Microbial community profiling for human microbiome projects: Tools, techniques, and challenges

Micah Hamady[1] and Rob Knight[2,3]

[1]*Department of Computer Science, University of Colorado, Boulder, Colorado 80309, USA;* [2]*Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado 80309, USA*

High-throughput sequencing studies and new software tools are revolutionizing microbial community analyses, yet the variety of experimental and computational methods can be daunting. In this review, we discuss some of the different approaches to community profiling, highlighting strengths and weaknesses of various experimental approaches, sequencing methodologies, and analytical methods. We also address one key question emerging from various Human Microbiome Projects: Is there a substantial core of abundant organisms or lineages that we all share? It appears that in some human body habitats, such as the hand and the gut, the diversity among individuals is so great that we can rule out the possibility that any species is at high abundance in all individuals: It is possible that the focus should instead be on higher-level taxa or on functional genes instead.

The human microbiota (the collection of microbes that live on and inside us) consists of about 100 trillion microbial cells that outnumber our "human" cells 10 to 1 (Savage 1977), and that provide a wide range of metabolic functions that we lack (Gill et al. 2006). If we consider ourselves as supraorganisms encompassing these microbial symbionts (Lederberg 2000), by far the majority of genes in the system are microbial. In this sense, completing the human genome requires us to characterize the microbiome (the collection of genes in the microbiota) (Turnbaugh et al. 2007). Currently, there are two main methods for performing this characterization that do not rely on growing organisms in pure culture: small-subunit ribosomal RNA (rRNA) studies, in which the 16S rRNA gene sequences (for archaea and bacteria) or the 18S rRNA gene sequences (for eukaryotes) are used as stable phylogenetic markers to define which lineages are present in a sample (Pace 1997), and metagenomic studies, in which community DNA is subject to shotgun sequencing (Rondon et al. 2000). Small subunit rRNA-based studies are sometimes also considered to be "metagenomic" in that they analyze a heterogeneous sample of community DNA. Community profiling, or determining the abundance of each kind of microbe, is much cheaper using rRNA because only one gene out of each genome is examined, but metagenomic profiles are essential for understanding the functions encoded in those genomes. Techniques that probe gene expression directly such as metatranscriptomics and metaproteomics (analysis of the transcripts or proteins in a community, respectively), although useful in simpler microbial communities such as acid mine drainage (Lo et al. 2007; Frias-Lopez et al. 2008), are just beginning to be applied to human-associated microbial communities (Verberkmoes et al. 2008).

Through the use of metagenomic and rRNA-based techniques, much progress has been made in characterizing the human microbiome and its role in health and disease in the past few years, especially with the advent of high-throughput sequencing. These studies are challenging because of the scale and complexity of the microbiome and because of the unexpected variability between individuals. In this review, we cover the combination of experimental and analytical techniques used to characterize the microbiomes of humans and of other mammals. In particular, we describe how recent advances in technology and experimental techniques, together with computational methods that draw on the long tradition of community analysis in large-scale ecological studies, are essential for uncovering large-scale trends that relate the microbiomes of many individuals.

One fundamental question raised by the National Institutes of Health (NIH), European Union (EU), and other sponsored Human Microbiome Projects (HMPs) is whether there is a core human microbiome of genes or species that we all share (Fig. 1; Turnbaugh et al. 2007, 2009). If there is a substantial core, the strategy for understanding the microbiome is clear: Identify the organisms that comprise the core using 16S rRNA analysis, sequence their genomes, and use these genomes as scaffolds for metagenomic, metatranscriptomic, and metaproteomic studies that provide information about small fragments of genes, transcripts, or proteins, respectively, but that require assembly against known sequences (Turnbaugh et al. 2007; Zaneveld et al. 2008). However, if there is a minimal core or no core at all, alternative strategies will need to be developed because new genes and species will continue to be found in each new person examined.

Another key question is whether changes in the relative abundance of members of human-associated microbial communities are generally important. For example, the proportional representation of the bacterial phyla Firmicutes, Actinobacteria, and Bacteroidetes in the gut is associated with obesity in both humans and mice (Ley et al. 2005, 2006c; Turnbaugh et al. 2006, 2008). However, although this observation establishes that changes in the abundance of broad bacterial groups such as entire phyla can be important and we also know that miniscule inoculations of particular pathogenic strains can cause disease, we know little about the physiological impacts of changes in microbial abundance at a given taxonomic level in general. Our power to detect particular species depends on depth of sequencing and on whether they can be selected by culture-based techniques. For example, we think of *Escherichia coli* as a classic gut bacterium, but the entire Gamma-proteobacteria phylum that contains it typically comprises much less than 1% of gut bacteria—rather, *E. coli* grows extremely well in culture and can thus be detected at low abundance. Most species found in the gut of a given individual are rare (Dethlefsen et al. 2007), which makes them difficult to detect,

[3]**Corresponding author.**
E-mail rob.knight@colorado.edu; fax (303) 492-7744.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.085464.108.

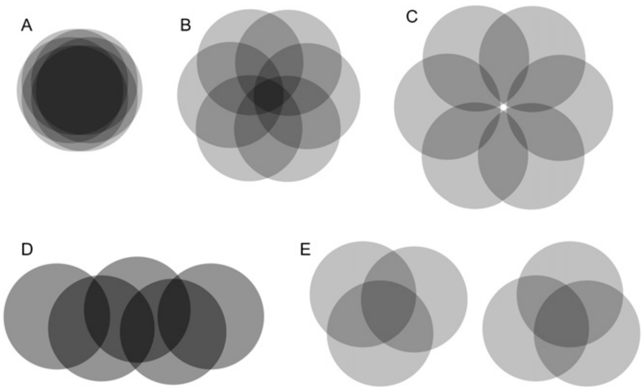

Figure 1. Models of a core microbiome. The circles represent the microbial communities in different individuals and can be thought of as either representing different taxa (species, genera, etc.) or representing different genes. (*A*) "Substantial core" model. Most individuals share most components of the microbiota. (*B*) "Minimal core" model. All individuals share a few components, and any individual shares many components with a few other individuals, but very little is shared across all individuals. (*C*) "No core" model. Nothing is shared by all individuals, and most diversity is unique to a given individual. (*D*) "Gradient" model. Individuals next to each other on a gradient, for example, age or obesity, share many components, but individuals at opposite ends share little or nothing. (*E*) "Subpopulation" model. Different subpopulations, for example, those defined by geography or disease, have different cores, but nothing is shared across subpopulations. Scenarios *C–E* would represent situations in which the strategy of identifying core species for sequencing, then using these as a scaffold for "omics" studies, would be problematic.

and a plethora of rare species has also been found in other ecosystems (Sogin et al. 2006; Huber et al. 2007). One possibility is that everyone shares the same microbial species but that the abundance of individual species varies by orders of magnitude in different people in ways that affect health and disease. If rare species are generally important, much deeper characterization of the microbiome may be required.

Community profiling with 16S rRNA: New life through deeper sequencing

Microbial community profiling using 16S rRNA is currently undergoing a renaissance (Tringe and Hugenholtz 2008) as high-throughput techniques such as barcoded pyrosequencing allow us to gain deep views into hundreds of microbial communities simultaneously (Hamady et al. 2008). These studies are made possible by the remarkable observation that a small fragment of the 16S rRNA gene is sufficient as a proxy for the full-length sequence for many community analyses, including those based on a phylogenetic tree (Liu et al. 2007, 2008; Wang et al. 2007). Although the phylogenetic trees produced from ~250-base reads from the current 454 Life Sciences (Roche) GS FLX instrument are relatively inaccurate, they are still vastly better than the so-called "star phylogeny," the phylogeny that assumes all species are equally related, that all nonphylogenetic methods for comparing communities implicitly use (e.g., by counting how many species are shared). However, such trees should only be used as a guide to community comparisons and not for inferring true phylogenetic relationships among reads. Rapid advances in sequencing technology, such as the recent availability of 400-base reads with the Titanium kit from Roche or, in the future, the availability of instruments providing 1500-base single-molecule reads, as reported by Pacific Biosciences (Korlach et al. 2008), will also improve the accuracy of existing methods for building phylogenetic trees and classifying functions of metagenomic reads. Similarly, the availability of many improved computational methods for comparing large numbers of microbial communities including UniFrac (Lozupone and Knight 2005; Lozupone et al. 2006), SONS (Schloss and Handelsman 2006), and network-based comparisons (Ley et al. 2008a) will allow very rapid progress to be made.

Sequence databases, especially rRNA sequence databases, are growing explosively (Medini et al. 2008), and the ability to see hundreds of samples at depth of coverage of many thousands of sequences per sample allows us to contemplate completely new types of analyses. At the same time, this flood of data poses formidable challenges in data analysis because many standard computational tools are not designed for input on this scale. Many investigators are already encountering the limitations of existing tools—for example, it is impossible to align the half-million sequences obtained from a single 454 FLX run with traditional tools such as ClustalW (Thompson et al. 1994) or even the publicly available versions of newer, rRNA-specific tools such as NAST (DeSantis Jr. et al. 2006b). These tools and databases must anticipate scaling up to thousands of samples and many millions of sequences over the next few years.

Key questions facing investigators

The wide array of sequencing technologies and analytical tools can be daunting. The path to a successful study is first to define what hypothesis is being tested and then to select the appropriate technology. For example, it would be unfortunate to spend months and millions of dollars performing a metagenomic study solely to find changes such as the shift in the Bacteroidetes: Firmicutes–Actinobacteria ratio in the gut of obese individuals when a much faster and cheaper assay could have provided the same result at a much lower cost. Such studies must be justified by additional analyses that can only be performed with metagenomic sequences (Turnbaugh et al. 2009). Here we cover some of the key questions facing investigators: whether to use sequencing or to use lower resolution but cheaper methods that allow more samples to be processed for the same cost, which type of sequencing to perform, and how the data should be analyzed. These decisions, especially with respect to data analysis, often differ between rRNA and metagenomic surveys. For example, phylogenetic methods are increasingly useful for rRNA surveys because this gene allows accurate reconstruction of phylogeny, whereas functional or taxon-based methods are typically more useful for metagenomic surveys because of the range of functions represented and because of the difficulty of reconstructing the phylogenies of small fragments of many gene families.

When is it necessary to obtain sequences, and when should cheaper approximate methods such as fingerprinting be used?

Although the cost of sequencing is dropping, fingerprinting techniques (techniques that provide limited information about the microbial community) are still orders of magnitude cheaper and faster to perform. Fingerprinting techniques include T-RFLP, DGGE, and TGGE: These methods have been reviewed comprehensively (Anderson and Cairney 2004). Briefly, they rely on amplification of a specific gene, typically but not always 16S rRNA, then separating different variants of the gene in the community sample by electrophoresis. These methods can be used to analyze large numbers of samples, including clustering of the banding patterns with statistical techniques such as Principal Coordinates Analysis (PCoA) (Dollhopf et al. 2001), but typically the dynamic range is limited (so only the few most abundant members of the community can be observed), and it is difficult to relate banding patterns to changes in particular species or lineages. It is also generally impossible to combine data from different studies into a single analysis. However, these techniques can be useful for checking for stability in the dominant members of a community and for clustering communities according to changes in the dominant members across large numbers of samples (Fierer and Jackson 2006).

The main advantages of sequencing studies over fingerprinting are that sequences can be classified according to taxonomy and function, that sequencing provides much greater dynamic range and ability to compare complex samples, and that sequences from different studies can be compared to one another and placed in the same phylogenetic tree (Lozupone and Knight 2007). Sequencing is especially useful when asking which specific genes or species contribute to differences among communities.

Two intermediate approaches between fingerprinting and sequencing are to use short sequence tags and to use DNA microarrays. These approaches provide access to large numbers of samples, as does fingerprinting, but also provide phylogenetic resolution, as does sequencing, albeit with limits in both dimensions.

Techniques that use short sequence tags include the ARISA technique that uses the intergenic spacer of the ribosomal RNA (Fisher and Triplett 1999) or sequencing of the V6 hypervariable region of the 16S rRNA (Sogin et al. 2006). These methods allow comparisons among samples and measurements of diversity, and, when the database of full-length sequences is sufficiently complete, assessment of the taxonomic distribution. However, they are less useful in cases in which new lineages with no close relatives are dominant.

DNA microarrays such as the PhyloChip (Wilson et al. 2002) and GeoChip (He et al. 2007) provide a convenient way of screening 16S rRNA and functional gene sequence libraries, respectively. These tools have the potential to be much cheaper and provide much greater dynamic range than sequencing studies but require that the sequences be known in advance so that they can be printed on the chip (DeSantis et al. 2007). For example, this approach was recently used to track development of the human gut microbiota in infancy (Palmer et al. 2007). An additional advantage of microarrays is that probes with broad but defined specificity, for example, at the family or phylum level, can be used to estimate the abundance of the group as a whole, even if probes for more specific taxa are missing. Normalization can be an issue and is probe dependent: This issue is analogous to issues with primer bias that affect sequencing and fingerprinting, and the most important consideration is to use the same technique for all samples that are to be analyzed together.

Both the sequence tag approach and the microarray approach can be used to generate data suitable for the community analysis techniques described below, essentially by mapping each sequence tag (or spot on the microarray) onto the closest full-length sequence in the database and using that sequence as a proxy. Sequencing studies are thus most useful when the samples have been poorly characterized and the discovery of many new gene or species lineages is anticipated. 16S rRNA sequencing studies are especially useful for characterizing which kinds of organisms are present in a wide range of samples (especially when differences at or above the genus level distinguish the samples), whereas metagenomic sequencing studies are especially useful for characterizing microbial assemblages at a functional level (see below).

How should I perform my sequencing?

What are my choices for sequencing technology?

There are several choices of sequencing technology. Capillary (Sanger) sequencing currently produces longer reads of up to 800 bases, which are very useful for inferring gene functions for metagenomics (Wommack et al. 2008). However, pyrosequencing (Ronaghi et al. 1996, 1998; Margulies et al. 2005) is orders of magnitude cheaper and faster and also eliminates the laborious step of preparing clone libraries. The benefit of a large number of short reads clearly outweighs the drawbacks of short read lengths for many kinds of rRNA-based community analysis: 200-base reads, accounting for ~12% of the data in the 16S rRNA gene, yield community clustering results as accurate as those obtained using 70% of the original number of full-length sequences in a comparison across different habitats (Liu et al. 2007), and Sanger and pyrosequencing data provided comparable results in fecal samples from both rhesus macaques (McKenna et al. 2008) and in lean and obese humans (Turnbaugh et al. 2009). Given that Sanger sequencing is at least an order of magnitude more expensive than pyrosequencing, requires that DNA templates be clonable into a common host (*E. coli*), and has markedly lower throughput/instrument, the latter is clearly the most cost-effective option for testing hypotheses about the distribution of microbial diversity among samples at this point.

A key pyrosequencing innovation currently used in both 16S rRNA and metagenomic studies is multiplexing. Because far more sequences are generated in a single pyrosequencing run than are needed for many kinds of community analyses, it is often desirable to split a single run across many samples. Two general strategies are physical separation of samples and barcoded pyrosequencing: These approaches can be used in combination as each part of the plate may be used to run many barcoded samples.

Barcoded pyrosequencing uses molecular barcoding techniques that were initially developed in the 1980s (Church and Kieffer-Higgins 1988; Shoemaker et al. 1996). Sequences in each sample are tagged with a unique barcode either by ligation or, for amplicon sequencing, by using a barcoded primer when amplifying each sample by PCR (Fig. 2). Several different barcoding strategies have been used with pyrosequencing (Binladen et al. 2007; Hoffmann et al. 2007; Huber et al. 2007; Parameswaran et al. 2007; Fierer et al. 2008; Hamady et al. 2008). Of these, we recommend the use of error-detecting and error-correcting codes that use formal mathematical techniques to define barcodes in such a way that a certain number of errors can be detected and corrected: For example, using Hamming codes of eight bases can detect all double-bit errors and correct all single-bit errors (Hamady et al. 2008), and Golay codes of 12 bases can correct all triple-bit errors and detect all quadruple-bit errors. Both types of barcodes have been used to sequence several hundred samples per run (Fierer et al. 2008; Hamady et al. 2008) and can theoretically accommodate thousands. One important issue with barcoding is variability in reads per sample; however, the sources of this

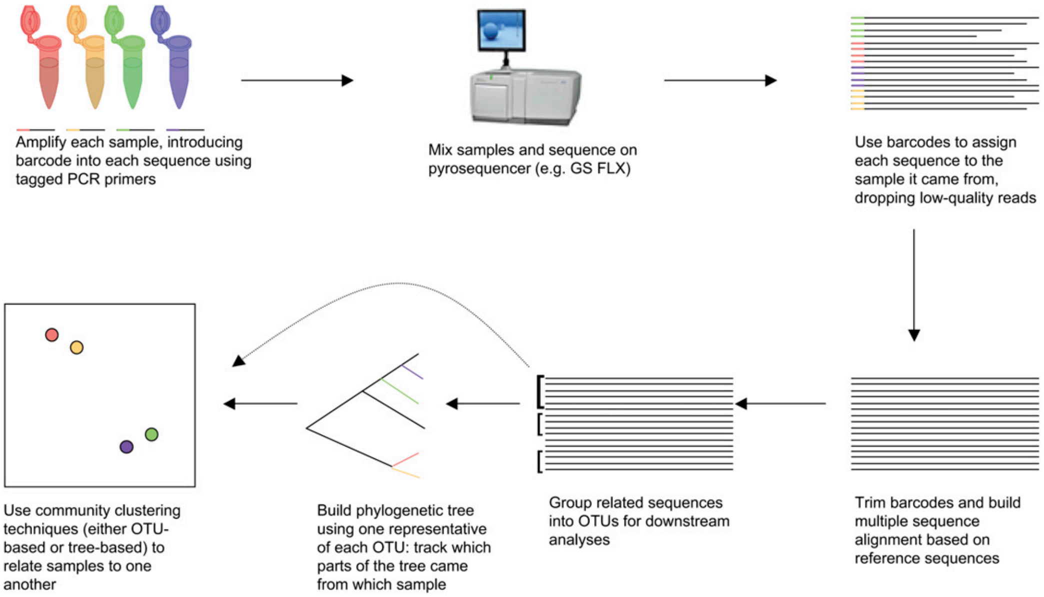
Amplify each sample, introducing barcode into each sequence using tagged PCR primers
Mix samples and sequence on pyrosequencer (e.g. GS FLX)
Use barcodes to assign each sequence to the sample it came from, dropping low-quality reads
Trim barcodes and build multiple sequence alignment based on reference sequences
Group related sequences into OTUs for downstream analyses
Build phylogenetic tree using one representative of each OTU: track which parts of the tree came from which sample
Use community clustering techniques (either OTU-based or tree-based) to relate samples to one another

Figure 2. Overview of barcoded pyrosequencing workflow. The sample-specific barcodes are introduced into each sample during the PCR step (for amplicon sequencing), or through ligation (for metagenomics). After sequencing, individual sequences can then be traced back to individual samples using the barcodes they contain. The sequences from each sample are then separated, aligned, and then either used directly for taxa-based analyses or used to build trees for phylogenetic analyses. OTU, operational taxonomic unit.

variability are relatively poorly understood at this point, and existing studies of bias due to the use of barcoded primers (Binladen et al. 2007), while useful, have been limited. Systematic studies of the effects of specific barcodes on community structure remain to be performed, although the available evidence suggests that technical or biological replicates sequenced using different barcodes typically cluster together (Hamady et al. 2008; Turnbaugh et al. 2009).

In principle, all of these barcoding techniques can be used for metagenomics by ligating a barcode to fragmented DNA sequences in each sample (Meyer et al. 2008b), although this is still very much an emerging technique.

What read length should I aim for?

For 16S rRNA studies 250-base reads can be essentially as good as full-length sequences for many microbial community comparisons and can even be useful for taxonomy assignment, provided that the region of the 16S rRNA is carefully chosen, for example, the V2 or V4 regions (Liu et al. 2007, 2008; Wang et al. 2007). The increased number of reads makes these types of studies an exceptional bargain compared to Sanger sequencing of full-length amplicons. However, to define new bacterial phyla, or in cases in which the sequences obtained are highly divergent from related sequences in the reference databases, obtaining the full-length sequence (e.g., by repeating the PCR using a very specific primer designed to match the short read and a universal primer near the other end of the 16S rRNA) is essential. For metagenomics, the read length is much more important owing to the difficulty of identifying the function and/or species of each gene from relatively short reads. The increase in read length from 250 to 400 bases with the Titanium kit should make a large difference to the fraction of reads that can be accurately assigned (Huson et al. 2007; Mavromatis et al. 2007; Dalevi et al. 2008; Krause et al. 2008; Wommack et al. 2008). However, the increase to 400 bases has relatively little effect on 16S rRNA taxonomic assignment (Liu et al. 2007).

What sampling depth is needed?

The number of sequences required to characterize a sample depends on the goal of the study, the diversity of species in the sample, the read length, and, for amplicons, the choice of gene and region for sequencing.

If the goal is to estimate the major bacterial phyla in each sample, relatively few sequences per sample are needed. For example, in 22 human gut samples with depth of coverage of at least 350 sequences/individual, communities averaged 75% Firmicutes and 18% Bacteroidetes (Fig. 3; Ley et al. 2008b). It would thus take little sequencing effort to conclude that these are the two dominant phyla in this habitat: With only 50 sequences, we would already conclude that the total proportion of sequences in the remaining phyla was 14% ± 4%. As with all statistical questions, large differences can be detected with smaller samples. For example, sequence jackknifing showed that only 17 sequences were needed to reliably cluster microbial assemblages in two oligotrophic subtropical seawater samples together, and to separate these samples from other marine microbial assemblages from sediments, terrestrially impacted seawater, and sea ice (Lozupone and Knight 2005). As few as 100 sequences per sample were sufficient to detect the major patterns of variation among the microbial communities in the guts of diverse mammals (Ley et al. 2008a). Thus, large-scale patterns can be recovered with shallow sampling, even when this sampling only scratches the surface of the diversity in the communities (see also Fig. 4 for comparison of phylogenetic and taxon-based techniques at different depths of coverage). For pyrosequencing studies of 16S rRNA, depth of coverage of about 1000 sequences/sample seems to provide a good balance between number of samples and depth of sampling. This

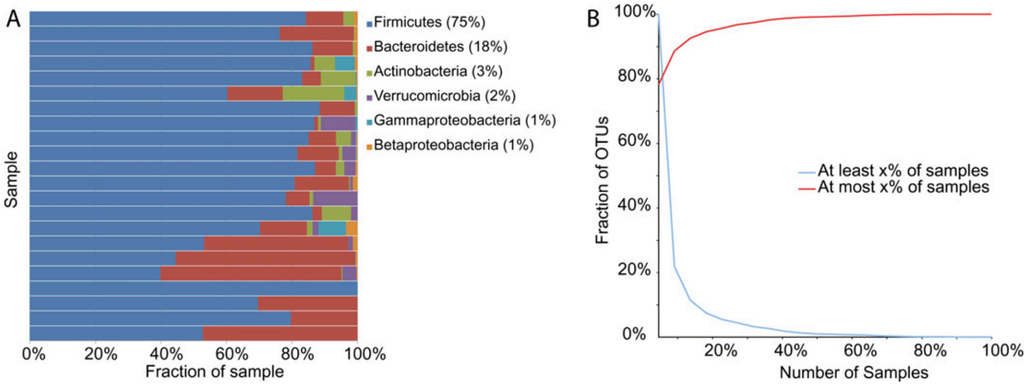

Figure 3. (*A*) Phylum-level abundance and (*B*) shared "species" (represented here as 97% OTUs, approximately species level) in 22 human gut samples with depth of coverage of at least 350 sequences per individual. These data are taken from a meta-analysis (Ley et al. 2008a) covering several large Sanger-sequencing studies of humans in different populations (Suau et al. 1999; Hayashi et al. 2002a,b, 2003; Eckburg et al. 2005; Ley et al. 2006c; Nagashima et al. 2006). Interestingly, the results are very consistent with results from both Sanger sequencing and pyrosequencing within a North American population of lean and obese twins (Turnbaugh et al. 2009). Note: No species-level OTUs were shared across all samples with 350 sequences per sample; 1813 OTUs were only present in one sample; the total number of OTUs was 2320.

depth of coverage allows us to infer the frequencies of species at 1% abundance with reasonable accuracy (we expect to observe 10 sequences at this level, with standard deviation of 3.1 sequences), although it will miss many of the rare species.

If the goal is complete characterization of all sequences in a sample, vast numbers of sequences may be required if many species are rare or if the diversity is high, such as in seawater and soils (Sogin et al. 2006; Roesch et al. 2007). Such a tiny fraction of the total number of cells is sampled (a full 454 run currently recovers ~5 × 10^5 sequences, so if there are ~10^{14} microbes in the gut, only about five cells in every billion are sampled, assuming one rRNA molecule per cell) that characterizing the full diversity is not a reasonable goal.

It is important to note that there is no abundance threshold below which we can disregard microbes as unimportant: Many pathogens are rare and can be detected at low levels using PCR-based assays in clinically relevant settings. Thus, the appropriate trade-off between depth of coverage and number of samples is likely to vary depending on the goal of the study.

It is likely that the frequent practice of using a full FLX run for metagenomic characterization is inefficient: If the goal is to infer the frequencies of approximately 250 functional categories, a few tens of thousands of sequences (rather than hundreds of thousands of sequences) should suffice. Fewer sequences may be needed once longer reads are routinely available with the Titanium, because more of these longer reads can be assigned functions. Details about rare genes, or about the coverage of particular functions by rare organisms, will be missed by this shallow coverage, but clustering of communities based on overall functions should still be possible. The prospect of extending metagenomic studies to dozens of samples per run could transform our understanding of how functional communities are assembled from different components.

Which region of the rRNA should I sequence?

Because the full 16S rRNA cannot be sequenced using high-throughput methods, a shorter region of the sequence must be selected to act as proxy. Currently, there is no consensus on a single "best" region, and consequently different groups are sequencing different regions (or multiple regions). This diversity of methods hinders direct comparisons among studies, and we recommend standardization on a single region. Of the nine variable regions (Neefs et al. 1990), several of the more popular regions include the regions surrounding V2, V4, and V6: In general, a combination of variable and moderately conserved regions seems to be optimal for performing analyses at different phylogenetic depths. Wang et al. (2007) and Liu et al. (2008) report that V2 and V4 give the lowest error rates when assigning taxonomy, and these regions are also suitable for community clustering (Liu et al. 2007). However, several other regions are also in use (Baker et al. 2003; Edwards et al. 2006; Sogin et al. 2006; Huse et al. 2007; Roesch et al. 2007; Andersson et al. 2008).

Both the choice of region and the design of the primers are critical, and poor choice of primers can lead to radically different biological conclusions (Andersson et al. 2008; Liu et al. 2008). It is well known that primer bias due to differential annealing leads to the over- or underrepresentation of specific taxa, and some groups can be missed entirely if they match the consensus sequence poorly (von Wintzingerode et al. 1997; Kanagawa 2003). Solutions proposed include using mixtures of large numbers of primers (Frank et al. 2008), although this approach introduces substantial additional complexity and free parameters (e.g., the ratios in which the primers are mixed). Issues of primer bias can be important. For example, although some widely used primers such as 8F (Meyer et al. 2004; Edwards et al. 2006; Frank et al. 2008), 337F (Huse et al. 2007), 338R (Harris et al. 2004; Hamady et al. 2008), 515F (Marcille et al. 2002; Meyer et al. 2004), 915F (Marcille et al. 2002), 930R (Marcille et al. 2002), 1046R (Sogin et al. 2006), and 1061R (Andersson et al. 2008) match >95% of the sequences in RDP from all of the major bacterial phyla in the gut (Firmicutes, Bacteroidetes, Actinobacteria, Verrucomicrobia, and Proteobacteria), others miss specific divisions: 784F (Andersson et al. 2008) is biased against Verrucomicrobia, 967F (Sogin et al. 2006) matches <5% of Bacteroidetes, and 1492R (Meyer et al. 2004) matches 61% of Actinobacteria, 54% of Proteobacteria, and fewer than half of the other divisions. These matches were measured using ProbeMatch on the Ribosome Database Project (RDP) site (http://rdp.cme.msu.edu/probematch), using a search restricted to "good" sequences with coverage over the primer range and allowing up to two mismatches. Comparisons of relative abundance among

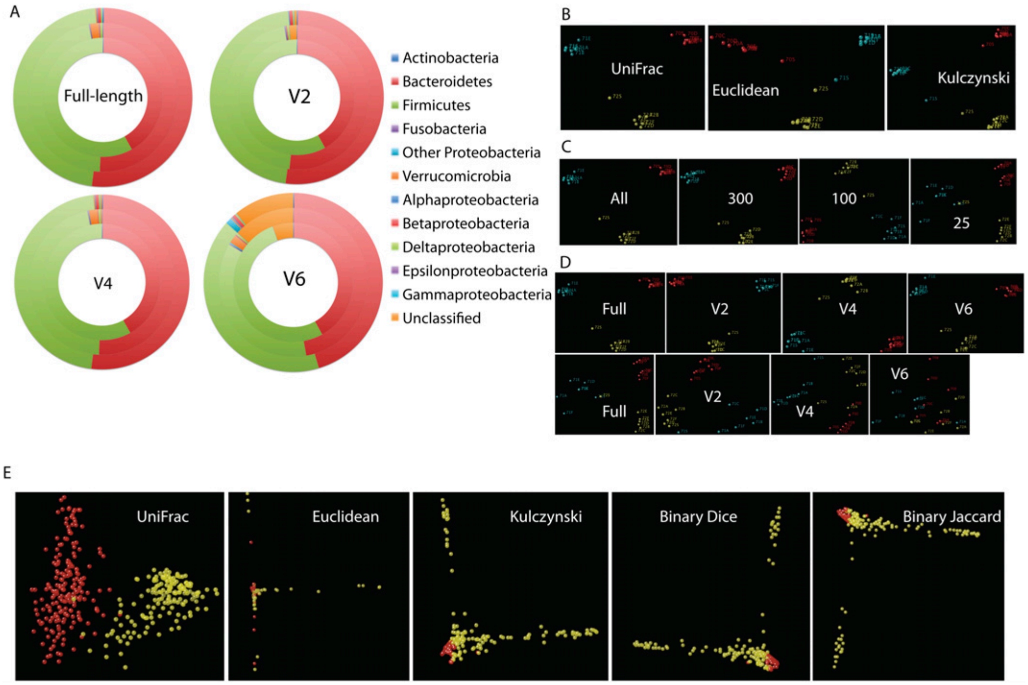

Figure 4. Comparison of phylogenetic and nonphylogenetic methods for comparing communities. (*A–D*) Sequences are from stool and six different biopsy sites along the distal gut from three unrelated healthy human subjects (Eckburg et al. 2005); (*E*) sequences are from 162 free-living communities and 159 vertebrate gut communities (Ley et al. 2008b). Fragments are labeled as either full-length, V2 or V4 (250-nt reads ending at 338R or starting at 515F, respectively), or V6 (80-nt reads ending at 1046R). (*A*) Effect of fragment on phylogenetic assignment: Each circle is one of the three individual human subjects, pooling sequences from all sites. Note increase in unclassified reads produced by V6; results from V2 and V4 are very similar to those obtained from the full-length sequences. Assignments performed using RDP. (*B*) Effect of three different distance measures for principal coordinates analysis on the full-length 16S rRNA sequence data: UniFrac (a phylogenetic method), and Euclidean and Kulczynski distances on the sample by OTU matrix (two examples of taxon-based methods). Only the relative positions of and distances between points are relevant: The choice of direction along each axis is a mathematical artifact. Individual points are samples, colored according to the three subjects that the samples came from (i.e., the three colors represent three subjects: The same color scheme is used for panels *C* and *D*). In this data set, all methods give broadly equivalent results and cluster the samples by individual, not by sample location (stool or individual sites along the distal gut mucosa). (*C*) Effect of reducing the number of sequences per sample on the UniFrac clustering, comparing the results obtained using all sequences to results obtained using a random sample of sequences. (*Right* panel) Clustering is still good, as measured by the consistency of clustering together the samples from the same individual as in panel *B*, at 25 sequences per sample, although there is more scatter as the number of sequences per sample decreases. (*D*) Effect of the different regions on clustering with UniFrac using either (*top* row) all sequences or (*bottom* row) 25 sequences/sample. For this analysis, we take each full-length sequence, computationally clip out the part of the sequence corresponding to each region to simulate 454 data, and repeat the analysis: The analysis thus includes the effect of the region sequenced, but not the effect of primer bias that may differentially amplify specific taxa. Again, we expect the samples from each individual to cluster together, and a mixture of samples from different individuals indicates poor performance. V6 is especially affected at low sample coverage, and V2 is especially unaffected. (*E*) Effect of different clustering measures, indicated on each panel, on the data set from Ley et al. (2008a), showing only the (yellow) vertebrate gut and (red) free-living samples. This data set is very heterogeneous and includes many samples with low numbers of sequences per sample or where nonoverlapping regions of the 16S rRNA were chosen for sequencing. In this data set, UniFrac, which is a phylogenetic metric, performs very well, separating the samples into two groups; in contrast, the other three methods, which are all taxon based, perform poorly with obvious clustering artifacts such as spikes leading off at right angles from one another, and fail to separate the two types of samples into two discrete clusters. Note that this figure is not based on the Arb parsimony insertion tree used in Ley et al. (2008a) but rather on a tree constructed de novo from the NAST-aligned sequences using Clearcut (Sheneman et al. 2006). The artifacts in the taxon-based methods are due to lack of overlap at the species level among different kinds of samples. An exploration of primer effects in a subset of these data shows that sample type is more important than region sequenced or length of amplicon (Liu et al. 2007).

different studies should thus be treated with caution. However, meta-analysis of presence/absence data from different studies is extremely useful for revealing broad trends, even when different studies use different primers (Lozupone and Knight 2007; Ley et al. 2008b).

With so many new bacterial phyla being discovered by culture-independent methods (Rappe and Giovannoni 2003; Ley et al. 2006a), it is very likely that we are still missing many important components of the biosphere with current primer sets. As more sequence data and better taxonomic assignments become available, improved primer sets, with better coverage (including primers for archaea and eukaryotes), will likely provide a substantial advance over current degenerate primer techniques. In particular, 16S rRNA reads from metagenomic studies provide a source of sequences that is not subject to PCR primer bias (although other biases may, of course, be present) and therefore covers taxa that are missed by existing but popular primer sets. Another promising approach is the use of miniprimers (Isenbarger et al. 2008), which, together with an engineered DNA polymerase, may allow greater coverage of desired groups.

Should I analyze individual samples or pool samples before sequencing?

One widespread approach in clinical and environmental studies is to reduce variability due to idiosyncratic effects by pooling DNA from the samples before PCR. The hypothesis is that this pooling will make the differences between groups more apparent and also reduce the overall cost of the analysis. With the availability of barcoded pyrosequencing, we strongly recommend against pooling in favor of tagging each sample with a unique barcode. The pooling can always be performed computationally at the end of the analysis, and, given the variability in 16S rRNA lineages observed among samples in many mammalian-associated body habitats including the gut (Eckburg et al. 2005; Ley et al. 2005, 2006c, 2008a; Turnbaugh et al. 2006, 2008; Frank et al. 2007; Li et al. 2008), vagina (Coolen et al. 2005; Hyman et al. 2005; Zhou et al. 2007), vulva (Brown et al. 2007), breast milk (Martin et al. 2007), the mouth (Aas et al. 2005; Nasidze et al. 2009), and skin (Gao et al. 2007; Grice et al. 2008), it is crucial to know whether the apparent differences between groups are due to a consistent shift in each sample or to very large shifts in a small subpopulation of samples. When heterogeneity is high, less accurate data about a large number of samples will be much more informative than more accurate data about a small number of samples, or about pooled samples, especially when developing biomarkers for classification of healthy and diseased individuals.

How should I analyze my data?

Once the sequences are collected, the next challenge is data analysis. Especially with pyrosequencing, many established tools for alignment, phylogenetic inference, and defining taxonomic groups by sequence similarity (OTUs, or operational taxonomic units) cannot handle the vast data sets produced.

How should I filter out low-quality reads?

There are two main issues with low-quality data: errors in the sequence and chimeras resulting from recombination between sequences. Although error rates for pyrosequencing were relatively high with the older GS 20 instrument, they are lower for the GS FLX (Droege and Hill 2008). On a known template, the error rate was 0.4% overall, but 96% of these errors were insertions or deletions in homopolymer runs rather than nucleotide substitutions in high-complexity regions, leading to a substitution rate of 0.042% (Quinlan et al. 2008). Most errors are concentrated in a few exceptionally bad reads; thus the usual, conservative practice is to discard sequences that contain any errors in the primer (including the barcode, if present) and sequences where the average quality score is below 25 (Huse et al. 2007). An update of the Huse et al. (2007) study for the FLX and Titanium methods would be a valuable addition to the literature.

Chimeras can be detected in small 16S rRNA data sets using several techniques (Huber et al. 2004; Ashelford et al. 2006), although no solution yet exists for the large data sets produced by pyrosequencing. Publicly available databases are unfortunately rife with chimeric sequences (DeSantis et al. 2006a), so some caution with taxonomy assignment is warranted. However, although chimeras can affect estimates of diversity within a sample, because they are generated uniquely within each sample, they have relatively little effect on patterns of similarities and differences among communities (Ley et al. 2008a).

Should I perform taxon-based or phylogenetic analyses?

In general, ecological analyses of diversity can be split along three major axes (Magurran 2004; Ley et al. 2008b). First, an analysis can examine either "alpha diversity" (how many kinds of taxa or lineages are in one sample) or "beta diversity" (how taxa or lineages are shared among samples, e.g., along a gradient). Second, an analysis can be either "qualitative," examining only presence-absence data, or "quantitative," also taking into account relative abundance. (Qualitative analyses and quantitative analyses are also called analyses of community membership and community structure, respectively, although community structure is sometimes also considered to include spatial or temporal structure.) Third, an analysis can be either "phylogenetic," making use of a phylogenetic tree to relate the sequences, or "taxon based," treating all taxa at a given rank (e.g., species) as phylogenetically equivalent. In practice, most sequences come from uncultured microbes that have not been formally described, and so taxa are operationally defined by sequence similarity. For example, 97% of OTUs contain sequences that have 97% sequence identity. In this classification, the widely used Chao1 index (Chao 1984) for estimating the minimum number of species in a sample is a quantitative, taxon-based, alpha diversity metric; unweighted UniFrac (Lozupone and Knight 2005) is a qualitative, phylogenetic, beta diversity metric. In general, taxon-based and phylogenetic methods provide different but equally useful insights, and both should be performed.

Taxon-based analyses are especially useful for asking how many different "species" (or other taxonomic units) are likely to be in a sample (Chao 1984; Schloss and Handelsman 2005), for comparing which OTUs are shared among particular subsets of samples (Schloss and Handelsman 2006), or for building networks that relate species and samples to one another (Ley et al. 2008a). (But see below for discussion of how the OTU selection method can affect the results.) Incidence matrices recording which OTUs occurred in each sample can also be used as input to standard community clustering methods (Magurran 2004), although, in our experience, phylogenetic methods tend to be more illuminating for community clustering when samples are extremely heterogeneous and when the number of sequences per sample is low (Fig. 4). The main reason for the increased power of phylogenetic methods in this context is that taxon-based methods are not free of assumptions about phylogeny; rather, they implicitly assume the so-called "star phylogeny," in which all taxa are equally related to one another. This assumption is problematic because it ignores the correlation between evolutionary relatedness and ecological similarity. Although errors in phylogenetic reconstruction can affect the clustering results, regardless of reconstruction method, a tree will provide a more accurate model of the real data than will the star phylogeny. In practice, different phylogenetic reconstruction methods give similar results (Lozupone et al. 2007), so speed of reconstruction is usually the overriding concern (especially with large data sets). Phylogenetic methods are also useful for asking how much evolutionary history is unique to a particular sample and for identifying samples likely to contain additional unique deep-branching lineages (Lozupone and Knight 2007).

What level of taxonomic classification is important?

Depending on the question to be answered, taxonomic differences at different depths might be important. For detecting pathogens, 16S rRNA has important limitations: Approaches that cover many genes, such as MLST (multi-locus sequence typing) (Maiden et al. 1998), can be critical for resolving drug-resistant or virulent strains from their close but less harmful relatives. However, these techniques typically rely on detailed knowledge of which organisms in the sample are of interest, and are less useful for broad phylogenetic characterization or discovery of new phyla or other high-level taxa. One advantage of the 16S rRNA gene is that it contains both fast- and slow-evolving regions, so it can be used to resolve phylogenetic relationships at different depths. However, with short sequence reads, only specific sections can be chosen for sequencing, and some regions recapture the patterns obtained with the full-length sequences much better than others. In general, however, it is not yet known whether we should concern ourselves with strain-level, species-level, genus-level, or higher-order differences among samples when searching for differences in functions. The human gut microbiota consists of a relatively small number of deep-branching taxa with enormous diversity at the tips (Ley et al. 2006b). On the other hand, given that the gut microbiota of different humans show very different patterns of abundance even at the phylum level (Fig. 3), pursing these large differences seems an obvious first step. It is likely that only additional data about microbiomes in health and disease will resolve this issue.

The above discussion assumes that we can identify the taxa accurately and integrate them into an existing taxonomy. However, there are several competing taxonomies that differ substantially (DeSantis et al. 2006a), and several different algorithms for grouping sequences by similarity into OTUs that differ radically in their results (Schloss and Handelsman 2005). Two popular methods for selecting OTUs are the nearest-neighbor algorithm (in which a sequence is added to an OTU if it is similar to any sequence already in that OTU), and the furthest-neighbor algorithm (in which a sequence is added to an OTU if it is similar to all other sequences already in that OTU). In general, these algorithms can produce remarkably different results (Fig. 5). In pyrosequencing data, the nearest-neighbor algorithm often produces a single huge OTU, although the basis for this effect is still unknown (M. Hamady and R. Knight, unpubl.).

The final issue is to choose a method for inserting new, unclassified sequences into the taxonomy. There are several general methods for performing this task: Sequences can be matched by similarity to an existing sequence in the database by BLAST, pairwise alignment, or by the count of oligonucleotide frequencies; or they can be inserted into a phylogenetic tree and then assigned to the group that they fall into. In practice, different taxonomies produce large differences in the representation of different bacterial groups in a sample, but once a taxonomy is chosen, the various methods for inserting sequences into that taxonomy usually give consistent results (Liu et al. 2008). The RDP classifier, which matches sequences to groups using oligonucleotide frequencies, is especially fast (Wang et al. 2007), although the Greengenes classifier (DeSantis et al. 2006a), which uses a BLAST-based approach, may provide comparable or better accuracy. (The SILVA classification workflow [Pruesse et al. 2007], which uses alignment and nearest-neighbor matching or tree insertion, was released after this comparison was performed, but is also likely to be useful.) These methods are typically used to make a pie chart or bar graph showing the representation of each phylum in each sample, as in Figure 3. One major emerging issue is the massive amount of uncharacterized data in the public databases: When trying to classify newly sequenced data by BLAST, query sequences are far more likely to hit sequences with vague annotations such as "uncultured soil bacterium" rather than a cultured isolate with well-characterized genetics and biochemistry. Overcoming the challenges associated with storing metadata about each sample, especially using a consistent structured vocabulary and ontology for sample annotation, will be critical for making the data collected by the HMPs a useful resource (Garrity et al. 2008).

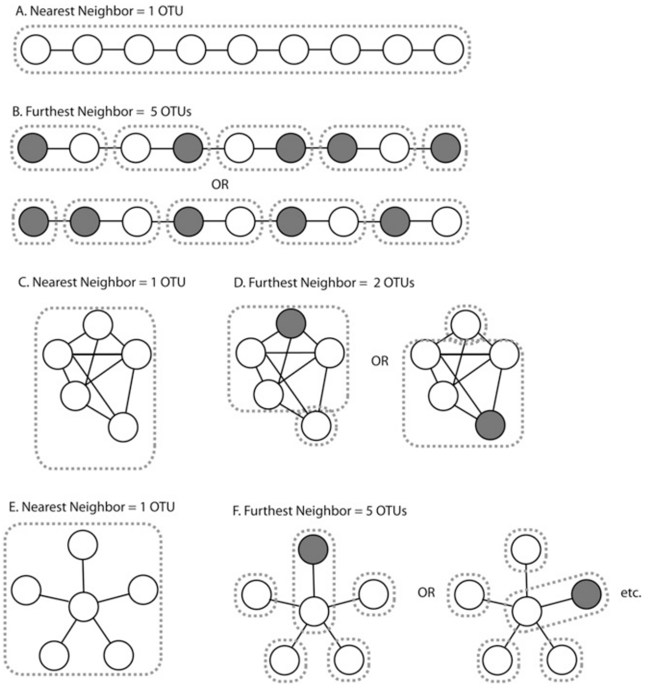
A. Nearest Neighbor = 1 OTU
B. Furthest Neighbor = 5 OTUs
OR
C. Nearest Neighbor = 1 OTU
D. Furthest Neighbor = 2 OTUs
OR
E. Nearest Neighbor = 1 OTU
F. Furthest Neighbor = 5 OTUs
OR
etc.

Figure 5. Different methods for selecting OTUs produce different results. (*A*,*B*) If sequences are arranged so that each sequence has a neighbor within the OTU threshold (e.g., 97%) but these neighbors are not similar to one another, that is, the variation is not in the same direction, (*A*) the nearest-neighbor algorithm will produce one OTU because every sequence is connected to every other sequence through a chain of neighbors within threshold, but (*B*) the furthest-neighbor algorithm will produce a series of OTUs where all members of each OTU are similar to one another. (OTU boundaries are indicated by dashed lines.) Note that the precise OTUs produced by the furthest-neighbor algorithm will vary every time owing to the choice of the randomly chosen seed sequence for each (gray) OTU. (*C*,*D*) If sequences are arranged so that there is one outlier that is within threshold of only one of the other sequences, (*C*) the nearest-neighbor algorithm will produce a single OTU, (*D*) but the furthest-neighbor algorithm will produce two OTUs in which one of the two most distant sequences is excluded from the main OTU at random. (*E*,*F*) If sequences are arranged so that all sequences are within threshold of a central sequence but are outside threshold from each other, (*E*) the nearest-neighbor algorithm will again produce one OTU, (*F*) but the furthest-neighbor algorithm will group one sequence with the central sequence at random and break the other sequences into their own OTUs.

At the metagenomic level, a typical approach is to produce a heat map showing the abundance of each function or each taxonomic group in each metagenomic sample, and use standard (nonphylogenetic) clustering techniques to relate the samples to one another according to the functions they contain (Tringe et al. 2005; Turnbaugh et al. 2006, 2008, 2009; Huson et al. 2007; Dinsdale et al. 2008; Schloss and Handelsman 2008). Extension of phylogenetic analysis techniques to these kinds of data (Lozupone et al. 2008) is likely to allow improved resolution, as they have for 16S rRNA analyses. Metagenomic sequences also provide important opportunities for emerging techniques such as metabolic network reconstruction (Markowitz et al. 2008; Meyer et al. 2008a) and multivariate analyses that relate changes in pathway representation to environmental gradients (Gianoulis et al. 2009): Combining multiple approaches is essential at this point.

Is there a core? Key themes emerging from recent studies of the microbiome

The major challenge facing the various Human Microbiome Projects is: How can we best relate differences in community composition to differences in function, especially for relating microbial changes to human health and disease?

New techniques, including high-throughput phylogenetic methods and network-based analyses, allow us to answer large-scale questions about placing human microbiomes in context that previously could not be addressed. In particular, the availability of hundreds of samples covered at thousands of sequences/sample allows us to characterize the variability among different humans, which is essential for providing a baseline for studies of microbes associated with specific diseases.

Different people harbor remarkably dissimilar microbiota in their guts (Fig. 2; see also Eckburg et al. 2005; Frank et al. 2007; Ley et al. 2008a; Turnbaugh et al. 2009), in their saliva (Nasidze et al. 2009), and on their hands (Fierer et al. 2008; Grice et al. 2008), both in terms of which species are present and in terms of the relative ratios of the bacterial phyla. This amazing degree of variability among individuals will greatly complicate the Human Microbiome Project's ability to deliver on the promise of identifying microbes that are biomarkers for specific diseases. Studies of the guts of different individual mice (Ley et al. 2005) and macaques (Hoffmann et al. 2007) reinforce the point that this variability within species is likely to be a recurring theme. In particular, comparative studies must be very cautious in attributing observed differences in the microbiota to differences among species rather than to differences among individuals, although when dozens of species are studied, the broad-scale trends will be apparent even with one individual per species (Ley et al. 2008a).

Our current depth of coverage allows us to rule out the possibility that all humans share any species (approximated here as 97% OTUs) in the gut at the 1% level of abundance, and that all hands share any bacterium at the 2% level of abundance. This calculation is performed as follows (Turnbaugh et al. 2008). Assume that the true level of abundance of a species *s* in all samples is *x*%. We can then calculate the probability that in *n* sequences we missed *s* completely using Poisson sampling statistics: Pr(missed) = e^{-xn}. If we then have *N* samples of equal size, we can calculate the probability that we only observed the species in at most *n* of the *N* samples using the binomial distribution. We can then vary *x* until the probability of missing *s* in the actual number of samples *N* is a specified significance level α, say, 0.05. If sample sizes are unequal, the binomial distribution cannot be used, but an empirical distribution can be obtained by simulation. Using this approach on the human gut samples from Ley et al. (2008a), evenly sampled, we can rule out the possibility that any species is at more than 0.9% in the gut of all humans studied with depth of coverage of at least 350 sequences/sample. This result remains consistent using pyrosequencing data at several thousand amplicon sequences/fecal sample (Turnbaugh et al. 2009). Using this approach on human hand samples (Fierer et al. 2008), we can rule out the possibility that any species is at more than 2% on all hands sampled. This latter result is especially surprising because on average one species comprises 37% of the community on a given hand. It is possible that (1) we all share a much greater proportion of these species at much lower abundance that could be detected with deeper sequencing or with qPCR or culture-based techniques, and (2) many species are very abundant in individual samples. However, we can already rule out the possibility that there are species that we all share at high abundance. Nonetheless, at the phylum level, we all share the same few groups in a given body habitat (Ley et al. 2006b; Dethlefsen et al. 2007), albeit at radically different levels of abundance. At what taxonomic level, if any, will we start to see a shared core among humans, and/or a set of human-specific lineages that differentiates us from other mammals?

One intriguing recent result is that species-level variability appears to be associated with extensive functional redundancy, in which completely different microbial communities converge on the same functional state. For example, at the metagenomic level, different habitats (soils, lakes, etc.) converge on similar functional gene repertoires (Dinsdale et al. 2008). Using the same set of samples from lean and obese twins, completely different species assemblages appear to lead to very similar functional profiles, as measured by the representation of KEGG pathways (Turnbaugh et al. 2009). This result has a clear parallel in macroecosystems: For example, a grassland in North America and a grassland in Africa will share many obvious ecological similarities (with respect to, say, forests) yet will have none of their species in common and may have some ecosystem functions, such as pollination, performed by phylogenetically independent guild members (e.g., bees versus bats). Exploration of this relationship between species assemblages and ecosystem function will be a key result of microbiome studies and may provide new insights into assembly of a wide range of ecosystems.

Thus, we can rule out the possibility that there is a large core microbiome at the species level, although at higher-order taxonomic levels (e.g., phylum), the communities begin to resemble one another more (although there is still immense variability in, e.g., the ratio of Firmicutes to Bacteroidetes). There may be a consistent functional signature, however, and discovering these relationships to metabolic function (Li et al. 2008; Turnbaugh et al. 2009) will be an especially important outcome of HMPs. Untangling the relationships among these very high-dimensional data sets will also require methods developed to handle millions or even billions of sequences from thousands to millions of samples. Key limiting factors will be algorithms to reduce computational complexity and ontologies to allow convenient retrieval of relevant data sets from vast numbers of community samples. Interdisciplinary studies combining ecology, microbiology, advanced computational methods, genomics, culture-based studies, and so on will be essential (Ley et al. 2008b). At this stage, we need to be judicious in interpreting microbiome results because so much variability is being uncovered, but the prospects for discovery and impacts on human health are inspiring.

Acknowledgments

We thank Cathy Lozupone, Jeffrey Gordon, Ruth Ley, Peter Turnbaugh, Elizabeth Costello, Daniel McDonald, and Justin Kuczynski for helpful discussions and comments on drafts of the manuscript; and Sean Eddy and Manuel Lladser for helpful discussions about the statistical analysis. M.H. was supported by an NSF EAPSI fellowship (OISE0812861) and the NIH Molecular Biophysics Training Grant (T32GM065103). The ideas in this review arose from work in the Knight and Gordon laboratories supported by the NIH (P01DK078669), the Crohn's and Colitis Foundation of America, the Bill and Melinda Gates Foundation, and from discussion at the Banbury Center Conference "Living on Human Beings: Metagenomic Approaches and Challenges," March 2–5, 2008.

References

Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. 2005. Defining the normal bacterial flora of the oral cavity. *J Clin Microbiol* **43:** 5721–5732.

Anderson IC, Cairney JW. 2004. Diversity and ecology of soil fungal communities: Increased understanding through the application of molecular techniques. *Environ Microbiol* **6:** 769–779.

Andersson AF, Lindberg M, Jakobsson H, Backhed F, Nyren P, Engstrand L. 2008. Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One* **3:** e2836. doi: 10.1371/journal.pone.0002836.

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. 2006. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl Environ Microbiol* **72:** 5734–5741.

Baker GC, Smith JJ, Cowan DA. 2003. Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods* **55:** 541–555.

Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, Nielsen R, Willerslev E. 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One* **2:** e197. doi: 10.1371/journal.pone.0000197.

Brown CJ, Wong M, Davis CC, Kanti A, Zhou X, Forney LJ. 2007. Preliminary characterization of the normal microbiota of the human vulva using cultivation-independent methods. *J Med Microbiol* **56:** 271–276.

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

Church GM, Kieffer-Higgins S. 1988. Multiplex DNA sequencing. *Science* **240:** 185–188.

Coolen MJ, Post E, Davis CC, Forney LJ. 2005. Characterization of microbial communities found in the human vagina by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes. *Appl Environ Microbiol* **71:** 8729–8737.

Dalevi D, Ivanova NN, Mavromatis K, Hooper SD, Szeto E, Hugenholtz P, Kyrpides NC, Markowitz VM. 2008. Annotation of metagenome short reads using proxygenes. *Bioinformatics* **24:** i7–i13.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006a. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72:** 5069–5072.

DeSantis, TZ Jr, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, Phan R, Andersen GL. 2006b. NAST: A multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res* **34:** W394–W399.

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.

Dethlefsen L, McFall-Ngai M, Relman DA. 2007. An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature* **449:** 811–818.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, et al. 2008. Functional metagenomic profiling of nine biomes. *Nature* **452:** 629–632.

Dollhopf SL, Hashsham SA, Tiedje JM. 2001. Interpreting 16S rDNA T-RFLP data: Application of self-organizing maps and principal component analysis to describe community dynamics and convergence. *Microb Ecol* **42:** 495–505.

Droege M, Hill B. 2008. The Genome Sequencer FLX System—longer reads, more applications, straight forward bioinformatics and more complete data sets. *J Biotechnol* **136:** 3–10.

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. 2005. Diversity of the human intestinal microbial flora. *Science* **308:** 1635–1638.

Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC Jr, Rohwer F. 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7:** 57. doi: 10.1186/1471-2164-7-57.

Fierer N, Jackson RB. 2006. The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci* **103:** 626–631.

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* **105:** 17994–17999.

Fisher MM, Triplett EW. 1999. Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol* **65:** 4630–4636.

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* **104:** 13780–13785.

Frank JA, Reich CI, Sharma S, Weisbaum JS, Wilson BA, Olsen GJ. 2008. Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl Environ Microbiol* **74:** 2461–2470.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, Delong EF. 2008. Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci* **105:** 3805–3810.

Gao Z, Tseng CH, Pei Z, Blaser MJ. 2007. Molecular analysis of human forearm superficial skin bacterial biota. *Proc Natl Acad Sci* **104:** 2927–2932.

Garrity GM, Field D, Kyrpides N, Hirschman L, Sansone SA, Angiuoli S, Cole JR, Glockner FO, Kolker E, Kowalchuk G, et al. 2008. Toward a standards-compliant genomic and metagenomic publication record. *OMICS* **12:** 157–160.

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I, Yamada T, Paccanaro A, Jensen LJ, Snyder M, et al. 2009. Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci* **106:** 1374–1379.

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. 2006. Metagenomic analysis of the human distal gut microbiome. *Science* **312:** 1355–1359.

Grice EA, Kong HH, Renaud G, Young AC, Bouffard GG, Blakesley RW, Wolfsberg TG, Turner ML, Segre JA. 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.

Harris JK, Kelley ST, Pace NR. 2004. New perspective on uncultured bacterial phylogenetic division OP11. *Appl Environ Microbiol* **70:** 845–849.

Hayashi H, Sakamoto M, Benno Y. 2002a. Fecal microbial diversity in a strict vegetarian as determined by molecular analysis and cultivation. *Microbiol Immunol* **46:** 819–831.

Hayashi H, Sakamoto M, Benno Y. 2002b. Phylogenetic analysis of the human gut microbiota using 16S rDNA clone libraries and strictly anaerobic culture-based methods. *Microbiol Immunol* **46:** 535–548.

Hayashi H, Sakamoto M, Kitahara M, Benno Y. 2003. Molecular analysis of fecal microbiota in elderly individuals using 16S rDNA library and T-RFLP. *Microbiol Immunol* **47:** 557–570.

He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, Huang Z, Wu W, Gu B, Jardine P, et al. 2007. GeoChip: A comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME J* **1:** 67–77.

Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, Bushman FD. 2007. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* **35:** e91. doi: 10.1093/nar/gkm435.

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML. 2007. Microbial population structures in the deep marine biosphere. *Science* **318:** 97–100.

Huber T, Faulkner G, Hugenholtz P. 2004. Bellerophon: A program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20:** 2317–2319.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8:** R143. doi: 10.1186/gb-2007-8-7-r143.

Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Res* **17:** 377–386.

Hyman RW, Fukushima M, Diamond L, Kumm J, Giudice LC, Davis RW. 2005. Microbes on the human vaginal epithelium. *Proc Natl Acad Sci* **102:** 7952–7957.

Isenbarger TA, Finney M, Rios-Velazquez C, Handelsman J, Ruvkun G. 2008. Miniprimer PCR, a new lens for viewing the microbial world. *Appl Environ Microbiol* **74:** 840–849.

Kanagawa T. 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng* **96:** 317–323.

Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL, Roitman DB, Pham TT, Otto GA, Foquet M, Turner SW. 2008. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc Natl Acad Sci* **105:** 1176–1181.

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J. 2008. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res* **36:** 2230–2239.

Lederberg J. 2000. Infectious history. *Science* **288:** 287–293.

Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. 2005. Obesity alters gut microbial ecology. *Proc Natl Acad Sci* **102:** 11070–11075.

Ley RE, Harris JK, Wilcox J, Spear JR, Miller SR, Bebout BM, Maresca JA, Bryant DA, Sogin ML, Pace NR. 2006a. Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. *Appl Environ Microbiol* **72:** 3685–3695.

Ley RE, Peterson DA, Gordon JI. 2006b. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124:** 837–848.

Ley RE, Turnbaugh PJ, Klein S, Gordon JI. 2006c. Microbial ecology: Human gut microbes associated with obesity. *Nature* **444:** 1022–1023.

Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, 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.

Li M, Wang B, Zhang M, Rantalainen M, Wang S, Zhou H, Zhang Y, Shen J, Pang X, Wei H, et al. 2008. Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci* **105:** 2117–2122.

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. doi: 10.1093/nar/gkm541.

Liu Z, DeSantis TZ, Andersen GL, Knight R. 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* **36:** e120. doi: 10.1093/nar/gkn491.

Lo I, Denef VJ, Verberkmoes NC, Shah MB, Goltsman D, DiBartolo G, Tyson GW, Allen EE, Ram RJ, Detter JC, et al. 2007. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446:** 537–541.

Lozupone C, Knight R. 2005. UniFrac: A new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71:** 8228–8235.

Lozupone CA, Knight R. 2007. Global patterns in bacterial diversity. *Proc Natl Acad Sci* **104:** 11436–11440.

Lozupone C, Hamady M, Knight R. 2006. UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7:** 371. doi: 10.1186/1471-2105-7-371.

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.

Lozupone CA, Hamady M, Cantarel BL, Coutinho PM, Henrissat B, Gordon JI, Knight R. 2008. The convergence of carbohydrate active gene repertoires in human gut microbes. *Proc Natl Acad Sci* **105:** 15076–15081.

Magurran AE. 2004. *Measuring biological diversity.* Blackwell, Oxford, UK.

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, et al. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci* **95:** 3140–3145.

Marcille F, Gomez A, Joubert P, Ladire M, Veau G, Clara A, Gavini F, Willems A, Fons M. 2002. Distribution of genes encoding the trypsin-dependent lantibiotic ruminococcin A among bacteria isolated from human fecal microbiota. *Appl Environ Microbiol* **68:** 3424–3431.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I, et al. 2008. IMG/M: A data management and analysis system for metagenomes. *Nucleic Acids Res* **36:** D534–D538.

Martin R, Heilig HG, Zoetendal EG, Jimenez E, Fernandez L, Smidt H, Rodriguez JM. 2007. Cultivation-independent assessment of the bacterial diversity of breast milk among healthy women. *Res Microbiol* **158:** 31–37.

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, et al. 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* **4:** 495–500.

McKenna P, Hoffmann C, Minkah N, Aye PP, Lackner A, Liu Z, Lozupone CA, Hamady M, Knight R, Bushman FD. 2008. The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathog* **4:** e20. doi: 10.1371/journal.ppat.0040020.

Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R. 2008. Microbiology in the post-genomic era. *Nat Rev Microbiol* **6:** 419–430.

Meyer AF, Lipson DA, Martin AP, Schadt CW, Schmidt SK. 2004. Molecular and metabolic characterization of cold-tolerant alpine soil *Pseudomonas sensu stricto*. *Appl Environ Microbiol* **70:** 483–489.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, et al. 2008a. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9:** 386. doi: 10.1186/1471-2105-9-386.

Meyer M, Stenzel U, Hofreiter M. 2008b. Parallel tagged sequencing on the 454 platform. *Nat Protocols* **3:** 267–278.

Nagashima K, Mochizuki J, Hisada T, Suzuki S, Shimomura K. 2006. Phylogenetic analysis of 16S ribosomal gene sequences from human fecal microbiota and improved utility of terminal restriction fragment length polymorphism profiling. *Biosci. Microflora* **25:** 99–107.

Nasidze I, Li J, Quinque D, Tang K, Stoneking M. 2009. Global diversity in the human salivary microbiome. *Genome Res*. doi: 10.1101/gr.084616.108.

Neefs JM, Van de Peer Y, Hendriks L, De Wachter R. 1990. Compilation of small ribosomal subunit RNA sequences. *Nucleic Acids Res* (Suppl.) **18:** 2237–2317.

Pace NR. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276:** 734–740.

Palmer C, Bik EM, Digiulio DB, Relman DA, Brown PO. 2007. Development of the human infant intestinal microbiota. *PLoS Biol* **5:** e177. doi: 10.1371/journal.pbio.0050177.

Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, Ronaghi M, Fire AZ. 2007. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* **35:** e130. doi: 10.1093/nar/gkm760.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO. 2007. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35:** 7188–7196.

Quinlan AR, Stewart DA, Stromberg MP, Marth GT. 2008. Pyrobayes: An improved base caller for SNP discovery in pyrosequences. *Nat Methods* **5:** 179–181.

Rappe MS, Giovannoni SJ. 2003. The uncultured microbial majority. *Annu Rev Microbiol* **57:** 369–394.

Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1:** 283–290.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242:** 84–89.

Ronaghi M, Uhlen M, Nyren P. 1998. A sequencing method based on real-time pyrophosphate. *Science* **281:** 365. doi: 10.1126/science.281.5375.363.

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, et al. 2000. Cloning the soil metagenome: A strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* **66:** 2541–2547.

Savage DC. 1977. Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol* **31:** 107–133.

Schloss PD, Handelsman J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71:** 1501–1506.

Schloss PD, Handelsman J. 2006. Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl Environ Microbiol* **72:** 6773–6779.

Schloss PD, Handelsman J. 2008. A statistical toolbox for metagenomics: Assessing functional diversity in microbial communities. *BMC Bioinformatics* **9:** 34. doi: 10.1186/1471-2105-9-34.

Sheneman L, Evans J, Foster JA. 2006. Clearcut: A fast implementation of relaxed neighbor joining. *Bioinformatics* **22:** 2823–2824.

Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW. 1996. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet* **14:** 450–456.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere.". *Proc Natl Acad Sci* **103:** 12115–12120.

Suau A, Bonnet R, Sutren M, Godon JJ, Gibson GR, Collins MD, Dore J. 1999. Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Environ Microbiol* **65:** 4799–4807.

Thompson JD, Higgins DG, Gibson TJ. 1994. ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673–4680.

Tringe SG, Hugenholtz P. 2008. A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11:** 442–446.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, et al. 2005. Comparative metagenomics of microbial communities. *Science* **308:** 554–557.

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.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The human microbiome project. *Nature* **449:** 804–810.

Turnbaugh PJ, Backhed F, Fulton L, Gordon JI. 2008. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe* **3:** 213–223.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, et al. 2009. A core gut microbiome in obese and lean twins. *Nature* **457:** 480–484.

Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, Lefsrud MG, Apajalahti J, Tysk C, Hettich RL, et al. 2008. Shotgun metaproteomics of the human distal gut microbiota. *ISME J* **3:** 179–189.

von Wintzingerode F, Gobel UB, Stackebrandt E. 1997. Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* **21:** 213–229.

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73:** 5261–5267.

Wilson KH, Wilson WJ, Radosevich JL, DeSantis TZ, Viswanathan VS, Kuczmarski TA, Andersen GL. 2002. High-density microarray of small-subunit ribosomal DNA probes. *Appl Environ Microbiol* **68:** 2535–2541.

Wommack KE, Bhavsar J, Ravel J. 2008. Metagenomics: Read length matters. *Appl Environ Microbiol* **74:** 1453–1463.

Zaneveld J, Turnbaugh PJ, Lozupone C, Ley RE, Hamady M, Gordon JI, Knight R. 2008. Host–bacterial coevolution and the search for new drug targets. *Curr Opin Chem Biol* **12:** 109–114.

Zhou X, Brown CJ, Abdo Z, Davis CC, Hansmann MA, Joyce P, Foster JA, Forney LJ. 2007. Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J* **1:** 121–133.