Exploring Microbial Diversity and Taxonomy Using SSU rRNA

Hypervariable Tag Sequencing

Susan M. Huse, Les Dethlefsen, Julie A. Huber, David Mark Welch, David A. Relman, Mitchell L. Sogin

Abstract

Background

Massively-parallel pyrosequencing of hypervariable regions of small subunit ribosomal RNA (SSU rRNA) genes can sample a microbial community two or three orders of magnitude more deeply per dollar and per hour than capillary sequencing of full-length SSU rRNA. As with full-length rRNA surveys, each sequence read is a tag surrogate for a single microbe. However, rather than assigning taxonomy by creating gene trees de novo that include all experimental sequences and certain reference taxa, we compare the hypervariable region tags to an extensive database of rRNA sequences and assign taxonomy based on the best match in a Global Alignment for Sequence Taxonomy (GAST) process. The resulting taxonomic census provides information on both composition and diversity of the microbial community.

Methodology / Principal Findings

To determine the effectiveness of using only hypervariable region tags for assessing microbial community membership, we compared the taxonomy assigned to the V3 and V6 hypervariable regions with the taxonomy assigned to full-length SSU rRNA
sequences from sequences isolated from both the human gut and a deep-sea hydrothermal vent.

Conclusions / Significance

The hypervariable region tags and full-length rRNA sequences provided equivalent taxonomy and measures of relative abundance of microbial communities. The greater sampling depth per dollar afforded by massively-parallel pyrosequencing reveals many more members of the “rare biosphere” than does capillary sequencing of the full-length gene. In addition, tag sequencing eliminates cloning bias and the sequences are short enough to be completely sequenced in a single read, maximizing the number of organisms sampled in a run, minimizing chimera formation. This technique allows the cost-effective exploration of changes in microbial community structure, including the rare biosphere, over space and time, and can be applied immediately to initiatives, such as the Human Microbiome Project.

Introduction

The biosphere contains between $10^{30}$ and $10^{31}$ microbial genomes, at least 2-3 orders of magnitude more than the number of plant and animal cells combined [1]. Microbes control global utilization of nitrogen through nitrogen fixation, nitrification, and nitrate reduction, and drive the bulk of sulfur, iron and manganese biogeochemical cycles [2]. They regulate the composition of the atmosphere, influence climates, recycle nutrients, and decompose pollutants. Without microbes, multi-cellular life on earth would not have evolved and biology as we know it would not be sustainable.
The diversity of microbial communities and their ecologic and metabolic functions are being explored across a great range of natural environments: in soils [3-5], air [6] and seas [7-10], on plants [11] and in animals [12,13] and in extreme environments such as the arctic [14], deep-sea vents[15], uranium-contaminated soil [16], and waste-water treatment discharge [17]. In recognition of the role marine microbes play in the biogeochemical processes that are critical to life in all environments on Earth including carbon and nitrogen cycling, the International Census of Marine Microbes (ICoMM: http://icomm.mbl.edu) has launched an international effort to catalogue the diversity of microbial populations in the oceanic, coastal, and benthic waters. Microbes associated with human health will be intensely studied through two recent large-scale initiatives: the Human Microbiome Project sponsored by the NIH (http://nihroadmap.nih.gov/hmp/) and MetaHIT sponsored by the EU (http://www.metahit.edu), which seek to characterize the composition, diversity and distribution of human-associated microbial communities. Other recent human health studies include microbes in breast milk[18], chronic wounds [19], human gut [20], dental caries[21], childcare facilities[22]

Microbes associated with the human body outnumber human cells by at least a factor of ten [23]. Some microbes cause disease, but the overwhelming majority are either innocuous or play a role in human physiology, including immune response, digestion and vitamin production. As recently as the late 1980’s, descriptions of human-associated microbiota were constrained by cultivation technologies. Over the last twenty years, sequencing surveys of amplified regions of small subunit ribosomal RNA (SSU rRNA)
genes have revealed that microbial diversity is much greater than the 5,000 microbial species described using phenotypic features in Bergey’s taxonomic outline [24], and that microbial communities are far more complex than initially thought. For instance, *E. coli*, once thought to be a dominant species in the human gut, is clearly a minor member relative to various members of the phyla *Bacteroidetes* and *Firmicutes*. It is now evident that microbiologists have been successful in culturing fewer than one percent of the different kinds of single cell organisms from most microbial communities[25]. Even for well-studied communities, such as the human distal gut, only 20-40% of the microbes have been cultured. Deeper surveys with new approaches are revealing ever-greater diversity. Even these studies, with hundreds of thousands of microbes sampled, have not been extensive enough to provide a complete picture of the diversity (richness) and relative abundance (evenness) of microbial communities.

To explore these questions, microbiologists must be able to compare microbial communities within or across individuals, in different states of health or disease, and over time. The first step in these community analyses is to develop detailed descriptions of each population, including low abundance taxa which comprise the “rare biosphere” [9]. Exploration of the human microbiome can leverage methods used to explore the microbiome of other environments such as soil, the deep sea, and other vertebrate microbiomes.

By necessity, microbiologists have historically focused their efforts on the dominant components of microbial communities. Recognizing the importance of gathering
information about high, medium and very low abundance taxa, Sogin et al. [9] introduced the use of massively-parallel DNA sequencing of short hypervariable regions of SSU rRNA to characterize microbial populations. In a subsequent study that collected nearly one million short hypervariable region tags, Huber et al. [15] demonstrated that there are over ~40,000 different kinds of bacteria and archaea in a few liters of hydrothermal vent fluid. Rarefaction data from this study and others show that in many environments, even this level of sequencing is insufficient to fully describe microbial diversity [5,26].

The lower cost and higher throughput of pyrosequencing employed in these studies allows for sampling efforts that are orders of magnitude greater than traditional capillary dideoxy sequencing of cloned SSU rRNA amplicons [27]. Although the method cannot be used to produce full-length SSU rRNA sequence (current technology limits read length to ~250nt and forward and reverse reads cannot be generated from the same amplicon), it can produce reads spanning any hypervariable region of SSU rRNA. These reads are too short to produce accurate gene trees and the hundreds of thousands of reads produced in a single experiment far exceeds the limitations of current phylogenetic software. The reads are also too short for current implementations of Bayesian classifiers such as the Ribosomal Database Project Classifier (RDP) [28]. However, each read represents a hypervariable region tag of an SSU rRNA gene present in the sample. We developed an tag search engine, Global Alignment for Sequence Taxonomy, GAST, [9] which utilizes existing databases of full-length SSU rRNA genes and their pre-computed phylogeny for high-throughput taxonomic analysis of microbial communities using hypervariable region tag sequences.
The use of a single, small hypervariable region tag for assigning taxonomy presents several challenges. The information content in a short hypervariable region sequence may not be sufficient for inferring taxonomic affinity. BLAST [29] alone is insufficient to identify the sequences in molecular databases that are the closest matches to tag queries. Here we analyze the reliability of assigning taxonomic identifiers based solely on tags, specifically using the V3 and the V6 hypervariable regions of SSU rRNA. Using SSU rRNA genes from the human gut and deep-sea vents, we compare the taxonomic assignments of the full-length sequences with the taxonomic assignments of their V3 and V6 regions, excised in-silico. We then examine microbial populations of the human gut in greater detail, using both massively-parallel pyrosequencing of hypervariable region tag and Sanger-generated full-length sequences to determine if any differences in sampling and taxonomic assignment exist with these two sequencing strategies. In our companion paper (Dethlefsen et al. 2008), we use the GAST method to explore the effects of Ciproflaxin on the human gut microbiota.

**Results**

*Assessment of hypervariable region specificity in the RefSSU database*

Tags from a hypervariable region must map to a full-length SSU rRNA with minimal ambiguity to serve as reliable phylogenetic markers; i.e. phylogenetically distinct lineages should not contain identical tags. Many of the redundant SSU rRNA sequences in the database are from the same or very similar organisms. This redundancy does not
interfere with the use of the hypervariable region tags for taxonomy, but rather strengthens the assignment. We examined the number and taxonomy of full-length sequences that correspond to each unique V3 and V6 sequence. As shown in Table 1, of the 59,830 unique V6 reference sequences, 74% mapped to one SSU rRNA sequence, 84% mapped to no more than 2, and only 5% mapped to 7 or more. The V3 region showed slightly better resolution: 82% mapped to one SSU rRNA, 90% mapped to no more than 2, and only 3% mapped to 7 or more. Although only a small percentage of tags map to more than a few SSU rRNA sequences, this included some that mapped to a very large number of different SSU rRNA sequences. For example, 3 V6 reference sequences and 8 V3 reference sequences each mapped to more than 1000 different SSU rRNA sequences.

Since RefSSU sequences that give rise to identical hypervariable region tags are generally from the same or highly similar organisms, V6 and V3 tags can be unambiguously mapped to the genus level 97% and 99% of the time, respectively. Even if we examine only the subset of V3 sequences that mapped to multiple SSU rRNA sequences, 95% map uniquely to genus, 98% to family, and 99% to order, class and phylum. Similarly for the V6 region, 91% of tags derived from multiple SSU rRNA entries map uniquely to genus, 96% to family, 97% to order and 99% to class and phylum. Not only do most of the reference tags in the database have only one SSU rRNA source, even those from multiple SSU rRNA still represent almost exclusively one taxon.
Comparison of taxonomy from long SSU rRNA sequences vs. in-silico generated hypervariable region tags in the human gut and deep-sea vent microbiomes

We compared the taxonomy of full-length SSU rRNA sequences assigned by RDP to the taxonomy of both their V3 and V6 regions (generated in-silico from full-length sequences) as assigned by GAST. We used two independent datasets of full-length sequences: 7215 sequences from the human gut and 1058 sequences from deep-sea vents. The taxonomy assigned by GAST to either the V3 or V6 hypervariable regions from the human microbiome and vent datasets matched the RDP taxonomy of the parent SSU rRNA sequences at the genus level 99+% and 98% of the time, respectively (see Table 2). Both datasets provided agreement between hypervariable region tags and full-length sequences in more than 96% of instances: the human microbiome data were classified consistently in 99% or more instances. The human microbiome tags were represented in the reference database than were those from deep-sea vents. For the human tags, 91% were exact matches to a reference tag and 99.7% of them were within 10% similarity of the nearest reference tag. Only 51% of the deep-sea vent tags had an exact match in the reference database and only 90% were within 10% similarity of the nearest match. Despite this greater divergence from the reference sequences, the GAST process still credibly mapped the taxonomy of the deep-sea vent V6 tags.

In addition, we compared the use of GAST to assign taxonomy with the use of the top BLAST match (see Table 2). While the top BLAST match was consistent with GAST to the order level, it was less accurate for family and genus for the V6 region. The RDP classifier does not assign taxonomy below genus, so we could not compare species
identification. In addition, the top BLAST match was often not the best GAST match in cases where the taxonomic assignment was the same (Table 3). A comparison of the BLAST rank vs. GAST distance did not show any significant correlation (results not shown).

Comparison of population sampling using Sanger-generated full-length and pyrosequencing-generated V3 and V6 tags for the human gut microbiome

We compared taxonomic assignments and their frequencies for the human gut microbiome data sampled with full-length SSU rRNA (n=7,215), V3 tags (n=422,992) and V6 tags (n=441,894). V6 sampling detected 60 genera and V3 sampling detected 74 genera that were not detected using the full-length sequencing. V3 sampling revealed all but one taxon represented in the full-length sequences. V6 sampling missed the family TM7, represented by only one full-length sequence and 4 V3 sequences. One hundred forty-two genera were identified in both datasets; 26 genera were identified by V6 tags that were not detected with V3, and 42 genera were identified by V3 tags that were not detected with V6. However, the taxa represented in only one of the variable region datasets were all of very low abundance, with only 3 taxa occurring more than 20 times. V3 tags identified 21 Lonpinella and 34 Megasphaera that were not identified by the V6 tags. V6 tags identified 46 Paenibacillus that were not identified by the V3 tags. Each of the classes and phyla discovered by only one hypervariable region were represented by fewer than 10 sequences. Only two orders or families containing more than 10 sequences
(11 Chromatiales Chromatiaceae, and 16 Acholeplasmatales Acholeplasmataceae) were present in only one sample set (V6).

The frequency of each taxonomic assignment at each level (phylum, class, order, family, and genus) show a linear correlation between the full-length sequencing and each of the hypervariable region tag sequencing sets with an $R^2$ correlation value of 99 (Figure 2). Taxonomic abundance levels revealed by each of the two hypervariable region tag sequencing datasets also correlated strongly against each other ($R^2=.99$) at all taxa levels. The x-intercept of the data in Figures 1a and 1b shows the level to which uncommon taxa were sampled by tag sequencing but missed by full-length sequencing. Figure 3 visually compares the relative abundances of the dominant taxa in each sample, and the numbers of rare taxa detected by each.

**Discussion**

The goal of fully describing microbial diversity in the natural world, including the human microbiome, remains elusive. The challenge posed by the unseen species is analogous to the ‘dark matter’ problem in astrophysics and is made more difficult by the unevenness of natural communities, especially in the intestinal tract of animals, and the uncertainty about the potential importance of rare species and hence the depth of coverage necessary to understand the behavior of the system. New surveying tools are needed to meet this challenge. Towards this aim, the low cost per read and high throughput of massively
parallel sequencing provide a means for deeply sampling these environments. Although the read length of pyrosequencing is increasing, it is unlikely that the technology will produce full-length rRNA sequences in the near future. While next generation sequence technologies may not be appropriate for generating full-length SSU rRNA sequences for traditional analyses, they are particularly well-suited for a tag sequencing strategy, where each read represents a short amplicon and can be used as a tag to match the sequence to known full-length sequences.

Microbial ecology requires not only identification of organisms, but also consistent and reproducible sampling of populations. We sampled the human gut microbiota with conventional long SSU rRNA sequences and with V3 and V6 tag sequencing, comparing both the types of organisms and their relative abundances (see Figure 1). At all taxonomic ranks and for both hypervariable regions, the composition of the microbiota as revealed using long sequences or tag sequences correlated for 99% of the taxa. These results applied to all taxonomic ranks (genus to phylum) and to both hypervariable regions (V3 and V6). The greater sampling depth provided by the massively-parallel pyrosequencing of hypervariable tags provides a greater window onto the breadth of taxonomy in the human gut than the use of longer full-length sequences. This is illustrated by the large x-intercept of the data in Figures 1a and 1b. Each set of tag sequences showed measurable quantities of taxa (x > 1) that were lost in the full-length sequencing (y=0). The x-intercept comparing the two tag sequencing experiments (Figure 1c) is at the origin, implying that the two hypervariable regions were comparable in elucidating the rare biosphere.
For tag sequencing of hypervariable regions to be effective for mapping taxonomy, specific sequences must match unambiguously to source organisms. If it is common for two divergent organisms to have the same or highly similar V3 or V6 regions, tag sequencing will not be an accurate means for assigning taxonomy. In our reference database of over 500,000 rRNA sequences, we found that hypervariable region tags map to individual taxa with high fidelity. In only a few cases for either the V3 or the V6 region did we find sequences that exactly match two or more distinct taxa.

Approximately three-quarters of our reference V6 tags map uniquely to a long sequence in RefSSU and have an unambiguous taxonomic assignment based on that long. In most cases where a tag maps to more than one SSU rRNA source in our reference database, the multiple sources have the same taxonomy. The reference databases are replete with highly-studied bacteria, and multiple copies of a hypervariable region for these organisms does not imply any taxonomic ambiguity. The V3 region was 99% accurate to the genus level. The V6 region was only slightly less resolved than the V3, still providing a 97% accuracy in assignment to the genus level, 98+% accuracy to the level of family, and 99% accuracy at the level of order. These levels of accuracy show that in theory, and based on current knowledge of the rRNA genes, hypervariable region tags contain adequate information to accurately map taxonomy of both bacterial and archaeal organisms.

To assign taxonomy to our long rRNA gene sequences, both in our SSURef database and in our experimental sequences, we used the Ribosomal Database Project classifier (RDP). RDP is not 100% accurate and some of the ambiguities in the reference database could be
attributable to limitations with the classifier. For tag sequencing with the GAST process, however, we are assessing the utility of a tag as a surrogate for the longer SSU rRNA sequences via a look-up and distance matrix. Can we consistently assign the correct taxonomy to both a SSU rRNA sequence and its constituent hypervariable regions independently? The RDP taxonomy provides a consistent taxonomic classification (Bergey’s taxonomy), facilitating our analysis. Slight inaccuracies in RDP will not be an important factor in whether a tag sequence can be used as a surrogate for a full-length rRNA sequence.

We conducted an in-silico experiment assessing taxonomic assignment using tags of hypervariable regions extracted from full-length sequences and compared the tags directly to the full-length sequences. We used two independent datasets, one of human gut microbiota, which should be relatively well represented in the reference databases of rRNA genes, and one of deep-sea vent microbes, which are less well studied and therefore less well represented in the reference databases. We examined the use of both the V3 and the V6 region tags to assign taxonomy for both groups of microbes. In all four cases, we found excellent correspondence between the use of the GAST process for assigning taxonomy to short hypervariable region tags and the use of RDP for assigning taxonomy to the full-length sequences. For both variable regions from the human microbiota, the taxonomic assignments of the tags agreed with the long sequences at a rate consistently greater than 99%. The deep-sea vent data agreed in 97% of instances at the genus level.
The two variable regions mapped taxonomy with virtually identical fidelity. The greater
difference was not in choice of variable region, but in the microbes examined. Of the
human gut microbes sampled, 91% of all tags had exact matches in the reference
database and virtually all had matches within a 10% sequence match. Of the deep-sea
vent microbes, on the other hand, only 51% had exact matches in the reference database
and only 90% were within a 10% sequence match of a reference sequence. Despite the
fact that as many as 10% of the deep-sea vent tags did not have a close match in the
reference database, the GAST process only mis-assigned 3% of the tags. These results
imply that hypervariable region tag sequencing and the GAST process are excellent tools
for assigning taxonomy, but they cannot overcome basic gaps in knowledge of the under-
explored areas of the microbiome. As more is learned about these organisms, and their
rRNA genes are added to the reference databases, hypervariable region tags sequencing
projects will directly benefit, and taxonomy will improve.

The top BLAST matches for the human gut microbiome mapped taxonomy better than
the top BLAST matches for the deep-sea vents. This is likely because the human
microbiome data are better represented in the reference database. Since 91% of the
human microbiome tags had exact matches in the reference database these should be
consistently identified by BLAST as the best match. For the deep-sea samples where
only 51% had exact matches, a top BLAST hit may find only a local match within the
sequence rather than a globally-weighted match as with GAST. This seems particularly
true for the V6 region, where 11% of the time the top BLAST hit did not correctly
identify the genus. The GAST process did not show a similar drop in taxonomic
assignment. The V6 region may have some local variations within the sequences that can be misleading but which are averaged out along the length of the sequence. An increase in distance from the tag to the nearest reference sequence using GAST did not correlate with a lower BLAST rank. The magnitude of divergence from the reference database does not explain the difference between V3 and V6 regions.

The V3 tags were noticeably more divergent from their top BLAST hit than are the V6 tags, despite the larger dataset of reference V3 sequences. This did not adversely affect the ability to identify tags to the genus level. Since the RDP taxonomy is restricted to the genus level, we could not review the BLAST ranks for species-level. The top BLAST hit may be losing the finer sequence or species detail that could be picked up with GAST. For example, some health benefits are associated with only certain strains of probiotic species in the human gut [30]. Hypervariable region tag sequencing in conjunction with a GAST analysis could potentially differentiate these strains.

Both the V3 and the V6 hypervariable regions uncovered multiple genera that were missed by the full-length sequencing (148 and 59 respectively). The only organism discovered with the full-length rRNA sequences that was not discovered with the V6 tag sequencing was rare, with only one sequence uncovered by the full-length and 4 uncovered by the V3 sequencing. The absence of TM7 from the V6 tag set could easily be explained by the stochastic nature of rare biosphere sampling rather than a V6-specific bias. The hypervariable region tag sequencing did not introduce any strong biases against the discovery of common taxa or the relative abundance of these taxa in this experiment.
As predicted, the hypervariable region tag sequencing did provide a much greater breadth and depth of sampling.

Although the level of sampling with tag sequencing is orders of magnitude greater than with traditional methods, a single pyrosequencing run (with > 400,000 sequences) is still insufficient to fully sample the rare biota in the human distal gut. The sampling limitation of this experiment can be seen in the small but distinct number of taxa that appeared in one but not the other tag sequencing experiment. All were of low abundance and are dispersed throughout the microbial world rather than clustering in one specific taxon. No common taxa were omitted by sequencing in either hypervariable region. Some variation in taxa and their abundance can be caused by the selection of primers, although this did not appear to greatly affect the experiment.

Conclusions/Significance:
Hypervariable region tag sequencing using either the V3 or the V6 region, and presumably other hypervariable regions, is an effective means for assigning taxonomy and provides great advantages over traditional sampling. An tag mapping process such as GAST with an extensive database of rRNA genes such as our RefSSU derived from SILVA can map tag sequences to the same taxonomy as their source genes at better than a 99% correlation rate for commonly studied environments such as the human microbiome and better than 96% for less commonly studied environments such as deep-sea vents. The V3 and V6 regions have only minimal ambiguity in mapping to the SSU rRNA gene all the way to the genus level. While tags can map to more than one SSU
rRNA source, these cross-mappings between taxa are infrequent and do not compromise the overall methodology. We show that these short hypervariable region tags contain adequate information to uniquely and accurately map the phylogeny to a very high degree even without an exact match in the reference database or with potential multiple copies in the database. The consistently high correspondence of the hypervariable region tags vs. long SSU rRNA taxonomies shows the robustness of the GAST process and the use of tags as surrogates for the full-length rRNA genes even for microbial environments that are not well-represented in the reference databases.

Massively-parallel pyrosequencing of tags can be used to great advantage over traditional sequencing of full-length rRNA genes to explore both the diversity and relative abundance of microbial populations. Further research into hypervariable region tag sequencing may uncover advantages of one region over another, such as the relative levels of microvariation, length of sequence, density of homopolymers which can lead to pyrosequencing errors, ability to identify to the species level, or the merits of different sequencing primers. Tag sequencing provides more reads, representing more organisms, avoids assembly, and costs less per read than conventional sequencing of full-length SSU rRNA genes. Both methods yield equivalent taxonomy and relative abundance. The great advantage of hypervariable region tag sequencing is that it can take advantage of massively-parallel pyrosequencing, sampling to depths several orders of magnitude greater than previously achieved, facilitating the exploration of the vast diversity of microbial populations and the rare biosphere.
Material and Methods:

Creating the reference database of full-length, V3, and V6 SSU rRNA sequences

We downloaded 503,971 aligned small subunit rRNA sequences from the SILVA database, version 92 [31]. Using the SILVA quality assessments, we eliminated low-quality sequences (sequence quality $\leq 50$, alignment quality $\leq 50$, pintail score $\leq 40$). Identical SSU rRNA sequences were flagged as redundant. The resultant dataset included 417,433 unique sequences. Although the sequences vary in length and coverage of the full-length SSU rRNA gene, we refer to these sequences as “long” or “full-length” sequences for the purposes of this paper, and the dataset of these sequences as RefSSU.

From all aligned RefSSU sequences, we extracted the V3 and V6 hypervariable regions, defined as homologous positions between positions 338 and 533 of the *E. coli* SSU rRNA sequence (U00096) for V3, and 967 to 1046 for V6. Sequences shorter than 50 nt or containing ambiguous bases were culled. We removed all gap characters to create a set of 293,265 V3 reference tags (RefV3 database) and 195,344 V6 reference tags (RefV6 database). These databases include 123,206 unique V3 tag sequences and 59,830 unique V6 tag sequences.

We classified all bacterial and archaeal long sequences directly with the Ribosomal Database Project Classifier (RDP) (Cole et al., 2007). We used only RDP classifications with a bootstrap value of $\geq 80\%$. If the bootstrap value was $< 80\%$, the taxonomic assignment was moved to a higher classification level until an 80% or better bootstrap value was achieved. For example, if the genus assignment had a bootstrap value of 70%,
but the family had a value of 85%, that sequence would be assigned only as far as family and not to genus. RDP Classifier does not classify sequences below the genus level.

**Sampling of Human Gut Microbiome**

Detailed methods are described in the companion paper (Dethlefsen et al., 2008).

Briefly, we extracted DNA from fecal samples of three individuals before, during, and after a 5-day course of the antibiotic ciprofloxacin, then performed PCR with primers designed to amplify the SSU rRNA gene (hereafter referred to as “full-length”), as well as the V3 and V6 hypervariable regions of the full-length gene. The forward primers for the full length product were 90% bacterial primer 8F (5’-AGAGTTTGATCMTGGCTCAG-3’) and 10% 8F-Bif targeting *Bifidobacteria* (5’-AGGGTTCGATTCTGGCTCAG-3’), and were paired with the 3 domain reverse primer 1391R (5’-GACGGGCGGTGTGTRCA-3’). Re-conditioning PCR reactions were performed for the full-length gene amplifications. Full-length amplicons were cloned and sequenced with Sanger dideoxy methods, assembled with PHRAP [32], aligned with NAST [33] and evaluated for chimeras with Bellerophon [34] (GenBank Accession numbers: pending). The V3 and V6 amplicon libraries were sequenced with a GS FLX using pyrosequencing.

The primers spanning V3 were 338F (5’-ACT CCT ACG GGA GGC AGC AG-3’) and 533R (5’-TTA CCG CGG CTG CTG GCA C-3’). The primers spanning V6 were a combination of 967F primers (5’-CAA CGC GAA GAA CCT TAC C-3’ and 5’-ATA
CGC GA[AG] GAA CCT TAC C-3’) and a combination of 1046R primers (5’-AGG TGN TGC ATG GCT GTC G-3’ and 5’-AGG TGN TGC ATG GTT GTC G-3’).

*Generation of deep-sea vent microbe long sequences*

Detailed methods are described elsewhere [15]. Briefly, DNA was extracted from deep-sea vent samples and PCR was conducted using primers designed to amplify a 1000 bp region of SSU rRNA. The primers used for amplification of the 1000 bp amplicons were sequenced bidirectionally using primers T3 (5’- ATT AAC CCT CAC TAA AGG GA) and T7 (5’- TAA TAC GAC TCA CTA TAG GG). Sequencing was performed on an Applied Biosystems 3730XL capillary sequencer (GenBank Accession numbers DQ919170-DQ910173).

*Generation of in-silico hypervariable region tag subsequences from full-length sequences*

We aligned each dataset of full-length sequences with MUSCLE [35] (default parameters). We located the position of the V6 primers (967F and 1046R), and the V3 primers (338F and 533R) in each alignment and extracted, in silico, the hypervariable regions from each of the aligned full-length sequences.

*Assigning taxonomic classification to hypervariable region tags through GAST*

In Sogin et al. [9], we proposed a tag mapping methodology, GAST (Global Alignment for Sequence Taxonomy) to assign a taxonomic classification to environmental V6 tags. We used BLAST [36] to generate a set of best local matches to the reference database (RefV3 or RefV6) for each tag. Because the top BLAST hit may not have the highest
overall similarity to the tag sequence, particularly because edge-effects in the short region being compared can be pronounced, we aligned the tag sequence to the reference hypervariable region tags corresponding to the top 100 BLAST hits. We used MUSCLE (with parameters –diags and -maxiters 2 to reduce processing time) [35] because it is well suited to high-throughput experiments. We calculated the global distance from the sample tag to each of the aligned reference sequence tags as the number of insertions, deletions and mismatches divided by the length of the tag. We considered the reference sequence or sequences with the minimum global distance to be the top GAST match(es). The top BLAST hit was generally the best global match; however, for 5% to 25% of tags the best global match was to a reference sequence with a lower BLAST score.

For each tag, we identified all of the reference long sequences in RefSSU that contained the exact hypervariable sequence of the top GAST match(es). We compared the taxonomic classification of all corresponding SSU rRNA sequences (with RDP bootstrap values >= 80) and generated a consensus taxonomy. If two-thirds or more of the full-length sequences shared the same assigned genus, the tag was assigned to that genus. If there was no such agreement, we proceeded up one level to family. If there was a two-thirds or better consensus at the family level, we assigned this taxonomy to the tag, and if not, we continued to proceed up the tree. Occasionally, a tag could not be assigned taxonomic classification at the domain level. This was because the RDP Classifier could not assign a domain with an adequate bootstrap value, rather than a tag mapping to full-length sequences from different domains. Sample tags that did not have a BLAST match in the RefSSU database were not given a taxonomic assignment. We chose to use a 66%
majority although other values or a distributional vs. strict percentage approach can be implemented. We reviewed over 10 million tags in our sequencing database (primarily of the V6 region) from a wide range of studies using the 66% majority as the threshold for assignment. The full-length source sequences had unanimous agreement on taxonomy 67% of the time. Less than 10% of the tags had a voting majority of 75% and essentially 0% resolved with a majority as low as 66%.

**Acknowledgments**

We thank Richard Fox for computing facility support and Hilary G. Morrison for contributions to the creation and maintenance of the reference database RefSSU.

**Funding**

Woods Hole Center for Oceans and Human Health from the National Institutes of Health and National Science Foundation (NIH/NIEHS 1 P50 ES012742-01 and NSF/OCE 0430724-J Stegeman PI to HGM and MLS). Doris Duke Charitable Trust (Distinguished Clinical Scientist Award to DAR) and National Institutes of Health (Director’s Pioneer Award DP1OD000964 to DAR).

**References**

1. Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A 95: 6578-6583.

2. Atlas RM, Bartha R (1993) Microbial Ecology: Fundamentals and Applications. Redwood City: The Benjamin/Cummings Publishing Company, Inc.
3. Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, et al. (2007) Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil. Appl Environ Microbiol 73: 7059-7066.

4. Schloss PD, Handelsman J (2006) Toward a Census of Bacteria in Soil. PLoS Computational Biology 2: e92.

5. Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, et al. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1: 283-290.

6. Brodie EL, DeSantis TZ, Parker JPM, Zubietta IX, Piceno YM, et al. (2007) Urban aerosols harbor diverse and dynamic bacterial populations. PNAS 104: 299-304.

7. Alonso-Saez L, Gasol JM (2007) Seasonal Variations in the Contributions of Different Bacterial Groups to the Uptake of Low-Molecular-Weight Compounds in Northwestern Mediterranean Coastal Waters. Appl Environ Microbiol 73: 3528-3535.

8. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, et al. (2008) Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A 105: 3805-3810.

9. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences 103: 12115-12120.

10. Stevens H, Ulloa O (2008) Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. Environmental Microbiology 10: 1244-1259.
11. Fang M, Kremer RJ, Motavalli PP, Davis G (2005) Bacterial Diversity in Rhizospheres of Nontransgenic and Transgenic Corn. Appl Environ Microbiol 71: 4132-4136.

12. Michalke K, Schmidt A, Huber B, Meyer J, Sulkowski M, et al. (2008) Role of Intestinal Microbiota in Transformation of Bismuth and Other Metals and Metalloids into Volatile Methyl and Hydride Derivatives in Humans and Mice. Appl Environ Microbiol 74: 3069-3075.

13. Yu Z, Garcia-Gonzalez R, Schanbacher FL, Morrison M (2008) Evaluations of Different Hypervariable Regions of Archaeal 16S rRNA Genes in Profiling of Methanogens by Archaea-Specific PCR and Denaturing Gradient Gel Electrophoresis. Appl Environ Microbiol 74: 889-893.

14. Stoeck T, Kasper J, Bunge J, Leslin C, Ilyin V, et al. (2007) Protistan Diversity in the Arctic: A Case of Paleoclimate Shaping Modern Biodiversity? PLoS ONE 2: 16.

15. Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, et al. (2007) Microbial Population Structures in the Deep Marine Biosphere. Science 318: 97-100.

16. Barns SM, Cain EC, Sommerville L, Kuske CR (2007) Acidobacteria Phylum Sequences in Uranium-Contaminated Subsurface Sediments Greatly Expand the Known Diversity within the Phylum. Appl Environ Microbiol 73: 3113-3116.

17. Wakelin SA, Colloff MJ, Kookana RS (2008) Effect of Wastewater Treatment Plant Effluent on Microbial Function and Community Structure in the Sediment of a Freshwater Stream with Variable Seasonal Flow. Appl Environ Microbiol 74: 2659-2668.
18. Delgado S, Arroyo R, Martin R, Rodriguez J (2008) PCR-DGGE assessment of the bacterial diversity of breast milk in women with lactational infectious mastitis. BMC Infectious Diseases 8: 51.

19. Dowd S, Sun Y, Secor P, Rhoads D, Wolcott B, et al. (2008) Survey of bacterial diversity in chronic wounds using Pyrosequencing, DGGE, and full ribosome shotgun sequencing. BMC Microbiology 8: 43.

20. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635-1638.

21. Aas JA, Griffen AL, Dardis SR, Lee AM, Olsen I, et al. (2008) Bacteria of Dental Caries in Primary and Permanent Teeth in Children and Young Adults. J Clin Microbiol 46: 1407-1417.

22. Lee L, Tin S, Kelley S (2007) Culture-independent analysis of bacterial diversity in a child-care facility. BMC Microbiology 7: 27.

23. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host-Bacterial Mutualism in the Human Intestine. Science 307: 1915-1920.

24. Guerrero R (2001) Bergey's manuals and the classification of prokaryotes. Int Microbiol 4: 103-109.

25. Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276: 734-740.

26. Ashby MN, Rine J, Mongodin EF, Nelson KE, Dimster-Denk D (2007) Serial Analysis of rRNA Genes and the Unexpected Dominance of Rare Members of Microbial Communities. Appl Environ Microbiol 73: 4532-4542.
27. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

28. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, et al. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Research 33: D294-296.

29. Alterovitz G, Jiwaji A, Ramoni MF (2008) Automated programming for bioinformatics algorithm deployment. Bioinformatics 24: 450-451.

30. Dethlefsen L, McFall-Ngai M, Relman DA (2007) An ecological and evolutionary perspective on human, Bbacteriome mutualism and disease. Nature 449: 811-818.

31. Ludwig W, Strunk O, Westram R, Richter L, Meier H, et al. (2004) ARB: a software environment for sequence data. Nucleic Acids Research 32: 1363-1371.

32. Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Research 8: 186-194.

33. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72: 5069-5072.

34. Huber T, Faulkner G, Hugenholtz P (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics: bth226.

35. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792-1797.

36. Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403-410.
Figure Legends:

FIGURE 1  Correspondence of genera found by each method

The Venn Diagram shows the extent of overlap between the V3, V6 and full-length sequencing of the human gut microbiome. The V3 tag sequencing found the most genera (116), the V6 found 103 genera, and full-length sequencing found only 43 genera.

FIGURE 2 – Correlation of taxonomic assignments based on full-length SSU rRNA, V3 and V6 variable regions.

At each taxonomic level, the number of sequences from a particular taxon using one sequencing strategy(e.g., V6 tags) is plotted against the number of sequences from that same taxon using a second sequencing strategy (e.g., full-length 16S genes). For instance, classifying to the genus level, the Clostridial genus *Ruminococcus* occurred 186 times in the full-length sequences and 19,332 times in the V6 tags. The order Clostridiales occurred 3,613 times in the full-length sequences, and 217,482 times in the V6 tags. *Note:* correlations are linear, although the axes use log scales for clarity.

Figure 1a compares V6 tags with full-length sequences, Figure 1b compares V3 tags with full-length sequences, and Figure 1c compares V6 and V3 tags.

FIGURE 3  Comparison of taxonomic assignments at the genus level

Assignment to genus and their relative abundances are distinctly similar for each of the three methods, V3 tag, V6 tag and full-length sequencing of SSU rRNA genes. Only sequences classified to the genus level are included. The tag sequencing approach, however, reveals many more rare taxa than does conventional full-length sequencing.
Table 1a: Hypervariable region, V3.

| # of Taxa | Genus  | Family | Order   | Class  | Phylum   |
|-----------|--------|--------|---------|--------|----------|
| 1         | 99.11% / 69686 | 99.62% / 88015 | 99.88% / 99682 | 99.93% / 109352 | 99.96% / 114328 |
| 2         | 0.07% / 495 | 0.34% / 297 | 0.12% / 113 | 0.07% / 77 | 0.04% / 42 |
| 3         | 0.09% / 64 | 0.04% / 34 | 0.00% / 4 | 0.00% / 2 | 0.00% / 42 |
| 4         | 0.05% / 35 | 0.00% / 3 | 0 | 0 | 0 |
| 5+        | 0.00% / 29 | 0.00% / 3 | 0 | 0 | 0 |

Table 1b: Hypervariable region V6

| # of Taxa | Genus  | Family | Order   | Class  | Phylum   |
|-----------|--------|--------|---------|--------|----------|
| 1         | 97.33% / 31463 | 98.77% / 40454 | 99.30% / 45579 | 99.68% / 51795 | 99.83% / 54728 |
| 2         | 2.07% / 670 | 1.08% / 444 | 0.62% / 285 | 0.30% / 158 | 0.17% / 94 |
| 3         | 0.35% / 112 | 0.11% / 45 | 0.05% / 23 | 0.01% / 6 | <0.01% / 1 |
| 4         | 0.15% / 50 | 0.03% / 11 | 0.01% / 5 | 0 | 0 |
| 5+        | 0.01% / 31 | 0.01% / 5 | 0.01% / 6 | 0 | 0 |

Table 1: Percent of hypervariable region tags from the Ref16S database that map to one or more taxa.

For HVTs to accurately identify microbial taxa, the same tag sequence must not be present in different taxa. Table 1 shows the percentage of hypervariable tags (V3 or V6) present in the Ref16S reference genes that map to single or multiple taxa. The results are displayed both as a percentage of tags and as a total number of tags.
### Table 2a: Human microbiome sequences

| Human gut microbiome | V6 region | V3 region |
|----------------------|-----------|-----------|
|                      | Count | Same | %Same | Count | Same | %Same |
| Superkingdom         | 7215  | 7215  | 100.00% / 100.00% | 7208  | 7208  | 100.00% / 100.00% |
| Phylum               | 7175  | 7152  | 99.68% / 99.82% | 7168  | 7145  | 99.68% / 99.82% |
| Class                | 7040  | 7009  | 99.56% / 99.52% | 7033  | 7002  | 99.56% / 99.52% |
| Order                | 7026  | 6994  | 99.54% / 99.56% | 7019  | 6987  | 99.54% / 99.57% |
| Family               | 5731  | 5693  | 99.34% / 98.93% | 5726  | 5688  | 99.34% / 98.93% |
| Genus                | 5183  | 5158  | 99.52% / 97.97% | 5178  | 5153  | 99.52% / 97.99% |

### Table 2b: Deep-sea vent sequences

| Deep-sea Vents | V6 region | V3 region |
|----------------|-----------|-----------|
|                | Count | Same | %Same | Count | Same | %Same |
| Superkingdom   | 1058  | 1058  | 100.00% / 99.53% | 963   | 963   | 100.00% / 100.00% |
| Phylum         | 1008  | 986   | 97.82% / 90.00% | 920   | 901   | 97.93% / 97.68% |
| Class          | 970   | 939   | 96.80% / 90.03% | 884   | 857   | 96.95% / 99.55% |
| Order          | 881   | 847   | 96.14% / 87.06% | 807   | 784   | 97.15% / 100.00% |
| Family         | 833   | 814   | 97.72% / 87.23% | 764   | 746   | 97.64% / 98.44% |
| Genus          | 766   | 749   | 97.78% / 88.52% | 701   | 686   | 97.86% / 99.10% |

### Table 2: Comparison of taxonomic assignments using full-length and V3 and V6 hypervariable region tags

Treating the V3 and the V6 regions independently, we counted the number of assignments the GAST process made at each taxonomic rank and the number and percent of times those assignments were the same as the assignment given to the full-length source sequence. The second percent value is the rate at which the top BLAST match predicted the same assignment as the full-length source.
| BLAST rank | Human Microbiome | Deep-Sea Vents |
|------------|------------------|---------------|
|            | V6 region        | V6 region     |
|            | V3 region        | V3 region     |
| 1          | 94.90%           | 81.89%        |
|            | 83.70%           | 75.17%        |
| 2          | 4.34%            | 7.16%         |
|            | 13.48%           | 5.27%         |
| > 2        | 0.76%            | 10.95%        |
|            | 2.82%            | 19.57%        |

**Table 3: BLAST ranks for top GAST hits.** The reported percentages represent the frequency with which the top GAST match corresponds to the top BLAST match, the second best BLAST match, or any other BLAST match.
FIGURE 1

- V3 only (39)
- V3 & V6 (35)
- Full-Length & V3 & V6 (39)
- V6 only (28)

Legend:
- FL & V3 (3)
- FL & V6 (1)
**Figure 2a** Full-length vs V6 Tag Sequences

Correlation (R^2)
- Genus: 0.988
- Family: 0.991
- Order: 0.998
- Class: 0.998
- Phylum: 0.999

**Figure 2b** Full-length vs V3 Tag Sequences

Correlation (R^2)
- Genus: 0.991
- Family: 0.992
- Order: 0.992
- Class: 0.992
- Phylum: 0.993

**Figure 2c** V3 vs V6 Tag Sequences

Correlation (R^2)
- Genus: 0.995
- Family: 0.997
- Order: 0.992
- Class: 0.991
- Phylum: 0.993

**FIGURE 2**
V3 Tag Sequencing
N = 299,044
Other genera = 99

V6 Tag Sequencing
N = 322,971
Other genera = 82

Full-Length Sequencing
N = 5,519
Other genera = 26

FIGURE 3