bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Phylogenetically novel uncultured microbial cells dominate Earth microbiomes

Authors: Karen G. Lloyd[1*], Joshua Ladau[2], Andrew D. Steen[3], Junqi Yin[4], Lonnie Crosby[4]

Affiliations:

[1]Department of Microbiology, University of Tennessee, Knoxville, TN 37996
[3]Gladstone Institutes, University of California San Francisco, San Francisco, CA 94158
[2]Department of Earth and Planetary Sciences, University of Tennessee, Knoxville TN 37996
[4]Joint Institute for Computational Sciences, University of Tennessee, Knoxville, TN 37996

*Correspondence to: klloyd@utk.edu.

Running title: Quantifying the abundance of uncultured microbes

Abstract word count: 250

Text word count: 3,312

Importance 94 words. In the past few decades, it has become apparent that most of the microbial diversity on Earth has never been characterized in laboratory cultures. We show that these unknown microbes, sometimes called "microbial dark matter", are numerically dominant in all major environments on Earth, with the exception of the human body, where most of the microbes have been cultured. We also show that about a quarter of microbial cells on Earth belong to phyla with no cultured relatives, suggesting that, if we can discover their novel functions, they might be important to ecosystem functions.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

To unequivocally determine a microbe's physiology, including its metabolism, environmental roles, and growth characteristics, it must be grown in a laboratory culture. Unfortunately, many phylogenetically-novel groups have never been cultured, so their physiologies have only been inferred from genomics and environmental characteristics. Although the diversity, or number of different taxonomic groups, of uncultured clades has been well-studied, their global abundances, or number of cells in any given environment, have not been assessed. We quantified the degree of similarity of 16S rRNA gene sequences from diverse environments in publicly-available metagenome and metatranscriptome databases, which we show are largely free of the culture-bias present in primer-amplified 16S rRNA gene surveys, to their nearest cultured relatives. Whether normalized to scaffold read depths or not, the highest abundance of metagenomic 16S rRNA gene sequences belong to phylogenetically novel uncultured groups in seawater, freshwater, terrestrial subsurface, soil, hypersaline environments, marine sediment, hot springs, hydrothermal vents, non-human hosts, snow and bioreactors (22-87% uncultured genera to classes and 0-64% uncultured phyla). The exceptions were human and human-associated environments which were dominated by cultured genera (45-97%). We estimate that uncultured genera and phyla could comprise 7.3×10^{29} (81%) and 2.2×10^{29} (25%) microbial cells, respectively. Uncultured phyla were over-represented in metatranscriptomes relative to metagenomes (46-84% of sequences in a given

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

55 environment), suggesting that they are viable, and possibly more active than cultured
56 clades. Therefore, uncultured microbes, often from deeply phylogenetically divergent
57 groups, dominate non-human environments on Earth, and their undiscovered
58 physiologies may matter for Earth systems.
59
60 **Introduction**
61 Direct sequencing of environmental DNA has shown that most microbial lineages
62 have not been isolated in pure culture (1–3). However, cellular abundances and viability
63 states of uncultured microbes at different levels of phylogenetic divergence from their
64 closest cultured relative are unknown. Cellular abundance and viability may, in some
65 cases, signify importance to current ecosystem functions, as opposed to members of the
66 rare biosphere which become important for ecosystem functioning when conditions
67 change (4). With the exception of keystone species which can have great ecosystem
68 importance even at low biomass concentrations, prokaryotic abundance and viability is
69 generally an indicator for participation in current ecosystem functions (1).
70 Quantifying the cellular abundance of all microbial taxa in any sample is
71 challenging. Fluorescent in situ hybridization (FISH) allows fluorescent tagging of a
72 taxonomic group whose cells can then be counted under a microscope (2). However,
73 FISH requires developing probes for phylogenetic groups one-by-one, which is
74 impractical for quantifying highly diverse natural samples that are often comprised of
75 thousands of species (3). Furthermore, FISH techniques are not always quantitative in all
76 environments, due to taxon-specific biases in probe efficacy (4, 5). Quantitative PCR has
77 the same low-throughput limitations, since individual measurements must be made for

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

each taxon, and primer bias makes them not absolutely quantitative (4). However, understanding the total cellular abundance of uncultured clades of archaea and bacteria in all environments on Earth is an essential question in microbiology, so we approximated it using the data available in public databases.

Genes encoding the 16S rRNA small subunit of the ribosome are the most commonly-used taxonomic and phylogenetic identifier for bacteria and archaea, and most scientific journals make publication contingent on the posting of 16S rRNA gene sequences to public databases. Therefore, the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov) houses a nearly-complete database of full length 16S rRNA gene sequences. This database is subject to biases since the gene entries have undergone exponential amplification from their initial abundances, and small mismatches between DNA primers and different taxa are magnified during this amplification (5). However, we examined it here since it incorporates microbial phylogenetic information from thousands of different research studies. Assembled metagenomes provide a less-biased accounting of 16S rRNA genes from a given environment. Here all DNA is chemically extracted from a sample, purified, sequenced in a small-read high throughput platform, and then bioinformatically assembled to contigs. Full length 16S rRNA genes can be identified in these contigs using hidden Markov model-based programs like RNAmmer (6). If the sequencing depth is great enough, quantifying read recruitment to each 16S rRNA gene provides the closest approximate quantification of individual 16S rRNA genes currently available.

Cellular activity, however, is as important to environmental functions as cellular abundance (1). In cultured cells, rRNA content correlates with cellular activity (18),

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

101 although no universally predictive relationship between those two parameters has been
102 identified (19). Metatranscriptomes, in which 16S rRNA transcripts are converted to
103 cDNA and sequenced without the use of primers, provide an estimate of which cells
104 contained ribosomes, and therefore were at least poised for activity in the environment
105 (19).

106 We determined the percent similarity of nearly all 16S rRNA gene sequences
107 from public databases, to get a first estimate of the global abundance of microbial clades
108 at different levels of similarity to their nearest cultured relative in different environments.
109 The metagenomic and metatranscriptomic datasets show that uncultured clades dominate
110 cellular abundance of non-human Earth environments. Knowing the global abundance of
111 cells from uncultured taxa is crucial for estimating the importance of uncultured lineages
112 to ecosystem functions, determining the appropriateness of using cultured microbes as
113 model systems for natural environments, and predicting the causes of unculturability.
114

115 **Materials and Methods**

116 Primer-amplified sequences were obtained from www.arb-silva.de Silva123Ref
117 (5), which contains chimera-checked, high quality, >900 bp (for archaea) and >1200 bp
118 (for bacteria) 16S rRNA gene sequences, almost all of which were Sanger-sequenced
119 clone inserts from primer-amplified PCR products. This yielded 952,509 bacterial and
120 51,608 archaeal sequences from 4,743 studies that employed a wide variety of primers.
121 Genes annotated as 16S rRNA and >900bp were collected from the Joint Genome
122 Institute IMG/M for metagenomes larger than 1 GB total or metatranscriptomes larger
123 than 60 Mb total (6). Too few metatranscriptomes were available from humans, human-

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

adjacent environments, rock, snow, hydrothermal vents, hypersaline environments, or marine sediments to be included. Scaffold read depths were available for metagenomes, but not metatranscriptomes.

Metagenomes and metatranscriptomes are prone to chimera production during assemblies of short reads along the highly conserved 16S rRNA gene (7). We therefore implemented uChime (8) in mothur (9) with the Silva Gold alignment to identify and remove a further 1.3% and 0.6% of possible chimeras from metagenomes and metatrancriptomes, respectively. Further chimera checks are described below. Taxonomic identifications were made for each sequence in the metagenomic and metatranscriptomic datasets in mothur (9) for alignment, pre-clustering, and classification to silva.nr_v132 (10). Sequences identifying as chloroplasts, mitochondria, or eukaryotes (<1% of sequences) were removed.

BLASTn was used to determine the percent identity of each sequence to its single most-closely related 16S rRNA gene sequence from cultured archaea (4,170 sequences) or bacteria (22,150 sequences) obtained from Arb-Silva. Only cultured archaea and bacteria with official names from the International Journal of Systematic Bacteriology or the International Journal of Systematic and Evolutionary Microbiology were included, excluding candidatus organisms or enrichments. Rather than relying on annotations to separate archaea and bacteria in metagenomes and metatrancriptomes, sequences were queried against a database with bacteria and archaea combined to get the top hit. We used a BLASTn implementation parallelized for high performance computation, HPC-BLAST (11), on the Beacon cluster (12) at the Joint Institute for Computational Sciences. The alignment results of HPC-BLAST are compatible with those of NCBI BLAST.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A few metagenomic and metatranscriptomic 16S sequences did not yield BLASTn hits, so were not considered further. For sequences with query alignment lengths <300 bp, percent identity increased with decreasing alignment length, suggesting that these were partial hits to small conserved regions, so they were removed from the analysis. Short query alignment lengths could also signify chimeras. Therefore sequences with < 90% alignment length to their closest cultured relative were aligned with BLASTn to the SilvaNR database, containing environmental DNA sequences. Sequences with <90% alignment to both the cultured and Silva NR databases were considered to be chimeric and were removed from analysis. This removed 6% of the metagenomic database, leaving 39,426 bacterial and 13,404 archaeal sequences from 1,504 metagenomes, as well as 7% of the metatranscriptomic database, leaving 9,396 bacterial and 3,863 archaeal sequences from 381 metatranscriptomes. Each remaining sequence was manually categorized into one of 14 environment types, based on user-provided metadata (Tables S1 and S2).

16S rRNA gene sequences that shared more than 96.6% sequence identity with a cultured organism were considered to be in the same genus, and sequences that shared at least 86% sequence similarity were considered to be in the same phylum (13). These create “similarity bins” of cultured species to genus, uncultured genus to class, and uncultured phyla and higher. For primer-amplified, metagenomic, and metatranscriptomic datasets, the fraction of sequences in each similarity bin was calculated for a given environment. In metagenomes for which sequence read depth was available, the fraction in each similarity bin was calculated as the sum of sequence read

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

169 depths for each similarity bin within each metagenome. These values were averaged for
170 all metagenomes in each environment.
171
172 **Results and Discussion**
173 More than a third of primer-amplified 16S rRNA gene sequences were from the
174 same species or genus as a culture (37% for bacteria and 34% for archaea, Fig. 1), in
175 agreement with previous findings that primer-amplified databases skew toward cultured
176 organisms (14–16). However, even in the primer-amplified dataset, the majority of
177 sequences were from uncultured genera or higher taxonomic groups, including 17% and
178 44% from uncultured phyla in bacteria and archaea, respectively. This suggests that, as a
179 group, uncultured microbes, including those that are very highly divergent, are fairly
180 abundant when all full length 16S rRNA genes in public databases are considered.
181 Metagenomes had lower fractions of 16S rRNA gene sequences from cultured species
182 (Fig. 1), with 15% for both bacteria and archaea based on total sequences and 28% for
183 bacteria and 31% for archaea based on scaffold read depths. The rest of the 16S rRNA
184 gene sequences were from uncultured genera and higher taxonomic groups, with about a
185 third of total sequences from uncultured phyla (36% and 46% without read depths, 24
186 and 33% with read depths for bacteria and archaea). However, we recognize that it is
187 impossible to absolutely link 16S identity to taxonomic level, since phylogenetic
188 difference is inconsistently related to 16S sequence difference across lineages (13).
189 Therefore, these sequence similarity cutoffs are proxies for degrees of phylogenic novelty
190 rather than rigidly-defined taxonomic levels. By using published values for similarity
191 bins (13), our findings are comparable to other studies and serve as an estimate for

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

192 phylogenetic novelty informed by the available data. Therefore, 16S rRNA gene
193 sequences from uncultured cells were more abundant than those from cultured cells,
194 suggesting that uncultured microbial clades are not collectively relegated to the "rare
195 biosphere" (17), but are instead the most numerically dominant cells in the public
196 databases.

197 We found that highly divergent uncultured sequences were better represented in
198 metatranscriptomes than in metagenomes, with only 4% (bacteria) and 5% (archaea) of
199 total sequences from cultured species to genera, and 65% (bacteria) and 71% (archaea) of
200 total sequences from uncultured phyla. Therefore, cells from highly divergent uncultured
201 groups were alive *in situ*, and may even be more active in natural samples than cells from
202 cultured species and genera. A comparison between metagenomes and
203 metatranscriptomes both derived from the same samples in the Gulf of Mexico showed
204 that uncultured clades were indeed active relative to cultured clades (20).

205 Contributions from uncultured clades varied by environment (Fig. 2). The only
206 environments dominated by sequences from cultured species and genera were the human
207 body and human-adjacent environments (Fig. 2). This result was not due to primer bias,
208 since primer-amplified and metagenomic datasets contained mostly cultured species and
209 genera (45-97%, inclusive of bacteria and archaea). High culturability in human
210 environments likely benefits from a high frequency of culturing efforts, since all
211 culturing happens in the vicinity of humans, and since the study of human diseases has
212 driven much research (21). Uncultured clades were also present in humans and human-
213 adjacent environments, but very few were uncultured at a taxonomic cut-off above family
214 level.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Primer bias toward cultures was more severe in all other environments, where uncultured archaea and bacteria were much more abundant in metagenomic datasets than in primer-amplified datasets (Fig. 2). The exception was archaea in marine sediments, possibly indicating that commonly-used primers have good matches to the uncultured phyla that are abundant in these environments (22). To avoid primer bias and account for a high environmental abundance of closely related sequences, we used the metagenomic datasets with read depths to estimate quantifications (Fig. 3). Hypersaline environments were the next best-cultured environments after human environments, with nearly half of archaea and bacteria from cultured genera, and very few from uncultured phyla (Fig. 3). The next best-cultured group was archaea in bioreactors. All other environments had more sequences from uncultured phyla than from cultured genera. Hot springs and hydrothermal vents, in particular, had high frequencies of uncultured phyla in both bacteria and archaea. Even though human host environments were dominated by cultured groups, non-human hosts had as few sequences from cultured archaea and bacteria as did soil, seawater, freshwater, marine sediment, terrestrial subsurface, snow and bioreactors (for bacteria). This suggests that highly divergent uncultured microbes, possibly with novel functions, dominate non-human environments on Earth.

By using a large collection of publicly-available sequences that are as complete a sampling as possible, our sequence abundance quantifications can be extrapolated to global cell estimates, although this approach is biased against cells that are less amenable to DNA extraction and under-sampled environments. Copy numbers of 16S rRNA genes per cell can only be determined for completed genomes (means of 3.8 copies/genome for 1657 bacteria, and 1.8 copies/genome for 79 archaeal genomes on IMG

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

238 https://img.jgi.doe.gov/mer/, March 30, 2018). However, no complete genomes are yet
239 available for uncultured organisms. Applying the 16S rRNA copy numbers for completed
240 genomes to our estimations of total cells would increase our estimates of the abundance
241 of uncultured organisms, since archaea, which we found to be less well-cultured, would
242 be divided by the smaller number. Therefore, we use the conservative simplification of a
243 single 16S rRNA copy number per genome to estimate that 81% of microbial cells on
244 Earth are from uncultured genera or higher (7.3×10^{29} cells) and 25% are from
245 uncultured phyla (2.2×10^{29} cells) (Table 1). When abundance data are derived from
246 metatranscriptomes, uncultured cells increase to 98% (5.9×10^{29}), with uncultured phyla
247 contributing 69% (4.2×10^{29}) (Table 2). If the terrestrial subsurface datasets lack
248 contributions from the ultra-small uncultured cells missed in standard filtering methods
249 (23), or if DNA extraction favors cultured taxa, which may have more easily-lysed cell
250 membranes, then these values are underestimates for the abundance of uncultured cells
251 on Earth.

252 We tested whether only a few clades account for this global dominance of
253 uncultured microbes, suggesting that problem of culturability could be solved by just
254 getting a few important species into culture. On the contrary, each category of
255 phylogenetic novelty contained many different genera (Fig. 4). Also, genera at all levels
256 of phylogenetic novelty were distributed throughout the rank abundance curves in all
257 environments except for humans (Fig. 4). The taxonomic identities of the ten most
258 abundant genera differed between environments, and often included genera from newly-
259 named uncultured phyla such as Parcubacteria, Omnitrophica, Latescibacteria,
260 Patescibacteria, Bathyarchaeota, Woesearchaeota, Armatimonadetes, AC1,

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

261 Miscellaneous Euryarchaeotal Group, Saccharibacteria, WS6, Marinimicrobia, and FBP
262 (Fig.4). Despite having fewer overall sequences than bacteria, archaea were in the ten
263 most abundant genera in 8 of the 12 environments. Few of the top ten genera in
264 metagenomes were also in the top ten genera in metatranscriptomes. The exception was
265 Chloroflexi_Anaerolineaceae, which was present in the top ten genera in both datasets for
266 hot springs, terrestrial subsurface, and bioreactors. However, this could be an artifact,
267 since uncultured members of this group have not been taxonomically characterized to the
268 genus level, so these bins may lump together many different genera collectively labeled
269 as "uncultured". Some of the most abundant uncultured clades, such as *Candidatus*
270 *Pelagibacter sp.* in seawater, have actually been obtained in pure culture (24), but their
271 physiological requirements prevent them from meeting stringent requirements, such as
272 the ability to be grown out of stocks of cells preserved in glycerol at -80°C, required to
273 receive an official taxonomy. However, few other examples of such cryptically cultured
274 organisms occur in our dataset.

275 Many of the top ten genera were taxonomically identified as belonging to cultured
276 phyla, even though we found them to be <86% similar to their nearest cultured neighbor.
277 This is because taxonomic identification and phylogenetic identification are not identical
278 methods. Sequences that have low similarity to cultures can nonetheless be given a
279 taxonomic classification to a cultured phylum because the database used for classification
280 also contains many instances of uncultured sequences that have previously been named as
281 part of that phylum. When genomes become available, such groups are often re-assigned
282 as phyla (1). Our results suggest that rare and abundant taxa are both cultured and
283 uncultured, as well as bacterial and archaeal.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Our datasets likely include some amount of relic preserved DNA that can inflate diversity estimates (25). However, we do not calculate total diversity in a single sample, but rather occurrence frequency across many samples. Extracellular DNA from a particular taxonomic group is not likely to be abundant in the majority of samples, to the exclusion of intracellular DNA from that taxonomic group. In addition, in all environments, metatranscriptomes were characterized by higher fractions of sequences from uncultured groups than the metagenomic databases were, with particularly high contributions from uncultured phyla (Fig. 2). This suggests that the uncultured cells that dominate these datasets likely come from living organisms.

These results offer at least a partial explanation for "the great plate count anomaly", which states that <1% of environmental microbial cells are culturable with standard methods (26). To update this analysis, we examined 347 experiments in 26 studies from lakes, rivers, drinking water, seawater, marine and terrestrial subsurface, animal hosts, and soils, and found a median of 0.5% culturable cells (Table S3). The past several decades have seen considerable progress on novel culturing techniques, which have yielded higher fractions of culturable cells (25 ± 20%, n = 38) in fish guts (27), rice paddies (28), surface marine sediments (29, 30), agricultural soils (31), and eutrophic lakes (26). However, these studies expanded the set of cultured taxa only to novel families (29, 31), and we show that the percentages of cells from cultured families in these environments match these studies' percentages of culturable cells (Table S4). Therefore, we propose that these innovative methods likely were successful at culturing viable but non-culturable cells (VBNC), which are cells from previously-cultured cultured clades that are temporarily and reversibly culture-resistant (32). However, our

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

307 analysis shows that a considerable fraction of cells in non-human environments are
308 phylogenetically divergent, even belonging to novel phyla. We propose that
309 representatives of these phyla resist cultivation due to more fundamental reasons, making
310 them phylogenetically divergent non-cultured cells (PDNC). We roughly define PDNC as
311 cells from orders or higher with no cultured representatives. Unlike VBNC, PDNC are
312 not dormant close relatives of cultured species that can be expected to behave like known
313 cultures if given the correct combination of growth conditions. These entire lineages may
314 have physiologies that prevent growth in pure culture, such as dependences on syntrophic
315 interactions (33), precise chemical or physical parameters that are difficult to maintain
316 (29), extreme dependence on oligography (24, 34, 35), or very slow growth rates (36).
317 Two of the most recent uncultured phyla that have been brought into pure culture are
318 Nitrosopumilus sp., from the Thaumarchaeota phylum (34), and *Abditibacterium*
319 *utsteinense*, from the FBP phylum (37). These required extremely low nutrient
320 environments, incubation times of many months, and, in the case of *A. utsteine*nse, high
321 antibiotic levels to keep out competitors. Fundamentally novel culturing techniques,
322 possibly guided by insights on cell physiology derived from genomic studies, are likely
323 required to get more of these highly abundant and deeply-divergent clades in culture.
324 Given the substantial functional differences that often exist between closely-
325 related microbial species or strains, these uncultured lineages are likely to contain many
326 novel metabolic pathways, enzyme functions, cellular structures, physiologies (38). For
327 instance, uncultured clades of archaea and bacteria have more genes that are un-
328 annotatable with current databases than cultured clades (27 and 37%, vs. 19 and 31%,
329 respectively, Fig. S1). In addition, a rapidly growing number of studies are uncovering

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

potentially important functions of uncultured clades within specific environmental contexts (20, 39–41).

We conclude that uncultured taxa, often at very high levels of phylogenetic novelty, are abundant and alive in Earth's microbiome, and may harbor undiscovered functions that are important on an ecosystem level. The high proportion of sequences from uncultured groups in human-maintained bioreactors, animal and plant hosts, and soils, many of which were agricultural or municipal, shows that highly divergent novel clades are not only a feature of pristine wilderness environments, but are important in engineered environments with immediate human applications as well. This suggests that *ex situ* experiments on existing microbial cultures may not represent the functions of the majority of cells *in situ*. For environmentally important VBNC cells, novel culture techniques are showing great success in getting them into culture (35, 42). For PDNC, novel culture-independent techniques such as genomic inference (43), label incorporation (44–46), or tracking slow growth in a mixed population under different conditions (47), will allow the study of their physiology and ecology, and guide efforts to culture them.

References

1. **Rinke C**, **Schwientek P**, **Sczyrba A**, **Ivanova NN**, **Anderson IJ**, **Cheng J-F**, **Darling A**, **Malfatti S**, **Swan BK**, **Gies E a**, **Dodsworth J a**, **Hedlund BP**, **Tsiamis G**, **Sievert SM**, **Liu W-T**, **Eisen J a**, **Hallam SJ**, **Kyrpides NC**, **Stepanauskas R**, **Rubin EM**, **Hugenholtz P**, **Woyke T**. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature **499**:431–7.

2. **Huber JA**, **Mark Welch DB**, **Morrison HG**, **Huse SM**, **Neal PR**, **Butterfield DA**, **Sogin ML**. 2007. Microbial population structures in the deep marine biosphere. Science (80-) **318**:97–100.

3. **Parks DH**, **Rinke C**, **Chuvochina M**, **Chaumeil P**, **Woodcroft BJ**, **Evans PN**, **Hugenholtz P**, **Tyson GW**. 2017. Recovery of nearly 8,000 metagenome-

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

assembled genomes substantially expands the tree of life. Nat Microbiol **903**:1–10.

4. **Wang Y**, **Hatt JK**, **Tsementzi D**, **Rodriguez-R LM**, **Ruiz-Perez CA**, **Weigand MR**, **Kizer H**, **Maresca G**, **Krishnan R**, **Poretsky R**, **Spain JC**, **Konstantinidis KT**. 2017. Quantifying the importance of the rare biosphere for microbial community response to organic pollutants in a freshwater ecosystem. Appl Environ Microbiol **83**:1–19.

5. **Pruesse E**, **Quast C**, **Knittel K**, **Fuchs BM**, **Glo FO**, **Ludwig W**. 2007. SILVA : a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. October **35**:7188–7196.

6. **Chen I-M a**, **Markowitz VM**, **Chu K**, **Palaniappan K**, **Szeto E**, **Pillay M**, **Ratner A**, **Huang J**, **Andersen E**, **Huntemann M**, **Varghese N**, **Hadjithomas M**, **Tennessen K**, **Nielsen T**, **Ivanova N**, **Kyrpides NC**. 2017. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res **45**:507–516.

7. **Yuan C**, **Lei J**, **Cole J**, **Sun Y**. 2015. Reconstructing 16S rRNA genes in metagenomic data. Bioinformatics **31**:i35–i43.

8. **Edgar RC**. 2016. UCHIME2: improved chimera prediction for amplicon sequencing. bioRxiv 74252.

9. **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 DJ Van**, **Weber CF**. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol **75**:7537–7541.

10. **Yilmaz P**, **Parkrey LW**, **Yarza P**, **Gerken J**, **Pruesse E**, **Quast C**, **Schweer T**, **Peplies J**, **Ludwig W**, **Glöckner FO**. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res **42**:D643–D648.

11. **Sawyer SE**, **Rekepalli B**, **Horton MD**, **Brook RG**. 2015. HPC-BLAST: Distributed BLAST for Xeon Phi Clusters. Proc 6th ACM Conf Bioinformatics, Comput Biol Heal Informatics 512–513.

12. **Brook R**, **Heinecke A**, **Costa AB**, **Peltz Jr. P**, **Betro VC**, **Baer T**, **Bader M**, **Dubey P**. 2015. Beacon: Deployment and application of Intel Xeon Phi coprocessors for scientific computing. Comput Sci Eng **17**:65–72.

13. **Yarza P**, **Yilmaz P**, **Pruesse E**, **Glöckner FO**, **Ludwig W**, **Schleifer K**, **Whitman WB**, **Euzéby J**, **Amann R**, **Rosselló-móra R**. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol **12**:635–645.

14. **Eloe-Fadrosh E a**, **Ivanova NN**, **Woyke T**, **Kyrpides NC**. 2016. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. Nat Microbiol **1**:15032.

15. **Elshahed MS**, **Youssef NH**, **Sheik C**, **Najar FZ**, **Sukharnikov LO**, **Roe BA**, **Davis JP**, **Schloss PD**, **Bailey VL**, **Krumholz LR**. 2008. Novelty and uniqueness

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

patterns of rare members of the soil biosphere. Appl Environ Microbiol **74**:5422–5428.

16. **Karst SM**, **Dueholm MS**, **Mcilroy SJ**, **Kirkegaard RH**, **Nielsen PH**, **Albertsen M**. 2018. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nat Biotechnol **advance on**.

17. **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 ." Sci York.

18. **Kemp PF**, **Lee S**, **Laroche J**. 1993. Estimating the growth rate of slowly growing marine bacteria from RNA content. Appl Environ Microbiol **59**:2594–601.

19. **Blazewicz SJ**, **Barnard RL**, **Daly RA**, **Firestone MK**. 2013. Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. ISME J **7**:2061–2068.

20. **Thrash JC**, **Seitz KW**, **Baker BJ**, **Temperton B**, **Gillies LE**, **Rabalais NN**, **Henrissat B**, **Mason U**. 2017. Metabolic roles of uncultivated bacterioplankton lineages in the Northern Gulf of Mexico "Dead Zone." MBio **8**:1–20.

21. **Browne HP**, **Forster SC**, **Anonye BO**, **Kumar N**, **Neville BA**, **Stares MD**, **Goulding D**, **Lawley TD**. 2016. Culturing of "unculturable" human microbiota reveals novel taxa and extensive sporulation. Nature **533**:543–546.

22. **Teske A**, **Sørensen KB**. 2008. Uncultured archaea in deep marine subsurface sediments: have we caught them all? ISME J **2**:3–18.

23. **Luef B**, **Frischkorn KR**, **Wrighton KC**, **Holman HN**, **Birarda G**, **Thomas BC**, **Singh A**, **Williams KH**, **Siegerist CE**, **Tringe SG**, **Downing KH**, **Comolli LR**, **Banfield JF**. 2015. Diverse uncultivated ultra-small bacterial cells in groundwater. Nat Commun **6**:1–8.

24. **Rappe MS**, **Connon SA**, **Vergin KL**, **Giovannoni SJ**. 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature **418**:0–3.

25. **Carini P**, **Marsden PJ**, **Leff JW**, **Morgan EE**, **Strickland MS**, **Fierer N**. 2016. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. Nat Microbiol **2**:1–6.

26. **Staley JT**, **Konopka A**. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol **39**:321–346.

27. **Yano Y**, **Nakayama A**, **Yoshida K**. 1997. Distribution of polyunsaturated fatty acids in bacteria present in intestines of deep-sea fish and shallow-sea poikilothermic animals. Appl Environ Microbiol **63**:2572–2577.

28. **Chin K**, **Hahn D**, **Hengstmann ULF**, **Liesack W**, **Janssen PH**. 1999. Characterization and identification of numerically abundant culturable bacteria from the anoxic bulk soil of rice paddy microcosms. Appl Environ Microbiol **65**:5042–5049.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29. **Kaeberlein T**, **Lewis K**, **Epstein SS**. 2002. Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment. Science **296**:1127–9.

30. **Wilms R**, **Engelen B**, **Cypionka H**, **Sass H**. 2005. Microbial diversity in coastal subsurface sediments: a cultivation approach using various electron acceptors and substrate gradients. Appl Environ Microbiol **71**:7819–7830.

31. **Janssen PH**, **Yates PS**, **Grinton BE**, **Taylor PM**, **Sait M**. 2002. Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia. Appl Environ Microbiol **68**:2391–2396.

32. **Xu H-S**, **Roberts N**, **Singleton FL**, **Attwell RW**, **Grimes DJ**, **Colwell RR**. 1982. Survival and viability of nonculturable Escherichia coli and Vibrio cholerae in the estuarine and marine environment. Microb Ecol **8**:313–323.

33. **Knittel K**, **Boetius A**. 2009. Anaerobic Oxidation of Methane: Progress with an Unknown Process. Annu Rev Microbiol **63**:311–334.

34. **Könneke M**, **Bernhard AE**, **Torre R De**, **Walker CB**, **Waterbury JB**, **Stahl DA**. 2005. Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature **437**:543–546.

35. **Henson MW**, **Pitre DM**, **Weckhorst JL**, **Lanclos VC**, **Webber AT**, **Thrash JC**. 2016. Artificial seawater media facilitate cultivating members of the microbial majority from the Gulf of Mexico. mSphere **1**:1–11.

36. **Hoehler TM**, **Jørgensen BB**. 2013. Microbial life under extreme energy limitation. Nat Rev Microbiol **11**:83–94.

37. **Tahon G**, **Tytgat B**, **Lebbe L**, **Carlier A**, **Willems A**. 2018. Abditibacterium utsteinense sp. nov., the first cultivated member of candidate phylum FBP, isolated from ice-free Antarctic soil samples. Syst Appl Microbiol **accpeted**.

38. **Hug LA**, **Baker BJ**, **Anantharaman K**, **Brown CT**, **Probst AJ**, **Castelle CJ**, **Butter CN**, **Hernsdorf AW**, **Amano Y**, **Ise K**, **Suzuki Y**, **Dudek N**, **Relman DA**, **Finstad KM**, **Amundson R**, **Thomas BC**, **Ban JF**. 2016. A new view of the tree of life. Nat Microbiol **1**:1–6.

39. **Nobu MK**, **Dodsworth JA**, **Murugapiran SK**, **Rinke C**, **Gies EA**, **Webster G**, **Schwientek P**, **Kille P**, **Parkes RJ**, **Sass H**, **Jørgensen BB**, **Weightman AJ**, **Liu W**, **Hallam SJ**. 2015. Phylogeny and physiology of candidate phylum independent genomics 1–14.

40. **Youssef NH**, **Rinke C**, **Stepanauskas R**, **Farag I**, **Woyke T**, **Elshahed MS**. 2014. Insights into the metabolism , lifestyle and putative evolutionary history of the novel archaeal phylum " Diapherotrites ." ISME J **9**:447–460.

41. **Spang A**, **Caceres EF**, **Ettema TJG**. 2017. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. Science (80-) **357**:eaaf3883.

42. **Solden LM**, **Hoyt DW**, **Collins WB**, **Plank JE**, **Daly RA**, **Hildebrand E**, **Beavers TJ**, **Wolfe R**, **Nicora CD**, **Purvine SO**, **Carstensen M**, **Lipton MS**,

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Spalinger DE, **Firkins JL**, **Wolfe BA**, **Wrighton KC**. 2016. New roles in hemicellulosic sugar fermentation for the uncultivated Bacteroidetes family BS11. ISME J **11**:691–703.

43. **Lloyd KG**, **Schreiber L**, **Petersen DG**, **Kjeldsen KU**, **Lever MA**, **Steen AD**, **Stepanauskas R**, **Richter M**, **Kleindienst S**, **Lenk S**, **Schramm A**, **Jørgensen BB**. 2013. Predominant archaea in marine sediments detrital proteins. Nature **496**:215–218.

44. **Morono Y**, **Terada T**, **Nishizawa M**, **Ito M**, **Hillion F**, **Takahata N**, **Sano Y**, **Inagaki F**. 2011. Carbon and nitrogen assimilation in deep subseafloor microbial cells. Proc Natl Acad Sci U S A **108**:18295–300.

45. **Hatzenpichler R**, **Scheller S**, **Tavormina PL**, **Babin BM**, **Tirrell DA**, **Orphan VJ**. 2014. In situ visualization of newly synthesized proteins in environmental microbes using amino acid tagging and click chemistry. Environ Microbiol **16**:2568–2590.

46. **Kopf SH**, **Sessions AL**, **Cowley ES**, **Reyes C**, **Sambeek L Van**, **Hu Y**, **Orphan VJ**, **Kato R**, **Newman DK**. 2015. Trace incorporation of heavy water reveals slow and heterogeneous pathogen growth rates in cystic fibrosis sputum. Proc Natl Acad Sci USA **113**:E110–E116.

47. **Kevorkian R**, **Bird JT**, **Shumaker A**, **Lloyd KG**. Estimating population turnover rates from relative quantification methods reveals microbial dynamics in marine sediment. Appl Environ Microbiol (in Press AEM.01443-17.

48. **Kallmeyer J**, **Pockalny R**, **Adhikari RR**, **Smith DC**, **D'Hondt S**. 2012. Global distribution of microbial abundance and biomass in subseafloor sediment. Proc Natl Acad Sci U S A **109**:16213–16216.

49. **Whitman WB**, **Coleman DC**, **Wiebe WJ**. 1998. Prokaryotes: The unseen majority. Proc Natl Acad Sci USA **95**:6578–6583.

50. **Lindow SE**, **Brandl MT**. 2003. Microbiology of the phyllosphere. Appl Environ Microbiol **69**:1875–1883.

51. **Kieft TL**, **Simmons KA**. 2015. Allometry of animal – microbe interactions and global census of animal-associated microbes. Proc R Soc B **282**.

52. **Bowman JP**, **Mccammon SA**, **Gibson JAE**, **Robertson L**, **Nichols PD**. 2003. Prokaryotic Metabolic Activity and Community Structure in Antarctic Continental Shelf Sediments **69**:2448–2462.

53. **Su J**, **Engelen B**, **Cypionka H**, **Sass H**. 2004. Quantitative analysis of bacterial communities from Mediterranean sapropels based on cultivation-dependent methods **51**:109–121.

54. **Freitag TE**, **Klenke T**, **Krumbein WE**, **Gerdes G**, **Prosser JI**. 2003. Effect of anoxia and high sulphide concentrations on heterotrophic microbial communities in reduced surface sediments (Black Spots) in sandy intertidal flats of the German Wadden Sea. FEMS Microbiol Ecol **44**:291–301.

55. **Olsen RA**, **Bakken LR**. 1987. Viability of soil bacteria: optimization of plate-

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

counting technique and comparison between total counts and plate counts within different size groups. Microb Ecol **13**:59–74.

56. **Razumov A**. 1932. The direct method of calculation of bacteria in water: comparison with the Koch method. Mikrobiologiia **1**:131–146.

57. **Pedersen K**, **Ekendahl S**. 1990. Distribution and activity of bacteria in deep granitic groundwaters of Southeastern Sweden. Microb Ecol **20**:37–52.

58. **Bone TL**, **Balkwill DL**. 1988. Morphological and cultural comparison of microorganisms in surface soil and subsurface sediments at a pristine study site in Oklahoma. Microb Ecol **16**:49–64.

59. **Hirsch P**, **Rades-Rohkohl E**. 1988. Some special problems in the determination of viable counts of groundwater microorganisms. Microb Ecol **16**:99–113.

60. **Ludvigsen L**, **Ringelberg DB**, **Ekelund F**, **Christensen TH**. 1999. Distribution and composition of microbial populations in a landfill leachate contaminated aquifer (Grindsted, Denmark). Microb Ecol **37**:197–207.

61. **Sass AM**, **Sass H**, **Coolen MJL**, **Cypionka H**, **Overmann J**. 2001. Microbial communities in the chemocline of a hypersaline deep-sea basin (Urania Basin, Mediterranean Sea). Appl Environ Microbiol **67**:5392–5402.

62. **Huber JA**, **Butterfield DA**, **Baross JA**. 2002. Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge subseafloor habitat. Appl Environ Microbiol **68**:1585–1594.

63. **Bruns A**, **Cypionka H**, **Overmann J**. 2002. Cyclic AMP and Acyl homoserine lactones increase the cultivation efficiency of heterotrophic bacteria from the Central Baltic Sea. Appl Environ Microbiol **68**:3978–3987.

64. **Junge K**, **Imhoff F**, **Staley T**, **Deming JW**. 2002. Phylogenetic diversity of numerically important Arctic sea-ice bacteria cultured at subzero temperature. Microb Ecol **43**:315–328.

65. **Jannasch HW**, **Jones GE**. 1959. Bacterial populations in sea water as determined by different methods of enumeration. Limnol Oceanogr **4**:128–139.

66. **Cragg B**, **Harvey S**, **Fry J**, **Herbert R**, **Parkes R**. 1992. Bacterial Biomass and Activity in the Deep Sediment Layers of the Japan Sea, Hole 798B. Proc Ocean Drill Progr **127/128**:761–776.

67. **Cragg BA**, **Parkes RJ**, **Fry JC**, **Weightman AJ**, **Rochelle PA**, **Maxwell JR**. 1996. Bacterial populations and processes in sediments containing gas hydrates (ODP Leg 146: Cascadia Margin). Earth Planet Sci Lett **139**:497–507.

68. **Wellsbury P**, **Goodman K**, **Cragg B a**, **Parkes RJ**. 2000. The geomicrobiology of deep marine sediments from Blake Ridge containing methane hydrate (Sites 994, 995, and 997). Proc Ocean Drill Progr Sci Results **164**:379–391.

69. **Wellsbury P**, **Mather I**, **Parkes RJ**. 2002. Geomicrobiology of deep, low organic carbon sediments in the Woodlark Basin, Pacific Ocean. FEMS Microbiol Ecol **42**:59–70.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

70. **Bartscht K**, **Cypionka H**, **Overmann J**. 1999. Evaluation of cell activity and of methods for the cultivation of bacteria from a natural lake community. FEMS Microbiol Ecol **28**:249–259.

71. **Nicomrat D**, **Dick WA**, **Tuovinen OH**. 2006. Microbial populations identified by Fluorescence In Situ Hybridization in a constructed wetland treating acid coal mine drainage. J Environ Qual **35**:1329–1337.

Acknowledgments: This work was supported by (1) the National Science Foundation under grant numbers OCE-1431598 (KL), 1137097 (LC and JY), and 0711134 (LC and JY), (2) the NSF Center for Dark Energy Biosphere Investigations (OCE-0939564) contribution # *to be filled in* (KL), (3) the Alfred P. Sloan Foundation Fellowship (FG-2015-65399, KL), (4) the Simons Foundation (404586, KL), and (5) NASA Exobiology (NNX16AL59G), (6) the University of Tennessee, Knoxville College of Arts and Sciences (LC and JY), (7) the Joint Institute for Computational Sciences and the Beacon Project (LC and JY), (8) Intel Corporation through an Intel Parallel Computing Center award to support development of HPC-BLAST (LC and JY), and (9) the Oak Ridge National Laboratory Science Alliance. Andrew D. Steen helped with the R code and Tatiana Vishnivetskaya performed the Russian to English translations. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the University of Tennessee, or Intel Corporation.

Author Contributions

K. G. L. conceived of and obtained funding for the project, analyzed data, and led the writing of the paper; J. L. and A. D. S. provided supplementary data analysis and helped write the paper; J. Y. parallelized the BLAST analyses and helped write the paper; L. C. obtained funding, advised J. Y., and helped write the paper.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Metagenome-based estimates of global microbial cell abundances from uncultured archaea and bacteria, based on 16S rRNA gene sequence read depths. Estimated fractions are in parentheses. Environments with fewer microbial cells were excluded.

	Cells x 10^{26}			
Environment	**Total microbial cells**	**Cultured species to genera[#]**	**Uncultured genera to classes[#]**	**Uncultured Phyla and higher[#]**
Marine sediment (48)	2,900	390 (13%)	1,921 (66%)	590 (20%)
Soil (49)	2,560	454 (18%)	1,268 (50%)	839 (33%)
Terrestrial subsurface (49)	2,500	702 (28%)	1,211 (48%)	587 (23%)
Seawater (49)	1,010	143 (14%)	640 (63%)	229 (23%)
Freshwater (49)	1.3	0.1 (11%)	0.8 (64%)	0.3 (25%)
Plant hosts (50)	1	0.5 (49%)	0.4 (37%)	0.1 (14%)
Animal hosts (51)	0.2	0.1 (49%)	0.1 (37%)	0.0 (14%)
Sum	**8,974**	**1,689 (19%)**	**5,050 (56%)**	**2,245 (25%)**

[#]Cut-offs are the upper 95% confidence interval of the median 16S rRNA gene identity for each taxonomic level (13).

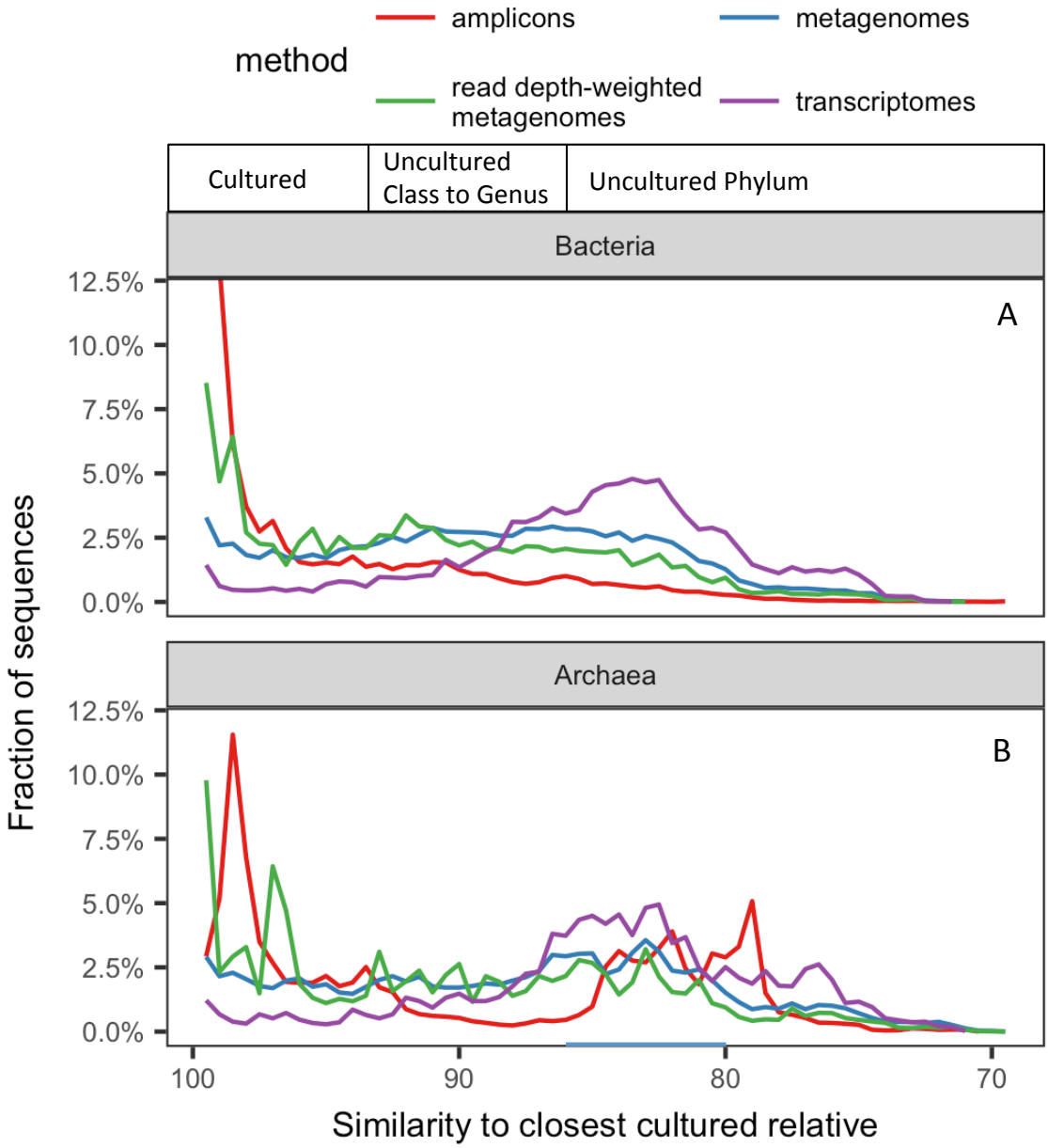
Table 2. Metatranscriptome-based estimates of global microbial cell abundances from uncultured archaea and bacteria, based on 16S rRNA gene sequence numbers. Estimated fractions are in parentheses. Environments with fewer microbial cells were excluded.

	Cells x 10^{26}			
Environment	**Total microbial cells**	**Cultured species to genera[#]**	**Uncultured genera to classes[#]**	**Uncultured Phyla and higher[#]**
Marine sediment (48)	NA	NA	NA	NA
Soil (49)	2,560	49 (2%)	758 (30%)	1,753 (69%)
Terrestrial subsurface (49)	2,500	45 (2%)	597 (24%)	1,858 (74%)
Seawater (49)	1,010	36 (4%)	389 (38%)	587 (58%)
Freshwater (49)	1.3	0.0 (3%)	0.5 (40%)	0.7 (56%)
Plant hosts (50)	1	0.2 (18%)	0.3 (33%)	0.5 (49%)
Animal hosts (51)	0.2	0.0 (18%)	0.1 (33%)	0.1 (49%)
Sum	**6,074**	**129 (2%)**	**1,744 (29%)**	**4,200 (69%)**

[#]Cut-offs are the upper 95% confidence interval of the median 16S rRNA gene identity for each taxonomic level (13).

NA means not applicable because too few metatranscriptome data exist from this environment to be included.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



606
607 **Figure 1. Fraction of 16S rRNA genes from bacteria (A) and archaea (B) in public**
608 **databases from primer-amplified, metagenomes (with and without read depths),**
609 **and metatranscriptomes at different percent identities with their closest cultured**
610 **relative.** Box along the top shows estimated cut-offs for different taxonomic level of
611 novelty relative to all cultures (13). Primer-amplified bacterial sequences go up to 30% at
612 100% similar to their closest cultured relative, but were removed for clarity.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

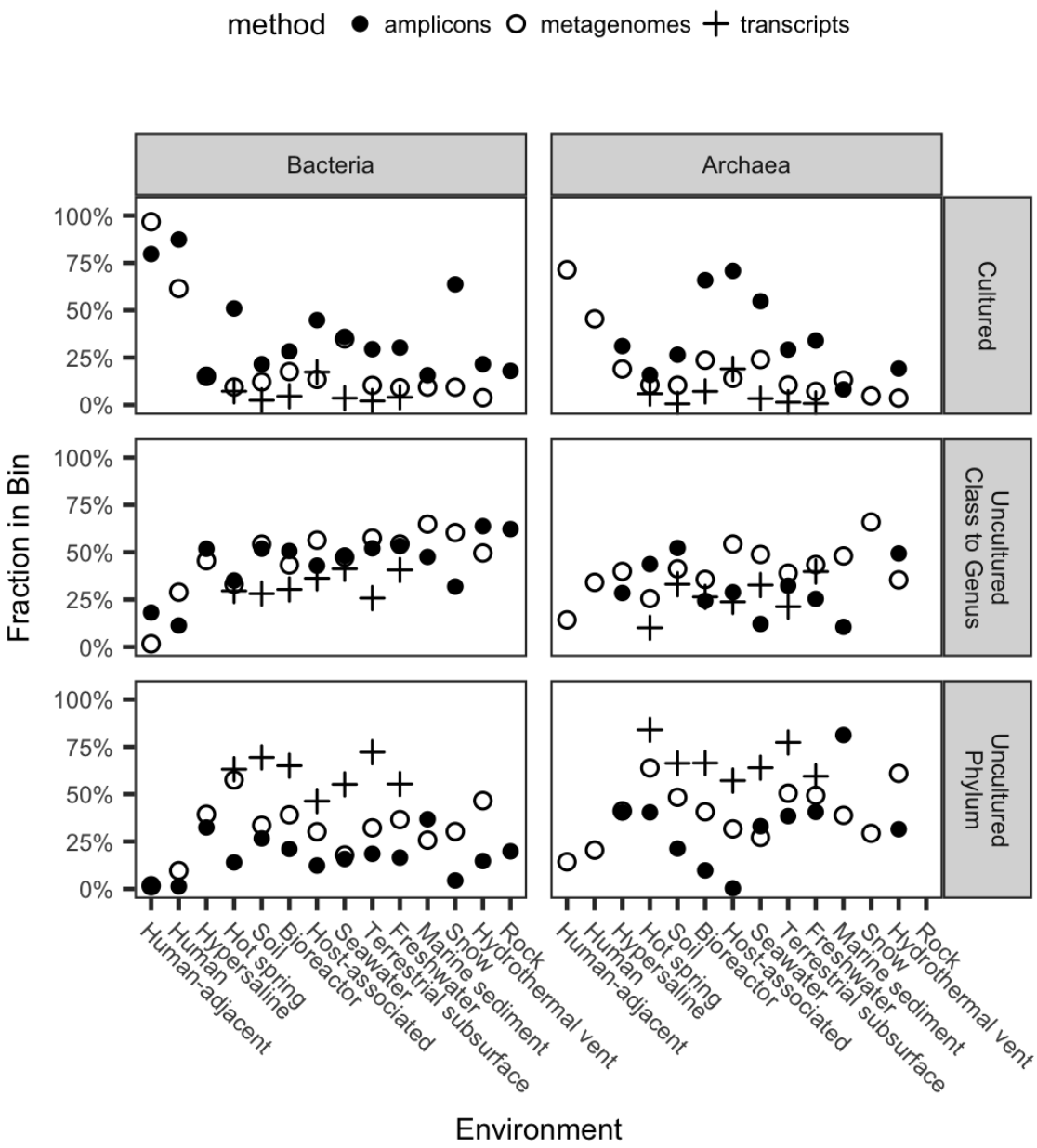
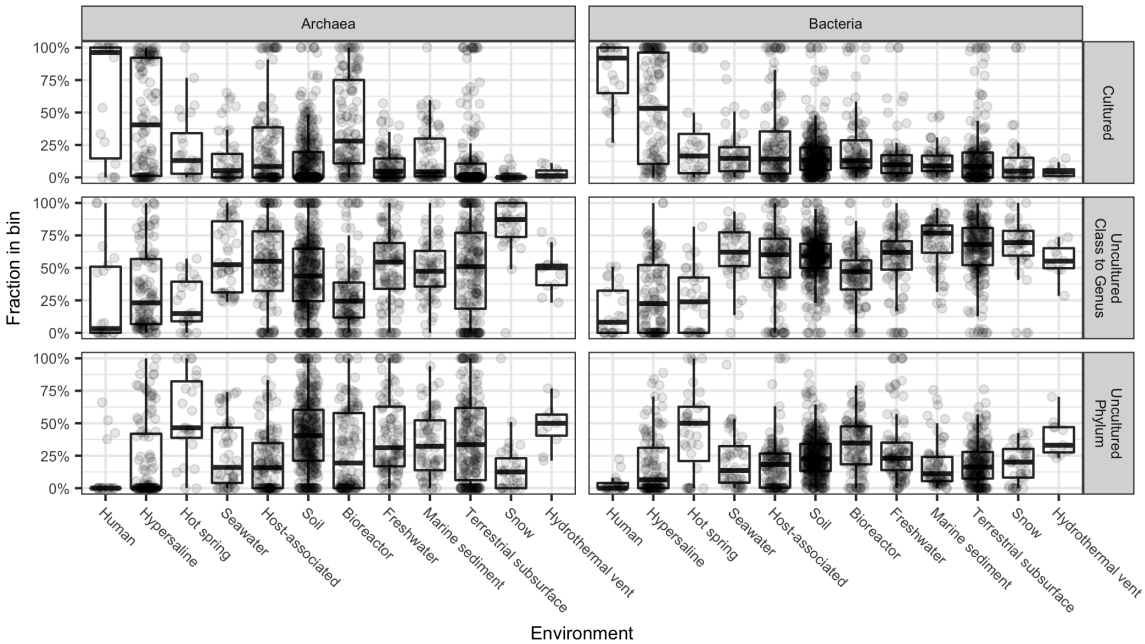


Figure 2. Proportion of 16S rRNA sequences in each category of phylogenetic novelty relative to cultures for each environment, by amplicons, metagenomes (without scaffold read depth), and metatranscriptomes. Closed circles are primer-amplified amplicons, open circles are metagenomes, and crosses are transcriptomes.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

623

624



625

626 **Fig. 3 Proportion of 16S rRNA gene sequences by scaffold read depth averaged**
627 **across all metagenomes.** Each single datapoint represents the abundance of reads in that
628 similarity bin from a single metagenome. Rows represent different similarity bins.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

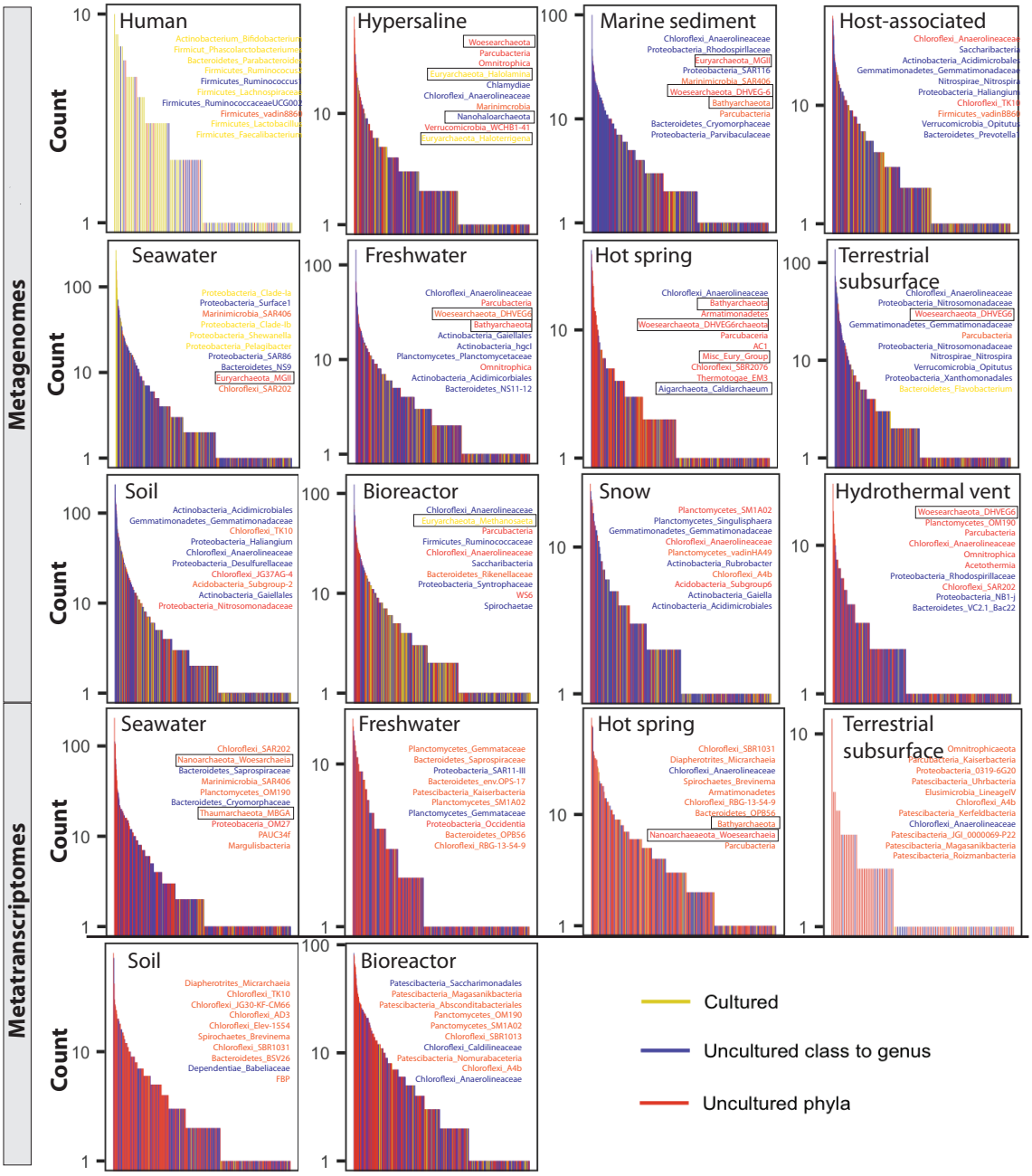


Fig. 4 Rank abundance plots by taxonomic genus assignments for metagenomic data (top three rows) and metatranscriptomic data (bottom two rows). Listed in each box are the top ten most abundant genera for that environment in the format of phylum_lowest identified taxonomic group, with archaea outlined in black. Data are colored for uncultured phyla (orange), uncultured class to genus (blue) or cultured (yellow). Taxonomic-based genera that had sequences from multiple phylogeny-based percent identity bins were labeled with the color of the bin with the most sequences.

bioRxiv preprint doi: https://doi.org/10.1101/303602; this version posted April 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

638

639
640