NAR Genomics and Bioinformatics, 2025, **7**, lqaf061
https://doi.org/10.1093/nargab/lqaf061
Standard Article

CABO-16S—a Combined Archaea, Bacteria, Organelle 16S rRNA database framework for amplicon analysis of prokaryotes and eukaryotes in environmental samples

Eryn M. Eitel[1,*,†], Daniel R. Utter[1,2,†], Stephanie A. Connon[1], Victoria J. Orphan[1,2], Ranjani Murali[2,3]

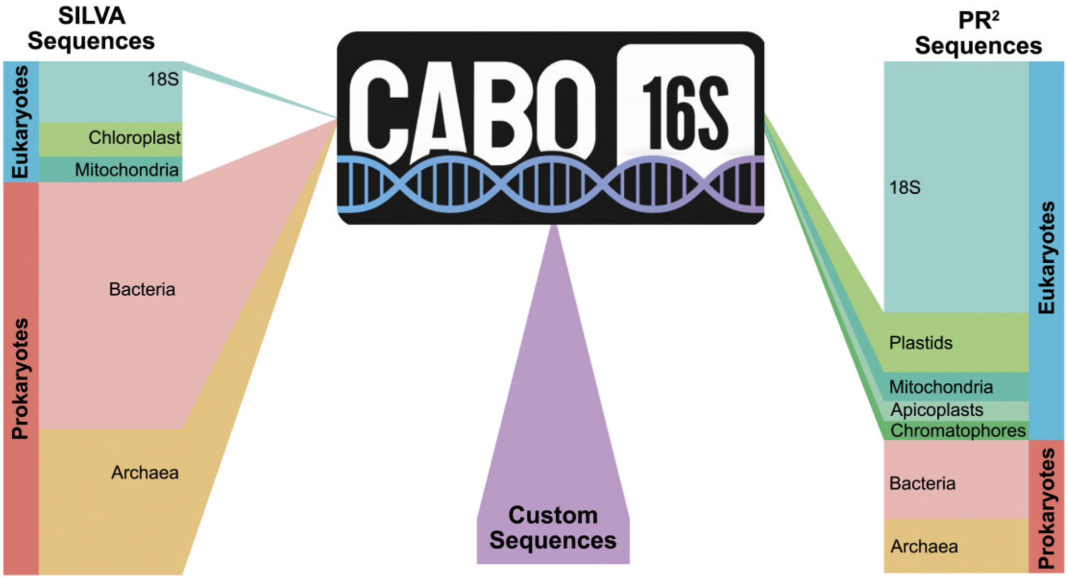
[1]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, United States
[2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, United States
[3]School of Life Sciences, University of Nevada, Las Vegas, NV 89154, United States

[*]To whom correspondence should be addressed. Email: eeitel@caltech.edu
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Identification of both prokaryotic and eukaryotic microorganisms in environmental samples is currently challenged by the need for additional sequencing to obtain separate 16S and 18S ribosomal RNA (rRNA) amplicons or the constraints imposed by "universal" primers. Organellar 16S rRNA sequences are amplified and sequenced along with prokaryote 16S rRNA and provide an alternative method to identify eukaryotic microorganisms. CABO-16S combines bacterial and archaeal sequences from the SILVA database with 16S rRNA sequences of plastids and other organelles from the PR2 database to enable identification of all 16S rRNA sequences. Comparison of CABO-16S with SILVA 138.2 results in equivalent taxonomic classification of mock communities and increased classification of diverse environmental samples. In particular, identification of phototrophic eukaryotes in shallow seagrass environments, marine waters, and lake waters was increased. The CABO-16S framework allows users to add custom sequences for further classification of underrepresented clades and can be easily updated with future releases of reference databases. Addition of sequences obtained from Sanger sequencing of methane seep sediments and curated sequences of the polyphyletic SEEP-SRB1 clade resulted in differentiation of syntrophic and non-syntrophic SEEP-SRB1 in hydrothermal vent sediments. CABO-16S highlights the benefit of combining and amending existing training sets when studying microorganisms in diverse environments.

Graphical abstract



Received: November 27, 2024. **Revised:** April 22, 2025. **Editorial Decision:** April 24, 2025. **Accepted:** May 11, 2025

Introduction

Life on Earth is composed of bacteria, archaea, and eukaryotes. All three branches of domains significantly influence ecosystem function [1], and while the impact and response of humans and other macro-organisms to their environment can often be directly observed [2–4], the role of unseen microorganisms is equally important [5–7]. Microorganisms generally refer to any organism not individually visible to the human eye and are usually smaller than 50 μm. This includes bacteria, archaea, viruses, fungi, and protists, with protists referring to unicellular eukaryotes that can be free-living or form colonies, but do not have differentiated cellular functions. Over the last 20 years, high-throughput sequencing of small subunit (SSU) ribosomal RNA (rRNA) has enabled study of the microbial ecology in terrestrial [8] and marine [9] environments and facilitated understanding of the microbiome in plants [10] and animals [11]. Major research efforts such as the Earth Microbiome Project (EMP) [12], Tara Oceans [13], and long-term human microbiome studies [14] have extensively used amplicon sequencing to characterize microbial communities. While metagenomic analysis can provide functional and strain-level insight, sequencing of SSU rRNA remains a fundamental tool for taxonomic assignment due to its widespread adoption, cost-effectiveness, and established protocols. Identification of prokaryotes was pioneered using the 16S rRNA gene [15], while the homologous 18S rRNA gene was optimized for eukaryote identification [16], and the internal transcribed spacer region proved best for fungi [17].

Amplification of both prokaryotic and eukaryotic microorganisms with a single polymerase chain reaction (PCR) would be ideal and could reduce amplicon library preparation costs two- to three-fold compared to the use of separate primers for 16S and 18S rRNA gene sequencing. Although some “universal” primers (515f/926r) can amplify eukaryotic 18S rRNA along with 16S rRNA [18], simultaneous and accurate analysis of both eukaryotic and prokaryotic microorganisms is challenging. First, mismatches between the primers and their template may be more common when attempting to amplify a broader target group, and indeed, in mock communities, only a single mismatch with the reverse primer resulted in a three- to eight-fold underestimation of eukaryotes [19]. Second, 18S sequences are typically 160–180 bp longer than 16S sequences and both PCR and sequencing are biased against longer amplicons [20]. Environmental samples may contain long 18S sequences [21] or higher proportions of dinoflagellates that tend to have mismatches, suggesting that underestimation of eukaryotes with 515f/926r primers may be significantly worse in those cases [19]. Finally, 16S rRNA gene copy numbers range from 1 to 15 in most bacteria and average only one copy in a majority of archaeal phyla [22, 23], while 18S gene copy numbers can vary between 1 and 1 800 000 in phytoplankton [24, 25]. Although 18S gene counts may be significantly correlated with biovolumes, they cannot be reliably used to determine relative taxonomic abundances [26].

Plastids, eukaryotic organelles originating from ancestral cyanobacterial endosymbionts, contain 16S rRNA copies that are taxonomically linked to their host organism rather than their cyanobacterial past. Plastid 16S rRNA gene copies are more constrained than 18S rRNA gene copies. The number of plastids, such as chloroplasts, varies less with cell size than 18S rRNA. Although environmental factors can still influence chloroplast counts, particularly in large taxa [27–29], within the chloroplasts there are typically one to two copies of the 16S rRNA gene. Together, this results in similar ranges of plastidal 16S rRNA gene counts and bacterial or archaeal 16S rRNA gene counts. Furthermore, although primer choices impacted early studies comparing 16S and 18S rRNA analysis [30], more recent works have found phytoplankton tthat taxonomy determined with 18S rRNA gene sequences is comparable to 16S rRNA [29, 31], and in some cases, plastid 16S rRNA may provide better phylogenetic resolution than 18S rRNA [32, 33].

Complete analysis of the 16S rRNA gene, including plastidal sequences, appears to provide a pathway for identification of bacterial, archaeal, and most photosynthetic eukaryotic microorganisms relevant in marine and terrestrial environments without the use of multiple primers or having to account for the nonuniform biases introduced during attempts to quantify both 16S and 18S rRNA with “universal” primers. As a part of routine 16S rRNA analysis, plastidal 16S is automatically amplified and sequenced along with prokaryote 16S rRNA. However, identification of plastid sequences is not currently possible, as a curated reference database that includes bacterial, archaeal, and plastidal sequences for taxonomic identification is lacking. Tedious manual verification is required to demonstrate that unidentified sequences are in fact plastids. If not annotated, plastid sequences can complicate sample analysis by artificially inflating the fraction of unidentified taxa in a sample, potentially leading to erroneous inference of “novel” taxa. The SILVA database (v138.2) [34] contains a comprehensive record of 16S rRNA gene sequences and is continuously updated to include recently identified taxa from novel environments; however, only broad identification of chloroplasts is possible with no ability to provide further taxonomic identification. On the other hand, to the best of our knowledge, the database PR2 [35] contains the most comprehensive plastidal 16S rRNA record, but with a focus on protists, limited bacterial and archaeal sequences are supplied. With the CABO-16S (Combined Archaeal, Bacterial, and Organelle for 16S) database, we combine the bacterial and archaeal sequences from SILVA with the 16S sequences of plastids and other organelles from PR2 to enable efficient identification of all 16S sequences. A framework and clearly annotated scripts are provided for CABO-16S to be updated with future releases of SILVA and PR2, or users can include sequences particular to their own area of interest. Taxonomic classification is compared between CABO-16S and SILVA-132.1 in mock microbial communities and published datasets from mammalian guts, deep-sea seep sites, seagrass beds, alkaline lakes with high abundances of microalgae, and terrestrial soils. Finally, addition of custom sequences to CABO-16S can allow identification of specialized taxonomic groups, even increasing our understanding of polyphyletic groups that are difficult to constrain within the current SILVA structure. This may allow deeper understanding of complex environments, such as deep-sea hydrothermal vents.

Materials and methods

Aggregation and training

To build both the CABO-16S database and the simplified SILVA 138.2 database used as a comparison in this study, sequences from the most recent version of SILVA available (138.2) were downloaded

(SILVA_138.2_SSURef_NR99_tax_silva.fasta.gz) along with mapped taxonomy (taxmap_slv_ssu_ref_nr_138.2.txt.gz) and quality values (SILVA_138.2_SSURef_Nr99.quality.gz). All sequences with a pintail value <50 or an alignment quality value <75 were removed. The prokaryotic taxonomy at the species level was refined to remove naming schemes based on organism host, sample collection, unclear bacterium groupings, or repetition of genus (i.e. "Genus sp."). For both CABO-16S and the SILVA 138.2 training set, sequences identified as mitochondria and eukaryotes were removed, except for 100 randomly selected eukaryotes, which were retained as an outgroup. The taxonomy of the eukaryotic outgroup was only identified at the phylum level. For the CABO-16S database, sequences identified as chloroplast were also removed. Plastid, apicoplast, mitochondrion, and chromatophore sequences were added to the CABO-16S database from the PR2 database (v 5.0.0) with acquisition using the R package "pr2database" (https://pr2database.github.io/pr2database/articles/pr2database.html). To match the seven rank orders found in SILVA taxonomy, the ranks of supergroup and subdivision were dropped from PR2 sequences. Finally, custom 16S rRNA sequences obtained from Sanger sequencing of methane seeps (https://doi.org/10.6084/m9.figshare.27288090) and a curated list of representative SEEP-SRB1 sequences [36] were combined with the selections from SILVA and PR2 to form the basis of the CABO-16S dataset. The number of sequences included in CABO-16S and the simplified SILVA 138.2 is provided in Supplementary Table S1. Construction of the simplified SILVA 138.2 database and the CABO-16S database is provided in the R script "combining_CABO-16S.Rmd." The script can be easily updated with future releases of SILVA or PR2 and allows users to amend CABO-16S with their own custom sequences.

The CABO-16S and simplified SILVA 138.2 training sets were made using recommendations from DECIPHER [37] and based on the IDTAXA algorithm [38]. Details are provided in the R script "training_CABO-16S.Rmd" such that future iterations of the CABO-16S database can be easily trained. Briefly, oversampled groups were randomly subset to 100 sequences before training over three iterations with the LearnTaxa function. Kmer length was set to 8 nt to match RDP and QIIME2 defaults. Note that full-length 16S rRNA reference sequences were used for training; truncation to the amplicon window may slightly improve accuracy (e.g. [39]) but at the cost of potentially generating ambiguity [40]. Thus, we present full-length sequences and perform comparisons from full-length sequences and leave the choice of truncation to users.

Classification of benchmarking datasets

Taxonomic classifications using both CABO-16S and SILVA-132.1 were compared in published 16S rRNA sequences from a broad range of sources, including both mock communities of known bacterial isolates and environmental samples. For all compared samples, the V4–V5 region of the 16S rRNA gene was amplified using archaeal/bacterial primers (515f/926r) and sequenced on the Illumina MiSeq platform. It should be pointed out that the specific extraction and PCR methods may impact the proportion of organellar, bacterial, and archaeal 16S rRNA sequences recovered [41–43]. In environmental datasets with more than five samples, an arbitrary set of subsamples was selected for taxonomic classification comparison. Published metadata and accession numbers or data sources are combined in Supplementary Table S2; full methodological details regarding the generation of the datasets can be found in the original papers.

All downloaded raw sequences were processed identically, except for Needham and Fuhrman [29] data, for which the already-analyzed OTU (operational taxonomic unit) sequences and observation matrices were downloaded and used directly. The full details and parameters used for generating amplicon sequence variants (ASVs) from the raw FASTQ files available on NCBI SRA are provided in "dada2_CABO-16S.Rmd." Briefly, primers were removed using Cutadapt [44], and then sequences were trimmed (240f/200r), merged with a 12-bp overlap, denoised, and aligned using DADA2 [45]. Chimeras were removed and taxonomy was assigned by the IdTaxa function from IDTAXA [38] via the DECIPHER package [37]. IdTaxa parameters were default except for using a confidence threshold of 40%, as this produced classifications most directly comparable to the default QIIME2 classifications based on unpublished comparisons. ASV counts and taxonomic assignments are reported in Supplementary Table S3, and ASV sequences are in Data_ASVs.fa (https://doi.org/10.6084/m9.figshare.27288090). Metrics for comparison of CABO-16S with SILVA 138.2 are provided in "benchmarking_CABO-16S.Rmd" with the anticipation that benchmarking can be easily done by users who amend their own custom sequences to CABO-16S or with the future releases of SILVA and PR2. Future updates to CABO-16S will be hosted at https://github.com/emelissa3/CABO-16S.

Results and discussion

CABO-16S is a practical unification of commonly used 16S rRNA databases to provide a single database that can be easily expanded by users to incorporate database updates or incorporate unpublished sequences. The 389 144 bacterial and 19 213 archaeal 16S rRNA sequences from SILVA 138.2 [34] were used as the initial backbone of CABO-16S database, along with a random 100 eukaryote sequences from SILVA retained as an outgroup. These were combined with 8540 16S rRNA sequences from organellar 16S rRNA genes from the PR2 database [35]. Finally, custom sequences can be combined to maximize resolution into target groups; here we added a set of unpublished full-length 16S rRNA sequences obtained from Sanger sequencing of methane seep sediments and a curated list of representative SEEP-SRB1 sequences [36].

CABO-16S annotates previously unclassified ASVs

CABO-16S was compared against SILVA 138.2 using a combination of previously published datasets representing diverse systems, including both mock communities and environmental samples (Supplementary Table S2). The dataset consists of well-characterized benchmarking sets based on mock communities [46, 47], mammalian guts and residential soils [46], boreal forest soils [48], seagrass roots, leaves, and surrounding sediment [49], deep-sea sediments from cold methane seeps [50], hydrothermal vent sediment [51], marine water with abundant phytoplankton communities [29], and finally, water from a closed basin lake dominated by the microalgae *Picocystis* [52]. The combined dataset consists of 64 402 ASVs, with individual datasets ranging from 45 to 32 090 ASVs.

The CABO-16S database outperformed the unamended SILVA 138.2 database in terms of the total number of ASVs

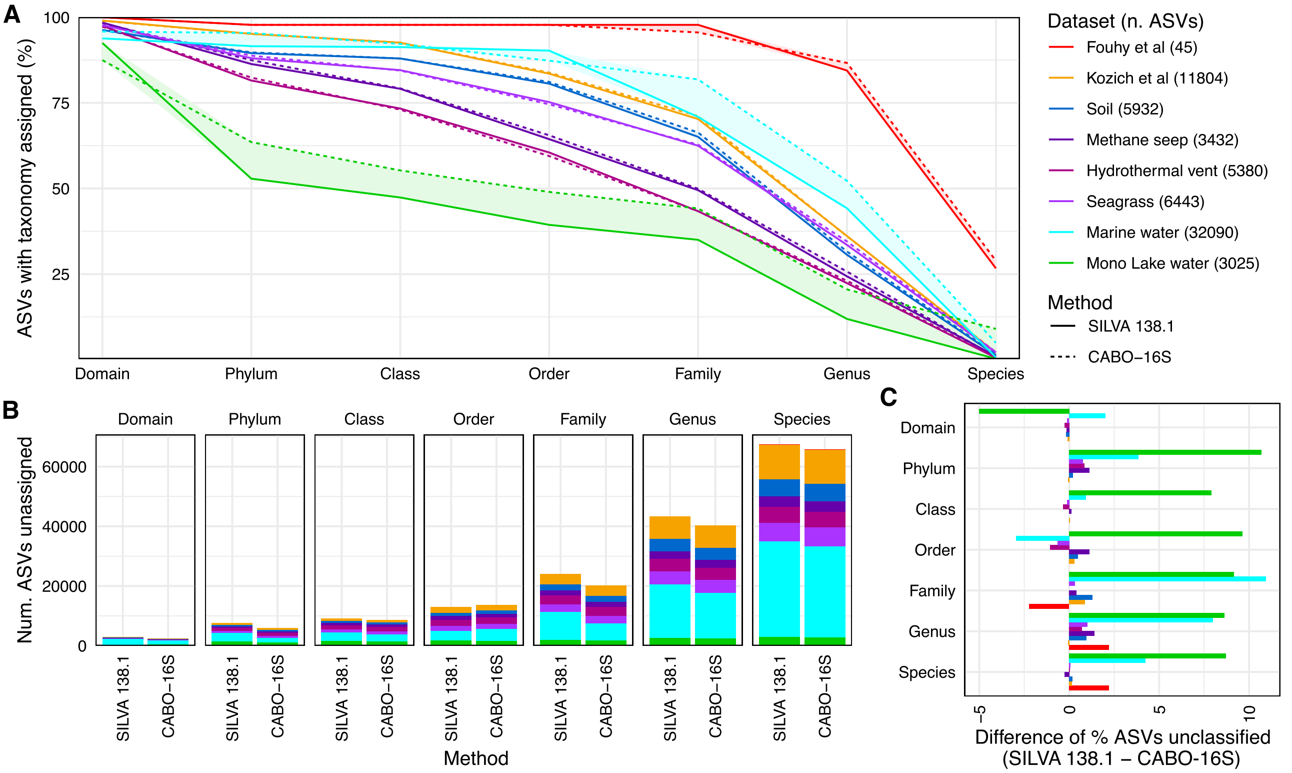


Figure 1. The CABO-16S database improves taxonomic classification in environmental datasets with phototrophic eukaryotes. The Fouhy *et al.* datasets represent mock communities, while the Kozich *et al.* dataset contains mock communities as well as samples from mammalian guts and residential soils. (**A**) Percentage of ASVs classified at a given taxonomic level (*x*-axis) by dataset (colored lines) for SILVA 138.2 (dashed) and CABO-16S (solid). The lines continuously decrease as ASVs lacking annotation at higher ranks, e.g. domain, are by definition also lacking annotation for lower ranks, e.g. species. (**B**) Absolute number of ASVs lacking classification, as in panel (A). (**C**) The difference in percentage unclassified between the two databases. Positive percentages reflect CABO-16S annotating more ASVs than SILVA 138.2, and vice versa for negative percentages.

receiving taxonomic assignment across all taxonomic levels (Fig. 1). The largest differences were in the datasets with the most phototrophic eukaryotes, such as the shallow seagrass environments, marine waters, and lake waters. For example, in the marine water dataset, CABO-16S allowed classification of ~10% more ASVs than SILVA 138.2 at the phylum level (Fig. 1C). Other datasets differed little, suggesting that the inclusion of the PR2 database's organelle 16S sequences did not meaningfully impact the ability of the classifier to continue to accurately predict bacterial and archaeal taxonomy. The only notable exception was at the order level in the marine water dataset, where SILVA 138.2 had slightly higher classification rates than CABO-16S, which we attribute to the increased number of phytoplankton orders from PR2 versus the singular "Chloroplast" label in SILVA. The vast majority of ASVs could not be classified at the species level in either dataset, although this may be partly due to relatively few reference sequences being labeled down to the species level, particularly for microbes endemic to non-human environments.

Both CABO-16S and SILVA classifications reveal a distinction between two types of ambiguity, which can prevent taxonomic annotation. Ambiguity is the most commonly considered form of precision, where a sequence is placed intermediate between two or more reference taxa and thus cannot be assigned to a single taxon at a chosen confidence threshold (40% in this study). IDTAXA and other similar classifiers handle such occurrences by classifying the sequence to the lowest common level of the competing reference taxa, and sometimes adding a prefix of "unclassified" to the conflicted taxonomic rank. Conversely, a sequence may be confidently assigned to a single taxon; however, taxonomy may still be lacking at a given rank if the reference sequence lacks annotation at that rank. Such a scenario affects many uncultured lineages, e.g. the candidate phylum radiation family SR1 (phylum Patescibacteria) has no genus or species assignments in SILVA 138.2, all 121 sequences are annotated only to the family level. Thus, an SR1 ASV lacking a genus classification is not due to classifier uncertainty but rather taxonomic uncertainty. Furthermore, some lineages can include both sources of uncertainty; e.g. in SILVA 138.2, the family Desulfosarcinacae has 53 sequences labeled to the species level, 676 labeled to the genus level, and 345 labeled to the family level. Therefore, Desulfobacteraceae ASVs lacking genus-level annotation could be due to a close similarity to a set of sequences labeled only to the family level (taxonomic ambiguity) or by being indistinguishable from different genera (classifier ambiguity). Thus, we distinguish between the two using IDTAXA's convention of prepending "unclassified_" to situations of classifier ambiguity and an additional convention of prepending "unspecified_" to the lowest taxon level for situations of reference sequence ambiguity.

Of the classified sequences, CABO-16S and SILVA produced similar community compositions in most datasets (Fig. 2 and Supplementary Fig. S1). Indeed, as the non-cyanobacterial portion of CABO-16S archaeal and bacterial sequences is exactly shared with SILVA 138.2, this agreement

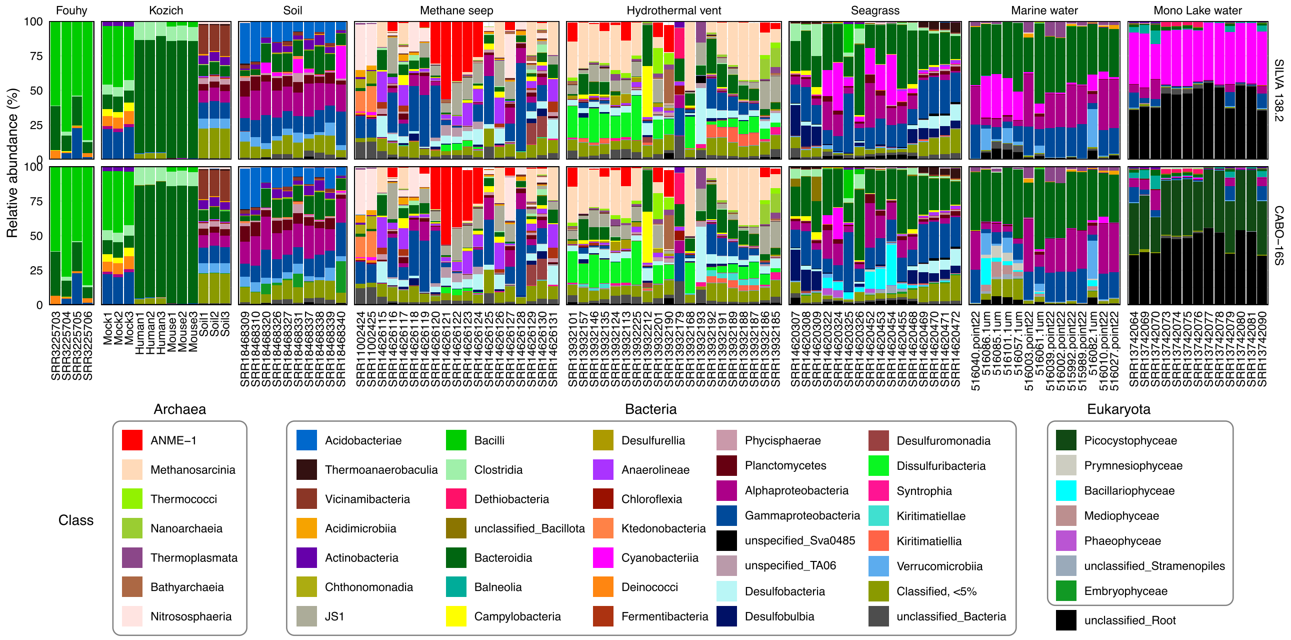


Figure 2. Composition of each dataset with CABO-16S versus original SILVA 138.2. ASVs were aggregated to the class level (colors). Classes with at least 5% in any sample are shown. ASVs that could not be assigned a specific class were similarly aggregated at the lowest annotated rank. Remaining ASVs with <5% relative abundance were grouped into a single category.

is expected. In datasets with little to no plastids (Fouhy *et al.*, Kozich *et al.*, methane seep, and hydrothermal vents), there was good correspondence in the taxonomic composition determined by SILVA and CABO-16S (average $R^2 = 0.99$ at the class level; Supplementary Fig. S2). However, in datasets with phototrophic eukaryotes (e.g. soil, seagrass, marine, and lake water column datasets), the CABO-16S database allowed classification of eukaryotic chloroplasts, which accounted for nearly 50% of the reads in some samples (Mono Lake dataset; Fig. 2) and contributed to more disagreement between the taxonomic composition determined by the different databases (R^2 ranging from 0.30 to 0.89 at the class level). Besides the photosynthetic eukaryotes, relatively few taxonomic assignments were changed at the class level, confirming that the inclusion of PR2 sequences did not negatively impact the taxonomic assignment of non-organellar 16S rRNA sequences (Supplementary Fig. S2). The marine water dataset had by far the greatest phytoplankton diversity of any of the example sets [29]; much of this diversity could be assigned a taxonomic label with CABO-16S. In the Mono Lake dataset, the remaining unclassified diversity could be attributed to phytoplankton mitochondrial sequences by manual blasting versus NCBI. While the current PR2 database includes ~1842 mitochondrial sequences, the vast majority (1782 or 96.7%) belong to Opisthokonta, with only 22 sequences spanning Archaeplastida (plants and many algae). Although mitochondria are not found in all eukaryotic cells [53], we anticipate that the future expansion of PR2 to include more mitochondrial 16S from plant and algal lineages will ameliorate this problem.

Enhanced identification of phototrophic eukaryotes

We tracked ASV classification through each dataset for phototrophic eukaryotes to elucidate points of difference between the databases (Fig. 3). For the coastal marine water dataset [29] undergoing a eukaryotic phytoplankton bloom, SILVA 138.2 was able to classify the bacterial community accurately, but a large fraction of the reads, presumably eukaryotic, were not assigned at the domain level or simply annotated as chloroplast at the family level. However, with the CABO-16S dataset, these same plastid ASV sequences received further taxonomic assignment (Fig. 3). For example, at the order level, ~4700 ASVs were identified as "Chloroplast" by SILVA 138.2, but CABO-16S provided ~3700 of these ASVs with order-level eukaryotic designations (Fig. 3). At the phylum level, CABO-16S provided plastidal taxonomy to just over 6000 ASVs, in contrast to the just over 5000 ASVs identified as cyanobacteria by SILVA 138.2. This dataset includes some non-chloroplast cyanobacteria (Cyanobiaceae), and both databases provide the exact same number of ASVs assigned to Cyanobiaceae ($n = 342$). Notably, the diversity of sequences did not always allow for unambiguous taxonomic assignment to lower levels; i.e. tracing the taxonomic assignment in Fig. 3 reveals that many ambiguities arise at the class or order level. Some taxon ranks include "_X" suffixes, which are placeholder intermediates used by PR2 akin to "*Incertae Sedis*" used in other taxonomies. What taxonomic assignment can be obtained through plastid classification is useful, however, as the major phytoplankton groups (e.g. diatoms, dinoflagellates, cryptophytes, etc.) are distinguished.

Custom sequences increase identification of polyphyletic clades

Addition of custom sequences to CABO-16S can increase taxonomic classification of species not currently included in either the SILVA or PR2 databases. We added sequences obtained from Sanger sequencing of methane seep sediments and a curated list of representative SEEP-SRB1 sequences [36]. SEEP-SRB1 is a polyphyletic clade of sulfate-reducing bacteria [54, 55], including known syntrophic partners of ANME during the anaerobic oxidation of methane (AOM), such as SEEP-SRB1a and SEEP-SRB1g, and other non-syntrophic members (SEEP-SRB1b, SEEP-SRB1c, SEEP-SRB1d, SEEP-SRB1e, and

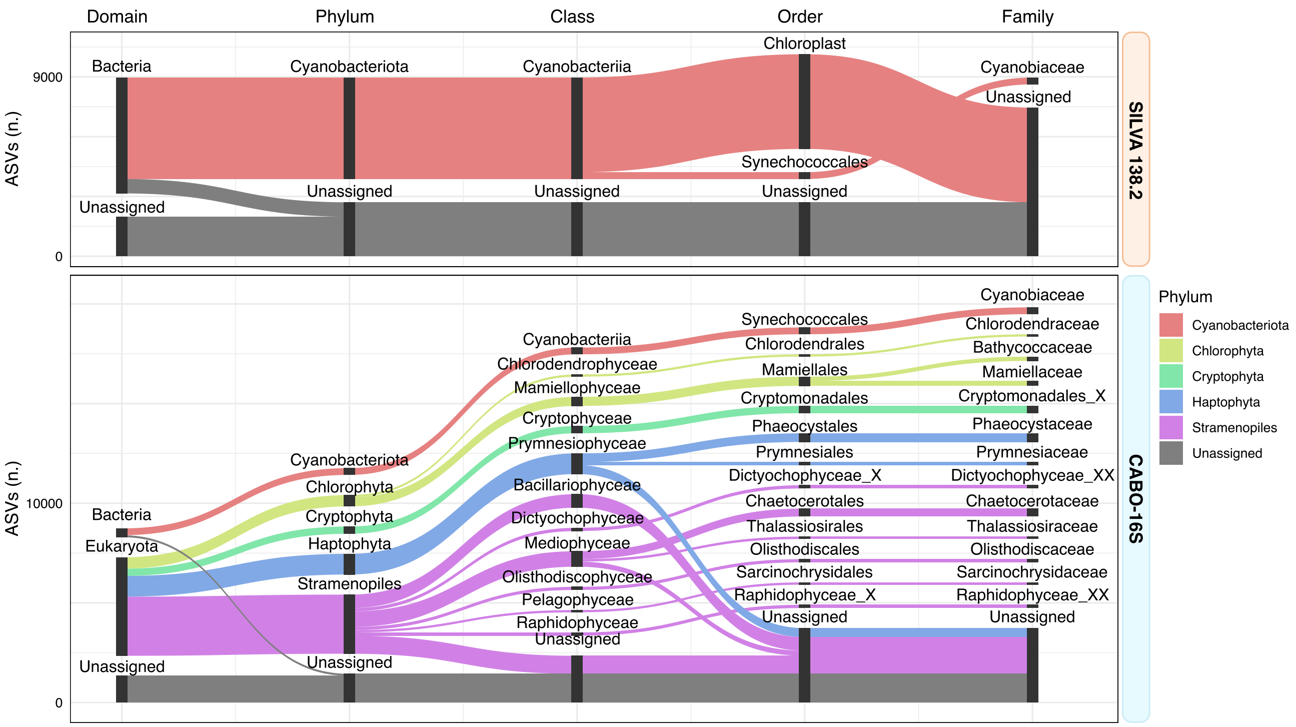


Figure 3. CABO-16S resolves both eukaryotic and bacterial phytoplankton in the coastal marine dataset. Alluvial plots for SILVA 138.2 (top panel) and CABO-16S (bottom panel) tracking ASV classification across taxonomic ranks from domain (leftmost) to family (rightmost) for families with at least 100 ASVs. ASVs related to cyanobacteria or plastid sequences based on either database along with ASVs unassigned at the domain level are shown. For each rank, bars represent the different taxa classified, with the size of each bar scaled to reflect the number of ASVs. The "unassigned" category is the union of "unclassified" and "unspecified" designations to simplify identification of the point where taxonomic assignment stops. White space between bars is for ease of visualization. Flows connecting ranks are colored based on the phylum-level classification of those ASVs.

SEEP-SRB1f). Although currently identified as a genus-level clade in SILVA 138.2, this is an overly simplified grouping of these organisms. Indeed, while some members, such as SEEP-SRB1g and SEEP-SRB1c, have been described as species-level clades, others such as SEEP-SRB1a are more accurately described as genus-level clades. Further complicating SEEP-SRB taxonomy is the asymmetric phylogenetic distance between SEEP-SRB1 subgroups; for example, SEEP-SRB1g and SEEP-SRB1a may reside in different orders based on genomic trees [36, 56–58]. While rectifying phylogenetic distance with taxonomic classification is beyond the scope of this work, we note that such conflicts between historical naming conventions are unfortunately common in environmental microbiology and difficult to resolve. However, expanding databases with precisely named groups offers a means to circumvent these discrepancies. Thus, we added the representative sequences for SEEP-SRB subgroups as a "species" of SEEP-SRB1, except for SRB1g, which was added as a "species" of *Desulfosudis*, the taxonomic designation from SILVA 138.2 for sequences most similar to SRB1g.

The inclusion of these additional SEEP-SRB sequences in the CABO-16S database resolved a portion of the environmental SRB1 group ASVs to their respective subgroups (Fig. 4). In the methane seep and hydrothermal vent datasets, the improved resolution revealed varying distributions of the different SEEP-SRB1 subgroups (Fig. 4A). In the vent dataset, only a subset of samples harbored the syntrophic Seep-SRB1a versus the non-syntrophic Seep-SRB1d, a distinction that was not resolvable with the default SILVA 138.2 database. Further differences in taxonomic assignment become clear when tracking how ASVs annotated in SILVA as SEEP-SRB1, unclassified_Desulfosarcinaceae, or *Desulfosudis* are classified by CABO-16S compared to SILVA 138.2 in methane seep and hydrothermal vent datasets (Fig. 4B). While a relatively small proportion of the total ASVs from the two combined datasets were differently classified by CABO-16S (Fig. 4B), based on Fig. 4A, the differences in classification are significant in particular environments such as the sedimented hydrothermal vents included in this study. Interestingly, a few ASVs classified as Desulfosarcinaceae with SILVA 138.2 were unclassified at higher ranks with CABO-16S (e.g. unclassified_Desulfobacterales) or differently classified (Fig. 4B). We attribute such variance to inter-run variance in the IDTAXA algorithm, as it randomly subsamples kmers for each run, and so a small portion of ASVs at the edge of classification confidence threshold receive different classifications between runs [38]. In support of this observation, flipping the analysis to include ASVs classified as SEEP-SRB1, Desulfosarcinaceae, or LCP-80 with CABO-16S produced a similar result of overall agreement, with some ASVs classified with CABO-16S as Desulfosarcinaceae being unclassified with SILVA 138.2.

Challenges associated with adding custom sequences

While annotation for Seep-SRB1 subtypes could be achieved by the addition of known sequences with specific annotation, there remains a discrepancy between taxonomy, the hierarchical nomenclature system, and phylogeny, the evolutionary history of these clades. Other groups may require even more complicated approaches than the one we employed for SEEP-SRB1. The genera *Shigella* and *Escherichia* are emblematic of

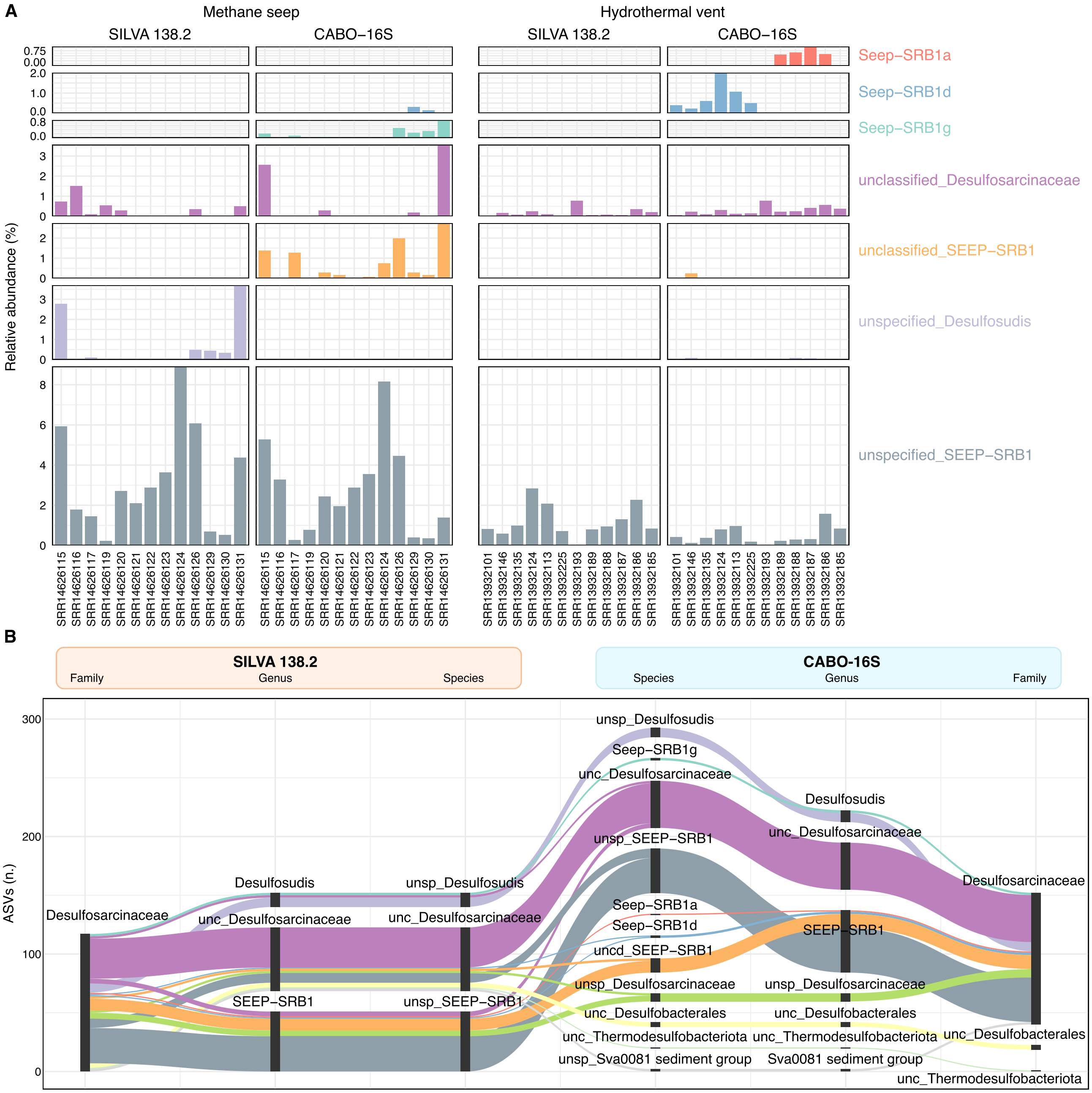


Figure 4. CABO-16S allowed classification of SEEP-SRB1 at higher resolution by adding in specific sequences. (**A**) Relative abundance of SEEP-SRB1 and related taxa (colors, also subplot rows) at the lowest classified level. For the methane seep and hydrothermal vent datasets, the left subpanels show SILVA 138.2 classifications versus CABO-16S classifications in the right subpanels. *Y*-axis is in % of each sample's total number of reads. (**B**) Alluvial plot showing classification of the same ASV sequences across databases. Each column is a different rank (family through species), with flows colored by the species assigned at the species rank with CABO-16S. Flow height reflects the number of ASVs. Abbreviations: unc, unclassified; unsp, unspecified; and sed, sediment. Note that we distinguish between lack of annotation due to classification conflicts (unclassified) versus due to incomplete annotation of reference taxa (unspecified), discussed above in the context of Fig. 1.

this conflict, as both are deeply convolved evolutionarily, yet the taxonomic legacy has continued to complicate reference database hierarchies [59, 60]. Other environmental groups like *Synechococcus* have similarly posed challenges to rectifying taxonomy and phylogeny for decades [61–63]. For such groups, adding sequences with a specific, phylogenetically correct hierarchy is unlikely to improve classification as the LCA method of resolving ambiguities assumes all sequences share the same hierarchy. Thus, all existing sequences would require similar reclassification following the desired phylogenetic framework and necessitate additional curation to ensure compatibility of the new taxonomic hierarchy with sequence similarity. Ultimately, the feasibility of rectifying phylogeny and taxonomy is limited by the signal embedded in the 16S rRNA gene, and while genome-based phylogenies and 16S rRNA phylogenies largely agree, they are not identical [64].

An additional barrier to improving resolution is errors or inconsistencies in taxonomic assignments, i.e. similar sequences with conflicting names; such errors are estimated to account for 1.5%–17% of sequences in SILVA [65, 66]. Based on the assumption that the majority of sequences are correctly and consistently labeled, approaches like IDTAXA incorporate tools to identify and drop individual sequences that conflict with the majority of similarly named sequences during training [38], and stand-alone tools also exist [66]. However, such approaches work best for taxa represented by many sequences, which is not always the case for environmental lineages in need of improved resolution. Classifier resolution and accuracy can also be improved by constraining the database to include only microbes specific to the habitat sampled as has been successfully done for many animal microbiomes [40, 67]. Such habitat-specific training sets are undoubtedly the best for focused research on a specific system. However, understanding the environmental context of specific taxa with broad distributions, like SEEP-SRB, necessitates approaches that maximize the resolution possible with general databases like SILVA.

Conclusion

CABO-16S successfully combines bacterial and archeal 16S rRNA sequences from SILVA 138.2 and organellar 16S rRNA sequences from the PR2 database with customly selected sequences, resulting in increased taxonomic assignment of ASVs compared to SILVA 138.2. Specifically, with the addition of plastidal sequences from PR2, CABO-16S excels at identification of phototrophic eukaryotes in marine and lake waters without additional sequencing of both 16S and 18S primer sets. Although some 16S sequences, such as mitochondria from plants and algae, are still minimal and may impact classification of specific environments, CABO-16S reduces the number of unassigned phototrophs such that the remaining abundant sequence can be rapidly quarried. CABO-16S is a ready-to-use database; however, with the extensive documentation and provided scripts, we also provide a tool that allows for the easy addition of custom sequences and benchmarking methods necessary to test those additions. This was verified with the addition of sequences from the polyphyletic clade of SEEP-SRB1, where we saw increased taxonomic differentiation in hydrothermal vent sediment samples. This could help determine the likelihood of syntrophy in particular environments and increase our understanding of the communities that contribute to AOM. Although the addition of custom sequences must be done cautiously, considering the number of unspecified sequences within SILVA and the difficulty of constraining polyphyletic clades to the current taxonomy structure, this function of CABO-16S gives users the freedom to customize 16S taxonomic classification and potentially increase the understanding of specific environments. Finally, CABO-16S provides a framework that can be easily updated with the release of future versions of the SILVA and PR2 databases.

Acknowledgements

The authors would like to thank all the members of the Orphan lab for useful discussions and Dmitri Bilyk for logo design. We acknowledge members of the Agouron International Geobiology Course from 2021 and 2022 and Alex Sessions in particular for encouraging work at Mono Lake, which initiated some of this research. We would like to thank the Agouron Institute and the Simons Foundation for funding the International Geobiology Course. Finally, we would like to thank the SILVA and PR2 teams for their work creating and updating those databases.

Author contributions: Eryn M. Eitel (Conceptualization, Methodology, Software Analysis Validation, Visualization, Writing), Daniel R. Utter (Conceptualization, Methodology, Software Analysis Validation, Visualization, Writing), Stephanie A. Connon (Data Curation), Victoria J. Orphan (Supervision, Funding acquisition), and Ranjani Murali (Data Curation, Supervision, Funding acquisition).

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

This work is supported by the National Science Foundation [2126631] to D.U. as part of the Ocean Sciences Postdoctoral Research Fellowship; the Simons Foundation [602126] to E.E. as part of the Postdoctoral Fellowships in Marine Microbial Ecology; and the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research [DE-SC0020373].

Data availability

Custom sequences, scripts, and other files are hosted permanently on Figshare (https://doi.org/10.6084/m9.figshare.27288090). Future updates to CABO-16S will be made available at https://github.com/emelissa3/CABO-16S.

References

1. Hooper DU, Chapin FS III, Ewel JJ *et al.* Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecol Monogr* 2005;**75**:3–35. https://doi.org/10.1890/04-0922
2. Tilman D, Lehman C. Human-caused environmental change: impacts on plant diversity and evolution. *Proc Natl Acad Sci USA* 2001;**98**:5433–40. https://doi.org/10.1073/pnas.091093198
3. Pecl GT, Araújo MB, Bell JD *et al.* Biodiversity redistribution under climate change: impacts on ecosystems and human well-being. *Science* 2017;**355**:eaai9214. https://doi.org/10.1126/science.aai9214
4. Johnson CN, Balmford A, Brook BW *et al.* Biodiversity losses and conservation responses in the Anthropocene. *Science* 2017;**356**:270–5. https://doi.org/10.1126/science.aam9317
5. Cavicchioli R, Ripple WJ, Timmis KN *et al.* Scientists' warning to humanity: microorganisms and climate change. *Nat Rev Microbiol* 2019;**17**:569–86. https://doi.org/10.1038/s41579-019-0222-5
6. Graham EB, Knelman JE, Schindlbacher A *et al.* Microbes as engines of ecosystem function: when does community structure enhance predictions of ecosystem processes? *Front Microbiol* 2016;**7**:214. https://doi.org/10.3389/fmicb.2016.00214
7. Fuhrman JA. Microbial community structure and its functional implications. *Nature* 2009;**459**:193–9. https://doi.org/10.1038/nature08058
8. Van Der Heijden MGA, Bardgett RD, Van Straalen NM. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecol Lett* 2008;**11**:296–310. https://doi.org/10.1111/j.1461-0248.2007.01139.x
9. Fuhrman JA, Cram JA, Needham DM. Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol* 2015;**13**:133–46. https://doi.org/10.1038/nrmicro3417

10. Agler MT, Ruhe J, Kroll S *et al.* Microbial hub taxa link host and abiotic factors to plant microbiome variation. *PLoS Biol* 2016;**14**:e1002352. https://doi.org/10.1371/journal.pbio.1002352
11. Ley RE, Lozupone CA, Hamady M *et al.* Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* 2008;**6**:776–88. https://doi.org/10.1038/nrmicro1978
12. Thompson LR, Sanders JG, McDonald D *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 2017;**551**:457–63. https://doi.org/10.1038/nature24621
13. Sunagawa S, Coelho LP, Chaffron S *et al.* Structure and function of the global ocean microbiome. *Science* 2015;**348**:1261359. https://doi.org/10.1126/science.1261359
14. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. *Genome Med* 2016;**8**:51. https://doi.org/10.1186/s13073-016-0307-y
15. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* 1977;**74**:5088–90. https://doi.org/10.1073/pnas.74.11.5088
16. Amaral-Zettler LA, McCliment EA, Ducklow HW *et al.* A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One* 2009;**4**:e6372. https://doi.org/10.1371/journal.pone.0006372
17. Schoch CL, Seifert KA, Huhndorf S *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proc Natl Acad Sci USA* 2012;**109**:6241–6. https://doi.org/10.1073/pnas.1117018109
18. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* 2016;**18**:1403–14. https://doi.org/10.1111/1462-2920.13023
19. Yeh Y-C, McNichol J, Needham DM *et al.* Comprehensive single-PCR 16S and 18S rRNA community analysis validated with mock communities, and estimation of sequencing bias against 18S. *Environ Microbiol* 2021;**23**:3240–50. https://doi.org/10.1111/1462-2920.15553
20. Kittelmann S, Seedorf H, Walters WA *et al.* Simultaneous amplicon sequencing to explore co-occurrence patterns of bacterial, archaeal and eukaryotic microorganisms in rumen microbial communities. *PLoS One* 2013;**8**:e47879. https://doi.org/10.1371/journal.pone.0047879
21. Obiol A, Giner CR, Sánchez P *et al.* A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Mol Ecol Resour* 2020;**20**:718–31. https://doi.org/10.1111/1755-0998.13147
22. Klappenbach JA, Saxman PR, Cole JR *et al.* rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res* 2001;**29**:181–4. https://doi.org/10.1093/nar/29.1.181
23. Pan P, Gu Y, Sun D-L *et al.* Microbial diversity biased estimation caused by intragenomic heterogeneity and interspecific conservation of 16S rRNA genes. *Appl Environ Microbiol* 2023;**89**:e02108–22. https://doi.org/10.1128/aem.02108-22
24. Zhu F, Massana R, Not F *et al.* Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* 2005;**52**:79–92. https://doi.org/10.1016/j.femsec.2004.10.006
25. Yarimizu K, Sildever S, Hamamoto Y *et al.* Development of an absolute quantification method for ribosomal RNA gene copy numbers per eukaryotic single cell by digital PCR. *Harmful Algae* 2021;**103**:102008. https://doi.org/10.1016/j.hal.2021.102008
26. de Vargas C, Audic S, Henry N *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* 2015;**348**:1261605. https://doi.org/10.1126/science.1261605
27. Bedoshvili YD, Popkova TP, Likhoshway YV. Chloroplast structure of diatoms of different classes. *Cell Tiss Biol* 2009;**3**:297–310. https://doi.org/10.1134/S1990519X09030122
28. Tomas CR. *Identifying Marine Phytoplankton*. San Diego, CA: Academic Press, 1997.
29. Needham DM, Fuhrman JA. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat Microbiol* 2016;**1**:16005. https://doi.org/10.1038/nmicrobiol.2016.5
30. Shi XL, Lepère C, Scanlan DJ *et al.* Plastid 16S rRNA gene diversity among eukaryotic picophytoplankton sorted by flow cytometry from the South Pacific Ocean. *PLoS One* 2011;**6**:e18979. https://doi.org/10.1371/journal.pone.0018979
31. Hamilton M, Mascioni M, Hehenberger E *et al.* Spatiotemporal variations in antarctic protistan communities highlight phytoplankton diversity and seasonal dominance by a novel cryptophyte lineage. *mBio* 2021;**12**:e02973–21. https://doi.org/10.1128/mBio.02973-21
32. Choi CJ, Jimenez V, Needham DM *et al.* Seasonal and geographical transitions in eukaryotic phytoplankton community structure in the Atlantic and Pacific oceans. *Front Microbiol* 2020;**11**:542372. https://doi.org/10.3389/fmicb.2020.542372
33. Bachy C, Sudek L, Choi CJ *et al.* Phytoplankton surveys in the Arctic Fram Strait demonstrate the tiny eukaryotic alga *Micromonas* and other picoprasinophytes contribute to deep sea export. *Microorganisms* 2022;**10**:961.
34. Quast C, Pruesse E, Yilmaz P *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**:D590–6. https://doi.org/10.1093/nar/gks1219
35. Guillou L, Bachar D, Audic S *et al.* The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* 2013;**41**:D597–604. https://doi.org/10.1093/nar/gks1160
36. Murali R, Yu H, Speth DR *et al.* Physiological potential and evolutionary trajectories of syntrophic sulfate-reducing bacterial partners of anaerobic methanotrophic archaea. *PLoS Biol* 2023;**21**:e3002292. https://doi.org/10.1371/journal.pbio.3002292
37. Wright ES. Using DECIPHER v2.0 to analyze big biological sequence data in R. *The R Journal* 2016;**8**:7. https://doi.org/10.32614/RJ-2016-025
38. Murali A, Bhargava A, Wright ES. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* 2018;**6**:140. https://doi.org/10.1186/s40168-018-0521-5
39. Werner JJ, Koren O, Hugenholtz P *et al.* Impact of training sets on classification of high-throughput bacterial 16S rRNA gene surveys. *ISME J* 2012;**6**:94–103. https://doi.org/10.1038/ismej.2011.82
40. Escapa IF, Huang Y, Chen T *et al.* Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. *Microbiome* 2020;**8**:65. https://doi.org/10.1186/s40168-020-00841-w
41. Feinstein Larry M, Sul Woo J, Blackwood Christopher B. Assessment of bias associated with incomplete extraction of microbial DNA from soil. *Appl Environ Microbiol* 2009;**75**:5428–33. https://doi.org/10.1128/AEM.00120-09
42. Beckers B, Op De Beeck M, Thijs S *et al.* Performance of 16S rDNA primer pairs in the study of rhizosphere and endosphere bacterial microbiomes in metabarcoding studies. *Front Microbiol* 2016;**7**:650.
43. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. *eLife* 2019;**8**:e46923. https://doi.org/10.7554/eLife.46923
44. Martin M . Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**:3.
45. Callahan BJ, McMurdie PJ, Rosen MJ *et al.* DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;**13**:581–3. https://doi.org/10.1038/nmeth.3869
46. Kozich JJ, Westcott SL, Baxter NT *et al.* Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* 2013;**79**:5112–20. https://doi.org/10.1128/AEM.01043-13

47. Fouhy F, Clooney AG, Stanton C *et al.* 16S rRNA gene sequencing of mock microbial populations—impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiol* 2016;**16**:123. https://doi.org/10.1186/s12866-016-0738-z
48. Porter TM, Smenderovac E, Morris D *et al.* All boreal forest successional stages needed to maintain the full suite of soil biodiversity, community composition, and function following wildfire. *Sci Rep* 2023;**13**:7978. https://doi.org/10.1038/s41598-023-30732-7
49. Kardish MR, Stachowicz JJ. Local environment drives rapid shifts in composition and phylogenetic clustering of seagrass microbiomes. *Sci Rep* 2023;**13**:3673. https://doi.org/10.1038/s41598-023-30194-x
50. Marlow JJ, Hoer D, Jungbluth SP *et al.* Carbonate-hosted microbial communities are prolific and pervasive methane oxidizers at geologically diverse marine methane seep sites. *Proc Natl Acad Sci USA* 2021;**118**:e2006857118. https://doi.org/10.1073/pnas.2006857118
51. Speth DR, Yu FB, Connon SA *et al.* Microbial communities of Auka hydrothermal sediments shed light on vent biogeography and the evolutionary history of thermophily. *ISME J* 2022;**16**:1750–64. https://doi.org/10.1038/s41396-022-01222-x
52. Phillips AA, Wu F, Sessions AL. Sulfur isotope analysis of cysteine and methionine via preparatory liquid chromatography and elemental analyzer isotope ratio mass spectrometry. *Rapid Commun Mass Spectrom* 2021;**35**:e9007. https://doi.org/10.1002/rcm.9007
53. Karnkowska A, Vacek V, Zubáčová Z *et al.* A eukaryote without a mitochondrial organelle. *Curr Biol* 2016;**26**:1274–84. https://doi.org/10.1016/j.cub.2016.03.053
54. Schreiber L, Holler T, Knittel K *et al.* Identification of the dominant sulfate-reducing bacterial partner of anaerobic methanotrophs of the ANME-2 clade. *Environ Microbiol* 2010;**12**:2327–40. https://doi.org/10.1111/j.1462-2920.2010.02275.x
55. Knittel K, Boetius A, Lemke A *et al.* Activity, distribution, and diversity of sulfate reducers and other bacteria in sediments above gas hydrate (Cascadia margin, Oregon). *Geomicrobiol J* 2003;**20**:269–94. https://doi.org/10.1080/01490450303896
56. Parks DH, Chuvochina M, Chaumeil P-A *et al.* A complete domain-to-species taxonomy for bacteria and archaea. *Nat Biotechnol* 2020;**38**:1079–86. https://doi.org/10.1038/s41587-020-0501-8
57. Skennerton CT, Chourey K, Iyer R *et al.* Methane-fueled syntrophy through extracellular electron transfer: uncovering the genomic traits conserved within diverse bacterial partners of anaerobic methanotrophic archaea. *mBio* 2017;**8**:e00530-17. https://doi.org/10.1128/mBio.00530-17
58. Metcalfe KS, Murali R, Mullin SW *et al.* Experimentally-validated correlation analysis reveals new anaerobic methane oxidation partnerships with consortium-level heterogeneity in diazotrophy. *ISME J* 2021;**15**:377–96. https://doi.org/10.1038/s41396-020-00757-1
59. Escobar-Páramo P, Giudicelli C, Parsot C *et al.* The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. *J Mol Evol* 2003;**57**:140–8. https://doi.org/10.1007/s00239-003-2460-3
60. Parks DH, Chuvochina M, Reeves PR *et al.* Reclassification of*Shigella*species as later heterotypic synonyms of *Escherichia coli*in the Genome Taxonomy Database. bioRxiv, https://doi.org/10.1101/2021.09.22.461432, 22 September 2021, preprint: not peer reviewed.
61. Robertson BR, Tezuka N, Watanabe MM. Phylogenetic analyses of *Synechococcus* strains (cyanobacteria) using sequences of 16S rDNA and part of the phycocyanin operon reveal multiple evolutionary lines and reflect phycobilin content. *Int J Syst Evol Microbiol* 2001;**51**:861–71. https://doi.org/10.1099/00207713-51-3-861
62. Dvořák P, Casamatta DA, Poulíčková A *et al.* *Synechococcus*: 3 billion years of global dominance. *Mol Ecol* 2014;**23**:5538–51. https://doi.org/10.1111/mec.12948
63. Salazar VW, Tschoeke DA, Swings J *et al.* A new genomic taxonomy system for the *Synechococcus* collective. *Environ Microbiol* 2020;**22**:4557–70. https://doi.org/10.1111/1462-2920.15173
64. Parks DH, Chuvochina M, Waite DW *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;**36**:996–1004. https://doi.org/10.1038/nbt.4229
65. Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases. *PeerJ* 2018;**6**:e5030. https://doi.org/10.7717/peerj.5030
66. Kozlov AM, Zhang J, Yilmaz P *et al.* Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Res* 2016;**44**:5022–33. https://doi.org/10.1093/nar/gkw396
67. Newton I, Roeselers G. The effect of training set on the classification of honey bee gut microbiota using the naïve Bayesian classifier. *BMC Microbiol* 2012;**12**:221. https://doi.org/10.1186/1471-2180-12-221

Received: November 27, 2024. **Revised:** April 22, 2025. **Editorial Decision:** April 24, 2025. **Accepted:** May 11, 2025