Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample

J. Gregory Caporaso[a], Christian L. Lauber[b], William A. Walters[c], Donna Berg-Lyons[b], Catherine A. Lozupone[a], Peter J. Turnbaugh[d], Noah Fierer[b,e], and Rob Knight[a,f,1]

[a]Department of Chemistry and Biochemistry, [b]Cooperative Institute for Research in Environmental Sciences, [c]Department of Molecular, Cellular, and Developmental Biology, and [e]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309; [d]Harvard FAS Center for Systems Biology, Cambridge, MA 02138; and [f]Howard Hughes Medical Institute, Boulder, CO 80309

Edited by Jeffrey I. Gordon, Washington University School of Medicine, St. Louis, MO, and approved April 30, 2010 (received for review February 27, 2010)

The ongoing revolution in high-throughput sequencing continues to democratize the ability of small groups of investigators to map the microbial component of the biosphere. In particular, the coevolution of new sequencing platforms and new software tools allows data acquisition and analysis on an unprecedented scale. Here we report the next stage in this coevolutionary arms race, using the Illumina GAIIx platform to sequence a diverse array of 25 environmental samples and three known "mock communities" at a depth averaging 3.1 million reads per sample. We demonstrate excellent consistency in taxonomic recovery and recapture diversity patterns that were previously reported on the basis of meta-analysis of many studies from the literature (notably, the saline/nonsaline split in environmental samples and the split between host-associated and free-living communities). We also demonstrate that 2,000 Illumina single-end reads are sufficient to recapture the same relationships among samples that we observe with the full dataset. The results thus open up the possibility of conducting large-scale studies analyzing thousands of samples simultaneously to survey microbial communities at an unprecedented spatial and temporal resolution.

human microbiome | microbial community analysis | microbial ecology | next-generation sequencing

High-throughput sequencing technologies have opened new frontiers in microbial community analysis by providing a cost-effective means of identifying the microbial phylotypes that are present in samples. These studies have revolutionized our understanding of the microbial communities in our bodies (1, 2) and on our planet (3–5). This revolution in sequencing technology, combined with the development of advanced computational tools that exploit metadata to relate hundreds of samples to one another in ways that reveal clear biological patterns, has reinvigorated studies of the 16S rRNA gene (6). Studies of 16S rRNA provide a view of which microbial taxa are present in a given sample because it is an excellent phylogenetic marker (7). Although alternative techniques, such as metagenomics, provide insight into all of the genes (and potentially gene functions) present in a given community, 16S rRNA-based surveys are extraordinarily valuable given that they can be used to document unexplored biodiversity and the ecological characteristics of either whole communities or individual microbial taxa. Perhaps because 16S rRNA phylogenies tend to correspond well to trends in overall gene content (8), the ability to relate trends at the species level to host or environmental parameters has proven immensely powerful (9).

New technologies have led to astonishing decreases in the cost of sequencing: at the scale of the whole human genome, the price per megabase has decreased by approximately an order of magnitude per year since 2001 (10). This rapid increase in sequencing capacity has led to a process almost akin to a coevolutionary "arms race" in which newer sequencing platforms generate datasets of unprecedented scale that break existing software tools: new software is then developed that exploits these massive datasets to produce new biological insight, but in turn the availability of these software tools prompts new experiments that could not previously have been considered, which lead to the production of the next generation of datasets, starting the process again. However, we would argue that the situation is not precisely that of a "Red Queen" coevolutionary process (in which one must run faster and faster to remain in the same place), because each advance really does provide a new level of insight into a range of biological phenomena. The increase in number of sequences per run from parallel pyrosequencing technologies such as the Roche 454 GS FLX (5×10^5) to Illumina GAIIx (1×10^8) is on the order of 1,000-fold and greater than the increase in the number of sequences per run from Sanger (1×10^3 through 1×10^4) to 454. The transition from Sanger sequencing to 454 sequencing has opened new horizons in microbial community analysis by making it possible to collect hundreds of thousands of sequences spanning hundreds of samples. A transition to the Illumina platform will similarly allow for deeper sequencing than has previously been feasible, with the possibility of detecting even phylotypes that are very rare (11). By using a variant of the barcoding strategy used for 454 (12–14) with the Illumina platform (Fig. 1), thousands of samples could be analyzed in a single run, with each of the samples analyzed in unprecedented depth.

In this study, we address the question of whether the Illumina technology is suitable for large-scale comparisons among microbial communities at different scales. One limitation of the Illumina platform is that it can currently only produce relatively short reads (75–100 bp in a single read—although paired reads can produce 150–200 bp from a single molecule). Previous work has suggested that fragments of the 16S rRNA as small as 100 bp can be sufficient for resolving microbial community differences (15). Whether the short and potentially error-prone reads produced by the Illumina GAIIx are suitable for large-scale community comparisons remains unknown, although the platform has been used to sequence 16S rRNA genes from a small number of samples from the oral microbiota (11). In this study, we use a simple mock community to determine whether Illumina sequencing accurately captures information about known communities. We also address

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Microbes and Health" held November 2–3, 2009, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and audio files of most presentations are available on the NAS Web site at http://www.nasonline.org/SACKLER_Microbes_and_Health.

Author contributions: J.G.C., C.L.L., N.F., and R.K. designed research; J.G.C., C.L.L., W.A.W., and D.B.-L. performed research; J.G.C. and P.J.T. contributed new reagents/analytic tools; J.G.C., C.A.L., P.J.T., and R.K. analyzed data; and J.G.C., N.F., and R.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Data have been submitted to the NCBI Sequence Read Archive under Study Number SRA012609.1.

[1]To whom correspondence should be addressed. E-mail: rob.knight@colorado.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1000080107/-/DCSupplemental.

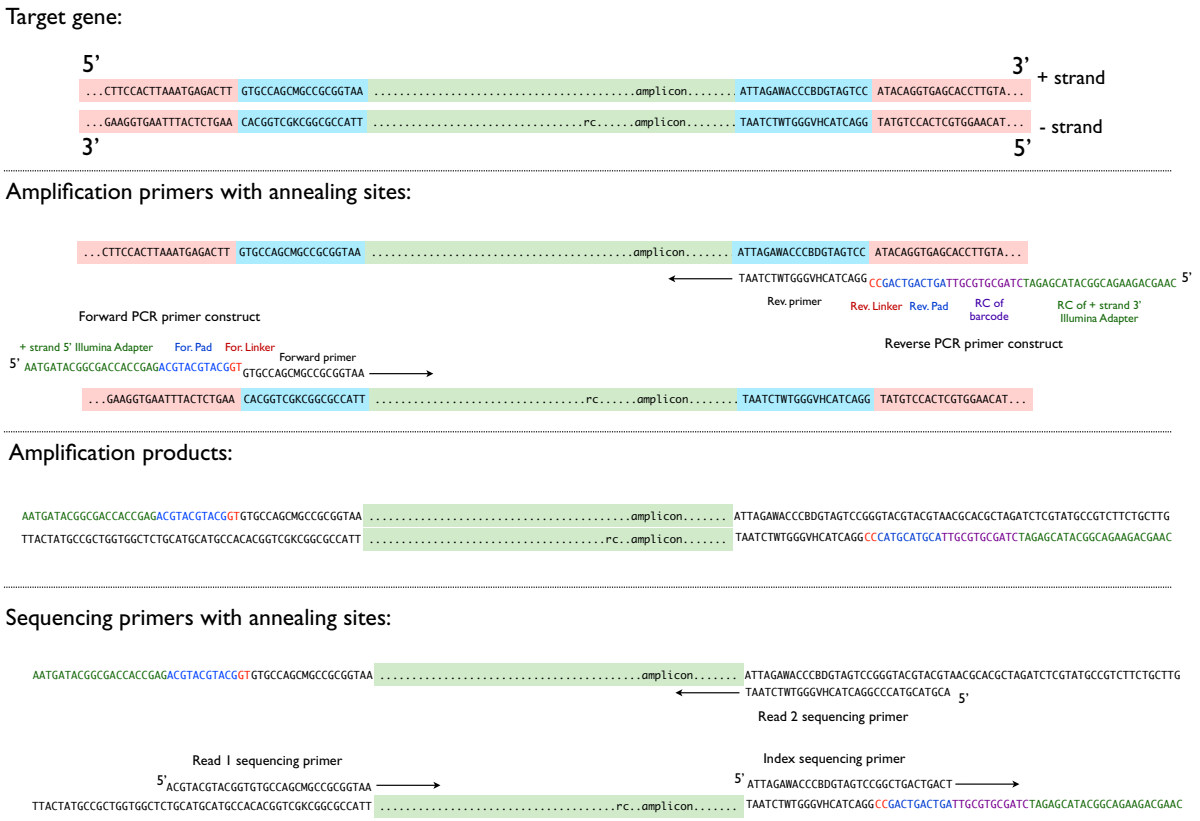

Fig. 1. Protocol for barcoded Illumina pyrosequencing. First, conserved regions within the target gene (in this case, 16S rRNA) are identified (blue), together with an amplicon that clipping studies along the lines of ref. 15 indicate are especially good for community sequence analysis (green). Second, PCR amplifications are performed, using primers that include a linker sequence not homologous to any 16S rRNA sequence at the corresponding positions, the barcode, and the Illumina adaptor. Thus, the match between the primer and the template sequence ends at the end of the black region of the primer, and the linker and adaptors (shown in color) do not match the template. This procedure yields a library of amplification products that contain the barcode and Illumina adaptors. Finally, three separate primers are used to yield the 5′ read, the 3′ read, and the index read (that yields the barcode sequence).

the question of whether this technology is suitable for large-scale comparisons among microbial communities at different scales by determining whether Illumina sequencing can recover the previous observation of an intriguing global pattern of bacterial distribution, with a partitioning of environmental sequences between saline and nonsaline habitats (16), and an even deeper partitioning between host-associated and free-living communities (17). Intriguingly, this former observation has been replicated independently in archaea (18). Accordingly, this study presents the results of sequencing barcoded PCR amplicons from environmental ($n = 25$) and mock community ($n = 3$) samples using one full plate (seven lanes plus phiX control lane) on the Illumina GAIIx platform.

Results and Discussion

Surmounting the Bioinformatics Challenge. As has been the case with other advances in sequencing technology, the unprecedented sequencing depth provided by our Illumina run posed considerable challenges for microbial community analysis software: we obtained 87,507,177 paired-end reads of exactly 100 nucleotides in length from each end before quality filtering. We therefore developed a new protocol, facilitated by the Quantitative Insights Into Microbial Ecology (QIIME) toolkit (19), which can perform standard microbial community analysis techniques on sequence sets of this size, including quality filtering of reads, efficient operational taxonomic unit (OTU) picking, taxonomy assignment, computation of α and β diversity measures, and other analyses.

Because the Illumina platform has to date been primarily used for genome sequencing and resequencing, there is no literature that we are aware of discussing and comparing quality-filtering strategies for community 16S rRNA reads. A custom strategy was therefore developed to quality filter the reads by truncating each read at the point where it incurred two or more adjacent low-quality base calls. If a truncated read was shorter than 75 bases, it was discarded. Reads surviving this step were discarded if they contained ambiguous base calls (N characters) in their sequence or barcode. After this quality filter, 36,329,392 5′ reads and 22,177,779 3′ reads were retained for subsequent analysis. Variants of this quality filter and their effects on read counts are provided as Table S1: the approach described here is conservative. We are making these methods available as part of the open-source QIIME pipeline, allowing others to apply the same techniques.

The number of sequences was too great to divide the sequences into unique OTUs at the 97% level using cd-hit (20), which is the standard tool used for this task when handling pyrosequencing data. We have previously shown, however, that patterns that were observed with de novo tree-making methods could be captured equally well using a "BLAST to reference tree" protocol (21) and then calculating community differences with UniFrac. We thus processed the data after quality filtering by using a Trie prefix tree (22), followed by BLASTing each remaining sequence against the greengenes database filtered at 99% identity (the *greengenes reference collection*) and choosing the best BLAST match via a combination of percent identity, alignment length, and E-value (the results are filtered by E-value and percent identity, and then the longest alignment matching these criteria is chosen). Representative sequences were then chosen for each OTU by choosing the most abundant sequence from the original sequence collection. A phylogenetic tree that was computed for the greengenes

reference collection using fasttree (23) was used for the calculation of phylogeny-based α and β diversity metrics. The advantage of this approach of matching sequences against a known tree is that it greatly reduced the compute time from O(N log N) to O (N) in the number of sequences. The disadvantage is that any novel taxa (i.e., microbial taxa not present in the reference collection) would effectively be disregarded (however, we note that the short reads used are problematic for identifying novel lineages in the first place). Taxonomy was then assigned to all representative sequences using the Ribosomal Database Project (RDP) classifier. All of these steps were performed using the QIIME toolkit [via QIIME's parallel wrappers in the case of BLAST (24) and the RDP classifier (25)].

Taxonomic and Alpha Diversity Analysis of Mock Communities Reveals Excellent Consistency Across Replicates. We first applied these methods to the three mock community samples, representing genomic DNA from 67 bacterial isolates pooled at even concentrations. These samples have been recently analyzed through 454 FLX pyrosequencing (26) in order to quantify the noise introduced during PCR and sequencing and its potential contribution to the observed and estimated diversity. The three mock community samples were analyzed separately from the 25 environmental samples because, rather than providing insight into similarities and differences among communities, they provide information about the sequencing error rate. The sequencing of defined communities has been critical for quantifying sequencing error profiles for pyrosequencing (27, 28), but each new method requires validation through the sequencing of "control" samples in addition to the more comprehensive analysis of simulated datasets. The 5′ and 3′ reads were analyzed independently from one another. The analysis pipeline was identical for each of the four datasets. A brief description of the analysis pipeline follows (see *Methods* for full details).

Factors that may cause the observed taxa abundances (as depicted in Fig. 2) to differ from the expected results include sequencing error, PCR primer bias, and incorrect taxonomic assignment. To assess the ability of the Illumina sequencer to capture the actual members of each community, we tested whether abundance-filtered sequences (see below) represented the expected species distribution. Within-category replicates (i.e., mock5 vs. mock5 or mock3 vs. mock3) Bray-Curtis distances were significantly lower than between-category replicates (i.e., mock5 vs. mock3) Bray-Curtis distances both at the order level ($P < 0.0001$, two-tailed, two-sample t test) and at the genus level ($P < 0.0001$). The difference in Bray-Curtis distance between the mock5 replicates and the expected species distribution, and the mock3 replicates and the expected species distribution, was not significant at the order level ($P = 0.094$, two-tailed, two-sample t test). At the genus level, however, the mock5 data were significantly more similar to the expected sequence distribution than the mock3 data ($P < 0.0001$). Taken together, these data illustrate that taxonomy assignment was highly accurate and reproducible across replicates at the order level and highly reproducible at the genus level, illustrating that the Illumina platform is able to correctly and reproducibly identify the actual members of a microbial community (although, as expected, the order-level taxa are more correctly and reproducibly assigned than the genus-level taxa by BLAST owing to the short read lengths) (Fig. 2). All between-sample Bray-Curtis distances are provided in Table S2.

We next tested whether the Illumina technique could correctly quantify the within-community (α) diversity, in terms of OTU richness. We first compared the phylogenetic diversity (PD), Chao1, and observed OTUs computed on the expected mock community sequences with the sequencing results at several minimum abundance thresholds. The three mock community samples were technical replicates designed to contain 67 OTUs (at the 97% level, corresponding to taxonomically valid species) at even abundances (26). We computed the expected α diversity of these samples after following the same workflow as applied to the sequencing results in terms of OTU picking and taxonomy assignment. The consistency between replicates was excellent (Fig. 3): only in the completely unfiltered data do the lines diverge toward the right-hand side of the graph [the dashed horizontal line shows the expected α diversity (PD) or richness (Chao1, observed species) under each measure, which should be the same at all sequence abundance thresholds examined]. The α diversities of the mock 5′ and mock 3′ datasets were greatly inflated when compared with the expected species richness. Only when a minimum abundance threshold of 10,000 was applied, meaning that sequences were only considered to

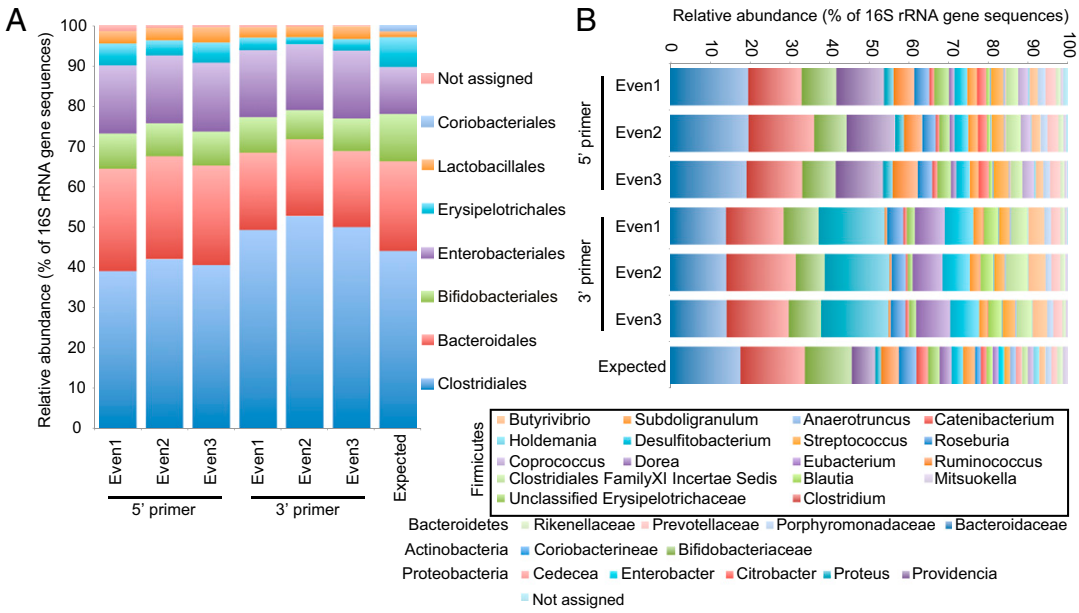

Fig. 2. Reproducibility of taxon assignment at the order level and the genus level. Reproducibility of taxon assignment at the order level (*A*) is excellent; reproducibility at the genus level (*B*) is extremely consistent within a region, although the 5′ and 3′ regions lead to somewhat different assignments.

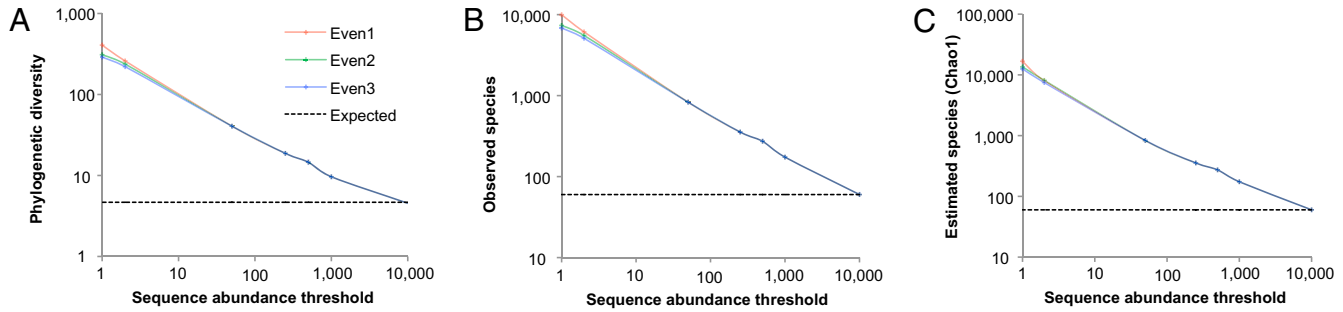

Fig. 3. Comparison of α diversity measures in the mock community. (*A*) PD, or phylogenetic diversity, a measure showing the branch length on a phylogenetic tree that is covered by a given sample. (*B*) Number of observed species-level OTUs (at the 97% level) in each community. (*C*) Number of estimated species using the Chao1 estimator of species richness. The three replicates are shown in red, blue, and green; the true value for the mock community is shown as a dashed black line (note that the expected line shown in *C* is the true number of species, not the Chao1 estimate of this number, because Chao1 has no meaning when applied to a community that consists solely of singletons). Note the log scales on both axes. As with other sequencing platforms, aggressive quality filtering is required to correctly interpret diversity results from the Illumina platform.

represent real organisms if observed at least 10,000 times (i.e., as at least 0.01% of the total number of sequences in the run) did all three α diversity metrics closely match the true diversity/richness of the mock community (Fig. 3, dashed line).

Taken together, these results suggest that the Illumina platform holds promise for community sequencing but that erroneous reads or imperfect OTUs may currently make it difficult to identify rare taxa in a sample. An understanding of the types of errors that are likely to arise, and improvement of quality-filtering strategies similar to those developed for 454 sequences (27, 28), will be important next steps in the development of the Illumina Genome Analyzer as a platform for microbial community sequencing. Quality filtering and denoising will be particularly important for studies related to rare taxa.

Beta Diversity Analysis of Environmental Samples Confirms That Host-Associated Samples Are Especially Diverse, and the Deep Partitioning of Diversity Among Saline and Nonsaline Environmental Samples. Unlike α diversity estimates such as species richness, β diversity is a measure of the degree of similarity (e.g., phylogenetic relatedness) between pairs of communities. Thus phylogenetic β diversity metrics are useful for documenting shifts in the membership and/or structure of communities that may occur between sample categories or across environmental gradients (29). Previous work has shown that β diversity metrics tend to be far less sensitive to effects such as sequencing errors and chimeras (30) than are α diversity metrics. We therefore compared the environmental samples with one another using jackknifed UPGMA (unweighted pair group method with arithmetic mean) clustering based on the unweighted UniFrac distances between samples (29). For both the 5′ and 3′ sequences (Fig. 4 *A* and *B*, respectively), the results support prior observations regarding the global patterns in bacterial distributions that were derived from metaanalysis of published data collected using different sequencing methodologies. Samples differed in membership primarily on the basis of whether they are derived from feces or nonfeces (also aerobic vs. anaerobic, in this case). Among the nonfeces samples, there was a separation between samples derived from human body habitats and those from other habitats. In

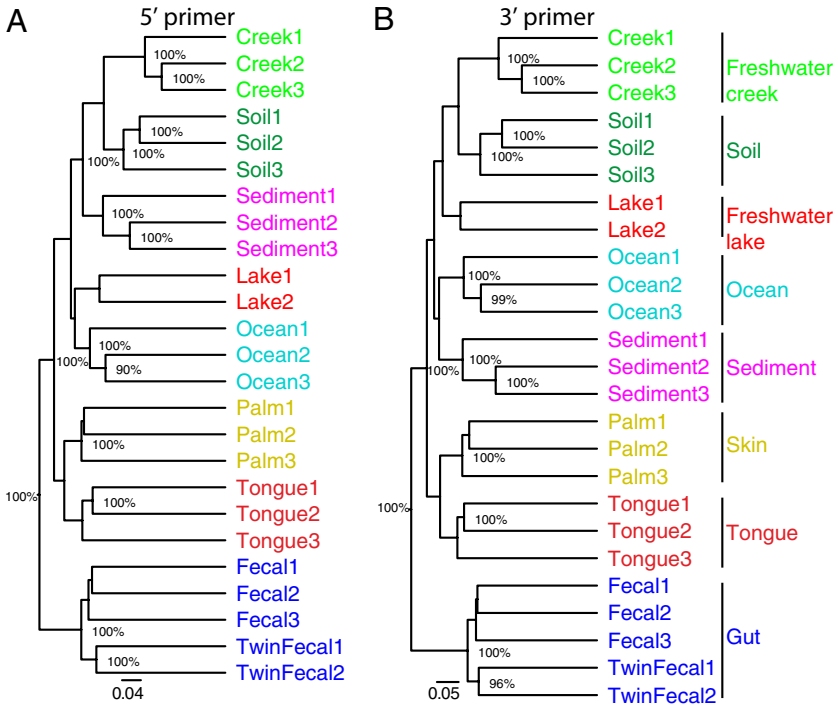

Fig. 4. UPGMA UniFrac clustering of the 5′ and 3′ reads from each environmental sample show that samples from a given environment type cluster together well. Samples are feces (blue), freshwater creek (bright green), freshwater lake (red), ocean (cyan), sediment (pink), skin (yellow), soil (dark green), and tongue (dark red). Jackknife-supported clusters showing >80% support are shown on the tree.

the human body–derived samples, the results match prior observations, with a split separating distal tongue and skin samples (1, 17). The clustering of the environmental samples differs slightly when comparing the 5′ and 3′ results. In the 3′ results, there is a split based on salinity confirming previous observations (16), whereas in the 5′ samples there is not a clear delineation between saline and nonsaline environments. The environmental delineation is clearer in the principal coordinates analysis (PCoA) plots than in the UPGMA clustering.

One key limitation to metaanalyses of 16S rRNA sequences in the public databases is that procedures for depositing representative OTUs, along with information on the relative abundances of those OTUs, are typically not represented in machine-readable form (16). High-throughput sequencing of many communities simultaneously avoids this limitation, because the Sequence Read Archive requires that all reads be deposited, allowing abundances to be calculated. When we incorporate abundance weighting into the results using the weighted UniFrac algorithm (31), we see that the clustering by sample type remains essentially the same, although somewhat more of the variance is explained. These results suggest that changes in community structure and community membership are both important in producing large-scale patterns of microbial diversity.

Conclusion

A valid concern regarding the use of the Illumina platform for environmental sequencing is that although there are many more reads, the single-end Illumina reads are less than half the length of 454 reads. The Illumina reads are between 75 and 100 bases, compared with 454 reads of 250–400 bases. The reproduction of results previously obtained using much longer Sanger reads suggest that, although the Illumina reads are considerably shorter than 454 reads, similar between-sample (β) diversity conclusions can be reached using the Illumina platform. This directly confirms the results of prior computational work on the necessary read length for accurate community comparisons (15).

Despite the success of this analysis, there are computational challenges specific to the Illumina platform that still need to be addressed to facilitate microbial community analyses on datasets of this size and larger. One specific obstacle to overcome is how to effectively take advantage of the paired-end reads in downstream analyses. Although the sequencing run presented herein used paired-end reads, the data analysis treated the 5′ and 3′ reads independently. Because the expected amplicons were longer than the 150 base pairs read, simply joining the reads in the correct orientations results in a large apparent deletion in the middle of sequence relative to the actual 16S sequence. This resulted in poor performance for sequence searching but could likely be addressed by using an alignment-based search tool with an alternative scoring system. However, a comparison of PCoA on unweighted Unifrac distance matrices shows that the conclusions reached when looking at the 5′ and 3′ reads independently are essentially identical in both community structure and community membership, as illustrated by Procrustes analysis on the first three principal coordinates ($M^2 = 0.023$, $P < 0.001$ for unweighted UniFrac; $M^2 = 0.021$, $P < 0.001$ for weighted UniFrac; Fig. 5). This suggests that if sequencing a well-chosen region of the 16S gene, 75–100 bases may suffice for drawing reliable conclusions from the data, circumventing the need for paired-end reads. Although this conclusion might seem surprising, the success of early phylogenetic studies using the 5S rRNA, which is typically less than 150 nucleotides, suggests a precedent (32).

The Illumina platform returns on the order of 100 million sequencing reads per flowcell and could thus potentially support either comparison of thousands of barcoded samples with thousands of sequences per sample, fewer barcoded samples with higher depth of coverage, or the simultaneous analysis of many markers for phylogenetic and/or functional genes. We could now, for example, contemplate analyzing samples from comprehensive time series, to quantify microbial community dynamics across many sites, or producing detailed 3D maps of microbial communities in environments ranging from soils to the mammalian gut. The increased sample analysis capacity made possible with the approach described here will make it feasible to address a vast range of questions that would be impossible to address using previous generations of sequencing technology. For example, we can now explore, in detail, whether changes in rare or abundant species are primarily responsible for differences in microbial communities associated with health and disease. Similarly, we can begin exploring the ecological characteristics of even rare microbial taxa (for example, novel uncultivated phyla) by examining how changes in environmental conditions influence the structure of microbial communities across time and space in a wide range of environments.

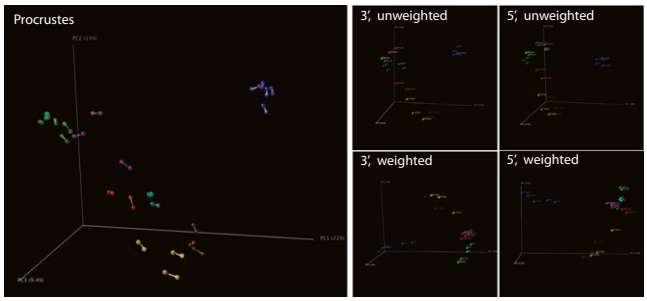

Fig. 5. PCoA of the samples using sequences from each region. Samples are feces (blue), freshwater creek (bright green), freshwater lake (red), ocean (cyan), sediment (pink), skin (yellow), soil (dark green), and tongue (dark red). In the Procrustes analysis, using all reads, the samples derived from the 5′ end and the 3′ end are linked with a bar: in every case, the distance between the 5′ and 3′ reads of the same samples is much smaller than the distance between samples, highlighting the robustness of UniFrac analysis relative to the taxonomic analysis shown in Fig. 2. The smaller panels, using only 2,000 randomly chosen sequences per sample, show the weighted and unweighted UniFrac results from the 5′ and 3′ reads individually: the pattern of samples is highly reproducible (note that the direction of each axis is arbitrary, only the relative position of the points matters rather than whether a particular sample appears to the left or the right of the plot). As seen in ref 17, axis 1 is host associated/free living and axis 3 is saline/nonsaline environment.

Methods

Datasets. Several independently compiled sample collections were included in this analysis. These include samples from human feces ($n = 3$), skin ($n = 3$), and the dorsal tongue surface ($n = 3$) (1); fecal samples from human twins (31) ($n = 2$); soil ($n = 3$) (5, 34); freshwater and freshwater sediment ($n = 5$), ocean ($n = 3$), and marine sediment ($n = 3$) samples; and "mock community" samples ($n = 3$) from genomic DNA isolated from 67 bacterial strains and pooled at even abundances (26). Counts of the number of reads associated with each sample are provided in Table S3.

All individuals were made aware of the nature of the experiment and gave written informed consent in accordance with the sampling protocol approved by the University of Colorado Human Research Committee (protocol 0708.12), except for the twin specimens, which were repurposed from a previous institutional review board–approved study at Washington University (33).

Data Analysis. The QIIME software package was applied to analyze the results of this run. Because of the number of reads, several significant performance enhancements were made to QIIME and contributed back to the open-source project.

Primers. Five primers, two for PCR and three for sequencing, were developed for this analysis (Fig. 1). The PCR primers (F515/R806) were developed against the V4 region of the 16S rRNA, which we determined would yield optimal community clustering with reads of this length using a procedure similar to that of ref. 15. [For reference, this primer pair amplifies the region 533–786 in the *Escherichia coli* strain 83972 sequence (greengenes accession no. prokMSA_id:470367).] The reverse PCR primer is barcoded with a 12-base error-correcting Golay code to facilitate multiplexing of up to ≈1,500 samples per

lane, and both PCR primers contain sequencer adapter regions. The three sequencing primers include two for reading in from each end of the amplicon and a third for reading the barcode. Because of technical limitations at the sequencing facility, only part of the barcode was sequenced, so we were unable to exploit the error-correcting properties fully; however, even with partial barcodes we were able to resolve the samples, demonstrating the robustness of the approach. It is important to note that this primer collection allows for sequencing of paired-end reads, but the downstream data analyses are not yet capable of supporting paired-end reads. Our results illustrate interesting and correlated patterns based on analysis of the unpaired reads (i.e., α and β diversity evaluations based on the 5′ only and 3′ only reads independently achieve similar results, suggesting that 100 bases in this region of the 16S gene can allow for successful screening and comparison of microbial communities). The reads generated from these PCR primers are both identified as "recommended" regions by Liu et al. (15).

Polymerase Chain Reaction. Sample preparation was performed similarly to that described by Costello et al. (1). Briefly, each sample was amplified in triplicate, combined, and cleaned using the MO BIO 96 htp PCR clean up kit. PCR reactions contained 13 μL MO BIO PCR water, 10 μL 5 Prime Hot Master Mix, 0.5 μL each of the forward and reverse primers (10 μM final concentration), and 1.0 μL genomic DNA. Reactions were held at 94°C for 3 min to denature the DNA, with amplification proceeding for 35 cycles at 94 °C for 45 s, 50 °C for 60 s, and 72 °C for 90 s; a final extension of 10 min at 72 °C was added to ensure complete amplification. Cleaned amplicons were quantified using Picogreen dsDNA reagent in 10 mM Tris buffer (pH 8.0). A composite sample for sequencing was created by combining equimolar ratios of amplicons from the individual samples, followed by gel purification and ethanol precipitation to remove any remaining contaminants and PCR artifacts. The sample, along with aliquots of the three sequencing primers, was sent to Illumina for sequencing.

Quality Filtering of Reads. The quality scores associated with each base call for each read were used to determine the portion of each read that was of acceptable quality. The 100 base reads were truncated when they achieved two or more consecutive base calls with quality scores below 1e-5. Truncation was applied to include the bases through the last position that achieved a quality score of greater than 1e-5. For a read to be included in downstream analyses, it was required to have a minimum length of 75 bases, and truncated reads were discarded if they contained an N character in their sequence or barcode. This filtering step reduced the 87,507,177 paired-end reads to 36,329,392 5′ reads and 22,177,779 3′ reads. The quality filtering parameters used here were determined empirically, and details on other parameter settings are provided in Table S1. Development of an error model for the Illumina platform is in order, and we expect that it would improve the results, as has been shown for the 454 platform.

OTU Picking. To facilitate OTU picking on so many sequences, OTUs were chosen in a two-step process. First, sequences were clustered into OTUs using the Trie algorithm, which groups sequences that are exact prefixes of another sequence. Representative sequences were then selected on the basis of these "Trie-picked" OTUs. The second step of OTU picking was performed by BLASTing the representative sequences against a reference database. A filtered version of the greengenes database was used as the reference database, and sequences were clustered by their representative BLAST match. Representative BLAST matches were chosen on the basis of three criteria. First, the BLAST match must achieve an E-value less than 1e-10. Next, the percent sequence identity of the alignment between a BLAST match and the read must be greater than or equal to the OTU selection threshold (0.97 here, corresponding to species-like OTUs). Of the remaining BLAST matches, the match that achieves the longest alignment to the read is chosen as the representative BLAST match. For each resulting OTU, a representative sequence was chosen as the most abundant postquality filtering read.

Taxonomy Assignment. Taxonomy for the environmental sequences was assigned to the representative sequence of each OTU using QIIME's parallel wrappers for the RDP classifier. The most detailed taxonomic level assigned to an OTU's representative sequence at confidence of greater than or equal to 0.80 was taken as the taxon of the OTU.

Taxonomy assignments for the mock community alone were generated by BLASTing the mock5 and mock3 representative sequences against the known full-length 16S sequences from the mock community to assign taxonomy to each representative sequence. Each representative sequence was assigned the taxonomy of the best BLAST hit with an E-value less than 0.001, and the relative abundances of each taxon were then computed for each mock community sample at each taxonomic level.

Comparisons of taxonomic assignments between the mock5, mock3, and known full-length 16S sequences data were performed by computing quantitative Bray-Curtis distances within and between the different samples, where total sequence counts were normalized to 10,000 sequences per sample. Bray-Curtis "categories" were defined by taxonomic classification at the order and genus levels. Distributions of Bray-Curtis distances were compared with two-tailed, two-sample *t* tests.

Alpha Rarefaction and Beta Diversity. Alpha rarefaction was performed using the Phylogenetic Diversity, Chao1, and observed species metrics. Ten sampling repetitions were performed, without replacement, at each sampling depth, and the error bars in Fig. 3 indicate the range of α diversity scores achieved at a given sampling depth.

Beta diversity was estimated by computing weighted and unweighted UniFrac distances between samples using QIIME. Samples were clustered based on their between-samples distances using UPGMA, and jackknifing was performed by resampling 100 times with replacement at a depth of 100,000 sequences per sample.

To compare the results of the env5 and env3 datasets, principal coordinates analysis was applied to reduce the dimensionality of the resulting distance matrices. The patterns of β diversity arising from the env5 and env3 were compared using Procrustes analysis. Principal coordinates were transformed with Procrustes using PyCogent to facilitate direct comparison of trends in community structure. Additionally, Monte Carlo simulations were performed by shuffling each set of coordinates in one of the principle coordinates matrices and recomputing M^2 to estimate the probability of seeing a pair of coordinate matrices with an M^2 value as high or higher than the actual M^2 value achieved by Procrustes analysis. M^2 and P values presented herein are based on the first three principal coordinates only for each sample, although similar results were found when looking at all coordinates (Table S4).

To evaluate the ability of the GAIIx platform to recapture similar between-community patterns when applied at a much lower sampling depth (e.g., when including more barcoded samples), 2,000 sequences were randomly sampled without replacement from each sample in the env5 and env3 datasets. Weighted and unweighted UniFrac were applied to these datasets, followed by PCoA, as was done for the full env5 and env3 datasets. The PD, Chao1, and observed OTU scores are provided in Table S5, along with variants on the expected scores computed from the expected reads only, the expected amplicons only, and the full-length control sequences.

ACKNOWLEDGMENTS. We thank Eric Roden (University of Wisconsin-Madison), Jennifer Martiny (University of California, Irvine), Trina McMahon (University of Wisconsin-Madison), Steve Allison (University of California, Irvine), Jeffrey I. Gordon (Washington University), and Adam Martiny (University of California, Irvine) for graciously providing us with DNA from their environmental samples; Craig Pierson, Jeremy Pierce, Courtney McCormick, and Christian Haudenschild at Illumina for coordinating and conducting the sequencing run described here; the Illumina Corporation for donating the run itself; and Elizabeth Costello, Dan Knights, Justin Kuczynski, and Jesse Stombaugh for feedback on drafts of the manuscript. This work was funded with grants from the US Department of Agriculture and the National Science Foundation (to N.F. and R.K.), from the National Institutes of Health, the Bill and Melinda Gates Foundation, and the Crohn's and Colitis Foundation of America (to R.K.), and the Howard Hughes Medical Institute.

1. Costello EK, et al. (2009) Bacterial community variation in human body habitats across space and time. *Science* 326:1694–1697.
2. Grice EA, et al.; NISC Comparative Sequencing Program (2009) Topographical and temporal diversity of the human skin microbiome. *Science* 324:1190–1192.
3. Roesch LFW, et al. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* 1:283–290.
4. Sogin ML, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci USA* 103:12115–12120.
5. Lauber CL, 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. *Appl Environ Microbiol* 75:5111–5120.
6. Tringe SG, Hugenholtz P (2008) A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* 11:442–446.
7. Pace NR (1997) A molecular view of microbial diversity and the biosphere. *Science* 276: 734–740.
8. Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* 102:2567–2572.

9. Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res* 19:1141–1152.
10. Pushkarev D, Neff NF, Quake SR (2009) Single-molecule sequencing of an individual human genome. *Nat Biotechnol* 27:847–852.
11. Lazarevic V, et al. (2009) Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods* 79:266–271.
12. Binladen J, et al. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2:e197.
13. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 5:235–237.
14. Huber JA, et al. (2007) Microbial population structures in the deep marine biosphere. *Science* 318:97–100.
15. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* 35:e120.
16. Lozupone CA, Knight R (2007) Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* 104:11436–11440.
17. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI (2008) Worlds within worlds: Evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* 6:776–788.
18. Auguet JC, Barberan A, Casamayor EO (2010) Global ecological patterns in uncultured Archaea. *ISME J* 4:182–190.
19. Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*, 10.1038/nmeth.f.303.
20. Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.
21. Hamady M, Lozupone C, Knight R (2010) Fast UniFrac: Facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* 4:17–27.
22. Fredkin E (1960) Trie memory. *Commun ACM* 3:490–499.
23. Price MN, Dehal PS, Arkin AP (2009) FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 26:1641–1650.
24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
25. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267.
26. Turnbaugh PJ, et al. (2010) Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci USA* 107: 7503–7508.
27. Quince C, et al. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 6:639–641.
28. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* 12:118–123.
29. Lozupone C, Knight R (2005) UniFrac: A new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71:8228–8235.
30. Ley RE, et al. (2008) Evolution of mammals and their gut microbes. *Science* 320: 1647–1651.
31. Lozupone CA, Hamady M, Kelley ST, Knight R (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* 73:1576–1585.
32. Fox GE, Woese CR (1975) The architecture of 5S rRNA and its relation to function. *J Mol Evol* 6:61–76.
33. Turnbaugh PJ, et al. (2009) A core gut microbiome in obese and lean twins. *Nature* 457:480–484.
34. Fierer N, Jackson RB (2006) The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* 103:626–631.