

Comprehensive single-PCR 16S and 18S rRNA community analysis validated with mock communities, and estimation of sequencing bias against 18S

Yi-Chun Yeh ^{*}, Jesse McNichol ,
David M. Needham,[†] Erin B. Fichot, Lyria Berdjeb and
Jed A. Fuhrman

Department of Biological Sciences, University of
Southern California, CA, Los Angeles, 90089-0371.

Summary

Universal primers for SSU rRNA genes allow profiling of natural communities by simultaneously amplifying templates from Bacteria, Archaea, and Eukaryota in a single PCR reaction. Despite the potential to show relative abundance for all rRNA genes, universal primers are rarely used, due to various concerns including amplicon length variation and its effect on bioinformatic pipelines. We thus developed 16S and 18S rRNA mock communities and a bioinformatic pipeline to validate this approach. Using these mocks, we show that universal primers (515Y/926R) outperformed eukaryote-specific V4 primers in observed versus expected abundance correlations (slope = 0.88 vs. 0.67–0.79), and mock community members with single mismatches to the primer were strongly underestimated (threefold to eightfold). Using field samples, both primers yielded similar 18S beta-diversity patterns (Mantel test, $p < 0.001$) but differences in relative proportions of many rarer taxa. To test for length biases, we mixed mock communities (16S + 18S) before PCR and found a twofold underestimation of 18S sequences due to sequencing bias. Correcting for the twofold underestimation, we estimate that, in Southern California field samples (1.2–80 μm), there were averages of 35% 18S, 28% chloroplast 16S, and 37% prokaryote 16S rRNA genes. These data demonstrate the potential for universal primers to generate comprehensive microbiome profiles.

Introduction

Bacteria, archaea, and eukaryotes make up dynamic, diverse communities that interact with one another and their environment. Studying all these components simultaneously is essential for understanding how the ecosystem functions as a whole (Fuhrman *et al.*, 2015; Needham *et al.*, 2018; Chénard *et al.*, 2019), though the individual components are mostly studied separately, due in part to the perception that separate assays are required for each. Since high-throughput DNA sequencing was introduced, SSU rRNA sequencing has been widely used for analyzing microbial community structure - especially for prokaryotes by targeting the 16S rRNA gene (Sogin *et al.*, 2006). Analyses focusing on eukaryotic communities with 18S rRNA sequencing, however, are not as common partly because early sequencing lengths could not fully capture diversity in longer hyper-variable regions (Amaral-Zettler *et al.*, 2009). With advances in sequencing capacities, two regions (V4 and V9) have become commonly used for planktonic eukaryotic community profiles (Amaral-Zettler *et al.*, 2009; Stoeck *et al.*, 2010; Balzano *et al.*, 2015; De Vargas *et al.*, 2015).

Despite these methodological developments, the question of how well the entire sequencing and analysis pipeline recovers the true abundance of rRNA genes found in the natural community has received less attention. In pelagic marine environments, studies have underscored the importance of careful primer design for accurately resolving natural communities, e.g. correcting the severe underestimate of the SAR11 clade that occurred with one of the most popular primers (Caporaso *et al.*, 2012; Apprill *et al.*, 2015; Parada *et al.*, 2016). In addition, validation and inter-comparison of primer performance have also been facilitated by the development and application of microbial internal standards or ‘mock communities’ to PCR amplicon analysis (Wear *et al.*, 2018) (hereafter ‘mocks’). The application of mocks to the PCR amplification, sequencing, and analysis protocol has demonstrated that even well-designed primers (515Y/926R vs. 515Y/806R) differ in terms of their ability to recover natural abundance patterns

Received 6 December, 2019; revised 12 April, 2021; accepted 30 April, 2021. *For correspondence. E-mail yichuny@usc.edu; Tel: 213-740-5759. [†]Present address: GEOMAR Helmholtz Centre for Ocean Research, Kiel 24148, Germany.

(Parada *et al.*, 2016). Including mocks in a sequencing run can also verify instrument performance, thus avoiding improper ecological conclusions. For example, inclusion of mocks in a previous study revealed that an unknown technical issue affecting a single sequencing run inexplicably caused an entire major taxon to be missing in output data and altered abundances of other taxa (Yeh *et al.*, 2018). More recently, it has been shown that amplicon methods can be made even more quantitative by the addition of internal DNA standards (i.e. added to samples before extraction and purification of DNA; Lin *et al.* (2019)). This allows normalization of amplicon data closer to true abundances found in seawater (except for lysis efficiency variations) and was found to be consistent with other, extensively validated methods (Lin *et al.*, 2019).

Bioinformatic methods used for amplicon sequence analysis have also evolved considerably, with initial efforts focusing on how well algorithms resolve true biological sequences by clustering sequences into operational taxonomic units (OTUs) at a certain similarity threshold. This effort has culminated in the development of 'denoising' algorithms that are designed to recover true underlying biological sequences to the individual base (i.e. amplicon sequence variants; ASVs) by endeavoring to eliminate sequencing and PCR errors (Eren *et al.*, 2015; Callahan *et al.*, 2016; Amir *et al.*, 2017). Moreover, unlike OTU clustering that analyze sequences into often vaguely defined or 'fuzzy' units that change study-by-study, denoising methods aim to better account for batch effects across multiple sequencing runs and are able to analyze sequences either sample-by-sample (Deblur) or run-by-run (DADA2), which greatly reduces computational demand compared to OTU clustering algorithms that analyze sequences all together (Callahan *et al.*, 2016).

Collectively, these studies show how PCR amplicons can generate quantitative data that allow microbial community composition to be measured alongside other oceanographic variables. However, choosing an appropriate sequencing strategy remains a significant challenge given the diverse primers and sequencing technologies currently available. In order to maximize overall utility, it is highly desirable to keep costs low while generating data with high phylogenetic resolution. Parada *et al.* (2016) have previously described a universal primer set (515Y/926R, modified from Quince *et al.* (2011)) that simultaneously amplifies 16S and 18S rRNA in a single PCR reaction. Because of their universal nature, these primers measure both eukaryotes and prokaryotes and can provide insights into processes such as predation, parasitism and mutualism (Needham and Fuhrman, 2016; Needham *et al.*, 2018).

However, analyzing data generated from the universal 515Y/926R primer set has several potential challenges.

First, mixed 16S and 18S amplicon sequences present bioinformatic challenges since the two types of amplicons must be analyzed differently. This is because current Illumina read lengths are not long enough to allow the forward and reverse reads to overlap for the longer 18S amplicon (575–595 bp). If they do not overlap by at least 12 bases (according to standard methods), they cannot be merged, and if they cannot be merged, the entire amplicon cannot be generated and analyzed as is typical for 16S amplicon analysis. Second, PCR and sequencing both discriminate against longer amplicons (Kittelman *et al.*, 2013), yet the degree of PCR and sequencing biases against longer 18S amplicons is unknown. These biases can potentially be detected via mock community analysis, specifically collections of known 16S or 18S rRNA gene fragments (Bradley *et al.*, 2016; Parada *et al.*, 2016; Needham *et al.*, 2017; Wear *et al.*, 2018; Catlett *et al.*, 2020). Yet to our knowledge, there have not been tests with mixed mock communities consisting of both 16S and 18S rRNA genes.

In this study, we present results from mock communities designed to validate the 515Y/926R primer set with particular emphasis on its performance with 18S sequences in comparison to commonly used 18S-specific primer sets, V4F/V4R and V4F/V4RB (Stoeck *et al.*, 2010; Balzano *et al.*, 2015). We also present a bioinformatics workflow designed for mixed 16S and 18S amplicons that generate ASVs differing by as little as a single base, and reproducibly recovers the exact known sequences from the mock communities. This workflow, which uses common tools such as cutadapt (Martin, 2011), bbtools (<http://sourceforge.net/projects/bbmap/>), DADA2 (Callahan *et al.*, 2016), deblur (Amir *et al.*, 2017) and QIIME 2 (Bolyen *et al.*, 2018), simplifies sequence analysis for mixed 16S and 18S amplicons. We also rigorously examined biases between 16S and 18S amplicons at the PCR and sequencing steps. Lastly, we analyzed natural marine samples collected from the San Pedro Ocean Time-series (SPOT) using the validated workflow to compare the performance of different primer sets.

Results and discussion

Comparison of universal primers (515Y/926R) and eukaryote-specific primers (V4F/V4R and V4F/V4RB) with 18S mock communities

Our 18S mock communities are mixtures of a number of nearly full-length 18S rRNA genes designed to represent the major eukaryotic groups found in marine environments. Among them, a clone in the *Prymnesiales* (haptophyta) has a single mismatch to the reverse primer V4R (at the 3' end), and three *Dinophyta* species

(*Lingulodinium*, Dino-Group-II_b and *Gymnodinium*) have a single mismatch to the reverse primer 926R. Separate mock communities were developed with members in equal or staggered concentrations to allow for deeper assessment of PCR, sequencing, or bioinformatic pipelines. As the abundances of taxa in mock communities are known *a priori*, they can be used to test which primer set and denoising algorithm recover the community composition most accurately.

For 18S even mock communities, V4F/V4R (Stoeck *et al.*, 2010) underestimated *Prymnesiales* (haptophyta) by ~fourfold, presumably because of a single base mismatch on the 3' end of the reverse primers (Fig. 2A). On

the other hand, the V4F/V4RB (Balzano *et al.*, 2015) primers that do not have any mismatches overestimated *Prymnesiales* (haptophyta) by ~fourfold (Fig. 2B) while the 515Y/926R primers produced a community composition similar to that expected (Fig. 2C).

For 18S staggered mock communities, similar results were found. V4F/V4R underestimated *Prymnesiales* (haptophyta) by ~fivefold (Fig. 3A), and V4F/V4RB overestimated *Prymnesiales* (haptophyta) by ~threefold (Fig. 3B). 515Y/926R underestimated three *Dinophyta* species (with single primer mismatches) to varying degrees (*Lingulodinium*, ~eightfold; Dino-Group-II_b, ~threefold; *Gymnodinium*, ~fourfold) (Fig. 3C). However,

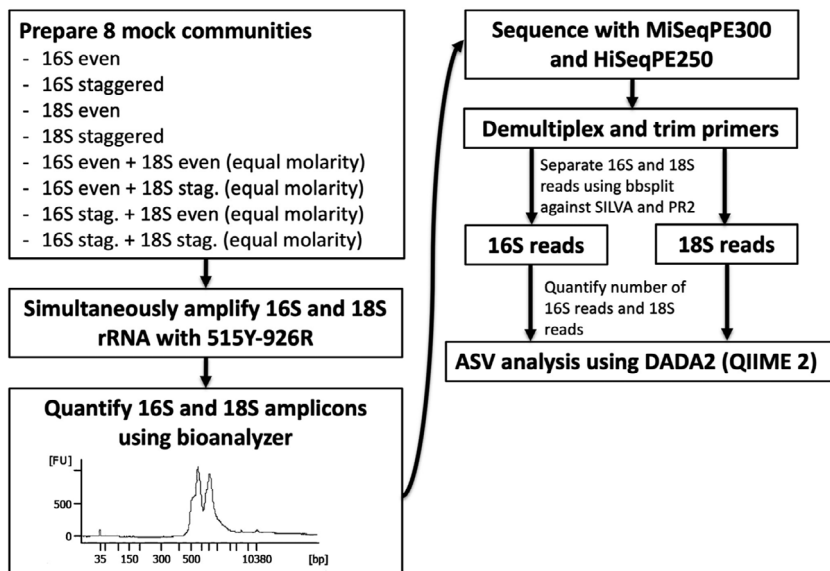


Fig. 1. Experimental design. Eight mock communities were amplified using the 515Y/926R primers. The quantity of amplicon DNA was measured with a Bioanalyzer, allowing quantification of PCR bias against longer 18S amplicons. After sequencing, 16S and 18S reads were then separated through an *in silico* sorting step, and the number of 16S and 18S reads were counted to quantify the sequencing bias against 18S. The 16S and 18S reads were then denoised separately using DADA2.

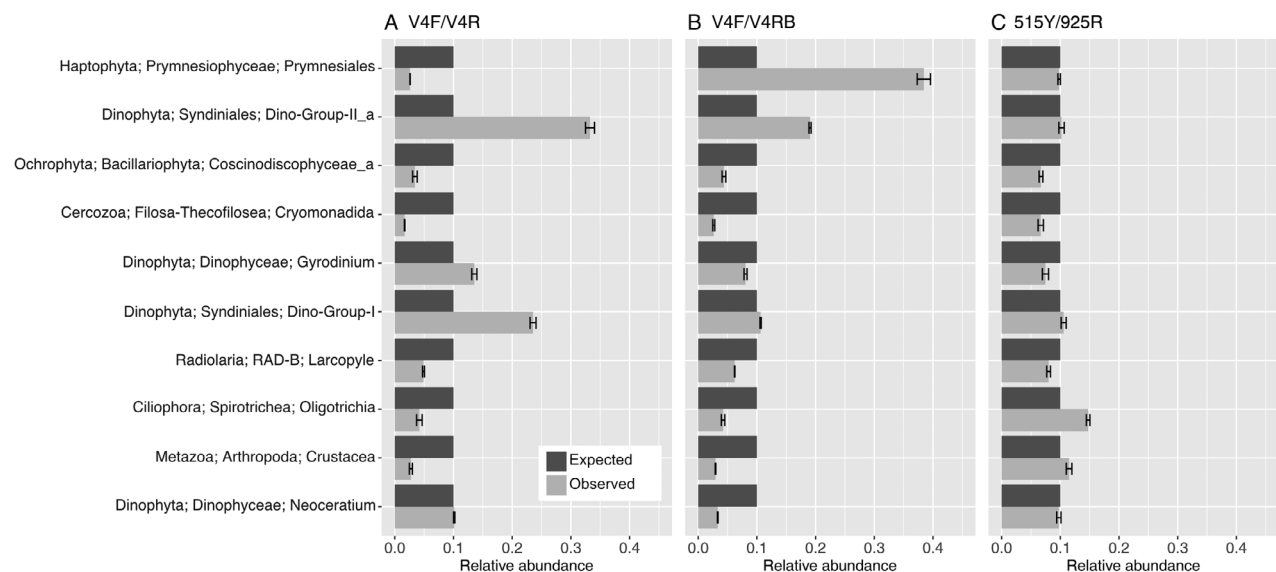


Fig. 2. Comparison of the proportions added (expected) vs. proportions observed in the sequence output of even 18S mock communities amplified with V4F/V4R (A), V4F/V4RB (B) and 515Y/926R (C).

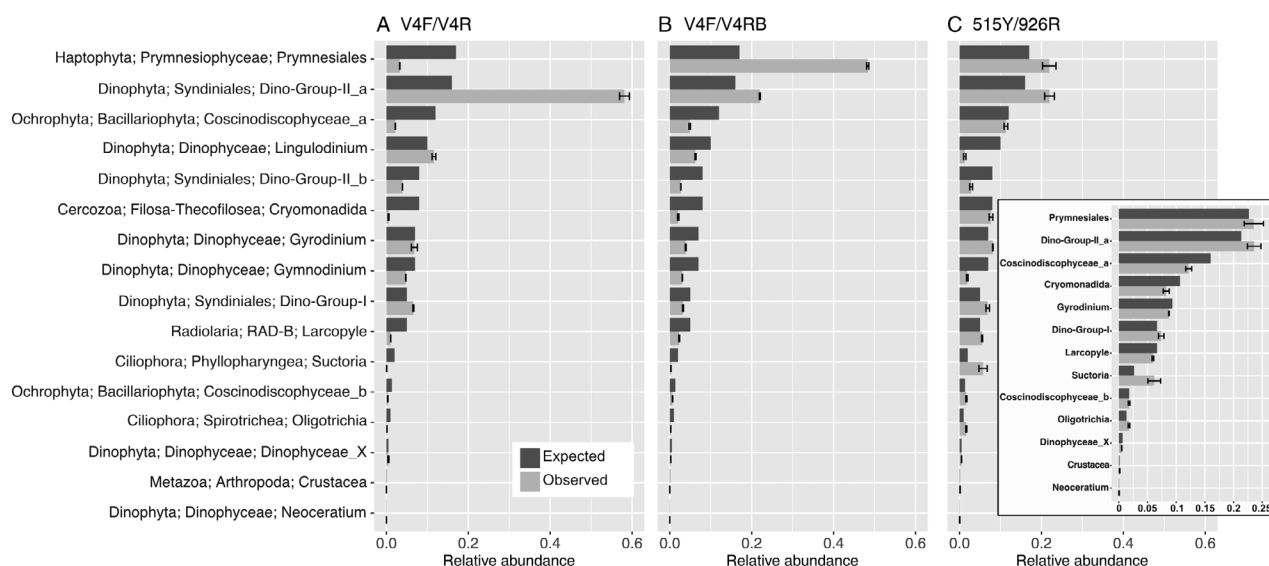


Fig. 3. Comparison the proportions added (expected) vs. proportion observed in the sequence outputs of staggered 18S mock communities amplified with V4F/V4R (A), V4F/V4RB (B) and 515Y/926R (C). The inset in (C) shows re-normalized staggered 18S mock community amplified with 515Y/926R after removing three mismatched *Dinophyta* species.

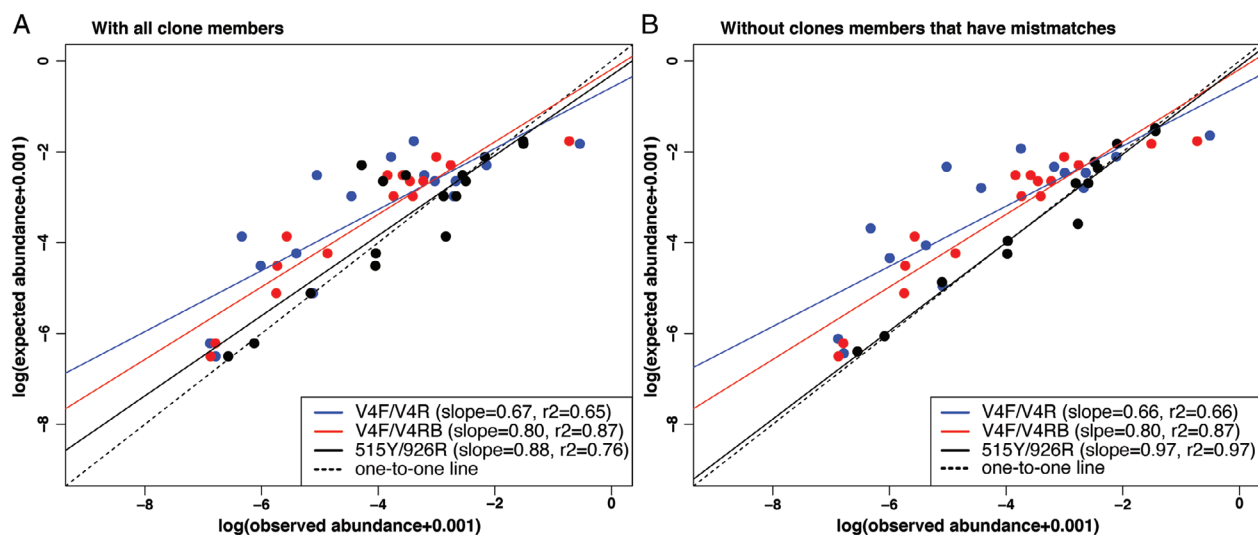


Fig. 4. Expected staggered 18S mock community abundance of each component plotted against observed staggered 18S mock community abundance from samples amplified with different primers pairs (A), and excluding clone members that have mismatches on the given primer pairs (B). Note the near-perfect slope (0.97) and r^2 (0.97) of the 515Y/926R primer pair for clones with no mismatches. [Color figure can be viewed at wileyonlinelibrary.com]

there was no relationship between degree of underestimation and locations of primer mismatch (*Lingulodinium*, -11 bases from the 3' end; *Dino-Group-II_b*, -12 bases from the 3' end; *Gymnodinium*, -2 bases from the 3' end).

Overall, the observed 18S mock community composition was more similar to the expected with 515Y/926R (slope = 0.88, r^2 = 0.76), especially after removing three mismatched *Dinophyta* species (slope = 0.97, r^2 = 0.97), followed by V4F/V4RB (slope = 0.79, r^2 = 0.87) and

V4F/V4R (slope = 0.67, r^2 = 0.65) (Fig. 4). With mixed mock communities, 16S and 18S mock communities were also recovered accurately (Fig. S1). These findings, together with the results of Parada *et al.* (2016), indicate that 515Y/926R primers recover both 16S and 18S mock communities quantitatively regardless of whether examined as separate or in combination.

In addition, our results indicated a threefold to eightfold underestimation when there was a primer mismatch. The

same issue was previously found with the original EMP primers (515C/806R, V4) that underestimated SAR11 by eightfold (Apprill *et al.*, 2015). Bru *et al.* (2008) found that underestimation generally increased as mismatches were closer to the 3' end of the primer, yet there was no predictable relationship between the position of mismatch and the degree of underestimation, which is consistent with our findings. The worst mismatches are at the 3' end of the primers, as occurs with the V4R primer (Stoeck *et al.*, 2010) for many common haptophytes. This observation was the rationale for the creation of the V4RB primer with a 3' degeneracy (Balzano *et al.*, 2015) that greatly improves recovery of haptophytes that are often dominant in seawater (Berdjeb *et al.*, 2018). However, we found that, instead of recovering 18S mock communities as expected, V4F/V4RB overestimated haptophytes by threefold to fourfold. Since there was no primer mismatch bias and all the amplicons were analyzed in the same sequencing run, a possible source of such bias might be differences in PCR protocols (1-step PCR with 515Y/926R vs. 2-step PCR with V4F/V4RB), but it is not understood why such a strong positive bias would occur with haptophytes and not the other taxa we examined.

Estimation of PCR and sequencing bias against 18S amplicons using mixed mock communities

To test for length-based PCR bias against longer 18S reads, 18S mock communities were mixed with 16S mock communities in equimolar amounts prior to PCR amplification. The mixed mock communities were then PCR amplified, products analyzed on an Agilent 2100 Bioanalyzer and then sequenced (Fig. 1). Based on bioanalyzer traces that separately quantify the abundance of 16S and 18S amplicons, there was little systematic PCR bias (about 0.7–1.3-fold) against 18S PCR products when using the 18S even mock communities that have no primer mismatches to 515Y/926R (Fig. 5, circle and asterisk, x-axis only). When the 18S staggered mocks were included (with three *Dinophyta* species that have one mismatch to the reverse primer, 926R), there was considerably more PCR bias, up to threefold (Fig. 5, triangle and square, x-axis only). The mixed amplicons were then sequenced and split into 16S and 18S reads pools by an *in silico* sorting step. By comparing ratios in the bioanalyzer outputs and the raw read counts after *in silico* sorting, we observed that there was typically a twofold sequencing discrimination against 18S reads (Fig. 5), which is consistent regardless of community types (even, staggered) and sequencing runs. That suggests sequencing bias due to length differences is a consistent property of the Illumina sequencing platform, yet PCR bias due to primer mismatches is much less predictable. Thus, an evaluation of primer coverage across

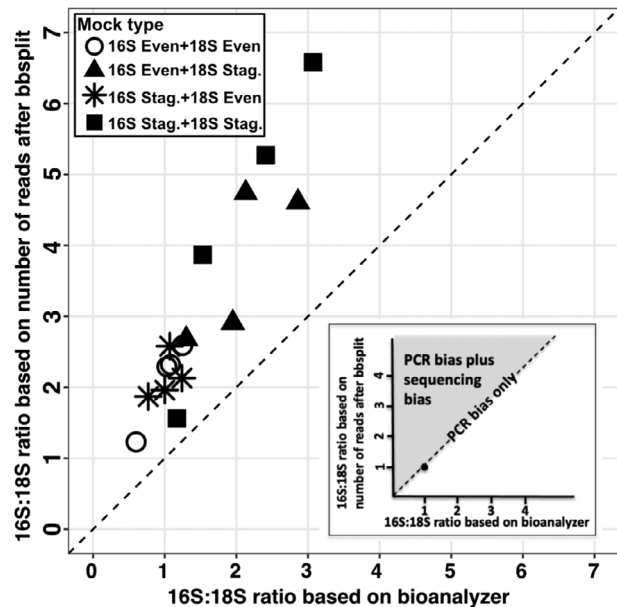


Fig. 5. Comparison of PCR and sequencing biases among four mixed 16S and 18S mock communities, each combined in 1:1 M ratios. The x-axis shows ratios in the PCR products, and the y-axis shows the ratios in the final sequences, including biases from PCR plus sequencing. Hypothetically, if there is no bias, all the mixed mocks would be located at a single point (1,1). If there is only PCR bias, all the data points would be at the one-to-one line. If there are PCR and sequencing bias, all the data points would be located above the one-to-one line (gray area). The slope indicates the sequencing bias. The data points all occur above the dashed one-to-one line, indicating most biases are from sequencing. Note for 18S even mocks (circle and asterisk), none of which have primer mismatches, the PCR products have a bioanalyzer output ratio near 1, indicating little PCR bias. The staggered 18S mocks (triangle and square) include three members with primer template mismatches and correspondingly more PCR bias visible on the x-axis. In all cases the final reads show about twofold more bias than the PCR biases alone, suggesting a twofold sequencing bias against 18S.

three domains, in actual field samples, may help better account for the PCR bias. Parada *et al.* (2016) found that 515Y/926R perfectly matches 86% of eukaryotes, 87.9% of bacteria and 83.9% of archaea in the SILVA database, but we note that in actual practice the extent of mismatches in field samples depends on the particular taxa present and their proportions (McNichol *et al.*, in press). We should also note that our 18S mock communities are very rich in alveolates such as dinoflagellates (3 of 10 in even, 7 of 16 in staggered) that tend to have mismatches to the 515Y/926R primers; hence they probably overestimate the biases expected in most field samples.

Comparison of universal primers (515C/926R) and eukaryote-specific primers (V4F/V4RB) with field samples

A previously published daily time series was used to compare outcomes with different primers amplifying

either 16S + 18S or 18S alone from the same DNA extracts (Needham and Fuhrman, 2016; Berdjeb *et al.*, 2018). This time series covered a spring bloom through summer at the San Pedro Ocean Time-Series off of Southern California, thus the comparison was evaluated under a wide range of environmental conditions and biological diversity.

We note that the universal primer in that paper (515C) was slightly different from that tested here with 18S mock communities (the fourth base on the 5' end of the forward primer was C instead of Y, where Y is a mixture of C and T). We thus used a recently reported dataset (McNichol *et al.*, in press) to compare the primer coverage. McNichol *et al.* (in press) have compared 515Y with SSU rRNA sequences retrieved from several marine metagenomic datasets (BioGEO TRACES, Malaspina, MBARI and TARA). Their results showed that 88% of eukaryotic 18S rRNA sequences and 99% of cyanobacteria and chloroplast 16S rRNA sequences perfectly matched 515Y. We further examined whether these sequences perfectly matched 515C or 515T. The results showed that >97% of the sequences perfectly matched 515C, and the incremental improvement of also considering a T at that position only yielded 0%–0.1% additional perfect matches (Table S3), suggesting the results from both primers should be comparable. Note that the 515Y primer simply adds a single degeneracy (primer versions with a

C and T at that position are equally present), so will perfectly match better than 515C alone.

To examine how alpha diversity differed with primer sets, we first rarefied sequences to the sample with the fewest 18S sequences (1155 reads) and repeated this 100 times for each primer set. The mean rarefied richness (i.e. number of observed ASVs) of the samples amplified with V4F/V4RB was significantly higher than those amplified with 515C/926R (Welch's *t*-test, $p < 0.001$; 30–217 for 515C/926R vs. 104–318 for V4F/V4RB; Fig. 6). The mean rarefied Shannon index values, however, were similar between these primer sets (Welch's *t*-test, $p > 0.05$; 0.97–4.79 for 515C/926R vs. 1.70–5.05 for V4F/V4RB; Fig. 6). We next evaluated the primer effects on beta diversity; cluster analysis showed that 515C/926R and V4F/V4RB detected a significantly similar temporal variation (Mantel test, $r = 0.95$, $p < 0.001$), i.e. similar overall clustering, in community composition. Starting from spring bloom in early March, followed by a post-bloom period in late March, a transition during April and May to summer in July (Fig. 7).

To test the extent that these two primer sets (515C/926R and V4F/V4RB) amplify similar communities at the ASV level, the 220 bp overlapping region of the forward reads was used for detailed examination (noting the forward primers are offset by four bases, with V4F going four bases further towards the 3' end vs. 515C). After

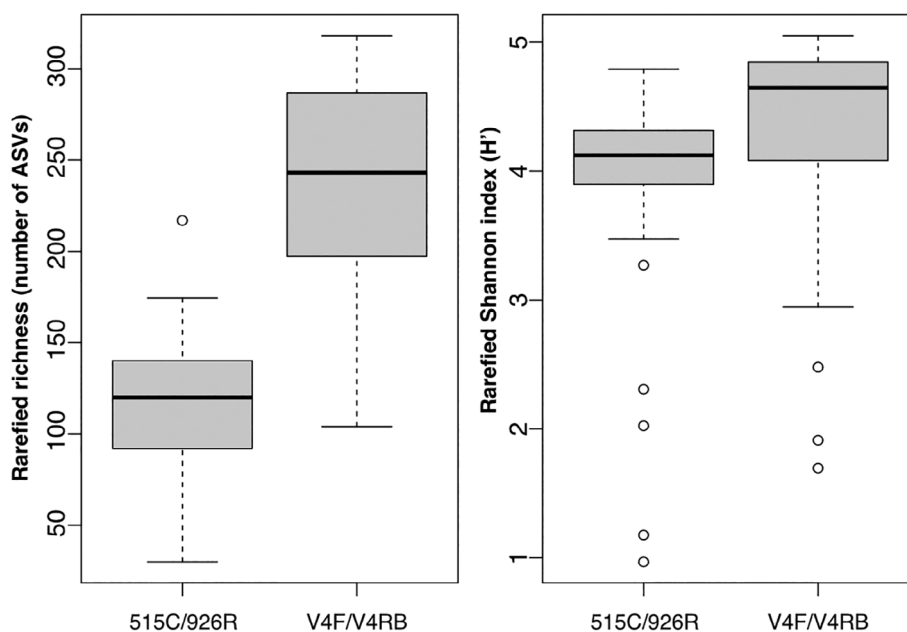


Fig. 6. The mean rarefied richness (number of ASVs, left) and Shannon Index (H' , right) of eukaryotic communities from a March–July daily time series off of Southern California. Sequences from each sample were rarefied to the sample with fewest 18S sequences (1681 reads) and repeated 100 times to obtain mean rarefied richness and Shannon Index.

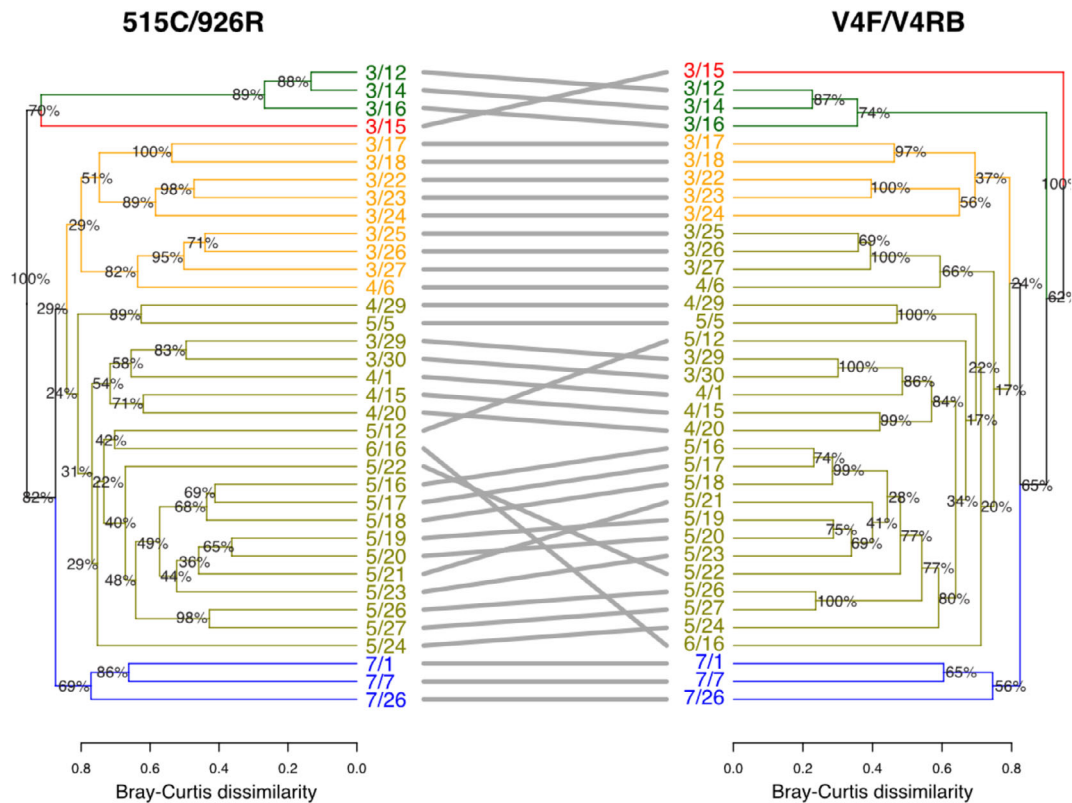


Fig. 7. Bootstrapped average-linkage clustering of eukaryotic communities from a March–July time-series off of Southern California. Eukaryotic communities were amplified by universal primers (left) and eukaryote-specific primers (right), denoised using DADA2 and then clustered based on the Bray–Curtis dissimilarity. The crossing lines indicate samples that shifted in clustering order; note that shifts are generally in deep branches and do not greatly change the overall community clustering patterns. Different colors represent different clusters. Sampling dates are shown as month/day. [Color figure can be viewed at wileyonlinelibrary.com]

rarefying samples amplified with 515C/926R and V4F/V4RB to the sample with fewest 18S reads (1681 reads) 100 times, on average 2741 ASVs were detected across the time series, and 1131 ASVs were shared between the primer sets (Fig. S2). These shared ASVs contributed to 80%–100% of the total sequences in the communities amplified with 515C/926R and 87%–97% of sequences in communities amplified with V4F/V4RB. A total of 254 ASVs were unique to the samples amplified with 515C/926R, and 1412 ASVs were only found in the samples amplified with V4F/V4RB (Fig. S2). A direct comparison showed 515C/926R and V4F/V4RB detected the same abundant ASVs with similar relative abundances, while there were more differences in rare ASVs, with the V4F/V4RB typically detecting more of these taxa (Fig. 8 and Fig. S3). The relative abundances of ASVs missed by 515C/926R were all found to be rare (less than 1.5%) by V4F/V4RB, whereas the relative abundances of some ASVs missed by V4F/V4RB were more abundant (more than 5%) by 515C/926R (Fig. 8). A taxonomic comparison shows that under the same sampling effort, there were differences at the order level between primer sets (Fig. S2). Three orders (Cryptomonas,

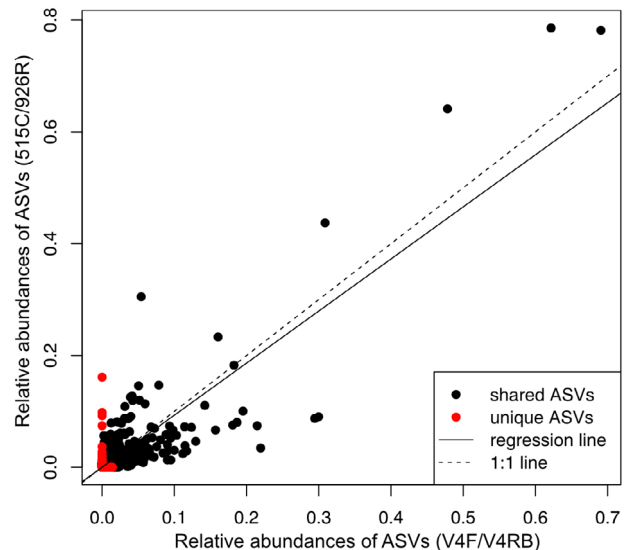


Fig. 8. The mean relative abundances of each 18S ASV amplified with V4F/V4RB plotted against the relative abundances of the same ASV amplified with 515C/926R, based only on the 220 bp overlapping region from universal (515C/926R) and eukaryote-specific primers (V4F/V4RB). Sequences from each sample were rarefied to the sample with fewest 18S reads (1681 reads) repeated 100 times. [Color figure can be viewed at wileyonlinelibrary.com]

Discicristata, MAST-6) were uniquely found in the samples amplified with 515C/926R, and two orders (Coccolithales, Hemiselmis) were unique to the samples amplified with V4F/V4RB. Thus, while the V4 primers tended to yield more of the rarer ASVs, neither primer set was much more comprehensive taxonomically than the other, and the two together yielded a broader overall diversity than either alone.

Clustering of all these overlapping sequences analyzed together showed a similar spring–summer transition (Fig. S4), with samples usually clustering by date rather than by primer pair. However, individual dates typically had a 30%–50% Bray–Curtis distance between data analyzed by the two primers, indicating significant differences in quantitative composition. While lacking independent knowledge of the actual taxa distribution, we note observed distributions within mock communities more closely matched the expected outcome when using the universal primers, especially for sequences with no mismatches (Fig. 4).

Our results indicated that beta diversity may be less sensitive to primer bias than alpha diversity, which has been previously reported (Caporaso *et al.*, 2012; He *et al.*, 2013; Tremblay *et al.*, 2015). Comparing the 220 bp overlapping region amplified by both primer sets demonstrated that the variation in community composition due to primer sets comes mostly from the rare taxa, perhaps in part from PCR and/or sequencing errors (He *et al.*, 2013). Notably, the V4F/V4RB amplification requires a two-step PCR amplification, with more opportunities for errors and/or biases (Yu *et al.*, 2015).

The application of universal primers (515Y/926R) to three-domain amplicon analysis

Quantitative 16S/18S biases determined by mixed mock community analyses were ‘corrected’ in field samples to make current best-guess estimates of the true relative proportions of 18S, chloroplast and prokaryotic 16S gene abundances. With the mock communities, we found an overall twofold bias against 18S at the sequencing step, and we can use that as a starting point for making corrections. Applying this twofold bias, data from the protist-enriched 1.2–80 µm fraction of the spring–summer SPOT time-series samples would yield an average of 35% 18S, 28% chloroplast 16S and 37% prokaryote 16S rRNA gene amplicons (Fig. S5) - in other words, an almost even split in these three categories. Future work will help determine to what extent the twofold bias applies in general, but because some 18S sequences are much longer than others (Obiol *et al.*, 2020); it is quite possible the biases are worse in some samples and for some taxa, compared to others. A direct measure of average biases from each sample should be possible by quantifying DNA

in the 16S and 18S amplicons before sequencing and then comparing the actual 16S and 18S sequences in the final outputs. The read composition (Fig. S5) and rarefaction curves (Fig. S6) constructed for each field sample together indicated that for the 1.2–80 µm size fraction collected from the SPOT location, about 15 000–70 000 total sequences are required to effectively sample the true richness of marine planktonic prokaryotes, phytoplankton and heterotrophic eukaryotes in a single PCR reaction (Fig. S6, detailed calculation is described in the figure legend). For studies collecting whole seawater (>0.2 µm) that include a larger fraction of prokaryotes, we also estimated the required sequencing depth using a previously reported dataset from the BioGEOTRACES GA03 trans-Atlantic expedition (Biller *et al.*, 2018; McNichol *et al.*, in press). The McNichol *et al.* study amplified DNA collected from 100 ml of whole seawater using the universal primers (515Y/926R), analyzed using DADA2. With the read composition and the rarefaction curves of the Atlantic euphotic seawater samples, we calculated that about 28 000–110 000 total sequences are required to capture the diversity of prokaryotes and eukaryotes (Fig. S7, detailed calculation is described in the figure legend). However, the number would vary with different locations, size fractions, sampling volumes, extraction methods, and analysis pipelines.

Conclusions and future prospects

This study shows that the three-domain universal primer (515Y/926R) can resolve community composition for 16S and 18S rRNA in a single PCR reaction, with biases we could quantify and manage. We were able to investigate the biases relevant to the use of these primers in a natural setting through the use of 18S mock communities, separately and in concert with 16S mocks. With field samples, the universal primers (515C/926R) detected similar community composition and beta-diversity patterns as commonly used eukaryote-specific primers (V4F/V4RB). However, the abundance of several taxa varied with primer set (notably with the V4F/V4RB primers yielding more rare eukaryotic taxa), though without independent data, we cannot assume that reporting more taxa is necessarily more accurate.

Comprehensive simultaneous three-domain analysis has three potential advantages over single-domain analyses for determining microbial community composition. First, there is the obvious advantage of directly comparing 16S and 18S sequence abundances, which can now be corrected (to some extent) for biases as we have described. Even without absolute corrected gene counts, results allow for consistent comparisons of ratios between all taxa, across samples and even sample types; i.e. even without bias corrections, the relative

ratios are robust (McLaren *et al.*, 2019). Second, it provides an independent analysis of phototrophic protists. As chloroplast 16S rRNA gene databases are constantly growing, the chloroplast 16S genes amplified with 515Y/926R can help identify (or verify the identities of) phototrophic eukaryotes, providing a way to characterize phytoplankton communities independent of 18S and known wide variability in 18S per-cell copy number variations, which range over 10 000-fold (see also Needham and Fuhrman (2016)). Chloroplast 16S data may more closely reflect phytoplankton biomass distributions than do 18S data and are being increasingly used in biological oceanographic studies, sometimes with higher phylogenetic resolution than 18S (Needham and Fuhrman, 2016; Bennke *et al.*, 2018; Bolaños *et al.*, 2020; Choi *et al.*, 2020). Lastly, a single universal amplification reduces some major costs associated with amplicon analysis. As sequencing continues to drop in price per base, the major expense per sample comes from PCR enzymes, clean-up beads, and labor required for quantification, dilution, gel imaging, etc. Compared with single-PCR 16S and 18S rRNA community analysis, using separate primers for 16S and 18S assays (noting V4 18S needs two-step PCR) increases amplicon library preparation costs twofold to threefold, which can exceed the costs of increased coverage in a single universal assay to yield the desired number of 18S sequences. Overall, this method provides a feasible path for making quantitative rRNA gene-based assessments of microbial communities across three domains using amplicon data, when proper validation such as from mock communities is employed.

Materials and methods

Mock community preparation. For 16S mock communities, nearly full-length marine 16S rRNA genes were prepared as previously described (Parada *et al.*, 2016; Yeh *et al.*, 2018). For 18S mock communities, nearly full-length 18S rRNA clone libraries were prepared from the large size fraction (1.2–80 µm) of seawater sample collected from the SPOT location. The detailed preparation is described in the supplementary information. To mimic natural marine communities consisting of both eukaryotes and prokaryotes, 16S and 18S mock communities were mixed in four combinations (Fig. 1). Each mixed mock community was pooled at equal molarity after taking lengths into account; the average length of 16S mocks is 1425 bp and the average length of 18S mocks is 1770 bp, so resulting amplicons are internal to these lengths and therefore shorter.

Field samples

A daily-to-weekly time series used samples collected from the 5 m depth at the SPOT location from March 12 to July 26 in 2011. The methods and sequencing data have been previously published under accession numbers PRJEB12215 (universal primers) and PRJEB10834 (18S primers) (Needham and Fuhrman, 2016; Berdjeb *et al.*, 2018).

PCR and sequencing. To pool multiple samples in a single Illumina paired-end sequencing platform, a dual-index sequencing strategy was used with forward primer (*A-I-NNNN-barcode-loci specific forward primer*) and reverse primer (*A-index-I-loci specific reverse primer*), where *A* is the Illumina sequencing adapter, *I* is the Illumina primer, and barcode and index are sample-specific tags (5 bp barcode and 6 bp index). A detailed protocol is available at doi.org/10.17504/protocols.io.vb7e2m. To compare 16S/18S universal primers with eukaryote-specific primers, mock communities were amplified with 515Y (5'-GTGYCAGCMGCCGCGTAA-3') and 926R (5'-CCGYCAATYMTTTRAGTTT-3'), V4F (5'-CCAGCASCYCGCGTAATTCC-3') and V4R (5'-ACTTTTCGTTCTTGATYRA-3') and V4F and V4RB (5'-ACTTTCGTTCTTGATYRR-3') (Stoeck *et al.*, 2010; Balzano *et al.*, 2015; Parada *et al.*, 2016). The only difference between V4F/V4R and V4F/V4RB is the last base on the 3' end of the reverse primer (A to R), which corrects a mismatch, allowing V4F/V4RB amplify haptophytes and some other taxa better (Balzano *et al.*, 2015). The amplification conditions for each primer pair are described in the supplementary information. Purified PCR products were quantified with PicoGreen and sequenced on Illumina HiSeq 2500 in PE250 mode and MiSeq PE300.

In silico processing of amplicon sequences. Sequences were demultiplexed by forward barcodes and reverse indices allowing no mismatches using QIIME 1.9.1 `split_libraries_fastq.py`. The fully demultiplexed forward and reverse sequences were then split into per-sample fastq files using QIIME 1.9.1 `split_sequence_file_on_sample_ids.py` and submitted to the EMBL database under accession number PRJEB35673.

Scripts necessary to reproduce the following analysis are available at github.com/jcmcnch/eASV-pipeline-for-515Y-926R. Demultiplexed amplicon sequences were trimmed with `cutadapt`, discarding any sequence pairs not containing the forward or reverse primer. We allowed an error rate of up to 20% to retain amplicons with mismatches to the primer. Similar to the workflow proposed Mike Lee (https://astrobiomike.github.io/amplicon/16S_and_18S_mixed), mixed amplicon sequences were split

into 16S and 18S pools using `bbsplit.sh` from the `bbtools` package (<http://sourceforge.net/projects/bbmap/>) against curated 16S/18S databases derived from SILVA 132 (Quast *et al.*, 2013) and PR2 (Guillou *et al.*, 2013). The splitting databases used are available at <https://osf.io/e65rs/>. The two amplicon categories were then analyzed in parallel using `qiime2` (Bolyen *et al.*, 2019) or DADA2 implemented as the standalone R package (Callahan *et al.*, 2016) as described in the supplementary information. The results were all based on DADA2 (QIIME 2) that worked best for each type of sequence after comparing several denoising algorithms (a detailed comparison is described in the supplementary information).

PCR and sequencing bias estimation. 16S and 18S mixed mock communities amplified with 515Y/926R were run on an Agilent 2100 Bioanalyzer to quantify concentrations of 16S and 18S PCR products in each mixed mock community. Amplicons were analyzed with the High-sensitivity DNA assay kit according to the manufacturer's instructions. Due to the length differences between 16S and 18S amplicons, the concentrations of amplicons were measured by quantifying peak areas on an Agilent 2100 Bioanalyzer using automatic peak detection without altering the instrument-determined baseline. The 16S:18S ratio of molarity was used to determine PCR bias. Sequence pre-processing (i.e. `bbsplit.sh`) split reads into 16S and 18S pools. The 16S:18S ratio based on the number of reads was used to determine sequencing and PCR bias. The slope of the line derived from plotting the 16S:18S ratio from Bioanalyzer traces against 16S:18S ratio based on the number of reads after the `bbsplit` step was used to define sequencing bias.

Acknowledgements

This work was supported by NSF OCE 1737409, Gordon and Betty Moore Foundation Marine Microbiology Initiative grant 3779 and Simons Foundation Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems (CBIOMES) grant 549943. We thank Julio C. Ignacio-Espinoza, Alexandra Santora, Colette Fletcher-Hoppe, Delaney Nolin, Jake Weissman, Melody Aleman, Shengwei Hou and Sarah Laperriere for help with sampling, laboratory work and feedback on the manuscript.

References

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., and Huse, S.M. (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.

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Xu, Z.Z., *et al.* (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems* **2**: e00191–16

Apprill, A., McNally, S., Parsons, R., and 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.

Balzano, S., Abs, E., and Leterme, S.C. (2015) Protist diversity along a salinity gradient in a coastal lagoon. *Aquat Microb Ecol* **74**: 263–277.

Benneke, C., Pollehne, F., Müller, A., Hansen, R., Kreikemeyer, B., and Labrenz, M. (2018) The distribution of phytoplankton in the Baltic Sea assessed by a prokaryotic 16S rRNA gene primer system. *J Plankton Res* **40**: 244–254.

Berdjeb, L., Parada, A., Needham, D.M., and Fuhrman, J.A. (2018) Short-term dynamics and interactions of marine protist communities during the spring–summer transition. *ISME J* **12**: 1907.

Billar, S.J., Berube, P.M., Dooley, K., Williams, M., Satinsky, B.M., Hackl, T., *et al.* (2018) Marine microbial metagenomes sampled across space and time. *Sci data* **5**: 1–7.

Bolaños, L.M., Karp-Boss, L., Choi, C.J., Worden, A.Z., Graff, J.R., Haëntjens, N., *et al.* (2020) Small phytoplankton dominate western North Atlantic biomass. *ISME J* **14**: 1–12.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., and Asnicar, F. (2018) QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints*.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**: 852–857.

Bradley, I.M., Pinto, A.J., and Guest, J.S. (2016) Design and evaluation of Illumina MiSeq-compatible, 18S rRNA gene-specific primers for improved characterization of mixed phototrophic communities. *Appl Environ Microbiol* **82**: 5878–5891.

Bru, D., Martin-Laurent, F., and Philippot, L. (2008) Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. *Appl Environ Microbiol* **74**: 1660–1663.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**: 581–583.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., *et al.* (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**: 1621–1624.

Catlett, D., Matson, P.G., Carlson, C.A., Wilbanks, E.G., Siegel, D.A., and Iglesias-Rodriguez, M.D. (2020) Evaluation of accuracy and precision in an amplicon sequencing workflow for marine protist communities. *Limnol Oceanogr Methods* **18**: 20–40.

Chénard, C., Wijaya, W., Vaultot, D., dos Santos, A.L., Martin, P., Kaur, A., and Lauro, F.M. (2019) Temporal and spatial dynamics of Bacteria, Archaea and protists in equatorial coastal waters. *Sci Rep* **9**: 1–13.

- Choi, C.J., Jimenez, V., Needham, D., Poirier, C., Bachy, C., Alexander, H., et al. (2020) Seasonal and geographical transitions in eukaryotic phytoplankton community structure in the Atlantic and Pacific Oceans. *Front Microbiol* **11**: 2187.
- De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., et al. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.
- Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., and Sogin, M.L. (2015) Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* **9**: 968–979.
- Fuhrman, J.A., Cram, J.A., and Needham, D.M. (2015) Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol* **13**: 133–146.
- Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., et al. (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41**: D579–D604.
- He, Y., Zhou, B.-J., Deng, G.-H., Jiang, X.-T., Zhang, H., and Zhou, H.-W. (2013) Comparison of microbial diversity determined with the same variable tag sequence extracted from two different PCR amplicons. *BMC Microbiol* **13**: 1–8.
- Kittelmann, S., Seedorf, H., Walters, W.A., Clemente, J.C., Knight, R., Gordon, J.I., and Janssen, P.H. (2013) Simultaneous amplicon sequencing to explore co-occurrence patterns of bacterial, archaeal and eukaryotic microorganisms in rumen microbial communities. *PLoS One* **8**: e47879.
- Lin, Y., Gifford, S., Ducklow, H., Schofield, O., and Cassar, N. (2019) Towards quantitative microbiome community profiling using internal standards. *Appl Environ Microbiol* **85**: e02634–e02618.
- Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**: 10–12.
- McLaren, M.R., Willis, A.D., and Callahan, B.J. (2019) Consistent and correctable bias in metagenomic sequencing experiments. *Elife* **8**: e46923.
- McNichol, J.C., Berube, P.M., Biller, S.J., and Fuhrman, J.A. (in press) Evaluating and Improving SSU rRNA PCR Primer Coverage for Bacteria, Archaea, and Eukaryotes Using Metagenomes from Global Ocean Surveys. *mSystems*.
- Needham, D.M., Fichot, E.B., Wang, E., Berdjeb, L., Cram, J.A., Fichot, C.G., and Fuhrman, J.A. (2018) Dynamics and interactions of highly resolved marine plankton via automated high-frequency sampling. *ISME J* **12**: 2417.
- Needham, D.M., and Fuhrman, J.A. (2016) Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat Microbiol* **1**: 16005.
- Needham, D.M., Sachdeva, R., and Fuhrman, J.A. (2017) Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows micro-diversity matters. *ISME J* **11**: 1614–1629.
- Obiol, A., Giner, C.R., Sánchez, P., Duarte, C.M., Acinas, S. G., and Massana, R. (2020) A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Mol Ecol Resour* **20**: 718–731.
- Parada, A.E., Needham, D.M., and Fuhrman, J.A. (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.
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590–D596.
- Quince, C., Lanzen, A., Davenport, R.J., and Turnbaugh, P. J. (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12**: 38.
- Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., et al. (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proc Natl Acad Sci U S A* **103**: 12115–12120.
- Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D., Breiner, H.W., and Richards, T.A. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* **19**: 21–31.
- Tremblay, J., Singh, K., Fern, A., Kirton, E.S., He, S., Woyke, T., et al. (2015) Primer and platform effects on 16S rRNA tag sequencing. *Front Microbiol* **6**: 771.
- Wear, E.K., Wilbanks, E.G., Nelson, C.E., and Carlson, C.A. (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.
- Yeh, Y.-C., Needham, D.M., Sieradzki, E.T., and Fuhrman, J.A. (2018) Taxon disappearance from microbiome analysis reinforces the value of mock communities as a standard in every sequencing run. *MSystems* **3**: e00023–e00018.
- Yu, G., Fadrosh, D., Goedert, J.J., Ravel, J., and Goldstein, A.M. (2015) Nested PCR biases in interpreting microbial community structure in 16S rRNA gene sequence datasets. *PLoS One* **10**: e0132253.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1: Supporting Information.