

ORIGINAL ARTICLE

Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data

Micah Hamady^{1,5}, Catherine Lozupone^{2,3,5} and Rob Knight^{2,4}

¹Department of Computer Science, University of Colorado, Boulder, CO, USA; ²Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO, USA; ³Center for Genome Sciences, Washington University School of Medicine, St Louis, MO, USA and ⁴Howard Hughes Medical Institute, Chevy Chase, MD, USA

Next-generation sequencing techniques, and PhyloChip, have made simultaneous phylogenetic analyses of hundreds of microbial communities possible. Insight into community structure has been limited by the inability to integrate and visualize such vast datasets. Fast UniFrac overcomes these issues, allowing integration of larger numbers of sequences and samples into a single analysis. Its new array-based implementation offers orders of magnitude improvements over the original version. New 3D visualization of principal coordinates analysis results, with the option to view multiple coordinate axes simultaneously, provides a powerful way to quickly identify patterns that relate vast numbers of microbial communities. We show the potential of Fast UniFrac using examples from three data types: Sanger-sequencing studies of diverse free-living and animal-associated bacterial assemblages and from the gut of obese humans as they diet, pyrosequencing data integrated from studies of the human hand and gut, and PhyloChip data from a study of citrus pathogens. We show that a Fast UniFrac analysis using a reference tree recaptures patterns that could not be detected without considering phylogenetic relationships and that Fast UniFrac, coupled with BLAST-based sequence assignment, can be used to quickly analyze pyrosequencing runs containing hundreds of thousands of sequences, showing patterns relating human and gut samples. Finally, we show that the application of Fast UniFrac to PhyloChip data could identify well-defined subcategories associated with infection. Together, these case studies point the way toward a broad range of applications and show some of the new features of Fast UniFrac.

The ISME Journal (2010) 4, 17–27; doi:10.1038/ismej.2009.97; published online 27 August 2009

Subject Category: microbial population and community ecology

Keywords: beta diversity; community ecology; multiplex pyrosequencing of 16S rDNA; PhyloChips

Introduction

Understanding beta diversity is critical for studies of microbial ecology because of the enormous variation among microbial communities even when those communities are sampled from similar environment types (Lozupone and Knight, 2007). In contrast to alpha diversity, which measures how many kinds of organism are in a single community, beta diversity measures how community membership varies over time or space, and is especially important for finding trends in large numbers of samples (a

problem that significance tests for differences between each pair of communities differs cannot address). For example, Human Microbiome Projects (Turnbaugh *et al.*, 2007) and related efforts to study microbial communities occupying various human body habitats are showing a surprising amount of diversity among individuals in skin (Fierer *et al.*, 2008; Grice *et al.*, 2008), gut (Turnbaugh *et al.*, 2009), and mouth ecosystems (Nasidze *et al.*, 2009). As all current methods of surveying microbial communities using culture-independent methods introduce inherent biases in DNA extraction and/or amplification of small subunit rRNA genes, patterns that relate different communities may be more meaningful than estimates of diversity or of taxon abundance within a single community (Kanagawa, 2003). Measures of beta diversity can be either taxon-based (using overlap in lists of species, genera, OTUs, and so on) or phylogenetic (using

Correspondence: R Knight, Department of Chemistry and Biochemistry, University of Colorado, Boulder, UCB 215, CO 80309, USA.

E-mail: rob.knight@colorado.edu

⁵These authors contributed equally to this work.

Received 9 June 2009; revised 30 July 2009; accepted 30 July 2009; published online 27 August 2009

overlap on a phylogenetic tree). Phylogenetic beta diversity measures, such as UniFrac (Lozupone and Knight, 2005; Lozupone *et al.*, 2006), are especially important because, unlike taxon-based measures, they exploit the similarities and differences among species (Graham and Fine, 2008; Lozupone and Knight, 2008). This additional information makes phylogenetic beta diversity measures more effective at showing ecological patterns than taxon-based methods (Lozupone and Knight, 2008).

Considerable insight has been gained from applying beta diversity methods to microbes in different environments. For example, to date >70 papers have used UniFrac to compare microbial assemblages. These include bacterial (Rawls *et al.*, 2006; Hartman *et al.*, 2008; Hsu and Buckley, 2009), archaeal (Harrison *et al.*, 2009), eukaryotic (Porter *et al.*, 2008; Alexander *et al.*, 2009), and viral (Desnues *et al.*, 2008; Marhaver *et al.*, 2008) assemblages important for understanding human health and disease (Turnbaugh *et al.*, 2006; Frank *et al.*, 2007; Li *et al.*, 2008; Osman *et al.*, 2008; Wen *et al.*, 2008), bioremediation (Hiibel *et al.*, 2008), and basic ecology and evolution (Fraune and Bosch, 2007; Balakirev *et al.*, 2008; Bryant *et al.*, 2008). Applications of UniFrac have focused both on 16S rRNA sequence sets from Sanger sequencing and pyrosequencing (Fierer *et al.*, 2008; Hamady *et al.*, 2008; Turnbaugh *et al.*, 2009) and on sequences from genes with other functions (Elifantz *et al.*, 2008; Lozupone *et al.*, 2008; Hsu and Buckley, 2009; Lauber *et al.*, 2009).

Despite the clear advantages of phylogenetic beta diversity approaches, the challenges inherent in building and analyzing trees with thousands to millions of sequences have thus far limited the broad application of these techniques. For example, many recent pyrosequencing studies in a range of environments have used taxon-based methods to compare samples (Sogin *et al.*, 2006; Huber *et al.*, 2007; Roesch *et al.*, 2007), primarily because of these challenges. Similarly, to our knowledge, no PhyloChip studies have yet exploited phylogenetic beta diversity techniques, despite the potential of the PhyloChip to collect data from dozens or hundreds of samples in a cost-effective manner. Here, we make these techniques available to the broader community by presenting Fast UniFrac, a much faster version of UniFrac, which allows analysis of much larger datasets. In addition to a stand-alone version, the online version includes more advanced visualizations to facilitate rapid identification of patterns in large and complex datasets. These visualizations include 3D views of any combination of the first 10 principal coordinates, and parallel coordinates plots that plot the position of each sample along each of the first 10 principal coordinates, showing which coordinates discriminate among groups of samples. Parallelization of the resampling techniques, such as jackknifing, makes it more feasible to test whether particular clusters are robust to sampling effort. Together with the ability to accept pyrosequencing

and PhyloChip datasets as input, Fast UniFrac should greatly expand our insight into a wide range of microbial processes.

Materials and methods

Performance enhancements

With the goal of supporting very large datasets, including pyrosequencing and PhyloChip datasets produced in association with the Human Microbiome Project, we have redesigned UniFrac so that calculations on the phylogenetic tree are performed using an array-based implementation instead of a tree-based one. In the original implementation of UniFrac, environment data are stored associated with objects representing the nodes of a tree. To calculate the UniFrac value for a specific comparison, it is necessary to traverse the tree, assign the states (environments) for the internal nodes based on presence/absence (unweighted) or the sum of the counts (weighted) in the child nodes, and traverse the tree once again to perform the calculations. In the new implementation, we store the environment states in an array, and use accelerated vector operations in the numpy package to propagate states down the tree and to multiply the states by the branch lengths (Figure 1) (in addition, we cache the tree structure implicitly using a nested list of arrays for speed). There are several advantages to this new approach: (i) environment states can be propagated using the cache, which is much faster than using custom tree objects; (ii) by using logical and numerical operations, the whole array of environments or specific pairs of environments, can be isolated as array slices, saving the expensive traversal step; (iii) the tree does not need to be pruned for branches absent from the chosen pair of environments because the branch lengths for those branches get multiplied by 0 (being absent from all environments) and do not contribute to the overall result; and (iv) because the array of counts of each sequence in each environment is contiguous, jackknifing can be performed rapidly. This re-conceptualization also leads to potential future improvements, such as using MPI or other parallelization toolkits and/or GPUs to accelerate the comparisons further. The array-based implementation also uses far less memory and storage space than the tree-based implementation, allowing the same hardware to process much larger datasets. Finally, parallelization of the Monte Carlo operations such as the *P*-test (Martin, 2002) and sequence jackknifing greatly improves the performance of significance tests, and allows larger numbers of replicates so that *P*-values for rarer events can be estimated. These speed enhancements produce the same final result, but have allowed us to increase the default limits from 5000 unique sequences, 50 samples, and 100 permutations in the original UniFrac web application to 100 000 unique

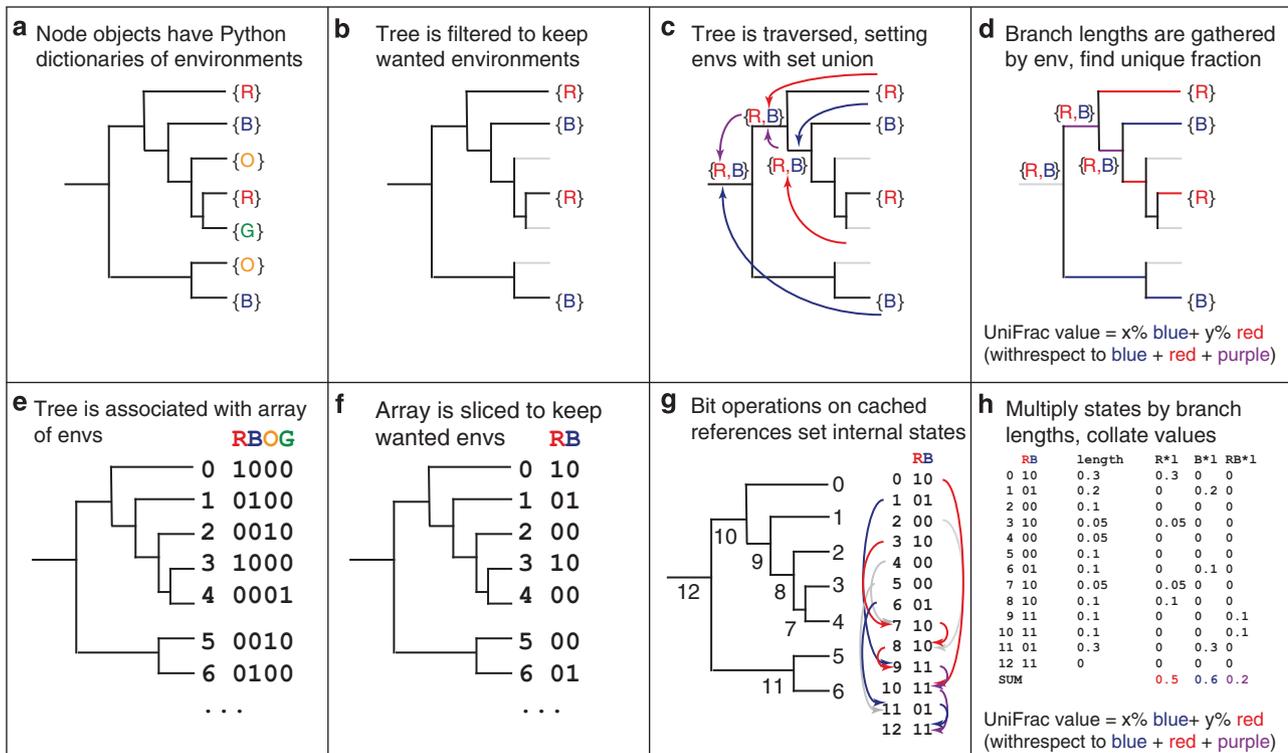


Figure 1 Difference in procedure between the original UniFrac and the new Fast UniFrac (for clarity, only the unweighted UniFrac algorithm is shown here, but similar principles apply to weighted UniFrac). In the original procedure, (a) environments are stored as sets in a tree object, (b) the tree is pruned to include only the branches leading to wanted environments, (c) the sets of environments are compared using set algorithms, states are assigned to each internal node, and (d) the result is calculated by another tree traversal. In the new procedure, (e) the environments are stored as an array of tip x environment counts, (f) selected environments are chosen by slicing this array, (g) internal states are calculated using array operations on slices of the array, and (h) the products of the incidence array and the branch lengths of nodes leading to either or both of the environments are summed, allowing calculation of the UniFrac value. The array-based approach allows substantial gains in efficiency.

sequences, 200 samples, and 1000 permutations in the Fast UniFrac web interface.

New features

BLAST-based phylogeny generation. The application of UniFrac to large sequence sets, such as those generated with pyrosequencing, is also limited by the computational power needed to make a *de novo* phylogenetic tree using standard methods, such as neighbor joining, likelihood, or parsimony methods. We show below that the analysis of such large sequence sets is possible by assigning them to their closest relative in a phylogeny of the Greengenes core set (DeSantis *et al.*, 2006) using BLAST's megablast protocol (Altschul *et al.*, 1990). The Greengenes core set reference tree is given as a drop down menu option during upload of data, and a detailed protocol and python script has been provided in the Fast UniFrac tutorial for the generation of a BLAST-based sample mapping file that corresponds to the Greengenes core set or any other reference tree.

Visualization enhancements

As the size and complexity of microbial datasets rapidly increase, so does the difficulty associated

with interpreting the results and identifying ecologically meaningful patterns. New ways of exploring and visualizing results are thus essential. Fast UniFrac introduces several powerful tools to assist in visualizations of the results of principal coordinates analysis (PCoA), such as in 3D using the Java KiNG viewer (<http://kinemage.biochem.duke.edu/software/king.php>). These tools include (i) the ability to color large collections of samples using different user-defined subcategories (for example, coloring environmental samples according to temperature or pH), (ii) automatic scaled/unscaled views, which accentuate dimensions that explain more variance, (iii) the ability to interactively explore hundreds of points (and user-configurable labels) in 3D, (iv) parallel coordinates displays that allow the dimensions that separate particular groups of environments to be readily identified, and (v) scree plots that help researchers more easily discern the number of important dimensions and thus assist in inferring biological significance in complex datasets (Zhu and Ghodsi, 2006).

PhyloChip support

Another new feature is support for PhyloChip data (Wilson *et al.*, 2002; DeSantis *et al.*, 2007) using the

UniFrac export option of the PhyloTrac software (<http://phylotrac.org>). In the PhyloChip interface, a reference tree allows the comparison of multiple PhyloChip runs: all that is required is a combined mapping file containing abundance information from all of the PhyloChip samples, together with an additional mapping file relating each sample to study meta-data.

Usability

Finally, we added important usability enhancements that allow multiple user-defined category mappings to be uploaded, along with sample descriptions that permit easier and more rapid exploration of the dataset broken down by a range of different parameters or categories. For example, one might want to color a set of mammalian gut samples by diet, by species, by taxonomic order, by continent of origin, and so on to determine which factors were most important in structuring the communities. The ‘category mapping’ file can also be automatically generated in the Fast UniFrac web interface. When this option is selected, an example category mapping file is generated with a single real subcategory called *Envs* containing values identical to the *sample IDs* provided in the *sample ID mapping file*. In addition to the real subcategory, several placeholder subcategories are created that act as a template for users when the file is downloaded and modified for future runs. Error checking and error correction for problems with the input trees and other input data has been substantially expanded, and numerous other performance-related optimizations substantially accelerate the overall workflow.

Sources of data

Data for testing and validation of Fast UniFrac came from four main sources: (1) a large meta-analysis of Sanger-sequencing data from a wide range of different host-associated and free-living environments (Ley *et al.*, 2008b); (2) an analysis of how gut bacterial populations change in obese humans on fat-restricted and carbohydrate-restricted diets (Ley *et al.*, 2006); (3) pyrosequencing studies of the human hand (Fierer *et al.*, 2008), and of fecal microbiota of lean and obese twin pairs and their mothers (Turnbaugh *et al.*, 2009); and (4) a PhyloChip study of citrus pathogens (Sagaram *et al.*, 2009). These studies were chosen as they represent some of the largest datasets for their respective types of analyses. A reference tree was assembled from the Greengenes core set (DeSantis *et al.*, 2006): both this tree and the PhyloChip G2 reference tree are available from the Fast UniFrac web site.

Phylogenetic methods

The application of UniFrac to large datasets, such as those generated by pyrosequencing, has been lim-

ited by the ability to make *de novo* trees using standard tree building methods. Although programs such as ARB’s parsimony insertion algorithm (Ludwig *et al.*, 2004) have been used to analyze datasets with almost 100 000 sequences (Ley *et al.*, 2008b), this technique is very time consuming, and cannot be automated or enhanced by parallelization on high performance clusters for the larger datasets that pyrosequencing produces. We show that using BLAST’s (Altschul *et al.*, 1990) megablast method to find the nearest neighbor of each short read in an existing library (in this case the Greengenes core set), recaptures the same patterns detected using the parsimony insertion method of ARB, and that these methods can be applied to pyrosequencing data with hundreds of thousands of sequences. The method of BLASTing sequence reads to an existing phylogeny can be extended to work with any gene and any existing phylogeny.

Megablast protocol

The Greengenes core set was downloaded from (http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/11-Aug_2007) and made into a BLAST database using formatdb. The Global Environment dataset (Ley *et al.*, 2008b)(99 801 sequences), the human obesity dataset (Ley *et al.*, 2006)(18 348 sequences), and all unique pyrosequences from studies of the human hand, and the fecal microbiota of lean and obese twins (Fierer *et al.*, 2008; Turnbaugh *et al.*, 2009) (232 165 unique sequences from 680 000 initial reads) were then searched against the Greengenes core set using megablast. The hit tables were parsed to make sample ID mapping files, in which each sequence was mapped to its closest hit in the core set. Query sequences that had no hit below an e-value threshold of either $1e-50$ or $1e-30$ were excluded from the analysis (255 sequences were excluded with a $1e-50$ criterion for the Global Environment dataset and 4789 unique sequences for the human pyrosequencing datasets. No sequences were excluded from the human obesity dataset with a $1e-30$ criterion). A script for performing this analysis is available in the tutorial at the Fast UniFrac web site.

A tree containing the same set of sequences as in the Greengenes core set FASTA file was obtained by downloading the most recent ARB database available at the Greengenes site (http://greengenes.lbl.gov/Download/Sequence_Data/arb_databases/greengenes236469.arb.gz). The database is annotated with a ‘coreset’ field, and searching for sequences with value ‘1’ in that field produced a list approximating the core set. As the overlap with the core set FASTA file was imperfect (both extra and missing sequences), the missing sequences in the core set FASTA file were added using ARB’s parsimony insertion, and then extra sequences were marked and pruned from the tree. The resulting reference tree, that we call ‘Greengenes Core’ is available for download from the tutorial and as a drop down

menu option in the Fast UniFrac web site. In addition, to assess the impact that accounting for phylogenetic relationships, as opposed to shared 'best hit' information alone, had on the results, we also performed analyses on the Greengenes tree represented as a 'star phylogeny,' which was produced by attaching all sequences in the core set to a root node with a branch length of 1.

Category mapping files were created and the data analyzed through the Fast UniFrac web interface. The category mapping allows for the samples to be grouped by any number of criteria for coloration and dynamic visualization of PCoA analysis results in the 3D visualization using the Java KiNG viewer (<http://kinemage.biochem.duke.edu/software/king.php>). Sample and experiment descriptions were also added in this file that are displayed on the samples throughout the interface upload and results pages, aiding in results interpretation.

Global environment ARB parsimony insertion protocol
Sequences were parsimony inserted into the Greengenes core set in ARB as described earlier (Ley *et al.*, 2008b). In this analysis, the sequence sets from each sample were dereplicated by the DivergentSet method (Widmann *et al.*, 2006) and only one divergent sequence from each sample was used. The environment file from the original analysis was edited so that the sample names conformed to the Fast UniFrac interface conventions (for example, to remove underscores and other characters with special meanings in the Fast UniFrac web interface). The resulting file was analyzed using the Fast UniFrac web interface, using the same category mapping file as for the megablast to Greengenes dataset.

PhyloChip/PhyloTrac protocol

PhyloTrac was downloaded from <http://phylotrac.org>. The CEL data and PhyloTrac thresholds were obtained from a previously published study (Sagarin *et al.*, 2009), in which microbial communities from citrus trees infected with the Huanglongbing pathogen and controls were assessed by PhyloChip and reanalyzed. A Fast UniFrac sample ID mapping file (environment file) was exported from PhyloTrac and uploaded to Fast UniFrac. The G2 PhyloChip was selected as the reference tree, and the category mapping file auto-generated. This file was then downloaded and modified to use the same categories as in the paper, using metadata kindly provided by the authors. Finally, the results were analyzed using the Fast UniFrac web interface.

Availability

The Fast UniFrac Python code is now available in the 1.3 release of the open-source PyCogent package (Knight *et al.*, 2007), available at <http://sourceforge.net/projects/pycogent> in the cogent/maths/unifrac directory, and the web interface is available at <http://www.bmf.colorado.edu/fastunifrac>.

Results

Comparing the ARB and BLAST protocols using the global environmental survey dataset

The ARB parsimony insertion protocol and the megablast protocol gave similar results for the global environment survey, both at the broad level and in detail (Figures 2a, b, d). The amount of variation explained by the principal axes is about the same (PC1 is 7.3% for ARB, 10.0% for BLAST) and the pairwise UniFrac distances between samples were highly correlated for the two protocols (Figure 2d). Perhaps more importantly, the overall clustering patterns are very similar and would yield the same ecological inferences. Samples from the vertebrate gut (blue) clearly separated from free-living environments (magenta, green) along PC1, with the termite gut (orange) and human mouth and skin (particularly from the vulva) (red) having intermediate values. Free-living assemblages separated into saline (magenta) vs non-saline (green) environments along PC3, with mixed habitats (grey) such as estuaries intermediate between the two. In contrast, the results of using megablast to the Greengenes coresets, but using a star phylogeny instead of the core set phylogeny, looked quite different. The amount of variation explained by PC1 is less (4.3%) and the clustering forms a star pattern with less clear separation between samples and environments as in the other two methods. The pairwise UniFrac distances between samples for the star phylogeny and the ARB parsimony insertion protocol were far less correlated (Figure 2e). Overall this shows that the 'megablast to the Greengenes core set' protocol is a good alternative to ARB parsimony insertion for making a tree because it produces essentially the same result in a dataset in which accounting for phylogenetic relationships affects the results.

Fifty of the 464 samples in the global environment dataset were also subsampled to provide a simpler example for the tutorial at the Fast UniFrac website and to test the robustness of the conclusions to the number of samples used. Despite using only ~10% of the samples, the same major patterns emerged, with PC1 again separating the vertebrate gut from free-living samples and the termite gut intermediate. Salinity was again an important factor, with saline water separating from non-saline soils and sediment along PC2. This subset shows the robustness of the global environment survey result, and also provides an example dataset for exploring the functionality of the web interface. As PCoA results can be affected by the number of samples from different groups in the study, redoing the

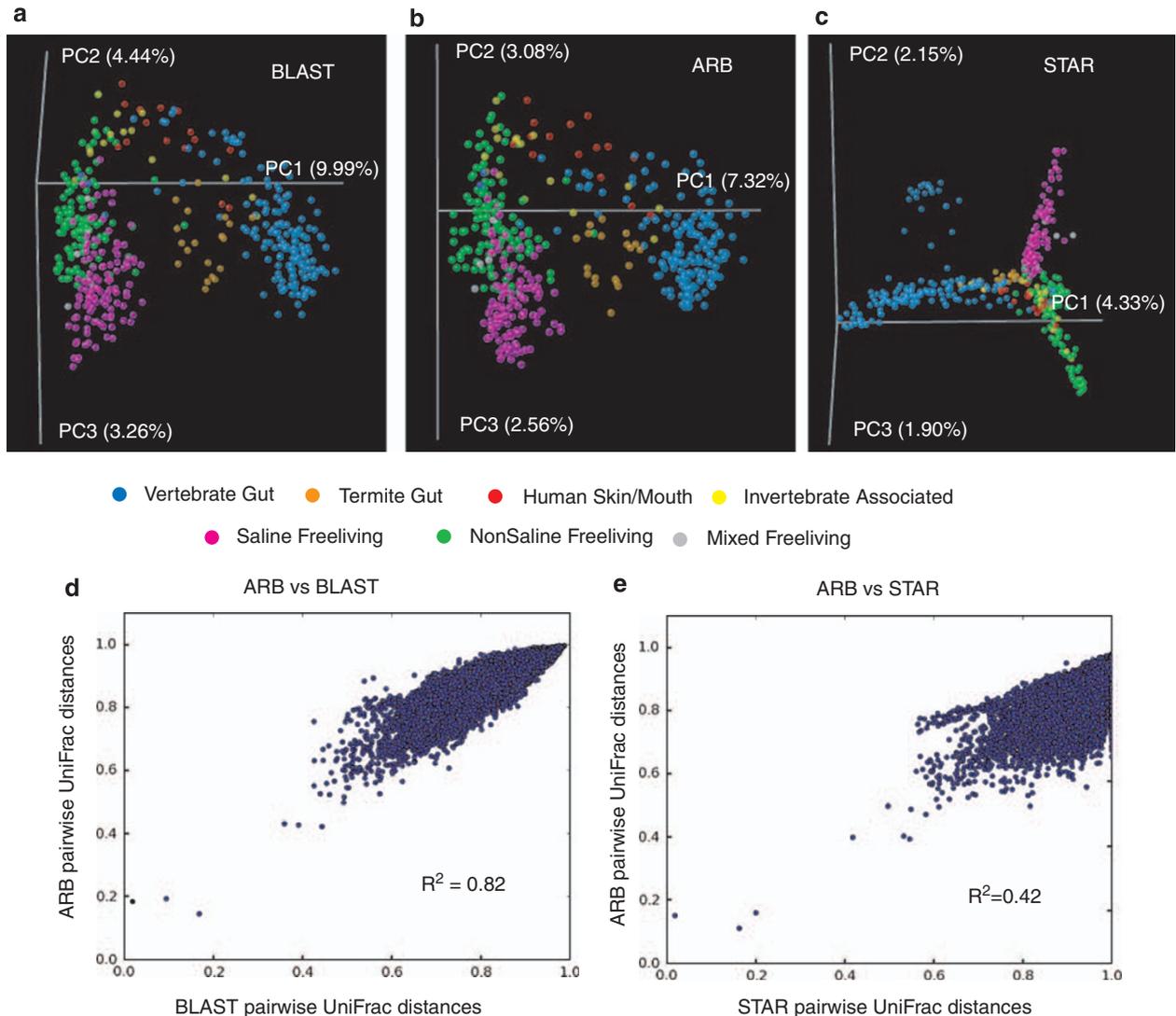


Figure 2 Global Environmental Survey dataset (Ley *et al.*, 2008b) analyzed using PCoA of unweighted pairwise UniFrac distances with trees generated using (1) megablast mapping to the Greengenes core set tree (a), (2) an ARB parsimony insertion tree (b), and (3) megablast mapping to the Greengenes core set represented as a star phylogeny (that is a phylogeny in which all taxa are treated as equally related, ignoring the actual phylogenetic information) (c). All plots show the first three principal axes as visualized in the 3D viewer. Scatterplots of the pairwise UniFrac distances (d and e), as well as the PCoA analysis, show that megablast to the Greengenes core set produced similar results as ARB parsimony insertion, but only when the phylogenetic relationships in the Greengenes core set are considered.

analysis with random subsets of samples is a good way to test the robustness of the results.

Comparing the ARB and BLAST protocols using the human obesity dataset

The global environment dataset contained samples from extremely different environments. However, UniFrac is also useful for exploring closely related samples. We thus also tested an example dataset consisting of closely related microbial communities to illustrate that the resolution of the megablast protocol is sufficient for the dynamic monitoring of the same community over time. We repeated the UniFrac analysis reported in Figure 1a of Ley *et al.* (2006). Here, Ley *et al.* sequenced the bacteria in

stool samples from 11 obese individuals who followed either a fat-restricted ($n=5$) or carbohydrate-restricted ($n=6$) diet for 3–4 timepoints over the course of a year. Hierarchical clustering based on UniFrac analysis of an ARB parsimony insertion tree showed that the bacterial lineages were remarkably constant within individuals over time, because samples from the same person generally clustered with each other rather than with samples from other people (Ley *et al.*, 2006). Repeating this analysis with the megablast to greengenes protocol and Fast UniFrac as described above yielded trees that differed somewhat in the details of the topology, but for which the samples clustered equally well by individual (see Supplementary information). Thus, the megablast protocol provides sufficient resolu-

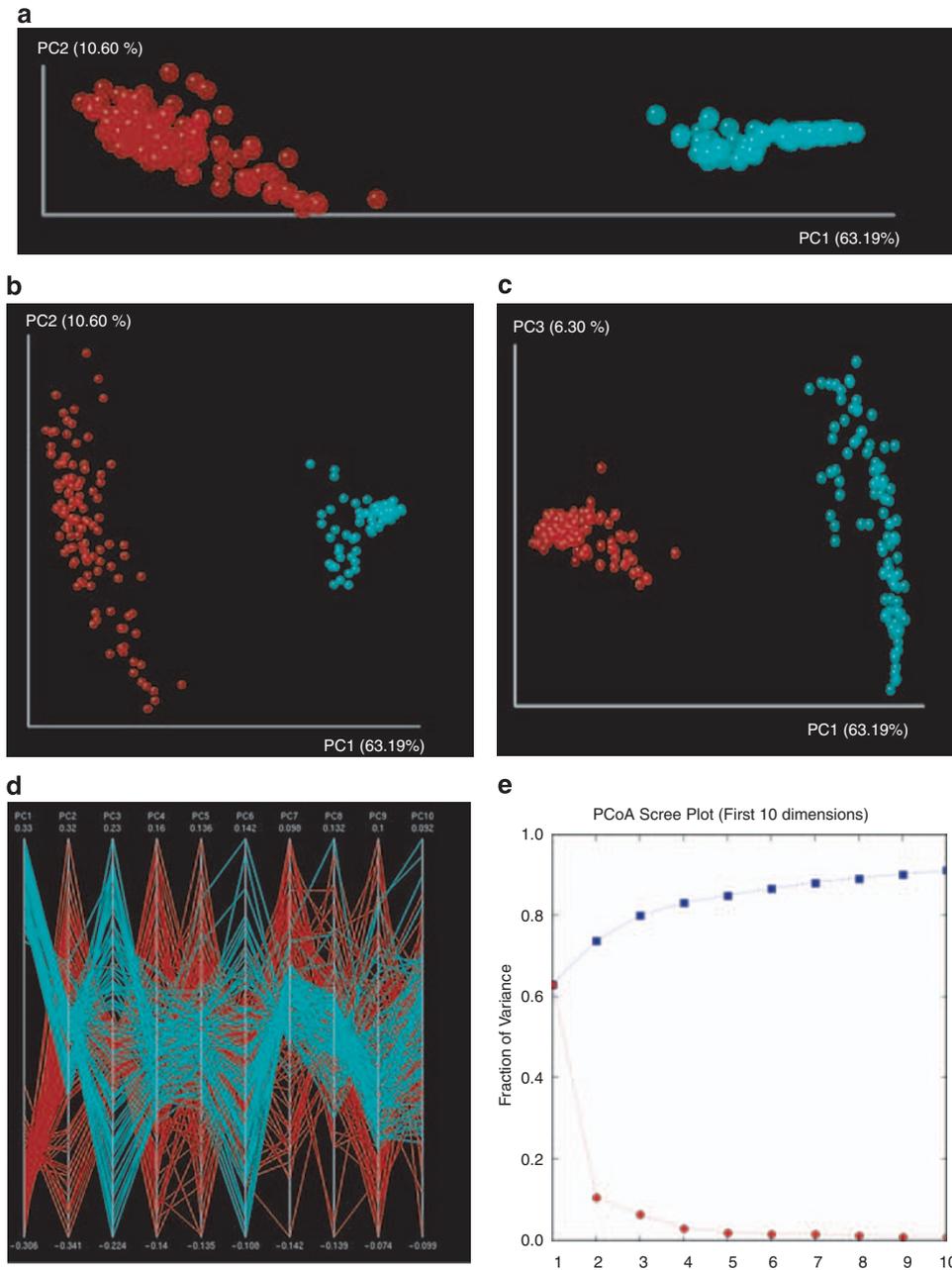


Figure 3 Principal coordinates analysis of Weighted UniFrac values between hand (blue) and gut (red) pyrosequencing datasets with the axes scaled by the percentage of the variance that they contain (a) or unscaled (b and c). Panel b plots PC1 vs PC2 and panel c plots PC1 vs PC3. A parallel coordinates plot (d) allows visualization of which of the first 10 PC axes the hand vs. gut samples are varying across: in this display, the position of each sample along each of the first 10 axes is plotted (for example, the hand samples score high on PC1 and the gut samples score low, so on the first line, for PC1, the hand samples have high values and the gut samples have low values). A scree plot (e) allows for easy visualization of the % fraction of the variance explained by the first 10 PC axes, both individually (red) and cumulatively (blue).

tion for the analysis of similar as well as dissimilar sample types.

Combining the hand and gut pyrosequencing datasets
The combination of hand and gut datasets provides the largest combined pyrosequencing 16S rRNA dataset analyzed to date, encompassing 680 000 sequences. PCoA analysis of pairwise weighted

UniFrac values shows that, as expected, the difference between the hand and gut samples accounts for the majority of the variation among these samples (63.2%) (Figure 3). Gut samples differentiate along PC2 (10.6%) and skin along PC3 (6.3%), forming two separate gradients, one within the hand samples and one within the gut samples that are orthogonal to each other (Figure 3). The relative importance of the hand–gut differences is most easily viewed

when the axes are scaled by the % of the variation explained in the 3D viewer (Figure 3a). However, the separation between the gut and hand samples in PC axes 2 and 3 can be most easily seen using an unscaled view (Figures 3b and c). The parallel coordinates plot, which is also accessed in the 3D viewer (Figure 3d), allows for easy visualization for which of the first 10 PCoA axes the hand and gut samples vary across, and the scree plot, which is displayed directly in the web interface, allows for easy visualization on the relative and cumulative importance of the different axes (Figure 3e). The major pattern, with orthogonal gradients in hand and gut, is visually immediately obvious but was unsuspected before the datasets were combined.

This analysis of these pyrosequencing reads would have taken approximately two orders of magnitude longer to perform using the original version of UniFrac on a single CPU. To compare the performance of Fast UniFrac to the original implementation, we sampled 1000–10 000 unique nodes from the reference tree from the hand/gut dataset (225 samples, ~680 000 sequences) in steps of 1000 (Figure 4). On average, each number of nodes corresponds to a much larger number of sequences because many sequences are abundant across samples. Both implementations were compared on the same set of trees: 10 trees were created for each sample size, and the average is displayed on a log scale. In general, the new implementation is 10–100 times faster than the original implementation, and the large difference in performance

between weighted and unweighted UniFrac in the original implementation is eliminated.

Analysis of PhyloChip data

The 24 PhyloChip dataset used was from a study in which leaf samples from citrus trees infected with the Huanglongbing pathogen from several different groves were analyzed using the PhyloChip (Sagaram *et al.*, 2009). The entire analysis of 24 PhyloChip samples took Fast UniFrac a matter of minutes after exporting the data from PhyloTrac (Figure 5). Similar to the original study, we found no significant clustering of the overall community by grove or disease status although the clustering does look suggestive and larger sample sizes could make the patterns more conclusive: the clear arch effect, in which samples are spread along a curve, strongly suggests that there is a single underlying gradient that explains much of the variation in the community, and the scree plot shows that most (>80%) of the variance in the data is explained by the first three principal coordinate axes. Additional collection of metadata about the individual plants may help explain the major unmeasured sources of variation in the dataset and allow more subtle patterns associated with infection to be detected. The ability to see results colored by different metadata categories in the context of the full dataset is extremely useful for exploratory analyses, and can direct additional sample collection efforts once the overall patterns are clear.

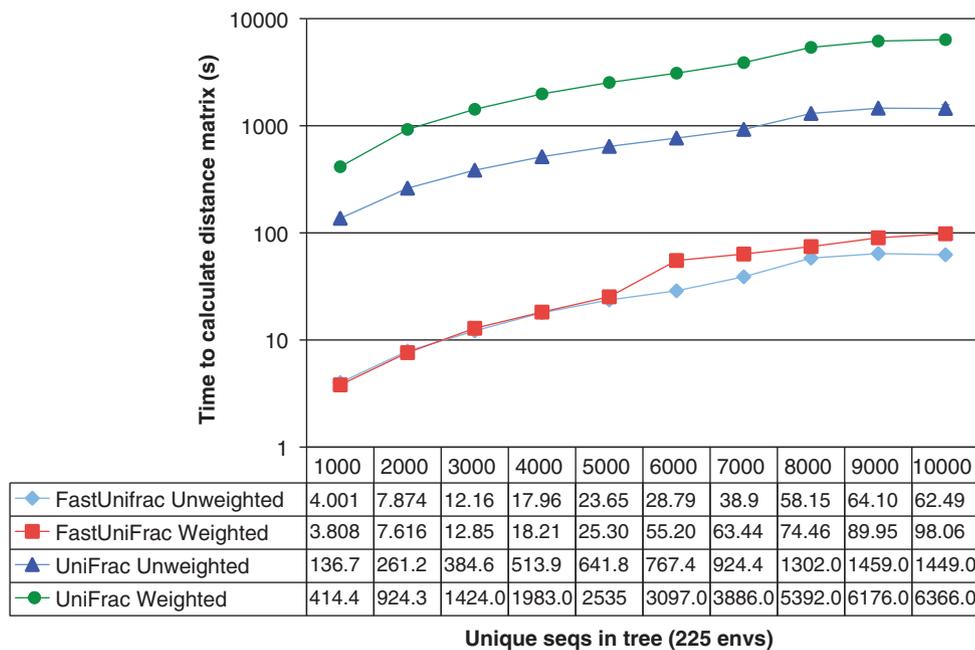


Figure 4 Performance of Fast UniFrac vs original implementation on sample sizes ranging from 1000 to 10 000 sequences. Fast UniFrac implementation is consistently about two orders of magnitude faster, and largely eliminates the difference in time to calculate weighted and unweighted UniFrac metrics.

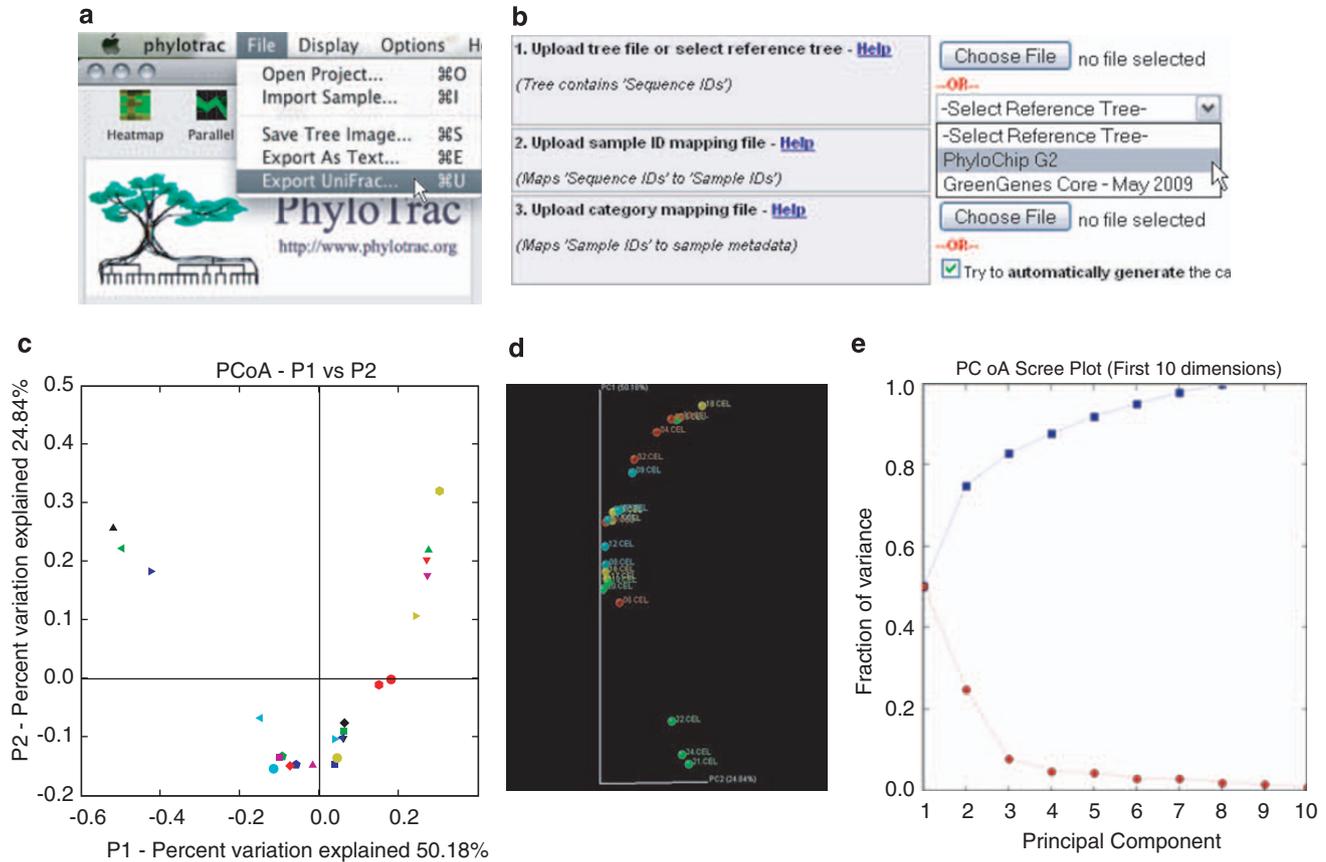


Figure 5 Example PhyloChip analysis performed using Phylotracc and Fast UniFrac. (a) Exporting the environment file from Phylotracc, (b) uploading to Fast UniFrac, (c) viewing weighted Fast UniFrac PCoA results in the web interface directly (in this display, each point is a sample, and we see a 2D projection of the first two principal coordinates obtained by PCoA; the relatively smooth curve suggests that there is a gradient connecting the samples), (d) viewing unweighted Fast UniFrac ordination results in the linked 3D viewer: again, each point is a sample and the distances are calculated by PCoA of the UniFrac distances, but in this case three dimensions are shown, and (e) a scree plot showing how much of the variation is explained singly or cumulatively by each of the first 10 principal coordinates, allowing the user to see that, for example, the first three principal coordinates together explain over 80% of the variance in the samples. As reported in the original study, no clear patterns are readily seen using ordination, but shows the speed and ease with which this sort analysis can now be performed.

Discussion

Our results indicate that the performance increase achieved with Fast UniFrac, and the corresponding ability to perform analyses and meta-analyses of large numbers of samples using readily available techniques (for example, BLAST and Phylotracc), will greatly enhance a wide range of studies of microbial ecology. In general, the speedup by two orders of magnitude in processing time, the ability to rapidly color samples according to different criteria and to display more than the first 3D for rapid profiling, and the ability to reproduce previous results using a standardized pipeline based on familiar tools will allow many groups to integrate large pyrosequencing and/or PhyloChip studies, thus providing key cyberinfrastructure for Human Microbiome Projects and related efforts.

Some of the biological findings presented here are intriguing in their own right, although detailed follow-up is beyond the scope of the present paper. We note that the saline/non-saline split in environ-

mental samples (Lozupone and Knight, 2007) and the even deeper split between environmental and host-associated samples (Ley *et al.*, 2008b) have now been recaptured using a range of methodologies and appear to be robust. The levels of intra- and interpersonal variability observed within and between human body habitats (Frank *et al.*, 2007; Fierer *et al.*, 2008; Ley *et al.*, 2008a; Turnbaugh *et al.*, 2009) suggest that large sample sizes, including time series analyses, will be especially critical for understanding whether or not observed community structures are significantly associated with physiologic or pathophysiologic states. Our re-analysis of the PhyloChip data associated with Huanglongbing pathogen-infected citrus (Sagaram *et al.*, 2009) reinforces this point: although we see intriguing differences in intrinsic variability of the leaf communities in different groves, much larger numbers of samples would be required to establish these patterns conclusively. However, the decreasing cost of the PhyloChip and, especially, of barcoded multiplex pyrosequencing (Hamady *et al.*, 2008)

should provide the statistical power required to observe subtle biomarkers of disease.

We note that significance tests such as the *P*-test (Martin, 2002) and the UniFrac significance tests become decreasingly useful as the depth of coverage and the number of samples increases. For example, essentially all pairs of pyrosequencing-derived samples we examined in this study are significantly different by the *P*-test (data not shown), as statistical power increases with sampling effort. Performing many pairwise significance tests in studies with many samples, however, has limited meaning because (1) corrections for multiple comparisons, such as the Bonferroni correction, make it difficult to detect real differences because of a high Type II error rate (β errors or false negatives), (2) the number of randomizations that are needed to detect differences becomes prohibitively large, and (3) no information is gained on variation in the degree of difference between sample pairs as significance is a factor of both degree of difference and sampling effort. We recommend a shift in emphasis from testing whether each pair of samples is *significantly* different to using multivariate methods, such as PCoA and hierarchical clustering to detect broad trends of similarities and differences that relate all samples (a broad suite of statistical techniques, such as the Mantel test, ANOSIM, PERMANOVA, and so on, already exists to test for significant differences among categories). If samples are really drawn from a single distribution, as the *P*-test and UniFrac significance test assume as their null hypothesis, then no large-scale trends will be observed. In contrast, if sample clustering does exist, the ability to relate large-scale differences in community to specific biological observables, such as sample type, pH, salinity, or other variables becomes essential. By allowing the UniFrac distance matrices to be exported for analysis in third-party packages such as R and PRIMER, and by allowing the same principal coordinates projection to be colored many different ways according to different user-supplied categorical variables, Fast UniFrac facilitates insight into the specific variables associated with sample clustering. Similarly, the ability to perform lineage-specific analyses by including only a subset of the tree allows insight into the specific lineages responsible for associations with ecologically important variables.

In conclusion, we have shown that Fast UniFrac provides order-of-magnitude improvements in speed over the original version, together with many user interface enhancements and connections to other data sources that greatly increase the throughput of analyses. Contribution of the Fast UniFrac code to open-source efforts such as PyCogent (Knight *et al.*, 2007) and the Human Microbiome Project Data Analysis and Coordination Center (<http://www.hmpdacc.org/>) will provide key cyber-infrastructure as the field moves beyond clone libraries to analyses of hundreds to thousands of

PhyloChips or massively parallel sequencing efforts that yield millions of reads.

Acknowledgements

We thank Jeffrey I Gordon, Ruth Ley, Noah Fierer, Brian Muegge, Jesse Stombaugh, Daniel McDonald, and Christian Lauber for valuable feedback on the paper. This work was supported in part by NIH Grants 1R01HG004872-01, 1U01HG004866-01, and P01DK 078669, by the Crohn's and Colitis Foundation of America, and by a Bill and Melinda Gates Foundation Mal-ED Network Discovery Project.

References

- Alexander E, Stock A, Breiner HW, Behnke A, Bunge J, Yakimov MM *et al.* (2009). Microbial eukaryotes in the hypersaline anoxic L'Atalante deep-sea basin. *Environ Microbiol* **11**: 360–381.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.
- Balakirev ES, Pavlyuchkov VA, Ayala FJ. (2008). DNA variation and symbiotic associations in phenotypically diverse sea urchin *Strongylocentrotus intermedius*. *Proc Natl Acad Sci USA* **105**: 16218–16223.
- Bryant JA, Lamanna C, Morlon H, Kerkhoff AJ, Enquist BJ, Green JL. (2008). Colloquium paper: microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proc Natl Acad Sci USA* **105**(Suppl 1): 11505–11511.
- DeSantis TZ, Brodie EL, Moberg JP, Zubieta IX, Piceno YM, Andersen GL. (2007). High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microb Ecol* **53**: 371–383.
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.
- Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M *et al.* (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452**: 340–343.
- Elifantz H, Waidner LA, Michelou VK, Cottrell MT, Kirchman DL. (2008). Diversity and abundance of glycosyl hydrolase family 5 in the North Atlantic Ocean. *FEMS Microbiol Ecol* **63**: 316–327.
- Fierer N, Hamady M, Lauber CL, Knight R. (2008). The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci USA* **105**: 17994–17999.
- Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* **104**: 13780–13785.
- Fraune S, Bosch TC. (2007). Long-term maintenance of species-specific bacterial microbiota in the basal metazoan Hydra. *Proc Natl Acad Sci USA* **104**: 13146–13151.
- Graham CH, Fine PV. (2008). Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecol Lett* **11**: 1265–1277.

- Grice EA, Kong HH, Renaud G, Young AC, Bouffard GG, Blakesley RW *et al.* (2008). A diversity profile of the human skin microbiota. *Genome Res* **18**: 1043–1050.
- 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.
- Harrison BK, Zhang H, Berelson W, Orphan VJ. (2009). Variations in archaeal and bacterial diversity associated with the sulfate-methane transition zone in continental margin sediments (Santa Barbara Basin, California). *Appl Environ Microbiol* **75**: 1487–1499.
- Hartman WH, Richardson CJ, Vilgalys R, Bruland GL. (2008). Environmental and anthropogenic controls over bacterial communities in wetland soils. *Proc Natl Acad Sci USA* **105**: 17842–17847.
- Hiibel SR, Pereyra LP, Inman LY, Tischer A, Reisman DJ, Reardon KF *et al.* (2008). Microbial community analysis of two field-scale sulfate-reducing bioreactors treating mine drainage. *Environ Microbiol* **10**: 2087–2097.
- Hsu SF, Buckley DH. (2009). Evidence for the functional significance of diazotroph community structure in soil. *ISME J* **3**: 124–136.
- Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA *et al.* (2007). Microbial population structures in the deep marine biosphere. *Science* **318**: 97–100.
- Kanagawa T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng* **96**: 317–323.
- Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC *et al.* (2007). PyCogent: a toolkit for making sense from sequence. *Genome Biol* **8**: R171.
- Lauber CL, Sinsabaugh RL, Zak DR. (2009). Laccase gene composition and relative abundance in oak forest soil is not affected by short-term nitrogen fertilization. *Microb Ecol* **57**: 50–57.
- Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS *et al.* (2008a). Evolution of mammals and their gut microbes. *Science* **320**: 1647–1651.
- Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. (2008b). Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* **6**: 776–788.
- Ley RE, Turnbaugh PJ, Klein S, Gordon JI. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* **444**: 1022–1023.
- Li M, Wang B, Zhang M, Rantalainen M, Wang S, Zhou H *et al.* (2008). Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci USA* **105**: 2117–2122.
- Lozupone C, Hamady M, Knight R. (2006). UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7**: 371.
- Lozupone C, Knight R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228–8235.
- Lozupone CA, Hamady M, Cantarel BL, Coutinho PM, Henrissat B, Gordon JI *et al.* (2008). The convergence of carbohydrate active gene repertoires in human gut microbes. *Proc Natl Acad Sci USA* **105**: 15076–15081.
- Lozupone CA, Knight R. (2007). Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104**: 11436–11440.
- Lozupone CA, Knight R. (2008). Species divergence and the measurement of microbial diversity. *FEMS Microbiol Rev* **32**: 557–578.
- Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.
- Marhaver KL, Edwards RA, Rohwer F. (2008). Viral communities associated with healthy and bleaching corals. *Environ Microbiol* **10**: 2277–2286.
- Martin AP. (2002). Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol* **68**: 3673–3682.
- Nasidze I, Li J, Quinque D, Tang K, Stoneking M. (2009). Global diversity in the human salivary microbiome. *Genome Res* **19**: 636–643.
- Osman S, La Duc MT, Dekas A, Newcombe D, Venkateswaran K. (2008). Microbial burden and diversity of commercial airline cabin air during short and long durations of travel. *ISME J* **2**: 482–497.
- Porter TM, Skillman JE, Moncalvo JM. (2008). Fruiting body and soil rDNA sampling detects complementary assemblage of Agaricomycotina (Basidiomycota, Fungi) in a hemlock-dominated forest plot in southern Ontario. *Mol Ecol* **17**: 3037–3050.
- Rawls JF, Mahowald MA, Ley RE, Gordon JI. (2006). Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell* **127**: 423–433.
- Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD *et al.* (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283–290.
- Sagaram US, DeAngelis KM, Trivedi P, Andersen GL, Lu SE, Wang N. (2009). Bacterial diversity analysis of Huanglongbing pathogen-infected citrus, using PhyloChip arrays and 16S rRNA gene clone library sequencing. *Appl Environ Microbiol* **75**: 1566–1574.
- Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored ‘rare biosphere’. *Proc Natl Acad Sci USA* **103**: 12115–12120.
- Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.
- Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. (2007). The human microbiome project. *Nature* **449**: 804–810.
- Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**: 1027–1031.
- Wen L, Ley RE, Volchkov PY, Stranges PB, Avanesyan L, Stonebraker AC *et al.* (2008). Innate immunity and intestinal microbiota in the development of Type 1 diabetes. *Nature* **455**: 1109–1113.
- Widmann J, Hamady M, Knight R. (2006). DivergentSet, a tool for picking non-redundant sequences from large sequence collections. *Mol Cell Proteomics* **5**: 1520–1532.
- Wilson KH, Wilson WJ, Radosevich JL, DeSantis TZ, Viswanathan VS, Kuczumarski TA *et al.* (2002). High-density microarray of small-subunit ribosomal DNA probes. *Appl Environ Microbiol* **68**: 2535–2541.
- Zhu M, Ghodsi A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Comput Stat Data Anal* **51**: 918–930.

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