

# The status of the microbial census: an update

Running title: The microbial census

Patrick D. Schloss<sup>1†</sup>, Rene Girard<sup>2</sup>, Thomas Martin<sup>2</sup>, Joshua Edwards<sup>2</sup>, and J. Cameron Thrash<sup>2†</sup>

† To whom correspondence should be addressed: [pschloss@umich.edu](mailto:pschloss@umich.edu) and [thrashc@lsu.edu](mailto: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

## 1 **Abstract**

2 A census is typically carried out for people at a national level; however, microbial ecologists  
3 have implemented a molecular census of bacteria and archaea by sequencing their 16S rRNA  
4 genes. We assessed how well the microbial census of full-length 16S rRNA gene sequences is  
5 proceeding in the context of recent advances in high throughput sequencing technologies. Among  
6 the 1,411,234 and 53,546 full-length bacterial and archaeal sequences, 94.5% and  
7 95.1% of the bacterial and archaeal sequences, respectively, belonged to operational taxonomic  
8 units (OTUs) that have been observed more than once. Although these metrics suggest that the  
9 census is approaching completion, 29.2% of the bacterial and 38.5% of the archaeal OTUs have  
10 been observed more than once. Thus, there is still considerable microbial diversity to be explored.  
11 Unfortunately, the rate of new full-length sequences has been declining and new sequences are  
12 primarily being deposited by a small number of studies. Furthermore, sequences from soil and  
13 aquatic environments, which are known to be rich in bacterial diversity, only represent 7.8 and  
14 16.5% of the census while sequences associated with zoonotic environments represent 55.0%  
15 of the census. Continued use of traditional approaches and new technologies such as single  
16 cell genomics and short read assembly are likely to improve our ability to sample rare OTUs if  
17 it is possible to overcome this sampling bias. The success of ongoing efforts to use short read  
18 sequencing to characterize microbial communities requires that researchers strive to expand the  
19 depth and breadth of the microbial census.

## 20 **Importance**

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

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

## 31 Introduction

32 The effort to quantify the number of different organisms in a system remains fundamental to  
33 understanding ecology (1, 2). At the scale of microorganisms, small physical sizes, morphological  
34 ambiguity, and highly variable population sizes complicate this process. Furthermore, creating  
35 standards for delimiting what makes one microbe “different” from another has been contentious (3,  
36 4). In spite of these challenges, we continue to peel back the curtain on the microbial world with  
37 the aid of more and more informative, if still limited, technologies like cultivation, 16S rRNA gene  
38 surveys, single cell technologies, and metagenomics.

39 Generating a comprehensive understanding of any system with a single gene may seem a fool’s  
40 errand, yet we have learned a considerable amount regarding the diversity, dynamics, and natural  
41 history of microorganisms using the venerable 16S rRNA gene. In 1983, the full-length 16S rRNA  
42 gene sequence of *Escherichia coli* (accession J01695) was deposited into NCBI’s GenBank making  
43 it the first of what is now more than 10 million 16S rRNA gene sequences to be deposited into the  
44 database (5). 16S rRNA gene accessions represent nearly one-third of all sequences deposited in  
45 GenBank, making it the best-represented gene. As Sanger sequencing has given way to so-called  
46 “next generation sequencing” technologies, hundreds of millions of 16S rRNA gene sequences  
47 have been deposited into the NCBI’s Sequence Read Archive. The expansion in sequencing  
48 throughput and increased access to sequencing technology has allowed for more environments to  
49 be sequenced at a deeper coverage, resulting in the identification of novel taxa. The ability to obtain  
50 sequence data from microorganisms without cultivation has radically altered our perspective of their  
51 role in nearly every environment from deep ocean sediment cores (e.g. accession AY436526) to  
52 the International Space Station (e.g. accession DQ497748).

53 Previously, Schloss and Handelsman (6) assigned the 56,215 partial 16S rRNA gene sequences  
54 that were available in the Ribosomal Database Project to operational taxonomic units (OTUs) and  
55 concluded that the sampling methods of the time were insufficient to identify the previously estimated  
56  $10^7$  to  $10^9$  different species (7, 8). That census called for a broader and deeper characterization of  
57 all environments. Refreshingly, this challenge was largely met. There have been major investments  
58 in studying the Earth’s microbiome using 16S rRNA gene sequencing through initiatives such as the

59 Human Microbiome Project (9), the Earth Microbiome Project (10), and the International Census of  
60 Marine Microorganisms (11). But most importantly, the original census was performed on the cusp  
61 of radical developments in sequencing technologies. That advancement has moved the generation  
62 of sequencing throughput from large sequencing centers to individual investigators and leveraged  
63 their diverse interests to expand the representation of organisms and environments represented in  
64 public databases.

65 It is disconcerting that the increase in sequencing volume has come at the cost of sequence  
66 length. The commonly used MiSeq-based sequencing platform from Illumina is extensively used to  
67 sequence the approximately 250 bp V4 hypervariable region of the 16S rRNA gene; other schemes  
68 have used different parts of the gene that are generally shorter than 500 bp. The number of OTUs  
69 that are sampled when using different regions within the 16S rRNA gene can vary considerably and  
70 the genetic diversity within these regions typically has only a modest correlation the genetic diversity  
71 of the full-length sequence (12, 13). Thus, it remains unclear to what degree richness estimates  
72 from short read technologies over or underestimate the numbers from full-length sequences.  
73 Furthermore, we likely lack the references necessary to adequately classify the novel biodiversity  
74 we are sampling when we generate 100-times the sequence data from a community than we did  
75 using full-length sequencing.

76 Here we update the status of the microbial census with full-length 16S rRNA gene sequences. In  
77 the 13 years since the collection of data for Schloss and Handelsman's initial census, the number  
78 of full-length sequences has grown exponentially, despite the overwhelming contemporary focus by  
79 most researchers on short-read technologies. This update to the census allows us to evaluate the  
80 relative sampling thoroughness for different environments and clades and make an argument for  
81 the continued need to collection full-length sequence data from many systems that have a long  
82 history of study. As researchers consider coalescing into a Unified Microbiome Initiative (14), it  
83 will be important to balance the need for mechanism-based studies with the need to generate  
84 full-length reference sequences from a diversity of environments.

## 85 **Results and Discussion**

86 ***The status of the bacterial and archaeal census.*** To assess the field's progress in characterizing  
87 the biodiversity of bacteria and archaea, we assigned each 16S rRNA gene sequence to OTUs  
88 using distance thresholds that varied between 0 and 20%. Although it is not possible to link a  
89 specific taxonomic level (e.g. species, genus, family, etc.) to a specific distance threshold, we  
90 selected distances of 0, 3, 5, 10, and 20% because they are widely regarded as representing the  
91 range of genetic diversity of the 16S rRNA gene within each domain. By rarefaction, it was clear that  
92 the ongoing sampling efforts have started to saturate the number of current OTUs. After sampling  
93 1,411,234 near full-length bacterial 16S rRNA gene sequences we have identified 217,645, 108,950,  
94 66,819, 15,743, and 3,731 OTUs at the respective thresholds (Figure 1A, Table 1). Using only  
95 the OTUs generated using a 3% threshold, we calculated a 94.5% Good's coverage (percent of  
96 sequences belonging to OTUs that have been observed more than once), but only 29.2% OTU  
97 coverage (percent of the OTUs that have been observed more than once). Paralleling the bacterial  
98 results, after sampling 53,546 archaeal 16S rRNA gene sequences we have identified 11,040,  
99 4,252, 2,364, 812, and 110 OTUs (Figure 1B, Table 1). Using only the OTUs generated with a  
100 3% threshold, we calculated a 95.1% Good's coverage, but only 38.5% OTU coverage. These  
101 results indicate that regardless of the domain, continued sampling with the current strategies for  
102 generating full-length sequences will largely reveal OTUs that have already been observed, even  
103 though a large fraction of OTUs have only been sampled once. Considering more than 70.8% of  
104 the OTUs have only been observed once, it is likely that an even larger number of OTUs have yet  
105 to be sampled for both domains.

106 ***Sequencing efforts are a source of bias in the census.*** One explanation for the large number  
107 of OTUs that have only been observed once is that with the the broad adoption of sequencing  
108 platforms that generate short sequence reads, the rate of full-length sequence generation has  
109 declined. In fact, since 2009 the number of new bacterial sequences generated has slowed to an  
110 average of 189,960 sequences per year (Figure 2A). Although this is still an impressive number of  
111 sequences, since 2007 the number of new bacterial OTUs has plateaued at an average of 11,184  
112 new OTUs per year (Figure 2B). Given the expense of generating full-length sequences using the

113 Sanger sequencing technology and the transition to other platforms at that time, we expected that  
114 the large number of sequences were being deposited by a handful of large projects. Indeed, when  
115 we counted the number of submissions responsible for depositing 50% of the sequences, we found  
116 that with the exception of 2006 and 2013, eight or fewer studies were responsible for depositing the  
117 majority of the full-length sequences each year since 2005 (Figure 2C). Between 2009 and 2012,  
118 908,190 total sequences were submitted and 6 submissions from 5 studies were responsible for  
119 depositing 550,274 (60.6% of all sequences). These studies generated sequences from the human  
120 gastrointestinal tract (15), human skin (16, 17), murine skin (18), and hypersaline microbial mats  
121 (19). The heavy zoonotic focus is reflected in the rarefaction curve for this category (Figure 1C). In  
122 contrast to recent years, between 1995 and 2006, an average of 39.3 studies were responsible for  
123 submitting more than half of the sequences each year. Although the recent deep surveys represent  
124 significant contributions to our knowledge of bacterial biogeography, their small number and lack of  
125 environmental diversity is indicative of the broader problems in advancing the bacterial census.

126 The depth of sequencing being done to advance the archaeal census has been 26-times less  
127 than that of the bacterial census (Table 1). The annual number of sequences submitted has  
128 largely paralleled that of the bacterial census with a plateau starting in 2009 and an average of  
129 7,075 sequences each year since then. The number of new archaeal OTUs represented by these  
130 sequences began to slow in 2005 with an average of 355.5 new OTUs per year. With the exception  
131 of 2012 and 2014, the number of submissions responsible for more than 50% of the archaeal  
132 sequences submitted per year has varied between 2 and 11 submissions per year. The clear bias  
133 towards sequencing bacterial 16S rRNA genes has limited the ability to more fully characterize the  
134 biodiversity of the archaea, which is clearly reflected in the relatively meager sampling effort across  
135 habitats, compared to bacteria (Figure 1D),

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

142 was published (20). Based on the representation of sequences within the SILVA database, in 2006  
143 there were 61 bacterial and 18 phyla. Since then there have been 4 new bacterial (CKC4, OC31,  
144 S2R-29, and SBYG-2791) and 2 new archaeal candidate phyla (Ancient Archaeal Group and  
145 TVG8AR30). Relative to the overall sequencing trends before and after 2006, several phyla stand  
146 out for being over and underrepresented in sequence submissions (Figure 3). Among the bacterial  
147 phyla with at least 1,000 sequences, Atribacteria and Kazan-3B-09 were sequenced 4-fold more  
148 often while Deinococcus-Thermus and Tenericutes were sequenced 2-fold less often than would  
149 have been expected since 2006. Among the archaeal phyla with at least 1,000 sequences, the  
150 Thaumarchaeota were sequenced 2.0-fold more often and the Crenarchaeota were sequenced  
151 6.7-fold less often than expected. Together, these results demonstrate a change in the phylum-level  
152 lineages represented in the census from before and after 2006 and encouragingly, show that some  
153 underrepresented phyla are becoming better sampled.

154 ***Focusing the census by environment.*** We were able to assign 89.3 and 95.1% of the sequences  
155 to one of seven broad environmental categories based on the metadata that accompanied the  
156 SILVA database (Tables 1). Across these broad categories there was wide variation in the number  
157 of sequences that have been sampled. Among bacterial sequences, the three best represented  
158 groups were from zoonotic (N=804,585), aquatic (N=214,085), and built environment (N=108,799)  
159 sources. Among the archaeal sequences the three best represented groups were the same,  
160 but ordered differently: aquatic (N=34,400), built environment (N=7,286), and zoonotic (N=5,597)  
161 (Figure 1C,D)). For both domains, soil samples were the fourth most represented category (bacteria:  
162 74,870; archaea: 2,517). The orders of these categories was surprising considering soil and aquatic  
163 environments harbor the most microbial biomass and biodiversity (21). In spite of wide variation in  
164 sequencing depth and coverage (Table 1), the interquartile range across the fine-level categories  
165 for the bacterial OTU coverage only varied between 34.5 to 40.0 (median coverage=36.7%).  
166 The interquartile range in the OTU coverage by environment for the archaeal data was 41.5 to  
167 53.1 (median coverage=44.9%). The archaeal coverage was higher than that of the bacterial  
168 OTU coverage for all categories except the food-associated, plant surface, and other invertebrate  
169 categories. Across all categories, the bacterial and archaeal sequencing data represented a  
170 limited number of phyla (Figure 4). Among the bacterial data, the fine-scale categories were



171 dominated by Proteobacteria (N=24), Firmicutes (N=2), and Actinobacteria (N=1) and among  
172 the archaeal data, they were dominated by Euryarchaeota (N=16), Thaumarchaeota (N=10), and  
173 Aenigmarchaeota (N=1). Regardless, there were clear phylum-level signatures that differentiated  
174 the various categories. Within each of the bacterial and archaeal phyla, there was considerable  
175 variation in the relative abundance of each across the categories confirming that taxonomic  
176 signatures exist to differentiate different environments even at a broad taxonomic level.

177 ***The cultured census.*** In the 2004 bacterial census, there was great concern that although  
178 culture-independent methods were significantly enhancing our knowledge of microbial life, there  
179 were numerous bacterial phyla with no or only a few cultured representatives. To update this  
180 assessment, we identified those sequences that came from cultured and uncultured organisms.  
181 Overall, 18.9% of bacterial sequences and 6.8% of archaeal sequences have come from isolated  
182 organisms. Comparing the fraction of sequences deposited during and before 2006 from isolates to  
183 those collected after 2006, we found that culturing rates lag by 2.4 and 2.5-fold for bacteria and  
184 archaea, respectively. Among the 65 bacterial phyla, 24 have no cultured representatives and 14 of  
185 the 20 archaeal phyla have no cultured representatives. This lag is likely due to the differences  
186 in throughput of culture-dependent and -independent approaches. Of the phyla with at least one  
187 cultured representative, the median percentage of sequences coming from a culture was only 2.8%  
188 for the bacterial phyla and 1.7% for the archaeal phyla (Figure 5). Even though many phyla have  
189 cultured representatives, there is still a skew in the representation of most phyla found in cultivation  
190 efforts.

191 Considering the possibility that large culture-independent sequencing efforts may only be  
192 re-sequencing organisms that already exist in culture, we asked what percentage of OTUs had at  
193 least one cultured representative. We found that 16.9% of the 117,385 bacterial OTUs and 13.1%  
194 of the 4,574 archaeal OTUs had at least one cultured representative (Figure 5). Comparing the  
195 percentage of sequences with cultured representatives to the percentage of OTUs containing a  
196 sequence from a cultured representative revealed a strong cultivation bias within the Firmicutes,  
197 which had a higher percentage of sequences generated by cultivated representatives than would be  
198 expected based on the number of cultured organisms represented by OTUs (Figure 5). This likely  
199 reflects the extremely high number of cultivated biomedically relevant cultivars from genera such

200 as *Bacillus*, *Streptococcus*, *Lactobacillus*, *Staphylococcus*, and others. Conversely, many phyla,  
201 including Cyanobacteria, Actinobacteria, Bacteroidetes, and Nitrospirae, had a lower percentage  
202 of sequences belonging to cultivated representatives than would be expected based on the  
203 percentage of OTUs that have sequences from cultured organisms, indicating that the cultivation  
204 efforts in these clades are relatively inefficient with regards to available diversity. Nevertheless, it is  
205 clear that the majority of OTUs from any phylum remain uncultivated, to say nothing of the diversity  
206 of organisms that may be encapsulated within the 97% sequence identity cutoff.

207 ***New technologies to access novel biodiversity.*** Given the shift from Sanger sequencing to  
208 platforms that offer higher throughput but shorter reads, there is concern that our ability to harvest  
209 full-length sequences from communities will remain stalled. Several culture-independent methods  
210 have been developed that offer the ability to obtain full-length sequences of the 16S rRNA gene  
211 and even the complete genome. These have included single cell genomics (22) and assembly  
212 of short 16S rRNA gene fragments using data generated from PCR amplicons or metagenomic  
213 shotgun sequence data with the Expectation-Maximization Iterative Reconstruction of Genes from  
214 the Environment (EMIRGE) algorithm (23, 24). To test the ability of these technologies to expand  
215 our knowledge of microbial diversity beyond that of traditional approaches, we compared the overlap  
216 of OTUs found using each of the new methods with the traditional approaches (Figure 6). Utilizing  
217 the 16S rRNA gene sequences extracted from the single-cell genomes available on the Integrated  
218 Microbial Genomes (IMG) system (25), we identified 311 bacterial and 70 archaeal sequences,  
219 which were assigned to 115 and 27 bacterial and archaeal OTUs, respectively. Interestingly, only  
220 8.7 and 3.7% of the bacterial and archaeal single celled OTUs, respectively, had not been observed  
221 by previous efforts. Next, we identified six studies that utilized EMIRGE to assemble 16S rRNA  
222 gene sequences from metagenomic sequences (23, 26–30). Together these studies assembled  
223 599 bacterial and 9 archaeal full-length sequences, which were assigned to 335 and 7 bacterial  
224 and archaeal OTUs, respectively. Only 40.6 and 60.3% of the bacterial OTUs generated by this  
225 approach were previously identified by this traditional cultivation and PCR-based approaches,  
226 respectively. Although the application of this approach to Archaea has been limited, it was still  
227 surprising that 85.7 and 85.7% of the archaeal OTUs had been previously recovered by traditional  
228 cultivation and PCR-based approaches, respectively. Finally, we pooled 76,080 bacterial sequences

229 from five studies that utilized EMIRGE to assemble 16S rRNA gene sequences from fragmented  
230 amplicons (24, 31–34). These sequences were assigned to 40,213 OTUs. We were surprised that  
231 only 7.6% of these OTUs were previously found by a more traditional approach. Although these  
232 PCR-based EMIRGE results may be valid, the high degree of novelty that was observed suggests  
233 that the error of the assembled reads may be too high for generating reference sequences. Each of  
234 these methods represent promising opportunities to continue the bacterial census using full-length  
235 sequences as well as genomic information.

## 236 **Conclusions**

237 It is clear that considerable biodiversity has been discovered since the first census in 2004. However,  
238 much of it has been biased towards particular phyla and environments. Our analysis suggests  
239 that 94.5% of new full-length bacterial and archaeal sequences are likely to have already been  
240 seen. Meanwhile, 29.2% of bacterial and 38.5% of archaeal OTUs have only been observed once.  
241 In spite of current estimates suggesting the global bacterial species richness may be as high as  
242  $10^{12}$  species (35), the current census based on full-length 16S rRNA gene sequences suggests  
243 that existing sampling methods will prevent us from acquiring full-length sequences for that level of  
244 diversity. As we have shown, current strategies repeatedly sample the same OTUs and do a poor  
245 job of resampling rarer populations. Given this low level of OTU coverage, it is likely that there are  
246 many more bacterial and archaeal populations yet to be sampled.

247 There are several additional reasons to suspect that the current census should be considered  
248 conservative. First, we found that most sequences recently deposited into public databases  
249 are being made by a small number of projects that have deeply sampled similar environments,  
250 and the number of full-length reads deposited into the databases has stalled. Second, it is  
251 widely acknowledged that 16S rRNA gene primers are biased; these biases are amplified when  
252 designing primers to amplify subregions used in sequencing short reads (36). Assembly of  
253 metagenomic data has shown the presence of introns in the 16S rRNA genes of organisms  
254 within the so-called “Candidate Phyla Radiation” (e.g. Saccharibacteria (TM7), Peregrinibacteria,  
255 Berkelbacteria (ACD58), WWE3 Microgenomates (OP11), Parcubacteria (OD1), et al.) that would

256 preclude detection with standard PCR-based approaches (37, 38). Third, the willingness of  
257 researchers to contribute their sequences and the metadata describing the environment that the  
258 sequences were sampled from is critical for assessing the progress of the census and to accrue the  
259 benefits from having full-length sequences in the databases. Interestingly, the first 16S sequence  
260 was published in 1978, but was not available in a database until 1983. Similarly, only 5 of the 11  
261 studies that used the EMIRGE algorithm deposited their sequences in GenBank. This makes the  
262 sequences from the other studies effectively invisible to the search algorithms used by 16S rRNA  
263 gene-specific databases to harvest sequences. As assembly and long read technologies advance,  
264 a mechanism is needed to assess the quality of the consensus sequences and to make them easily  
265 accessible to the 16S rRNA gene-specific databases.

266 Efforts to census microbial life using short read technology such as the International Census  
267 of Marine Microbes, the Earth Microbiome Project, and the Human Microbiome Project have  
268 significantly advanced our knowledge of microbial biogeography; however, these analyses have  
269 demonstrated the limitations of databases and taxonomies that are based on sequences from  
270 common and abundant organisms. During the period prior to the introduction of massively  
271 parallelized high throughput sequencing, it was common for a study to generate dozens or hundreds  
272 of sequences per sample. The existing databases that are used for classifying sequences are based  
273 on those sequences, which represent organisms that are generally abundant. We hypothesize that  
274 recent difficulties obtaining adequate classification for short sequences captured from more rare  
275 organisms are because our databases do not contain full-length references for those sequences.  
276 We fear that these trends will worsen unless researchers can leverage new sequencing and  
277 cultivation technologies to generate large numbers of full-length sequences from a large number of  
278 diverse samples.

279 Novel technologies such as single-cell genomics, metagenomics, and algorithms to recover  
280 full-length sequences from new sequencing platforms have demonstrated promise in circumventing  
281 previous limitations in identifying new OTUs. Using EMIRGE to assemble fragmented 16S rRNA  
282 gene amplicons may allow us to obtain deep coverage of communities; however, it is still unclear  
283 how faithful the assembled sequence is to that of the original organism. Additional sequencing  
284 technologies also offer the ability to directly generate full-length sequences, such as PacBio and

285 potentially Oxford Nanopore. Initial application of PacBio to sequencing full-length fragments  
286 suggests that the sequences suffer from a high error rate (39). To obtain a more direct investigation  
287 of rare organisms, microbiologists are developing novel cultivation and single cell genomics  
288 techniques (???, 40–42). The ability to enrich or select for specific populations using these  
289 approaches could limit the need for redundant brute force sequencing. These approaches are  
290 still in active development, and we hope that through continuous refinement, they may allow us to  
291 significantly improve the coverage of OTUs in public databases.

## 292 **Materials and Methods**

293 ***Sequence data curation.*** The July 19, 2015 release of the ARB-formatted SILVA small subunit  
294 (SSU) reference database (SSU Ref v.123) was downloaded from [http://www.arb-silva.de/fileadmin/  
295 silva\\_databases/release\\_123/ARB\\_files/SSURef\\_123\\_SILVA\\_19\\_07\\_15\\_opt.arb.tgz](http://www.arb-silva.de/fileadmin/silva_databases/release_123/ARB_files/SSURef_123_SILVA_19_07_15_opt.arb.tgz) (43). This  
296 release is based on the EMBL-EBI/ENA Release 123, which was released in March 2015. The  
297 SILVA curators identify potential SSU sequences using keyword searches and sequence-based  
298 search using RNAmmer (<http://www.arb-silva.de/documentation/release-123/>). The SILVA curators  
299 then screened the 7,168,241 resulting sequences based on a minimum length criteria (<300 nt),  
300 number of ambiguous base calls (>2%), length of sequence homopolymers (>2%), presence of  
301 vector contamination (>2%), low alignment quality value (<75), and likelihood of being chimeric  
302 (Pintail value < 50). Of the remaining sequences, the bacterial reference set retained those  
303 bacterial sequences longer than 1,200 nucleotides and the archaeal reference set retained those  
304 archaeal sequences longer than 900 nucleotides. The aligned 1,515,024 bacterial and 59,240  
305 archaeal sequences were exported from the database using ARB along with the complete set  
306 of metadata. Additional sequence data was included from single-cell genomes available on the  
307 Integrated Microbial Genomes (IMG) system (25), many of which were recently obtained via the  
308 GEBA-MDM effort in Rinke et al. (22). “SCGC” was searched on the IMG database March 12,  
309 2015 to download the bacterial (N=249) and archaeal (N=46) 16S rRNA gene sequences and their  
310 associated metadata. Further, sequences generated from amplicon and shotgun metagenomic  
311 data using the EMIRGE program were also included (23, 24). The IMG and EMIRGE sequences

312 were aligned against the respective SILVA-based reference using mothur (44). The aligned bacterial  
313 and archaeal sequence sets were pooled and processed in parallel. Using mothur, sequences were  
314 further screened to remove any sequence with more than 2 ambiguous base calls and trimmed  
315 to overlap the same alignment coordinates. The sequences in the resulting bacterial dataset  
316 overlapped bases 113 through 1350 of an *E. coli* reference sequence (V00348) and had a median  
317 length of 1,233 nt. The sequences in the resulting archaeal dataset overlapped positions 362 to 937  
318 of a *Sulfolobus solfataricus* reference sequence (X03235) and had a median length of 580 nt. The  
319 archaeal sequences were considerably shorter than their initial length because it was necessary to  
320 find a common overlapping region across the sequences. The final datasets contained 1,411,234  
321 bacterial and 53,546 archaeal 16S rRNA gene sequences. Sequences were assigned to OTUs  
322 using the average neighbor clustering algorithm (45).

323 **Metadata curation.** The metadata that was contained within the SSU Ref database was used  
324 to expand our analysis beyond a basic count of sequences and the number of OTUs in each  
325 domain. The environmental origins of the 16S rRNA gene sequences were manually classified  
326 using seven broad “coarse” categories, and further refined to facilitate additional analyses with  
327 twenty-six more specific “fine” categories (Table S1). These were assigned based on manual  
328 curation of the “isolation\_source” category within the ARB database associated with each of  
329 the sequences. For source definitions that were not identifiable by online searches, educated  
330 guesses were made or they were placed into the coarse “Other” category. There were 151,669  
331 bacterial and 2,565 archaeal sequences where an “isolation\_source” term was not collected. We  
332 ascertained whether a sequence came from a cultured organism by including those sequences that  
333 had data in their “strain” or “isolate” fields within the database and excluded any sequences that  
334 had “Unc” as part of their database name as this is a convention in the database that represents  
335 sequences from uncultured organisms. Complete tables containing the ARB-provided metadata,  
336 taxonomic information, OTU assignment, and our environmental categorizations are available at  
337 FigShare for the bacterial (<https://dx.doi.org/10.6084/m9.figshare.2064927>) and archaeal (<https://dx.doi.org/10.6084/m9.figshare.2064942>) data.

339 **Calculating coverage.** Sequencing coverage ( $C_{\text{Sequence}}$ ) was quantified by two methods. The first  
340 was to use Good’s coverage according to

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

341 where  $n_1$  is the number of OTUs represented by only one sequence and  $N_t$  is the total number  
342 of sequences (46). Although Good's coverage provides information about the success of the  
343 sequencing effort in sampling the most abundant organisms in a community, it does not directly  
344 provide information about the success of the sequencing effort in recovering previously unobserved  
345 OTUs. To quantify the ability of sequencing to identifying novel OTUs or, in other words, to quantify  
346 the "distance" in the peak of the rarefaction curves to their hypothetical asymptote, we defined  
347 "OTU coverage" ( $C_{OTU}$ ) as

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

348 where  $S_t$  is the total number of OTUs. Whereas Good's coverage estimates the probability that a  
349 new sequence will have already been seen, OTU coverage estimates the probability that a new  
350 OTU will match an existing one. It is therefore an extension of Good's coverage in that it quantifies  
351 the probability that, for any given set of sequences clustered into an OTU, that OTU will have  
352 already been seen. Thus, high Good's coverage means that any new sequence is unlikely to be  
353 novel, and high OTU coverage means that any new OTU is unlikely to be novel.

354 **Data analysis.** Our analysis made use of ARB (OS X v.6.0) (43), mothur (v.1.37.0) (44), and R  
355 (v.3.2.2) (47). Within R we utilized the knitr (v.1.10.5), wesanderson (v.0.3.3.99), and openxlsx (v.  
356 2.4.0) packages. A reproducible version of this manuscript including data extraction and processing  
357 is available at [https://www.github.com/SchlossLab/Schloss\\_Census2\\_mBio\\_2016](https://www.github.com/SchlossLab/Schloss_Census2_mBio_2016).

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

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

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

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

381 **Figure 5. The rate that sequences and OTUs are generated from bacterial and archaeal**  
382 **cultures relative to all sequences and OTUs by phylum.** Phyla with greater than 1,000 sequences  
383 are listed by domain. Open circles indicate the percentage of sequences in the database that match  
384 cultured organisms. Closed circles indicate the percentage of OTUs in this analysis that contain  
385 sequences belonging to a cultured organism. The data for all bacterial and archaeal phyla are



386 available in Supplemental Tables 6 and 7, respectively.

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

## 390 References

- 391 1. **McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Bence HK, Dornelas M,**  
392 **Enquist BJ, Green JL, He F, Hurlbert AH, Magurran AE, Marquet PA, Maurer BA, Ostling A,**  
393 **Soykan CU, Ugland KI, White EP.** 2007. Species abundance distributions: Moving beyond single  
394 prediction theories to integration within an ecological framework. *Ecology Letters* **10**:995–1015.  
395 doi:[10.1111/j.1461-0248.2007.01094.x](https://doi.org/10.1111/j.1461-0248.2007.01094.x).
- 396 2. **Hubbell SP.** 2001. *A Unified Theory of Biodiversity and Biogeography*. Princeton University  
397 Press, Princeton.
- 398 3. **Konstantinidis KT, Ramette A, Tiedje JM.** 2006. The bacterial species definition in  
399 the genomic era. *Philosophical Transactions of the Royal Society B: Biological Sciences*  
400 **361**:1929–1940. doi:[10.1098/rstb.2006.1920](https://doi.org/10.1098/rstb.2006.1920).
- 401 4. **Oren A, Garrity GM.** 2013. Then and now: A systematic review of the systematics of prokaryotes  
402 in the last 80 years. *Antonie van Leeuwenhoek* **106**:43–56. doi:[10.1007/s10482-013-0084-1](https://doi.org/10.1007/s10482-013-0084-1).
- 403 5. **Brosius J, Palmer ML, Kennedy PJ, Noller HF.** 1978. Complete nucleotide sequence of a  
404 16S ribosomal RNA gene from *Escherichia coli*. *Proceedings of the National Academy of Sciences*  
405 **75**:4801–4805. doi:[10.1073/pnas.75.10.4801](https://doi.org/10.1073/pnas.75.10.4801).
- 406 6. **Schloss PD, Handelsman J.** 2004. Status of the microbial census. *Microbiology and Molecular*  
407 *Biology Reviews* **68**:686–691. doi:[10.1128/mubr.68.4.686-691.2004](https://doi.org/10.1128/mubr.68.4.686-691.2004).
- 408 7. **Dykhuizen DE.** 1998. Santa Rosalia revisited: Why are there so many species of bacteria?  
409 *Antonie van Leeuwenhoek* **73**:25–33. doi:[10.1023/a:1000665216662](https://doi.org/10.1023/a:1000665216662).
- 410 8. **Curtis TP, Sloan WT, Scannell JW.** 2002. Estimating prokaryotic diversity and its limits.  
411 *Proceedings of the National Academy of Sciences* **99**:10494–10499. doi:[10.1073/pnas.142680199](https://doi.org/10.1073/pnas.142680199).
- 412 9. **The Human Microbiome Consortium.** 2012. Structure, function and diversity of the healthy

413 human microbiome. *Nature* **486**:207–214. doi:[10.1038/nature11234](https://doi.org/10.1038/nature11234).

414 10. **Gilbert JA, Jansson JK, Knight R.** 2014. The Earth Microbiome Project: Successes and  
415 aspirations. *BMC Biology* **12**:69. doi:[10.1186/s12915-014-0069-1](https://doi.org/10.1186/s12915-014-0069-1).

416 11. **Amaral-Zettler L, Artigas LF, Baross J, P.A. LB, Boetius A, Chandramohan D, Herndl G,**  
417 **Kogure K, Neal P, Pedrós-Alió C, Ramette A, Schouten S, Stal L, Thessen A, Leeuw J de,**  
418 **Sogin M.** 2010. A global census of marine microbes, pp. 221–245. *In* *Life in the worlds oceans*.  
419 Wiley-Blackwell.

420 12. **Youssef N, Sheik CS, Krumholz LR, Najjar FZ, Roe BA, Elshahed MS.** 2009. Comparison  
421 of species richness estimates obtained using nearly complete fragments and simulated  
422 pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Applied*  
423 *and Environmental Microbiology* **75**:5227–5236. doi:[10.1128/aem.00592-09](https://doi.org/10.1128/aem.00592-09).

424 13. **Schloss PD.** 2010. The effects of alignment quality, distance calculation method, sequence  
425 filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol*  
426 **6**:e1000844. doi:[10.1371/journal.pcbi.1000844](https://doi.org/10.1371/journal.pcbi.1000844).

427 14. **Alivisatos AP, Blaser MJ, Brodie EL, Chun M, Dangl JL, Donohue TJ, Dorrestein PC,**  
428 **Gilbert JA, Green JL, Jansson JK, Knight R, Maxon ME, McFall-Ngai MJ, Miller JF, Pollard**  
429 **KS, Ruby EG, Taha SA.** 2015. A unified initiative to harness Earth's microbiomes. *Science*  
430 **350**:507–508. doi:[10.1126/science.aac8480](https://doi.org/10.1126/science.aac8480).

431 15. **Li E, Hamm CM, Gulati AS, Sartor RB, Chen H, Wu X, Zhang T, Rohlf FJ, Zhu W, Gu**  
432 **C, Robertson CE, Pace NR, Boedeker EC, Harpaz N, Yuan J, Weinstock GM, Sodergren E,**  
433 **Frank DN.** 2012. Inflammatory bowel diseases phenotype, *textitC. difficile* and NOD2 genotype  
434 are associated with shifts in human ileum associated microbial composition. *PLoS ONE* **7**:e26284.  
435 doi:[10.1371/journal.pone.0026284](https://doi.org/10.1371/journal.pone.0026284).

436 16. **Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, Nomicos E, Polley EC,**  
437 **Komarow HD, Murray PR, Turner ML, Segre JA.** 2012. Temporal shifts in the skin microbiome  
438 associated with disease flares and treatment in children with atopic dermatitis. *Genome Research*

439 **22:850–859.** doi:[10.1101/gr.131029.111](https://doi.org/10.1101/gr.131029.111).

440 **17. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Bouffard GG, Blakesley**  
441 **RW, Murray PR, Green ED, Turner ML, Segre JA.** 2009. Topographical and temporal diversity of  
442 the human skin microbiome. *Science* **324**:1190–1192. doi:[10.1126/science.1171700](https://doi.org/10.1126/science.1171700).

443 **18. Grice EA, Snitkin ES, Yockey LJ, Bermudez DM, Liechty KW, Segre JA, Mullikin J,**  
444 **Blakesley R, Young A, Chu G, Ramsahoye C, Lovett S, Han J, Legaspi R, Fuksenko T,**  
445 **Reddix-Dugue N, Sison C, Gregory M, Montemayor C, Gestole M, Hargrove A, Johnson**  
446 **T, Myrick J, Riebow N, Schmidt B, Novotny B, Gupti J, Benjamin B, Brooks S, Coleman H,**  
447 **Ho S-I, Schandler K, Smith L, Stantripop M, Maduro Q, Bouffard G, Dekhtyar M, Guan X,**  
448 **Masiello C, Maskeri B, McDowell J, Park M, Thomas PJ.** 2010. Longitudinal shift in diabetic  
449 wound microbiota correlates with prolonged skin defense response. *Proceedings of the National*  
450 *Academy of Sciences* **107**:14799–14804. doi:[10.1073/pnas.1004204107](https://doi.org/10.1073/pnas.1004204107).

451 **19. Harris JK, Caporaso JG, Walker JJ, Spear JR, Gold NJ, Robertson CE, Hugenholtz**  
452 **P, Goodrich J, McDonald D, Knights D, Marshall P, Tufo H, Knight R, Pace NR.** 2012.  
453 Phylogenetic stratigraphy in the guerrero negro hypersaline microbial mat. *The ISME Journal*  
454 **7**:50–60. doi:[10.1038/ismej.2012.79](https://doi.org/10.1038/ismej.2012.79).

455 **20. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ.**  
456 2006. Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of*  
457 *the National Academy of Sciences* **103**:12115–12120. doi:[10.1073/pnas.0605127103](https://doi.org/10.1073/pnas.0605127103).

458 **21. Whitman WB, Coleman DC, Wiebe WJ.** 1998. Prokaryotes: The unseen majority.  
459 *Proceedings of the National Academy of Sciences* **95**:6578–6583. doi:[10.1073/pnas.95.12.6578](https://doi.org/10.1073/pnas.95.12.6578).

460 **22. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A,**  
461 **Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T,**  
462 **Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T.** 2013.  
463 Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**:431–437.

464 doi:[10.1038/nature12352](https://doi.org/10.1038/nature12352).

465 23. **Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF.** 2011. EMIRGE: Reconstruction  
466 of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol*  
467 **12**:R44. doi:[10.1186/gb-2011-12-5-r44](https://doi.org/10.1186/gb-2011-12-5-r44).

468 24. **Miller CS, Handley KM, Wrighton KC, Frischkorn KR, Thomas BC, Banfield JF.** 2013.  
469 Short-read assembly of full-length 16S amplicons reveals bacterial diversity in subsurface sediments.  
470 *PLoS ONE* **8**:e56018. doi:[10.1371/journal.pone.0056018](https://doi.org/10.1371/journal.pone.0056018).

471 25. **Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang**  
472 **J, Woyke T, Huntemann M, Anderson I, Billis K, Varghese N, Mavromatis K, Pati A, Ivanova**  
473 **NN, Kyrpides NC.** 2013. IMG 4 version of the integrated microbial genomes comparative analysis  
474 system. *Nucleic Acids Research* **42**:D560–D567. doi:[10.1093/nar/gkt963](https://doi.org/10.1093/nar/gkt963).

475 26. **Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins**  
476 **MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF.** 2012. Fermentation,  
477 hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**:1661–1665.  
478 doi:[10.1126/science.1224041](https://doi.org/10.1126/science.1224041).

479 27. **Rocha UN da, Cadillo-Quiroz H, Karaoz U, Rajeev L, Klitgord N, Dunn S, Truong V,**  
480 **Buenrostro M, Bowen BP, Garcia-Pichel F, Mukhopadhyay A, Northen TR, Brodie EL.** 2015.  
481 Isolation of a significant fraction of non-phototroph diversity from a desert biological soil crust. *Front*  
482 *Microbiol* **6**:277. doi:[10.3389/fmicb.2015.00277](https://doi.org/10.3389/fmicb.2015.00277).

483 28. **Hamilton TL, Jones DS, Schaperdoth I, Macalady JL.** 2015. Metagenomic insights into  
484 S(0) precipitation in a terrestrial subsurface lithoautotrophic ecosystem. *Front Microbiol* **5**:756.  
485 doi:[10.3389/fmicb.2014.00756](https://doi.org/10.3389/fmicb.2014.00756).

486 29. **Handley KM, VerBerkmoes NC, Steefel CI, Williams KH, Sharon I, Miller CS, Frischkorn**  
487 **KR, Chourey K, Thomas BC, Shah MB, Long PE, Hettich RL, Banfield JF.** 2012. Biostimulation  
488 induces syntrophic interactions that impact c, s and n cycling in a sediment microbial community.

- 489 The ISME Journal **7**:800–816. doi:[10.1038/ismej.2012.148](https://doi.org/10.1038/ismej.2012.148).
- 490 **30. Gladden JM, Allgaier M, Miller CS, Hazen TC, VanderGheynst JS, Hugenholtz P,**  
491 **Simmons BA, Singer SW.** 2011. Glycoside hydrolase activities of thermophilic bacterial  
492 consortia adapted to switchgrass. *Applied and Environmental Microbiology* **77**:5804–5812.  
493 doi:[10.1128/aem.00032-11](https://doi.org/10.1128/aem.00032-11).
- 494 **31. Brooks B, Firek BA, Miller CS, Sharon I, Thomas BC, Baker R, Morowitz MJ, Banfield JF.**  
495 2014. Microbes in the neonatal intensive care unit resemble those found in the gut of premature  
496 infants. *Microbiome* **2**:1. doi:[10.1186/2049-2618-2-1](https://doi.org/10.1186/2049-2618-2-1).
- 497 **32. Wilkins MJ, Wrighton KC, Nicora CD, Williams KH, McCue LA, Handley KM, Miller CS,**  
498 **Giloteaux L, Montgomery AP, Lovley DR, Banfield JF, Long PE, Lipton MS.** 2013. Fluctuations  
499 in species-level protein expression occur during element and nutrient cycling in the subsurface.  
500 *PLoS ONE* **8**:e57819. doi:[10.1371/journal.pone.0057819](https://doi.org/10.1371/journal.pone.0057819).
- 501 **33. Handley KM, Wrighton KC, Miller CS, Wilkins MJ, Kantor RS, Thomas BC, Williams KH,**  
502 **Gilbert JA, Long PE, Banfield JF.** 2014. Disturbed subsurface microbial communities follow  
503 equivalent trajectories despite different structural starting points. *Environ Microbiol* **17**:622–636.  
504 doi:[10.1111/1462-2920.12467](https://doi.org/10.1111/1462-2920.12467).
- 505 **34. Alessi DS, Lezama-Pacheco JS, Janot N, Suvorova EI, Cerrato JM, Giammar DE, Davis**  
506 **JA, Fox PM, Williams KH, Long PE, Handley KM, Bernier-Latmani R, Bargar JR.** 2014.  
507 Speciation and reactivity of uranium products formed during in situ bioremediation in a shallow  
508 alluvial aquifer. *Environmental Science & Technology* **48**:12842–12850. doi:[10.1021/es502701u](https://doi.org/10.1021/es502701u).
- 509 **35. Locey KJ, Lennon JT.** 2015. Scaling laws predict global microbial diversity. *PeerJ PrePrints*.  
510 doi:[10.7287/peerj.preprints.1451v1](https://doi.org/10.7287/peerj.preprints.1451v1).
- 511 **36. Parada AE, Needham DM, Fuhrman JA.** 2015. Every base matters: Assessing small subunit  
512 rRNA primers for marine microbiomes with mock communities, time series and global field samples.  
513 *Environ Microbiol* n/a–n/a. doi:[10.1111/1462-2920.13023](https://doi.org/10.1111/1462-2920.13023).
- 514 **37. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton**

- 515 **KC, Williams KH, Banfield JF.** 2015. Unusual biology across a group comprising more than 15%  
516 of domain bacteria. *Nature* **523**:208–211. doi:[10.1038/nature14486](https://doi.org/10.1038/nature14486).
- 517 **38. Eloë-Fadrosh EA, Ivanova NN, Woyke T, Kyrpides NC.** 2016. Metagenomics  
518 uncovers gaps in amplicon-based detection of microbial diversity. *Nat Microbiol* 15032.  
519 doi:[10.1038/nmicrobiol.2015.32](https://doi.org/10.1038/nmicrobiol.2015.32).
- 520 **39. Highlander; PDSSLWMLJSK.** 2015. Sequencing 16S rRNA gene fragments using the pacBio  
521 SMRT DNA sequencing system. *PeerJ PrePrints*. doi:[10.7287/peerj.preprints.778v1](https://doi.org/10.7287/peerj.preprints.778v1).
- 522 **40. Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A, Kanigan T, Lewis**  
523 **K, Epstein SS.** 2010. Use of iChip for high-throughput in situ cultivation of “uncultivable” microbial  
524 species. *Applied and Environmental Microbiology* **76**:2445–2450. doi:[10.1128/aem.01754-09](https://doi.org/10.1128/aem.01754-09).
- 525 **41. Buerger S, Spoering A, Gavrish E, Leslin C, Ling L, Epstein SS.** 2012. Microbial  
526 scout hypothesis, stochastic exit from dormancy, and the nature of slow growers. *Applied and*  
527 *Environmental Microbiology* **78**:3221–3228. doi:[10.1128/aem.07307-11](https://doi.org/10.1128/aem.07307-11).
- 528 **42. Das N, Tripathi N, Basu S, Bose C, Maitra S, Khurana S.** 2015. Progress in the  
529 development of gelling agents for improved culturability of microorganisms. *Front Microbiol* **6**:698.  
530 doi:[10.3389/fmicb.2015.00698](https://doi.org/10.3389/fmicb.2015.00698).
- 531 **43. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO.** 2007. SILVA:  
532 A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data  
533 compatible with ARB. *Nucleic Acids Research* **35**:7188–7196. doi:[10.1093/nar/gkm864](https://doi.org/10.1093/nar/gkm864).
- 534 **44. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA,**  
535 **Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Horn DJV, Weber CF.**  
536 2009. Introducing mothur: Open-source, platform-independent, community-supported software  
537 for describing and comparing microbial communities. *Applied and Environmental Microbiology*  
538 **75**:7537–7541. doi:[10.1128/aem.01541-09](https://doi.org/10.1128/aem.01541-09).
- 539 **45. Westcott SL, Schloss PD.** 2015. De novo clustering methods outperform reference-based  
540 methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3**:e1487.

541 [doi:10.7717/peerj.1487](https://doi.org/10.7717/peerj.1487).

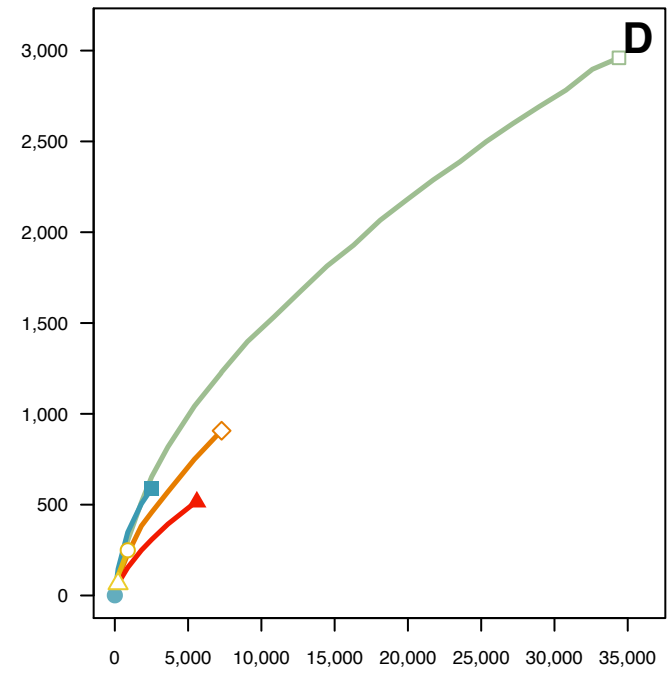
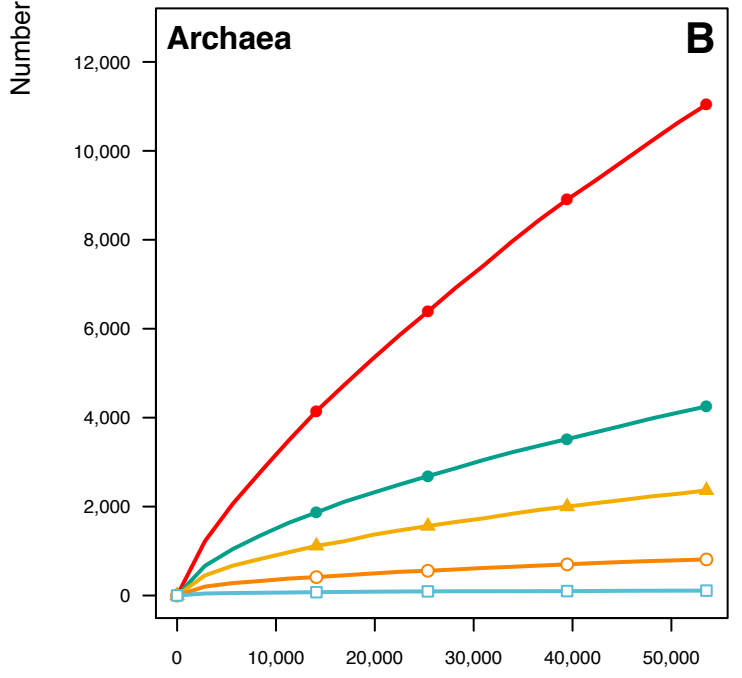
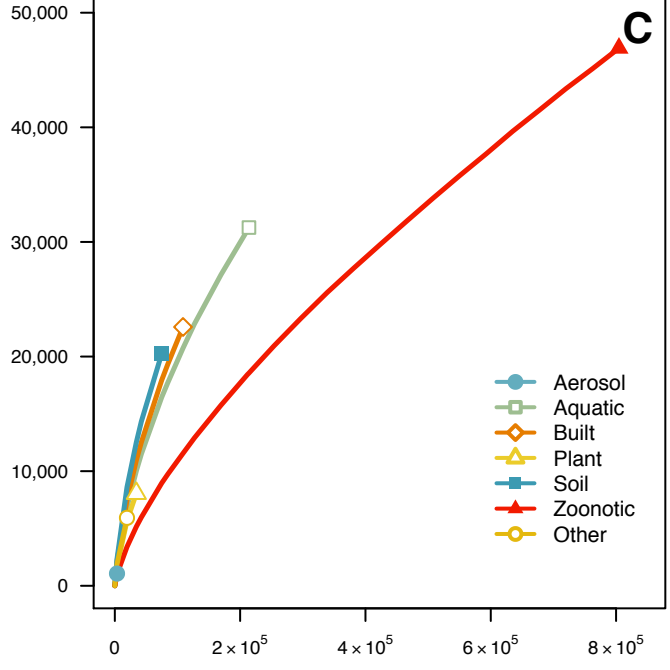
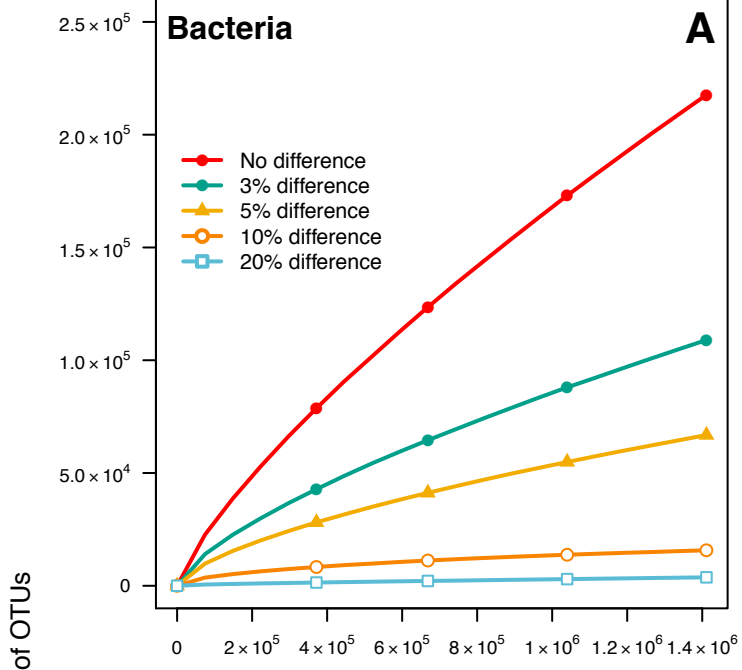
542 46. **Good IJ**. 1953. The population frequencies of species and the estimation of population  
543 parameters. *Biometrika* **40**:237–264. doi:[10.1093/biomet/40.3-4.237](https://doi.org/10.1093/biomet/40.3-4.237).

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

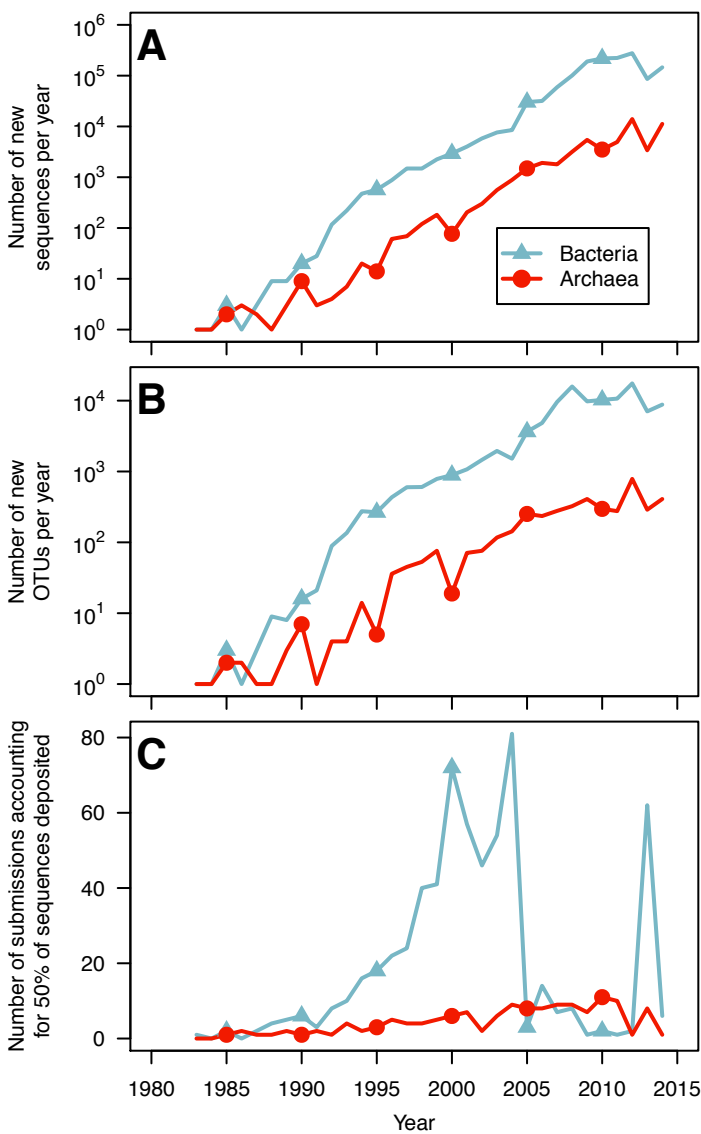


**Table 1. Status of microbial census by habitat classifications and domain.** The isolation\_source field from the SILVA reference database was manually curated to assign bacterial and archaeal sequences coarse and fine scale habitat classifications. We calculated the number of sequences and OTUs observed and the percent coverage on a sequence or OTU basis for each classification and domain. Descriptions of each category are provided in Table S1.

Coarse	Fine	Bacteria				Archaea			
		Sequences (N)	OTUs (N)	% Seq. Coverage	% OTU Coverage	Sequences (N)	OTUs (N)	% Seq. Coverage	% OTU Coverage
Aerosol		3,472	1,068	79.5	33.2	2	1	100.0	100.0
Aquatic	Brackish	1,094	646	54.6	23.1	1,368	314	87.4	44.9
	Brackish sediment	390	243	54.4	26.7	525	208	76.8	41.3
	Freshwater	21,647	6,689	80.8	37.7	1,540	439	84.7	46.5
	Freshwater sediment	6,733	3,549	63.0	29.8	1,324	488	79.3	43.9
	Marine	134,727	14,287	94.3	46.7	10,983	830	95.8	44.5
	Marine sediment	27,801	9,567	79.6	40.8	14,049	1,507	95.0	53.7
	Hydrothermal vent	10,860	4,216	75.4	36.5	3,797	734	90.4	50.3
	Ice	2,073	936	71.2	36.1	42	5	95.2	60.0
	Other	8,760	3,802	71.7	34.8	772	313	80.7	52.4
Built	Digesters	33,152	8,949	82.9	36.8	4,764	483	93.6	36.4
	Food-associated	11,813	1,632	92.0	41.9	117	40	80.3	42.5
	Industrial/mining	16,582	6,099	76.6	36.3	1,245	336	84.4	42.3
	Pollution associated	38,696	10,602	84.1	41.9	716	249	79.2	40.2
	Other	8,556	2,730	79.1	34.7	444	111	90.8	63.1
Plant associated	Root	19,695	5,052	84.3	38.7	200	61	85.5	52.5
	Surface	4,892	1,385	82.7	38.8	0	0	NA	NA
	Other	9,753	3,217	78.8	35.8	22	7	90.9	71.4
Soil	Agriculture	10,051	4,017	73.6	34.0	146	56	80.8	50.0
	Desert	3,042	1,280	73.7	37.5	245	79	77.6	30.4
	Permafrost	1,922	870	73.0	40.3	39	20	64.1	30.0
	Other	59,855	17,166	82.9	40.4	2,087	516	89.1	55.8
Zoological	Vertebrate	773,045	42,497	96.1	29.5	5,389	454	95.1	41.6
	Arthropod	13,209	3,688	81.8	34.7	87	52	58.6	30.8
	Other invertebrate	7,476	2,626	78.0	37.3	67	30	73.1	40.0
	Other	10,855	1,754	89.2	33.4	54	17	87.0	58.8
Other		19,414	5,930	81.6	39.9	882	249	84.2	44.2
No source data		151,669	14,144	94.9	45.6	2,565	559	88.6	47.6
Total		1,411,234	108,950	94.5	29.2	53,546	4,252	95.1	38.5



Number of Sequences

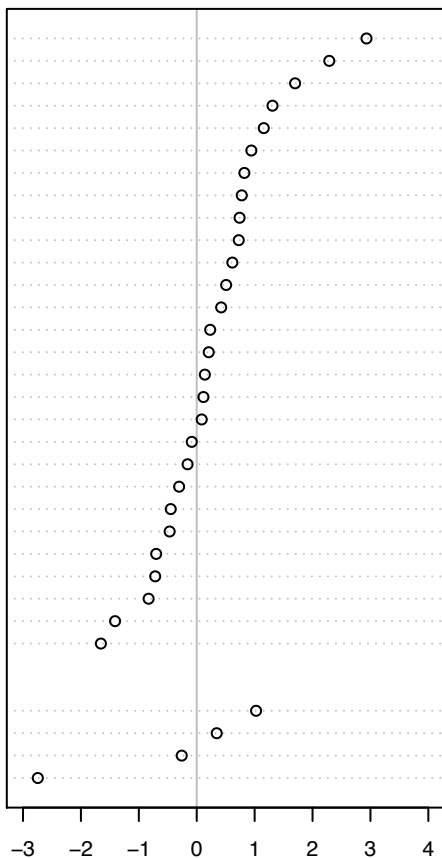


## Bacteria

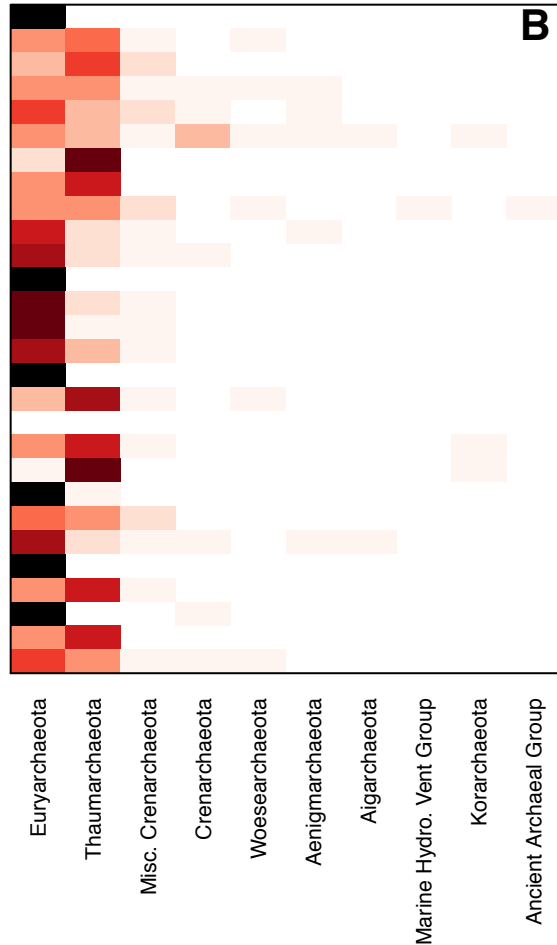
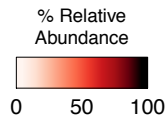
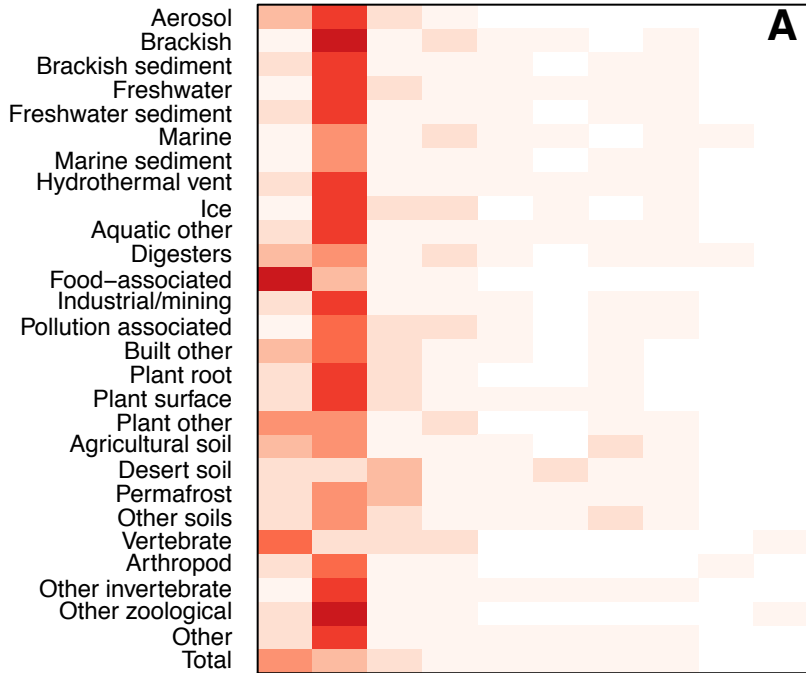
Kazan-3B-09  
Atribacteria  
Aminicenantes  
Gracilibacteria  
Lentisphaerae  
Deferribacteres  
Fusobacteria  
Marinimicrobia  
Chloroflexi  
Actinobacteria  
Saccharibacteria  
Fibrobacteres  
Planctomycetes  
Firmicutes  
Verrucomicrobia  
Gemmatimonadetes  
Synergistetes  
Armatimonadetes  
Acidobacteria  
Cyanobacteria  
Proteobacteria  
Bacteroidetes  
Parcubacteria  
Nitrospirae  
Chlorobi  
Spirochaetae  
Deinococcus-Thermus  
Tenericutes

## Archaea

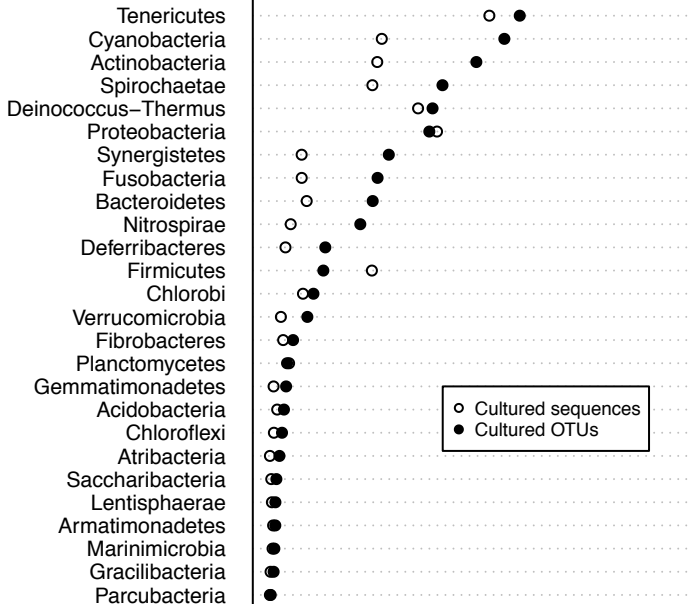
Thaumarchaeota  
Misc. Crenarchaeota  
Euryarchaeota  
Crenarchaeota



Fold difference in sequencing effort before and after 2006 (log2 scale)



## Bacteria



## Archaea

