Global patterns in bacterial diversity

Catherine A. Lozupone* and Rob Knight†‡

*Departments of Molecular, Cellular, and Developmental Biology and †Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309

Edited by Norman R. Pace, University of Colorado, Boulder, CO, and approved May 29, 2007 (received for review December 22, 2006)

Microbes are difficult to culture. Consequently, the primary source of information about a fundamental evolutionary topic, life's diversity, is the environmental distribution of gene sequences. We report the most comprehensive analysis of the environmental distribution of bacteria to date, based on 21,752 16S rRNA sequences compiled from 111 studies of diverse physical environments. We clustered the samples based on similarities in the phylogenetic lineages that they contain and found that, surprisingly, the major environmental determinant of microbial community composition is salinity rather than extremes of temperature, pH, or other physical and chemical factors represented in our samples. We find that sediments are more phylogenetically diverse than any other environment type. Surprisingly, soil, which has high species-level diversity, has below-average phylogenetic diversity. This work provides a framework for understanding the impact of environmental factors on bacterial evolution and for the direction of future sequencing efforts to discover new lineages.

environmental distribution | microbial ecology | phylogenetic diversity | UniFrac

A global picture of microbial diversity has remained elusive, yet it is critical to understanding microbial adaptation to different environments and their function in those environments. Sequencing of 16S rRNA genes from environmental samples has revolutionized our understanding of microbial systematics and diversity, revealing how far we are from cataloguing the vast diversity of microorganisms on Earth (1–4). Integrating information from these environmental surveys, however, has thus far been a formidable obstacle to a global understanding of microbial ecology. Determining physical and chemical factors, such as temperature, pH, or geography, that correlate with differences between diverse microbial communities will reveal how easily microbes tolerate different kinds of environmental change and will increase our understanding of microbial ecology and evolution. In addition, determining the environment types that contain the most phylogenetic diversity will reveal where new sequencing efforts to catalog global bacterial diversity will be most efficient at uncovering deep-branching lineages. Because of inconsistencies in how diversity is measured in individual studies, e.g., how operational taxonomic units (OTUs) are selected or which region of the rRNA gene is sequenced, it is only by integrating information from these studies into a single phylogenetic context that these important questions can be addressed.

Results

Toward a Global Survey of Natural Environments. We created an environmentally annotated tree of the bacteria including 21,752 sequences from 202 environmental samples compiled from 111 studies of diverse, globally distributed natural environments. We chose published studies that sequenced the most 16S rRNA clones, surveyed natural environments, and used primers sufficiently general to amplify all bacteria. The samples represent a vast diversity of environments, ranging from "normal" environments such as soil, seawater, and sediments to environments at the extremes of temperature (hot springs, hydrothermal vents, marine ice), salinity (hypersaline basins, lakes and mats), acidity (acidic springs and rocks, alkaline lakes), and nutrient availability (oligotrophic caves) [Table 1 and supporting information (SI) Data Set 1]. To normalize sampling effort across studies that used different techniques [e.g., by using restriction fragment length polymorphism (RFLP) patterns to screen for unique clones], we chose OTUs from each sample using a 97% identity threshold (5), including one sequence from each OTU in the analysis (see *Materials and Methods*).

Salinity Is the Major Factor Relating Microbial Communities. We clustered the environmental samples by the phylogenetic lineages that they contain by applying principal coordinates analysis (PCoA) (6, 7) (Figs. 1 and 2) and hierarchical clustering (8, 9) (and see SI Fig. 4) to a matrix of UniFrac distances by using the UniFrac web interface (10). UniFrac measures the distance between two communities as the fraction of branch length in a phylogenetic tree that leads to descendants of members of either community but not both (11). It thus captures the amount of environment-specific evolution in a single phylogenetic tree. Surprisingly, the major division is by salinity (Fig. 1 and SI Fig. 4). Almost all nonsaline environments (Fig. 1, pink circles), even those with extreme temperature and pH such as hot springs and acidic endolithic communities, cluster to the left of the diverse saline environments (Fig. 1, green triangles) along principal coordinate (PC) 1. Samples where saline and nonsaline water mix (blue squares) have intermediate values. The saline environments include marine samples, lakes, and springs: note that determinations of salinity in this study are qualitative and based on the habitat descriptions rather than on direct measurements of salt concentration. Remarkably few samples deviate from this trend, and those that do are illustrative. Two nonsaline samples cluster with the saline group: one is a microbial mat from a chemautolithotrophic cave community involved in mineral deposition, which may be locally saline (12); the other is from an anoxic rice paddy soil (13), where salinization is a common agricultural problem. One saline sample clusters with nonsaline: this is a coastal ocean sample from a study that also sampled the adjacent river and estuary (14), raising the possibility of contamination.

Environments of the same type also cluster together, in both the hierarchical cluster (SI Fig. 4) and PCoA plots (Fig. 2), even though each type includes diverse environments (Table 1). For example, nonsaline water samples (blue pentagons, Fig. 2) have high PC2 values, and surface soils (Fig. 2, purple inverted triangles) and sediments (Fig. 2, yellow sidewise triangles) have low PC2 values, indicating that substrate type (water vs. sediment) is the second most important factor for explaining community differences. Soils and sediments cluster separately, and

Author contributions: R.K. designed research; C.A.L. performed research; C.A.L. analyzed data; and C.A.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: OTU, operational taxonomic unit; PCoA, principal coordinates analysis; PD, phylogenetic diversity; G, gain in phylogenetic diversity; SSU, small subunit; NJ, neighbor-joining.

‡To whom correspondence should be addressed. E-mail: rob@spot.colorado.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0611525104/DC1.

Table 1. Summary of the 15 groups into which we binned the 202 samples

Group name	Description	#S	#O	*G* Resid	PD Resid	SN	ET
Nonsaline cultured (Nc)	Cultured from diverse nonsaline environments including soil, lake water, lake sediment, and air	11	545	−0.88**	−1.87**		
Soil surface (Nso)	Soils of diverse types (agricultural, rainforest, temperate forest, grass pasture, and desert) and geographical regions; some polluted (PCB, HC)	31	3,560	−0.76**	−0.61		
Nonsaline submerged (Nsu)	Soils that are submerged and potentially anoxic including subsurface soils, a rice paddy, and polluted and pristine wetland soils, aquifers, and sediments from a cave	10	490	0.41	0.30		
Nonsaline sediment (Nse)	Sediment from nonsaline lakes and reservoirs	7	708	0.60*	1.77**		
Nonsaline water (Nw)	Rivers and lakes from diverse geographical regions and trophic level. Samples taken from various depths	28	1,208	−0.34	0.26		
Nonsaline endolithic (Nen)	Scraped from cave walls with and without artificial lighting, epi and endolithic limestone in Mexico, and an acidic endolithic community in Yellowstone National Park	7	239	0.20	0.03		
Nonsaline springs (Nsp)	Thermophilic springs from Yellowstone National Park and Thailand (sediment and growth slide) and a microbial mat in a cave sulfidic spring	5	150	0.56*	0.02		
Saline cultured (Sc)	Cultured from diverse saline environments including marine ice, sediment, and coastal water, a salt marsh, and a hypersaline stromatolite	9	279	−0.23	−1.0		
Saline sediment (Sse)	Sediment from diverse saline environments including meromictic lakes, coastal and deep sea sediments, brackish to hypersaline water, active and inactive hydrothermal vent sites, gas hydrate mounds, salterns, and springs	38	1,979	0.54**	0.70**		
Saline water-anoxic (Swa)	Water from anoxic saline environments including meromictic Mono Lake (California), the anoxic zone of the Cariaco Basin, and deep hypersaline basins in the Mediterranean Sea	8	402	0.19	0.69*		
Saline water-subsurface (Swb)	Water from subsurface samples between 10 and 4,000 M depth, from diverse geographical locations and mostly open ocean	16	999	−0.057	−0.08		
Saline water-surface (Sws)	Water from surface saline water, mostly from coastal samples of diverse geographical location but also the Sargasso Sea	10	713	−0.15	−0.77		
Saline ice (Smi)	Marine ice from the Arctic and Antarctic.	2	78	−0.002	−0.30		
Saline-misc (So)	Miscellaneous saline environments including within gas hydrate mounds, salt marsh grasses, stromatolites, hypersaline mats, basalt, and hydrothermal vent colonizers	16	1,464	0.83**	0.14		
Mixed (M)	Environments with mixing of water from saline and nonsaline sources including estuaries and an intertidal hotspring	4	170	−0.33	0.29		

#S is the number of samples representing each group, #O is the number of OTUs represented in the samples, and *G* Resid and PD Resid are the average residuals for regression of *G* and PD values against sampling effort (Fig.3 and SI Fig. 5). A negative/positive residual means that the point fell below/above the overall regression line, and is indicative of low/high comparative diversity. Significantly different residual averages are marked with ''**'' (''*'' indicates that the value is sufficiently different that it would likely become significant with a larger sample size). The symbols that represents the samples in Fig. 1 (SN) and Fig. 2 (ET) are also indicated. Detailed information on the samples in each group is in SI Data Set 1.

submerged soils and aquifers (Fig. 2, gray diamonds) generally cluster with sediments. Interestingly, even hot springs (Fig. 2, cyan sidewise triangles) partition by substrate type along PC2. Hot spring sediments (Nsp_1, Nsp_7: see SI Data Set 1 for label descriptions) cluster with nonsaline sediments along PC2, and communities that colonized glass slides placed in the microbial mats (Nsp_93, Nsp_94) cluster near nonsaline water.

As we showed previously in marine environments (11), cultured samples from different environments (Fig. 2, pink circles and hexagons) generally cluster together rather than with their environment types. Cultured samples separate by salinity, however, both in the hierarchical cluster (SI Fig. 4) and along PC1 (Fig. 1). Although cultured samples do not separate from other water samples when PC1 and PC2 alone are used, PC3 clearly separates these groups (Fig. 2*B*). A few samples still do not separate from the cultured isolates when the first three principal components are used. These samples include both uncultured marine ice samples (Fig. 2*B*, green circles), about half of the endolithic communities (Fig. 2*B*, green triangles), and a small proportion of the other environment types. We have previously

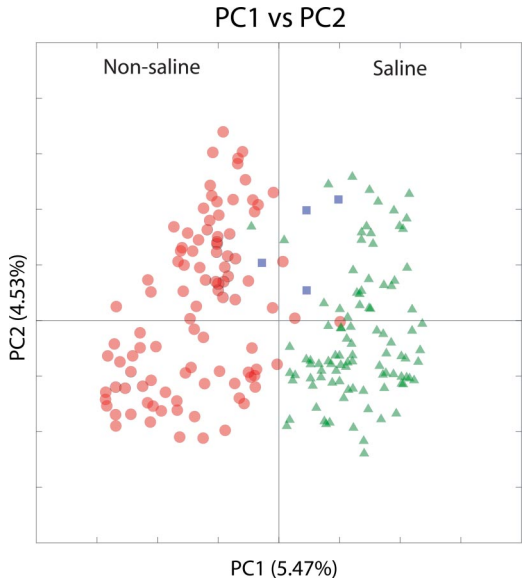

Fig. 1. Results of PCoA colored by salinity. Results of PCoA with a UniFrac distance matrix comparing the 202 samples summarized in Table 1 and SI Data Set 1. The scatterplot is of principal coordinate 1 (PC1) vs. principal coordinate 2 (PC2). The symbols are as described in Table 1: red circles indicate nonsaline environments, green triangles indicate saline environments, and blue squares indicate mixed environments. The percentage of the variation in the samples described by the plotted principal coordinates is indicated on the axes.

noted the similarity between uncultured marine ice communities and cultured isolates (11) and related it to the observation that most bacteria in marine ice can be cultured (15). The results suggest that the same may be true for many endolithic communities.

The saline environments separated along PC2 according to the same properties as the nonsaline environments, although clustering within each saline environment was looser. Hierarchical clustering (SI Fig. 4) and PCoA (Fig. 2) divided saline water samples into three subgroups: surface water, mostly in coastal regions (Fig. 2, blue inverted triangles); subsurface water, mostly in the open ocean (Fig. 2, gray sidewise triangles); and anoxic water from many locations (Fig. 2, cyan triangles; Table 1). The saline sediments (Fig. 2, purple circles) clustered together but overlapped other saline environments, including hypersaline mats, stromatolites, hydrothermal vent colonizers (Table 1, Saline–misc; Fig. 2, yellow squares), and anoxic saline water samples. Like nonsaline water and cultured isolates, surface/coastal water and cultures from saline environments separated from saline sediments along PC2. These results reinforce the suggestion that substrate type (water vs. sediment) is the second most important property for structuring diversity, perhaps because of differences in lineages adapted to planktonic vs. sessile lifestyles. However, because anoxic water samples cluster with sediments, oxygenation may also be important. For instance, clades of obligate anaerobes, such as the *Clostridia*, and clades with many planktonic representatives, such as filamentous α-proteobacteria, probably account for some of these community differences.

Environment Types Differ Substantially in Phylogenetic Diversity (PD). We also determined the PD of each sample, which is the branch length that remains when all other sequences are removed from the tree (16), and the PD gain (G), which is the branch length a sample adds to a tree containing sequences from all other samples (16). For example, if a new sample contained only sequences already found in other studies, adding that sample's

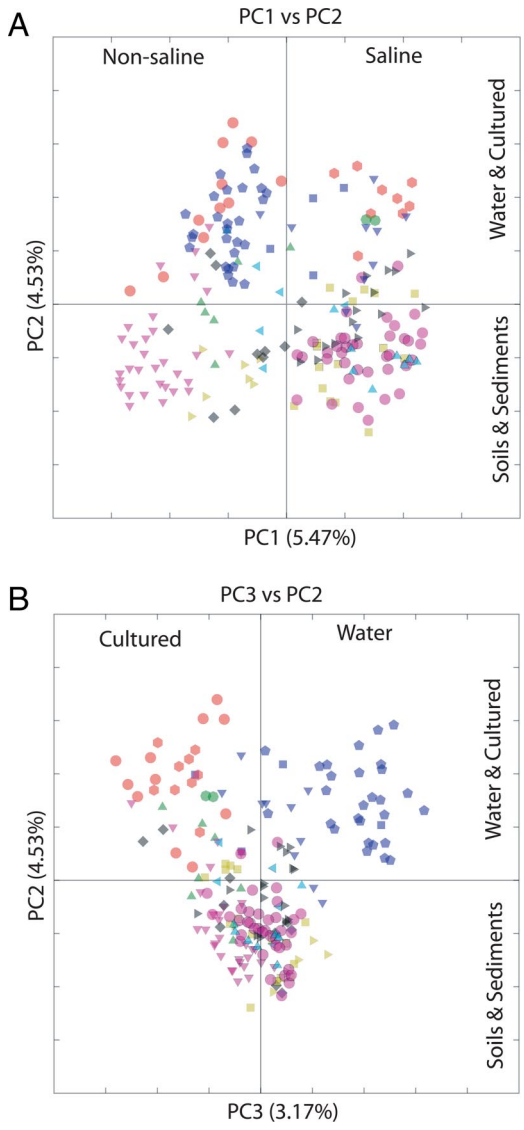

Fig. 2. Results of PCoA colored by environment type. A scatterplot of PC1 vs. PC2 (*A*) and PC3 vs. PC2 (*B*). The symbols represent the 202 samples and are as described in Table 1. A file of this scatterplot in which pop-up windows indicate which point corresponds to which sample name is available; see *SI Text*.

sequences to the tree would add no new branch length, and the G value would be 0. Environments with high G values are promising sites for discovering new, diverse microbial lineages. Samples with high PD and low G values have many phylogenetic lineages that are also found in other environments.

Because sequencing effort influences diversity estimates, we regressed both G (Fig. 3) and PD (SI Fig. 5) values on the number of OTUs in each sample. The relationships between sequencing effort and both PD and G are approximately linear (R^2 of 0.76 and 0.91, respectively), suggesting that deep sequencing of one environment uncovers as much new diversity as shallow sequencing of many related environments. Regressions for individual environment types indicated substantial differences in their contributions to known diversity (Fig. 3). We quantified these differences by calculating the residual of each sample from the regression of all samples (Fig. 3, blue line). Highly positive or negative residuals indicate high or low diversity respectively (Table 1; see Data Set 1 for individual sample results).

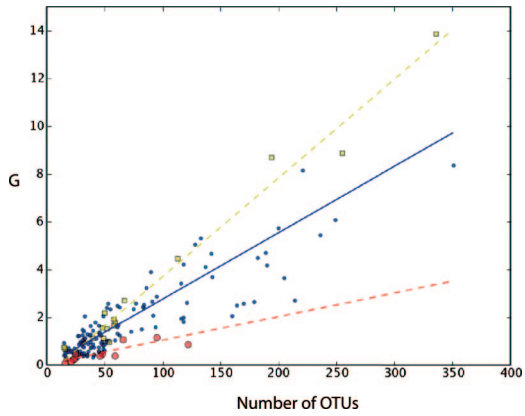

Fig. 3. Unique diversity (*G*) regression analysis. Plot of the amount of branch length that is added to the phylogenetic tree (*G* value) by each of the 202 samples, vs. the number of OTUs that represents each sample. The main regression line is shown in blue. Bacteria that were cultured from nonsaline environments (pink circles) generally fell below the main regression line (i.e., they had negative residuals for *G*), thus contributing less unique diversity. Samples in the saline–misc group (yellow squares) generally fell above the main regression line (i.e., they had positive residuals for *G*), thus contributing more unique diversity.

Soils Are Less Diverse Than Expected and Sediments and Hypersaline Mats Are More Diverse. Surprisingly, surface soils had significantly lower *G* values than other environments and negative average PD residuals (Table 1), even though soil is often described as one of the most diverse environment types on earth (17, 18). High estimates for soil diversity are based on the number of OTUs found in each sample (18) and reassociation kinetics (19) and not phylogenetic diversity. The high species diversity in soil may result from more closely related species persisting in the same sample, perhaps adapting to different niches by horizontal gene transfer (which would not affect phylogenetic relatedness measured by 16S rRNA).

The nonsaline cultured group also had significantly lower average PD and *G* residual values than the other environments (Table 1). This result is consistent with the observation that few lineages in these environments can be cultured (2). The saline-cultured environments also had negative average residuals for both total PD and *G*.

Saline sediment and saline-misc (Table 1) have significantly higher *G* values than other environment types. Nonsaline sediments and springs resembled saline sediments, but the sample sizes were too low for statistical significance (Table 1). Saline and nonsaline sediments also had high average PD residuals (Table 1). High diversity in sediments is consistent with previous observations and may stem from their highly stratified nature and chemical gradients (17). Nonsaline sediments are less thoroughly sampled than saline sediments and are thus especially good targets for future sequencing efforts. Interestingly, the miscellaneous saline and nonsaline spring groups had high *G* and low PD values, indicating that they, on average, contain relatively few, but highly divergent, lineages.

Some environment types clustered poorly, suggesting that they may not form natural groups. Residuals for individual samples are thus of interest (see SI Data Set 1 for values). The sample with the lowest *G* residual (Sws_M_163; −3.64 standard deviations from the mean) was from the Sargasso Sea (20), an environment known to have low diversity because of nutrient limitation and little spatial heterogeneity. The samples with the highest *G* residuals (So_Mm+_166 and So_Mm+_168; 5.08 and 3.71 standard deviations from the mean, respectively) were from different layers of the Guerrero Negro hypersaline mat, the molecular analysis of which introduced 15 previously unidentified candidate phyla, an unprecedented number for a single environment (21).

Discussion

The comprehensive analysis of the environmental distribution of bacteria has provided insights that were not apparent in the original studies. Because the analysis relies on a phylogeny of 16S rRNA sequences, the clear grouping of samples by environment type indicates a direct relationship between 16S rRNA lineages and environmental distribution. Thus, although processes such as horizontal gene transfer can be important factors for adaptation to new environments, they cannot obscure the overall evolutionary pattern, suggesting that bacteria make genomic trade-offs that prevent major changes in lifestyle simply through new gene acquisition. Some factors, such as salinity, seem especially to encourage such lineage-specific adaptations.

The results also add an interesting perspective to the study of extreme environments. Although organisms in environments at the extremes of temperature and pH are presumably under strong selective pressures, they still cluster by salinity and substrate type, indicating that the general properties of these environments still primarily determine which lineages can survive there.

The ability of comparisons of 16S rRNA data to reveal the effects of specific chemical and physical factors on microbial communities depends on the quality of information that has been measured for the source environments and the accessibility of this information in the public databases. Although we found clear patterns of variation between environment types, such as the split between saline and nonsaline environments, testing whether this split stems from ionic strength, osmolarity, availability of sulfate for reduction, or other factors remains unresolved, in part because detailed measurements were not available for many of the environmental samples. Another limitation is that, because the records do not include information on how many times each sequence was observed in each sample, it is not possible to compare samples by using quantitative measures of β diversity such as weighted UniFrac (22). Information about relative abundances is also required for almost all measurements of α diversity (total diversity of a sample) including Chao1, ACE, rarefaction analysis, and the Shannon and Simpson indices (reviewed in ref. 23). Thus, improved availability of environment information within structured, machine-readable fields in the database is a key requirement for future large-scale analyses of the factors influencing microbial diversity.

The overview that this analysis provides is useful for evaluating where to direct new sequencing efforts. The environmental clustering patterns allow us, at least in some cases, to define environment types based on the occurrence of similar bacterial lineages rather than arbitrary criteria. For instance, nonsaline lakes and rivers behave as a cohesive group but saline water does not. Evaluation of these environment types, as well as of individual environments, allows us to identify optimal targets for finding new diversity.

Materials and Methods

Selecting Relevant Environmental Samples. We extracted GenBank records from the April 15, 2006 release and identified small subunit (SSU) rRNA sequences and their associated publication titles. We identified SSU rRNA sequences as records that had any of the terms ("SSU," "16S," "18S," or "small subunit" and "rRNA," "rDNA," or "ribosomal RNA") (the search was case-insensitive). We extracted reference information for each record using a custom parser, and grouped the sequences that had the same title. The Excel spreadsheet that summarizes these large SSU rRNA surveys is available; see *SI Text*. The 267,731 putative SSU rRNA sequences were associated with 17,836 unique titles. Of these, 1,032 titles were associated with at least 50 sequences.

Surprisingly, fewer than half of all of the studies were associated with any publication, (485 of the 1,032 studies associated with at least 50 16S rRNA sequences). This underscores the importance of generating a standardized form for annotating sequences in the public databases with detailed information on the environments from which the sequences came.

Making the Phylogenetic Tree. We used NAST (24) to add sequences from the 111 selected studies to the standard Arb alignment (25). We then added the 21,752 sequences from the studies to a guide tree with >110,000 sequences using the Arb parsimony insertion tool. The guide tree was initially described in ref. 26 but was subsequently enhanced by the Pace lab (J. K. Harris and N. R. Pace, personal communication). We used a lanemask ("lanemaskPH") that is provided with the Hugenholz Arb database (27) available at the Ribosomal Database Project II (28), to exclude hypervariable regions from consideration while generating the tree. We chose a parsimony insertion algorithm rather than a *de novo* method such as neighbor joining (NJ) because it can relate sequences from different parts of the 16S rRNA molecule. This is essential because there is very little overlap in sequenced 16S rRNA regions when comparing all of the studies. For instance, only 6,552 of the 21,752 sequences (30%) were complete between positions homologous to 300 and 700 in *Escherichia coli* 16S rRNA and only 7,102 (33%) were complete for the region between *E. coli* positions 700 and 1,100. To test whether the Arb parsimony insertion tree gave similar results to a tree built *de novo*, we performed PCoA clustering on NJ trees of sequences from the 82 and 90 environments that had >15 sequences in the 300–700 region and the 700–1,100 region, respectively. The NJ trees were also made in Arb, by using the Jukes–Cantor model of nucleotide substitution. We compared the results to those from Arb parsimony insertion trees with the same set of sequences. For both regions, the results of PCoA clustering with the parsimony insertion and NJ trees were almost identical (data not shown). Clustering by using only the portion of the data that could be incorporated into the NJ trees recovered the saline/nonsaline split as the most important division in the data for both regions, although the coordinate axes were rotated slightly.

Selecting OTUs and Annotating the Tree with Environment Information. We divided the sequences into 225 environmental samples using annotations from the associated publications. By excluding 23 samples with <15 OTUs each, we produced a tree with 12,984 OTUs representing 202 samples. For each environmental sample, we chose OTUs with a 97% identity threshold using our Divergent Set software (5). We decided to dereplicate the sequence data for several reasons. First, dereplication of the data has little effect on clustering with UniFrac, because inclusion of near similar sequences will not change the amount of unique branch length in the tree. Removing near similar sequences thus produces a smaller tree that is more easily manipulated, without affecting the results. Second, because the inclusion of very small samples in a UniFrac analysis can produce spurious results, we wanted to exclude small environmental samples. Because some studies deposit near-identical sequences in GenBank, and others deposit sequences only after choosing OTUs, we needed to remove near-identical sequences from all studies to evaluate our sampling effort fairly. Finally, when we corrected the raw PD and G values for sampling effort, it was again essential to ensure that the results would be robust to the methodology used to choose OTUs in the original studies. We chose the 97% threshold because this is the most common threshold used for dereplication at the species level. Repetition of the analysis with all available sequences, i.e., without choosing OTUs at all, provided almost identical UniFrac clustering results (data not shown).

Statistical Analyses. We performed PCoA and hierarchical clustering in the UniFrac web interface (10), using the Arb tree and a file mapping sequence labels to environmental samples as input. PCoA is similar to principal coordinates analysis (PCA), except that the starting point is a matrix of distances between samples rather than a matrix of observations about each sample. We used the unweighted pair group method with arithmetic mean (UPGMA) hierarchical clustering algorithm, which produces clusters by finding the nearest pair of neighbors at each step, finding the midpoint between these neighbors, and adding a cluster consisting of the neighbors to a growing tree.

We also used the Arb tree for diversity analyses. We calculated PD for each sample by removing all sequences not from the sample from the tree and summing the remaining branch length. We determined G by removing only the sequences from that sample from the tree and summing the remaining branch length. We corrected each PD and G value for sampling effort by calculating the residual from the regression of PD and G vs. OTU count for all of the samples. We determined whether the average G and PD residuals for each environment type were significantly different from samples not in that environment type with a two-tailed Student's t test. These statistical analyses were performed by using custom code written in the Python language.

We thank Noah Fierer, Norman Pace, Jeffrey Gordon, Michael Yarus, Jesse Zaneveld, Kirk Harris, Jeff Walker, and Ruth Ley for valuable feedback on drafts of the manuscript. C.L. was supported by National Institutes of Health Predoctoral Training Grant T32 GM08759. This work was performed by using the Keck RNA Bioinformatics facility (Yale University, New Haven, CT).

1. Hugenholtz P, Goebel BM, Pace NR (1998) *J Bacteriol* 180:4765–4774.
2. Rappe MS, Giovannoni SJ (2003) *Annu Rev Microbiol* 57:369–394.
3. Schloss PD, Handelsman J (2004) *Microbiol Mol Biol Rev* 68:686–691.
4. Pace NR (1997) *Science* 276:734–740.
5. Widmann J, Hamady M, Knight R (2006) *Mol Cell Proteomics* 5:1520–1532.
6. Gower JC (1966) *Biometrika* 53:325–338.
7. Krzanowski WJ (2000) *Principles of Multivariate Analysis. A User's perspective* (Oxford Univ Press, Oxford).
8. Felsenstein J (2004) *Inferring Phylogenies* (Sinauer, Sunderland, MA).
9. Sokal RR, Michener CD (1958) *Univ Kansas Sci Bull* 38:1409–1438.
10. Lozupone C, Hamady M, Knight R (2006) *BMC Bioinformatics* 7:371–385.
11. Lozupone C, Knight R (2005) *Appl Environ Microbiol* 71:8228–8235.
12. Engel AS, Lee N, Porter ML, Stern LA, Bennett PC, Wagner M (2003) *Appl Environ Microbiol* 69:5503–5511.
13. Hengstmann U, Chin KJ, Janssen PH, Liesack W (1999) *Appl Environ Microbiol* 65:5050–5058.
14. Crump BC, Armbrust EV, Baross JA (1999) *Appl Environ Microbiol* 65:3192–3204.
15. Brinkmeyer R, Knittel K, Jurgens J, Weyland H, Amann R, Helmke E (2003) *Appl Environ Microbiol* 69:6610–6619.
16. Faith DP (1992) *Biol Conservation* 61:1–10.
17. Torsvik V, Ovreas L, Thingstad TF (2002) *Science* 296:1064–1066.
18. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, *et al.* (2005) *Science* 308:554–557.
19. Gans J, Wolinsky M, Dunbar J (2005) *Science* 309:1387–1390.
20. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, *et al.* (2004) *Science* 304:66–74.
21. Ley RE, Harris JK, Wilcox J, Spear JR, Miller SR, Bebout BM, Maresca JA, Bryant DA, Sogin ML, Pace NR (2006) *Appl Environ Microbiol* 72:3685–3695.
22. Lozupone CA, Hamady M, Kelley ST, Knight R (2007) *Appl Environ Microbiol* 73:1576–1585.
23. Magurran AE (2004) *Measuring Biological Diversity* (Blackwell, Oxford).
24. DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, Phan R, Andersen GL (2006) *Nucleic Acids Res* 34:394–399.
25. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, *et al.* (2004) *Nucleic Acids Res* 32:1363–1371.
26. Hugenholtz P, Huber T (2003) *Int J Syst Evol Microbiol* 53:289–293.
27. Hugenholtz P (2002) *Genome Biol* 3:1–8.
28. Maidak BL, Cole JR, Lilburn TG, Parker CT, Jr, Saxman PR, Farris RJ, Garrity GM, Olsen GJ, Schmidt TM, Tiedje JM (2001) *Nucleic Acids Res* 29:173–174.