with the same sampling intensity and low coverage (500 sequences each) using four different datasets; and c) randomly sampled microbial communities with increasing sequencing intensity comparing the datasets mentioned above to each other.

For microbial ecology purposes, the PCoA can be used to find clusters of samples that represent similar bacterial communities. Considering that our comparisons are made by taking sequences from the same dataset, one should expect to find a single cluster if the number of sequences are representative of the overall community. Considering the presence/absence of a particular OTU (unweighted UniFrac), the PCoA analysis showed that only those simulations with more than 10,000 sequences grouped together irrespective of the soil tested and the region of the 16S rRNA gene amplified (Fig. 4A). However, when the abundance of the OTUs was taken into account (weighted UniFrac), most of the communities clustered together (except the community made up with 100 sequences which did not cluster with the other samples) (Fig. 4B). These results showed that the most abundant sequences have a strong influence in the grouping analysis.

The PCoA analysis was also calculated with randomly sampled microbial communities with the same sampling intensity and low coverage (500 sequences each, Fig. 5). No clusters were observed in either qualitative or quantitative analysis in any of the datasets tested. These results support the idea that when the coverage is below 80% (see Table 1), the PCoA analysis does not perform well with very similar samples and a deeper sequencing would be necessary for obtaining a representative analysis.

One limitation of our simulations is that the sub-sets of sequences from the same dataset are supposed to be very similar. But the results obtained encourage us to test how low and high coverage can interfere in the PCoA analysis with very dissimilar types of soil samples. Hence, the question is which level of coverage is needed in order to discriminate microbial groups sampled from different environments? To address this question, three PCoA simulations were performed using sub-sets of sequences with 100, 500 and 20,000 sequences comparing the datasets from Everglades, Mauna Ulu, and Pu'u Puai (Fig. 6). The sets of 100, 500 and 20,000 sequences produced exactly the same clusters when considering only the presence/absence of taxonomic units. Nevertheless, the quantitative analysis showed a tendency of better discrimination, mainly between the Hawaiian samples, when a greater coverage was used to perform the analysis. This simulation suggests that deep sequencing was superfluous when comparing very dissimilar soil samples like those from an agricultural soil (Everglades, Florida) and unvegetated soils (Hawaii

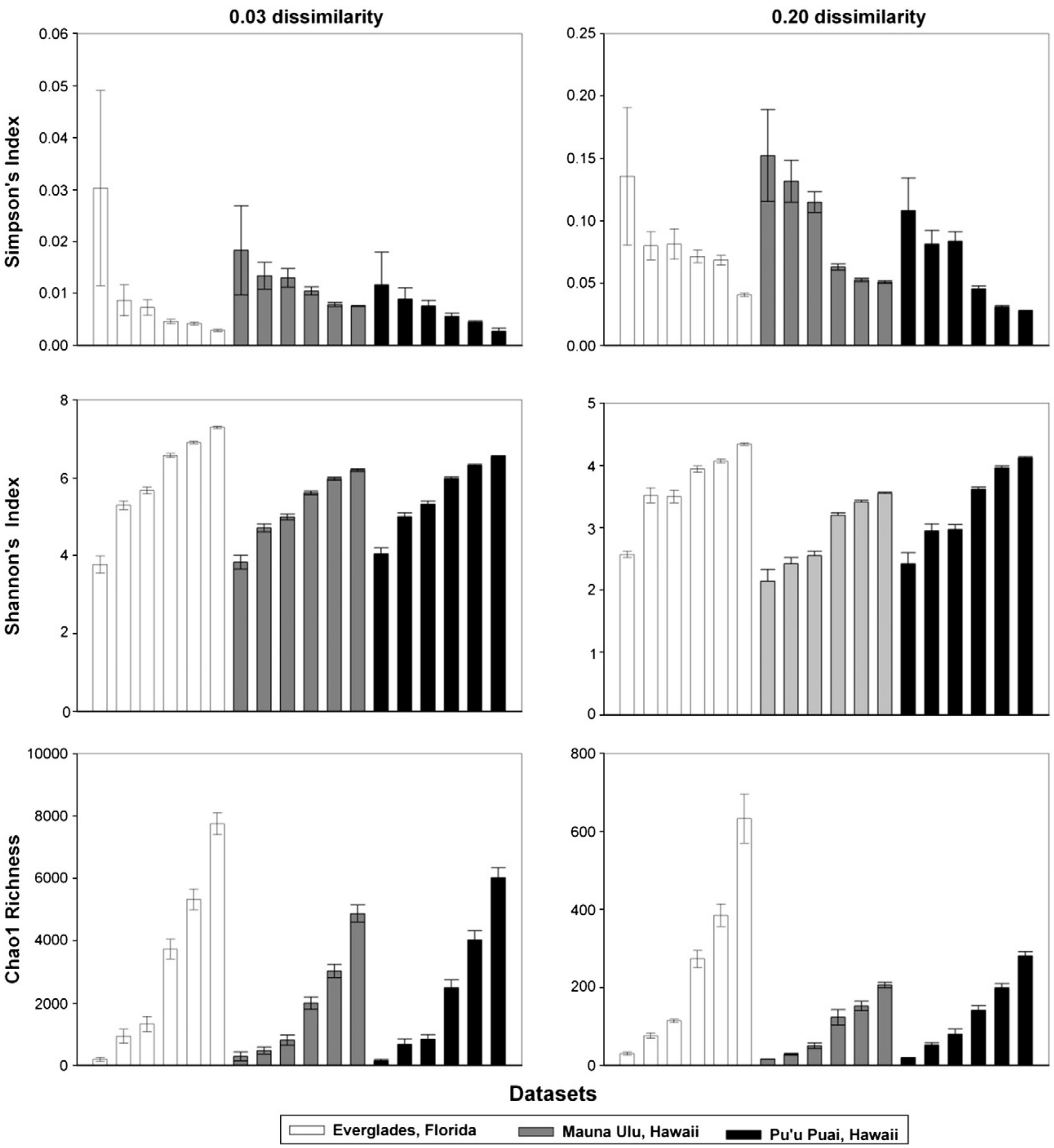

Fig. 2. Calculations of diversity indexes (Shannon and Simpson) and the richness estimation (Chao1) in each of the three different datasets tested at 0.03 and 0.20 dissimilarity as the cutoff distances. Each column represents a randomly sampled microbial community with 100, 500, 1000, 5000, 10,000 and 20,000 sequences (from the left to the right column respectively). The standard error about the mean is depicted in the error bar about the data columns.

Volcanoes National Park) collected across a broad continental scale. The Jackknife environment cluster analysis was also performed (Supplementary Figs. 1 to 10) and the results of both PCoA and Jackknife converge.

4. Discussion

One of the greatest problems concerning the study of microbial diversity based on sequencing approaches is sampling size. The major concern is about the reliability of the results when the number of sequences in the dataset is small. Are thousands of sequences needed to characterize microbial diversity? In this work, phylogenetic and taxon-based approaches were used to detect relevant bacterial patterns using sets of 16 rRNA sequences with small and high sequence coverage followed by analysis of those datasets using UniFrac and other tools. These simulations were based on data from soil samples, not simulated bacterial 16S rRNA gene datasets. The community analyses performed here illustrates that the methods of examining community structure should be chosen on the basis of sequence coverage and the origin of the environment analyzed.

To describe the similarity in membership and structure of the bacterial communities, sequences were identified that were either unique or shared by specific groups. The fraction of shared species was represented in Venn diagrams. This approach was developed in order to determine how many sequences are necessary to work with OTU-based approaches. To analyze the effect of sample size, this analysis

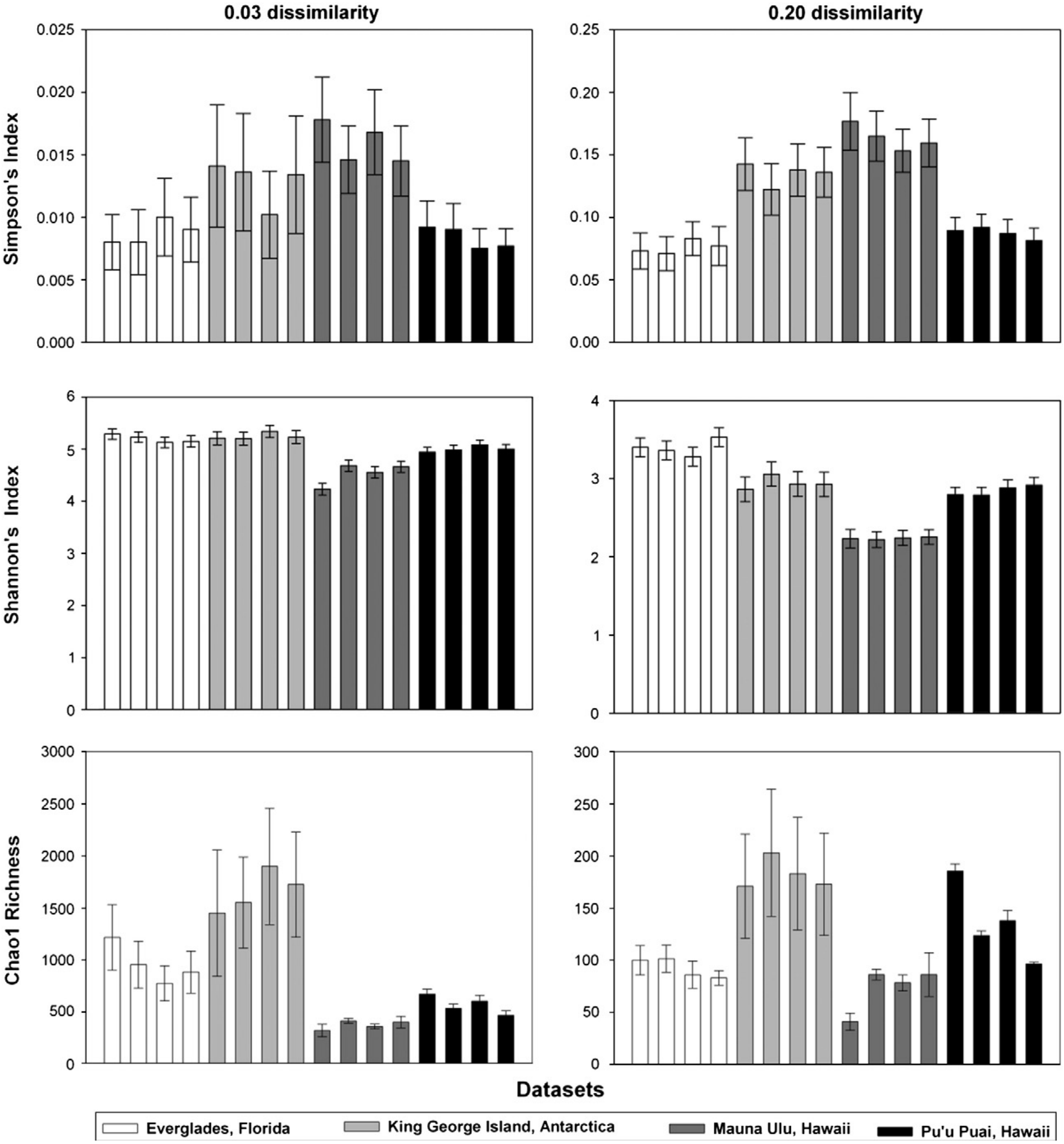

Fig. 3. Calculations of diversity indexes (Shannon and Simpson) and the richness estimation (Chao1) in each of the four different datasets tested at 0.03 and 0.20 dissimilarity as the cutoff distances. Each column represents a randomly generated microbial community with 500 sequences. The standard error about the mean is depicted in the error bar about the data columns.

was performed with low coverage and high coverage datasets. Considering that a coverage ranging from 51 to about 80% (500 sequences) would be representative of the community evaluated, the shared species among the four subsets should be relatively high. Taking into account that a very low fraction of sequences were shared among the datasets tested, this analysis would be reasonable only when the sequencing coverage were enough to detect most of the OTUs present (90% or more). The power for detecting overlapping species from multiple environments is strongly related to the sequencing intensity demonstrating the lack of OTU resolution based approaches for datasets with low sequence coverage. In general, bacterial species abundance exhibits lognormal distributions with very long tails (Martiny et al., 2006). In other words, a large percentage of species are present in extremely low numbers which challenges the analysis based in shared OTUs. Working with about 90% coverage improved the power of the analysis of shared OTUs.

Another observation highlighted by our results is associated with the uneven number of sequences among several samples. This is a common problem when dealing with multiple samples pooled in a single pyrosequencing run because of normalization issues. This is particularly important when dealing with diversity indices because communities represented by fewer sequences will appear artificially different. The problem associated with uneven sampling was noticed many years ago by Patil and Taillie (1982) but has been indiscriminately ignored by many researchers. In this work we applied Shannon and Simpson diversity indexes to compare bacterial diversity among subsamples with increasing number of sequences from the same dataset. The results obtained showed that irrespective of the level of

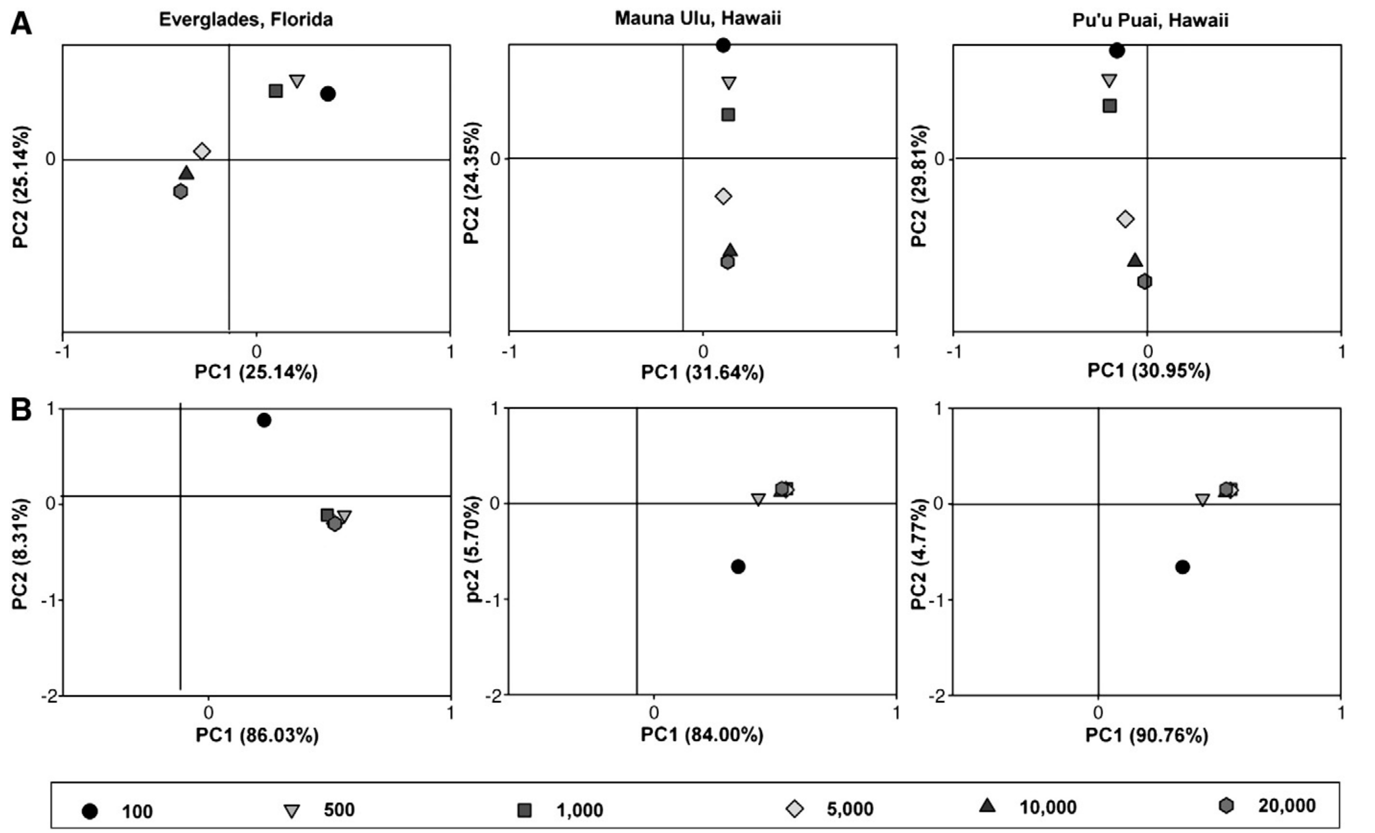

Fig. 4. Principal Coordinates analysis (PCoA) depicting the qualitative (A) (unweighted UniFrac) and quantitative (B) (weighted UniFrac) comparison within each of the three datasets tested with randomly sampled microbial community with 100, 500, 1000, 5000, 10000 and 20000.

similarity used, the diversity indices were always tied with sample size but when the number of sequences was normalized it was possible to apply those diversity indices in order to compare two or more environments. According to Chao et al. (2005), the classic Jaccard and Sørenson indices are also notoriously sensitive to sample size, especially for assemblages with numerous rare species. This happens because shared rare species are falsely scored as unique as they are detected in one sample but not in the other. Diversity index values increase with sample size making normalization of the number of sequences in all samples crucial. Despite these observations, well known and widely applied pipelines like RDP Pipeline (Cole et al., 2009) manipulate files containing the original number of sequences without taking into consideration the disparity between the number of sequences between samples. However, recently Giongo et al. (2010) normalized the datasets so that each sample contained the same number of reads prior to the Shannon diversity index calculation in PANGEA.

In order to estimate sample richness from the datasets the Chao1 richness estimator was used (Chao, 1984). Similar to the diversity indices, a strong correlation of this index with sampling size was observed. The effect of normalization of the datasets was also tested. Using subsamples of 500 sequences from the same dataset the Chao1 richness estimator presented large variation between values for each subsample. According to Colwell and Coddington (1994), Chao1 underestimates true richness at low sample sizes. Hughes et al. (2001) described this trend through the equation where the maximum value of Chao1 is $(S^2_{\mathrm{obs}}+1)/2$ (S_{obs} is the number of observed species) where one species in the sample is a doubleton and all others are singletons. Thus, Chao1 will strongly correlate with sample size until S_{obs} reaches at least the square root of twice the total richness. Clearly Chao1 would not be suitable for comparing environment richness among samples with both, low sequencing intensity and different sequencing intensity among environments. This agrees with Bent and Forney (2008) who state that the assessment of richness in complex communities is futile without extensive sampling.

An important finding of this study was the advantage of phylogenetic beta diversity approaches for examining the soil microbial communities with low sequence coverage. Here PCoA analysis was applied for quantitative and qualitative comparison of our datasets by using UniFrac (Lozupone and Knight, 2005). UniFrac compares the microbial communities overall for significant differences. Particularly, the PCoA can be applied for finding the most important axes along which the samples vary but it also can be used to find clusters of environments. Quantitative UniFrac accounts for changes in relative abundance of lineages between different communities, by weighting the branches in the phylogenetic tree-based when performing the calculations (Lozupone and Knight, 2008). It detects the changes in the number of sequences as well as the presence/absence of taxonomic units present so it is less dependent of the rare species. On the other hand, qualitative UniFrac does not account for changes in relative abundance in that case, duplicate sequences contribute no additional branch length to the tree (Lozupone and Knight, 2008). The number of sequences required to discriminate distinct environments (for example, soil samples from Florida and Hawaii) was relatively low. The same pattern could be obtained with 100 or 20,000 sequences. By using phylogenetic approaches, as few as 100 sequences per sample were also sufficient to detect variation among microbial communities in the guts of mammals (Ley et al., 2006) and between individuals with inflammatory bowel disease and healthy individuals (Frank et al., 2007). Patterns of variation between environment types, such as saline and nonsaline environments were also detected by using PCoA when comparing 202 samples ranging from 17 to 1000 sequences (Lozupone and Knight, 2007). Those are examples of comparison

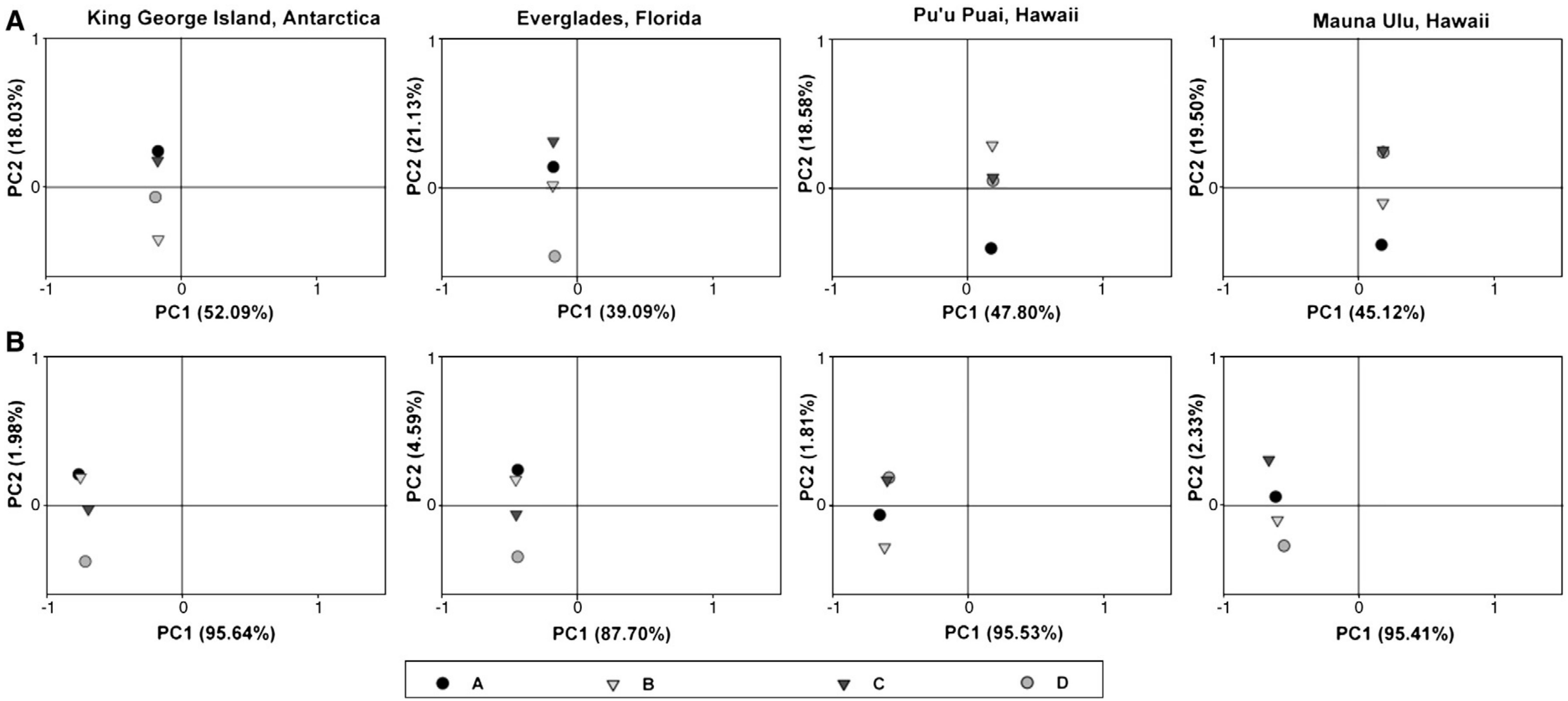

Fig. 5. Principal Coordinates analysis (PCoA) depicting the qualitative (A) (unweighted UniFrac) and quantitative (B) (weighted UniFrac) comparison of the bacterial communities obtained by re-sampling simulations with the low sequencing intensity (500 sequences) within the four datasets tested.

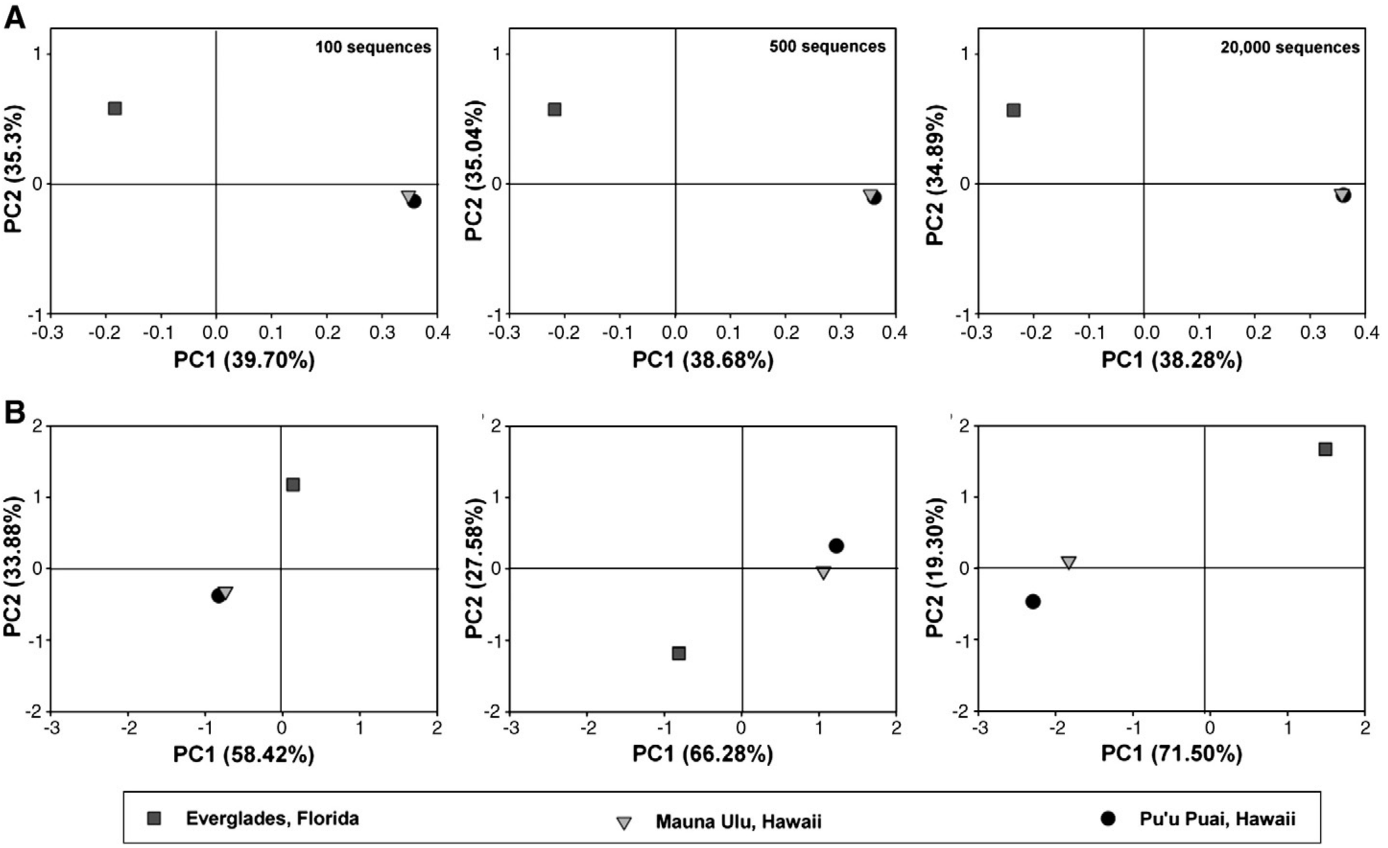

Fig. 6. Principal Coordinates analysis (PCoA) depicting the qualitative (A) (unweighted UniFrac) and quantitative (B) (weighted UniFrac) comparison for the three datasets tested with randomly sampled microbial community with 100, 500, and 20,000 sequences.

between samples collected from a large variety of environments. However, if the environments being compared were closely related (sub-sets of the same dataset as tested here), a deeper sequencing would be necessary to detect the variation in the microbial composition. At least 500 sequences were needed to detect microbial patterns in the quantitative PCoA and at least 10,000 when quantitative PCoA was applied.

For exploring ecological microbial patterns the methods based on phylogeny are useful to explore similarities and differences based on a phylogenetic tree (Hamady et al., 2010). OTU-based approaches need a rigid OTU definition based on a cutoff distance. Thus, phylogenetically parental sequences can be grouped differently than those based on OTU identification.

5. Conclusions

Phylogenetically-based approaches were the most useful approach for examining the soil microbial communities with either low or high sequence coverage. Taxon-based approaches are useful to detect shifts in specific OTUs but might need a deeper sequencing effort since many unique OTUs found may be due to undersampling of the datasets. It is possible to compare environments according to their microbial diversity even with a low sequencing effort but depending on the approach used, the sequencing effort required must be judged carefully. If the goal is to discriminate very different environments, such as aquatic versus terrestrial, as few as 100 sequences will suffice. But if the goal is to compare closely related communities thousands of reads might be necessary.

Supplementary materials related to this article can be found online at doi:10.1016/j.mimet.2011.03.014.

Acknowledgments

This work was supported by the Fundação de Amparo a Pesquisa do Estado do Rio Grande do Sul — FAPERGS (grant number 0901855) and the INCT-APA (CNPq process no. 574018/2008-5, FAPERJ E-26/170.023/2008), and supported by the Ministry of Science and Technology, and the secretariat for the Marine Resources Interministerial Committee (SECIRM). LFW Roesch and LN Lemos receive research fellowships from the CNPq (process number 503370/2009-6).

References

Bent, S.J., Forney, L.J., 2008. The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. The ISME Journal 2, 689–695.

Bowers, R.M., Lauber, C.L., Wiedinmyer, C., et al., 2009. Characterization of airborne microbial communities at a high-elevation site and their potential to act as atmospheric ice nuclei. Applied and Environmental Microbiology 75, 5121–5130.

Chao, A., 1984. Non-parametric estimation of the number of classes in a population. Scand J Stat 11, 265–270.

Chao, A., Chazdon, R.L., Colwell, R.K., Shen, T.J., 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecology Letters 8, 148–159.

Cole, J.R., Wang, Q., Cardenas, E., et al., 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research 37, D141–D145.

Colwell, R.K., Coddington, J.A., 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 345, 101–118.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 1–19.

Frank, D.N., Amand, A.L.S., Feldman, R.A., Boedeker, E.C., Harpaz, N., Pace, N.R., 2007. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proceedings of the National Academy of Sciences of the United States of America 104, 13780–13785.

Giongo, A., Crabb, D.B., Davis-Richardson, A.G., et al., 2010. PANGEA: pipeline for analysis of next generation amplicons. The ISME Journal 4, 852–861.

Good, I.J., 1953. The population frequencies of species and the estimation of the population parameters. Biometrika 40, 237–264.

Grice, E.A., Kong, H.H., Conlan, S., et al., 2009. Topographical and temporal diversity of the human skin microbiome. Science 324, 1190–1192.

Hamady, M., Knight, R., 2009. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Research 19, 1141–1152.

Hamady, M., Walker, J., Harris, J.K., Gold, N.J., Knight, R., 2008. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nature Methods 5, 235–237.

Hamady, M., Lozupone, C., Knight, R., 2010. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. The ISME Journal 4, 17–27.

Huber, J.A., Mark Welch, D., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A., Sogin, M.L., 2007. Microbial population structures in the deep marine biosphere. Science 318, 97–100.

Hughes, J.B., Hellmann, J.J., Ricketts, T.H., Bohannan, B.J.M., 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. Applied and Environmental Microbiology 67, 4399–4406.

Krzanowski, W.J., 2000. Principles of Multivariate Analysis: A User's Perspective, revised edition. Oxford University Press, Oxford.

Kunin, V., Engelbrektson, A., Ochman, H., Hugenholtz, P., 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environmental Microbiology 12, 118–123.

Lauber, C.L., Hamady, M., Knight, R., Fierer, N., 2009. Pyrosequencing-based assessment of soil ph as a predictor of soil bacterial community structure at the continental scale. Applied and Environmental Microbiology 75, 5111–5120.

Ley, R.E., Turnbaugh, P.J., Klein, S., Gordon, J.I., 2006. Microbial ecology — human gut microbes associated with obesity. Nature 444, 1022–1023.

Li, W.Z., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.

Lozupone, C., Knight, R., 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71, 8228–8235.

Lozupone, C.A., Knight, R., 2007. Global patterns in bacterial diversity. Proceedings of the National Academy of Sciences of the United States of America 104, 11436–11440.

Lozupone, C.A., Knight, R., 2008. Species divergence and the measurement of microbial diversity. FEMS Microbiology Reviews 32, 557–578.

Lozupone, C., Hamady, M., Knight, R., 2006. UniFrac — an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics 7.

Ludwig, J.A., Reynolds, J.F., 1988. Statistical Ecology. Wiley, New York.

Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., et al., 2006. Microbial biogeography: putting microorganisms on the map. Nature Reviews. Microbiology 4, 102–112.

Patil, G.P., Taillie, C., 1982. Diversity as a concept and its measurement. Journal of the American Statistical Association 77, 548–561.

Roesch, L.F., Fulthorpe, R.R., Riva, A., et al., 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. The ISME Journal 1, 283–290.

Schauer, R., Bienhold, C., Ramette, A., Harder, J., 2010. Bacterial diversity and biogeography in deep-sea surface sediments of the South Atlantic Ocean. The ISME Journal 4, 159–170.

Schloss, P.D., Handelsman, J., 2006. Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. Applied and Environmental Microbiology 72, 6773–6779.

Schloss, P.D., Westcott, S.L., Ryabin, T., et al., 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75, 7537–7541.

Simpson, E.H., 1949. Measurement of diversity. Nature 163, 688.

Sogin, M.L., Morrison, H.G., Huber, J.A., et al., 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences of the United States of America 103, 12115–12120.

Teixeira, L., Peixoto, R.S., Cury, J.C., Sul, W.J., Pellizari, V.H., Tiedje, J., Rosado, A.S., 2010. Bacterial diversity in rhizosphere soil from Antarctic vascular plants of Admiralty Bay, maritime Antarctica. The ISME Journal 4, 989–1001.