Comprehensive single-PCR 16S and 18S rRNA community analysis validated with mock communities and denoising algorithms

Yi-Chun Yeh[1], Jesse C. McNichol[1], David M. Needham[1,2], Erin B. Fichot[1], and Jed A. Fuhrman[1]

[1]Department of Biological Sciences, University of Southern California, Los Angeles, California 90089-0371, USA

[2]Current address: GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany, 24148

Correspondence:

Yi-Chun Yeh

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

Email: yichuny@usc.edu

Abstract

Universal SSU rRNA primers allow comprehensive quantitative profiling of natural communities by simultaneously amplifying templates from Bacteria, Archaea, and Eukaryota in a single PCR reaction. Despite the potential to show all rRNA gene relative gene abundances, they are rarely used due to concerns about length bias against 18S amplicons and bioinformatic challenges converting mixed 16S/18S sequences into amplicon sequence variants. We thus developed 16S and 18S rRNA mock communities and a bioinformatic pipeline to validate this three-domain approach. To test for length biases, we mixed eukaryotic and prokaryotic mocks before PCR, and found consistent two-fold underestimation of longer 18S sequences due to sequencing but not PCR bias. Using these mocks, we show universal V4-V5 primers (515Y/926R) outperformed eukaryote-specific V4 primers in observed vs. expected abundance correlations and sequences with single mismatches to the primer were strongly underestimated (3-8 fold). A year of monthly time-series data from a protist-enriched 1.2-80 µm size fraction yielded an average of 9% 18S, 17% chloroplast 16S, and 74% prokaryote 16S rRNA gene amplicons. These data demonstrate the potential for universal primers to generate quantitative and comprehensive microbiome profiles, although gene copy and genome size variability should be considered - as for any quantitative genetic analysis.

Introduction

Microbial communities of unicellular organisms are dynamic, diverse communities made up of bacteria, archaea, eukaryotes that interact with one another and their environment. Studying all these components together is essential for understand how the ecosystem functions as a whole (Fuhrman et al., 2015; Needham et

al., 2018; Chénard et al., 2019), though single components are most typically studied alone 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 hypervariable regions are mostly longer than early sequencing lengths (Amaral-Zettler et al., 2009). Until recently, with current sequencing capacities, two regions, V4 and V7-9, 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). To study metazoan host-dominated communities, however, a concern is that universal 16S&18S primers would mainly amplify 18S host sequences, overwhelming the 16S and non-host 18S microbiome, and thus blocking primers or universal non-metazoan primers are required in these studies (Del Campo et al., 2019).

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 (Apprill et al., 2015; Parada et al., 2016). In addition, validation and inter-comparison of primer performance has also been facilitated by the development and application of microbial internal standards or "mock communities" to PCR amplicon analysis (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 quantitatively recover natural abundance patterns (Parada et al., 2016). Including mocks in a sequencing run can also verify instrument performance, thus avoiding improper ecological conclusions. For example, mocks 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). 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). Unlike OTU clustering that must analyze sequences all together 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 (Callahan et al., 2016).

Collectively, these studies show how PCR amplicons can generate quantitative data that allows 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) that simultaneously amplifies 16S and 18S rRNA in a single PCR reaction. Because of their universal nature, these primers measure both eukaryotic microbes 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 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 with current Illumina technology, the forward and reverse reads do not overlap for the 18S amplicon (575-595 bp), meaning that they cannot be merged as is typical for 16S amplicon analysis. Second, PCR and sequencing both discriminate against longer amplicons (Kittelmann et al., 2013), yet we lack a quantitative understanding of PCR and sequencing biases against longer 18S amplicons. 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). 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. We also present a bioinformatics workflow designed for mixed 16S and 18S amplicons that generates ASVs differing by as little as a single base, and reproducibly recovers the known exact

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, and allowed us to rigorously test the performance of two different denoisers (DADA2 and deblur) with a variety of different data types. We also rigorously examined biases between 16S and 18S amplicons at the PCR and sequencing steps. Lastly, we analyzed natural marine samples collected from San Pedro Ocean Time-series (SPOT) using the same workflow to examine the application of 515Y/926R to environmental samples.

Results

Effects of trim length on 18S denoising

Since 18S amplicons amplified with 515Y/926R are too long (~575-595 bp) for forward and reverse reads to overlap (even with 2 x 300bp sequencing), we decided to trim reads to fixed lengths before denoising. Trimmed reads are concatenated either before (q2-dada2 and q2-deblur QIIME 2 plugins) or after denoising (standalone DADA2 R package). As quality profiles of reverse reads vary widely among runs, a trim length which worked equally well across runs needed to be determined. We therefore systematically decreased the trim length of reverse reads from 220 to 100 bp while fixing trim length of forward reads at 220 bp. Three criteria were then used to compare denoiser performance; 1) percent reads that perfectly matched *in silico* sequences, 2) R-squared values obtained by plotting the expected abundance of staggered mock community against the sequenced staggered mock community on a log (x+0.001) scale, and 3) percent reads removed after denoising (Table 1). Deblur successfully recovered staggered mock communities in the proportions expected regardless which trim length

was chosen, and we did not observe any sequences without an exact match to the known reference sequences. DADA2, however, produced up to 0.5 % of reads that did not perfectly match the mock communities, though it performed slightly better when concatenating reads after denoising (0.3 % of reads did not perfectly match *in silico* sequences). By blasting these reads against *in silico* sequences, we found that these reads could be accounted for by sample bleed-through or contamination as they had more than 1-mismatch to the *in silico* sequences, which were less likely produced by PCR/sequencing errors. Although deblur never produced such putative erroneous ASVs, it removed a large fraction of reads (~75%) compared with DADA2 (~25%), yielding fewer observations in the final ASV table (Table 1). According to the three criteria defined above, denoiser performance was relatively consistent among runs at a trim length for the reverse read of 200 bp. Therefore, this length was used for the rest of the analysis.

Comparison of 16S and 18S 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 with known concentrations, and were designed to represent the major eukaryotic groups found in marine environments, including Haptophyta, Dinophyta, Ochrophyta, Ciliophora, Cercozoa, Radiolaria, and Metazoa. Among them, 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. 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 closely to what is expected (Fig. 2).

For 18S even mock communities, V4F/V4R underestimated Prymnesiales (Haptophyta) by ~4-fold, presumably because of a single nucleotide mismatch on the 3' end of the reverse primers (Fig. 2a) (Stoeck et al., 2010). On the other hand, the V4F/V4RB primers that do not have any mismatches overestimated Prymnesiales (Haptophyta) by ~4-fold (Fig. 2b) (Balzano et al., 2015) while the 515Y/926R primers produced a community composition similar to that expected (Fig. 2c). The patterns were consistent among denoising pipelines (Fig. 2).

For 18S staggered mock communities, similar results were found. V4F/V4R underestimated Prymnesiales (Haptophyta) by ~5-fold (Fig. 3a), and V4F/V4RB overestimated Prymnesiales (Haptophyta) by ~3-fold (Fig. 3b). 515Y/926R underestimated three Dinophyta species (with single primer mismatches) to varying degrees (Lingulodinium, ~8-fold; Dino-Group-II_b, ~3-fold; Gymnodinium, ~4-fold) (Fig. 3c). However, 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). The patterns were consistent among denoising pipelines (Fig. 3). Overall, the observed 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).

Estimation of PCR and sequencing bias against 18S reads in mixed mock communities

To test for length-based PCR bias against 18S reads, 18S mock communities were mixed with 16S mock communities in equimolar amounts prior to PCR amplification.

189 The mixed mock communities were then PCR amplified, products analyzed on a Agilent
190 2100 Bioanalyzer, and then sequenced (Fig. 1). The bioanalyzer is a electrophoretic
191 instrument that accounts for differences in sequence length in estimating the molarity
192 (i.e., copies of DNA per unit volume) for DNA inputs. Based on bioanalyzer traces that
193 separately quantify the abundance of 16S and 18S amplicons, there was little systematic
194 PCR bias (about 0.7-1.3-fold) against 18S PCR products when using the 18S even mock
195 communities that have no primer mismatches to 515Y/926R (Fig 5, orange and blue
196 dots, x axis only). When the 18S staggered mocks were included (with three Dinophyta
197 species that have one mismatch to the reverse primer, 926R), there was considerably
198 more PCR bias, up to 3-fold (Fig 6 green and purple dots, x axis only). The mixed
199 amplicons were then sequenced and split into 16S and 18S reads pools by an *in silico*
200 sorting step. By comparing ratios in the bioanalyzer outputs and the raw read counts
201 after *in silico* sorting, we observed that there was typically a 2-fold sequencing
202 discrimination against 18S reads (Fig. 5), which is consistent regardless of community
203 types (even, staggered) and sequencing runs.
204
205 **The application of the 515Y/926R primer pair to field samples**
206 To examine the application of universal primers (515Y/926R) to natural
207 samples, surface seawater samples from a larger size fraction (1.2-80 μm) collected
208 from the San Pedro Ocean Time-series (SPOT) location in 2014 were analyzed. On
209 average, 9% of reads were 18S, 17% of reads were plastidal 16S reads, i.e. chloroplasts
210 in photosynthetic eukaryotes (excluding dinoflagellates whose chloroplasts are not
211 detected, Needham and Fuhrman (2016)), and 74% of reads were prokaryotic 16S (Fig.
212 6). A total of 2394 ASVs were identified across all samples (540 18S ASVs were affiliated
213 to 228 orders; 442 plastidal 16S ASVs were affiliated to 81 orders; 1412 prokaryotic

16S ASVs were classified into 85 orders). Since Metazoa in this size fraction were patchy (maximum of 2% reads), mainly dominated by copepods Maxillopoda (Fig. 7a), they were separately analyzed in the community composition of 18S reads (Fig. 7b). 18S reads were dominated by Dinophyceae, Spirotrichea, Syndiniales, Ciliophora, Mamiellophyceae, MAST, Bacillariophyta, RAD-B, Prymnesiophyceae, and Polycystinea (Fig. 7b). For plastidal 16S reads, phytoplankton communities were dominated by Prymnesiophyceae, Mamiellophyceae, Pelagophyceae, Dictyochophyceae, Bacillariophyta, Cryptophyceae, Chrysophyceae-Synurophyceae, Raphidophyceae, Prasinophyceae, Bolidophyceae, and Chrysophyceae (Fig. 7c). The prokaryotic 16S reads showed that prokaryotic communities were mainly dominated by SAR11, Synechococcales (i.e. Cyanobacteria), Flavobacteriales, Rhodobacterales, Actinomarinales, Verrucomicrobiales, Puniceispirillales, Rhodospirillales, Opitutales, and Cellvibrionales. (Fig. 7d). The phytoplankton and prokaryotic communities had similar seasonal patterns, whereas non-phytoplankton eukaryotic communities were less predictable.

Discussion

This study shows that the 3-domain universal primer (515Y/926R) can resolve community composition quantitatively 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 first applied here, separately and in concert with 16S mocks.

Unlike 16S rRNA sequencing, 18S rRNA sequencing using 515Y/926R is bioinformatically challenging because the amplicon is too long (~575-595 bp) for forward and reverse reads to overlap with current Illumina sequencing capacities. A

simplified solution in such a situation might be to use forward reads only instead of merged paired-end reads, because the quality of reverse reads is generally worse than forward reads, and errors near the 3' end cannot be corrected without overlapping paired-end reads. However, using only forward reads sacrifices phylogenetic resolution. Therefore, acquiring paired-end information without producing extra artifacts becomes critical for 18S rRNA processing. To do so, we developed a bioinformatic workflow which allowed us to split mixed 16S/18S amplicons into two sequence pools by mapping reads to a curated 16S/18S reference database derived from SILVA and PR2. The workflow was validated with mixed mock communities, showing that the *in silico* sorting step is able to successfully separate 16S and 18S mocks apart without changing their composition. We analyzed 18S mock communities by trimming reads to fixed lengths before denoising. In this way, we not only removed error bases but also kept 18S ASVs comparable between runs. Our analysis showed that we were able to recover 18S mock communities as expected even when the trim length of reverse reads was reduced to 100 bp, implying that this analytical strategy can be used even for poor-quality sequences (as we sometimes see).

While analyzing the quantitative recovery of our mock communities and comparing 515Y/926R with other commonly-used 18S specific primers, we found a 3-8 fold underestimation when there was an internal primer mismatch, as was the case with three Dinophyta included in our 18S mock community with mismatches to the reverse primer 926R. The same issue was previously found with the original EMP primers (515C/806R, V4) that underestimated SAR11 by 8-fold (Apprill et al., 2015) and Thaumarchaea by 1.5-3 fold (Parada et al., 2016). Consistent with our findings, 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. The worst mismatches are at the 3' end itself, 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).

For sequences that matched the primers exactly, we found that the 515Y/926R primers quantitatively recovered the mock communities abundances for both 16S and 18S mocks (r^2 = 0.97 for staggered mocks, observed vs. expected, Fig 3 and 4), indicating little preferential amplification of taxa that perfectly match primers. However, this did not apply to V4F/V4RB - although these primers perfectly matched all the clone members in the mock communities, they overestimated haptophytes by 3-4 fold. This discrepancy between results probably relates to methodological differences. V4RB has a considerably lower annealing temperature than V4F and a 2-step PCR is required (Stoeck et al., 2010). In contrast, amplification with the 515Y/926R primers can be done in a single step PCR reaction. This methodological difference may explain why the 515Y/926R primers more accurately recovered relative abundances of 18S taxa that have no mismatches versus the V4F/V4RB primers. These findings, together with the results of Parada et al. (2016), indicate that the 515Y/926R primers recover both 16S and 18S mock communities quantitatively when examined separately.

Since amplicon lengths of 16S and 18S rRNA gene are different, simultaneous amplification of 16S and 18S rRNA genes can lead to a length-based discrimination during PCR and sequencing stages, similar to what was previously reported to occur in general (Kittelmann et al., 2013). We endeavored to develop a quantitative understanding of the bias to better interpret results and to determine ideal sequencing depth for mixed communities. We therefore created mixed mock communities

289 consisting of both 16S and 18S rRNA in equal molarity, amplified with 515Y/926R, and
290 sequenced the resulting amplicons. The relative proportions of 16S and 18S amplicons
291 were quantified using a bioanalyzer to evaluate PCR bias, and the number of recovered
292 sequence reads were similarly counted (after splitting into 16S/18S pools in the
293 bioinformatic pipeline) to determine the cumulative bias (both PCR and sequencing
294 biases). We found very little PCR bias in even mocks that had no mismatches to the
295 primers. As expected, the biases were 1.5-3-fold higher when samples included 18S
296 staggered mock communities which contain three Dinophyta species that have one
297 mismatch to the reverse primer. Moreover, a 2-fold sequencing bias (in addition to the
298 aforementioned PCR biases) occurred in all combinations of mock community types.
299 That suggests sequencing bias due to length differences is a consistent property of the
300 Illumina sequencing platform, yet PCR bias due to primer mismatches is much less
301 predictable. Thus, an evaluation of primer coverage across three domains, in actual field
302 samples, may help better account for the PCR bias. Parada et al. (2016) found that
303 515Y/926R perfectly matches 86% of eukaryotes, 87.9% of bacteria, and 83.9% of
304 archaea in the SILVA database, but we note that in actual practice the extent of
305 mismatches in field samples depends on the particular taxa present and their
306 proportions. We should also note that our 18S mock communities are very rich in
307 alveolates such as dinoflagellates (3 of 10 in even, 7 of 16 in staggered) that tend to have
308 mismatches to the 515F/926F primers; hence they probably overestimate the biases
309 expected in most field samples.

310 Regarding pipeline recommendations, we found that all three pipelines (qiime2
311 q2-deblur plugin, qiime2 q2-dada2 plugin, and the standalone DADA2 R package) were
312 capable of recovering our mock communities exactly, although we note several
313 potential tradeoffs between the two algorithms we tested (deblur and DADA2) that

should be considered. Both DADA2 and Deblur can accurately recover the mocks under most conditions. However, we found DADA2 sometimes generated 1-mismatches to reference sequences when challenged with noisy sequence data (data not shown) or sequencing runs where PCR amplification used inconsistent methodological parameters (e.g. differing PCR cycles or input concentrations). These results reinforce the advisability of running some sort of control to account for PCR errors or run-to-run variability (mocks or duplicate samples spread across runs, see also Yeh et al. (2018)).

Although our work shows that DADA2 has the potential to generate false positives, we also note potential drawbacks to deblur. While deblur never produced 1-mismatches to the mock reference sequences, it removed a much larger fraction of total input sequences (~75%). This greatly reduces the sequencing yield, meaning that more sequencing is needed for the same coverage. In addition, the deblur algorithm differs significantly from DADA2 and there may be tradeoffs inherent in its design that are not apparent with the mocks. For example, deblur discards any reads deemed as errors whereas DADA2 attempts to correct sequencing errors (Callahan et al., 2016; Amir et al., 2017). Therefore, our evaluation with mock communities does not exclude the possibility that deblur may remove true biological variation deemed error sequences in natural samples where closely-related taxa coexist. Therefore, we recommend readers evaluate for themselves which denoiser is most appropriate for a given study, and consider the desired ultimate yield (of 16S and 18S sequences) when deciding sequencing depth.

To demonstrate the 515Y/926R full analytical pipeline with field samples, we analyzed monthly surface seawater samples collected from SPOT in 2014 from the 1.2-80 μm size fraction, which includes most protists, larger than average free-living prokaryotes and those attached to <80 μm particles. Even though this size fraction is

339 enriched in protists (free living bacteria which are mostly < 1 µm are excluded), we
340 found that 18S reads contributed an average of only to 9% of total reads, while
341 chloroplasts averaged 17% and prokaryotes 74%. This proportion of 18S is similar to
342 those reported from Southern California waters near SPOT (Needham and Fuhrman,
343 2016; Parada et al., 2016; Needham et al., 2018), but due to new results reported here
344 we now know the 18S has a sequencing bias that underestimates in their copy number
345 ~2-fold. And while quantitative interpretation of 18S copy numbers is greatly
346 complicated by the wide range of copies per genome (2 in some picoeukaryotes to
347 >50,000 in some dinoflagellates and others, Zhu et al. (2005), the 16S of chloroplasts
348 has a relatively smaller copy number variation and for phytoplankton the 16S plastid
349 data probably more closely reflects relative biomass, e.g. chloroplast count, than does
350 18S (Needham and Fuhrman, 2016), though dinoflagellate chloroplasts are missed. The
351 assay even detects the presence of multiple metazoa, in our case including copepods
352 and larvaceans, the latter of which are voracious bacterivores. However, since these are
353 relatively large individuals that may be patchily represented in samples, we
354 bioinformatically separated them and recalculated proportions of protists alone (Fig.
355 7b).

356 Overall, this universal 515Y/926R primer pair is able to simultaneously examine
357 the whole community structure across three domains, and we now have quantified the
358 extent of sequencing biases that can be expected. This will allow us to better estimate
359 cell abundances, biomasses, and overall inventories of all microbial (perhaps even some
360 metazoa) taxa in a sample, and will facilitate improved interpretation of interactions,
361 such as predation, parasitism, and mutualism. Although we now have quantified the
362 biases associated with primer mismatch (e.g. the underestimation of some Dinophyta
363 groups), the extent of underestimation is still unpredictable. Additionally, some protists

364 groups (Diplonemea, Kinetoplastida, and Discosea) are known to have particularly long
365 V4 regions (up to 800 bp) that are likely missed or seriously underestimated due to the
366 amplification biases (R. Massana pers. Comm.). Thus, the potential limitations of
367 interpretation of 18S results from this primer set, and probably any other primer set for
368 a similar purpose, should be recognized when attempting to comprehensively analyze
369 community composition.
370
371 **Materials and methods**
372 Mock community preparation
373 To generate even and staggered 16S mock communities, 11 and 27 clones of
374 marine 16S rRNA genes were respectively prepared as previously described (Parada et
375 al., 2016; Yeh et al., 2018). In the even 16S mock community, 11 clones were mixed with
376 equal molarity. For the staggered 16S mock community, 27 clones were mixed with
377 varying concentrations to roughly mimic marine prokaryote community composition
378 observed at our sampling site.
379 To create even and staggered 18S mock communities, nearly full-length 18S
380 rRNA clone libraries were prepared from the large size fraction (1.2-80 μm) of seawater
381 samples collected from SPOT location. DNA was amplified using universal eukaryote
382 primers Euk-A (5'-AACCTGGTTGATCCTGCCAGT-3') and Euk-B (5'-
383 GATCCTTCTGCAGGTTCACCTAC-3') (Countway et al., 2005). 25-μl PCR mixtures
384 contained 1.25X 5Prime Hot master mix (0.5 U Taq, 45mM KCl, 2.5 mM Mg2+, 200 μm
385 dNTPs; Quanta Bio), 0.5 μM primers, and 1 ng of DNA. PCR conditions were performed
386 as follows: an initial hot start step at 95°C for 2 min followed by 25 cycles of 95°C for 30
387 sec, 50°C for 30 sec and 68°C for 4 min. PCR products were then cloned using TOPO TA
388 cloning kit with pCR4-TOPO Vector and One Shot Top 10 competent cells according to

389 the manufacturer's protocols. The cloned PCR products were sequenced using Sanger
390 sequencing. Species identification was confirmed using BLASTn against the nt database.
391 16 clones were chosen to represent the major marine eukaryotic groups, including
392 haptophytes, dinoflagellates, diatoms, ciliates, cercozoa, radiolaria, and copepods.
393 Plasmids were purified using Qiagen plasmid plus 96 miniprep kit and then amplified
394 with Euk-A and Euk-B using the same conditions described above. In the even 18S mock
395 community, 10 clones were mixed with equal molarity. In the staggered 18S mock
396 community, 16 clones were mixed with different concentrations. To mimic natural
397 marine communities consisting of both eukaryotes and prokaryotes, 16S and 18S mock
398 communities were mixed in four combinations (Fig. 1). Each mixed mock community
399 was pooled with equal molarity after taking lengths into account (the average length of
400 16S mocks is 1425 bp, and the average length of 18S mocks is 1770 bp).
401
402 Sample collection and DNA extraction
403 Samples were collected from 5m depth at San Pedro Ocean Time-series (SPOT)
404 location in 2014. Approximately 12 L of seawater was sequentially filtered through 80-
405 µm mesh, a 1.2-µm A/E filter (Pall, Port Washington, NY), and a 0.2-µm Durapore filter
406 (ED Millipore, Billerica, MA). Filters were stored at -80°C until DNA extraction. This
407 study only used DNA extracted from A/E filters (1.2-80 µm), which consist of both
408 eukaryotic microbes and prokaryotes, for primer and pipeline testing purposes. DNA
409 was extracted from A/E filters using a NaCl/CTAB bead-beating extraction protocol as
410 described by Lie et al. (2013) with slight modification by adding an ethanol
411 precipitation step after lysis to reduce the volume of crude extract, which helps
412 minimize DNA loss during the subsequent purification.
413

PCR and sequencing

To compare 16S/18S universal primers with eukaryote-specific primers, 18S mock communities were amplified with V4F (5'-CCAGCASCYGCGGTAATTCC-3') and V4R (5'-ACTTTCGTTCTTGATYRA-3'), and V4F and V4RB (5'-ACTTTCGTTCTTGATYRR-3') (Stoeck et al., 2010; Balzano et al., 2015). The only difference between these two primer pairs is the last nucleotide on the 3' end of the reverse primer (A to R), which makes V4F/V4RB amplify haptophytes better (Balzano et al., 2015). Due to the considerably lower annealing temperature of the reverse primer, the full primers with indices and Illumina adaptors did not result in any PCR bands. Thus, 2-step PCR was required to obtain efficient amplification as is standard practice for this primer set (Stoeck et al., 2010; Balzano et al., 2015; Mahé et al., 2015; Pasulka et al., 2016). The first PCR mixtures contained 1X Phusion HF buffer (1.5 mM MgCl2), 300 µM dNTPs, 0.5 µM primers, 3% DMSO, 0.5 U Phusion High-Fidelity DNA polymerase (New England BioLabs Inc.), and 1 pg pure mock community. PCR cycles were as follows: 98°C for 1 min, 10 cycles of 98°C for 30 sec, 53°C for 30 sec, 72°C for 30 sec; and then 15 cycles of 98°C for 30 sec, 48°C for 30 sec, 72°C for 30 sec, and a final extension step of 72°C for 10 min. The second PCR reaction was performed with full primers that had barcoded indices and Illumina adaptors. PCR mixtures contained 1X Phusion HF buffer (1.5 mM MgCl2), 300 µM dNTPs, 0.5 µM primers, 3% DMSO, 0.5 U Phusion High-Fidelity DNA polymerase (New England BioLabs Inc.), and 2 µl of the PCR products from the first step. PCR cycles were as follows: 98°C for 1 min, 10 cycles of 98°C for 30 sec, 48°C for 30 sec, 72°C for 30 sec, and a final extension step of 72°C for 10 min.

With 16S/18S universal primers (515Y, 5'-GTGYCAGCMGCCGCGGTAA-3'; 926R, 5'-CCGYCAATTYMTTTRAGTTT-3'), as shown in Fig. 1, 8 different types of mock communities were amplified using single step PCR with full-length primers that had

barcoded indices and Illumina adaptors. 25-µl PCR mixtures contained 1.25X 5Prime Hot master mix (0.5 U Taq, 45mM KCl, 2.5 mM Mg2+, 200 µm dNTPs; Quanta Bio), 0.3 µM primers, and 1 pg of pure mock community. PCR conditions were as follows: 95°C for 2 min, 30 cycles of 95°C for 45 sec, 50°C for 45 sec and 68°C for 90 sec, and a final extension step of 68°C for 5 min. Environmental samples were amplified using the same condition described above but with 0.5 ng of DNA. PCR products were cleaned using 0.8X Ampure XP magnetic beads (Beckman Coulter). Purified PCR products were quantified with PicoGreen and sequenced on Illumina HiSeq 2500 in PE250 mode and MiSeq PE300. For each sequencing run, multiple blanks (i.e. PCR negative controls) were included as internal controls, meaning PCR water was amplified, cleaned and sequenced as environmental samples with the same conditions described above. After sequence processing, blanks were used to check for contamination that comes from sample bleed-through due to "index hopping".

Sequence demultiplexing

Sequences were demultiplexed by reverse indices allowing no mismatches using QIIME 1.9.1 split_libraries_fastq.py (Caporaso et al., 2010). Then, forward barcodes were extracted using QIIME 1.9.1 extract_barcoded.py. The sequences were demultiplexed by forward barcodes 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. The per-sample fastq files have been submitted to the EMBL database under accession number PRJEB35673.

In silico processing of amplicons

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. Mixed amplicon sequences were then 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 below.

16S processing

16S sequences were analyzed using the DADA2 R package (Callahan et al., 2016), the QIIME 2 q2-dada2 plugin, and the QIIME 2 q2-deblur plugin (Amir et al., 2017) to compare different denoising outputs. We ran DADA2 in both R and qiime2 platforms to compare version differences (standalone R DADA2 = v1.10.1; qiime2-2018.8). With DADA2, forward and reverse reads were trimmed and filtered after inspecting their quality profiles. Filtered reads were used to make parametric error models for forward and reverse reads independently. Then, filtered reads were denoised based on run-specific error models. Denoised reads were then merged and remove chimeric reads. Note that the DADA2 R package allows us to exclude blanks from error model training, but qiime2 q2-dada2 plugin does not have this capacity. With deblur, reads were merged with qiime2 VSEARCH and filtered using qiime2 q-score-joined plugin. Filtered reads were then processed through the qiime2 q2-deblur plugin. 16S ASVs were

489 classified with qiime2 classify-sklearn plugin against the SILVA 132 database subsetted
490 to the amplicon region. 16S ASVs classified as Chloroplast were extracted based on the
491 SILVA classifications and subsequently reclassified against the PhytoRef database.
492
493 18S processing
494 18S reads amplified with 515Y/926R are 575-595 bp, which is too long for
495 forward and reverse reads to overlap, so we chose to trim reads to fixed lengths before
496 the denoising step. While inspecting quality profiles, reverse reads were generally lower
497 quality than forward reads and quality profiles varied among sequencing runs. To find
498 trim lengths which worked equally well among runs, forward reads were trimmed to
499 220 bp and reverse reads were trimmed to varying lengths (100-220 bp). With the
500 DADA2 R package, trimmed forward and reverse reads were used to make parametric
501 error models independently. Then, trimmed reads were denoised based on run-specific
502 error models. Denoised reads were concatenated and chimeric reads were removed.
503 With the QIIME 2 q2-dada2 and q2-deblur plugins (which did not have an option for
504 independent denoising and subsequent merging at the time of writing), forward and
505 reverse reads were trimmed using bbduk.sh and concatenated using fuse.sh from
506 bbtools package (http://sourceforge.net/projects/bbmap/). Concatenated reads were
507 then processed as artificial single-end reads and chimeras were removed using QIIME 2
508 q2-dada2 and q2-deblur plugins. 18S ASVs were classified with qiime2 classify-sklearn
509 plugin against both SILVA 132 and PR2 databases. Once processed into exact amplicon
510 sequence variants (ASVs), 18S sequences were split into paired read files to indicate
511 their non-overlapping nature.
512
513 Validation of ASV algorithms by analysis of mock communities

514 ASVs were Blastn against *in silico* mock sequences to determine ASVs that
515 perfectly matched *in silico* sequences, artifacts (1-mismatch to the *in silico* sequences),
516 or contamination (more than 1-mismatch to the *in silico* sequences). To compare the
517 performance of each pipeline, we determined 1) the percent of reads that perfectly
518 matched *in silico* sequences, 2) the percent of reads removed after denoising, and 3) the
519 R-squared values of linear regression between observed and expected abundances on a
520 log (x+0.001) scale. The processing scripts in this study are available on Figshare at
521 https://doi.org/10.6084/m9.figshare.11320388
522
523 PCR and sequencing bias estimation
524 16S and 18S mixed mock communities amplified with 515Y/926R were run on a
525 Agilent 2100 Bioanalyzer to quantify concentrations of 16S and 18S PCR products in
526 each mixed mock community. Amplicons were analyzed with the High-sensitivity DNA
527 assay kit according to the manufacturer's instructions. Due to the length differences
528 between 16S and 18S amplicons, the concentration of each amplicons were measured
529 by checking peak area on Agilent 2100 Bioanalyzer using manual integration without
530 altering the instrument-determined baseline. The 16S:18S ratio of molarity was used to
531 determine PCR bias. Sequence pre-processing (i.e. bbsplit.sh) split reads into 16S and
532 18S pools. The 16S:18S ratio of the number reads was used to determine sequencing
533 and PCR bias. The slope of the line derived from plotting the 16S:18S ratio from
534 Bioanalyzer traces against 16S:18S ratio based on the number reads after the bbsplit
535 step was used to define sequencing bias.
536
537 **Acknowledgments**

We thank Mike Lee for selflessly taking the time to think and blog about how to split 16S from 18S informatically, which inspired the bbsplit method in our qiime2 workflow. 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.

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**: e00191-00116.

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. *Aquatic Microbial Ecology* **75**: 129-137.

Balzano, S., Abs, E., and Leterme, S.C. (2015) Protist diversity along a salinity gradient in a coastal lagoon. *Aquatic Microbial Ecology* **74**: 263-277.

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. *The ISME journal* **12**: 1907.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C., Al-Ghalith, G.A. et al. (2018) QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. In: PeerJ Preprints.

563 Bradley, I.M., Pinto, A.J., and Guest, J.S. (2016) Design and evaluation of Illumina MiSeq-
564 compatible, 18S rRNA gene-specific primers for improved characterization of mixed
565 phototrophic communities. *Appl Environ Microbiol* **82**: 5878-5891.
566 Bru, D., Martin-Laurent, F., and Philippot, L. (2008) Quantification of the detrimental
567 effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene
568 as an example. *Appl Environ Microbiol* **74**: 1660-1663.
569 Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P.
570 (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nature*
571 *methods* **13**: 581-583.
572 Chénard, C., Wijaya, W., Vaulot, D., dos Santos, A.L., Martin, P., Kaur, A., and Lauro, F.M.
573 (2019) Temporal and spatial dynamics of Bacteria, Archaea and protists in equatorial
574 coastal waters. *Scientific Reports* **9**: 1-13.
575 De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R. et al. (2015)
576 Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.
577 Del Campo, J., Pons, M.J., Herranz, M., Wakeman, K.C., Del Valle, J., Vermeij, M.J. et al.
578 (2019) Validation of a universal set of primers to study animal-associated
579 microeukaryotic communities. *Environmental microbiology* **21**: 3855-3861.
580 Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., and Sogin, M.L. (2015)
581 Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of
582 high-throughput marker gene sequences. *The ISME journal* **9**: 968-979.
583 Fuhrman, J.A., Cram, J.A., and Needham, D.M. (2015) Marine microbial community
584 dynamics and their ecological interpretation. *Nature Reviews Microbiology* **13**: 133-146.
585 Lin, Y., Gifford, S., Ducklow, H., Schofield, O., and Cassar, N. (2019) Towards quantitative
586 microbiome community profiling using internal standards. *Appl Environ Microbiol* **85**:
587 e02634-02618.

588 Mahé, F., Mayor, J., Bunge, J., Chi, J., Siemensmeyer, T., Stoeck, T. et al. (2015) Comparing
589 High-throughput Platforms for Sequencing the V4 Region of SSU-r DNA in
590 Environmental Microbial Eukaryotic Diversity Surveys. *Journal of Eukaryotic*
591 *Microbiology* **62**: 338-345.
592 Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput
593 sequencing reads. *EMBnet journal* **17**: 10-12.
594 Needham, D.M., and Fuhrman, J.A. (2016) Pronounced daily succession of
595 phytoplankton, archaea and bacteria following a spring bloom. *Nature microbiology* **1**:
596 16005.
597 Needham, D.M., Fichot, E.B., Wang, E., Berdjeb, L., Cram, J.A., Fichot, C.G., and Fuhrman,
598 J.A. (2018) Dynamics and interactions of highly resolved marine plankton via automated
599 high-frequency sampling. *The ISME journal* **12**: 2417.
600 Parada, A.E., Needham, D.M., and Fuhrman, J.A. (2016) Every base matters: assessing
601 small subunit rRNA primers for marine microbiomes with mock communities, time
602 series and global field samples. *Environmental microbiology* **18**: 1403-1414.
603 Pasulka, A.L., Levin, L.A., Steele, J.A., Case, D.H., Landry, M.R., and Orphan, V.J. (2016)
604 Microbial eukaryotic distributions and diversity patterns in a deep-sea methane seep
605 ecosystem. *Environmental microbiology* **18**: 3022-3043.
606 Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R. et al. (2006)
607 Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings*
608 *of the National Academy of Sciences* **103**: 12115-12120.
609 Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D., Breiner, H.W., and Richards, T.A.
610 (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly
611 complex eukaryotic community in marine anoxic water. *Molecular ecology* **19**: 21-31.

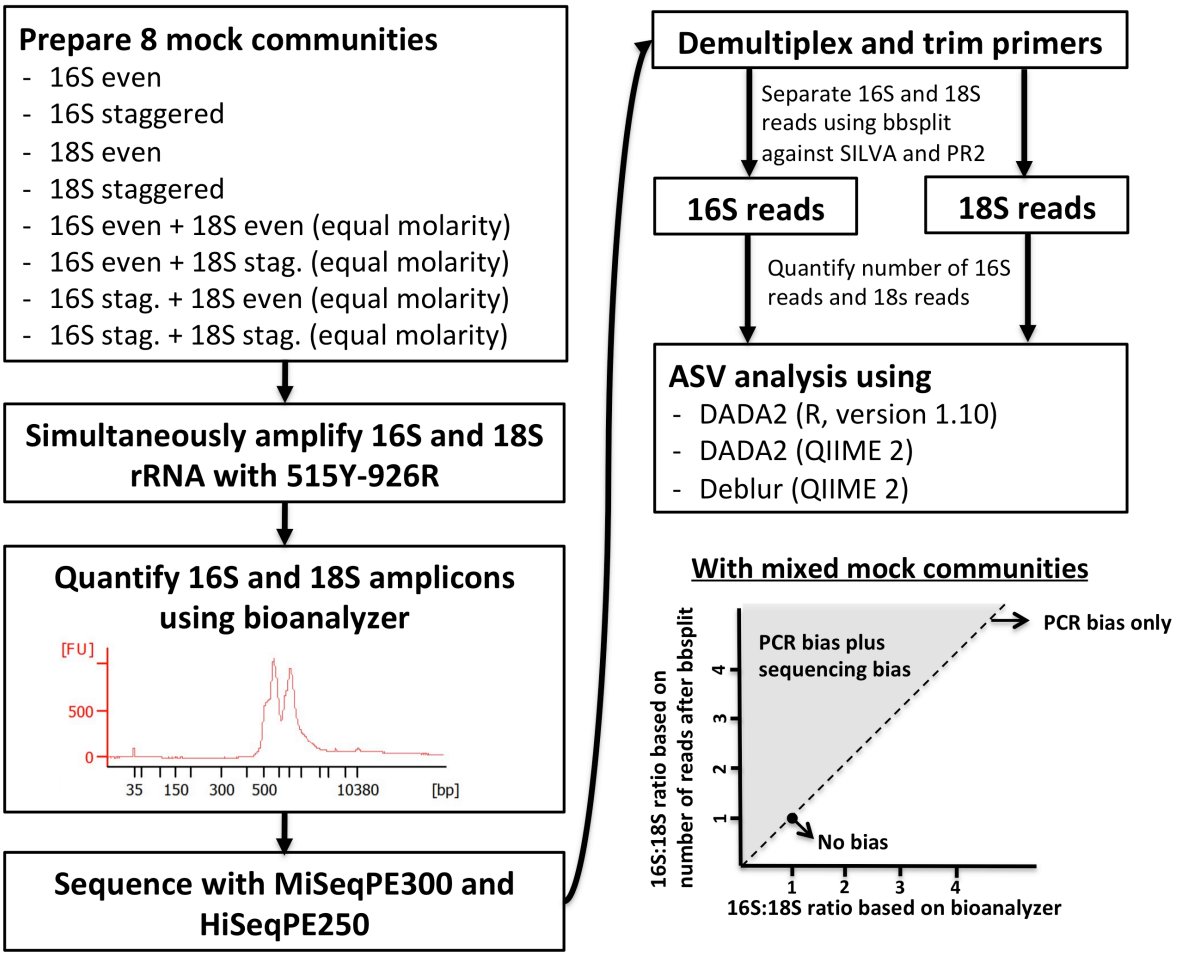
612 Yeh, Y.-C., Needham, D.M., Sieradzki, E.T., and Fuhrman, J.A. (2018) Taxon
613 Disappearance from Microbiome Analysis Reinforces the Value of Mock Communities as
614 a Standard in Every Sequencing Run. *MSystems* **3**: e00023-00018.

Table 1. Effects of trim length on 18S staggered mock community

Trim length of reverse reads (bp)	Deblur			DADA2 R package			DADA2 QIIME 2 version		
	Percent reads perfectly match in silico	r^2 expected vs. observed	Percent reads removed after denoising	Percent reads perfectly match in silico	r^2 expected vs. observed	Percent reads removed after denoising	Percent reads perfectly match in silico	r^2 expected vs. observed	Percent reads removed after denoising
220	100	0.77	73.5	99.5	0.76	35.8	99.4	0.77	20.0
210	100	0.77	72.6	99.6	0.74	31.2	99.4	0.77	29.3
200	100	0.77	72.3	99.7	0.76	28.3	99.5	0.77	27.3
190	100	0.77	71.2	99.8	0.76	25.8	99.6	0.77	25.1
180	100	0.77	70.2	99.8	0.76	24.3	99.6	0.77	23.4
170	100	0.76	71.5	99.9	0.76	23.3	99.7	0.76	22.4
160	100	0.76	70.4	99.9	0.76	22.5	99.6	0.76	21.6
150	100	0.76	69.3	99.9	0.76	21.7	99.7	0.76	20.7
140	100	0.76	70.2	99.9	0.76	20.9	99.8	0.76	19.5
130	100	0.76	69.2	100.0	0.76	20.4	99.9	0.76	18.9
120	100	0.76	68.3	100.0	0.76	19.8	99.8	0.76	18.1
110	100	0.76	67.4	100.0	0.76	19.3	99.8	0.76	17.6
100	100	0.76	66.4	100.0	0.76	18.8	99.9	0.76	16.9

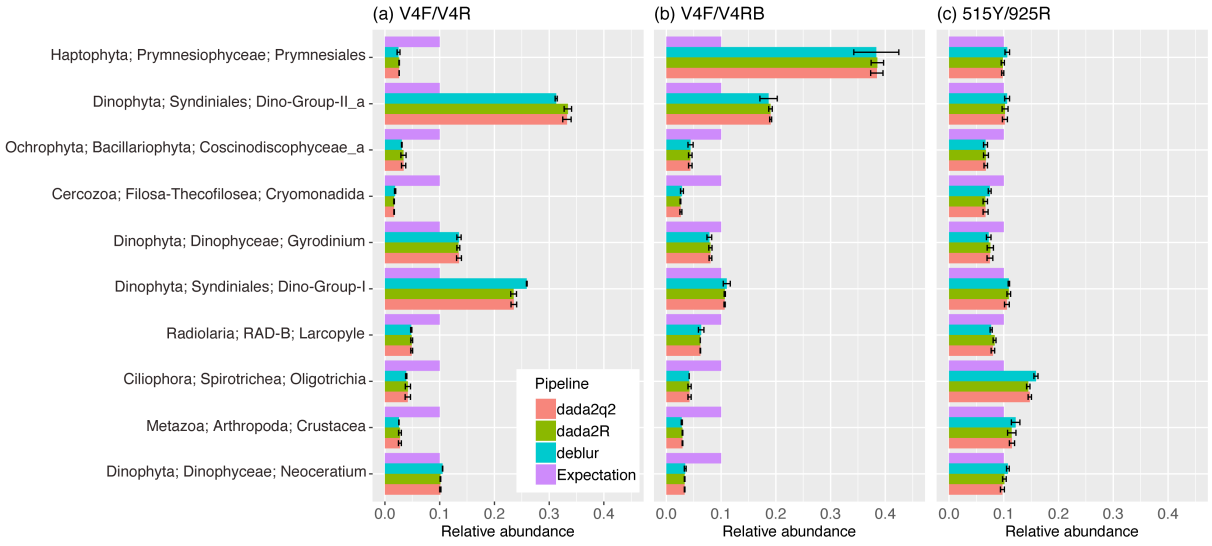
615

Figure 1. Experimental design. 8 mock communities were amplified using the 515Y/926R primers. The amplicons of mixed mock communities were analyzed using a Bioanalyzer to quantify the PCR bias against 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 counted to quantify the sequencing bias against 18S. Hypothetically, if there is no bias, all the mixed mocks are located at a single point (1,1). If there is only PCR bias, all the data points will be at the one-to-one line. If there are PCR and sequencing bias, all the data points will be located above the one-to-one line (gray area). The slope indicates the sequencing bias.



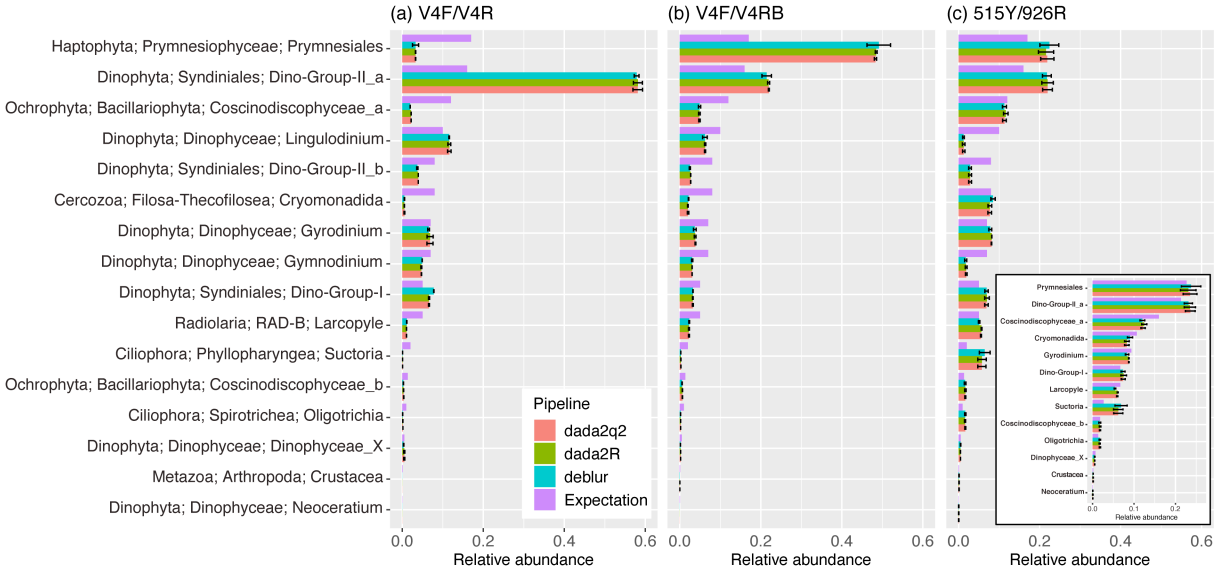
626 Figure 2. Comparison of even 18S mock communities amplified with V4F/V4R (a),

627 V4F/V4RB (b), and 515Y/926R (c).



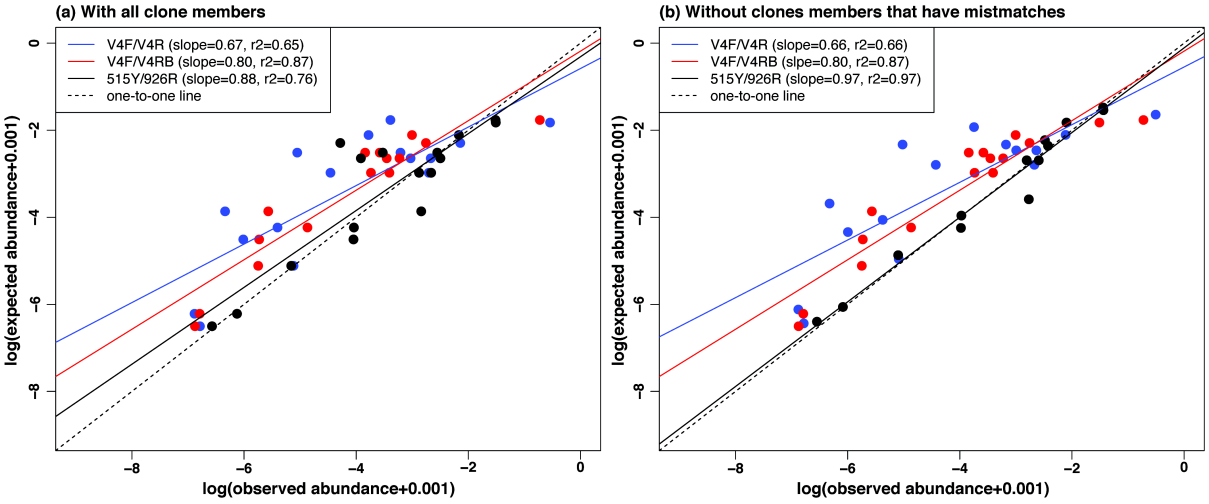
628

629 Figure 3. Comparison of staggered 18S mock communities amplified with V4F/V4R (b),
630 V4F/V4RB (b), and 515Y/926R (c). The insert in (c) shows only taxa with perfect
631 matches.



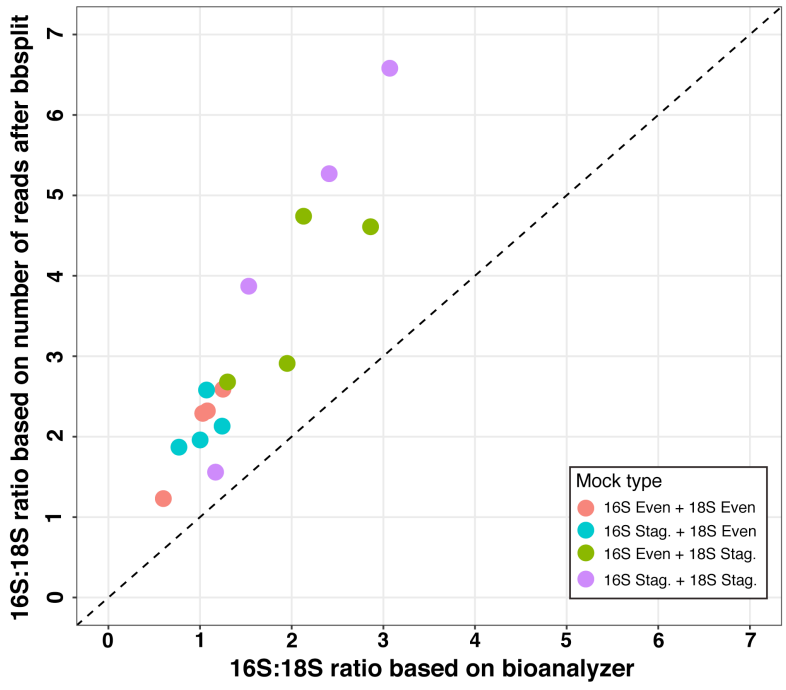
632
633

634 Figure 4. Expected staggered 18S mock communities plotted against observed
635 staggered 18S mock communities amplified with different primers pairs (a) and without
636 clone members that have mismatches on the given primer pairs (b).



637

638 Figure 5. Comparison of PCR and sequencing biases among four mixed 16S and 18S
639 mock communities, each combined in 1:1 molar ratios. The X axis shows ratios in the
640 PCR products, and the Y axis shows the ratios in the final sequences, including biases
641 from PCR plus sequencing. The data points all occur above the dashed 1:1 line,
642 indicating most biases are from sequencing. Note for 18S even mocks (orange and blue)
643 the PCR products have a bioanalyzer output ratio near 1, indicating little PCR bias. The
644 staggered 18S mocks (green and purple) include 3 members with primer-template
645 mismatches and correspondingly more PCR bias visible on the x axis. In all cases the
646 final reads show about 2-fold more bias than the PCR biases alone.



647

Figure 6. The composition of reads found in 1.2-80 µm size fraction of seawater samples collected from SPOT in 2014. Numbers on the X axis are months.

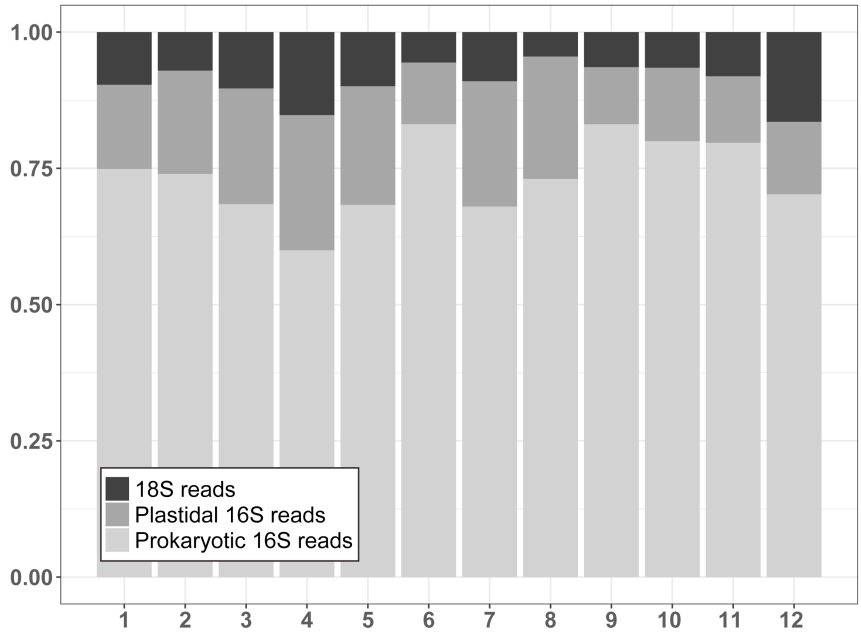


Figure 7. The monthly community composition of 1.2-80 μm size fraction of seawater samples collected from SPOT in 2014 at class level for eukaryotes and at order level for prokaryotes. Only the dominant ASVs (the average relative abundance is greater than 0.5%) were shown. Each box represents single ASV.