The status of the microbial census: an update

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

Abstract goes here.

Importance

Importance goes here.

Introduction

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 more than 10 million 16S rRNA gene sequences to be deposited into the database (1). GenBank accessions represent nearly one-third of all sequences deposited in the database making it the best-represented gene. As Sanger sequencing has given way to to-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). 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.

The increase in sequencing volume has come at the cost of sequence length. The commonly used MiSeq-based sequencing platform from Illumina is widely used 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. Perhaps most disconcerting about this development is the sense that the increased read depth is being gained using short read platforms rather than the full-length sequences. Because these short reads are used for classification to existing taxa, we are missing the opportunity to propose novel candidate taxa and vastly underappreciating the biodiversity of microbial life. 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.

Previously, Schloss and Handelsman (2) assigned the then available 56,215 partial 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 important, the original census was performed on the cusp of radical developments in sequencing technologies. That advancement has largely moved the bulk of sequencing throughput from large sequencing centers to individual investigators and leveraged their diversity to expand the representation of organisms and environments represented in public databases.

Here we update the status of the microbial census with nearly or completely full-length 16S rRNA gene sequences. In the 13 years since the collection of data for Schloss and Handelsman's analysis, the number of full-length sequences has grown exponentially, despite the overwhelming contemporary focus by most researchers on short-read technologies. 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. Although there has been a robust growth in the number of full-length sequences deposited to GenBank since its creation in 1983, the rate of growth has stalled over the past 5 years and the deposits have been dominated by a handful of research groups studying a limited number of environments. 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. In this, continued technological advances, such as the reconstruction of nearly full-length sequences from metagenomics with EMIRGE (Miller 2011) and inovations in culturing (references needed: iChip, Epstein dormancy, Thrash 2015; Gifford 2014), will be vital.

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 nt and the archaeal reference set retained those archaeal sequences longer than 900 nt. 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. The IMG sequences were aligned against the respective SILVA-based reference using mothur (**???**). The aligned bacterial and archaeal sequence sets from SILVA and IMG were pooled and processed in parallel in mothur. 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 113 through 1344 of an *E. coli* reference sequence (V00348) and had a median length of 1,227 nt. The sequences in the resulting archaeal dataset overlapped positions 362 to 937 of a *Sulfolobus solfataricus* reference sequence (X03235) and had a median length of 580 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 1,412,681 bacterial and 53,618 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 1). These were assigned based on manual curation of the "isolation_source" category within the ARB database associated with each of the sequences (Table 1). For source definitions that were identifiable by online searches, educated guesses were made or they were placed into the coarse "Other" category. There were 150,310 bacterial and 2,590 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 environemtnal categorizations are available at (**???**).

Data analysis. Our analysis made use of ARB (OS X v.6.0) (6), mothur (v.1.37.0) (7), and R (v.3.2.2) (8). Within R we utilized knitr (v.1.10.5) (9) and openxlsx (v. 2.4.0) (10) packages. A reproducible version of this manuscript including data extraction and processing is available at https://www.github.com/SchlossLab/Schloss_Census2_mBio_2015.

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 threshold varying 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 1,412,681 near full-length bacterial 16S rRNA gene sequences we have identified 239,622, 95,726, 54,268, 14,883, and 973 OTUs at the selected thresholds (Figure 1A, Table 1). Using on the OTUs generated using a 3% threshold, we found that 95.7% of the sequences belonged to OTUs that had been observed more than once. In contrast, only 36.6% of the OTUs that were observed had been seen more than once. Paralleling the bacterial results, after sampling 53,618 archaeal 16S rRNA gene sequences we have identified 7,543, 4,208, 2,351, 815, and 112 OTUs (Figure 1B, Table 1). Using on the OTUs generated using a 3% threshold, we found that 95.2% of the sequences belonged to OTUs that had been observed more than once. In contrast, only 38.8% of the OTUs that were observed had been seen more than once. The results for the bacterial and archaeal censuses indicate that regardless of the domain, if we continue sampling with the current strategies we will continue to sample OTUs that have already been observed even though a large fraction of OTUs have only been sampled once. Furthermore, considering more than 63.4% 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 highly parallelized 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 two an average of 191,390 sequences per year (Figure 2A). Although this is still an impressive number of sequences, since 2,007 the number of new bacterial OTUs has plateaued at an average of 9,646.9 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, 904,013 total sequences were submitted and 6 submissions from 5 studies were responsible

for depositing 548,274 (60.6% of all sequences). These studies generated sequences from the human gastrointestinal tract (11), human skin (12, 13), murine skin (14), and hypersaline microbial mats (15). In contrast to recent years, between 1995 and 2006, an average of 39.8 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 7,079.2 sequences each year since then. The number of new archaeal OTUs represented by these sequences began to slow in 2005 with an average of 359 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.

The ability to sample microbial life is taxonomically skewed (meh.) The Firmicutes, Proteobacteria, Actinobacteria, and Bacteroidetes represent 89.1% of the bacterial sequences and the Euryarchaeota and Thaumarchaeota 86.4% 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 and ushered in a radical change in how microbial communities are studied (16). In 2006 there were 62 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 phylum_effort.pdf). 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 94.5% of the sequences to one of seven broad environmental categories based on the metadata that accompanied the SILVA database. Across these broad categories there was wide variation in the number of sequences that have been sampled. Among the bacterial sequences the three best represented groups were from zoonotic (N=799,542), aquatic (N=226,070), and built environment (N=106,723) sources and among the archaeal sequences the three best represented groups were from aquatic (N=34,434), built environment (N=7,019), and zoonotic (N=5,597) sources. For both domains, soil samples were the fourth most represented category (bacteria: 73,804; archaea: 2,521). The orders of these categories was surprising considering soil and aquatic environments harbor the most microbial biomass and biodiversity (17). In spite of wide variation in sequencing depth and coverage (Table 1), the interquartile range across the fine-level categories for the bacterial OTU-based coverage only varied between 31.3 to 36.6 (median coverage=33.8%). The interquartile range in the OTU-based coverage by environment for the archaeal data was 38.5 to51.7 (median coverage=41.9%). 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 category_phylum_heatmap.pdf). Among the bacterial data, the fine-scale categories were dominated by Proteobacteria (N=22), Firmicutes (N=4), Actinobacteria (N=1), and Bacteroidetes (N=1) and among the archaeal data, they were dominated by Euryarchaeota (N=17) and Thaumarchaeota (N=10). 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, 18.6% 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 2.5 and 2.4-fold for bacteria and archaea, respectively. Among the 67 bacterial phyla, 19 have no cultured representatives and 20 of the 10 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 2.4% for the bacterial phyla and 1.7% for the archaeal phyla (Figure phylum_effort.pdf). 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 13.0% of the 95,734 bacterial OTUs and 9.1% of the 4,205 archeael OTUs had at least one cultured representative. Considering that the percentage of sequences from bacterial cultures is greater than the percentage of OTUs from bacterial cultures, we suspect that bacterial cultivation efforts are largely resampling the same diversity relative to culture-independent approaches. In contrast, the relationship is reversed among the archaea indicating that archaeal cultivation efforts continue to sample greater biodiversity than culture-independent approaches.

here

New technologies to access novel biodiversity Assembly of metagenomic and single cell shotgun sequence data offers the hope of identifying large fragments of genomic data from as yet uncultured organisms. To test the ability of single cell technologies to identify

- Single cell genomics
- Metagenomes

Conclusions

Future for PacBio in generating full-length sequences

The first 16S sequence was published in 1978, not deposited until 1983. A bit of an allegory for our time.

Impact of EMIRGE - http://www.genomebiology.com/2011/12/5/R44

Renewed call for cultivation

What are the most significant improvements since 2004, and where are we still lacking the most data?

Recent data suggests that a considerable diversity of microorganisms may be missing based on biases in existing 16S rRNA gene primers. This dataset does not include sequences from metagenomic assemblies but Brown et al. (2015) have used such assemblies to show evidence for introns in the 16S rRNA genes of organisms within the so-called “Candidate Phyla Radiation” (CPR- Saccharibacteria (TM7), Peregrinibacteria, Berkelbacteria (ACD58), WWE3 Microgenomates (OP11), Parcubacteria (OD1), et al.), that would preclude detection with standard cultivation-independent microbial surveys. Furthermore, many of these CPR organisms are very small and frequently pass through 0.2 µm filters (Luef, 2015). Thus, for many environments, the estimates within must be considered as lower bounds.

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. Indeed, continual community efforts to obtain 16S rRNA gene assessments of every environment possible have presented us with an ever-increasing estimate of total microbial diversity and the concomitant excitement of frontier science. While reliance on this gene subjects us to biases created by primer selection [REFS], differences in amplification strength [REFS] and fidelity [REFS], internal features which may disrupt traditional measurements [REFS], and potentially misleading classification due to infrequent horizontal gene transfer [REFS], the total data available from persistent collection of 16S rRNA

gene sequences nevertheless dwarfs that of any other genetic marker. Thus, an attempt to quantify how much of the microbial world has been revealed inevitably starts there.

Given the relatively low number of archaeal sequences that have been deposited for many of these categories, it is possible that the coverage for many categories may not be reliable (e.g. aerosol and plant surfaces). By decomposing the census by environmental categories, it is clear that even among the best sampled environments, our ability to claim anything but a basic characterization of microbial biodiversity is limited.

Figure 1. Sequences deposited by year

Pat, this is a placeholder for you

Figure 2. Sequence deposition by study

Pat, ibid.

Figure 3. Sampling by OTU cutoff

I only have an old code from Rene for this, and I don't think it reports all the sequences

Figure 4A. Bacterial sampling by environment

Figure 4B. Archaeal sampling by environment

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. 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 Earths microbiomes. Science **350**:507–508.

6. **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.

7. **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.

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

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

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

11. **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, c. difficile and NOD2 genotype are associated with shifts in human ileum associated microbial composition. PLoS ONE **7**:e26284.

12. **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.

13. **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.

14. **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.

15. **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.

16. **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.

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