AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

Evaluating and Improving Small Subunit rRNA PCR Primer Coverage for Bacteria, Archaea, and Eukaryotes Using Metagenomes from Global Ocean Surveys

Jesse McNichol,[a] Paul M. Berube,[b] Steven J. Biller,[b,c] Jed A. Fuhrman[a]

[a]Department of Biological Sciences, University of Southern California, Los Angeles, California, USA
[b]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[c]Department of Biological Sciences, Wellesley College, Wellesley, Massachusetts, USA

ABSTRACT Small subunit rRNA (SSU rRNA) amplicon sequencing can quantitatively and comprehensively profile natural microbiomes, representing a critically important tool for studying diverse global ecosystems. However, results will only be accurate if PCR primers perfectly match the rRNA of all organisms present. To evaluate how well marine microorganisms across all 3 domains are detected by this method, we compared commonly used primers with >300 million rRNA gene sequences retrieved from globally distributed marine metagenomes. The best-performing primers compared to 16S rRNA of bacteria and archaea were 515Y/926R and 515Y/806RB, which perfectly matched over 96% of all sequences. Considering cyanobacterial and chloroplast 16S rRNA, 515Y/926R had the highest coverage (99%), making this set ideal for quantifying marine primary producers. For eukaryotic 18S rRNA sequences, 515Y/926R also performed best (88%), followed by V4R/V4RB (18S rRNA specific; 82%)—demonstrating that the 515Y/926R combination performs best overall for all 3 domains. Using Atlantic and Pacific Ocean samples, we demonstrate high correspondence between 515Y/926R amplicon abundances (generated for this study) and metagenomic 16S rRNA (median R^2 = 0.98, n = 272), indicating amplicons can produce equally accurate community composition data compared with shotgun metagenomics. Our analysis also revealed that expected performance of all primer sets could be improved with minor modifications, pointing toward a nearly completely universal primer set that could accurately quantify biogeochemically important taxa in ecosystems ranging from the deep sea to the surface. In addition, our reproducible bioinformatic workflow can guide microbiome researchers studying different ecosystems or human health to similarly improve existing primers and generate more accurate quantitative amplicon data.

IMPORTANCE PCR amplification and sequencing of marker genes is a low-cost technique for monitoring prokaryotic and eukaryotic microbial communities across space and time but will work optimally only if environmental organisms match PCR primer sequences exactly. In this study, we evaluated how well primers match globally distributed short-read oceanic metagenomes. Our results demonstrate that primer sets vary widely in performance, and that at least for marine systems, rRNA amplicon data from some primers lack significant biases compared to metagenomes. We also show that it is theoretically possible to create a nearly universal primer set for diverse saline environments by defining a specific mixture of a few dozen oligonucleotides, and present a software pipeline that can guide rational design of primers for any environment with available meta'omic data.

KEYWORDS amplicon sequencing, marine microbiology, metagenomics, oceanography, Snakemake

Citation McNichol J, Berube PM, Biller SJ, Fuhrman JA. 2021. Evaluating and improving small subunit rRNA PCR primer coverage for bacteria, archaea, and eukaryotes using metagenomes from global ocean surveys. mSystems 6:e00565-21. https://doi.org/10.1128/mSystems.00565-21.

Editor Jack A. Gilbert, University of California San Diego

Address correspondence to Jesse McNichol, mcnichol@alum.mit.edu.

Received 6 May 2021
Accepted 7 May 2021
Published 1 June 2021

Amplicon sequencing is a powerful tool for understanding microbial community composition and dynamics in the oceans and other ecosystems (1), but the PCR amplification step is potentially biased due to both technical issues during amplification and mismatches to organisms found in natural environments (2–6). Despite these concerns, PCR amplicon sequencing retains several advantages that make it desirable to investigate and correct biases. First, it is a high-throughput and low-cost technique, making it suitable for large numbers of samples, e.g., for global surveys of sediment, water, animal-associated, and other microbial communities (1).

Second, the targeted nature of the PCR assay means relatively small numbers of sequences are sufficient for detecting rare organisms even when we have no genomes from any of their relatives, due to the conserved nature of the molecule and the existence of comprehensive small subunit (SSU) rRNA sequence databases. Being able to quantify the abundance and dynamics of rare organisms is important for understanding ecosystem function since many rare bacteria have impacts far greater than their low abundances might imply (7, 8). While PCR assays are targeted, we note that this does not imply they need to be taxonomically restricted. For example, there are some primer binding sites in the SSU rRNA molecule that are nearly universally conserved between *Archaea*, *Bacteria*, and *Eukarya* (discussed further below).

Third, using untargeted metagenomic sequencing for taxonomic profiling still has a number of significant disadvantages versus amplicon sequencing. Direct taxonomic assignment of randomly sheared metagenomic reads by recruitment to reference genomes has the potential to assign taxonomy at a very high resolution. However, the best currently available genome databases still recruit only ~40% of metagenomic reads for marine prokaryotes (9), meaning a significant fraction of untargeted metagenomic data sets are taxonomically uncharted. This problem is even more acute for environmental eukaryotes with large amounts of noncoding genomic DNA and fewer genomic references. It is also possible to extract SSU rRNA (or other marker genes) from metagenomes to generate a taxonomic profile. However, SSU rRNA fragments typically represent only a small fraction of a given data set (~0.1%), so this is a far more costly way to obtain a comprehensive community profile. In addition, because metagenomically retrieved marker genes are randomly sheared fragments covering both conserved and hypervariable regions and are in most cases too short to provide a unique match to reference sequences, they must be assembled or otherwise clustered using nondeterministic algorithms. Because of this, the taxonomic resolution of short-read metagenomic marker gene data is limited (e.g., near family or genus level for the best-known conserved markers like rRNA). In contrast, with rRNA amplicons, modern "denoising" algorithms produce exact amplicon sequence variants (ASVs) that are stable biogeographic markers that can be intercompared without reanalysis (10). The resulting data sets are also relatively simple to analyze since they consist of a single gene region and can be comprehensively classified with databases such as SILVA or RDP (11, 12).

Recent studies have shown that PCR amplicon sequencing of mock microbial communities can recover known relative and absolute abundances extremely well and thus provides accurate quantification of natural gene copy abundances (3, 13, 14). In turn, this accuracy allows amplicon data to serve as a ground-truth for modeling/ecological studies by quantifying community members and their dynamics. While well-designed primers have the potential to recover truly quantitative data, these studies also underscored the critical fact that a single mismatch between the primer and template sequences can have a dramatic effect on the measured community composition in complex natural mixtures. For example, a single internal mismatch in the popular Earth Microbiome Project primer set caused an ~10-fold bias against the most common bacteria in seawater (SAR11 cluster), and terminal 3′ mismatches can completely prevent amplification (2, 3, 6, 13). Since the extent of this bias is not easily predicted, PCR primers should incorporate degenerate bases or specific oligonucleotide variants so that all targeted organisms are perfectly matched without overly diluting the common perfect matches in primer mixtures. This is especially critical for abundant taxa

such as SAR11 as distortions in their relative abundances will skew the remainder of the community, but it is also important to consider for rare taxa which might have essential biogeochemical or ecological roles.

Previous studies have shown high coverage of natural taxa is possible by designing moderately degenerate primers without sacrificing specificity or PCR efficiency (3, 13). This primer design was accomplished by comparing oligonucleotide sequences to a SSU rRNA database such as SILVA or RDP (11, 12) and then checking for mismatches to organisms known to be abundant in the environment of interest. This approach led to marked improvements in primer design, for example by Apprill et al. (2) and Parada et al. (3) who reported that small modifications to existing primers could better quantify the dominant marine taxa SAR11 and *Thaumarchaea*. However, in these primer evaluations, the reliance on full-length references and giving each sequence in a database equal weight can lead to a distorted perspective of the actual extent of matches and mismatches expected in real samples since they do not take into account the highly unequal abundances in nature. In addition, some environments may have abundant taxa poorly represented or unrepresented in these curated reference databases.

Wear et al. (6) studied the effect of these potential biases empirically by testing 4 primer sets currently in broad use by marine microbiologists on a 16S rRNA mock microbial community derived from natural seawater communities near Santa Barbara, CA. These authors tested the effect of their analysis pipeline on recovery of mock community sequences for different primers and found their primer-pipeline combination had several sequence-specific biases, in some cases due to a primer mismatch. While this was an important step toward cross-comparing primers in an oceanographic context, Wear et al. (6) noted that results are specific to their mock microbial community, analyzed with their particular pipeline, and thus it remains unknown how representative their results are for other ecosystems and other pipelines. More specifically, it is currently unknown how many primer-mismatched rRNA sequences occur across diverse oceanographic environments, a key piece of information for designing field measurements and interpreting results.

To more fully evaluate the extent to which primers currently in broad use by the marine microbiology community perfectly match naturally occurring sequences, we developed a workflow to conduct *in silico* primer evaluations with metagenomic SSU rRNA fragments, under the general assumption that metagenomes have fewer methodological biases than PCR. While it is already known that some primers have better coverage for abundant marine organisms than do others (3, 13), a precise quantitative comparison of these primers with others that are in broad use has not yet been conducted.

We developed a new software pipeline for analyzing globally distributed marine metagenomic data sets (see Table S1 at https://osf.io/gr4nc/) to make a quantitative and objective evaluation of PCR primers for pelagic marine ecosystems and to suggest specific improvements based on this evidence. We further experimentally compared the quantitative performance of SSU rRNA amplicon sequencing, performed for this study, against the published shotgun metagenomic data from the same sample DNA from the BioGEOTRACES study (20). Our pipeline also allowed us to make intercomparisons between primer sets often not considered together. For example, we compared the performance of two "universal" (3-domain) PCR primers that amplify both 16S and 18S rRNA with primers designed to specifically target only 16S or 18S rRNA and exclude the other molecule.

Because our software pipeline is broadly applicable for any environment where metagenomic data exist, it will allow investigators to design environment-specific PCR probes that recover as many (or as few) targeted organisms as desired. It is also flexible, allowing design or improvement of primers from different variable regions either to retain stable intercomparisons with existing data or to target a region of the SSU rRNA molecule that has better taxonomic resolution for taxa of interest (15). Our approach has the added advantage of providing a quantitative, evidence-based framework to identify which taxa may have been affected by biases in existing data

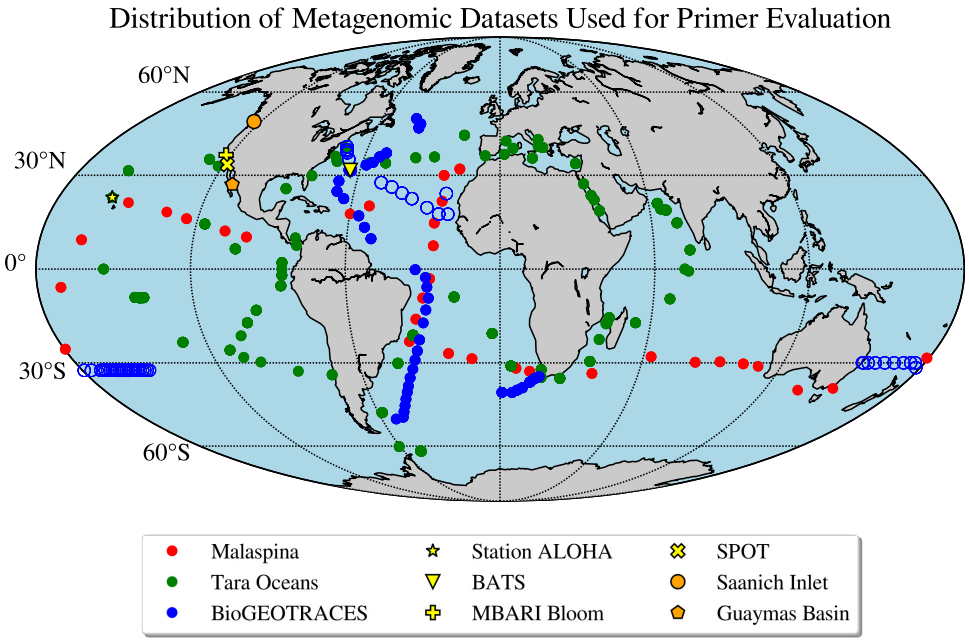

FIG 1 Distribution of metagenomic samples used in this study. BioGEOTRACES samples that were used in the metagenome/amplicon intercomparison are noted with open circles (GA03, North Atlantic; GP13, South Pacific). The map was produced using the open-source Cartopy package (https://scitools.org.uk/cartopy/docs/latest/).

sets generated with suboptimal primers—information that is often unknown or only recognized anecdotally by specific investigators.

RESULTS AND DISCUSSION

Overall primer coverage for pelagic ocean data sets. To evaluate the real-world extent of primer biases, we compared PCR primers currently in broad use by the marine microbiology community (see Table S2 at https://osf.io/gr4nc) to SSU rRNA sequences extracted from a geographically diverse set of pelagic ocean metagenomes (Fig. 1; see also Table S1 at https://osf.io/gr4nc). These analyses were based on the percentage of perfect matches between primers and metagenome sequences as a metric to determine potential PCR performance, since previous studies have shown that even a single mismatch can significantly bias results (3, 13).

We observed relatively even recruitment of metagenomic reads across the SSU rRNA molecule, which indicates that each different primer region has broadly similar potential to generate accurate taxonomic profiles of naturally occurring organisms (see supplemental results at https://osf.io/gr4nc). However, when we compared the precise sequences derived from the primer-binding regions to the oligonucleotide primer sequences, we observed that predicted median coverage of perfect matches varied widely among primer pairs (see Table S2 at https://osf.io/gr4nc), with 515Y/806RB and 515Y/926R showing the most consistently high incidence of perfect matches (Fig. 2). The ability of all primers to capture the underlying diversity in the metagenomic samples can be improved to some extent by increasing degeneracies, though this varied significantly between primer sets. Adding a single additional degeneracy fixes the majority of mismatches for some primers (e.g., 785R; discussed further below), whereas others have greater than 2 mismatches to metagenomic sequences and thus would require more extensive modifications (see Fig. S1 to S28 at https://osf.io/gr4nc). Below, we discuss the performance of these primers across 4 broad taxonomic groups (*Archaea*, *Bacteria*, *Cyanobacteria* + plastidal 16S rRNA, and *Eukarya*). We separated *Cyanobacteria* + phytoplankton plastidal 16S rRNA from bacterial 16S rRNA since these organisms are responsible for the vast majority of marine primary productivity and thus are important to quantify accurately.

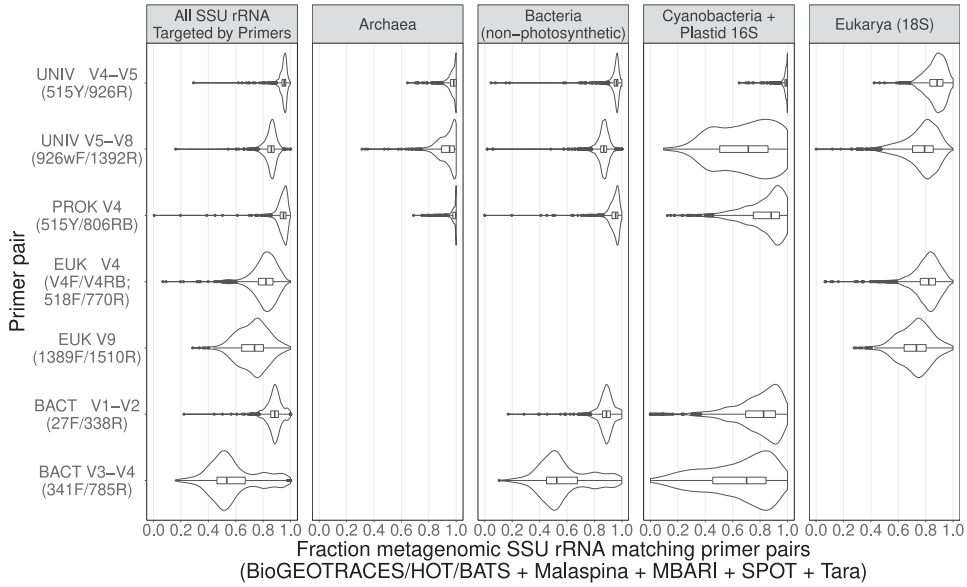

FIG 2 Oligonucleotide PCR primer coverage based on comparisons with globally distributed oceanic metagenomes. The shapes of the plots illustrate the distribution of primer coverage across all samples, where perfect correspondence of primers and metagenome sequences is represented by data piled up against "1.0" on the *x* axis, a situation approached by the UNIV V4-V5 primer for *Cyanobacteria* + plastid 16S rRNA, for example. Note that the "All SSU rRNA" panel represents all of the SSU rRNA targeted by a particular primer set (e.g., for 1389F/1510R this includes only *Eukarya* while for 926wF/1392R it includes all 4 taxonomic groups). Data are corrected for predicted taxonomic overlap between forward and reverse primers. Coordinates refer to locations of primer alignments to the *Escherichia coli* 16S rRNA gene (strain K-12, substrain MG1655).

For *Archaea*, the primer with the best median coverage was the EMP 515Y/806RB combination (coverage = 0.996), followed by 515Y/926R (0.982), and 926wF/1392R (0.952). We also note that our results suggest the 341F/785R primer combination could amplify most *Archaea* with the addition of several degeneracies in 341F (see Fig. S1 to S7 at https://osf.io/gr4nc).

For *Bacteria* (excluding *Cyanobacteria* and plastidal 16S rRNA sequences), the primer with the best median coverage was the 515Y/926R combination (0.961), followed by 515Y/806RB (0.955), 27F/338R (0.888), 926wF/1392R (0.869), and 341F/785R (0.525). The relatively poor performance of 341F/785R is mainly due to the reverse primer (median individual primer coverages: 341F = 0.887, 785R = 0.624), whereas the other two less-optimal primer pairs have more even individual performances (27F = 0.968, 338R = 0.913; 926wF = 0.916, 1392R = 0.943).

For *Cyanobacteria* and chloroplast 16S rRNA (derived from eukaryotic phytoplankton), the primer pair with the highest median coverage was 515Y/926R (0.992), followed by 515Y/806RB (0.881), 27F/338R (0.826), 926wF/1392R (0.714), and 341F/785R (0.702). We note that the vast majority of these mismatched sequences come from chloroplast 16S rRNA sequences—the common cyanobacterial groups *Prochlorococcus* and *Synechococcus* have high coverage with all primers listed above, so this is only a concern for those wishing to use chloroplast 16S rRNA data to quantify eukaryotic phytoplankton.

For *Eukarya*, the primer pair with the highest median coverage was 515Y/926R (0.883), followed by V4F/V4RB (0.822), 926wF/1392R (0.792), and 1389F/1510R (0.732). The distributions of coverage were considerably broader for *Eukarya* versus all other 16S rRNA categories, which may be due to a higher sequence heterogeneity among eukaryotic genomes. As above, all primer sets could be improved with the addition of some degeneracy (see Fig. S22 to S28 at https://osf.io/gr4nc), but approaching perfect coverage seems less practically achievable for eukaryotic SSU rRNA sequences. In addition, our analysis indicates that improvements in eukaryotic coverage are most likely to come from targeting universally conserved rRNA regions, such as those containing

the 515Y/V4F and the 1389F/1392R primers, respectively (see Fig. S22 to S28 at https://osf.io/gr4nc).

We note, however, that recognizable dinoflagellate chloroplast rRNA sequences were almost completely missing from the metagenomic sequences (for all primer regions, not just 515Y/926R) and thus may not appear in the resulting amplicons regardless of primer choice. This is possibly due to the unusual genomic organization/subcellular localization of chloroplast genes in dinoflagellates (16), or alternatively because we lack sufficiently accurate or comprehensive databases for identifying dinoflagellate chloroplast 16S rRNA (discussed further below). It is possible, however, to recover and confidently identify dinoflagellate 18S rRNA sequences from the 515Y/926R primer pair (17), making it feasible to reconstruct a holistic picture of phytoplankton community composition/abundance in combination with the chloroplast 16S rRNA.

In summary, the primer pairs with the highest environmental coverage were 515Y/926R (universal), 515Y/806RB (prokaryote-only), and V4F/V4RB (eukaryote-only). It is worth emphasizing that both universal primer sets have the potential to be equally or more comprehensive for *Eukarya* than the *Eukarya*-specific primer sets tested here. In addition, for those wishing to quantify the abundance of eukaryotic phytoplankton using chloroplast data, the 515Y/926R primer set nearly perfectly matches all environmental sequences. These results are a quantitative confirmation that the careful primer design previously reported for the 515Y/806RB/926R primers has resulted in very high overall coverage in oceanic ecosystems. On the other hand, certain primer sets (e.g., 341F/785R) have very low coverage due to mismatches to dominant organisms, despite being designed to maximize environmental coverage (18). Resulting data may thus be distorted, though we note that removing taxa with mismatches to primers from processed data will recover accurate relative abundances for the remaining taxa (13), due to the fact that biases are thought to be taxon specific (19). This relative abundance correction approach depends on our pipeline's specific identification of mismatched taxa and provides a way to recover useful quantitative information from legacy data sets that may have been biased during PCR. In combination with ecosystem-specific full-length 16S rRNA databases, this correction approach could represent an important tool for integrating historical amplicon data sets and modern long-read sequencing data into a single intercomparable framework for observing longer-term changes in ecosystem structure (15). In addition, by eliminating or at least controlling for primer bias, our approach could help better evaluate how well particular primer regions perform with respect to taxonomic profiling and classification since it would tease apart the effect of variable region from amplification artifacts.

A quantitative intercomparison between SSU rRNA amplicon sequencing and metagenomics. In order for amplicon sequencing to become useful as a quantitative tool for measuring natural gene abundances, it is desirable to benchmark amplicon abundances against an external reference that is not affected by potential PCR biases to gain a more objective determination of how well a primer set recovers true gene abundances. While shotgun metagenomics is not necessarily free of all biases, these biases are distinct from those in PCR-based amplicon studies; thus, if the two data types yield similar community composition patterns, we can be more confident about our overall conclusions. We tested whether amplicon-based strategies recover similar patterns as metagenomes by generating PCR amplicon sequences for a subset of 272 BioGEOTRACES (20) samples (cruises GA03 and GP13). We used the same DNA, eliminating extraction biases as a potential confounding factor in this intercomparison. We were thus able to make a direct "apples-to-apples" quantitative comparison between the relative abundances of organisms determined with amplicons and those determined with metagenomic reads.

To accomplish this, we compared exact amplicon sequence variants (ASVs) from the 515Y/926R primer pair with metagenomic reads from the same region of the SSU rRNA molecule. To make a more robust quantitative comparison, ASVs were clustered as necessary to account for the fact that short (150-bp) metagenomic reads could in some cases be attributed only to a broader phylogenetic group (e.g., *Prochlorococcus* and SAR11). The summed abundances of all taxonomic groups, as determined by ASVs

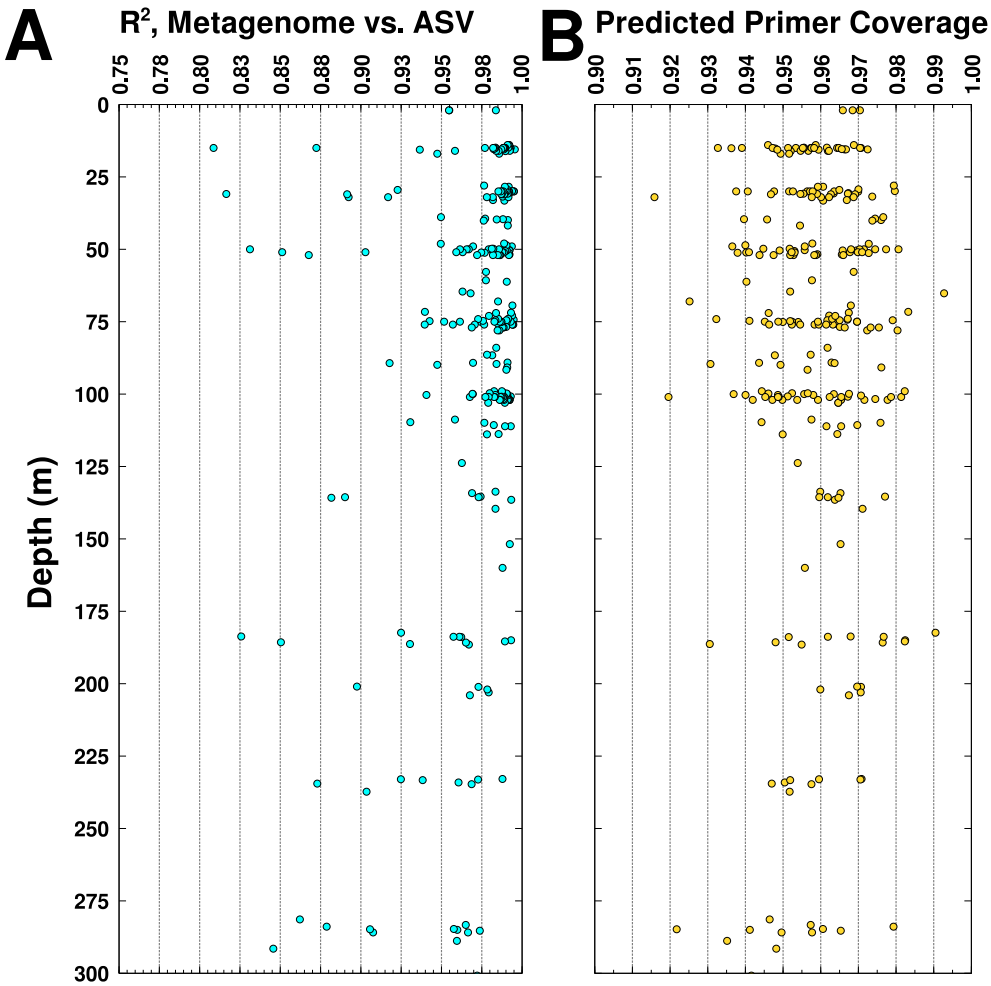

FIG 3 Metagenome-amplicon/oligonucleotide primer comparisons across the marine water column. (A) Metagenome-amplicon quantitative correspondence across all depths for the GA03 cruise (BioGEOTRACES), showing the R^2 value of the plot of the relative abundances of amplicon sequence variants (ASVs; 515Y/926R primers) versus relative abundances of metagenomic reads from the same organisms (see text for details of comparisons). (B) The fraction of exact matches of the 515Y and 926R amplicon primers compared to metagenomic reads of the priming regions for each sample, summarized as the "worst-case" coverage for the combined primers [i.e., $1 - (\text{missed}_{\text{fwd}} + \text{missed}_{\text{rev}})$].

and metagenomic reads, were then compared using a linear least-squares regression. For 16S rRNA sequences, we observed very similar community composition for both techniques, as shown by very high R^2 values and coverage in Fig. 3A (median $R^2 = 0.98$; range = 0.55 to 1.0). This indicated that the ASV sequences were recovering the same diversity of organisms present in metagenomic reads from the primer region and is consistent with the above-mentioned observation of even coverage of metagenomic reads across the SSU rRNA molecule.

These patterns were generally consistent across the surface-mesopelagic transition where there is a major change in community composition (Fig. 3). We do, however, note that at depths of >150 m about 38% of samples had R^2 values below 0.95 (Fig. 3A), versus 9.2% for samples at depths of <150 m. These lower correspondences are unlikely to be due to mismatches to the primer region, since predicted primer coverage is uniformly high across depth for the 515Y/926R primer pair across depth (Fig. 3B). Other factors such as DNA quality or quantity may have been responsible for this, but we cannot rule out the possibility that it involves characteristics of the sequences between the priming regions such as secondary structure or expanded loops known from some eukaryotes (21). Regardless, these results show that amplicon

sequencing and metagenomic profiling typically have a very high quantitative correspondence with the primers tested here (515Y/926R). Additionally, we provide a software toolbox for making further intercomparisons between arbitrary primer sets and metagenomes that could be useful for optimizing both PCR amplicon sequencing and metagenomics for diverse environments.

Improving existing primers. In addition to quantifying mismatches, our pipeline also identifies the specific primer region variants found in particular naturally occurring taxa and their coverage across data sets. This information can guide improvements for any of the primers tested here, and we identify two specific use cases. First, some of the primers identified as having major mismatches are dominated by a few organisms, and thus small modifications are all that is necessary to make them more appropriate for oceanic ecosystems. Second, we were interested to know whether for the better-performing primers it would be possible to identify a combination of oligonucleotides that would perform well across all marine environments tested here. Such a primer set could have the advantage of being able to monitor rare but biogeochemically important taxa that may become more abundant due to climate change, e.g., as a result of expanding oxygen minimum zones.

To illustrate how a primer with poor predicted performance could be improved, we use the case of 785R, which has a mismatch to the dominant organism SAR11. Because the SAR11 mismatches represent up to 40% of total 16S rRNA reads (see Table S3 at https://osf.io/gr4nc), this one mismatch will have a major effect on the resulting data but could be corrected by adding a single additional degeneracy at position 7 in the primer (5′-GACTAC**N**VGGGTATCTAATCC). However, 785R also has mismatches to a number of abundant chloroplast sequences (Fig. 2), which could be corrected with two additional degeneracies (5′-**R**A**Y**TACNVGGGTATCTAATCC). Compared to the original primer (5′-GACTACHVGGGTATCTAATCC), these primers would require the synthesis of a larger number of oligonucleotide combinations (12 and 48 versus 9 originally) but are well within the number of combinations currently in use for other environmental amplification strategies (22). Interestingly, despite this being the worst-performing primer for marine ecosystems due to mismatches to the dominant taxon SAR11, we predict that by adding the above degeneracies this primer would achieve nearly perfect coverage for *Bacteria* (see Fig. S15 to S21 at https://osf.io/gr4nc). This suggests the primer-binding region of 785R is highly conserved among *Bacteria*, consistent with the initial publication describing this primer as the best overall for prokaryotic taxa across diverse environments (18). This observation also underscores the value of comparing primer sequences with environmental data since it uncovered a significant, unexpected mismatch to an abundant taxon that can now be corrected.

We also investigated whether it is possible to improve primer sets that already perform well (e.g., 515Y/806RB/926R) to the point where they offer near-universal coverage of all taxa identified in this study. To do so, we identified data sets that had significant potential for coverage improvement (see Fig. S1 to S28 at https://osf.io/gr4nc) and certain rare order-level taxa that had many mismatches or are known to be biogeochemically important.

Our data showed that the 806RB primer performed extremely well overall for 16S rRNA sequences but did miss a number of 16S rRNA chloroplast sequences mostly from the *Mamiellales* clade, which includes small picoeukaryotes such as *Ostreococcus* and *Bathycoccus*. As shown in Fig. S8 to S14 at https://osf.io/gr4nc, these taxa were still missed when allowing for up to 2 mismatches, meaning that they are unlikely to be accurately quantified with the current primer design. Alterations would be possible, but we note that this may be an intentional design choice, since chloroplast amplification may be undesirable in certain environments (23) but is often very useful in euphotic zone marine water samples where it can be used to quantify primary producer abundance. Related to this, we note that a context-dependent advantage of 806RB is that it lacks an 18S rRNA binding site. This means that in combination with 515Y, nuclear 18S rRNA will not be amplified with 806RB. In other words, even though

the variable region covered by 515Y/806RB contains eukaryotic sequences, this primer set will not recover them as amplicons—demonstrating that primer design places a first-order constraint on how well a given amplicon product reflects the natural community. In contrast, the 515Y/926R combination will amplify these eukaryotic sequences. In a practical sense, amplifying 18S rRNA sequences with 16S rRNA can present problems, for example, in host-associated microbiome studies where 18S rRNA could overwhelm targeted 16S rRNA. This overwhelming of libraries by 18S rRNA sequences is, however, unlikely to be a major issue for the 515Y/926R primer pair for pelagic ocean samples. This is due to the fact that 18S rRNA sequences will generally be selected against in the sequencing process (thus reducing their overall abundances [13]) and the fact that marine environments are generally dominated by prokaryotic microorganisms (13) (see Fig. S29 at https://osf.io/gr4nc).

The 515Y primer was relatively easy to improve, given that its performance was already very robust across diverse data sets (see Fig. S1 to S28 at https://osf.io/gr4nc). We identified four positions where further degeneracies could be added (5′-GTG**B**CAGCM**SY**CGCGGT**M**A) and were sufficient to resolve the vast majority of mismatches including to rare taxa such as certain representatives of the *Patescibacteria* (also known as candidate phyla radiation [CPR] bacteria) which were previously identified as taxonomic "blind spots" in PCR amplicon analysis (5). However, incorporating these new degeneracies produced a relatively large number of oligonucleotide combinations compared to the original primer (48 versus 4 originally), and most of these new variants were not observed in the metagenomic data. We thus decided to redesign the primer based on a different approach, starting from a nondegenerate primer and adding in only variants that could be detected in the natural environment or in the SILVA database. This resulted in a specific mixture of 13 specific oligonucleotides that we term 515Yp-min ("p" for pelagic, "min" for minimal). This mixture of oligonucleotides is tailored specifically to the environmental sequences present in natural marine environments. It maximizes organismal coverage, while keeping the overall degeneracy at a relatively low level compared to other degenerate primers used in environmental microbiology (22). Another key advantage of this tailored approach is that it permits iterative improvements to a primer set over time. For example, we added a variant matching dinoflagellate chloroplast sequences from the PhytoRef database (24) that may potentially improve quantification of these organisms' plastidal 16S rRNA in combination with other modifications to 926R (discussed below). In the future, if we were able to identify other variants explaining why dinoflagellate chloroplast 16S sequences are conspicuously absent from 515Y/926R amplicon data, it would then be feasible to add these new oligonucleotides to the mixture without creating excessive sequence redundancy.

The 926R primer had mismatches to more taxa and thus required more extensive modifications. For example, Table S3 at https://osf.io/gr4nc shows that 926R has mismatches to both *Ectothiorhodospirales* (an uncultivated group related to purple sulfur bacteria [25]) and the *Rickettsiales* (an order that includes mitochondrial sequences in addition to free-living organisms [26]). While less abundant overall, we also noted mismatches in the Tara Oceans data set to *Brocadiales*, the order containing anammox bacteria (27). In the Guaymas Basin sediment samples, some of the dominant mismatches were to the *Campylobacteria*, dominant chemoautotrophs in many sulfidic systems (28), as well as unusual *Archaea* that are now recognized to be much more diverse than previously thought and missed by some PCR primers (29). Accurate quantification of these rare taxa would allow the monitoring of unusual but potentially significant changes in biogeochemistry. For example, the *Campylobacteria* are known to be abundant in certain low-oxygen microenvironments such as in sediment trap particles (30), and the *Brocadiales* are abundant in oxygen minimum zones (27). Being able to better quantify these "sentinel taxa" could allow us to monitor the dynamics of these low-oxygen environments that are likely to expand under global climate change.

Following the same approach discussed above for 515Yp-min, we developed a

mixture of 38 oligonucleotides ("926Rp-min") that contains perfect matches for above-mentioned taxa, primer variants that have >2% overall relative abundance across all environments, and 2 sequences from dinoflagellate chloroplasts that were not previously covered. This set of oligonucleotides is comprehensive for all the environments tested in our study and would likely result in major improvements in quantification of all of the above-mentioned taxa. To achieve the same organismal coverage by adding more degenerate sites would increase the total number of primer variants to at least 144 (compared to 16 in the original primer). Not only would this dilute perfectly matching oligonucleotides and be unlikely to result in efficient PCR amplification, it would also include many redundant oligonucleotides unnecessary with an explicit oligonucleotide mixture. The improvements in coverage for this new mixture may be particularly apparent for *Rickettsiales* templates, many of which have a single mismatch at the 3′ end of the primer that is known to completely prevent PCR amplification (6). Since the order *Rickettsiales* contains sequences from mitochondrial SSU rRNA, this modification will potentially allow for more effective quantification of protist abundances, similar to how chloroplasts quantify marine protistan phytoplankton.

Since it is beyond the scope of this paper to provide exhaustive recommendations for each primer set, we direct interested readers to an Open Science Foundation repository which contains all necessary output files, scripts, and a suggested workflow for improving primers using metagenomic data (https://osf.io/gr4nc/). Improvements should be possible for most taxon/primer combinations since a few mismatches typically dominated, with a long tail of rare variants that may derive from sequencing errors or pseudogenized SSU rRNA. This information would allow an interested reader to optimize one of the primers investigated here for particular taxa or environments of interest.

Moving forward, we recommend oceanographers consider applying these two new primer mixtures for pelagic water column surveys to maximize organismal coverage (515Yp-min, 926Rp-min; see Table S4 at https://osf.io/gr4nc/). By correcting for 3′ mismatches that lead to taxonomic "blind spots" and including exact matches to taxa found in low-oxygen regions, we should be able to better quantify the abundance and dynamics of organisms critical to marine carbon, nitrogen, and sulfur cycling. In addition, by representing primers as a specific mixture of oligonucleotides rather than a single sequence with degeneracies, it leaves open the possibility of small modifications in the future to improve coverage; the same is not true for fully degenerate primers, where each new degeneracy will multiply the total number of combinations by at least 2-fold and further dilute the other perfectly matching oligonucleotides. If proven to be as quantitative as the original 515Y/926R primers (3, 13), these new mixtures would provide an affordable, comprehensive (all organisms from *Archaea* to zooplankton), and scalable method to measure whole-community compositions across time and space in the world's oceans. While further validation is necessary, we believe our results and those of previous studies (3, 13–15) demonstrate the potential for PCR amplicon methods to accurately quantify natural marker gene abundances in a robust and accurate manner, either by recovering true relative abundances or in combination with internal standards to obtain absolute quantification (14). This would allow confident measurement of microbial community composition alongside well-developed methods such as those for quantifying phytoplankton photosynthetic pigments. In turn, this would help expand our understanding of how oceanic microbes interact with biogeochemical cycles and respond to global climate change.

Finally, our bioinformatic approach could be applied elsewhere to generate ecosystem-specific primer sets. For example, it remains unknown whether the patterns we observed for seawater would be true in other systems, such as terrestrial sediment, soil, freshwater, or animal/plant-associated microbiomes. Our reproducible workflow, combined with expert curation and broad environmental sampling, could ensure that taxa of interest are accurately quantified by oligonucleotide primers for any environment where deeply sampled metagenomes could be used as a database. In addition,

our approach could allow for the rational evaluation and improvement of new primers for next-generation long-read amplicon sequencing, or to improve the performance of primers used to identify animal species of economic or conservation importance. By evaluating primer coverage objectively and providing simple ways to iteratively improve existing primer sets, it would ensure that the potential of these techniques for monitoring microbiomes relevant to ecosystem functioning and human health is fully realized.

MATERIALS AND METHODS

Sample scope/biogeographic distribution and evaluated primers. We analyzed globally distributed metagenomic data sets that reflect a range of depths and nutrient regimes (Fig. 1; see also Table S1 at https://osf.io/gr4nc) and tested primers that are commonly used in environmental microbiology (see Table S2 at https://osf.io/gr4nc) (2–4, 18, 31–35). Metagenomic samples included worldwide sampling expeditions such as Tara Oceans (36), the Malaspina expedition (37), and BioGEOTRACES (20). Among these three campaigns, the Malaspina metagenomes focus largely on the deep sea, whereas Tara and BioGEOTRACES sample primarily the sunlit ocean. In addition, we used time-series data from the Hawaii Ocean Time-Series and the Bermuda Atlantic Time-Series (38, 39) (HOT/BATS; both analyzed with BioGEOTRACES as BiHOBA [BioGEOTRACES-HOT-BATS]), a bloom time-series in Monterey Bay (40), and the San Pedro Ocean Time-Series (41) (SPOT). We also included in the analysis two sites representing suboxic/anoxic systems—the Saanich Inlet time-series (42) and samples collected from hydrothermally altered sediments in Guaymas Basin (43). We also generated new 16S/18S rRNA universal amplicon sequences for two BioGEOTRACES cruise transects (GA03/GP13; noted with open circles in Fig. 1) which were used for making intercomparisons between metagenomes and amplicons.

Evaluation of primer coverage/biases using metagenomic reads. In order to evaluate the performance of our primer set compared to others, we used metagenomic reads to estimate the fraction of SSU rRNA fragments that could be successfully targeted by (i.e., perfectly match) each primer set. In brief, our pipeline retrieves SSU rRNA from fastq-formatted shotgun metagenomic samples, removes sequences containing uninformative repeats, sorts into four organismal categories (*Archaea*, *Bacteria*, *Cyanobacteria* [including plastidal 16S rRNA], and *Eukarya*), and aligns these sequences to a group-specific reference SSU rRNA sequence (e.g., *Saccharomyces cerevisiae* for *Eukarya*). To obtain sequences overlapping each oligonucleotide primer and to exclude nonprimer reads, coordinates provided in the config file were used to extract reads from alignments that overlapped the primer-binding region plus 5 leading or trailing bases. These reads, which represent SSU rRNA fragments with a primer-binding region, were then quality filtered and compared with primer sequences (see Table S2 at https://osf.io/gr4nc/) at 0-, 1-, and 2-mismatch thresholds (see Fig. S1 to S28 at https://osf.io/gr4nc) to identify matched and mismatched sequences. As an additional quality-control step, we excluded any aligned read falling in the primer region that did not have a match to the primer at a 6-mismatch threshold (e.g., misalignments or distantly related sequences). Both mismatched and matched reads were then classified using the SILVA 132 database (*Archaea*, *Bacteria*, *Eukarya*) and PhytoRef (11, 24) (*Cyanobacteria* + plastidal 16S rRNA). The specific software implementation is described fully in the supplemental material at https://osf.io/gr4nc/. For those wishing to inspect results further, complete plaintext summaries of output and summary plots are available online at https://osf.io/gr4nc/.

These computational steps were implemented with the Snakemake workflow engine (44) to create a documented and reproducible pipeline which is freely available (https://github.com/jcmcnch/MGPrimerEval). It automatically produces tabular output/summary plots and can be applied to new samples and primers by modifying template configuration files. The specific functioning of the three workflows is described in more detail in the supplemental material at https://osf.io/gr4nc/ and on the GitHub page.

PCR amplification and ASV generation. Two hundred seventy-two DNA samples from GEOTRACES cruises GA03 and GP13 were used to generate PCR amplicons and represent either surface (GP13) or surface plus upper mesopelagic water samples (GA03). GP13 is a longitudinal transect in the southern Pacific, whereas GA03 is a longitudinal transect in the North Atlantic. DNA used for PCR amplicons was the exact same DNA material used to produce the metagenomes described by Biller et al. (20) except that it was diluted to 0.5 ng/μl in low-EDTA Tris-EDTA (TE) buffer prior to amplification. DNA was amplified with the 515Y/926R primers (515Y, 5′-GTGYCAGCMGCCGCGGTAA/926R, 5′-CCGYCAATTYMTTTRAGTTT) (3) using 0.5 ng of DNA template in a 25-μl reaction mixture with a final primer concentration of 0.3 mM. Primers were part of a larger construct with Illumina adapter constructs/barcodes already included, allowing for single-step library preparation. Thermocycling was as follows: an initial denaturation at 95°C for 120 s followed by 30 amplification cycles of 95°C for 45 s, 50°C for 45 s, and 68°C for 90 s and a final elongation step at 68°C for 300 s. Laboratory methods are described in more detail at https://dx.doi.org/10.17504/protocols.io.vb7e2rn. Both 16S and 18S rRNA mock communities (staggered and even versions of both [3, 13]) were included in the sequencing run as quality controls. Amplicons were sequenced with HiSeq 2 × 250 RapidRun technology at the University of Southern California (USC) sequencing center in a run pooled with shotgun metagenomic samples. Raw basecall data were demultiplexed with bcl2fastq (-r 20 -p 20 -w 32 –barcode-mismatches 0) (https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html) using the 6-bp reverse indices and subsequently with the forward 5-bp barcode with cutadapt (48) according to the templates provided at https://github.com/jcmcnch/demux-notes. Raw data from this step of the pipeline are available on NCBI (see below).

Sequence data were analyzed with a custom *in silico* pipeline available at https://github.com/jcmcnch/eASV-pipeline-for-515Y-926R and based on qiime2 2019.4 (45). Briefly, 16S and 18S rRNA sequences were first partitioned using a custom 16S/18S rRNA-specific database derived from SILVA 132

and PR2 (11, 46) which is available at https://osf.io/e65rs/. After partitioning, primers were removed from amplicons allowing up to 20% mismatches to the primer, and raw sequences were imported into qiime2 and denoised with DADA2 (10). Taxonomy was assigned using a naive Bayesian classification algorithm with the SILVA 132 database subsetted to the amplicon region as a reference (47). Sequences identified as chloroplasts were reannotated with PhytoRef (24), and 18S rRNA sequences were additionally annotated with PR2 (46). A step-by-step protocol for these informatic steps is available at https://dx.doi.org/10.17504/protocols.io.vi9e4h6. After denoising and annotation, biom tables were exported with taxonomy as tsv files and converted to relative abundances which were used as input for the metagenome-amplicon comparison.

Data availability. Raw data from amplicon sequencing are available on NCBI under the umbrella BioProject PRJNA659851 and three subprojects PRJNA658608 (controls), PRJNA658384 (GA03), and PRJNA658385 (GP13).

ACKNOWLEDGMENTS

This work was supported by the Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems (CBIOMES) grant 549943 to J.A.F., the Gordon and Betty Moore Foundation Marine Microbiology Initiative grant 3779, and NSF grant OCE-1737409. DNA extraction was funded in part by grants to Sallie W. Chisholm (S.W.C.) at MIT that enabled the collection and processing of DNA samples from GEOTRACES cruises. These include the Simons Foundation (Life Sciences Project Award ID 337262, S.W.C.; SCOPE Award ID 329108, S.W.C.), the Moore Foundation (grant IDs GBMF495 and GBMF4511), and the National Science Foundation (OCE-1153588, OCE-1356460, and DBI-0424599).

We gratefully acknowledge the effort of the scientists who collected the BioGEOTRACES field samples analyzed here: Christel Hassler, Debbie Hulston, and Elizabeth Maas (GP13) and Jeremy Jacquot (GA03). We also acknowledge the efforts of Keven Dooley and Madeline Williams who conducted DNA extractions. We thank S.W.C. for donating the samples and supporting parts of the study.

REFERENCES

1. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi.org/10.1038/nature24621.
2. Apprill A, McNally S, Parsons R, Weber L. 2015. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquat Microb Ecol 75:129–137. https://doi.org/10.3354/ame01753.
3. Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol 18:1403–1414. https://doi.org/10.1111/1462-2920.13023.
4. Walters W, Hyde ER, Berg-Lyons D, Ackermann G, Humphrey G, Parada A, Gilbert JA, Jansson JK, Caporaso JG, Fuhrman JA, Apprill A, Knight R. 2016. Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. mSystems 1:e00009-15. https://doi.org/10.1128/mSystems.00009-15.
5. Eloe-Fadrosh EA, Ivanova NN, Woyke T, Kyrpides NC. 2016. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. Nat Microbiol 1:15032. https://doi.org/10.1038/nmicrobiol.2015.32.
6. Wear EK, Wilbanks EG, Nelson CE, Carlson CA. 2018. Primer selection impacts specific population abundances but not community dynamics in a monthly time-series 16S rRNA gene amplicon analysis of coastal marine bacterioplankton. Environ Microbiol 20:2709–2726. https://doi.org/10.1111/1462-2920.14091.
7. Campbell BJ, Yu L, Heidelberg JF, Kirchman DL. 2011. Activity of abundant and rare bacteria in a coastal ocean. Proc Natl Acad Sci U S A 108:12776–12781. https://doi.org/10.1073/pnas.1101405108.
8. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, Küsel K, Rillig MC, Rivett DW, Salles JF, van der Heijden MGA, Youssef NH, Zhang X, Wei Z, Hol WHG. 2017. Where less may be more: how the rare biosphere pulls ecosystems strings. ISME J 11:853–862. https://doi.org/10.1038/ismej.2016.174.
9. Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, Poulton NJ, Burkart MD, La Clair JJ, Chisholm SW, Stepanauskas R. 2019. Charting the complexity of the marine microbiome through single-cell genomics. Cell 179:1623–1635.e11. https://doi.org/10.1016/j.cell.2019.11.017.
10. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583. https://doi.org/10.1038/nmeth.3869.
11. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596. https://doi.org/10.1093/nar/gks1219.
12. Cole JR, Wang Q, Chai B, Tiedje JM. 2011. The Ribosomal Database Project: sequences and software for high-throughput rRNA analysis, p 313–324. *In* de Bruijn FJ (ed), Handbook of molecular microbial ecology I. John Wiley & Sons, Inc, Hoboken, NJ.
13. Yeh Y-C, McNichol JC, Needham DM, Fichot EB, Fuhrman JA. 2021. Comprehensive single-PCR 16S and 18S rRNA community analysis validated with mock communities, and estimation of sequencing bias against 18S. Environ Microbiol, in press. https://doi.org/10.1111/1462-2920.15553.
14. Lin Y, Gifford S, Ducklow H, Schofield O, Cassar N. 2019. Towards quantitative microbiome community profiling using internal standards. Appl Environ Microbiol 85:e02634-18. https://doi.org/10.1128/AEM.02634-18.
15. Dueholm MS, Andersen KS, McIlroy SJ, Kristensen JM, Yashiro E, Karst SM, Albertsen M, Nielsen PH. 2020. Generation of comprehensive ecosystem-specific reference databases with species-level resolution by high-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment (AutoTax). mBio 11:e01557-20. https://doi.org/10.1128/mBio.01557-20.
16. Laatsch T, Zauner S, Stoebe-Maier B, Kowallik KV, Maier U-G. 2004. Plastid-derived single gene minicircles of the dinoflagellate Ceratium horridum are localized in the nucleus. Mol Biol Evol 21:1318–1322. https://doi.org/10.1093/molbev/msh127.
17. Needham DM, Fuhrman JA. 2016. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. Nat Microbiol 1:16005. https://doi.org/10.1038/nmicrobiol.2016.5.
18. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res 41:e1. https://doi.org/10.1093/nar/gks808.

19. McLaren MR, Willis AD, Callahan BJ. 2019. Consistent and correctable bias in metagenomic sequencing experiments. Elife 8:e46923. https://doi.org/10.7554/eLife.46923.
20. Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, Hogle SL, Coe A, Bergauer K, Bouman HA, Browning TJ, De Corte D, Hassler C, Hulston D, Jacquot JE, Maas EW, Reinthaler T, Sintes E, Yokokawa T, Chisholm SW. 2018. Marine microbial metagenomes sampled across space and time. Sci Data 5:180176. https://doi.org/10.1038/sdata.2018.176.
21. Obiol A, Giner CR, Sánchez P, Duarte CM, Acinas SG, Massana R. 2020. A metagenomic assessment of microbial eukaryotic diversity in the global ocean. Mol Ecol Resour 20:718–731. https://doi.org/10.1111/1755-0998.13147.
22. Gaby JC, Buckley DH. 2012. A comprehensive evaluation of PCR primers to amplify the nifH gene of nitrogenase. PLoS One 7:e42149. https://doi.org/10.1371/journal.pone.0042149.
23. Thijs S, Op De Beeck M, Beckers B, Truyens S, Stevens V, Van Hamme JD, Weyens N, Vangronsveld J. 2017. Comparative evaluation of four bacteria-specific primer pairs for 16S rRNA gene surveys. Front Microbiol 8:494. https://doi.org/10.3389/fmicb.2017.00494.
24. Decelle J, Romac S, Stern RF, Bendif EM, Zingone A, Audic S, Guiry MD, Guillou L, Tessier D, Le Gall F, Gourvil P, Dos Santos AL, Probert I, Vaulot D, de Vargas C, Christen R. 2015. PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. Mol Ecol Resour 15:1435–1445. https://doi.org/10.1111/1755-0998.12401.
25. Oren A. 2014. The family Ectothiorhodospiraceae, p 199–222. *In* Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (ed), The prokaryotes. Springer, Berlin, Germany.
26. Fredricks DN. 2006. Introduction to the Rickettsiales and other intracellular prokaryotes, p 457–466. *In* Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E (ed), The prokaryotes. Springer, New York, NY.
27. Medina Faull L, Mara P, Taylor GT, Edgcomb VP. 2020. Imprint of trace dissolved oxygen on prokaryoplankton community structure in an oxygen minimum zone. Front Mar Sci 7:360. https://doi.org/10.3389/fmars.2020.00360.
28. Campbell BJ, Engel AS, Porter ML, Takai K. 2006. The versatile ε-proteobacteria: key players in sulphidic habitats. Nat Rev Microbiol 4:458–468. https://doi.org/10.1038/nrmicro1414.
29. Baker BJ, De Anda V, Seitz KW, Dombrowski N, Santoro AE, Lloyd KG. 2020. Diversity, ecology and evolution of Archaea. Nat Microbiol 5:887–900. https://doi.org/10.1038/s41564-020-0715-z.
30. Boeuf D, Edwards BR, Eppley JM, Hu SK, Poff KE, Romano AE, Caron DA, Karl DM, DeLong EF. 2019. Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. Proc Natl Acad Sci U S A 116:11824–11832. https://doi.org/10.1073/pnas.1903080116.
31. Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS One 4:e6372. https://doi.org/10.1371/journal.pone.0006372.
32. Wilkins D, van Sebille E, Rintoul SR, Lauro FM, Cavicchioli R. 2013. Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects. Nat Commun 4:2457. https://doi.org/10.1038/ncomms3457.
33. Balzano S, Abs E, Leterme S. 2015. Protist diversity along a salinity gradient in a coastal lagoon. Aquat Microb Ecol 74:263–277. https://doi.org/10.3354/ame01740.
34. Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner H-W, Richards TA. 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol 19:21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x.
35. Fortunato CS, Herfort L, Zuber P, Baptista AM, Crump BC. 2012. Spatial variability overwhelms seasonal patterns in bacterioplankton communities across a river to ocean gradient. ISME J 6:554–563. https://doi.org/10.1038/ismej.2011.135.
36. Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D, Karsenti E, Speich S, Troublé R, Dimier C, Searson S, Tara Oceans Consortium Coordinators. 2015. Open science resources for the discovery and analysis of Tara Oceans data. Sci Data 2:150023. https://doi.org/10.1038/sdata.2015.23.
37. Acinas SG, Sánchez P, Salazar G, Cornejo-Castillo FM, Sebastián M, Logares R, Sunagawa S, Hingamp P, Ogata H, Lima-Mendez G, Roux S, González JM, Arrieta JM, Alam IS, Kamau A, Bowler C, Raes J, Pesant S, Bork P, Agusti S, Gojobori T, Bajic V, Vaqué D, Sullivan MB, Pedrós-Alió C, Massana R, Duarte CM, Gasol JM. 2019. Metabolic architecture of the deep ocean microbiome. bioRxiv 635680. https://www.biorxiv.org/content/10.1101/635680v1.
38. Karl DM, Church MJ. 2014. Microbial oceanography and the Hawaii Ocean Time-series programme. Nat Rev Microbiol 12:699–713. https://doi.org/10.1038/nrmicro3333.
39. Steinberg DK, Carlson CA, Bates NR, Johnson RJ, Michaels AF, Knap AH. 2001. Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. Deep Sea Res 2 Top Stud Oceanogr 48:1405–1447. https://doi.org/10.1016/S0967-0645(00)00148-X.
40. Nowinski B, Smith CB, Thomas CM, Esson K, Marin R, Preston CM, Birch JM, Scholin CA, Huntemann M, Clum A, Foster B, Foster B, Roux S, Palaniappan K, Varghese N, Mukherjee S, Reddy TBK, Daum C, Copeland A, Chen I-MA, Ivanova NN, Kyrpides NC, Glavina del Rio T, Whitman WB, Kiene RP, Eloe-Fadrosh EA, Moran MA. 2019. Microbial metagenomes and metatranscriptomes during a coastal phytoplankton bloom. Sci Data 6:129. https://doi.org/10.1038/s41597-019-0132-4.
41. Sieradzki ET, Ignacio-Espinoza JC, Needham DM, Fichot EB, Fuhrman JA. 2019. Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. Nat Commun 10:1169. https://doi.org/10.1038/s41467-019-09106-z.
42. Hawley AK, Torres-Beltrán M, Zaikova E, Walsh DA, Mueller A, Scofield M, Kheirandish S, Payne C, Pakhomova L, Bhatia M, Shevchuk O, Gies EA, Fairley D, Malfatti SA, Norbeck AD, Brewer HM, Pasa-Tolic L, del Rio TG, Suttle CA, Tringe S, Hallam SJ. 2017. A compendium of multi-omic sequence information from the Saanich Inlet water column. Sci Data 4:170160. https://doi.org/10.1038/sdata.2017.160.
43. Dombrowski N, Seitz KW, Teske AP, Baker BJ. 2017. Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. Microbiome 5:106. https://doi.org/10.1186/s40168-017-0322-2.
44. Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28:2520–2522. https://doi.org/10.1093/bioinformatics/bts480.
45. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857. https://doi.org/10.1038/s41587-019-0209-9.
46. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, de Vargas C, Decelle J, Del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Logares R, Mahé F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet A-L, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R. 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res 41:D597–D604. https://doi.org/10.1093/nar/gks1160.
47. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Caporaso JG. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6:90. https://doi.org/10.1186/s40168-018-0470-z.
48. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12. https://doi.org/10.14806/ej.17.1.200.