environmental
microbiology

Environmental Microbiology (2016) **18**(5), 1403–1414 doi:10.1111/1462-2920.13023

Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples

Alma E. Parada, David M. Needham and Jed A. Fuhrman*
University of Southern California, Los Angeles, CA, USA

Summary

Microbial community analysis via high-throughput sequencing of amplified 16S rRNA genes is an essential microbiology tool. We found the popular primer pair 515F (515F-C) and 806R greatly underestimated (e.g. SAR11) or overestimated (e.g. Gammaproteobacteria) common marine taxa. We evaluated marine samples and mock communities (containing 11 or 27 marine 16S clones), showing alternative primers 515F-Y (5′-GTGYCAGCMGCCGCGGTAA) and 926R (5′-CCGYCAATTYMTTTRAGTTT) yield more accurate estimates of mock community abundances, produce longer amplicons that can differentiate taxa unresolvable with 515F-C/806R, and amplify eukaryotic 18S rRNA. Mock communities amplified with 515F-Y/926R yielded closer observed community composition versus expected ($r^2 = 0.95$) compared with 515F-Y/806R ($r^2 \sim 0.5$). Unexpectedly, biases with 515F-Y/806R against SAR11 in field samples (~4–10-fold) were stronger than in mock communities (~2-fold). Correcting a mismatch to Thaumarchaea in the 515F-C increased their apparent abundance in field samples, but not as much as using 926R rather than 806R. With plankton samples rich in eukaryotic DNA (> 1 μm size fraction), 18S sequences averaged ~17% of all sequences. A single mismatch can strongly bias amplification, but even perfectly matched primers can exhibit preferential amplification. We show that beyond *in silico* predictions, testing with mock communities and field samples is important in primer selection.

Introduction

Next-generation sequencing continues to make analysis of microbial diversity easier and less expensive. Therefore, the choice of primers to amplify 16S genes becomes crucial to take advantage of the sequence length and coverage made possible by improved sequencing technologies. In 2010, the Earth Microbiome Project (EMP) was established to create a catalogue of microbial diversity from habitats across the world (Gilbert *et al*., 2010) with the goal of creating a database of microbial samples analysed exactly the same way to facilitate global comparisons. The EMP proposed standard primers and protocols to permit comparisons of diversity across samples. The primers 515F/806R were chosen to maximize the global coverage of Bacteria and Archaea while also providing polymerase chain reaction (PCR) products of suitable length for sequencing with available Illumina platforms (Caporaso *et al*., 2011; 2012). Since it is commonly assumed that one mismatch in the middle of a primer will still allow binding and amplification of target templates, these primers appeared to have comprehensive coverage *in silico*. At around the same time, reviews of various group-specific and universal primers, such as Klindworth and colleagues (2013), performed mostly *in silico* analysis of hundreds of primers. Although Klindworth and colleagues (2013) did not examine the exact reverse primer used by EMP, they reported on similar primers that also had high apparent coverage if one mismatch is allowed. Thus, this 515F/806R primer pair seemed a reasonable choice.

We submitted marine plankton samples from several (5–890 m) depths to the EMP and were surprised to find the SAR11 cluster, relating to the Candidate genus *Pelagibacter*, was poorly represented in the results (typically ~3%). Other studies of these samples taken from the San Pedro Ocean Time Series (SPOT) analysed via the Automated Ribosomal Intergenic Spacer Analysis (Fisher and Triplett, 1999; Beman *et al*., 2011; Chow *et al*., 2013; Cram *et al*., 2015), as well as prior studies at this location by FISH (Ouverney, 1999), indicated the SAR11 clade is typically 20–40% of the bacterial community. This agrees with many analyses of marine plankton samples from around the world (Morris *et al*., 2002; Venter *et al*., 2004; Carlson *et al*., 2009; Brown *et al*., 2012; Gómez-Pereira *et al*., 2013; Logares *et al*., 2013; Needham *et al*., 2013; Vergin *et al*., 2013; Salter *et al*., 2015; Apprill *et al*., 2015). This suggested that the EMP PCR amplification was strongly biased against SAR11. A recent publication by

Received 20 June, 2015; revised 31 July, 2015; accepted 12 August, 2015. *For correspondence. E-mail: fuhrman@usc.edu; Tel. (+1) 213 740 5759; Fax (+1) 213 740 8123.

© 2015 Society for Applied Microbiology and John Wiley & Sons Ltd

Apprill and colleagues (2015) reports a mismatch in the 806R primer, which when corrected greatly increases the SAR11 abundances to more closely resemble FISH results.

Criteria for selecting PCR primers for small subunit rRNA amplicon sequencing include sequencing depth, high coverage of the taxa of interest (here all Bacteria and Archaea), the ability to compare results with prior studies, accuracy in relative abundances and also the phylogenetic resolution of the sequenced PCR products. Reducing primer biases is especially important in the case of applications such as association networks or predicting functional processes using programs like PICRUSt (Langille *et al.*, 2013). While comparing primers to the 16S rRNA database, we noted that the EMP 515F primer has a single mismatch to a majority of the globally important Thaumarchaea and Crenarchaea, which can be corrected with a single degeneracy, as noted by Quince and colleagues (2011). To take advantage of longer sequences now available, we considered utilizing an alternate reverse (926R) primer used by Quince and colleagues (2011) that has very high coverage of bacteria and archaea. The 515F-Y/926R primer pair encompasses the V4 and V5 hypervariable regions, while 515F-Y/806R encompasses only the V4. Therefore, the 515F-Y/926R primer pair yields a more informative product of suitable length (given current sequencing capabilities, > 2 × 250 bp) that overlaps with the product of the EMP primers, facilitating comparisons.

In this report, we analysed mock communities made of marine bacterial and archaeal 16S rRNA clones as well as natural marine samples. We found the primer pair 515F-Y/926R had better coverage of extremely common marine taxa missed by the original EMP primers (515F-C/806R), more accurately represented expected mock community abundances, and the added length allowed for better identification of the taxa present.

Results

In silico *primer comparisons*

Comparisons of the 515F-C (as used by the EMP) and 515F-Y (modified for this study from Quince *et al.*, 2011) primers showed an increase from 57% to 93% in the perfect matches to all known Archaeal taxa with the Y degeneracy, driven mainly by an increased detection of Thaumarchaea Marine Group I (MGI) taxa in the database (from 0.4% to 96.4%, Table S1). The primers 515F-Y/926R increased the percentage of perfectly matched SAR11 taxa from 3% to 96%, and matched all three domains. The perfect matches to individual SAR11 subclades increased when using 515F-Y/926R (Table S2). One mismatch was required with 515F-Y/806R to match Deep 1, Surface 2 and Surface 3 subclades. However, perfect matches to Surface 4 were similar between primer pairs.

Mock community comparisons

Clone abundances in the even mock community samples amplified with 515F-Y/926R were more similar to the expected than with 515F-Y/806R (Fig. 1A). Even though only SAR11 and SAR116_a clones have a mismatch to the 806R primer, seven clones had abundances < 2/3 of the expected 9.1% (Fig. 1A). Additionally, SAR86_a and Marine Group A were overrepresented > 2-fold in 515F-Y/806R-amplified samples.

Amplification from the staggered mock community DNA exhibited greater primer bias. The observed community composition was more similar to the expected with 515F-Y/926R than with 515F-Y/806R ($r^2 = 0.95$ versus 0.53; Fig. 1B and C). Additionally, the deviation from expectation (Observed ÷ Expected = 1) was low with 515F-Y/926R, but included significant under- and overestimates with 515F-Y/806R (top inset, Fig. 1A). Several templates were responsible for the deviations observed with 515F-Y/806R, such as SAR11, SAR116_a and all Gammaproteobacteria clones (Fig. 1B, Table 1). SAR116_a reads were almost absent in the 515F-Y/806R samples, but this discrepancy was likely due to a 3′ mismatch to that clone, as clones SAR116_b and c were overrepresented. Removing the SAR11 and SAR116_a clones from the analysis and rescaling the abundances of the remaining operational taxonomic units (OTUs) increased the 515F-Y/806R r^2 only to 0.69, and actually decreased it when only the SAR11 OTUs were removed ($r^2 = 0.48$). The largest deviation from expected abundances in the 515F-Y/926R staggered mock community dataset was underrepresentation of SAR202_b (Fig. 1C, Table 1).

The mock communities, when clustered together ‘blindly’ with the field samples, also allowed us to evaluate several clustering methods (UCLUST, USEARCH, mothur, SWARM; Supplementary Information). Though no method can be perfect, some methods inaccurately clustered more than 5% of the sequences, and we found the average-neighbour algorithm in mothur, with pre-clustering at 2 base similarity, produced mock community compositions closest to those expected compared with UCLUST and USEARCH (Table S3). We also found SWARM (Mahé *et al.*, 2014) in QIIME to give mock community compositions similar to the expected without pre-clustering.

MGI

Several field samples from SPOT and other marine sites (see Experimental Procedures) were used to compare the abundance of taxa when amplifying with different combinations of the primers. Amplification with 515F-Y/806R

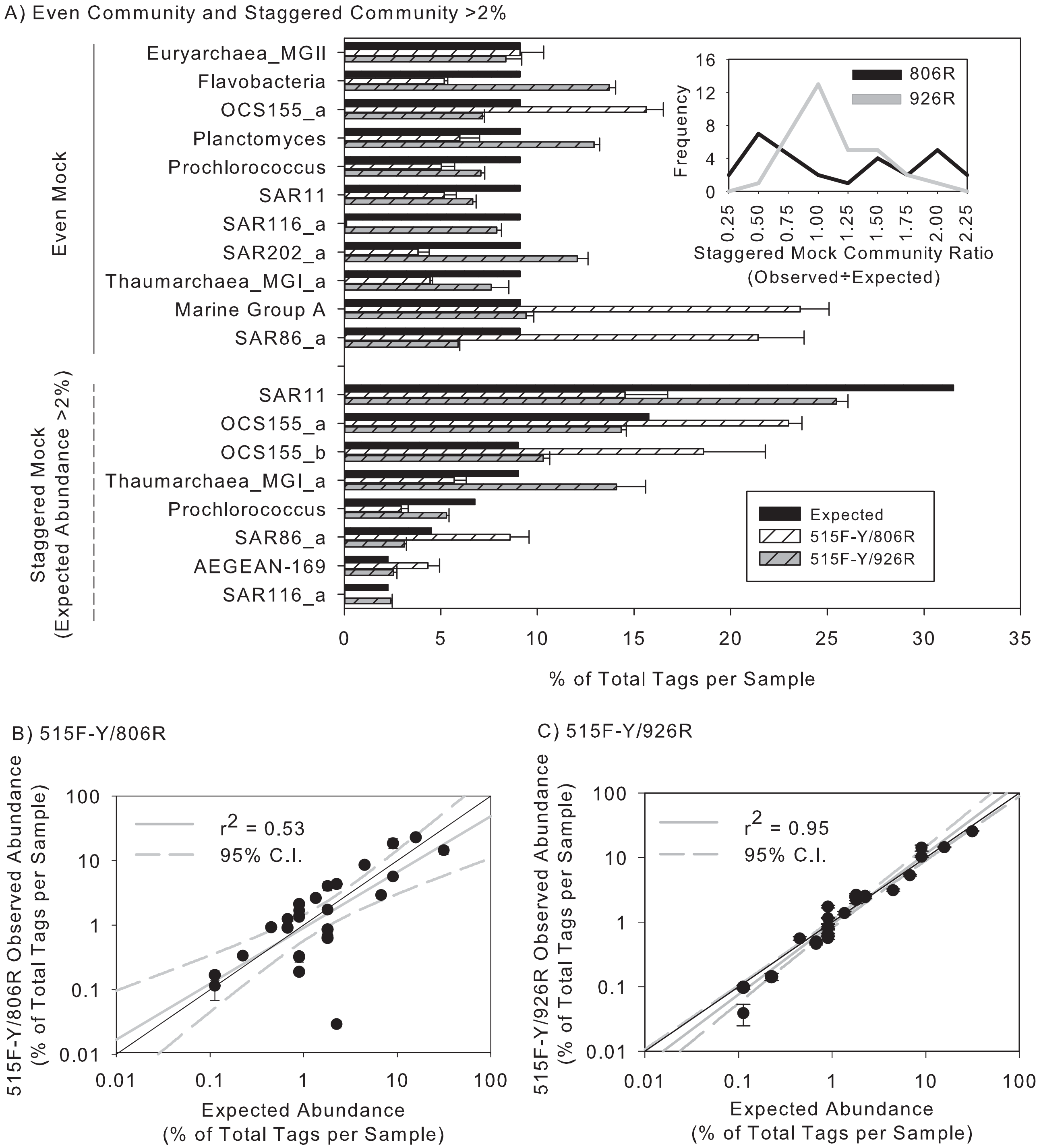

Fig. 1. Comparisons of even mock community (A) and staggered mock community clones (A–C) show tag abundances are closer to expected with 515F-Y/926R. The inset in (A) gives the observed ÷ expected ratio for the staggered mock clones for each reverse. Note that SAR116_a is virtually undetected with the 806R primer (A), and the difference in detection of the SAR11 clone between reverse primers is greater in the staggered community (A). Observed mock community profiles with different primer pairs (B and C); community profiles with 515F-Y/806R (in B) and 515F-Y/926R (in C), plotted against the expected staggered mock community profile. One-to-one line (solid black line) indicates the theoretical perfect match of observed and expected communities. Regression lines and 95% confidence intervals for each curve are included. The mean of four replicates is given as the abundance of each clone, with the error bars representing the standard error of the mean.

Table 1. Staggered mock community clone names, per cent expected and observed per cent abundance.

Clone name	Staggered per cent expected	515F-Y/806R observed abundance (mean ± SEM)	806R ratio difference (Obs ÷ Exp)	515F-Y/926R observed abundance (mean ± SEM)	926R ratio difference (Obs ÷ Exp)
SAR11 Surface 1 (Alpha)	31.5	14.5 ± 2.21	0.460	24.7 ± 0.62	0.784
OCS155_a (Actino)	15.8	23.0 ± 0.68	1.46	14.7 ± 0.27	0.930
OCS155_b (Actino)	9.01	18.6 ± 3.2	2.06	10.5 ± 0.34	1.17
Thaumarchaea MGI_a	9.01	5.70 ± 0.62	0.633	13.7 ± 1.5	1.52
Prochlorococcus	6.76	2.94 ± 0.36	0.435	5.41 ± 0.15	0.800
SAR86_a (Gamma)	4.5	8.60 ± 0.96	1.91	3.16 ± 0.12	0.702
AEGEAN-169 (Alpha)	2.25	4.34 ± 0.60	1.93	2.57 ± 0.18	1.14
SAR116_a (Alpha)	2.25	0.0294 ± 0.0037	0.0131	2.37 ± 0.080	1.05
Euryarchaea MGII	1.8	1.74 ± 0.13	0.967	2.23 ± 0.27	1.24
Flavobacteria	1.8	0.674 ± 0.062	0.374	2.74 ± 0.082	1.52
Planctomyces	1.8	0.859 ± 0.13	0.477	2.55 ± 0.067	1.42
SAR116_b (Alpha)	1.8	4.05 ± 0.64	2.25	2.29 ± 0.11	1.27
SAR202_a (Chloroflexi)	1.8	0.631 ± 0.048	0.351	2.61 ± 0.16	1.45
Marine Group A (aka SAR406)	1.35	2.64 ± 0.12	1.96	1.40 ± 0.077	1.04
Flavobacteria_Formosa	0.901	0.334 ± 0.031	0.371	1.65 ± 0.071	1.83
Flavobacteria_NS9	0.901	2.13 ± 0.044	2.36	1.17 ± 0.033	1.30
Pseudospirillum (Gamma)	0.901	1.68 ± 0.22	1.86	0.831 ± 0.086	0.922
SAR86_b (Gamma)	0.901	1.34 ± 0.057	1.49	0.560 ± 0.030	0.621
SAR92 (Gamma)	0.901	1.41 ± 0.12	1.56	0.645 ± 0.020	0.716
Thaumarchaea MGI_b	0.901	0.322 ± 0.056	0.357	0.823 ± 0.077	0.913
Verrucomicrobia	0.901	0.189 ± 0.017	0.210	0.830 ± 0.025	0.921
Rhodobacteriaceae (Alpha)	0.676	0.915 ± 0.094	1.35	0.493 ± 0.0073	0.729
SAR86_c (Gamma)	0.676	1.25 ± 0.096	1.85	0.449 ± 0.010	0.664
Flavobacteria_NS5	0.45	0.925 ± 0.11	2.06	0.565 ± 0.037	1.26
SAR86_d (Gamma)	0.22	0.337 ± 0.011	1.53	0.137 ± 0.016	0.623
SAR116_c (Alpha)	0.113	0.168 ± 0.030	1.49	0.0940 ± 0.0090	0.832
SAR202_b (Chloroflexi)	0.113	0.116 ± 0.048	1.03	0.0394 ± 0.014	0.349

Clones also used in the even mock community, each at 9.1% expected abundance, are bolded (results shown in Fig. 1). Observed abundances are given as the mean ± standard error of the mean (SEM). The ratios of the observed ÷ expected (Obs÷Exp) abundances are also given for each reverse primer. Broad group names of clones are in parentheses (Alpha and Gamma refer to Proteobacteria, Actino to Actinobacteria).

produced a statistically significant increase in MGI abundance from samples taken at a minimum depth of 150 m, in contrast to amplification with 515F-C/806R (Table S4). However, these samples showed no difference in MGI community composition (Bray–Curtis similarity = 83%, indistinguishable from technical replicates, Table S4). No statistically significant difference was observed in the MGI abundance or community composition when samples were amplified with 515F-C/926R or 515F-Y/926R (Table S4). The most significant difference in MGI abundance was a 1.91-fold increase ± 0.0493 SEM when amplifying with 515F-C/926R compared with 515F-C/806R ($P < 0.001$). There was also a statistically significant 1.55-fold increase ± 0.103 SEM in MGI abundance when amplifying with 515F-Y/926R compared with 515F-Y/806R ($P = 0.0318$); however, 5 of the 19 samples showed higher abundance with the 806R primer. Comparison of the MGI community composition of those five samples to the others did not indicate an obvious reason for the difference observed.

Field comparisons

Comparing SPOT samples from 5 and 890 m in October 2013 showed several differences between primer pairs similar to those observed with mock communities, but to a greater extent (Fig. 2). For example, the SAR11 clade was ~6-fold higher in abundance at 5 m with the 515F-Y/926R primer and ~4-fold higher at 890 m compared with amplification with 515F-Y/806R. Gammaproteobacteria were higher in abundance in both samples amplified with 515F-Y/806R, similar to that seen with the SAR86, Pseudospirillum and SAR92 clones in the staggered mock communities (Fig. 1 and Table 1), with some severalfold higher and one group (Other Oceanospirillales) only slightly so. Many SAR116 taxa were detected with 515F-Y/806R in field samples, even though SAR116_a was nearly absent in the mock communities. The abundance of Gammaproteobacteria with 515F-Y/806R was still higher than with 515F-Y/926R even when rescaling the abundances after removing SAR11 and SAR116 OTUs (1.32-fold ± 0.0291 SEM, $P < 0.001$, *t*-test).

The 515F-Y/926R matches 86% of eukaryotic 18S rRNA (0 mismatches, Table S1). However, our analytical pipeline that removes non-overlapping paired-end reads discards eukaryotic 18S sequences because they are typically 160–180bp longer. As a result, we found no 18S sequences in our merged reads. When we removed the requirement of overlapping paired ends, we found an average of ~1.5% (range 0.5–3.8%) 18S sequences of

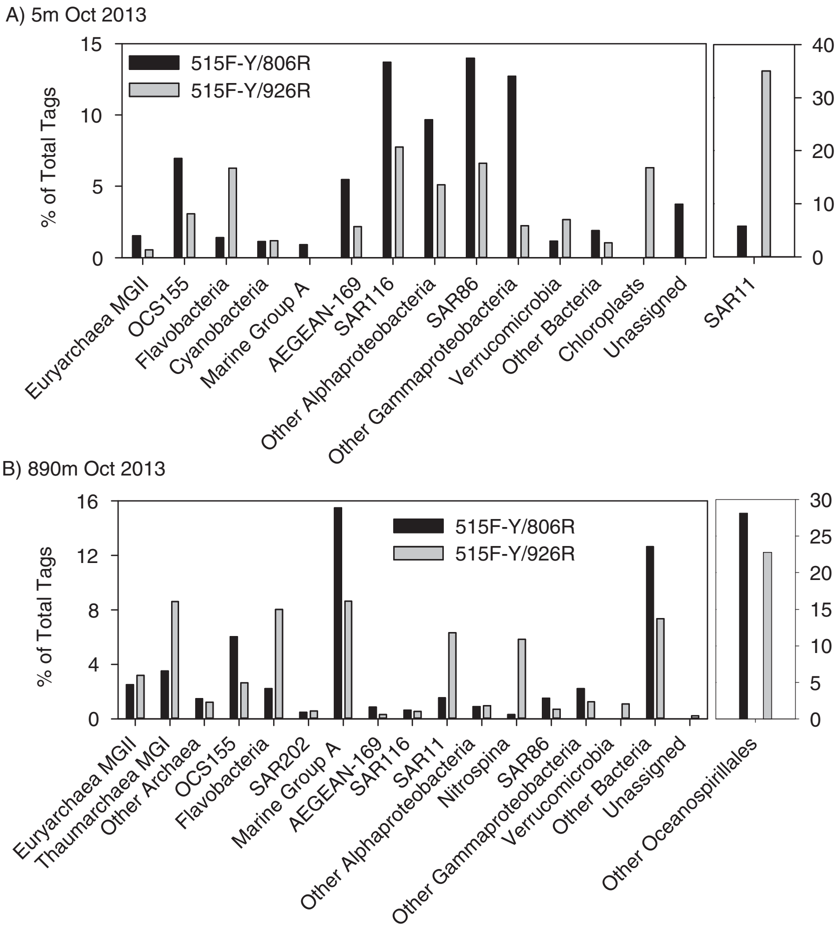

Fig. 2. Comparisons of two SPOT field samples amplified with the 806R or 926R show major differences in the taxa comprising 80% of each community. OTUs in order of decreasing abundance making up a total of 80% of the tags for each sample were grouped by taxonomy for each reverse primer. Data are from SPOT Oct 2013 samples at (A) 5 m and (B) 890 m.

picoeukaryotes in the 0.2–1 μm size fraction (Fig. S1, Table S6A). Preliminary analysis of marine plankton DNA from > 1 μm filters (some collected during a spring bloom), highly enriched in eukaryotes, yielded an average of 17% (range 8.6–35%) 18S sequences (Fig. S2, Table S6B).

The SAR11 in field samples were significantly (*t*-test, $P < 0.001$) and consistently higher (generally > 4×) with 515F-Y/926R compared with 515F-Y/806R, though different between subclades (Fig. 3). Some OTUs from the 515F-Y/926R samples were classified as Surface 3, though none were classified as such with 515F-Y/806R.

Differences in phylogenetic resolution

The additional sequence length provided by 515F-Y/926R added sequence variation not evident with 515F-Y/806R, often coinciding with apparent ecological differences. For example, several representative SAR11 OTU sequences from 515F-Y/926R-amplified samples were identical when trimmed to the 515F-Y/806R amplicon length, yet had distinct temporal and depth patterns (Fig. 4). SAR11 OTUs 2 and 3 had different abundances (Fig. 4C) and often varied inversely at 5 m. However, at the 515F-Y/806R length the representative sequences from those SAR11 OTUs were the same sequence (compare Fig. 4A and B). SAR11 OTUs 35, 163 and 13 (from the 515F-Y/926R dataset) had different patterns at each depth, sometimes varying inversely at 150 m (Fig. 4D), but all three would have been considered identical at the 515F-Y/806R length (Fig. 4B). A similar situation occurs with SAR11 OTUs 4000 and 264 (Fig. 4E).

Discussion

Primers for evaluating microbial communities by 16S rRNA gene amplification and sequencing are chosen to: (i) optimize the coverage of desired organisms with minimal biases in relative abundances, (ii) optimize the phylogenetic resolution, (iii) yield a high-quality product easily and inexpensively sequenced with the chosen sequencing platform and (iv) provide results generally comparable to other labs. Using these criteria, we evaluated the primers used initially in the EMP and an alternative set by amplification of both mock communities and diverse marine samples. Sequencing depth is also an

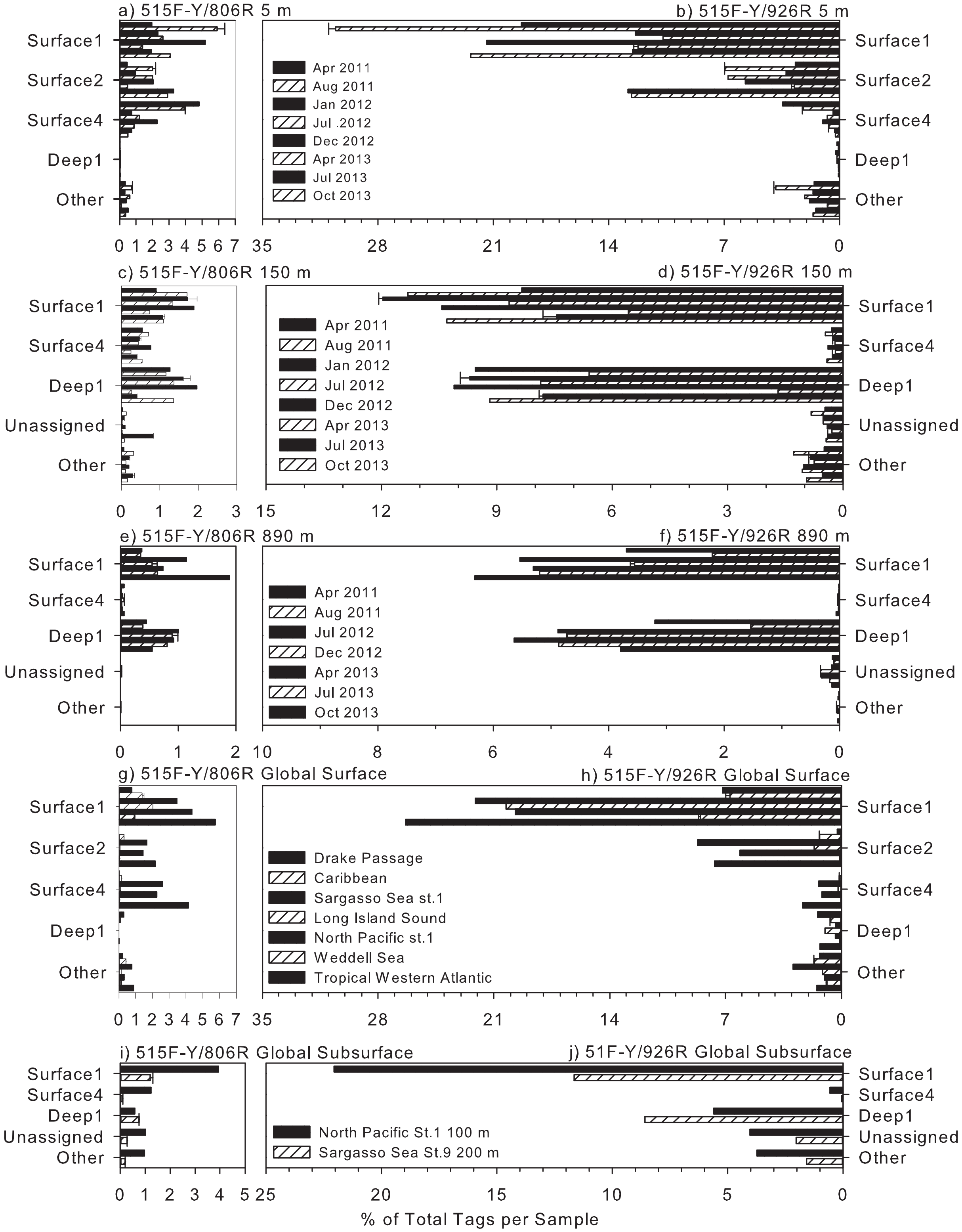

Fig. 3. Amplification with 515F-Y/926R yielded higher abundance of total SAR11 and most subclades in SPOT and global samples. Samples from (A, B) 5 m, (C, D) 500 m, (E, F) 890 m, (G, H) Global Surface and (I, J) Global Deep are given as bars of different patterns in chronological order, the same fill and order is used for the 806R and 926R panels. The per cent abundance of each clade in a sample is given on the x-axis (note scale is the same between reverse primers). The abundance of SILVA subclades Chesapeake Delaware-Bay, LD12, Surface 3 and Unassigned groups are combined at 5 m as Other. In deeper depths, Unassigned is shown separately, but Surface 2 was combined into Other. The mean abundance and the standard error of the mean are given where technical replicates were available.

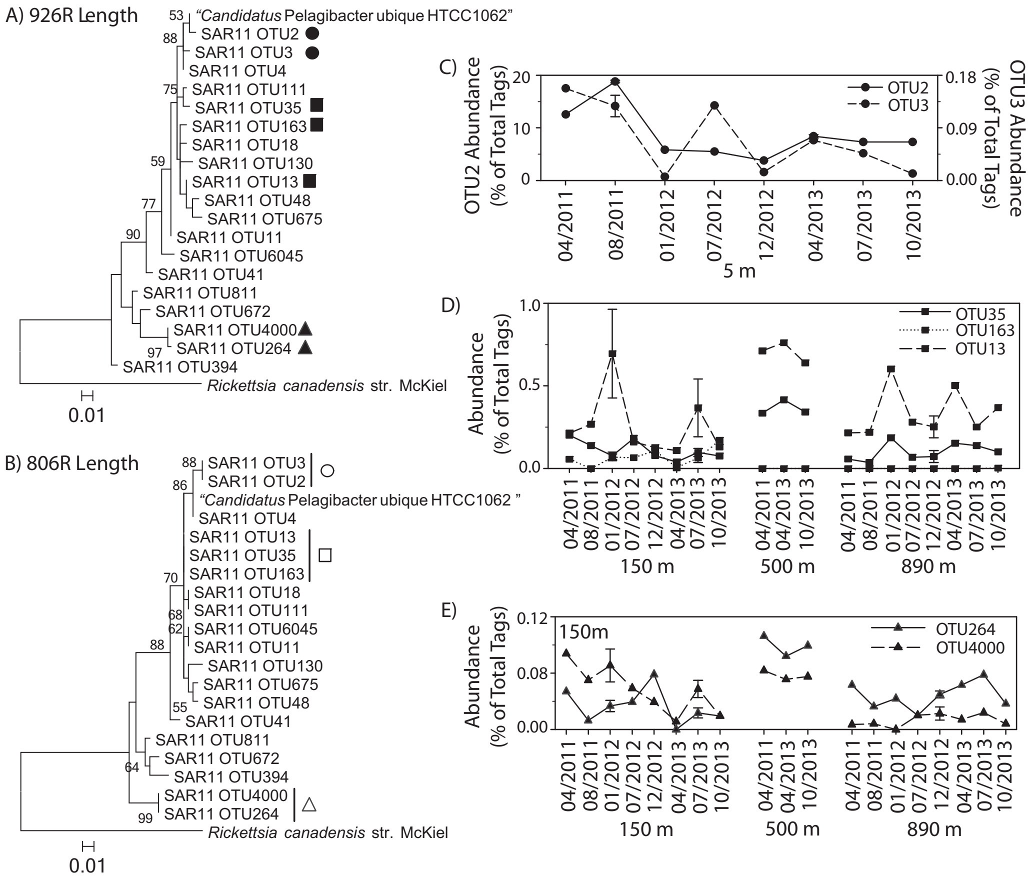

Fig. 4. Improved phylogenetic resolution with 515F-Y/926R shows ecological variations. The length of the 515F-Y/926R amplicon (A) resolved several SAR11 OTU representative sequences (closed symbols) with different ecological (time series) patterns (C, D, E), whereas the 515F-Y/806R length (B) of those amplicons classified them as identical sequences (open symbols). The time-series figures show OTUs resolvable only with 515F-Y/926R sometimes varying inversely, implying niche differentiation (E). Only bootstrap values ≥ 50% out of 1000 replicates are displayed on the trees. The subclade assigned to each OTU is given as S1 (Surface 1), D1 (Deep 1) or Unclassified.

important criterion for selecting primers, and newer higher throughput platforms can allow greater depth than the MiSeq 2 × 300 bp platform/chemistry we used, but there is no expectation that greater depth will reduce the quantitative biases we observed. Polymerase chain reaction optimization may reduce some biases inherent to each primer set, but we report results using PCR conditions similar to the EMP and published studies for the primers used. We reduced the number of cycles used by the EMP, but studies suggest this is unlikely to alter biases (Acinas *et al*., 2005; Sipos *et al*., 2007). Preliminary results similar to those presented in this study were supplied to the EMP, and alternative primer pair (515F-Y/926R) information is available on the EMP website.

Amplifying the staggered mock community demonstrated that the 515F-Y/926R primer pair produced communities much more similar ($r^2 = 0.95$) to the expected distribution than 515F-Y/806R ($r^2 = 0.53$). This was the result of significant overestimation of several clone taxa (notably Gammaproteobacteria, Actinobacteria, Marine Group A) and underestimation of several clones including SAR11 and SAR116_a. The total abundance of Gammaproteobacteria was also higher in field samples amplified with 515F-Y/806R compared with 515F-Y/926R. Removing SAR11 and SAR116 OTUs from both still showed greater total Gammaproteobacteria abundance with 515F-Y/806R. This suggests that differences observed were not due to missing SAR11 or SAR116 reads in the 515F-Y/806R dataset, but rather a bias for Gammaproteobacteria.

The use of mock communities allowed us to compare primer biases, but we also stress the importance of additionally comparing primers with field samples. For example, we found a less than twofold relative apparent bias between the two primer sets for SAR11 based on the mock communities (Fig. 1, Table 1); however, with field samples the SAR11 abundances with 515F-Y/926R were about 4–10-fold higher than with 515F-Y/806R (compare Figs 1–3). While in our study we did not have an absolute measure of SAR11 field abundances, another study that compared SAR11 FISH counts to the 515F-C/806R primers reported a > 10-fold bias against SAR11 in marine samples (Apprill *et al*., 2015). This indicates the importance of evaluating primers using field samples, in addition to *in silico* tests and amplification of mock communities.

Our focus in this study is on bacteria and archaea, so our standard samples were pre-filtered to remove the vast majority of eukaryotes; furthermore, our standard pipeline in practice removes 18S sequences. Modifying our pipeline to allow inclusion of 18S sequences showed that with 515F/926R < 1% of the amplicons were 18S (Fig. S1), and even in > 1 μm marine samples where chloroplast sequences greatly exceed those of bacteria and archaea, the 18S sequences averaged < 20% (Fig. S2). So while 18S amplification did not impact our study, it should be considered when these primers are used. A detailed analysis of the efficacy of 515F/926R for eukaryotic studies is beyond the scope of this report.

Our use of mock communities revealed apparent problems with typical OTU clustering protocols (Table S3). Due to our analyses, we chose to use mothur's average-neighbour algorithm with pre-clustering. The pre-clustering step may help explain the more congruent results observed, though it may mask natural sequence diversity by merging real variants. Other methods may require further optimization of available options to produce results closer to expectation, but this is beyond the scope of this study, and several studies have already evaluated many of these pipelines (Bonder *et al*., 2012; Pylro *et al*., 2014; Schmidt *et al*., 2014).

Analysis of field samples indicated that replacement of the 515F-C with 515F-Y results in a detectable, though small, increase in Thaumarchaea coverage when using 806R, not seen with 926R (Table S4). Thus, the modification of the 515F-Y may be more important when using 806R. Hugerth and colleagues (2014) also analysed changes to the 515F-C primer using the program DegePrime. Though they evaluated changing the C to a Y, they proceeded with using a B (C, G or T) and utilized a slightly different (805R) reverse primer. Our results suggest that this level of ambiguity may be unnecessary. We replaced the 'N' in the 515F primer used by Quince and colleagues (2011) with a 'Y' to reduce non-specificity that can conflict with some barcodes, potentially forming hairpin loops (W. Walters, pers. comm.). Therefore, greater ambiguity should be included only if it significantly increases detection of target organisms.

As Apprill and colleagues (2015) concluded, marine studies that used the original 515F-C/806R primer pair probably significantly underestimated the abundance of SAR11 in those samples (e.g. Paver *et al*., 2013; Taylor *et al*., 2014). Though the 806R modification presented by Apprill and colleagues (2015) reduced the bias against SAR11, it did not significantly alter proportions of other taxa. The 926R primer as we report here not only increases SAR11 coverage, but also appears to have more accurate estimates of multiple taxa and produces a longer amplicon that can improve phylogenetic resolution, and thus ecological analysis (Claesson *et al*., 2010; Schloss, 2010; Jeraldo *et al*., 2011; Kim *et al*., 2011; Ghyselinck *et al*., 2013).

In ecological research, it is ideal to measure microbial communities with high resolution and fidelity to the natural abundances. We found that beyond the initial *in silico* prediction of primer coverage, it is important to test primers with mock communities and examine further with field samples to fully evaluate the effectiveness of primers. We show that, compared with the 515F/806R primers, 515F-Y/926R gives an accurate and well-resolved picture of marine bacterial and archaeal communities.

Experimental procedures

Sampling sites and DNA extraction

Samples from the USC Microbial Observatory were collected at the SPOT station (33°33′N, 118°24′W) in 2011, 2012 and 2013 at various depths spanning surface to seafloor: 5 m, deep chlorophyll maximum layer, 150, 500 and 890 m (Table S5). Samples collected previously from different locations (global samples) were also analysed (Table S5, Fuhrman *et al*., 2008).

Water samples were filtered sequentially through a ~1 μm A/E filter (Pall) and 0.22 μm Durapore filter (ED Millipore). For this study, DNA from the 0.22 μm filter was analysed except when noted. The DNA was extracted by SDS lysis and purified by phenol-chloroform, as previously described (Fuhrman *et al*., 1988).

Primers and in silico *primer coverage analysis*

We compared the original 515F (515F-C) primer (5′-GTG**C**CAGCMGCCGCGGTAA, Caporaso *et al*., 2012) with one that replaces the C at the fourth position with a Y (515F-Y, 5′-GTG**Y**CAGCMGCCGCGGTAA, modified from Quince *et al*., 2011). We used reverse primers 806R (5′-GGACTAC**H**VGGGTWTCTAAT, Caporaso *et al*., 2012) and 926R (5′-CCGYCAATTYMTTTRAGTTT, Quince *et al*., 2011) to evaluate a subset of samples amplified with either the 515F-C or the 515F-Y. The 515F and 926R primers are similar to those originally published by Lane and colleagues (1985). Comparisons between reverse primers were performed only with 515F-Y because preliminary results demonstrated that samples amplified with 515F-C/806R or 515F-Y/806R gave similar results (data not shown). *In silico* primer coverage for primer pairs was analysed with zero or one mismatch using SILVA TestPrime 1.0 and individual primers were analysed using SILVA TestProbe 3.0. Both analyses used the SILVA Database SSU r123 (Quast *et al*., 2013).

DNA amplification

Triplicate 25 μl reaction mixtures contained 1 ng of DNA, 1.25× 5Prime Hot Master Mix (5Prime), 0.2 μM barcoded forward primer and 0.2 μM indexed reverse primer. Cycling conditions with the 806R primer followed the EMP temperature and time protocol, with a 3 min heating step at 94°C followed by 25 cycles of 94°C for 60 s, 50°C for 60 s, 72°C for 105 s, and a final extension of 72°C for 10 min. Cycling conditions with the 926R primer included a 3 min heating step at 95°C followed by 25 cycles of 95°C for 45 s, 50°C for 45 s, 68°C for 90 s, and a final extension of 68°C for 5 min. Triplicate reactions were pooled, and 5 μl used to check for amplification on a 2% agarose gel. The remaining 70 μl was cleaned and concentrated using 1× magnetic Agencourt AMPure XP beads (Beckman Coulter). Technical replicates for some samples and no template controls (blanks) were amplified and included in all analyses. Concentrated DNA was quantified by PicoGreen fluorescence assay (Life Technologies), pooled at equimolar concentrations then cleaned and concentrated with 0.8× SPRIselect magnetic beads (Beckman Coulter).

Sequencing and data processing

We used a combination of an inline (read on the first read) 5 bp barcode (at least 2 bases different) on the forward primers and unique 6 bp index (at least 2 bases different) on the reverse primer (read as an independent index read; Huse *et al*., 2014). Amplicons were sequenced using MiSeq Illumina 2 × 300 bp chemistry. Sequences were initially de-multiplexed by their reverse index allowing for one mismatch at the sequencing facility. The forward and reverse reads were merged using USEARCH v7, three mismatches were allowed across the overlapping region, choosing the higher quality base when a mismatch existed (Edgar, 2010). Sequences were then de-multiplexed by their forward barcode in QIIME 1.8, discarding any sequences with a mismatch to the barcode or primer (Caporaso *et al*., 2010). Sequences were discarded if the average quality score dropped below q33 across a 50 bp sliding window, if the sequence did not include the reverse primer, or contained any ambiguous bases. We also removed both the forward and reverse priming regions, excluding any sequences that did not contain the reverse primer. No mismatches to the reverse primer were allowed.

Pooled sequences were processed following the MiSeq SOP (Kozich *et al*., 2013) including alignment against the SILVA v119 database, and trimming to include only the overlapping regions. Sequences were then clustered *de novo* to form operational taxonomic units (OTUs) with mothur 1.34.4 at 99% similarity with the average-neighbour algorithm (Schloss *et al*., 2009), and pre-clustered at 2 (806R-amplified) or 3 (926R-amplified) base similarity to reduce the effects of sequencing errors. Chimera detection performed with UCHIME (Edgar *et al*., 2011) and classified with the default mothur classifier (Wang *et al*., 2007) using the SILVA v119 database at an 80% confidence cut-off (Quast *et al*., 2013). Samples with fewer than 10 000 sequences were not included in the analyses (results ranged from 10 134 to 96 492 sequences per sample). The samples were normalized by analysing the relative abundance for each OTU as the proportion of all sequences (tags) in a sample after all OTUs with fewer than six sequences across all samples was discarded. Other clustering approaches tested are detailed in Supplemental Materials and Methods in Supplementary Information.

All sequence data have been submitted to the EMBL database under accession number PRJEB10633.

Mock communities

Mock communities containing 11 or 27 clones were prepared from marine 16S rRNA clones (see Supplemental Material and Methods in Supporting Information). Suitably diluted DNA was treated like a sample throughout the process, and sequences clustered blindly with field samples.

Evaluation of phylogenetic resolution gained by 926R

To evaluate differences in phylogenetic resolution, a tree was made from 926R-amplified SAR11 OTU representative sequences and a separate tree made from trimming those sequences to the 806R length. The most abundant sequence from each SAR11 OTU from the 515F-Y/926R pool (371 bp) was aligned with ClustalW and a maximum-likelihood tree based on the Tamura–Nei model was constructed in MEGA6 (Tamura and Nei, 1993; Thompson *et al*., 1994; Tamura *et al*., 2013). Reference sequences '*Candidatus* Pelagibacter ubique HTCC1062' and *Rickettsia canadensis* str. Mckiel were also trimmed to the overlapping regions (Accession numbers NR_074224.1 and NR_074485.1, respectively). Sequences were trimmed to remove the 806R primer in mothur 1.34.4 (final length 255 bp). Each representative sequence is given as SAR11 OTU# in the trees (only a subset of the SAR11 OTUs were plotted).

Acknowledgements

We thank Jack Gilbert, Rob Knight, Tony Walters and Amy Apprill for helpful discussions; Troy Gunderson and the crew of the R/V Seawatch and R/V Yellowfin. Funding was provided by the Gordon and Betty Moore Foundation grant GMBF3779 and NSF Dimensions of Biodiversity grant 1136818. A.E. Parada and D.M. Needham were supported by National Science Foundation Graduate Research Fellowships.

References

Acinas, S.G., Sarma-rupavtarm, R., Klepac-Ceraj, V., and Polz, M.F. (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol* **71:** 8966–8969.

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.

Beman, J.M., Steele, J.A., and Fuhrman, J.A. (2011) Co-occurrence patterns for abundant marine archaeal and bacterial lineages in the deep chlorophyll maximum of coastal California. *ISME J* **5:** 1077–1085.

Bonder, M.J., Abeln, S., Zaura, E., and Brandt, B.W. (2012) Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* **28:** 2891–2897.

Brown, M., V, Lauro, F.M., DeMaere, M.Z., Muir, L., Wilkins, D., Thomas, T., *et al*. (2012) Global biogeography of SAR11 marine bacteria. *Mol Syst Biol* **8:** 595.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7:** 335–336.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., *et al*. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *PNAS* **108:** 4516–4522.

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.

Carlson, C.A., Morris, R., Parsons, R., Treusch, A.H., Giovannoni, S.J., and Vergin, K. (2009) Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J* **3:** 283–295.

Chow, C.-E.T., Sachdeva, R., Cram, J.A., Steele, J.A., Needham, D.M., Patel, A., *et al*. (2013) Temporal variability and coherence of euphotic zone bacterial communities over a decade in the Southern California Bight. *ISME J* **7:** 2259–2273.

Claesson, M.J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J.R., Ross, R.P., and O'Toole, P.W. (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res* **38:** e200.

Cram, J.A., Chow, C.T., Sachdeva, R., Needham, D.M., Parada, A.E., Steele, J.A., and Fuhrman, J.A. (2015) Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *ISME J* **9:** 563–580.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460–2461.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194–2200.

Fisher, M.M., and Triplett, E.W. (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol* **65:** 4630–4636.

Fuhrman, J.A., Comeau, D.E., Hagström, Å., and Chan, A.M. (1988) Extraction from natural planktonic microorganisms of DNA suitable for molecular biological studies. *Appl Environ Microbiol* **54:** 1426–1429.

Fuhrman, J.A., Steele, J.A., Hewson, I., Schwalbach, M.S., Brown, M.V., Green, J.L., and Brown, J.H. (2008) A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci USA* **105:** 7774–7778.

Ghyselinck, J., Pfeiffer, S., Heylen, K., Sessitsch, A., and De Vos, P. (2013) The effect of primer choice and short read sequences on the outcome of 16S rRNA gene based diversity studies. *PLoS ONE* **8:** e71360.

Gilbert, J.A., Meyer, F., Jansson, J.K., Gordon, J.I., Pace, N.R., Tiedje, J.M., *et al*. (2010) The Earth Microbiome Project: meeting report of the '1st EMP meeting on sample selection and acquisition' at Argonne National Laboratory. *Stand Genomic Sci* **3:** 249–253. October 6th 2010.

Gómez-Pereira, P.R., Hartmann, M., Grob, C., Tarran, G.A., Martin, A.P., Fuchs, B.M., *et al*. (2013) Comparable light stimulation of organic nutrient uptake by SAR11 and Prochlorococcus in the North Atlantic subtropical gyre. *ISME J* **7:** 603–614.

Hugerth, L.W., Wefer, H.A., Lundin, S., Jakobsson, H.E., Lindberg, M., Rodin, S., *et al*. (2014) DegePrime, a program for degenerate primer design for broad-taxonomic-range PCR in microbial ecology studies. *Appl Environ Microbiol* **80:** 5116–5123.

Huse, S.M., Young, V.B., Morrison, H.G., Antonopoulos, D.A., Kwon, J., Dalal, S., *et al*. (2014) Comparison of brush and biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human subjects. *Microbiome* **2:** 5.

Jeraldo, P., Chia, N., and Goldenfeld, N. (2011) On the suitability of short reads of 16S rRNA for phylogeny-based analyses in environmental surveys. *Environ Microbiol* **13:** 3000–3009.

Kim, M., Morrison, M., and Yu, Z. (2011) Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Methods* **84:** 81–87.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F.O. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41:** e1.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* **79:** 5112–5120.

Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., and Pace, N.R. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* **82:** 6955–6959.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., *et al*. (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31:** 814–821.

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F.M., Ferrera, I., Sarmento, H., *et al*. (2013) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol* **16:** 2659–2671.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2:** e593.

Morris, R.M., Rappe, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A., Carlson, C.A., and Giovannoni, S.J. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420:** 806–810.

Needham, D.M., Chow, C.-E.T., Cram, J.A., Sachdeva, R., Parada, A.E., and Fuhrman, J.A. (2013) Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. *ISME J* **7:** 1274–1285.

Ouverney, C.C. (1999) Dissecting the marine bacterioplankton ‘Black Box’ by type and function through FISH and STARFISH. Doctoral Dissertation. Los Angeles, CA: University of Southern California.

Paver, S.F., Hayek, K.R., Gano, K.A., Fagen, J.R., Brown, C.T., Davis-Richardson, A.G., *et al*. (2013) Interactions between specific phytoplankton and bacteria affect lake bacterial community succession. *Environ Microbiol* **15:** 2489–2504.

Pylro, V.S., Roesch, L.F.W., Morais, D.K., Clark, I.M., Hirsch, P.R., and Tótola, M.R. (2014) Data analysis for 16s microbial profiling from different benchtop sequencing platforms. *J Microbiol Methods* **107:** 30–37.

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.

Salter, I., Galand, P.E., Fagervold, S.K., Lebaron, P., Obernosterer, I., Oliver, M.J., *et al*. (2015) Seasonal dynamics of active SAR11 ecotypes in the oligotrophic Northwest Mediterranean Sea. *ISME J* **9:** 347–360.

Schloss, P.D. (2010) The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol* **6:** e1000844.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75:** 7537–7541.

Schmidt, T.S.B., Matias Rodrigues, J.F., and von Mering, C. (2014) Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale. *PLoS Comput Biol* **10:** e1003594.

Sipos, R., Székely, A.J., Palatinszky, M., Révész, S., Márialigeti, K., and Nikolausz, M. (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol* **60:** 341–350.

Tamura, K., and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10:** 512–526.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30:** 2725–2729.

Taylor, J.D., Cottingham, S.D., Billinge, J., and Cunliffe, M. (2014) Seasonal microbial community dynamics correlate with phytoplankton-derived polysaccharides in surface coastal waters. *ISME J* **8:** 245–248.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673–4680.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D.B., Eisen, J.A., *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66–74.

Vergin, K.L., Beszteri, B., Monier, A., Thrash, J.C., Temperton, B., Treusch, A.H., *et al*. (2013) High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J* **7:** 1322–1332.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73:** 5261–5267.

Supporting information

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

Fig. S1. Analysis of 18S sequences from the 0.2 μm to 1 μm size fraction from SPOT, with pipeline modified to allow detection of 18S rRNA sequences. These have longer PCR products than 16S so the paired ends do not overlap significantly. In the original pipeline, the vast majority of 18S sequences is undetectable because sequences are removed when the paired ends do not overlap. This modified pipeline removed that requirement and analysed only the sequences adjacent to the 515F primer. (A) shows the distribution of total tags in sequences adjacent to the 515F primer where the Eukarya range from 0.58% to 4.3% (mean 1.5%) of all sequences. (B) shows the distribution of major eukaryotic subdivisions via 18S, with most samples dominated by Chloroplastida (primarily Mamiellophyceae), Alveolata (primarily Syndiniales) and Stramenopiles (primarily MAST). The March samples included a spring phytoplankton bloom. The mean abundances for March 23 2011 replicate samples are reported, and the standard error of the mean given as error bars in Fig. S1A.

Fig. S2. Analysis of eukaryotes and attached or large bacteria, > 1 μm size fraction from SPOT, with pipeline modified to allow detection of 18S rRNA sequences, that have longer PCR products than 16S. This analysis uses the same modified pipeline as Fig. S1. Here we analysed separately the merged sequences and the sequences adjacent to the 515F primer. (A) shows the distribution of total tags where the Eukarya range from 8.6% to 35% (mean 17%). The inset shows the percent of Eukarya detected by the standard pipeline that includes merging the paired ends, and they were only detectable in one of the six samples at extremely low levels. (B) shows the distribution of major eukaryotic subdivisions via 18S, with most samples dominated by Metazoa, Rhizaria and Alveolata. The March samples included a spring phytoplankton bloom.

Table S1. *In silico* evaluation of coverage showing per cent hits to various taxa by individual and pairs of primers, analysed by SILVA TestProbe 3.0^{a} or SILVA TestPrime 1.0^{b} and SILVA dataset r123. Zero and one mismatch allowed as shown. Absolute differences greater than 10% between primer sets at zero mismatch are shown in bold.

Table S2. *In silico* primer coverage evaluation of SAR11 clades using SILVA TestPrime 1.0 and SILVA dataset r123.

Per cent of matches (zero or one mismatch allowed, as shown) to sequences in each SAR11 subgroup. Absolute differences > 10% between primers shown in bold.

Table S3. Evaluation of several clustering methods shows mothur's average-neighbour algorithm with pre-clustering, yields mock communities most similar to expected compared with commonly used methods. Each column designates a different clustering method and each row is the clone abundance as an average of four replicates, with a separate column for the standard error of the mean. The name of each clustering method is given in the column headers, indicating if default settings (default) or modified settings (mod) were used, as described in the Materials and Methods. All sequences including the simulated even and staggered fasta file were clustered together. When a taxon appears below the clones it is a different OTU from the OTUs that include the perfect sequences from the simulated mock community files.

Table S4. The difference in total abundance of Thaumarchaea Marine Group I (MGI) was statistically significant between 515F C and Y primers only when amplifying with the 806R primer. Due to low abundance of MGI at shallower depths, only samples from depths ≥ 150 m were evaluated. Both forward primers were used with each reverse primer. The mean ratio between primer combinations and standard error as well as the *P*-values of the Sign Tests are given. The mean and standard error of Bray–Curtis similarity values between MGI communities are also given.

Table S5. Sampling sites and depths.

Table S6. Per cent relative abundance per sample of top 150 eukaryotic OTUs and associated taxonomy for (A) the bacterial (0.2–1 μm) and (B) eukaryotic (> 1 μm) size fraction. Each number represents 100*(number of tags of each taxon ÷ total number of tags).