# Metagenomics uncovers gaps in amplicon-based detection of microbial diversity

Emiley A. Eloe-Fadrosh[1], Natalia N. Ivanova[1], Tanja Woyke[1], Nikos C. Kyrpides[1*]

[1]Joint Genome Institute, Walnut Creek, CA 94598, USA.

*Corresponding author. Mailing address: DOE Joint Genome Institute, Walnut Creek, CA 94598. Phone: 925-296-5718. E-mail: nckyrpides@lbl.gov.

Our understanding of the microbial world has made significant advances through the application of molecular approaches, particularly PCR-based amplification of the small-subunit ribosomal (SSU rRNA) gene[1]. The accumulation of SSU rRNA gene sequences has increased dramatically[2], yet the pace of discovery for new taxonomic lineages uncovered through PCR-based biodiversity surveys has seemingly slowed[3]. On the other hand, new candidate phyla have been identified using metagenomic and single-cell genomic techniques[4-6]. The above raises the question: Have we approached saturation or are there systematic biases in PCR-based surveys that preclude discovery of additional microbial lineages? Arguably, there is a wealth of microbial clades that are systematically under-represented or missed altogether, leaving major taxonomic “blind spots”[7]. PCR amplification biases, including primer mismatches, are well-recognized pitfalls in biodiversity surveys[8], yet a comprehensive analysis of the prevalence of such “blind spots” has not been undertaken[9]. Here, we systematically surveyed primer fidelity in SSU rRNA gene sequences recovered from over 6,000 assembled metagenomes sampled globally. Our findings show approximately 10% of environmental microbial sequences might be missed from classical PCR-based SSU rRNA gene surveys, mostly members of the Candidate Phyla Radiation (CPR)[4] and as-yet uncharacterized Archaea. These results underscore the extent of uncharacterized microbial diversity, and provide fruitful avenues for describing additional phylogenetic lineages.

We compared SSU rRNA gene sequences ≥ 1000 bp recovered from metagenomes (38,454 SSU rRNA) and isolate genomes (25,439 SSU rRNA) available through the Integrated Microbial Genomes with Microbiome Samples (IMG/M)[10] against commonly used PCR primer sets (515F-806R from the Earth Microbiome Project[2,11]; 515F-926R[12]; 341F-785R[13]; and 357F-926R from the Human Microbiome Project[14]), and generated a weighted score based on the number of mismatches (**Fig. 1a**; **Supplementary Table 1**; Methods). Unexpectedly, this analysis indicates that a minimum of 9.6% of environmental bacterial and/or archaeal sequences based on metagenomic data might not be recovered using a targeted PCR survey (**Fig. 1a**). The newly modified primers part of the Earth Microbiome Project appear to more fully capture SSU rRNA diversity (9.6% metagenomic sequences would likely be missed), while the relatively poor performance of the Human Microbiome Project primer set may be due to the narrow bacterial target range (22.4% bacterial metagenomic sequences would likely be missed). Notably, combining any two primer pairs does not significantly improve taxonomic coverage (**Supplementary Fig. 1**). Even when the four best-performing primer sets are combined together, 5.5% of sequences remain that would be missed based on metagenomic data, suggesting the use of multiple primer sets might slightly improve recovery but does not fully resolve the issue (**Supplementary Fig. 1**). Overall, these results are most likely an underestimate of diversity missed since our data consisted of assembled metagenomic contigs, primarily representing abundant organisms in a sample and neglecting less abundant phylotypes (e.g. the 'rare biosphere'). Taken together, we hypothesize that amplification-unbiased exploration of microbial diversity via metagenome and metatranscriptome sequencing will unquestionably improve our current view of the microbial tree of life.

An evaluation of base-specific biases for the commonly used PCR primer sets revealed a subset of bases contributing to the percentage of metagenomic SSU rRNA gene sequences that would likely be missed in PCR-based surveys (**Supplementary Fig. 2**). These bases, or "hot spots," could be candidates for increased degeneracy in

the current primers to capture a greater fraction of the microbial diversity or serve as guides in the design of new primer sets. Regardless, modifications of the specific bases contributing to the inefficiency of these primers would need to be experimentally validated prior to proposing improved primers.

It has been previously shown experimentally that primer mismatches can significantly affect species evenness[4]. In order to verify this, we evaluated the SSU rRNA gene sequences with primer mismatches from a set of eight matched metagenomic datasets and SSU rRNA surveys from a diversity of environmental locales to determine whether the predicted mismatches would impact recovery in SSU rRNA surveys (**Supplementary Table 2**). In all eight matched metagenomic datasets and SSU rRNA surveys, the computationally predicted missed sequences were not recovered in the SSU rRNA survey. We further evaluated the SSU rRNA gene sequences with primer mismatches from the large-scale Human Microbiome Project metagenomic datasets and compared them with their corresponding SSU rRNA surveys[14]. We searched more than 34 million reads from over 4,200 Human Microbiome Project SSU rRNA surveys against the 130 sequences from the HMP metagenomes predicted to be missed and found only 2,060 matches. Although we observe a small number of matches, these represent only 0.006% relative abundance in the SSU rRNA surveys and are in contrast to the two orders of magnitude greater abundance based on shotgun metagenome data. Together, these results are consistent with our computational predictions, and suggest that the primer mismatches would likely significantly reduce or prevent the recovery of taxonomic “blind spots” in PCR-based surveys.

Our analysis revealed phylogenetic patterns for those sequences that would presumably be missed with the widely used environmental primer set 515F-806R (overall primer weighted score ≥ 1) [2,11]. As anticipated, members of the recently described Candidate Phyla Radiation (CPR[4]; including Parcubacteria (OD1), Microgenomates (OP11), WWE3, Berkelbacteria (ACD58), Saccharibacteria (TM7),

WS6, Peregrinibacteria (PER), and Kazan phyla) collectively represented 70% of SSU rRNA gene sequences that would likely be missed in PCR-based surveys (**Fig. 1b; Supplementary Fig. 3**). The overall length of the SSU rRNA gene further compounds these findings for the CPR since the prevalence of encoded introns may hamper amplification fidelity[4]. Within the domain Archaea, more than half of the taxonomic "blind spots" are phylogenetically positioned outside of the recognized phylum-level lineages, revealing significant untapped archaeal diversity awaiting discovery (**Fig. 1b; Supplementary Fig. 3**). Recent efforts to resolve archaeal diversity through genome-resolved metagenomic analyses have yielded substantial progress towards a better understanding of archaeal evolutionary history[5,15].

Genomic mapping of the tree of life has been accelerated through application of both metagenomic and single-cell genomic sequencing of samples taken directly from the environment without the arduous task of cultivation. Within the last few years, significant advances in high-throughput single-cell genomics have provided some of the first genomic insight for a wealth of candidate phylogenetic lineages previously known only through SSU rRNA gene sequencing[6]. Further, deep sequencing of environmental samples, combined with improved metagenomic assembly and binning methods are yielding complete or near-complete genomes from many novel bacterial and archaeal lineages (see[4,5,16]). We performed a phylogenetic analysis of all SSU rRNA gene sequences derived from metagenomes (regardless of whether these sequences had primer mismatches) to identify the sequences that could not be placed with known bacterial or archaeal phyla. These data suggest that bacterial diversity has been charted extensively, with minimal SSU rRNA gene orphan sequences not assignable to any phylum (**Table 1**). This is in stark contrast to that of the Archaea, where significant diversity exists beyond the currently described major lineages. Additionally, there is likely to be sizable taxonomic novelty at more refined taxonomic levels, such as class and order.

Habitat distribution of those sequences unaffiliated with currently recognized phyla based on our phylogenetic analysis provides insight into the suite of environmental locales potentially hosting as-yet uncharacterized microbial life (**Fig. 2**). The habitats where more unclassified sequences are found include "extreme" habitats with unique environmental parameters (e.g. extremes in temperature, pressure, and chemical composition), favoring a distinct composition of microbial communities in these environments. Our data does suggest that these habitats may harbor more divergent phylogenetic groups, specifically within the archaea. On the other hand, environments such as marine cold seeps also comprise a wealth of uncharacterized microbial diversity with sampling challenges traditionally hampering their genomic exploration for the as-yet uncharacterized microbial life. More recent targeted sampling efforts have begun to shed light on these unique environments[4-6,15].

Taken together, we posit that more recent technologies including metagenomics and single-cell genomics are bringing into focus the taxonomic "blind spots" that have thus far eluded detection. While PCR-based SSU rRNA surveys provided a sizeable snapshot of the phylogenetic structure of the microbial world, we are poised to make substantial strides forward in our efforts to more comprehensively capture microbial diversity. We anticipate new sequencing technologies such as single-molecule sequencing will enable full recovery of genomic content for those bacterial and archaeal phyla that are known only through SSU rRNA, as well as candidate phyla discovered through systematic large-scale data mining efforts. Uncharted microbial diversity will soon find its place in the tree of life.

## Methods

All SSU rRNA gene sequences ≥ 1000 bp were retrieved from the Integrated Microbial Genomes with Microbiome Samples (IMG/M)[10] database and manually curated to remove mitochondrial and chloroplast sequences. A total of 3,003 mitochondrial and 668 chloroplast sequences were identified and culled from the metagenomic dataset.

PrimerProspector[17] was used to evaluate primer sets against SSU rRNA gene sequences ≥ 1000 bp recovered from metagenomes (38,454) and isolate genomes (25,439). Primers used in the analysis are listed in **Supplementary Table 1**. Weighted scores were calculated as follows: non-3' mismatches = 0.4; 3' mismatches = 1; non-3' gaps = 1; and 3' gaps = 3. Therefore, a weighted score ≥ 1 could consist of three non-3' mismatches and was considered a missed sequence. Validation of the eight matched metagenomes and corresponding SSU rRNA surveys (**Supplementary Table 2**) was performed using blat[18] with the metagenomic sequences trimmed to the 515F-806R region. Similarly, the Human Microbiome Project metagenomic datasets and the corresponding SSU rRNA surveys[14] were validated using the metagenomic sequences trimmed to the 357F-926R region. Taxonomic affiliation was determined by aligning all SSU rRNA genes using cmalign (--matchonly option)[19] against bacterial and archaeal hmm models and building a phylogenetic tree using RAxML[20] with reference sequences to identify coherent clades. Taxonomic affiliation was assigned as follows: Cultured phyla, percentage of metagenomic SSU rRNA gene sequences matching phyla from cultured representatives; Uncultured phyla, percentage of metagenomic SSU rRNA gene sequences matching candidate phyla without cultured representatives; and Unaffiliated with current phyla, percentage of metagenomic SSU rRNA gene sequences that do not match a known, named phylum. Primer sequence logos were generated using WebLogo[21].

## References


1. Pace, N. R. *Microbiol. Mol. Biol. Rev.* **73**, 565-576 (2009).
2. Caporaso, J. G. *et al. ISME J.* **6**, 1621-1624 (2012).
3. Yarza, P. *et al. Nature Rev. Microbiol.* **12**, 635-645 (2014).
4. Brown, C. T. *et al. Nature* **523**, 208-211 (2015).
5. Castelle, C. J. *et al. Current Biol.* **25**, 690-701 (2015).
6. Rinke, C. *et al. Nature* **499**, 431-437 (2013).
7. Lynch, M. D. J. & Neufeld, J. D. *Nature Rev. Microbiol.* **13**, 217-229 (2015).
8. Engelbrektson, A. *et al. ISME J.* **4**, 642-647 (2010).
9. Woyke, T. & Rubin, E. M. *Science* **346**, 698-699 (2014).

10. Markowitz, V. M. *et al. Nucleic Acids Res.* **42**, D568-D573 (2014).
11. Gilbert, J. A. *et al. Stand. Genomic Sci.* **3**, 249-253 (2010).
12. Parada, A., Needham, D. M. & Fuhrman, J. A. *Environ. Microbiol.*, doi: 10.1111/1462-2920.13023 (2015).
13. Klindworth, A. *et al. Nucleic Acids Res.* **41**, e1 (2013).
14. The Human Microbiome Consortium. *Nature* **486**, 215-221 (2012).
15. Spang, A. *et al. Nature* **521**, 173-179 (2015).
16. Albertsen, M. *et al. Nat. Biotechnol.* **31**, 533-538 (2013).
17. Walters, W. A. *et al. Bioinformatics* **27**, 1159-1161 (2011).
18. Kent, W. *Genome Res.* **12**, 656-664 (2002).
19. Nawrocki, E. P. & Eddy, S. R. *Bioinformatics* **29**, 2933-2935 (2013).
20. Stamatakis, A. *Bioinformatics* **22**, 2688-2690 (2006).
21. Crooks, G. E. *et al*. *Genome Res.* **14**, 1188-1190 (2004).

Correspondence and requests for materials should be addressed to N.C.K.

**Acknowledgments**

This work was conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under Contract No. DE-AC02-05CH11231 and used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

**Author Contributions**

E.A.E-F., N.N.I., T. W., and N.C.K. designed the project, analyzed the data, and wrote the manuscript. All authors discussed the results and commented on the manuscript.

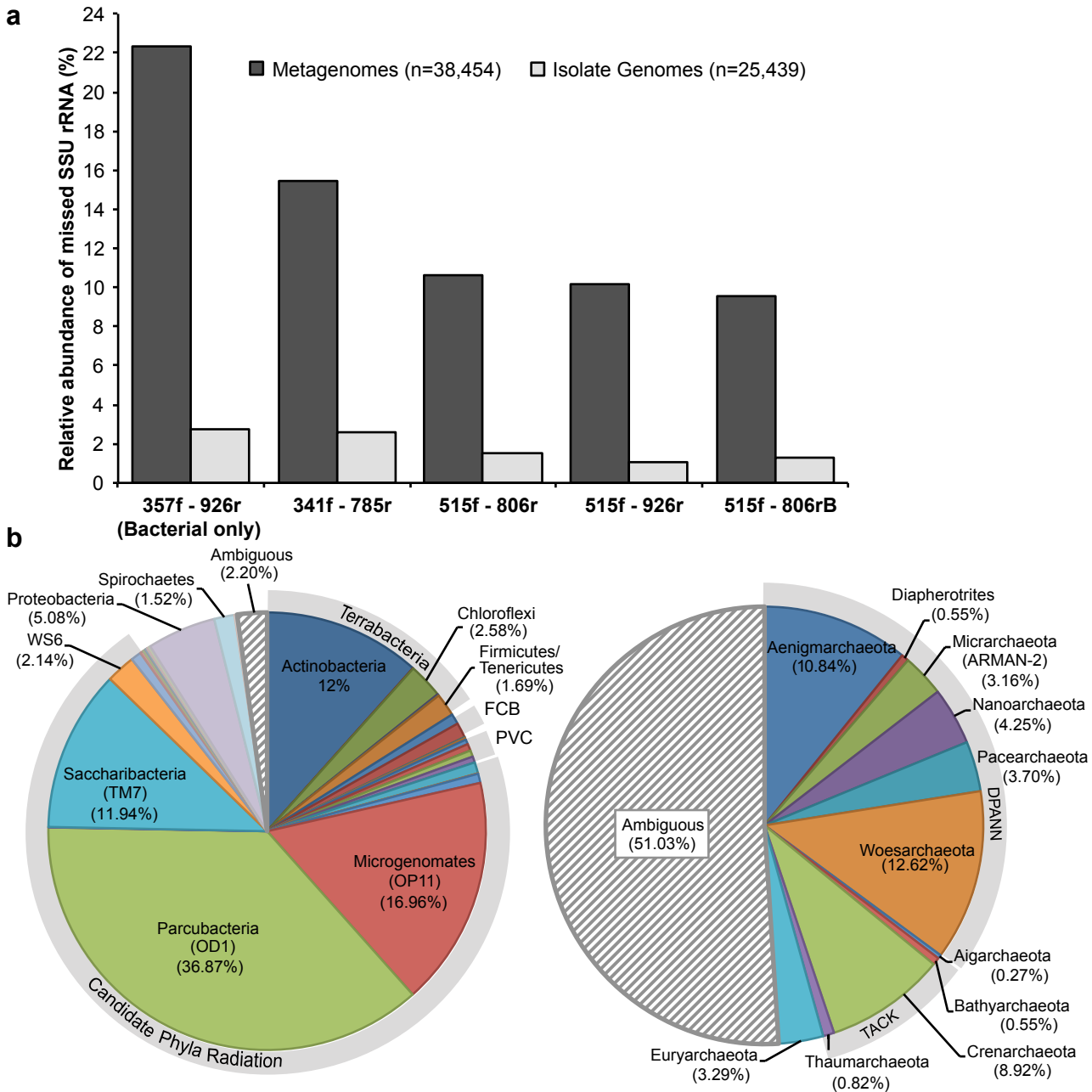
**Competing interests**

The authors declare no competing interests.

**Fig. 1. Primer fidelity for commonly used environmental PCR primers.** Primer sets were evaluated against 38,454 metagenomic and 25,439 isolate genome SSU rRNA gene sequences longer than 1000 bp. (**a**) Bar chart represents estimated percentage of

missed SSU rRNA gene sequences derived from metagenomes and isolate genomes. Details available in Methods. For HMP primer 357F-926R, only bacterial sequences were tested since they were not designed to include archaeal phylotypes. (**b**) Pie charts represents SSU rRNA gene sequences from metagenomes that would likely be missed with the widely used environmental primer set 515F-806R (overall primer weighted score ≥ 1). The phylum-level affiliation for bacteria (left; 3,366 total missed sequences) and archaea (right; 729 total missed sequences) are shown. Light gray bars along outside of pie charts represent superphyla as follows: Terrabacteria (Actinobacteria, Armatimonadetes (OP10), Chloroflexi, Cyanobacteria, Deinococcus-Thermus, Firmicutes and Tenericutes); FCB (Bacteroidetes, Caldithrix, Chlorobi, Cloacimonetes (WWE1), Fibrobacteres, Gemmatimonadetes, Kryptonia, Latescibacteria (WS3), Marinimicrobia (SAR406), Zixibacteria); PVC (Planctomycetes, Verrucomicrobia, Chlamydiae, Lentisphaerae, Omnitrophica (OP3)); Candidate Phyla Radiation (Berkelbacteria (ACD58), CPR2, CPR3, Gracilibacteria (GN02/BD1-5), Kazan, Microgenomates (OP11), Parcubacteria (OD1), Peregrinibacteria (PER), TM7, WS6, WWE3); DPANN (Diapherotrites, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaeota, Micrarchaeota (ARMAN-2), Parvarchaeota, Pacearchaeota, Woesearchaeota); and TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota, Bathyarchaeota).

**Fig. 2. Habitat distribution of taxonomic novelty.** Distribution of bacterial (black) and archaeal (gray) sequences recovered from metagenomes lacking definitive phylum-level affiliation. A total of 654 bacterial sequences and 816 archaeal sequences were analyzed based on environmental habitat metadata. Data was normalized based on the total number of SSU rRNA gene sequences ≥ 1000 bp recovered from metagenomes for each habitat.


a
Relative abundance of missed SSU rRNA (%)
Metagenomes (n=38,454)
Isolate Genomes (n=25,439)
357f - 926r (Bacterial only)
341f - 785r
515f - 806r
515f - 926r
515f - 806rB
b
Terrabacteria
Actinobacteria 12%
Chloroflexi (2.58%)
Firmicutes/ Tenericutes (1.69%)
FCB
PVC
Ambiguous (2.20%)
Spirochaetes (1.52%)
Proteobacteria (5.08%)
WS6 (2.14%)
Saccharibacteria (TM7) (11.94%)
Microgenomates (OP11) (16.96%)
Parcubacteria (OD1) (36.87%)
Candidate Phyla Radiation
Diapherotrites (0.55%)
Aenigmarchaeota (10.84%)
Micrarchaeota (ARMAN-2) (3.16%)
Nanoarchaeota (4.25%)
Pacearchaeota (3.70%)
DPANN
Ambiguous (51.03%)
Woesarchaeota (12.62%)
Aigarchaeota (0.27%)
Bathyarchaeota (0.55%)
TACK
Crenarchaeota (8.92%)
Thaumarchaeota (0.82%)
Euryarchaeota (3.29%)

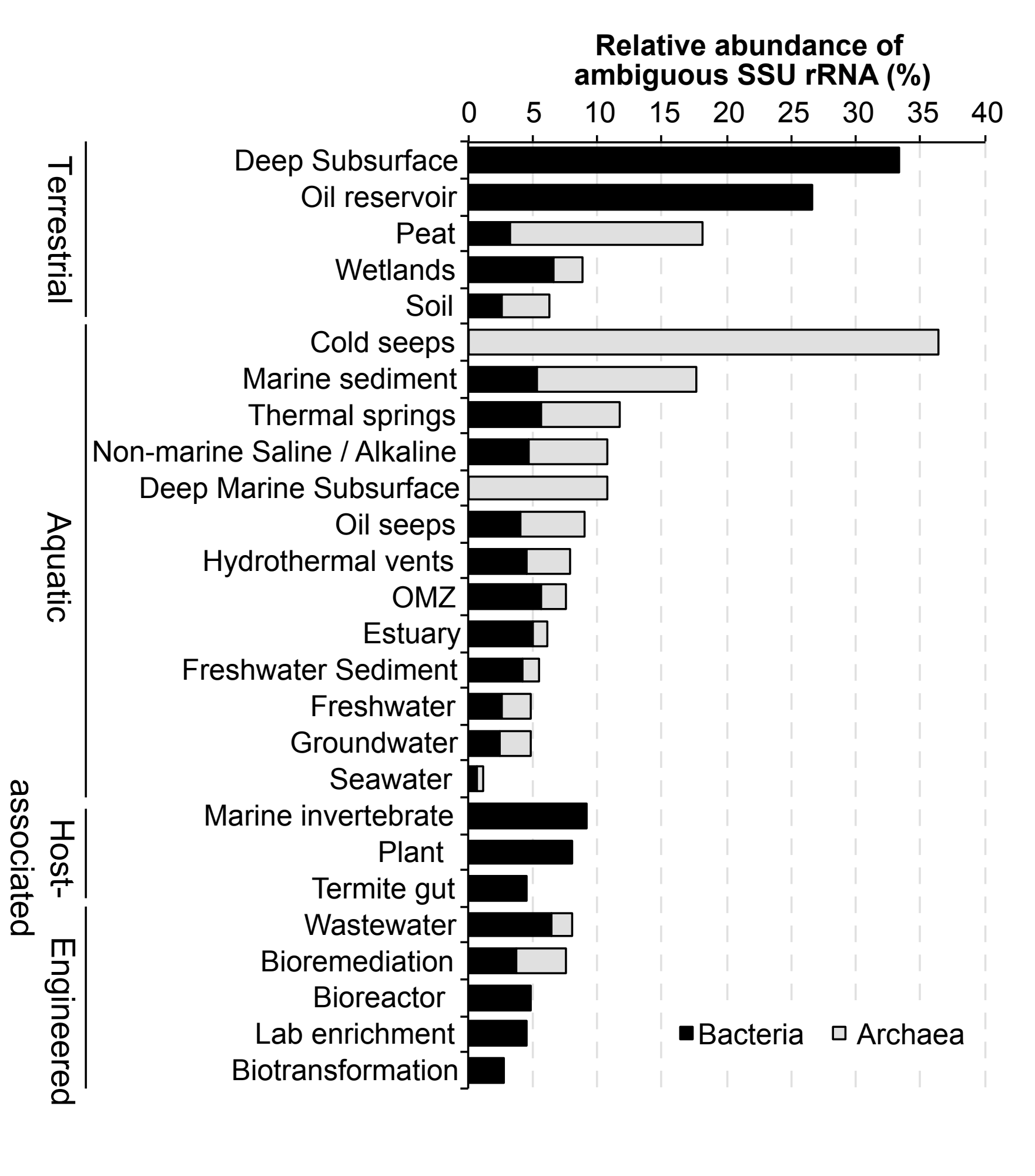

Relative abundance of ambiguous SSU rRNA (%)
0
5
10
15
20
25
30
35
40
Terrestrial
Aquatic
Host-associated
Engineered
Deep Subsurface
Oil reservoir
Peat
Wetlands
Soil
Cold seeps
Marine sediment
Thermal springs
Non-marine Saline / Alkaline
Deep Marine Subsurface
Oil seeps
Hydrothermal vents
OMZ
Estuary
Freshwater Sediment
Freshwater
Groundwater
Seawater
Marine invertebrate
Plant
Termite gut
Wastewater
Bioremediation
Bioreactor
Lab enrichment
Biotransformation
Bacteria
Archaea

**Table 1.** An estimate of taxonomic novelty derived from grouping metagenome-derived bacterial and archaeal SSU rRNA gene sequences into recognized phylum-level lineages.

| | Relative abundance of metagenome SSU rRNAs (%) | |
|---|---|---|
| | **Bacteria (n=34,596)** | **Archaea (n=3,858)** |
| Cultured Phyla | 83.53 | 46.73 |
| Uncultured Phyla | 14.58 | 32.12 |
| Unaffiliated with Current Phyla | 1.89 | 21.15 |