The status of the microbial census: an update

Running title: The microbial census

Patrick D. Schloss[1†], Rene Girard[2], Thomas Martin[2], Joshua Edwards[2], and J. Cameron Thrash[2†]

† To whom correspondence should be addressed: pschloss@umich.edu and thrashc@lsu.edu

1. Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109

2. Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803

Abstract

A census is a useful tool for assessing the characteristics of an entity, determining the best distribution of resources across individuals, and highlighting the strengths and weaknesses of the entity. Although a census is typically carried out for people at a national level, microbial ecologists have carried out a molecular census of bacteria and archaea by sequencing their 16S rRNA genes. The goal of this study was to update an earlier effort from 2004 by assessing how well the microbial census of full-length 16S rRNA gene sequences is proceeding in the context of recent advances in high throughput sequencing technologies. We found that nearly all full-length sequences belong to operational taxonomic units (OTUs) that have been previously observed, but that most bacterial and archaeal OTUs have only been observed once. This suggests that there is still considerable microbial diversity to be explored. We discuss the role that traditional culture-dependent and -independent approaches have had in advancing the census as well as the possible role that new technologies such as single cell genomics and short read assembly might have in allowing researchers to improve our ability to sample rare OTUs. We also show that environments that are known to be rich in bacterial diversity (e.g. soil and aquatic environments) have been poorly sampled. The ability to contextualize the biodiversity that is sampled using short read sequencing technologies depends on the ability to obtain similar sequence depths of full-length sequences from diverse environments. Furthermore, these sequences and their environmental metadata must be made available in public databases.

Importance

The biodiversity contained within the bacterial and archaeal domains dwarfs that of the eukaryotes and the services these organisms provide to the biosphere are critical. Surprisingly, we have done a relatively poor job of keeping track of the ongoing effort to characterize the biodiversity as represented in full-length 16S rRNA genes. By understanding how this census is proceeding, it is possible to suggest the best allocation of resources for advancing the census. We actually found that the ongoing effort has done an excellent job of sampling the most abundant organisms, but

struggles to sample the more rare organisms. Through the use of new sequencing technologies we should be able to obtain full-length sequences from these rare organisms. Furthermore, we suggest that by allocating more resources to sampling environments known to have the greatest biodiversity we will be able to make significant advances in our characterization of microbial diversity.

Introduction

The effort to quantify the number of different organisms in a system remains fundamental to understanding ecology (**???**). At the scale of microorganisms, small physical sizes, morphological ambiguity, and highly variable population sizes complicate this process. Furthermore, creating standards for delimiting what makes one microbe "different" from another has been contentious (**???**). In spite of these challenges, we continue to peel back the curtain on the microbial world with the aid of more and more informative, if still limited, technologies like cultivation, 16S rRNA gene surveys, single cell technologies, and metagenomics.

Generating a comprehensive understanding of any system with a single gene may seem a fool's errand, yet we have learned a considerable amount regarding the diversity, dynamics, and natural history of microorganisms using the venerable 16S rRNA gene. In 1983, the full-length 16S rRNA gene sequence of *Escherichia coli* (accession J01695) was deposited into NCBI's GenBank making it the first of what is now more than 10 million 16S rRNA gene sequences to be deposited into the database (1). 16S rRNA gene accessions represent nearly one-third of all sequences deposited in GenBank, making it the best-represented gene. As Sanger sequencing has given way to so-called "next generation sequencing" technologies, hundreds of millions of 16S rRNA gene sequences have been deposited into the NCBI's Sequence Read Archive. The expansion in sequencing throughput and increased access to sequencing technology has allowed for more environments to be sequenced at a deeper coverage, resulting in the identification of novel taxa. The ability to obtain sequence data from microorganisms without cultivation has radically altered our perspective of their role in nearly every environment from deep ocean sediment cores (e.g. accession AY436526) to the International Space Station (e.g. accession DQ497748).

Previously, Schloss and Handelsman (2) assigned the then available 56,215 partial 16S rRNA gene sequences to operational taxonomic units (OTUs) that were available in the Ribosomal Database Project and concluded that the sampling methods of the time were insufficient to identify the previously estimated 10^7 to 10^9 different species (3). That census called for a broader and deeper characterization of all environments. Refreshingly, this challenge was largely met. There have been major investments in studying the Earth's microbiome using 16S rRNA gene sequencing through

initiatives such as the Human Microbiome Project (**???**), the Earth Microbiome Project (**???**), and the International Census of Marine Microorganisms. But most importantly, the original census was performed on the cusp of radical developments in sequencing technologies. That advancement has moved the generation of sequencing throughput from large sequencing centers to individual investigators and leveraged their diverse interests to expand the representation of organisms and environments represented in public databases.

Disconcertingly, the increase in sequencing volume has come at the cost of sequence length. The commonly used MiSeq-based sequencing platform from Illumina is extensively leveraged to sequence the approximately 250 bp V4 hypervariable region of the 16S rRNA gene; other schemes have used different parts of the gene that are generally shorter than 500 bp. The number of OTUs that are sampled when using different regions within the 16S rRNA gene can vary considerably and the genetic diversity within these regions typically has only a modest correlation the genetic diversity of the full-length sequence [**???** et al. 2009; Schloss 200X]. Thus, it remains unclear to what degree richness estimates from short read technologies over or underestimate the numbers from full-length sequences. We likely lack the references necessary to adequately classify the novel biodiversity we are sampling when we generate 100-times the sequence data from a community than we did using full-length sequencing.

Here we update the status of the microbial census with full-length 16S rRNA gene sequences. In the 13 years since the collection of data for Schloss and Handelsman's initial census, the number of full-length sequences has grown exponentially, despite the overwhelming contemporary focus by most researchers on short-read technologies (Figure 1A). This update to the census allows us to evaluate the relative sampling thoroughness for different environments and clades and make an argument for the continued need to collection full-length sequence data from many systems that have a long history of study. As researchers consider coalescing into a Unified Microbiome Initiative (5), it will be important to balance the need for mechanism-based studies with the need to generate full-length reference sequences from a diversity of environments.

Results and Discussion

The status of the bacterial and archaeal census. To assess the field's progress in characterizing the biodiversity of bacteria and archaea, we assigned each 16S rRNA gene sequence to OTUs using distance thresholds that varied between 0 and 20%. Although it is not possible to link a specific taxonomic level (e.g. species, genus, family, etc.) to a specific distance threshold, we selected distances of 0, 3, 5, 10, and 20% because they are widely regarded as representing the range of genetic diversity of the 16S rRNA gene within each domain. By rarefaction, it was clear that the ongoing sampling efforts have started to saturate the number of current OTUs. After sampling 1,411,234 near full-length bacterial 16S rRNA gene sequences we have identified 217,645, 108,950, 66,819, 15,743, and 3,731 OTUs at the respective thresholds (Figure 1A, Table 1). Using only the OTUs generated using a 3% threshold, we calculated a 94.5% Good's coverage (percent of sequences belonging to OTUs that have been observed more than once), but only 29.2% OTU coverage (percent of the OTUs that have been observed more than once). Paralleling the bacterial results, after sampling 53,546 archaeal 16S rRNA gene sequences we have identified 11,040, 4,252, 2,364, 812, and 110 OTUs (Figure 1B, Table 1). Using only the OTUs generated with a 3% threshold, we calculated a 95.1% Good's coverage, but only 38.5% OTU coverage. These results indicate that regardless of the domain, continued sampling with the current strategies for generating full-length sequences will largely reveal OTUs that have already been observed, even though a large fraction of OTUs have only been sampled once. Furthermore, considering more than 70.8% of the OTUs have only been observed once, it is likely that an even larger number of OTUs have yet to be sampled for both domains.

The status of the bacterial census. One explanation for the large number of OTUs that have only been observed once is that with the the broad adoption of sequencing platforms that generate short sequence reads, the rate of full-length sequence generation has declined. In fact, since 2009 the number of new bacterial sequences generated has slowed to an average of 189,960 sequences per year (Figure 2A). Although this is still an impressive number of sequences, since 2007 the number of new bacterial OTUs has plateaued at an average of 11,184 new OTUs per year (Figure 2B). Given the expense of generating full-length sequences using the Sanger sequencing technology

and the transition to other platforms at that time, we expected that the large number of sequences were being deposited by a handful of large projects. Indeed, when we counted the number of submissions responsible for depositing 50% of the sequences, we found that with the exception of 2006 and 2013, eight or fewer studies were responsible for depositing the majority of the full-length sequences each year since 2005 (Figure 2C). Between 2009 and 2012, 908,190 total sequences were submitted and 6 submissions from 5 studies were responsible for depositing 550,274 (60.6% of all sequences). These studies generated sequences from the human gastrointestinal tract (6), human skin (7, 8), murine skin (9), and hypersaline microbial mats (10). The heavy zoonotic focus is reflected in the rarefaction curve for this category (Figure 1C). In contrast to recent years, between 1995 and 2006, an average of 39.3 studies were responsible for submitting more than half of the sequences each year. Although these deep surveys represent significant contributions to our knowledge of bacterial biogeography, their small number and lack of environmental diversity is indicative of the broader problems in advancing the bacterial census.

The status of the archaeal census. The depth of sequencing being done to advance the archaeal census has been 26-times less than that of the bacterial census (Table 1). The annual number of sequences submitted has largely paralleled that of the bacterial census with a plateau starting in 2009 and an average of 7,075 sequences each year since then. The number of new archaeal OTUs represented by these sequences began to slow in 2005 with an average of 355.5 new OTUs per year. With the exception of 2012 and 2014, the number of submissions responsible for more than 50% of the archaeal sequences submitted per year has varied between 2 and 11 submissions per year. The clear bias towards sequencing bacterial 16S rRNA genes has limited the ability to more fully characterize the biodiversity of the archaea, which is clearly reflected in the relatively meager sampling effort across habitats, compared to bacteria (Figure 1D),

The ability to sample microbial life is taxonomically skewed. The Firmicutes, Proteobacteria, Actinobacteria, and Bacteroidetes represent 89.2% of the bacterial sequences and the Euryarchaeota and Thaumarchaeota 86.5% of the archaeal sequences. We sought to understand how the representation of individual phyla has changed relative to the state of the census in 2006. We used 2006 as a reference point for calibrating the dynamics of the bacterial and archaeal censuses since that was the year that the first highly parallelized 16S rRNA gene sequence

dataset was published (11). In 2006 there were 61 bacterial and 18 phyla. Since then there have been 4 new bacterial (CKC4, OC31, S2R-29, and SBYG-2791) and 2 new archaeal candidate phyla (Ancient Archaeal Group and TVG8AR30). Relative to the overall sequencing trends before and after 2006, several phyla stand out for being over and underrepresented in sequence submissions (Figure 3). Among the bacterial phyla with at least 1,000 sequences, Atribacteria and Kazan-3B-09 were sequenced 4-fold more often while Deinococcus-Thermus and Tenericutes were sequenced 2-fold less often than would have been expected since 2006. Among the archaeal phyla with at least 1,000 sequences, the Thaumarchaeota were sequenced 2.0-fold more often and the Crenarchaeota were sequenced 6.7-fold less often than expected. Together, these results demonstrate a change in the phylum-level lineages represented in the census from before and after 2006.

Focusing the census by environment. We were able to assign 89.3 and 95.1% of the sequences to one of seven broad environmental categories based on the metadata that accompanied the SILVA database (Tables 1). Across these broad categories there was wide variation in the number of sequences that have been sampled. Among bacterial sequences, the three best represented groups were from zoonotic (N=804,585), aquatic (N=214,085), and built environment (N=108,799) sources. Among the archaeal sequences the three best represented groups were the same, but ordered differently: aquatic (N=34,400), built environment (N=7,286), and zoonotic (N=5,597) (Figure 1C,D)). For both domains, soil samples were the fourth most represented category (bacteria: 74,870; archaea: 2,517). The orders of these categories was surprising considering soil and aquatic environments harbor the most microbial biomass and biodiversity (12). In spite of wide variation in sequencing depth and coverage (Table 1), the interquartile range across the fine-level categories for the bacterial OTU coverage only varied between 34.5 to 40.0 (median coverage=36.7%). The interquartile range in the OTU coverage by environment for the archaeal data was 41.5 to53.1 (median coverage=44.9%). The archaeal coverage was higher than that of the bacterial OTU coverage for all categories except the food-associated, plant surface, and other invertebrate categories. Across all categories, the bacterial and archaeal sequencing data represented a limited number of phyla (Figure 4). Among the bacterial data, the fine-scale categories were dominated by (N=) and among the archaeal data, they were dominated by (N=). Regardless, there were clear

phylum-level signatures that differentiated the various categories. Within each of the bacterial and archaeal phyla, there was considerable variation in the relative abundance of each across the categories confirming that taxonomic signatures exist to differentiate different environments even at a broad taxonomic level.

The cultured census. In the 2004 bacterial census, there was great concern that although culture-independent methods were significantly enhancing our knowledge of microbial life, there were numerous bacterial phyla with no or only a few cultured representatives. To update this assessment, we identified those sequences that came from cultured and uncultured organisms. Overall, 18.9% of bacterial sequences and 6.8% of archaeal sequences have come from isolated organisms. Comparing the fraction of sequences deposited during and before 2006 from isolates to those collected after 2006, we found that culturing rates lag by 2.4 and 2.5-fold for bacteria and archaea, respectively. Among the 65 bacterial phyla, 24 have no cultured representatives and 20 of the 14 archaeal phyla have cultured representatives. This lag is likely due to the differences in throughput of culture-dependent and -independent approaches. Of the phyla with at least one cultured representative, the median percentage of sequences coming from a culture was only 2.8% for the bacterial phyla and 1.7% for the archaeal phyla (Figure 3). So, even though many phyla have cultured representatives, there is still a skew in the representation of most phyla found in cultivation efforts. Considering the possibility that large culture-independent sequencing efforts may only be re-sequencing organisms that already exist in culture, we asked what percentage of OTUs had at least one cultured representative. We found that 16.9% of the 117,385 bacterial OTUs and 13.1% of the 4,574 archeael OTUs had at least one cultured representative (Figure 5). Comparing the percentage of sequences with cultured representatives to the percentage of OTUs containing a sequence from a cultured representative revealed strong cultivation biases for several phyla associated with model/biomedically relevant organisms (Figure 5). The Tenericutes, Proteobacteria, Firmicutes, and Spirochaeatae all had considerably higher percentages of sequences generated by cultivated representatives than would be expected based on the number of cultured organisms represented by OTUs. This likely reflects the extremely high number of cultivated *Mycobacterium tuberculosis*, *Escherichia coli*, *Bacillus*, *Streptococcus*, *Lactobacillus*, *Staphylococcus*, *Borrelia*, and others. Conversely, clades such as the *Actinobacteria*, *Fusobacteria*, and *Bacteroidetes* had a lower

percentage of sequences belonging to cultivated representatives than would be expected based on the percentage of OTUs that have sequences from cultured organisms, indicating that the cultivation efforts in these clades are relatively efficient with regards to available diversity. Regardless of these observations, the majority of OTUs from any phylum remain uncultivated, to say nothing of the diversity of organisms that may be encapsulated within the 97% sequence identity cutoff.

New technologies to access novel biodiversity. Given the shift from Sanger sequencing to platforms that offer higher throughput but shorter reads, there is concern that our ability to harvest full-length sequences from communities will remain stalled. Several culture-independent methods have been developed that offer the ability to obtain full-length sequences of the 16S rRNA gene and even the complete genome. These have included single cell genomics and assembly of short 16S rRNA gene fragments using data generated from PCR amplicons or metagenomic shotgun sequence data using the Expectation-Maximization Iterative Reconstruction of Genes from the Environment (EMIRGE) algorithm (**???**). To test the ability of these technologies to expand our knowledge of microbial diversity beyond that of traditional culutre-independent amplification and cultivation-based approaches, we compared the overlap of OTUs found using each of the new methods with the traditional approaches (Figure 6). Utilizing the 16S rRNA gene sequences extracted from the single-cell genomes available on the Integrated Microbial Genomes (IMG) system (**???**), we identified 311 bacterial and 70 archaeal sequences that met our criteria, which were assigned to 115 and 27 bacterial and archaeal OTUs, respectively. Interestingly, only 8.7 and 3.7% of the bacterial and archaeal single celled OTUs, respectively had not been observed by cultivation or PCR-based efforts. Next, we identified six studies that utilized EMIRGE to assemble 16S rRNA gene sequences from metagenomic sequences. Together these studies assembled 599 bacterial and 9 archaeal full-length sequences, which were assigned to 335 and 7 bacterial and archaeal OTUs, respectively. Only 40.6 and 60.3% of the bacterial OTUs generated by this approach were previously identified by this traditional cultivation and PCR-based approaches, respectively. Although the application of this approach to Archaea has been limited, it was still surprising that 85.7 and 85.7% of the archaeal OTUs had been previously recovered by traditional cultivation and PCR-based approaches, respectively. Finally, we pooled 76,080 bacterial sequences from five studies that utilized EMIRGE to assemble 16S rRNA gene sequences from fragmented

amplicons (**???**). These sequences were assigned to 40,213 OTUs. We were surprised that only 7.6% of these OTUs were previously found by a more traditional approach. Although these PCR-based EMIRGE results may be valid, the high degree of novelty that was observed suggests that the error of the assembled reads may be too high for generating reference sequences. Each of these methods represent promising opportunities to continue the bacterial census using full-length sequences as well as genomic information.

Conclusions

In spite of current estimates suggesting the global bacterial species richness may be as high as 10^{11} species (**???**), the current census based on full-length 16S rRNA gene sequences suggests that our current methods and sampling schemes would never reach these estimates. As we have shown, current strategies repeatedly sample the same OTUs and do a poor job of resampling rarer populations. Given this low level of OTU coverage, it is likely that there are many more bacterial and archaeal populations yet to be sampled. The recent advances in sequencing technologies have done little to expand the deposition of full-length sequences representing microbial biodiversity. We found that most sequences deposited into public databases are being made by a small number of projects that have deeply sampled similar environments. Furthermore, the number of full-lenth reads deposited into the databases has stalled. We fear that these trends will worsen unless researchers can leverage new sequencing and cultivation technologies to generate large numbers of full-length sequences from a large number of diverse samples.

During the period prior to the introduction of massively parallelized high throughput sequencing, it was common for a study to to generate dozens or hundreds of sequences per sample. The existing databases that are used for classifying sequences are based on these sequences. Thus, the sequences from those studies represent organisms that are generally abundant. We hypothesize that recent difficulties obtaining adequate classification for short sequences captured from more rare organisms is because our databases do not contain full-length references for those sequences (**???**). As stated in the 2004 census, we need to expand the diversity of environments that are deeply sequenced to improve OTU coverage. Efforts to census microbial life using short read technology

such as the International Census of Marine Microbes, the Earth Microbiome Project, and the Human Microbiome Project have significantly advanced our knowledge of microbial biogeography; however, these analyses have demonstrated the limitations of databases and taxonomies that are based on sequences from common and abundant organisms.

It is clear that considerable biodiversity has been discovered since the first census in 2004. However, much of it has been biased towards particular phyla and environments. Our analysis suggests that 94.5% of new full-length bacterial and archaeal sequences are likely to have already been seen. Meanwhile, 29.2% of bacterial and 38.5% of archaeal OTUs have only been observed once. It appears that although brute force deep sequencing of communities may increase our coverage of OTUs, new methods are needed to more directly sample rare organisms. Nevertheless, novel technologies such as single-cell genomics, metagenomics, and algorithms to recover full-length sequences from new sequencing platforms have demonstrated promise in circumventing previous limitations in identifying new OTUs. Using EMIRGE to assemble fragmented 16S rRNA gene amplicons may allow us to obtain deep coverage of communities; however, it is still unclear how faithful the assembled sequence is to that of the original organism. Additional sequencing technologies also offer the ability to directly generate full-length sequences, such as PacBio and potentially Oxford Nanopore. Initial application of PacBio to sequencing full-length fragments suggests that the sequences suffer from a high error rate. To obtain a more direct investigation of rare organisms, microbiologists are developing novel cultivation and single cell genomics techniques [**???**, Epstein dormancy, Thrash 2015; Gifford 2014]. The ability to enrich or select for specific populations using these approaches could limit the need for redundant brute force sequencing. One strength of cultivation, single cell genomics, and metagenomics approaches is that not only do the methods yield a full 16S rRNA gene sequence, but they also allow researchers to obtain the complete genome of the organism. These approaches are still in active development, we hope that through continuous refinement, they may allow us to significantly improve the coverage of OTUs in public databases.

Our ability to assess the microbial census is limited by the quality of the data in the databases. First, it is widely acknowledged that the primers used to amplify 16S rRNA genes are biased; these biases are amplified when designing primers to amplify subregions used in sequencing short

reads (Furhman, 2015). Furthermore, assembly of metagenomic data has shown the presence of introns in the 16S rRNA genes of organisms within the so-called "Candidate Phyla Radiation" (e.g. Saccharibacteria (TM7), Peregrinibacteria, Berkelbacteria (ACD58), WWE3 Microgenomates (OP11), Parcubacteria (OD1), et al.), that would preclude detection with standard PCR-based approaches (**???**). Thus, for many environments, the estimates for the census must be considered as lower bounds. Although the primers used to amplify the entire gene are also biased, obtaining a better characterization of microbial diversity will help improve the design of primers targeting subregions of the gene. Second, the willingness of researchers to contribute their sequences and the metadata describing the environment that the sequences were sampled from is critical for assessing the progress of the census and to accrue the benefits from having full-length sequences in the databases. Interestingly, the first 16S sequence was published in 1978, but was not available in a database until 1983. Similarly, only XX of the 6 studies that used the EMIRGE algorithm deposited their sequences in GenBank. This makes the sequences from the other studies effectively invisible to the search algorithms used by 16S rRNA gene-specific databases to harvest sequences. As assembly and long read technologies advance, a mechanism is needed to assess the quality of the consensus sequences and to make them easily accessible to the 16S rRNA gene-specific databases.

The advance in sequencing technologies has significantly expanded our ability to peer into the depths of microbial communities. We now have unprecedented abilities to sequence large numbers of individual communities deeply for the same cost of generating hundreds of full length sequences. Yet, unless we invest in improving our full-length databases, it will continue to be difficult to contextualize this newly discovered biodiversity within the tree of life.

Materials and Methods

Sequence data curation. The July 19, 2015 release of the ARB-formatted SILVA small subunit (SSU) reference database (SSU Ref v.123) was downloaded from http://www.arb-silva.de/fileadmin/silva_databases/release_123/ARB_files/SSURef_123_SILVA_19_07_15_opt.arb.tgz (**???**). This release is based on the EMBL-EBI/ENA Release 123, which was released in March 2015. The

SILVA curators identify potential SSU sequences using keyword searches and sequence-based search using RNAmmer (http://www.arb-silva.de/documentation/release-123/). The SILVA curators then screened the 7,168,241 resulting sequences based on a minimum length criteria (<300 nt), number of ambiguous base calls (>2%), length of sequence homopolymers (>2%), presence of vector contamination (>2%), low alignment quality value (<75), and likelihood of being chimeric (Pintail value < 50). Of the remaining sequences, the bacterial reference set retained those bacterial sequences longer than 1,200 nucleotides and the archaeal reference set retained those archaeal sequences longer than 900 nucleotides. The aligned 1,515,024 bacterial and 59,240 archaeal sequences were exported from the database using ARB along with the complete set of metadata. Additional sequence data was included from single-cell genomes available on the Integrated Microbial Genomes (IMG) system (**???**), many of which were recently obtained via the GEBA-MDM effort in Rinke et al. (**???**). "SCGC" was searched on the IMG database March 12, 2015 to download the bacterial (N=249) and archaeal (N=46) 16S rRNA gene sequences and their associated metadata. Further, sequences generated from amplicon and shotgun metagenomic data using the EMIRGE program were also included (**???**). The IMG and EMIRGE sequences were aligned against the respective SILVA-based reference using mothur (**???**). The aligned bacterial and archaeal sequence sets were pooled and processed in parallel. Using mothur, sequences were further screened to remove any sequence with more than 2 ambiguous base calls and trimmed to overlap the same alignment coordinates. The sequences in the resulting bacterial dataset overlapped bases X113X through X1344X of an *E. coli* reference sequence (V00348) and had a median length of X1,227X nt. The sequences in the resulting archaeal dataset overlapped positions X362X to X937X of a *Sulfolobus solfataricus* reference sequence (X03235) and had a median length of X580X nt. The archaeal sequences were considerably shorter than their initial length because it was necessary to find a common overlapping region across the sequences. The final datasets contained 1,411,234 bacterial and 53,546 archaeal 16S rRNA gene sequences. Sequences were assigned to OTUs using the average neighbor clustering algorithm (**???**).

Metadata curation. The metadata that was contained within the SSU Ref database was used to expand our analysis beyond a basic count of sequences and the number of OTUs in each domain. The environmental origins of the 16S rRNA gene sequences were manually classified

using seven broad "coarse" categories, and further refined to facilitate additional analyses with twenty-six more specific "fine" categories (Table S1). These were assigned based on manual curation of the "isolation_source" category within the ARB database associated with each of the sequences. For source definitions that were not identifiable by online searches, educated guesses were made or they were placed into the coarse "Other" category. There were 151,669 bacterial and 2,565 archaeal sequences where an "isolation_source" term was not collected. We ascertained whether a sequence came from a cultured organism by including those sequences that had data in their "strain" or "isolate" fields within the database and excluded any sequences that had "Unc" as part of their database name as this is a convention in the database that represents sequences from uncultured organisms. Complete tables containing the ARB-provided metadata, taxonomic information, OTU assignment, and our environmental categorizations are available at (**???**).

Calculating coverage. Sequencing coverage ($C_{Sequence}$) was quantified by two methods. The first was to use Good's coverage according to

$$C_{Sequence} = 1 - \frac{n_1}{N_t}$$

where n_1 is the number of OTUs represented by only one sequence and N_t is the total number of sequences (**???**). Although Good's coverage provides information about the success of the sequencing effort in sampling the most abundant organisms in a community, it does not directly provide information about the success of the sequencing effort in recovering previously unobserved OTUs. To quantify the ability of sequencing to identifying novel OTUs or, in other words, to quantify the "distance" in the peak of the rarefaction curves to their hypothetical asymptote, we defined "OTU coverage" (C_{OTU}) as

$$C_{OTU} = 1 - \frac{n_1}{S_t}$$

where S_t is the total number of OTUs. Whereas Good's coverage estimates the probability that a new sequence will have already been seen, OTU coverage estimates the probability that a new

OTU will match an existing one. It is therefore an extension of Good's coverage in that it quantifies the probability that, for any given set of sequences clustered into an OTU, that OTU will have already been seen. Thus, high Good's coverage means that any new sequence is unlikely to be novel, and high OTU coverage means that any new OTU is unlikely to be novel.

Data analysis. Our analysis made use of ARB (OS X v.6.0) (13), mothur (v.1.37.0) (14), and R (v.3.2.2) (15). Within R we utilized the knitr (v.1.10.5) (16), wesanderson (v.0.3.3.99)(17), and openxlsx (v. 2.4.0) (18) packages. A reproducible version of this manuscript including data extraction and processing is available at https://www.github.com/SchlossLab/Schloss_Census2_mBio_2016.

Figure 1. Number of OTUs sampled among bacterial and archaeal 16S rRNA gene sequences for different OTU definitions and level of sequencing effort. Rarefaction curves for different OTU definitions of Bacteria (A) and Archaea (B). Rarefaction curves for the coarse environments in Table 1 for Bacteria (C) and Archaea (D). The number of bacterial and archaeal OTUs observed among the longest sequences in the SILVA database continues to grow at a rate too slow to ever reach estimates of 10^6 to 10^{11} bacterial species.

Figure 2. Progression of the microbial census since the first full-length 16S rRNA gene sequence was deposited into GenBank in 1983.* The number of bacterial and archaeal 16S rRNA gene sequences deposited (A) and the new OTUs they represent (B) has increased exponentially until the last several years when the rate of change has plateaued. For both bacterial and archaeal sequences, the number of studies that are responsible for depositing more than 50% of the sequences each year has been relatively small (C).

Figure 3. Relative rate of sequence deposition for each bacterial and archaeal phylum before and after 2006 relative to the sequencing of all bacteria. The figure shows the relative rates for those phyla with at least 1,000 sequences and the x-axis is on a log2 scale. The data for all bacterial and archaeal phyla are available in Supplemental Tables 2 and 3, respectively.

Figure 4. Heatmap depicting the relative abundance of the most common bacterial and archaeal phyla across different environments. Each environmental category exhibited a phylum-level signature although the bacterial census was dominated by sequences from the Firmicutes, Proteobacteria, Actinobacteria, and Bacteroidetes and the archaeal census was dominated by sequences from the Euryarchaeota and Thaumarchaeota. The ten most abundant phyla across all environmental categories are shown. The data for all bacterial and archaeal phyla are available in Supplemental Tables 4 and 5, respectively.

Figure 5. The rate that sequences and OTUs are generated from bacterial and archaeal cultures relative to all sequences and OTUs by phlum. Phyla with greater than 1,000 sequences are listed by domain. Open circles indicate the percentage of sequences in the database that match cultured organisms. Closed circles indicate the percentage of OTUs in this analysis that contain

sequences belonging to a cultured organism. The data for all bacterial and archaeal phyla are available in Supplemental Tables 6 and 7, respectively.

Figure 6. The percentage of bacterial and archaeal OTUs found by single cell genomics and EMIRGE using PCR or metagenomics that were also detected by other. The bars comparing a method to itself indicate the percentage of OTUs that were only detected by that method.

References

1. **Brosius J**, **Palmer ML**, **Kennedy PJ**, **Noller HF**. 1978. Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli.* Proceedings of the National Academy of Sciences **75**:4801–4805.

2. **Schloss PD**, **Handelsman J**. 2004. Status of the microbial census. Microbiology and Molecular Biology Reviews **68**:686–691.

3. **Dykhuizen DE**. 1998. Santa Rosalia revisited: Why are there so many species of bacteria? Antonie van Leeuwenhoek **73**:25–33.

4. **Curtis TP**, **Sloan WT**, **Scannell JW**. 2002. Estimating prokaryotic diversity and its limits. Proceedings of the National Academy of Sciences **99**:10494–10499.

5. **Alivisatos AP**, **Blaser MJ**, **Brodie EL**, **Chun M**, **Dangl JL**, **Donohue TJ**, **Dorrestein PC**, **Gilbert JA**, **Green JL**, **Jansson JK**, **Knight R**, **Maxon ME**, **McFall-Ngai MJ**, **Miller JF**, **Pollard KS**, **Ruby EG**, **Taha SA**. 2015. A unified initiative to harness Earth's microbiomes. Science **350**:507–508.

6. **Li E**, **Hamm CM**, **Gulati AS**, **Sartor RB**, **Chen H**, **Wu X**, **Zhang T**, **Rohlf FJ**, **Zhu W**, **Gu C**, **Robertson CE**, **Pace NR**, **Boedeker EC**, **Harpaz N**, **Yuan J**, **Weinstock GM**, **Sodergren E**, **Frank DN**. 2012. Inflammatory bowel diseases phenotype, textitC. difficile and NOD2 genotype are associated with shifts in human ileum associated microbial composition. PLoS ONE **7**:e26284.

7. **Kong HH**, **Oh J**, **Deming C**, **Conlan S**, **Grice EA**, **Beatson MA**, **Nomicos E**, **Polley EC**, **Komarow HD**, **Murray PR**, **Turner ML**, **Segre JA**. 2012. Temporal shifts in the skin microbiome

associated with disease flares and treatment in children with atopic dermatitis. Genome Research **22**:850–859.

8. **Grice EA**, **Kong HH**, **Conlan S**, **Deming CB**, **Davis J**, **Young AC**, **Bouffard GG**, **Blakesley RW**, **Murray PR**, **Green ED**, **Turner ML**, **Segre JA**. 2009. Topographical and temporal diversity of the human skin microbiome. Science **324**:1190–1192.

9. **Grice EA**, **Snitkin ES**, **Yockey LJ**, **Bermudez DM**, **Liechty KW**, **Segre JA**, **Mullikin J**, **Blakesley R**, **Young A**, **Chu G**, **Ramsahoye C**, **Lovett S**, **Han J**, **Legaspi R**, **Fuksenko T**, **Reddix-Dugue N**, **Sison C**, **Gregory M**, **Montemayor C**, **Gestole M**, **Hargrove A**, **Johnson T**, **Myrick J**, **Riebow N**, **Schmidt B**, **Novotny B**, **Gupti J**, **Benjamin B**, **Brooks S**, **Coleman H**, **Ho S-l**, **Schandler K**, **Smith L**, **Stantripop M**, **Maduro Q**, **Bouffard G**, **Dekhtyar M**, **Guan X**, **Masiello C**, **Maskeri B**, **McDowell J**, **Park M**, **Thomas PJ**. 2010. Longitudinal shift in diabetic wound microbiota correlates with prolonged skin defense response. Proceedings of the National Academy of Sciences **107**:14799–14804.

10. **Harris JK**, **Caporaso JG**, **Walker JJ**, **Spear JR**, **Gold NJ**, **Robertson CE**, **Hugenholtz P**, **Goodrich J**, **McDonald D**, **Knights D**, **Marshall P**, **Tufo H**, **Knight R**, **Pace NR**. 2012.

Phylogenetic stratigraphy in the guerrero negro hypersaline microbial mat. The ISME Journal **7**:50–60.

11. **Sogin ML**, **Morrison HG**, **Huber JA**, **Welch DM**, **Huse SM**, **Neal PR**, **Arrieta JM**, **Herndl GJ**. 2006. Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proceedings of the National Academy of Sciences **103**:12115–12120.

12. **Whitman WB**, **Coleman DC**, **Wiebe WJ**. 1998. Prokaryotes: The unseen majority. Proceedings of the National Academy of Sciences **95**:6578–6583.

13. **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 Research **35**:7188–7196.

14. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**, **Oakley BB**, **Parks DH**, **Robinson CJ**, **Sahl JW**, **Stres B**, **Thallinger GG**, **Horn DJV**, **Weber CF**. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology **75**:7537–7541.

15. **R Core Team**. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

16. **Xie Y**. 2013. Dynamic documents with R and knitr. Chapman; Hall/CRC, Boca Raton, Florida.

17. **Ram K**, **Wickham H**. Wesanderson: A wes anderson palette generator.

18. **Walker A**. 2015. Openxlsx: Read, write and edit xLSX files.