Evaluating and Improving SSU rRNA PCR Primer Coverage via Metagenomes from Global Ocean Surveys

Jesse McNichol[1], Paul M. Berube[2], Steven J. Biller[2,3], Jed A. Fuhrman[1]

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

2 - Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

3 - Department of Biological Sciences, Wellesley College, Wellesley, MA, USA

Abstract:

Small subunit ribosomal RNA (SSU rRNA) amplicon sequencing comprehensively profiles microbiomes, but results will only be accurate if PCR primers perfectly match environmental sequences. To evaluate whether primers commonly used in microbial oceanography match naturally-occurring organisms, we compared primers with > 300 million rRNA sequences retrieved from globally-distributed metagenomes. The best-performing 16S primers were 515Y/926R and 515Y/806RB which perfectly matched most sequences (~0.95). Considering Cyanobacteria/Chloroplast 16S, 515Y/926R had the highest coverage (0.99), making it ideal for quantifying phytoplankton. For 18S sequences, 515Y/926R performed best (0.88), followed by V4R/V4RB (18S-specific; 0.82). Using Atlantic and Pacific BioGEOTRACES field samples, we demonstrate high correspondence between 515Y/926R amplicons (generated as part of this study) and metagenomic 16S rRNA (median R^2=0.98, n=272), indicating amplicons can produce equally accurate community composition data versus

shotgun metagenomics. Since our pipeline identifies missed taxa, we suggest modifications to improve coverage of biogeochemically-important oceanic microorganisms - a strategy applicable to any environment with metagenomic data.

Significance Statement:

Quantification of taxonomically-informative marker genes using PCR amplification and high-throughput sequencing is a low-cost technique for monitoring distributions and changes of microbial communities across space and time. In order to maximize this procedure's effectiveness, it is essential that environmental organisms match PCR primer sequences exactly. In this study, we developed a software pipeline to evaluate how well commonly-used primers match primer-binding regions from globally-distributed short-read oceanic metagenomes. Our results demonstrate common primer sets vary widely in performance, and that including additional degenerate bases is a simple strategy to maximize environmental coverage. Written in the reproducible snakemake workflow language and publicly accessible, our pipeline provides a general-purpose tool to guide rational design of PCR primers for any environment with metagenomic data.

Introduction:

Amplicon sequencing is a powerful tool for understanding microbial community composition and dynamics in the oceans and other ecosystems (1), but the PCR amplification step is potentially biased due to both technical issues during amplification and mismatches to organisms found in natural ecosystems (2–5). Despite these potential issues, PCR amplicon sequencing retains several advantages that make it desirable to investigate and correct biases. First of all, it is a high-throughput and low cost

technique, making it suitable for large numbers of samples e.g. for global surveys of sediment, water, animal-associated, and other microbial communities (1). Secondly, the targeted nature of the assay means relatively small numbers of sequences are sufficient for detecting rare organisms even when we have no genomes from any of their relatives, due to the conserved nature of the molecule and comprehensive SSU rRNA sequence database. Thirdly, modern “denoising” methods produce exact amplicon sequence variants (ASVs) that can be used as stable and intercomparable biogeographic markers (6). Finally, amplicon datasets are simpler to analyze since they consist of a single gene. Because of this relative simplicity, the complete dataset can be classified with databases such as SILVA (7). In contrast, the complexity of metagenomic samples and current lack of reference genomes for many organisms makes it less possible to classify all sequences.

Recent studies have shown that PCR amplicon sequencing of mock microbial communities can recover known abundances extremely well (3, 8). Conversely, these studies have also underscored the key weakness of PCR-based methods - i.e. that a single mismatch between the primer and template sequences can have a dramatic effect on the measured community composition (e.g. ~10-fold bias against the taxon in question, from a single internal mismatch). In addition, the extent of this bias is not easily predicted if the mismatched base is not at the 3’ terminus of the primer, where biases are known to be severe (3, 8). Therefore, PCR primers should ideally incorporate degenerate bases to match all targeted organisms without sacrificing specificity or PCR efficiency. This is typically accomplished by comparing primer sequences to a SSU rRNA database such as SILVA and then checking for mismatches to organisms known to be abundant in the environment of interest. This general approach has led to marked improvements in primer design, for example by Apprill et al (2015) and Parada et al (2015) who reported that small modifications to existing primers could better quantify the dominant marine taxa SAR11 and *Thaumarchaea* (2, 3).

However, this approach remains limited because it relies on expert knowledge that may be incomplete and does not take into account the potential heterogeneity of natural systems. Wear et al. (2018) addressed this knowledge gap by testing 4 commonly-used primer sets on a 16S mock microbial community derived from natural seawater communities near Santa Barbara (9). These authors tested the effect of their analysis pipeline on recovery of mock community sequences for different primers, and found their primer-pipeline combination had several sequence-specific biases, in some cases due to a primer mismatch. While an important step towards cross-comparing primers in an oceanographic context, Wear et al. note that results are specific to their mock microbial community, analyzed with their particular pipeline, and thus it remains unknown how representative their results are for other ecosystems and other pipelines. More specifically, it is currently unknown how many primer-mismatched rRNA sequences occur across diverse oceanographic environments, a key piece of information for designing field measurements and interpreting results.

To more fully evaluate the extent to which commonly-used primers perfectly match naturally occurring sequences, we developed a workflow to conduct *in silico* primer evaluations with metagenomic SSU rRNA fragments, under the broad assumption that metagenomes have fewer methodological biases than PCR. We use the results from this pipeline to carry out a more objective evaluation of commonly used primers for pelagic marine ecosystems, suggest improvements, and also compare the fidelity of SSU rRNA amplicon sequencing versus shotgun metagenomic sequencing on these globally-distributed samples. Because this pipeline is broadly applicable for any environment where metagenomic data exists, it will allow investigators to design environment-specific PCR probes that optimally recover targeted organisms. Our approach has the added advantage of providing evidence-based data to correct datasets generated with sub-optimal primers in cases where new data cannot be regenerated due to a lack of archived sample material.

Results and Discussion:

Overall primer coverage for pelagic ocean datasets (BiHOBA, TARA, MSPINA, MBARI, SPOT):

To evaluate whether commonly used primer pairs are likely to return biased views of marine microbial communities, we applied our workflow to globally-distributed metagenomic datasets that reflect a range of depths and nutrient regimes (Table 1; Fig 1), testing primers that are commonly used in environmental microbiology (Table 2; (2–4, 10–15)). Metagenomic samples included worldwide sampling expeditions such as Tara oceans (TARA (16)), the Malaspina expedition (MSPINA (17)), and BioGEOTRACES (18). Among these three campaigns, the MSPINA metagenomes focus largely on the deep sea, whereas TARA and BioGEOTRACES sample primarily the sun-lit ocean. In addition, we used time-series data from the Hawaii Ocean Time-Series and the Bermuda Atlantic Time-Series (19, 20) (HOT/BATS; both analyzed with BioGEOTRACES as BiHOBA [BioGEOTRACES-HOT-BATS]), a bloom time-series in Monterey Bay (21) (MBARI), and the San Pedro Ocean Time-Series (22) (SPOT). We also included in the analysis two sites representing suboxic/anoxic systems - the Saanich inlet time-series (Saanich) (23), and samples collected from hydrothermally-altered sediments in Guaymas basin (24) (GuaySed). We also generated new 16S/18S universal amplicon sequences for two BioGEOTRACES cruise transects (GA03/GP13; noted with open circles on Fig 1) which were used for making intercomparisons between metagenomes and amplicons (discussed further below).

We chose the percentage of perfect matches as a metric to evaluate potential PCR performance since previous studies have shown that even a single mismatch can significantly bias results. To summarize data from pelagic ocean ecosystems, we plotted the fraction of metagenomic reads that perfectly matched primer pairs across all pelagic oceanic datasets (excluding GuaySed and

Saanich) with violin plots/boxplots to visualize the distribution of metagenomic primer coverage values (Fig 2). In the following discussion, we mostly use “median coverage” as a summary statistic to compare primer set performance.

Median coverage varied widely among primer pairs, with 515Y/806RB and 515Y/926R showing the most consistently high coverage (Fig 2). The ability of all primers to capture the underlying diversity in the metagenomic samples can be improved to some extent by increasing degeneracies, though this varied significantly between primer sets. Adding a single additional degeneracy fixes the majority of mismatches for some primers (e.g. 785R; discussed further below), whereas others have greater than 2 mismatches to metagenomic sequences and thus would require extensive modifications (Figures S1-S28). Below, we discuss the performance of these primers across 4 broad taxonomic groups (*Archaea*, *Bacteria*, *Cyanobacteria*+Plastidal 16S, and *Eukarya*). We separated *Cyanobacteria* + phytoplankton plastidal 16S from bacterial 16S since these organisms are responsible for the vast majority of marine primary productivity and thus are very important to quantify accurately.

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

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

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

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

In summary, the primer pairs with the highest environmental coverage were 515Y/926R (universal), 515Y/806RB (prokaryote-only), and V4F/V4RB (eukaryote-only). It is worth emphasizing that both universal primer sets have the potential to be equally or more comprehensive for *Eukarya* versus the *Eukarya*-specific primer sets tested here. In addition, for those wishing to quantify the abundance of eukaryotic phytoplankton using chloroplast data, the 515Y/926R primer set nearly perfectly matches all environmental sequences. We note, however, that recognizable dinoflagellate chloroplast rRNA sequences are almost completely missing from the metagenomic sequences (for all primer regions, not just 515Y/926R), and thus will likely not appear in the resulting amplicons regardless of primer choice. This is possibly due to the unusual genomic organization/subcellular

localization of chloroplast genes in dinoflagellates (25). However, dinoflagellate 18S sequences can be recovered from the 515Y/926R primer pair (26), making it possible to reconstruct their abundance *in situ* in combination with the chloroplast 16S. Overall, these results confirm that the careful primer design previously reported for the 515Y/806RB/926R primers has resulted in very high overall coverage in oceanic ecosystems. On the other hand, certain primer sets (e.g. 341F/785R) have very low coverage due to mismatches to dominant organisms. Resulting data may thus be distorted, though we note that removing taxa with mismatches to primers from processed data will recover accurate relative abundances from the remaining taxa (8). This relative abundance correction approach could also be combined with ecosystem-specific full-length 16S databases (27) to generate accurate intercomparisons between datasets derived from different primer sets. This approach will be important to "translate" data from different primer sets, both from legacy samples that cannot be reanalyzed and in the future as sequencing technology / primer design inevitably evolves.

A Quantitative Intercomparison between SSU rRNA Amplicon Sequencing and Metagenomics:

To determine whether amplicon-based strategies recover similar patterns as metagenomes, we generated PCR amplicon sequences for a subset of the BioGEOTRACES samples (cruises GA03 and GP13, using the same DNA), and thus were able to make an "apples to apples" quantitative comparisons between the relative abundances of organisms determined with amplicons vs. metagenomic reads.

To do so, we directly compared the abundance of exact amplicon sequence variants (ASVs) from the 515Y/926R primer pair with metagenomic reads from the same region of the SSU rRNA molecule. ASVs were clustered as necessary to account for the fact that short (150bp) metagenomic

reads could in some cases only be attributed to a broader phylogenetic group (e.g. *Prochlorococcus* and SAR11). The summed abundance of all taxonomic groups, as determined by ASVs and metagenomic reads, was then compared using a linear least-squares regression. For 16S sequences, the strength of the linear correlation was uniformly high across both cruise datasets (median R^2=0.98; range=0.55-1.0; Fig 3a). These patterns were generally consistent across depth for the GA03 cruise which sampled into the mesopelagic where community composition is markedly different from the surface. We do however note that at depths > 150m about 38% of samples had R^2 values below 0.95 (Fig 3a), vs. 9.2% for samples at depths < 150m. These lower correspondences are unlikely to be due to mismatches to the primer region, since predicted primer coverage is uniformly high across depth for the 515Y/926R primer pair across depth (Fig 3b). Other factors such as DNA quality or quantity may have been responsible for this, but we cannot rule out the possibility it involves characteristics of the sequences between the priming regions. Regardless, these results show that amplicon sequencing and metagenomic profiling typically have a very high quantitative correspondence with the primers tested here (515Y/926R).

Improving Existing Primers:

In addition to quantifying mismatches, our pipeline also identifies the specific primer region variants found in particular taxa and their coverage across datasets. This information can guide improvements for any of the primers tested here. In the sections below, we provide three case studies for improving primers using these metagenomic data.

To illustrate how a primer with poor predicted performance could be improved, we use the case of 785R which has a mismatch to the dominant organism SAR11. This could be corrected by adding a single additional degeneracy at position 7 in the primer GACTAC**N**VGGGTATCTAATCC.

However, 785R also has mismatches to a number of abundant chloroplast sequences (Fig 2), which could be corrected with two additional degeneracies **R**A**Y**TACNVGGGTATCTAATCC. Compared to the original primer (GACTACHVGGGTATCTAATCC), these primers would require the synthesis of a larger number of oligonucleotide combinations (12 and 48 vs. 9 originally) but are well within the number of combinations currently in use for other environmental amplification strategies (28). Interestingly, despite being the worst-performing primer for marine ecosystems due to mismatches to the dominant taxon SAR11, we predict that by adding the above degeneracies this primer would achieve nearly perfect coverage for *Bacteria* (Fig S15-S21). This suggests the primer-binding region of 785R is highly-conserved among *Bacteria,* consistent with the initial publication describing this primer as the best overall for prokaryotic taxa across diverse environments (10). This observation also underscores the value of comparing primer sequences with environmental data since it uncovered a significant, unexpected mismatch to an abundant taxon that can now be corrected.

We also investigated how primer sets that already perform well (515Y/806RB/926R) could be further improved. To do so, we identified datasets that had significant potential for coverage improvement (Figures S1-S28) and certain rare order-level taxa that had many mismatches or are known to be biogeochemically-important. The 515Y primer was relatively easy to improve, given that its performance was already very robust across diverse datasets (Fig S1-S28). An increase in degeneracies at four positions (GTG**B**CAGCM**SY**CGCGGT**M**A) was sufficient to resolve the vast majority of mismatches (including to rare taxa such as certain representatives of the *Patescibacteria* a.k.a. candidate phyla radiation (CPR) bacteria) which were previously identified as taxonomic “blind spots” in PCR amplicon analysis (5). These changes led to a primer with a degeneracy of 48, which we term 515Yp (“p” for “pelagic”).

Our data showed that the 806RB primer performed extremely well overall for 16S sequences, but did miss a number of 16S chloroplast sequences mostly from the *Mamielliales* clade, which

includes small picoeukaryotes such as *Ostreococcus* and *Bathycoccus*. As shown in Fig S8-S14, these taxa were still missed when allowing for up to 2 mismatches, meaning they are unlikely to be amplified with the current primer design. Alterations would be possible, but we note that this may be an intentional design choice, since chloroplast amplification may be undesirable in certain environments (29) but is often very useful in euphotic zone marine water samples where it can be used to quantify primary producer abundance. Related to this, we note that a context-dependent advantage of 806RB is that it lacks an 18S binding site. This means that in combination with 515Y(p), nuclear 18S will not be amplified with 806RB - a potential advantage for host-associated microbiome studies where 18S could overwhelm targeted 16S sequences. This is, however, unlikely to occur in general for pelagic ocean samples that are generally dominated by prokaryotic microorganisms (Fig S29), making the 926R primer a feasible choice.

The 926R primer had mismatches to more taxa, and thus required more extensive modifications. For example, Table S1 shows that 926R has mismatches to both *Ectothiorhodospirales* (an uncultivated group related to purple sulfur bacteria (30)), and the *Rickettsiales* (an order that includes mitochondrial sequences in addition to free-living organisms (31)). While less abundant overall, we also noted mismatches in the Tara Oceans dataset to *Brocadiales*, the order containing anammox bacteria (32). Resolving mismatches for all these three taxa as well as those having > 2% overall relative abundance, we arrive at a new primer CCGYCAATT**HH**TTTR**R**GTT**Y** (degeneracy=144), which could be further simplified by removing the trailing wobble base to CCGYCAATT**HH**TTTR**R**GTT (degeneracy=72). Removing the trailing base would allow more efficient amplification of *Rickettsiales* templates, which have a single mismatch at the 3’ end of the primer which was shown to completely prevent amplification of such sequences (9). For example, in the BiHOBA dataset, these modifications increase *Rickettsiales* coverage from ~73% to ~90%. Since the order *Rickettsiales* contains sequences from mitochondrial SSU rRNA, this modification will

potentially allow for more effective quantification of protist abundances, similar to how chloroplasts quantify marine protistan phytoplankton.

While this new version of 926R would increase coverage for all abundant groups sampled in the Tara/BioGEOTRACES expeditions (including eukaryotes), it would still miss some taxa important in low-oxygen regions. For example, the *Campylobacteria* are dominant chemoautotrophs in many sulfidic systems (33), and could be targeted by adding a single further degeneracy (CCGYC**W**ATTHHTTTRRGTT; degeneracy = 144). Since *Campylobacteria* are known to be abundant in certain rare low-oxygen microenvironments such as in sediment trap particles (34), they would be important to include for a primer set used across diverse marine pelagic systems.

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

Moving forward, we recommend oceanographers consider applying these two new degenerate primers for pelagic water column surveys to maximize organismal coverage (515Yp: GTGBCAGCMSYCGCGGTMA; 926Rp: CCGYCWATTHHTTTRRGTT [5’-3’]). If proven to be as quantitative as the original 515Y/926R primers (3, 8), they would provide an affordable, comprehensive (all organisms from Archaea - to Zooplankton), and scaleable method to measure whole community compositions across time and space in the world’s oceans. In turn, this would allow routine measurement of microbial community compositions alongside physical and chemical parameters on

global oceanographic surveys, helping expand our understanding of how oceanic microbes interact with global biogeochemical cycles and respond to global climate change.

Finally, we believe a similar approach could be applied elsewhere to generate ecosystem-specific primer sets. For example, while our revised primer set has high coverage for the vast majority of *Archaea* and *Bacteria* in the Guaymas Basin samples, it remains unclear whether this is broadly true for other marine sediment ecosystems. Our reproducible workflow, combined with expert curation and broader environmental sampling could ensure that sediment taxa of interest are accurately quantified by oligonucleotide primers, or indeed taxa from any environment where deeply-sampled metagenomes exist.

Materials and Methods:

Evaluation of primer coverage/biases using metagenomic reads:

In order to evaluate the performance of our primer set compared to others, we used metagenomic reads to estimate the fraction of SSU rRNA fragments that could be successfully targeted by (i.e. perfectly match) each primer set. In brief, our pipeline retrieves SSU rRNA from fastq-formatted shotgun metagenomic samples, removes sequences containing uninformative repeats, sorts into four organismal categories (*Archaea*, *Bacteria, Cyanobacteria* (including plastidal 16S), and *Eukarya*), and aligns these sequences to a group-specific reference SSU rRNA sequence (e.g. *Saccharomyces cerevisiae* for *Eukarya*). Aligned sequences are then subsetted by primer region, quality-filtered, and compared with primer sequences (Table 2) at 0, 1 and 2-mismatch thresholds (Fig S1-S28). Retrieved reads were then classified using the SILVA 132 database (*Archaea, Bacteria, Eukarya*) and PhytoRef (8, 35) (*Cyanobacteria* + plastidal 16S). The specific software implementation

is described fully in the supplementary material. For those wishing to inspect results further, complete plaintext summaries of output are available online at https://osf.io/gr4nc/ in addition to summary graphs.

These computational steps were implemented with the snakemake workflow engine (36) to create a documented and reproducible pipeline which is freely available (github.com/jcmcnch/MGPrimerEval). It automatically produces tabular output / summary plots and can be applied to new samples and primers by modifying template configuration files. The specific functioning of the three workflows are described in more detail in the supplemental material.

PCR amplification and ASV generation:

DNA used for PCR amplicons was the same DNA material used to produce metagenomes described by Biller et al (2018) (18) except that it was diluted to 0.5 ng/µL in low-EDTA TE buffer prior to amplification. DNA was amplified with the 515Y/926R primers (515Y: GTGYCAGCMGCCGCGGTAA / 926R: CCGYCAATTYMTTTRAGTTT) (3, 8) using 0.5ng of DNA template in a 25 µL reaction with a final primer concentration of 0.3mM. Primers were part of a larger construct with Illumina adapter constructs/barcodes already included, allowing for single-step library preparation. Thermocycling was as follows: an initial denaturation at 95°C for 120s followed by 30 amplification cycles of 95°C × 45s; 50°C × 45s; 68°C × 90s and a final elongation step at 68°C × 300s. Laboratory methods are described in more detail at https://dx.doi.org/10.17504/protocols.io.vb7e2rn. Both 16S and 18S mock communities (staggered and even versions of both (3, 8)) were included in the sequencing run as quality control. Amplicons were sequenced with HiSeq 2x250 RapidRun technology at the USC sequencing center in a run pooled with shotgun metagenomic samples. Raw basecall data was demultiplexed with bcl2fastq (-r 20 -p 20 -w 32

--barcode-mismatches 0) using the 6 basepair reverse indices and subsequently with the forward 5 basepair barcode with cutadapt according to the templates provided at github.com/jcmcnch/demux-notes. Raw data from this step of the pipeline is available on NCBI under the umbrella Bioproject PRJNA659851 and three sub-projects PRJNA658608 (controls), PRJNA658384 (GA03), and PRJNA658385 (GP13)**.**

Sequence data was analyzed with a custom *in-silico* pipeline available at github.com/jcmcnch/eASV-pipeline-for-515Y-926R and based on qiime2 2019.4 (35). Briefly, 16S and 18S sequences were first partitioned using a custom 16S/18S-specific databases derived from the SILVA 132 and PR2 (7, 36) which is available at osf.io/e65rs/. After partitioning, primers were removed from amplicons allowing up to 20% mismatches to the primer, and raw sequences were imported into qiime2 and denoised with DADA2 (6). For 18S sequences that do not overlap, they were denoised as single-end reads by trimming to a consistent length (forward=220bp, reverse=180bp) and concatenating the two reads. Taxonomy was assigned using a naïve Bayesian classification algorithm with the SILVA 132 database subsetted to the amplicon region as a reference (37). Sequences identified as chloroplasts were re-annotated with PhytoRef (38), and 18S sequences were additionally annotated with PR2 (36). A step-by-step protocol for these informatic steps is available at https://dx.doi.org/10.17504/protocols.io.vi9e4h6. After denoising and annotation, biom tables were exported with taxonomy as tsv files and converted to relative abundances which were used as input for the metagenome-amplicon comparison.

Acknolwedgements: This work was supported by the Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems (CBIOMES) grant 549943 to JAF, the Gordon and Betty Moore Foundation Marine Microbiology Initiative grant 3779 and NSF grant OCE1737409. We gratefully acknowledge the effort of the scientists who collected the

BioGEOTRACES field samples analyzed here: Christel Hassler, Debbie Hulston, and Elizabeth Maas (GP13) and Jeremy Jacquot (GA03). We also acknowledge the efforts of Keven Dooley and Madeline Williams who conducted DNA extractions, which was funded in part by grants to Sallie W. Chisholm at MIT that enabled the collection and processing of DNA samples from GEOTRACES cruises. These include the Simons Foundation (Life Sciences Project Award ID 337262, SCOPE Award ID 329108), the Moore Foundation (Grant IDs GBMF495 and GBMF4511), and the National Science Foundation (OCE-1153588, OCE-1356460, and DBI-0424599). We thank S.W.C. for donating the samples and supporting parts of the study.

References Cited:

1. L. R. Thompson, *et al.*, A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

2. A. Apprill, S. McNally, R. Parsons, L. Weber, Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology* **75**, 129–137 (2015).

3. A. E. Parada, D. M. Needham, J. A. Fuhrman, Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* **18**, 1403–1414 (2016).

4. W. Walters, *et al.*, Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. *mSystems* **1** (2016).

5. E. A. Eloe-Fadrosh, N. N. Ivanova, T. Woyke, N. C. Kyrpides, Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nature Microbiology* **1**, 15032 (2016).

6. B. J. Callahan, *et al.*, DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**, 581–583 (2016).

7. C. Quast, *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res.* **41**, D590–D596 (2013).

8. Y.-C. Yeh, J. C. McNichol, D. M. Needham, E. B. Fichot, J. A. Fuhrman, Comprehensive single-PCR 16S and 18S rRNA community analysis validated with mock communities and denoising algorithms. *bioRxiv*, 866731 (2019).

9. E. K. Wear, E. G. Wilbanks, C. E. Nelson, C. A. Carlson, Primer selection impacts specific population abundances but not community dynamics in a monthly time-series 16S rRNA gene amplicon analysis of coastal marine bacterioplankton. *Environmental Microbiology* **20**, 2709–2726 (2018).

10. A. Klindworth, *et al.*, Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucl. Acids Res.*, gks808 (2012).

11. C. S. Fortunato, L. Herfort, P. Zuber, A. M. Baptista, B. C. Crump, Spatial variability overwhelms seasonal patterns in bacterioplankton communities across a river to ocean gradient. *The ISME Journal* **6**, 554–563 (2012).

12. T. Stoeck, *et al.*, Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology* **19**, 21–31 (2010).

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

14. D. Wilkins, E. van Sebille, S. R. Rintoul, F. M. Lauro, R. Cavicchioli, Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects. *Nature Communications* **4**, 2457 (2013).

15. L. A. Amaral-Zettler, E. A. McCliment, H. W. Ducklow, S. M. Huse, A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes. *PLOS ONE* **4**, e6372 (2009).

16. S. Pesant, *et al.*, Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* **2**, 1–16 (2015).

17. S. G. Acinas, *et al.*, Metabolic Architecture of the Deep Ocean Microbiome. *bioRxiv*, 635680 (2019).

18. S. J. Biller, *et al.*, Marine microbial metagenomes sampled across space and time. *Scientific Data* **5**, 180176 (2018).

19. D. M. Karl, M. J. Church, Microbial oceanography and the Hawaii Ocean Time-series programme. *Nat Rev Micro* **12**, 699–713 (2014).

20. D. K. Steinberg, *et al.*, Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. *Deep Sea Research Part II: Topical Studies in Oceanography* **48**, 1405–1447 (2001).

21. B. Nowinski, *et al.*, Microbial metagenomes and metatranscriptomes during a coastal phytoplankton bloom. *Sci Data* **6**, 1–7 (2019).

22. E. T. Sieradzki, J. C. Ignacio-Espinoza, D. M. Needham, E. B. Fichot, J. A. Fuhrman, Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nat Commun* **10**, 1–9 (2019).

23. A. K. Hawley, *et al.*, A compendium of multi-omic sequence information from the Saanich Inlet water column. *Sci Data* **4**, 1–11 (2017).

24. N. Dombrowski, K. W. Seitz, A. P. Teske, B. J. Baker, Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. *Microbiome* **5**, 106 (2017).

25. T. Laatsch, S. Zauner, B. Stoebe-Maier, K. V. Kowallik, U.-G. Maier, Plastid-Derived Single Gene Minicircles of the Dinoflagellate *Ceratium horridum* Are Localized in the Nucleus. *Mol Biol Evol* **21**, 1318–1322 (2004).

26. D. M. Needham, J. A. Fuhrman, Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature Microbiology* **1**, 16005 (2016).

27. M. S. Dueholm, *et al.*, Generation of Comprehensive Ecosystem-Specific Reference Databases with Species-Level Resolution by High-Throughput Full-Length 16S rRNA Gene Sequencing and Automated Taxonomy Assignment (AutoTax). *mBio* **11** (2020).

28. J. C. Gaby, D. H. Buckley, A Comprehensive Evaluation of PCR Primers to Amplify the *nifH* Gene of Nitrogenase. *PLoS One* **7** (2012).

29. S. Thijs, *et al.*, Comparative Evaluation of Four Bacteria-Specific Primer Pairs for 16S rRNA Gene Surveys. *Front. Microbiol.* **8** (2017).

30. A. Oren, "The Family *Ectothiorhodospiraceae*" in *The Prokaryotes*, (Springer, Berlin, Heidelberg, 2014), pp. 199–222.

31. D. N. Fredricks, "Introduction to the *Rickettsiales* and Other Intracellular Prokaryotes" in *The Prokaryotes*, M. D. P. Dr, S. Falkow, E. Rosenberg, K.-H. Schleifer, E. Stackebrandt, Eds. (Springer New York, 2006), pp. 457–466.

32. L. Medina Faull, P. Mara, G. T. Taylor, V. P. Edgcomb, Imprint of Trace Dissolved Oxygen on Prokaryoplankton Community Structure in an Oxygen Minimum Zone. *Front. Mar. Sci.* **7** (2020).

33. B. J. Campbell, A. S. Engel, M. L. Porter, K. Takai, The versatile ε-proteobacteria: key players in sulphidic habitats. *Nature Reviews Microbiology* **4**, 458–468 (2006).

34. D. Boeuf, *et al.*, Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. *PNAS* **116**, 11824–11832 (2019).

35. E. Bolyen, *et al.*, Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* **37**, 852–857 (2019).

36. L. Guillou, *et al.*, The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41**, D597–D604 (2013).

37. N. A. Bokulich, *et al.*, Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90 (2018).

38. J. Decelle, *et al.*, PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Molecular Ecology Resources* **15**, 1435–1445 (2015).

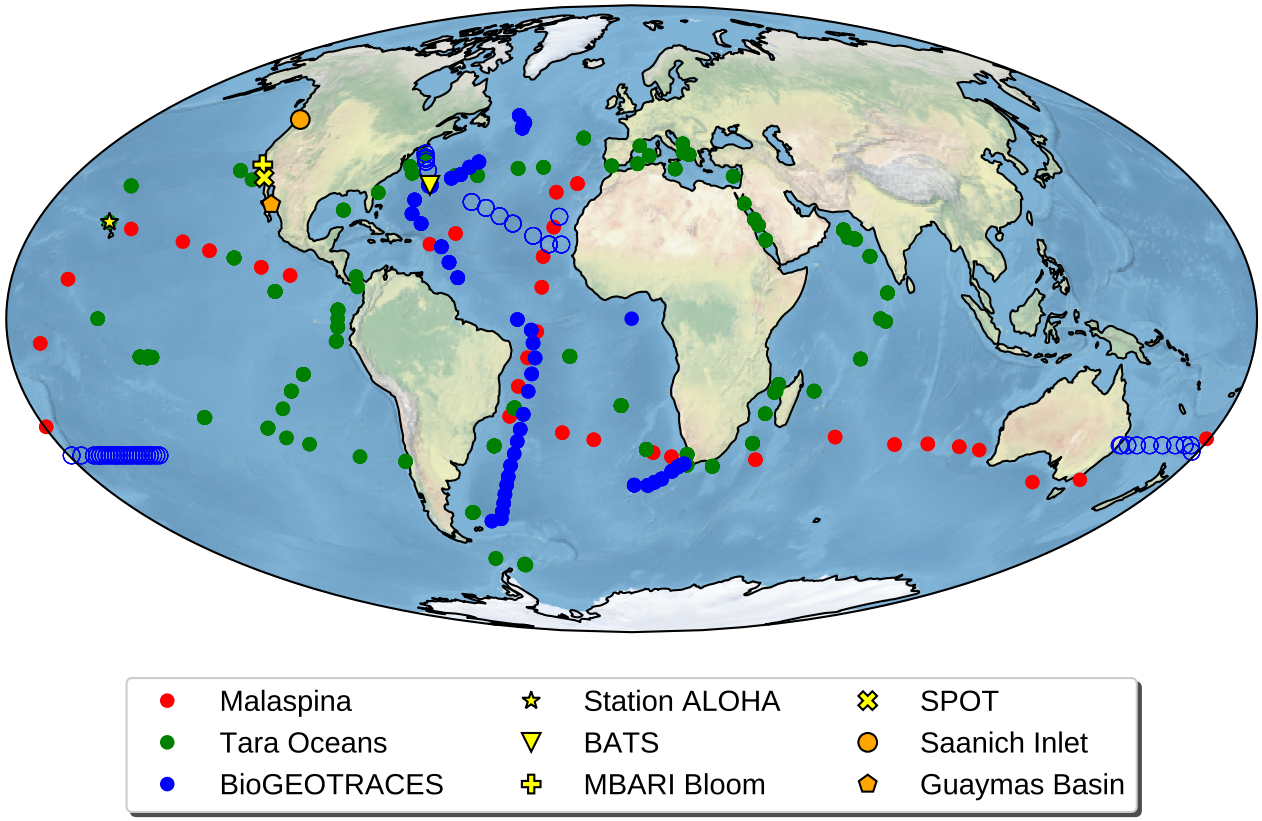

Figure 1: Distribution of metagenomic samples used in this study. BioGEOTRACES samples that were used in the metagenome/amplicon intercomparison are noted with open circles (GA03 - N. Atlantic; GP13 - S. Pacific).

							Median filtered sequences per sample/group/primer/direction**			
Dataset name (**abbreviation**)	Sample and environment description	Sample size fraction(s)	Sample volume (L)	Reference(s)	# samples tested	Total # SSU rRNA sequences	Archaea	Bacteria	Cyanobacteria & Chloroplast	Eukarya
BioGEOTRACES / HOT / BATS (**BiHOBA**)	Mostly euphotic and upper mesopelagic ocean waters; primarily oligotrophic	>0.2µm	0.1	18-20	610	3.16E+07	36.9	256.9	46.3	55.5
Tara Oceans (**TARA**)	Euphotic, deep chlorophyll maximum, and mesopelagic samples	All with upper size fraction <3µm*	≥100	16	458	1.32E+08	109.1	1148	141.7	155.4
Malaspina Deep Metagenomes (**MSPINA**)	Deep ocean metagenomes from samples collected >2000m in depth	0.2 - 0.8 µm (FL), 0.8 - 20 µm (PA)	120	17	58	3.46E+06	26.3	158.4	NA	32.6
MBARI bloom metagenome (**MBARI**)	Coastal surface seawater sampled during phytoplankton bloom across 52 days	0.22–5.0 µm	up to 1	21	84	1.32E+08	25.7	667.7	52.4	68.3
SPOT / Catalina metagenome (**SPOT**)	Surface seawater collected from the San Pedro Ocean Time-Series and Catalina island across 4 seasons	0.22-1 µm	20	22	8	5.04E+05	16.9	434.2	186.2	201.7
Saanich Inlet Metagenomes (**Saanich**)	Depths spanning the oxycline (10, 100, 120, 135, 150, and 200 m)	0.22-2.7 µm	1 or 20	23	87	1.37E+07	96.8	662.6	88.9	21.5
Guaymas sediment metagenomes (**GuaySed**)	Total DNA from > 10g hydrothermally-influenced sediment from 0-15cm depth	NA, whole sediment	NA, 10g sediment	24	4	1.95E+06	1047.3	3065.6	93	433.7

*<0.2µm; 0.2-1.6µm; 0.2-3µm; 0.8-3µm; **groups={1: Archaea,2:Bacteria less Cyanobacteria+Plastids,3:Chloroplast+Plastid,4:Eukarya}, direction=forward and reverse paired-end reads (analyzed separately)

Table 1: Data used in this study. SRA/ENA Accessions available on github.com/jcmcnch/MGPrimerEval in config files.

Primer	SSU rRNA region targeted	Organismal coverage	Primer sequence (5’ - 3’)	Reference(s)
27F	V1-V2	16S only	AGRGTTTGATYMTGGCTCAG	11
338R	V1-V2	16S only	TGCWGCCWCCCGTAGGWGT	11
341F	V3-V4	16S (Bacteria-only)	CCTACGGGNGGCWGCAG	10
785R	V3-V4	16S only	GACTACHVGGGTATCTAATCC	10
V4F	V4	18S only	CCAGCASCYGCGGTAATTCC	12
V4R	V4	18S only	ACTTTCGTTCTTGATYRA	12
V4RB	V4	18S only	ACTTTCGTTCTTGATYRR	13
515Y	V4/V4-V5	16S and 18S	GTGYCAGCMGCCGCGGTAA	2-4
806RB	V4	16S only	GGACTACNVGGGTWTCTAAT	2-4
926R	V4-V5	16S and 18S	CCGYCAATTYMTTTRAGTTT	2-4
926wF	V6-V8	16S and 18S	AAACTYAAAKGAATTGRCGG	14
1392R	V6-V8	16S and 18S	ACGGGCGGTGTGTRC	14
1389F	V9	18S only	TTGTACACACCGCCC	15
1510R	V9	18S only	CCTTCYGCAGGTTCACCTAC	15

Table 2: Primers tested in this study.

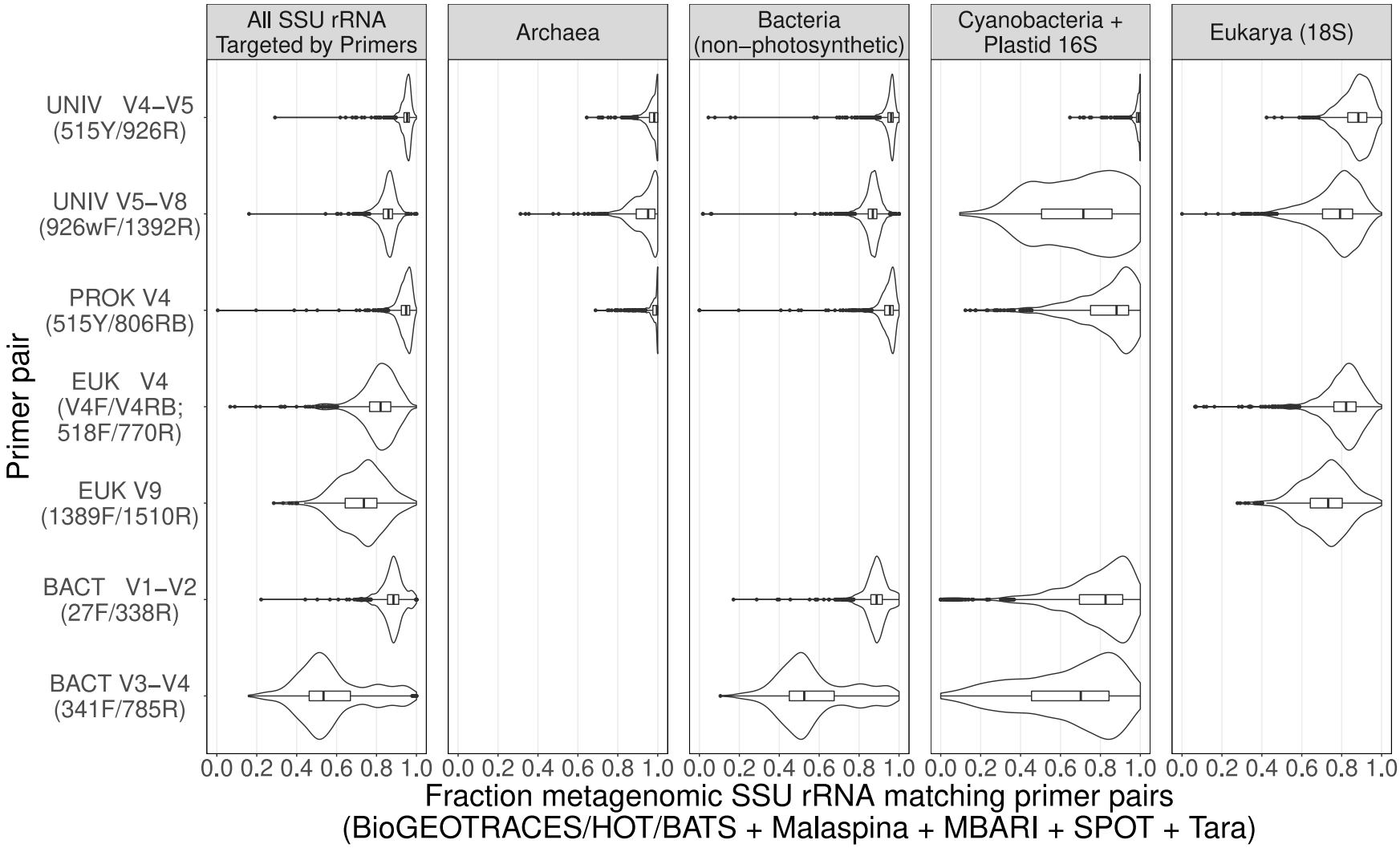

Figure 2: Violin plot with inlaid boxplot showing coverage for oligonucleotide PCR primers based on comparisons with globally-distributed oceanic metagenomes. The shapes of the violin plots illustrate the distribution of primer coverage across all samples. Note that the "All SSU rRNA" panel represents all of the SSU rRNA targeted by a particular primer set (e.g. for 1389F/1510R this includes only Eukarya *while for 926wF/1392R it includes all 4 taxonomic groups). Data are corrected for predicted taxonomic overlap between forward and reverse primers. Coordinates refer to locations of primer alignments to the* Escherichia coli *16S rRNA gene (strain K-12; substrain MG1655).*

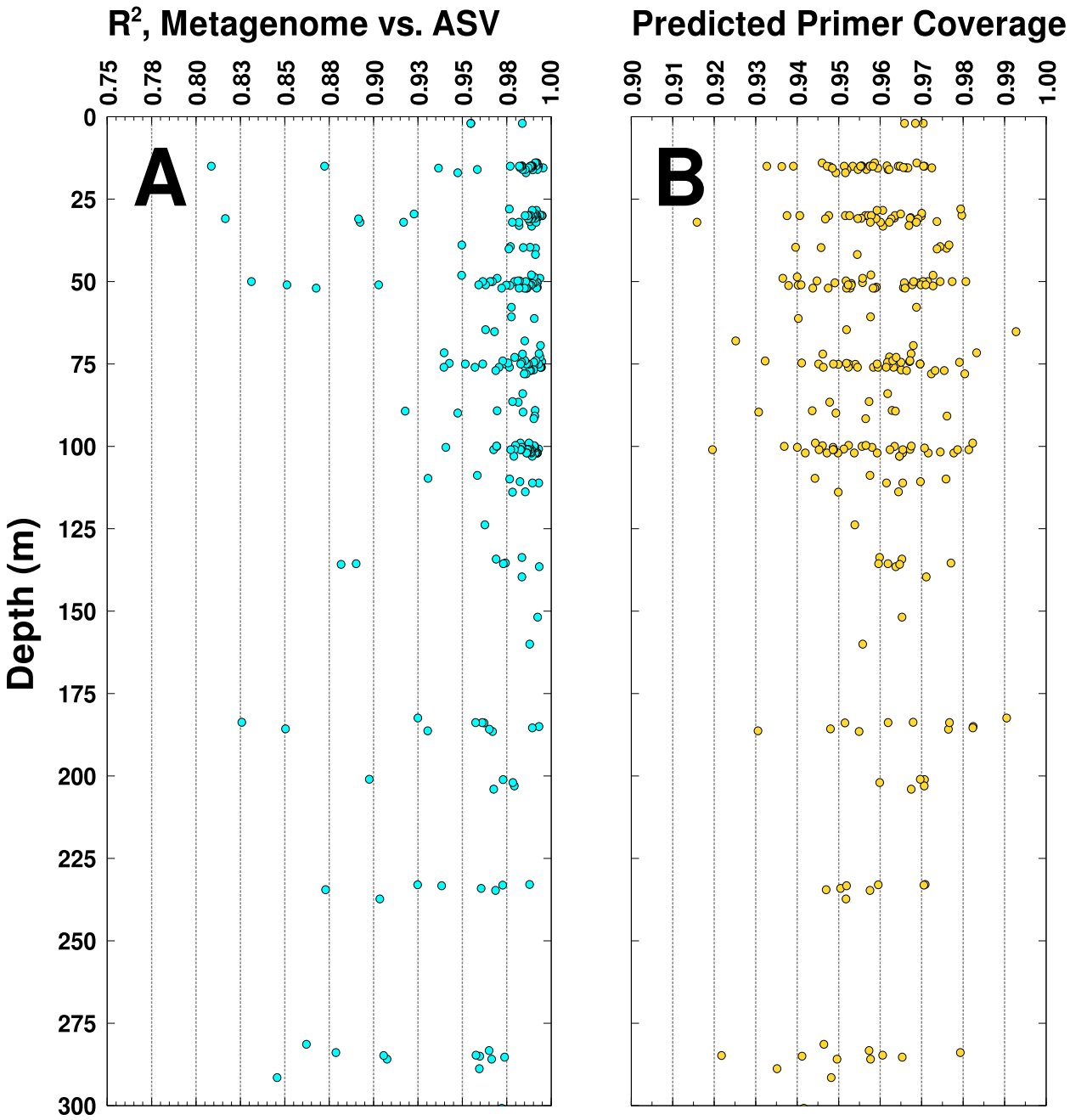

Fig 3: A) Metagenome-amplicon quantitative correspondence across all depths for the GA03 cruise (BioGEOTRACES), showing the R^2 value of the plot of the relative abundances of amplicon sequence variants (ASVs; 515Y/926R primers) vs relative abundances of metagenomic reads from the same organisms (see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

<u>Supplementary Information</u>:

Supplementary Methods:

Software Implementation: Software dependencies are handled by conda implemented in snakemake (1). Examples of how the pipeline can be run are available at github.com/jcmcnch/MGPrimerEval in the "runscripts" folder. Both the "classify" and "compare" workflows depend on the output of the "compute" workflow. The "classify" workflow is for detailed exploration of the taxonomy of retrieved sequences and the "compare" workflow allows for intercomparisons between SSU rRNA amplicon sequence variants (ASVs) and metagenomic SSU rRNA from the same amplicon region.

Compute workflow ("Snakefile-compute.smk"; Fig S30): This portion of the analysis calculates the coverage of primer sequences for metagenomic samples specified in a configuration file. Results are calculated as the fraction of total SSU rRNA reads matching the primer at 0, 1, and 2-mismatches for 4 taxonomic partitions as described below. A rule graph summarizing the steps is shown in Figure S30. Briefly, fastq reads are first cleaned using fastp's (v 0.14.1) default parameters and trimming 3 trailing and leading bases (2). To retrieve SSU rRNA sequences from metagenomes, we used the phyloFlash (v 3.3b1) package (3) which uses bbmap (v 38.69) (4) and a curated version of SILVA132 (5) to retrieve both 16S and 18S sequences. Since retrieved sequences contained a small fraction of low-complexity repeats that were not homologous to the SSU rRNA, we used komplexity (v 0.3.6) (6) to remove these likely spurious hits.

SSU rRNA sequences thus identified were then split into 4 different pools (Eukaryota, Archaea, Bacteria (non-cyanobacterial), and Cyanobacteria (including chloroplasts)) using bbsplit.sh v38.22 (4) comparing against curated subsets of the SILVA/PR2 reference databases (5, 7) (sequences

and a template script for creating databases are available at osf.io/e65rs/). Partial SSU rRNA sequences were then aligned using pyNAST v1.2.2 (8) to group-specific references available as part of the github repository. Primer-specific subsets were then parsed from the alignment using python/biopython (9), only outputting sequences that overlap the primer region plus 5 leading or trailing bases. The 1-based coordinates of each primer region (for each organismal reference) must be supplied by the user; several widely-used primers are already included in template configuration files.

To place all reads in the correct orientation, fastq files were reverse-complemented based on information found in pyNAST alignment headers. cutadapt (version 1.18) (10) was used to identify primer regions in fastq sequences with 0, 1, 2, and 6-mismatches to the primer. Only fastq sequences with 6 or fewer mismatches to the primer were further analyzed to exclude any spurious matches caused by pyNAST misalignments. Based on primer comparisons to references, this is likely to include even highly divergent taxa so long as they have a true primer binding site (NB: this is not true for some primers designed to specifically target a particular group - e.g. the 806RB primer does not contain a binding site for eukaryotic sequences). Primer coverage at each mismatch cutoff was then calculated using only sequences where all bases in the primer region plus 5 leading and trailing bases had a phred score of > 30. Plots shown in the main and supplemental text were generated with python/cartopy (11), ggplot2 (12), veusz (13), and snakemake (1) (Fig 1; Fig 2, Fig S1-S29; Fig 3; Fig S31-S32 respectively).

Classify workflow ("Snakefile-classify.smk"; Fig S31):

In order to determine which taxa were missed by a particular primer set and provide recommendations about how to further improve existing primers' performance, we implemented an optional subsequent workflow to taxonomically classify reads and provide summary statistics for each

group. Briefly, full-length fastq files from mismatched taxa are classified with VSEARCH 2.10.4 (14) and the SINTAX algorithm (15) against the SILVA132 nr99 database (5) or PhytoRef (16) databases (for cyanobacteria/plastids) that had been converted into ".udb" format. The same procedure was followed for matches except a maximum of 5,000 reads were randomly subsampled using "filterbyname.sh" (4) to reduce computation time.

Summary statistics and output files were then generated for each unique combination of group, primer, and mismatch threshold (e.g. EUK, 926R, 0-mismatch) at order-level resolution since informative taxonomic annotations could generally not be made at finer taxonomic levels. Summaries contain the fraction of each taxon not matching the primer at a given mismatch threshold, as well as the fraction of total sequences represented by these mismatches. In addition, multiple sequence alignments of mismatches and their relative abundances are generated, with summary plots automatically generated showing the relative abundances of particular mismatches both for the whole dataset and subsetted by individual taxa. In addition, the same output files are generated for specific order-level taxa, allowing an investigator to see whether a given primer covers a particular taxon of interest. For the datasets analyzed in this study, raw output files/plots can be downloaded at https://osf.io/gr4nc/.

Our pipeline generates coverage information at the individual primer level since metagenomic reads will rarely/never cover the entire amplicon region (depending on lengths of sequence reads and the amplicon region). However, it is advantageous to present results in terms of predicted coverage for the primer set. If the mismatches to the forward and reverse primer were completely non-overlapping (i.e. the mismatches were to completely different organisms) then the mismatched fractions would be simply additive ("worst case" as mentioned in main text and shown in Fig 3). In contrast, if organisms overlapped between the forward and reverse primer, the largest mismatched fraction for that organism could be used. We used the taxonomic assignments discussed

above to estimate this fraction of overlapping sequences ("rule normalize_summaries" and its dependent rules/scripts) and thus correct the primer coverage values as shown in Fig 2.

Compare workflow ("Snakefile-compare2tags-16s.smk" and "Snakefile-compare2tags-18s.smk"):

This workflow compares the abundance of taxa quantified by universal 16S and 18S amplicons using the 515Y/926R primers with abundances of taxa from metagenomic SSU rRNA in the same region of the molecule targeted by the PCR primers. The primer region is targeted by subsetting pyNAST alignments by primer coordinates. Only reads with at least 100bp overlap with the amplicon region are retained, and subsetted reads from the three prokaryotic categories are merged for the 16S comparison. For 18S reads that do not overlap (since the amplicon region is longer than the 2x250 sequencing length employed here), they were subsetted to the trim length of the partial amplicons (e.g. 220bp for forward read, 180 bp for reverse read).

These metagenomic sequences were then compared using BLASTn (17) (minimum 97% similarity, requiring 100% overlap) against a sample-specific database of exact amplicon sequence variants (ASVs) generated with DADA2 (18) that had a nonzero abundance in that particular sample. In order to make a direct comparison between the two data sources, the ASVs were clustered for each sample based on BLAST results with networkx (19) to account for the resolution mismatch between short metagenomic reads and longer, denoised amplicons. Total relative abundances of each cluster were plotted using seaborn (https://seaborn.pydata.org/) and sample-sample correspondence was evaluated using least squared linear regression with the stats.linregress function in scipy (20). The R^2 value of this regression was used to evaluate general trends in MG-ASV sample-sample

correspondence, and was plotted against both environmental covariates and the “worst-case” (simply additive) predicted primer performance (Fig 3b).

Amplicon Analysis Pipeline Supplementary Information:

SSU rRNA partitioning was carried out with the database mentioned in the main text (osf.io/e65rs/) using bbsplit.sh (4). These databases were constructed from SILVA132 (5) and PR2 (7) by first clustering with CDHIT (21) at 95% identity to reduce total data volume and subsequently curated to remove sequences that were extremely long or short relative to the average sequence length. Demultiplexing and trimming steps were carried out with cutadapt (10), and trimming prior to denoising (for 18S sequences) was carried out with bbduk.sh (4).

Supplementary Results:

Technical evaluation of pipeline:

After retrieval, quality-filtering, and subsetting to the primer region, dozens to hundreds of rRNA sequences were recovered for each group, sample, read direction (fwd & rev reads), and primer combination (Table 1). Considering the pelagic ocean datasets introduced above and discussed in detail below (BiHoBa, TARA, MSPINA, MBARI, SPOT), we retrieved between 158-1148 primer-matching regions for *Bacteria* for each combination (medians within datasets; Table 1), whereas lower numbers were retrieved for the remaining three groups of *Archaea*, *Cyanobacteria*+Plastids and *Eukarya* (on average 17-109, 46-186 and 32-202 respectively) most likely due to their lower relative abundances in the environmental DNA. Among the primers, smaller number of sequences were identified for the two

primers located near the end of the SSU rRNA molecule. For example, for Tara samples, a median of 455 sequences were retrieved per sample for 27F (close to end of molecule) for *Bacteria* vs. 1299 per sample for 338R (middle of molecule). Similarly, 58 eukaryotic sequences were retrieved per sample with 1510R vs. 136 per sample for 1389F.

With the exception of regions near the end of the SSU rRNA molecule, sequence coverage was mostly evenly distributed across the alignments to group-specific references. Some relatively small regions had lower coverage, most likely due to their absence in dominant environmental taxa. For example, the SAR11 isolate HTCC1062 is missing nucleotides (relative to the *E. coli* reference) corresponding to small portions of loops in the V1, V2, V3, and V9 regions (22). We note also that phyloFlash retrieved approximately double the total number of 18S sequences compared to a graftM (23) package based on an alignment of the greengenes database (data.ace.uq.edu.au/public/graftm/7/) which was employed in an earlier version of our pipeline. In addition, the coverage across the *S. cerevisae* 18S reference was much more even with phyloFlash-retrieved 18S rRNA vs. graftM. No such bias was noted for 16S sequences, suggesting this graftM package may be appropriate for 16S retrieval but not 18S.

Primer Limitations and Potential Improvements:

Table S1 shows a summary of the top 3 most dominant mismatched orders for each primer and group for the BiHoBa dataset. These samples mostly represent surface and upper mesopelagic environments across all organismal size classes (Table 1), making it useful for exploring general patterns of primer performance for oceanic ecosystems. Among the most prominent mismatches to *Archaea* are those between *Nitrosopumilales* MGI *Archaea* and primers 341F or 1389F. Neither of these primers were specifically designed for capturing archaeal sequences but our results show they

could be modified to include most *Archaea* as shown in Figs S1-S7 for 341F, and likely also for 1389F given that a universal primer with high coverage exists for the same general region (1392R). For other primers, the dominant mismatches include abundant groups such as MGII, and MGIII but we also note a smaller number of mismatches from the poorly-described *Aigarchaeales* and *Woesearchaeia.*

For *Bacteria*, many of the abundant mismatches were already discussed in the main text. Other patterns of particular note are mismatches to abundant members of the SAR86 clade and *Flavobacteriales* for 341F/338R primers. Together, these groups represent over 4% of the total sequences, meaning they would likely be desirable to correct in the future. Readers should note that the dominant taxa in Table S1 are the most abundant sequences, and there are a number of rare bacterial taxa that have very low coverage. These include *Poribacteria*, a number of members of the candidate phyla radiation or CPR, *Rubrobacterales* (*Actinobacteria*), and *Chlamydiales*. Interestingly, some members of these taxa (e.g. *Poribacteria*, CPR) had more than one mismatch to the 926R primer set, meaning they are likely very evolutionarily-divergent from other organisms and may require a more extensive primer redesign to quantify them fully. However, we note that these organisms are very rare in these datasets, so their absences are unlikely to bias overall community profiles in oxic marine environments and that the additional degeneracies proposed in the main text will increase coverage for some of these taxa.

With respect to the cyanobacterial/chloroplast fraction, while the 515Y/926R primer pair achieves nearly perfect coverage, the other primers generally have significant mismatches to one or more abundant chloroplast groups, including the *Mamelliophyceae*, *Prasinophyceae*, and *Prymnesiophyceae*. This means that care must be used in interpreting chloroplast results from some popular primer sets which may miss important taxa.

Among eukaryotic (nuclear 18S) sequences, the most commonly mismatched taxa included *Dinoflagellata*, *Discicristata*, *Protalveolata*, *Ochrophyta*, *Charophyta*, and *Metazoa.* As noted above,

the universal primer sets appear to perform nearly as well or better versus *Eukarya*-specific primers. Improvements described in the main text do increase 18S coverage for the 515Y/926R primers, but a fully comprehensive 18S primer may be less feasible since primer binding regions have a much higher sequence diversity versus 16S. This may be due to large copy numbers of SSU rRNA, some of which may become pseudogenized and undergo random evolutionary drift (24). In practical terms, this means that some of these more rare variants will inevitably be missed by primers since creating an overly-degenerate primer could cause problems with non-specific amplification during PCR. The same may be true for some mitochondrial 16S genes that have become pseudogenized after being incorporated into the host nuclear genome (25).

Similarly, one of the most challenging datasets for improving our primers came from the Guaymas basin (Table 1). Anoxic sediment environments are known to host unusual *Archaea* that are now recognized to be much more diverse than previously thought and are missed by some PCR primers (25). For 926R, we identified another degeneracy that combined with an additional 3' trim would target many of these rare *Archaea* (CCGYCWATTHH**Y**TTRRGT). However, this would increase total degeneracy to 288, and it is not clear whether this level of degeneracy would be desirable. Instead, it may be more effective to try to use our general approach to design an environment-specific primer for sediment microbiomes.

A Practical Consideration for the use of the 515Y/926R primer set:

One concern about using truly universal primers is the amplification of non-microbial sequences, e.g. in the case of host-associated microbiomes. This has led to the design of prokaryotic-specific primer sets (e.g. 515Y/806RB), that intentionally do not amplify 18S sequences. However, universal primers retain the advantage of quantifying a broad swathe of the entire biological

community together. Therefore, such primers could be potentially be very useful if 18S sequences do not overwhelm the SSU rRNA sequences in typical samples.

To test this, we quantified the total fraction of 18S (out of the total 16S + 18S) for the unfractionated (>0.2µM) BiHoBa metagenomes. For these data, we show 18S sequences rarely exceed ~1/3 of the total SSU rRNA sequences (Fig S29), indicating that universal primers will still be useful even when whole seawater samples are analyzed, providing quantification of both prokaryotic and eukaryotic communities simultaneously. The fraction of eukaryotic SSU rRNA relative to total 16S + 18S averaged 19.5 +/- 0.09 % (range = 0.024 - 0.7). Samples with large fractions of eukaryotes (>0.25) were dominated by *Metazoa*, mainly composed of taxa such as *Arthropoda, Platyhelminthes,* and *Hydrozoa*. These animal-derived sequences have the potential to overwhelm the signal from other environmental eukaryotes, especially for unfractionated samples. This can be alleviated to some extent by increasing overall sequence coverage such that sufficient non-metazoan 18S sequences can also be recovered.

Supplemental References Cited:

1. J. Köster, S. Rahmann, Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

2. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

3. H. R. Gruber-Vodicka, B. K. B. Seah, E. Pruesse, phyloFlash – Rapid SSU rRNA profiling and targeted assembly from metagenomes. *bioRxiv*, 521922 (2019).

4. B. Bushnell, BBMap short read aligner. *https://sourceforge.net/projects/bbmap/*.

5. C. Quast, *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res.* **41**, D590–D596 (2013).

6. E. Clarke, komplexity: A method of assessing sequence complexity based on kmer frequencies. *https://github.com/eclarke/komplexity*.

7. L. Guillou, *et al.*, The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41**, D597–D604 (2013).

8. J. G. Caporaso, *et al.*, PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**, 266–267 (2010).

9. P. J. A. Cock, *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

10. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

11. Met Office, *Cartopy: a cartographic python library with a matplotlib interface* (2010).

12. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2009) https:/doi.org/10.1007/978-0-387-98141-3.

13. J. Sanders, Veusz: A Scientific Plotting Package. *https://github.com/veusz/veusz*.

14. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

15. R. C. Edgar, SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, 074161 (2016).

16. J. Decelle, *et al.*, PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Molecular Ecology Resources* **15**, 1435–1445 (2015).

17. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

18. B. J. Callahan, *et al.*, DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**, 581–583 (2016).

19. A. A. Hagberg, D. A. Schult, P. J. Swart, Exploring Network Structure, Dynamics, and Function using NetworkX in *Proceedings of the 7th Python in Science Conference*, G. Varoquaux, T. Vaught, J. Millman, Eds. (2008), pp. 11–15.

20. P. Virtanen, *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

21. W. Li, A. Godzik, CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

22. P. Yarza, *et al.*, Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Micro* **12**, 635–645 (2014).

23. J. A. Boyd, B. J. Woodcroft, G. W. Tyson, GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes. *Nucleic Acids Res* **46**, e59–e59 (2018).

24. S. R. Santos, R. A. Kinzie, K. Sakai, M. A. Coffroth, Molecular Characterization of Nuclear Small Subunit (ISS)-rDNA Pseudogenes in a Symbiotic Dinoflagellate (Symbiodinium, Dinophyta). *Journal of Eukaryotic Microbiology* **50**, 417–421 (2003).

25. D. Bensasson, D.-X. Zhang, D. L. Hartl, G. M. Hewitt, Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology & Evolution* **16**, 314–321 (2001).

Table S1: The fraction of total identified sequences mismatched for the top 3 most numerically abundant mismatched orders in the BioHOBA dataset (when allowing 0 mismatches). Primers that have "NA" for particular groups often represents an intentional design choice to avoid off-target amplification.

Primer	Archaea	Bacteria (without Cyanobacteria and plastid 16S)	Cyanobacteria (including plastid 16S)	Eukarya (18S nuclear SSU rRNA only)
27F	NA, no apparent primer binding site	Rickettsiales (0.008), Ectothiorhodospirales (0.008), SAR11 clade (0.002)	Prasinococcales (0.011), Cyanobactera;Subsection I (0.003), Chrysophyceae-Synurophyceae (0.002)	NA, no apparent primer binding site
338R	NA, no apparent primer binding site	SAR86 clade (0.027), Flavobacteriales (0.015), SAR11 clade (0.009)	Prymnesiophyceae (0.123), Bacillariophyta (0.006), Prasinococcales (0.005)	NA, no apparent primer binding site
341F	Nitrosopumilales (0.51), MG II (0.41), MG III (0.07)	SAR86 clade (0.028), Flavobacteriales (0.016), SAR11 clade (0.016)	Prymnesiophyceae (0.121), Bacillariophyta (0.006), Prasinococcales (0.005)	NA, 5-mismatch to reference
785R	MG II (0.003), Nitrosopumilales (0.002), MG III (0.0004)	SAR11 clade (0.40), Marinimicrobia (SAR406_clade); unclassified order/class (0.006), Verrucomicrobiae;Arctic97B-4 (0.003)	Mamiellophyceae (0.103), Prymnesiophyceae (0.006), Cyanobacteria;SubsectionI (0.001)	NA, 3-mismatch to reference
V4F	NA, 5-mismatch to reference	NA, 3-mismatch to reference	NA, 4-mismatch to reference	Dinoflagellata (0.022), Metazoa (Animalia) (0.010), Protalveolata (0.007)
V4R	NA, 8-mismatch to reference	NA, 7-mismatch to reference	NA, many mismatches to reference	Dinoflagellata (0.081), Discicristata (0.027), Protalveolata (0.026)
V4RB	NA, 8-mismatch to reference	NA, 7-mismatch to reference	NA, many mismatches to reference	Dinoflagellata, (0.040), Discicristata (0.013), Protalveolata (0.012)
515Y	Nitrosopumilales (0.003), MG II (0.002), MG III (0.0003)	Rickettsiales (0.002), SAR11 clade (0.001), Chlamydiales (0.0006)	Cyanobacteria;Subsection I (0.003), Prymnesiophyceae (0.001), Pelagophyceae (0.000)	Dinoflagellata (0.021), Discicristata (0.014), Metazoa (Animalia) (0.010)
806RB	Nitrosopumilales (0.002), MG II (0.002), Aigarchaeales (0.001)	Marinimicrobia (SAR406_clade); unclassified order/ class (0.006), Flavobacteriales (0.005), SAR202 clade (0.002)	Mamiellophyceae (0.108), Prymnesiophyceae (0.014), Prasinophyceae (0.013)	NA, many mismatch to reference
926R	Nitrosopumilales (0.016), MG II (0.001), Woesearchaeia; unclassified order (0.0007)	Ectothiorhodospirales (0.007), Rickettsiales (0.004), SAR11 clade (0.002)	Cyanobacteria;Subsection I (0.003), Prymnesiophyceae (0.000), Dictyochophyceae (0.000)	Dinoflagellata (0.015), Protalveolata (0.012), Metazoa (Animalia) (0.010)
926wF	Nitrosopumilales (0.016), MG II (0.001), Woesearchaeia; unclassified order (0.0007)	Rhodospirillales (0.041), Ectothiorhodospirales (0.007), Rickettsiales (0.004)	Cyanobacteria;Subsection I (0.003), Prymnesiophyceae (0.000), Dictyochophyceae (0.000)	Metazoa (Animalia) (0.040), Ochrophyta (0.033), Dinoflagellata (0.016)
1392R	Nitrosopumilales (0.002), Woesearchaeia; unclassified order (0.002), MG II (0.001)	SAR11 clade (0.003), Clostridiales (0.001), Rickettsiales (0.001)	Mamiellophyceae (0.227), Dictyochophyceae (0.011), Bacillariophyta (0.002)	Dinoflagellata (0.014), Protalveolata (0.010), Metazoa (Animalia) (0.008)
1389F	Nitrosopumilales (0.53), MG II (0.37), MG III (0.05)	Dadabacteriales (0.002), Clostridiales (0.002), SAR11 clade (0.002)	Mamiellophyceae (0.167), Cyanobacteria;Subsection I (0.103), Trebouxiophyceae (0.055)	Dinoflagellata (0.013), Protalveolata (0.008), Metazoa (Animalia) (0.006)
1510R	NA, 4-mismatch to reference	NA, 4-mismatch to reference	Prasinophyceae (0.393), Pelagophyceae (0.375), Pavlovophyceae (0.054)	Metazoa (Animalia) (0.055), Charophyta (0.036), Protalveolata (0.012)

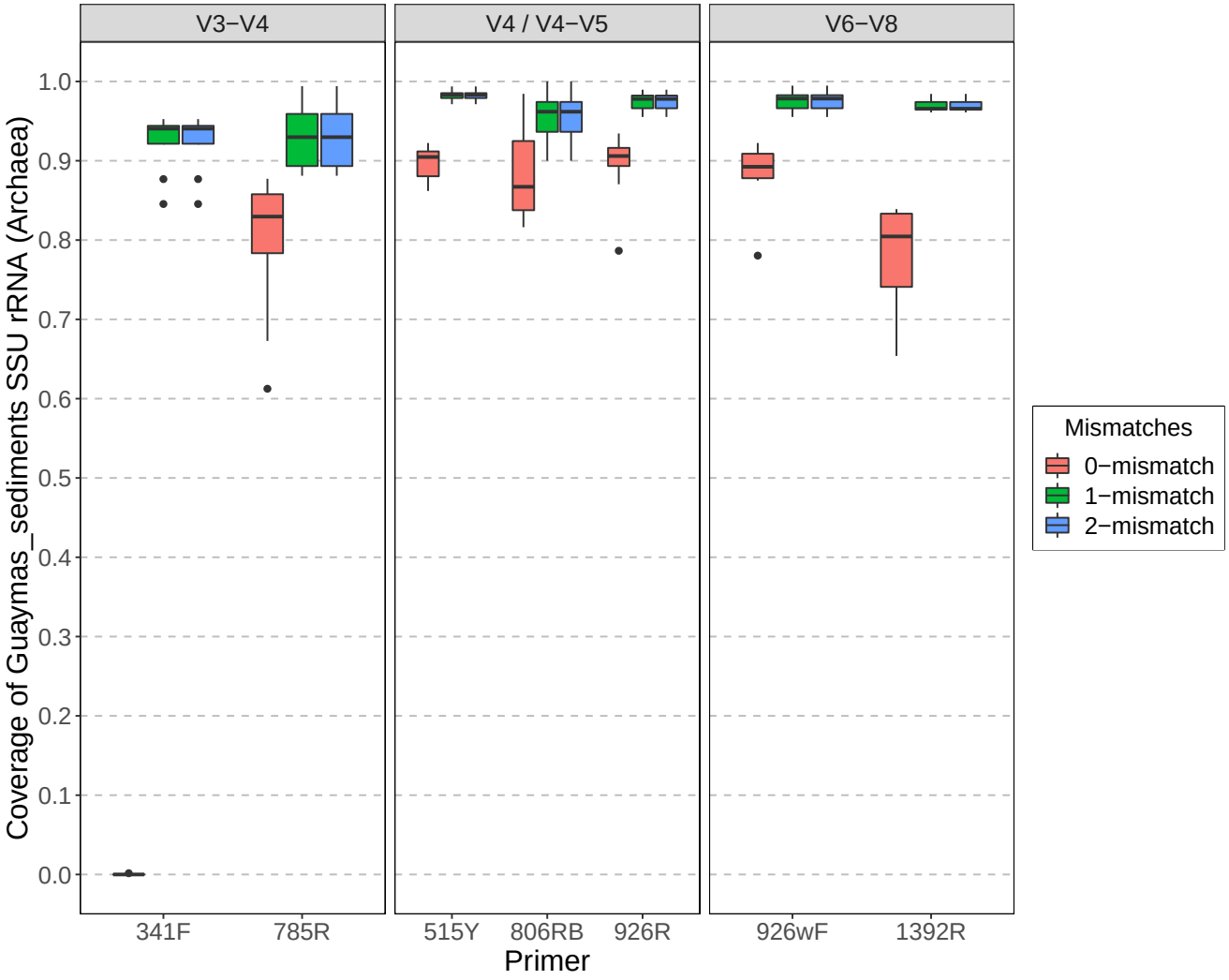

Fig S1: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Archaea* using metagenomes derived from Guaymas Basin sediments.

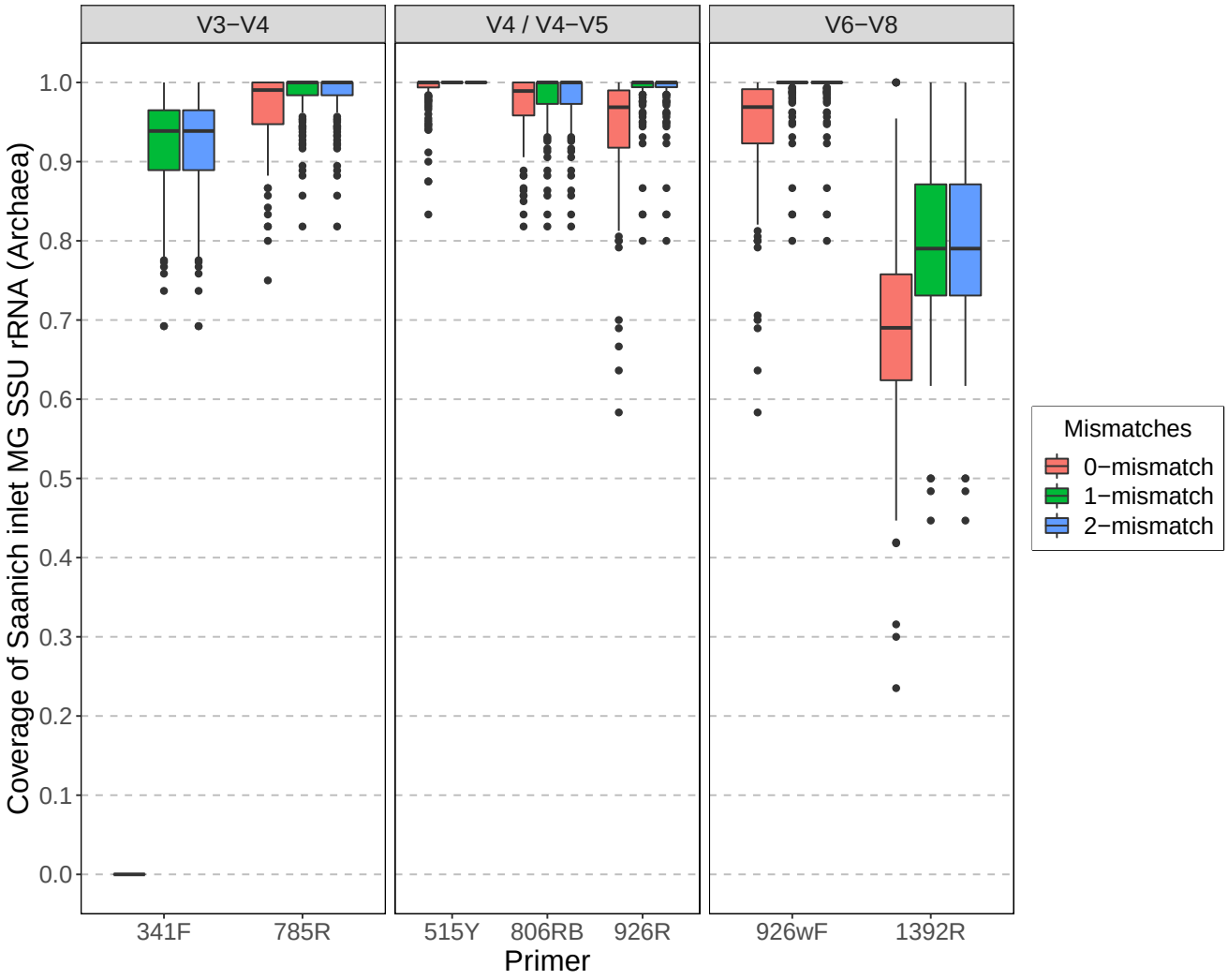

Fig S2: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Archaea* using metagenomes derived from Saanich Inlet water samples.

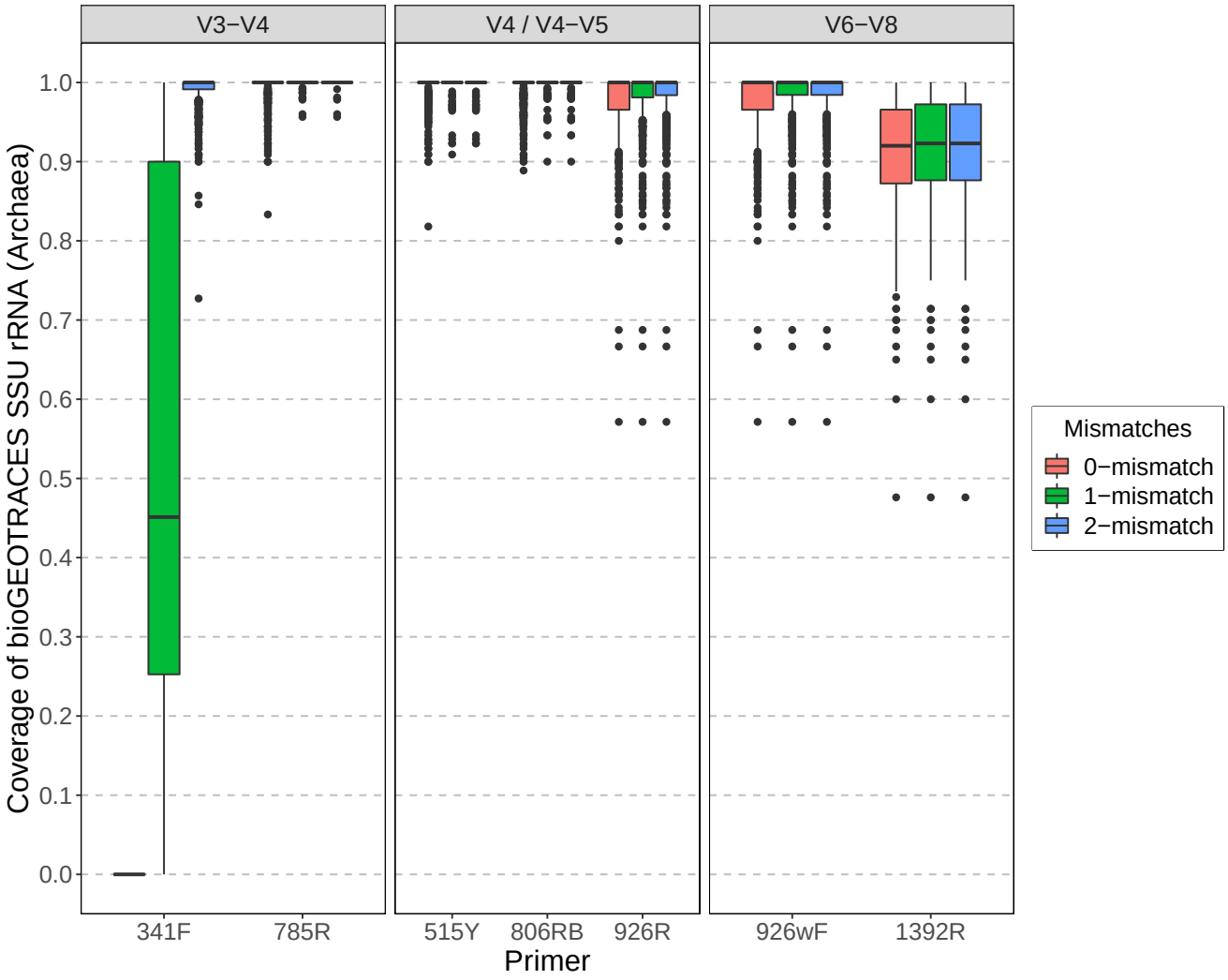

Fig S3: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Archaea* using metagenomes derived from BioGEOTRACES/BATS/HOT water samples.

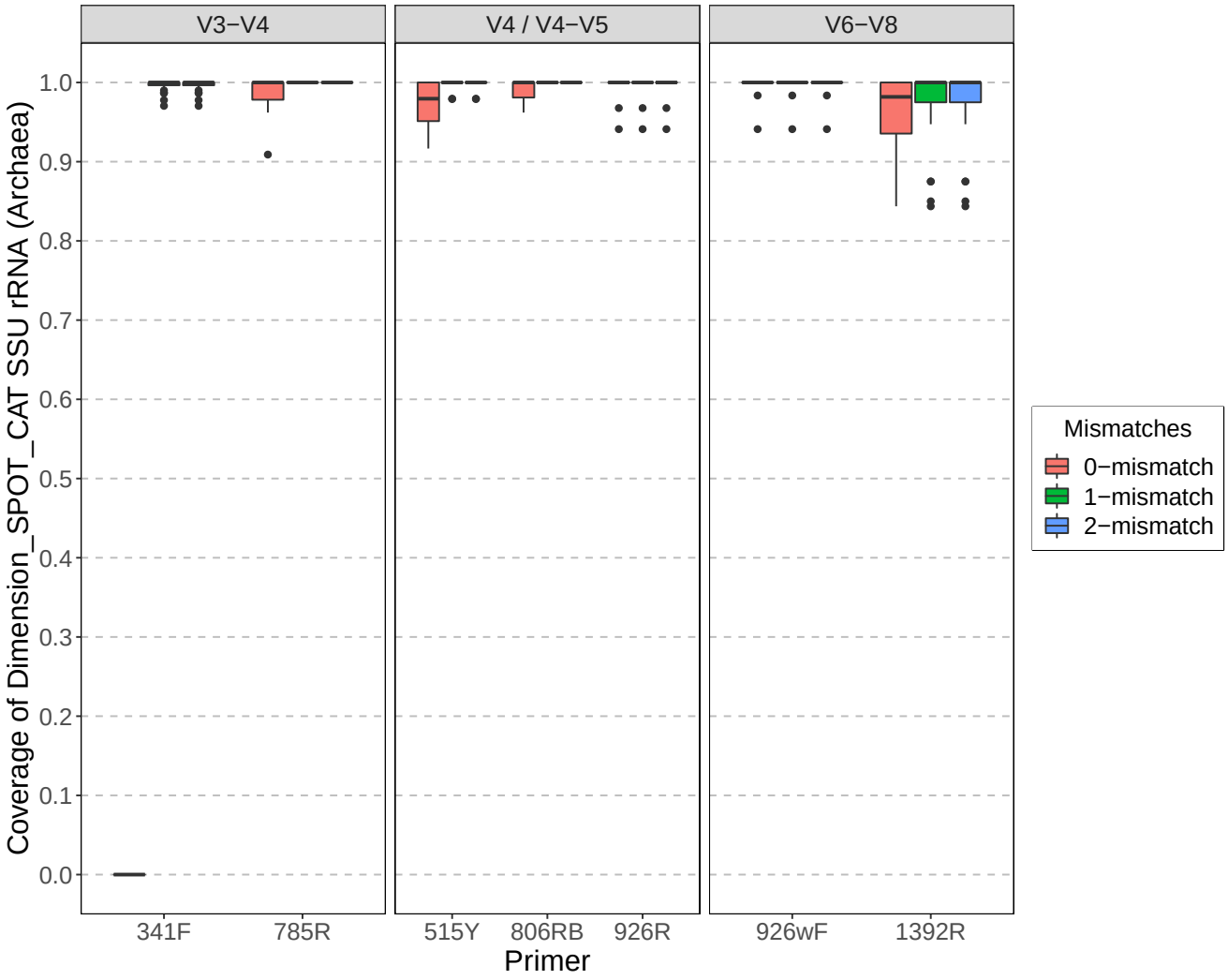

Fig S4: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Archaea* using metagenomes derived from SPOT/Catalina water samples.

Fig S5: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Archaea* using metagenomes derived from Malaspina deep-sea water samples.

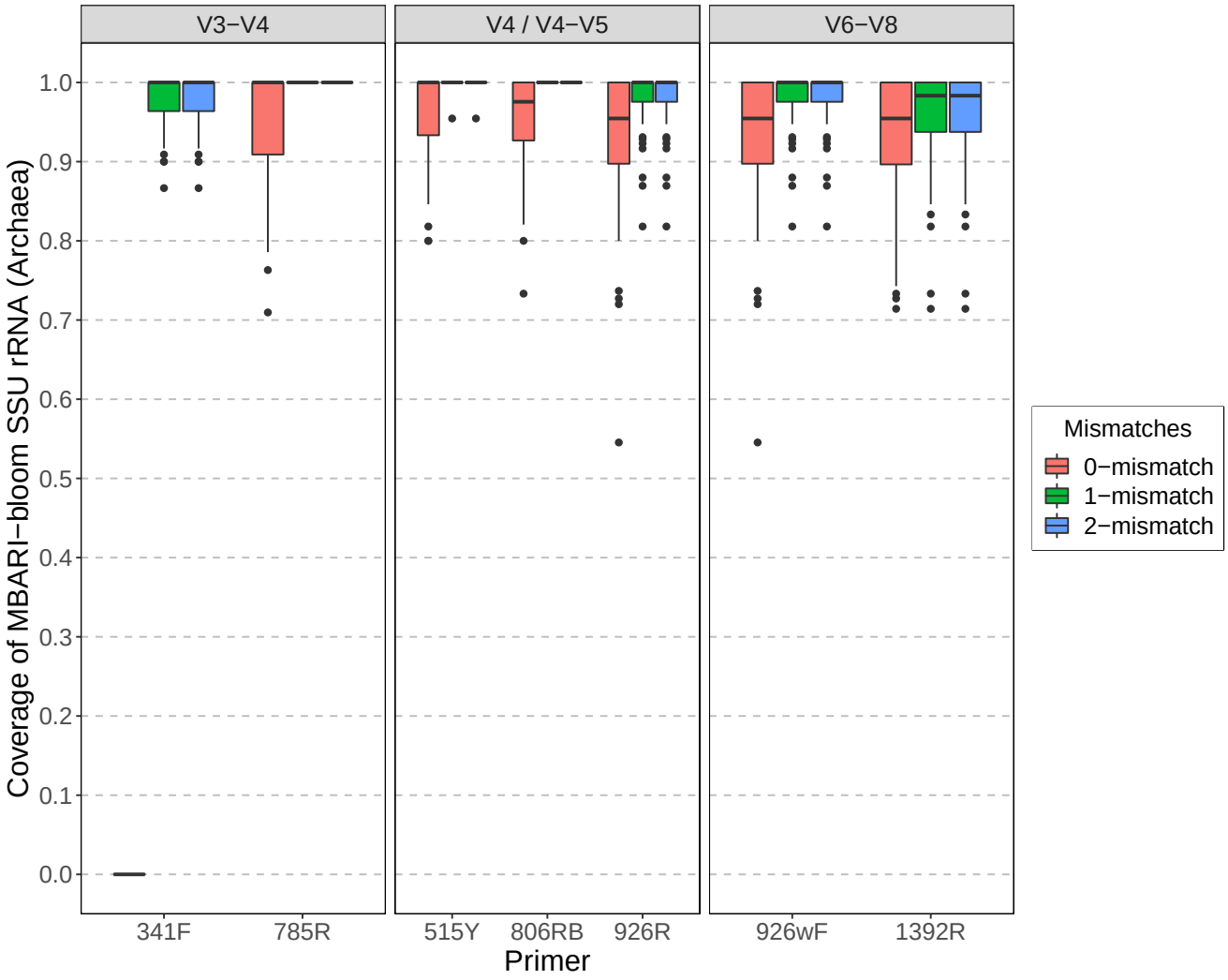

Fig S6: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Archaea* using metagenomes derived from MBARI bloom water samples.

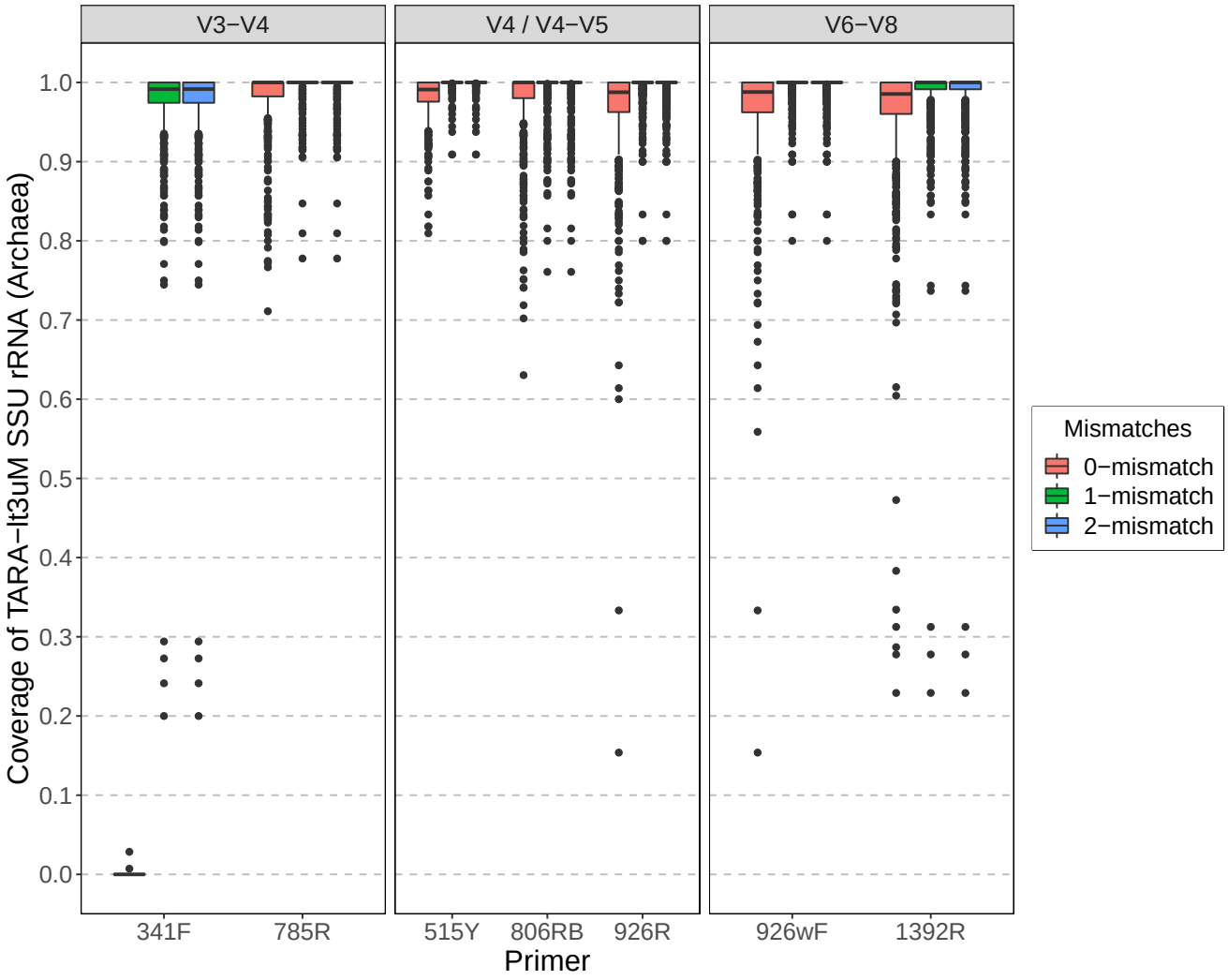

Fig S7: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Archaea* using metagenomes derived from Tara Oceans water samples (<3µm).

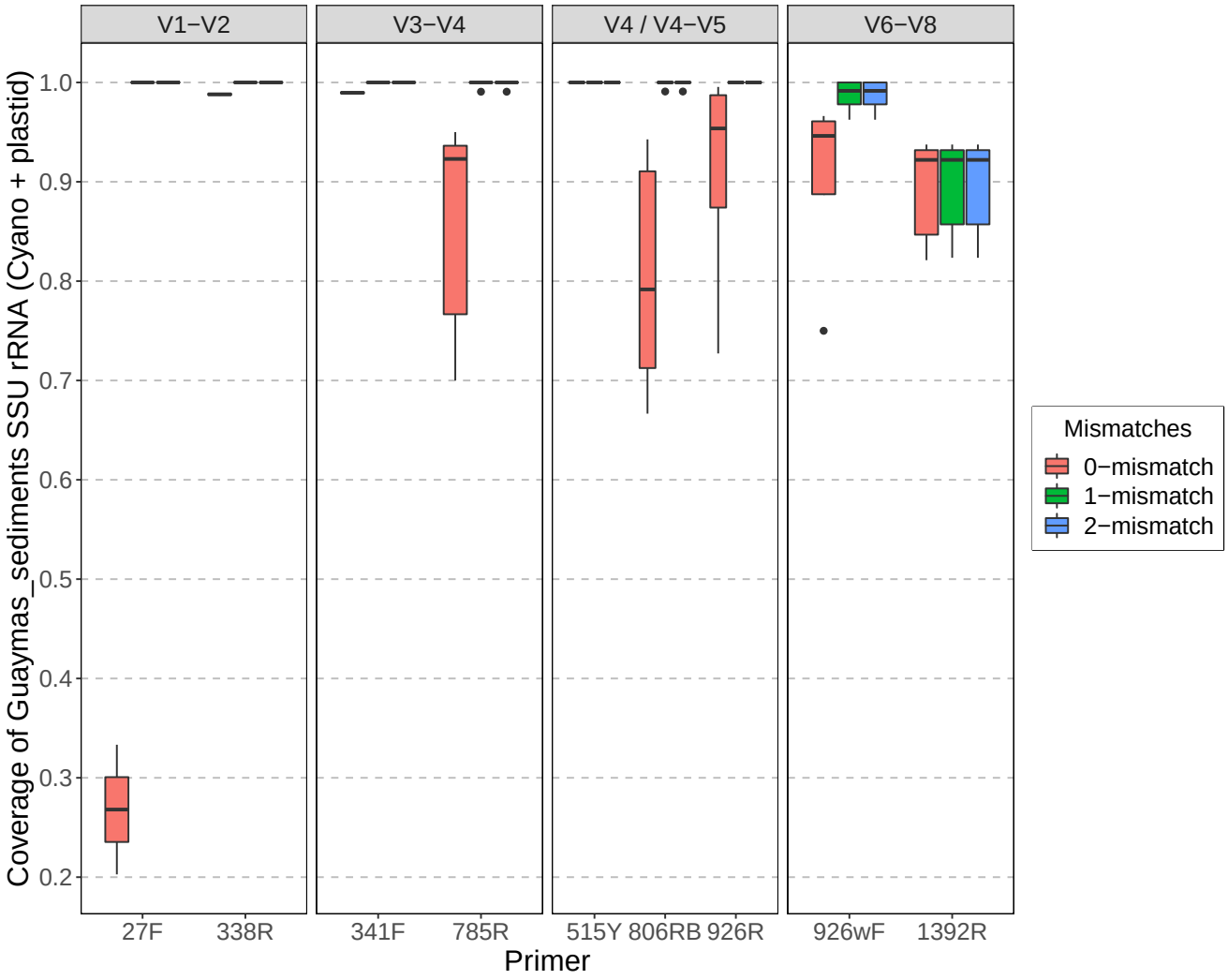

Fig S8: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Cyanobacteria* and plastids using metagenomes derived from Guaymas Basin sediments.

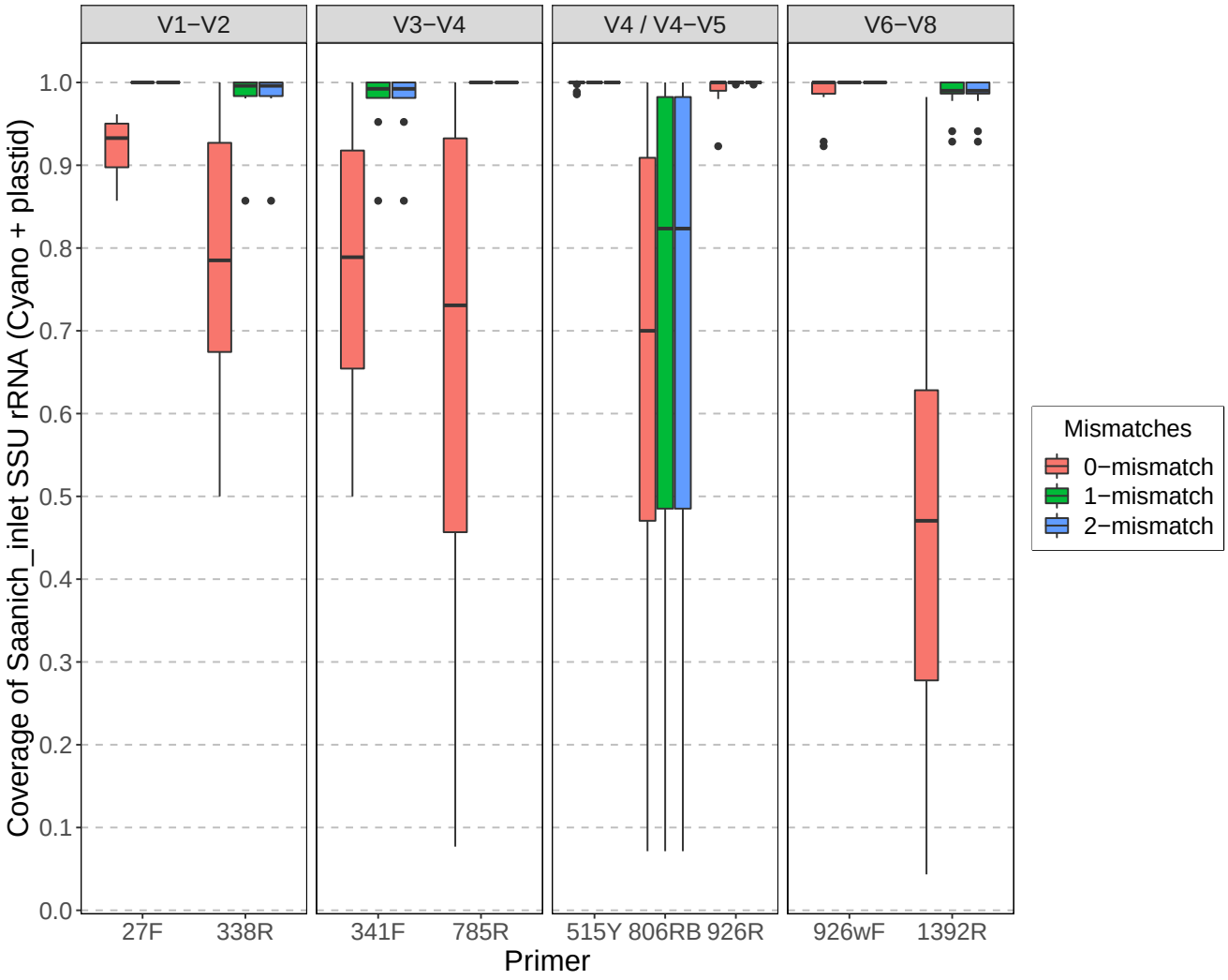

Fig S9: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Cyanobacteria* and plastids using metagenomes derived from Saanich Inlet water samples.

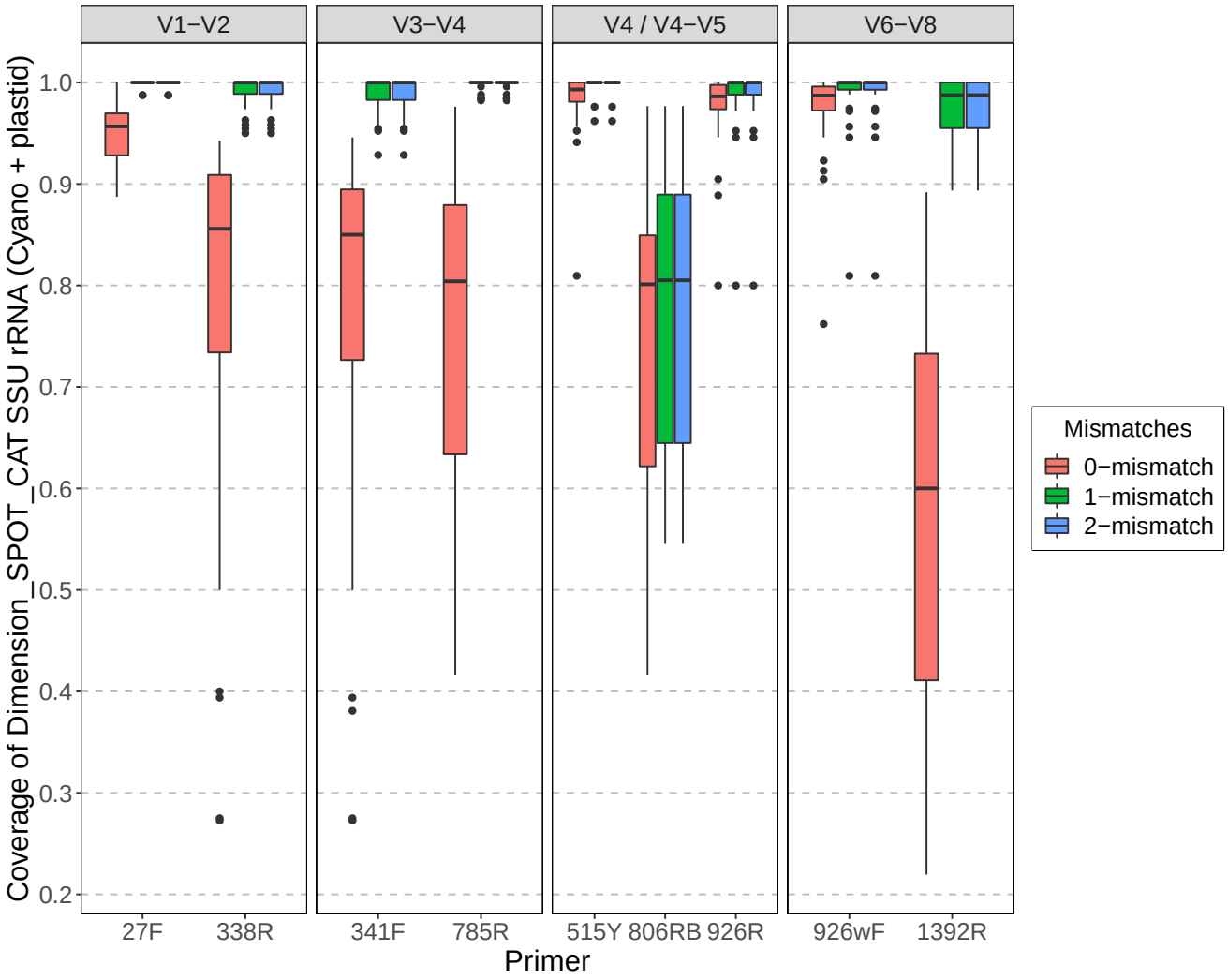

Fig S10: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Cyanobacteria* and plastids using metagenomes derived from SPOT/Catalina water samples.

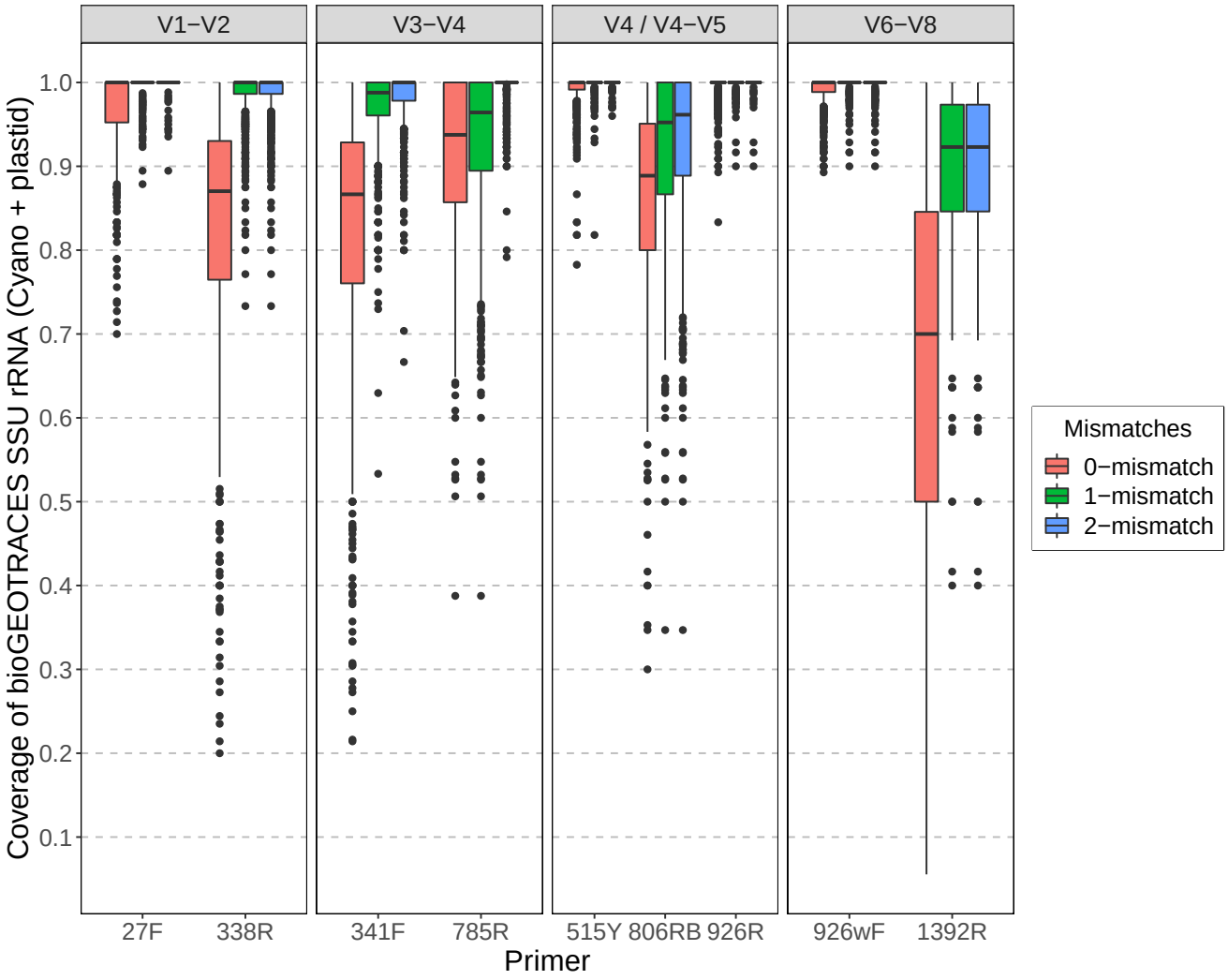

Fig S11: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Cyanobacteria* and plastids using metagenomes derived from BioGEOTRACES/BATS/HOT water samples.

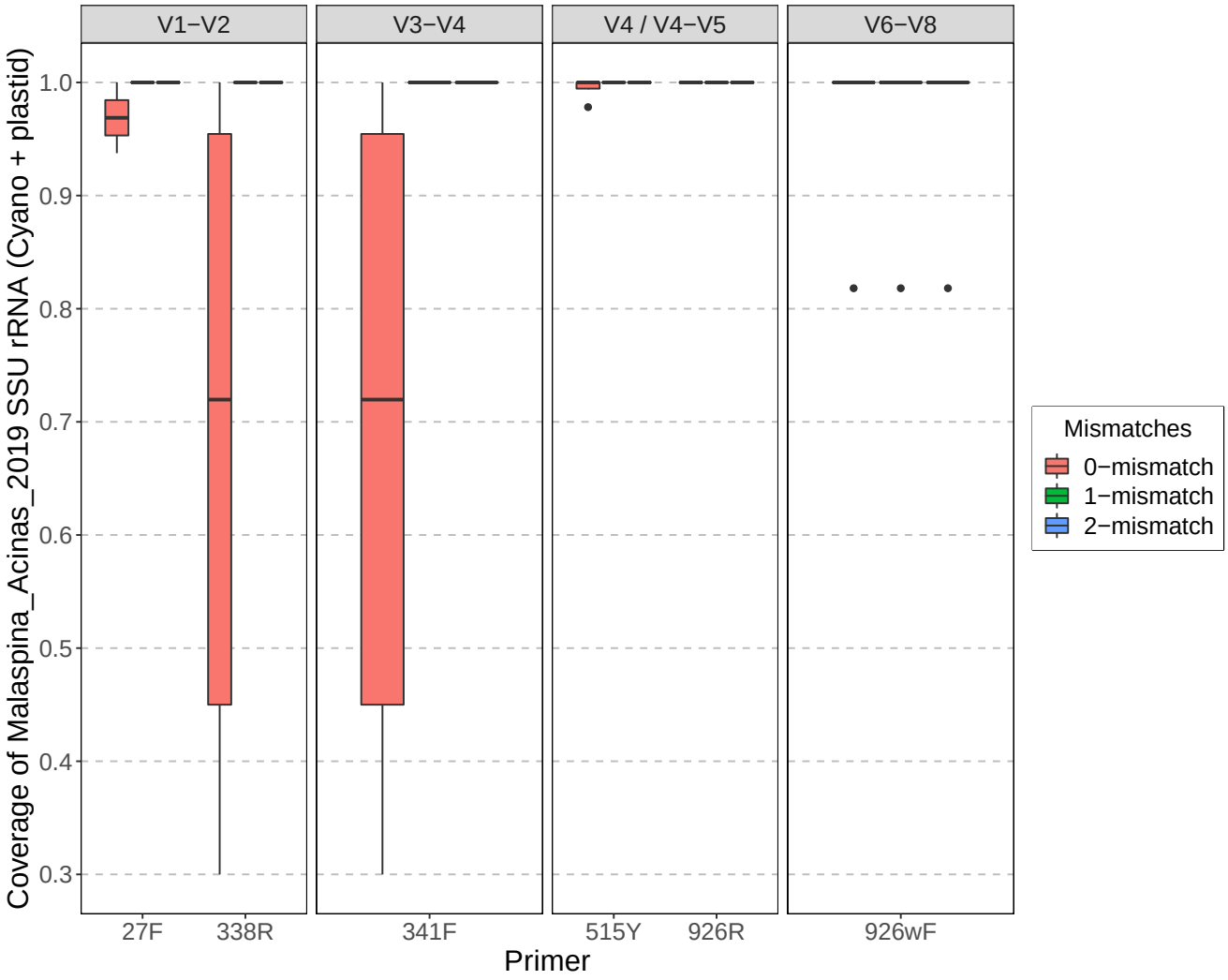

Fig S12: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Cyanobacteria* and plastids using metagenomes derived from Malaspina deep-sea water samples.

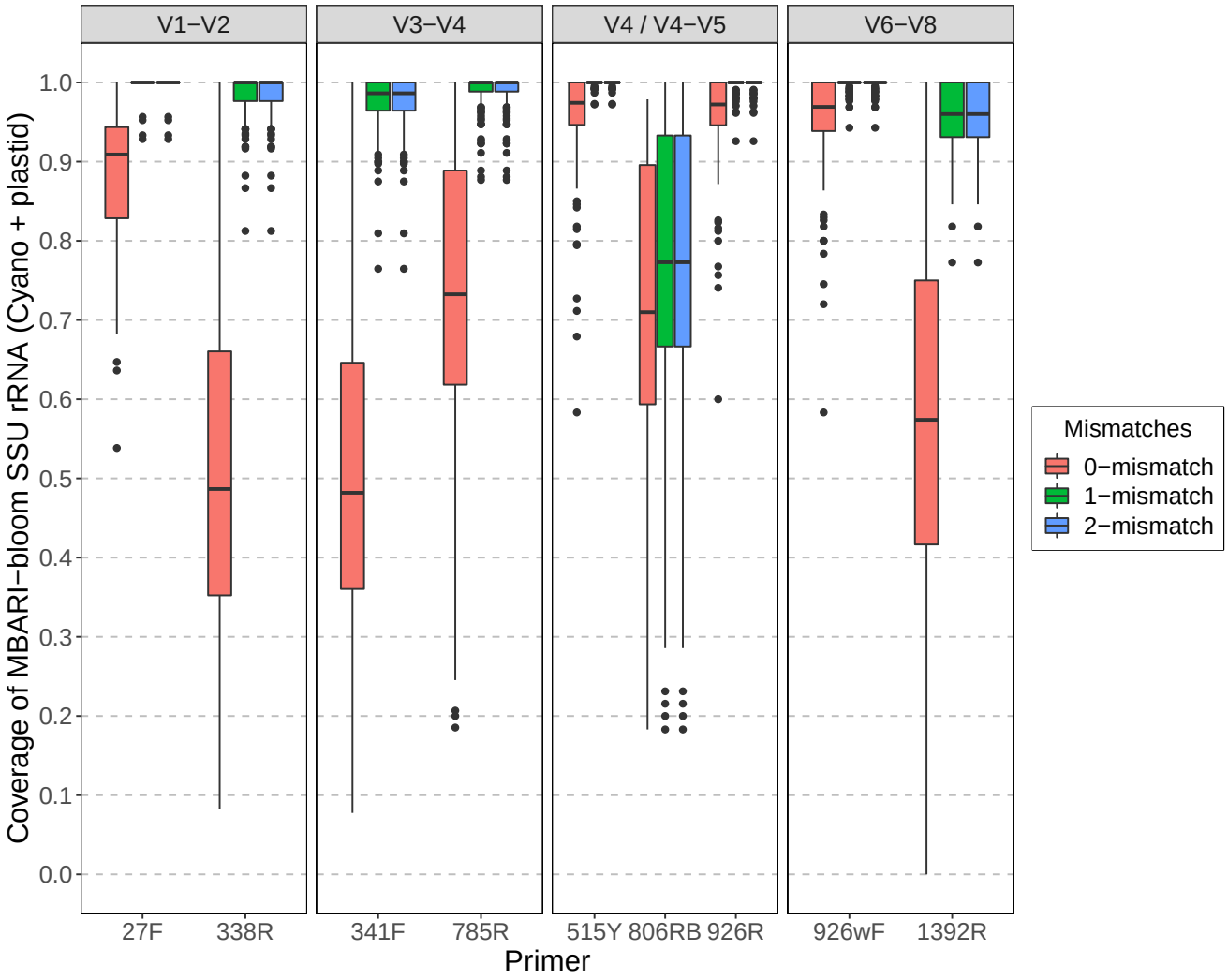

Fig S13: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Cyanobacteria* and plastids using metagenomes derived from MBARI bloom water samples.

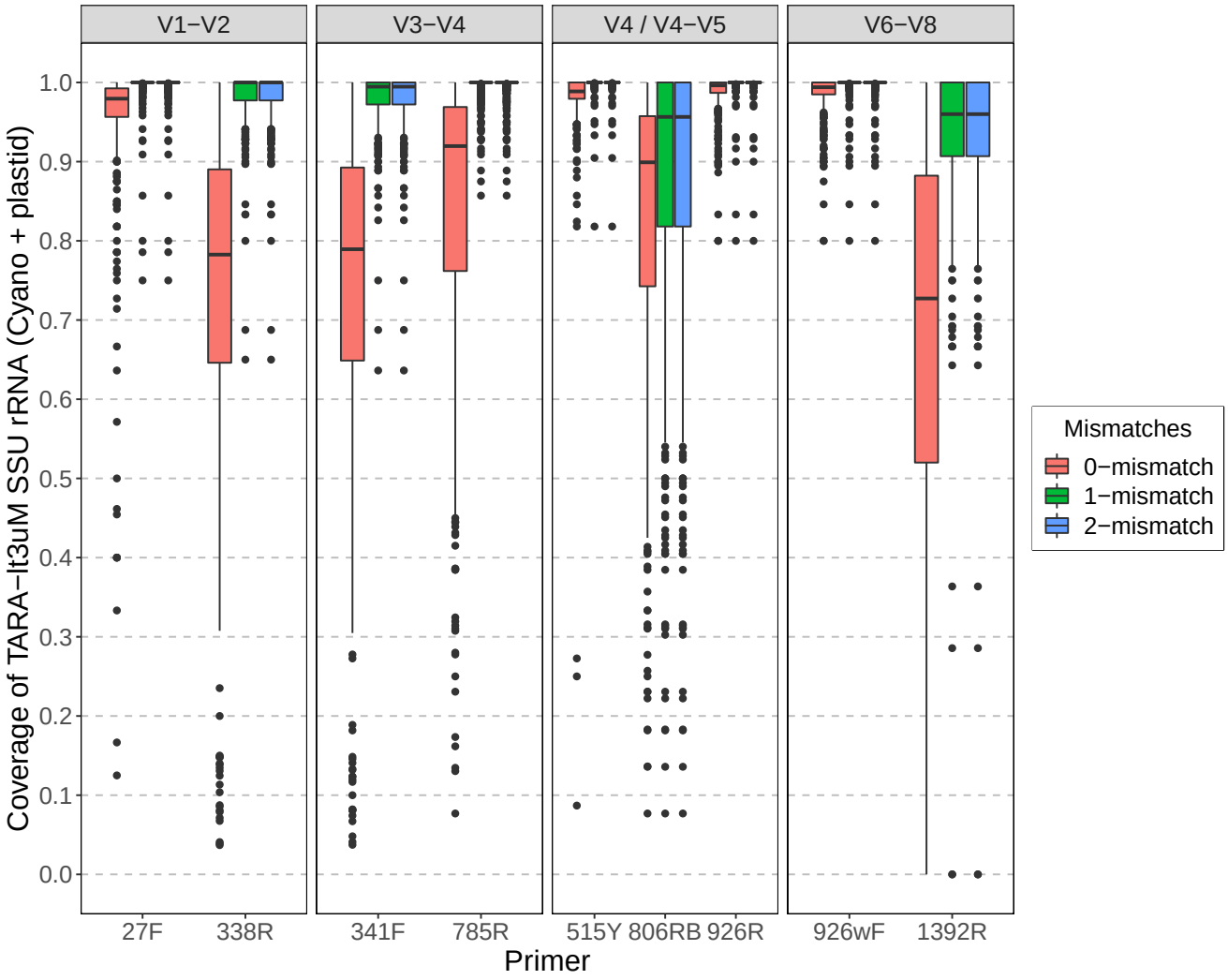

Fig S14: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Cyanobacteria* and plastids using metagenomes derived from Tara Oceans water samples (<3µm).

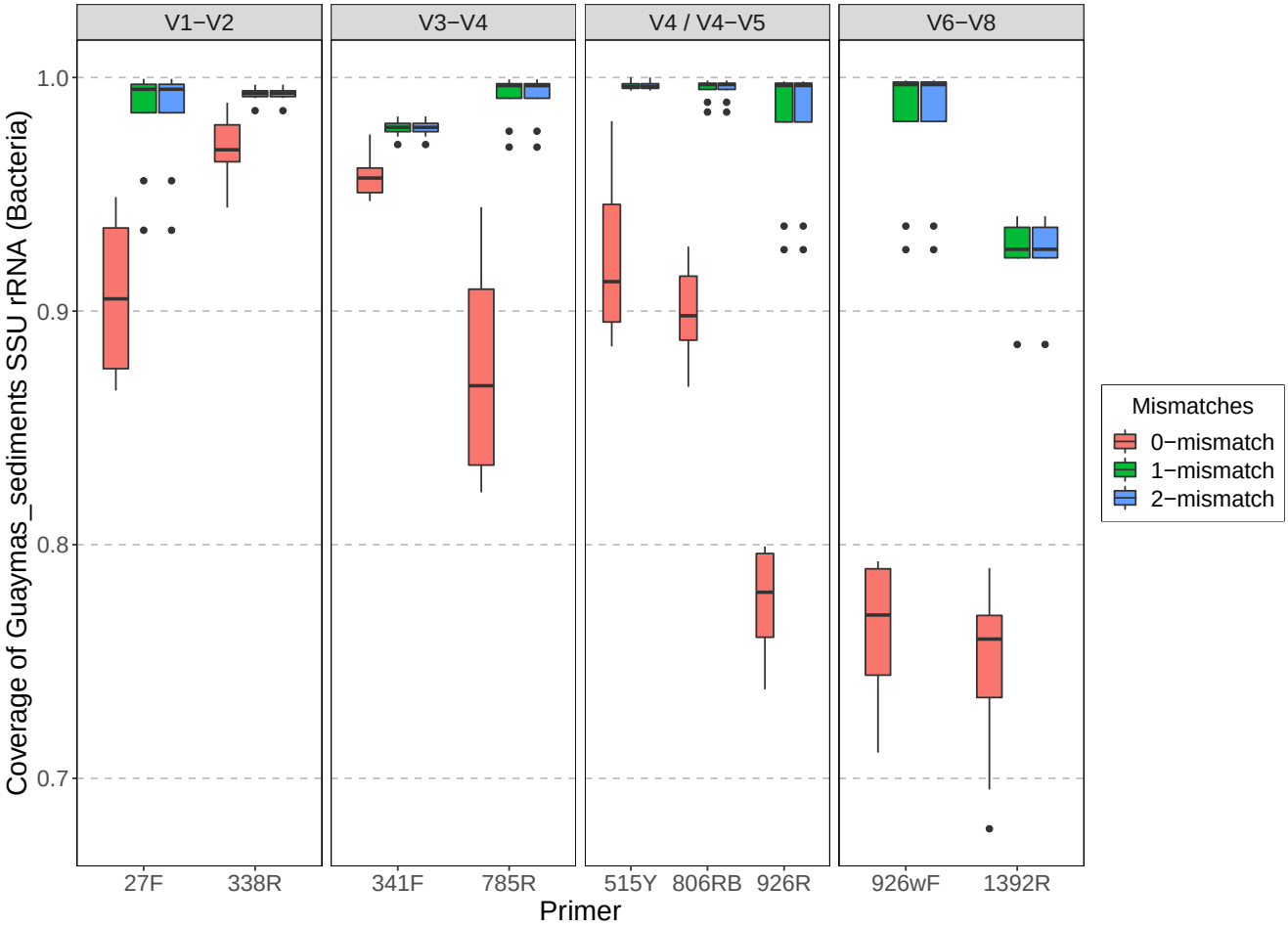

Fig S15: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Bacteria* (not including *Cyanobacteria* and plastids) using metagenomes derived from Guaymas Basin sediment samples.

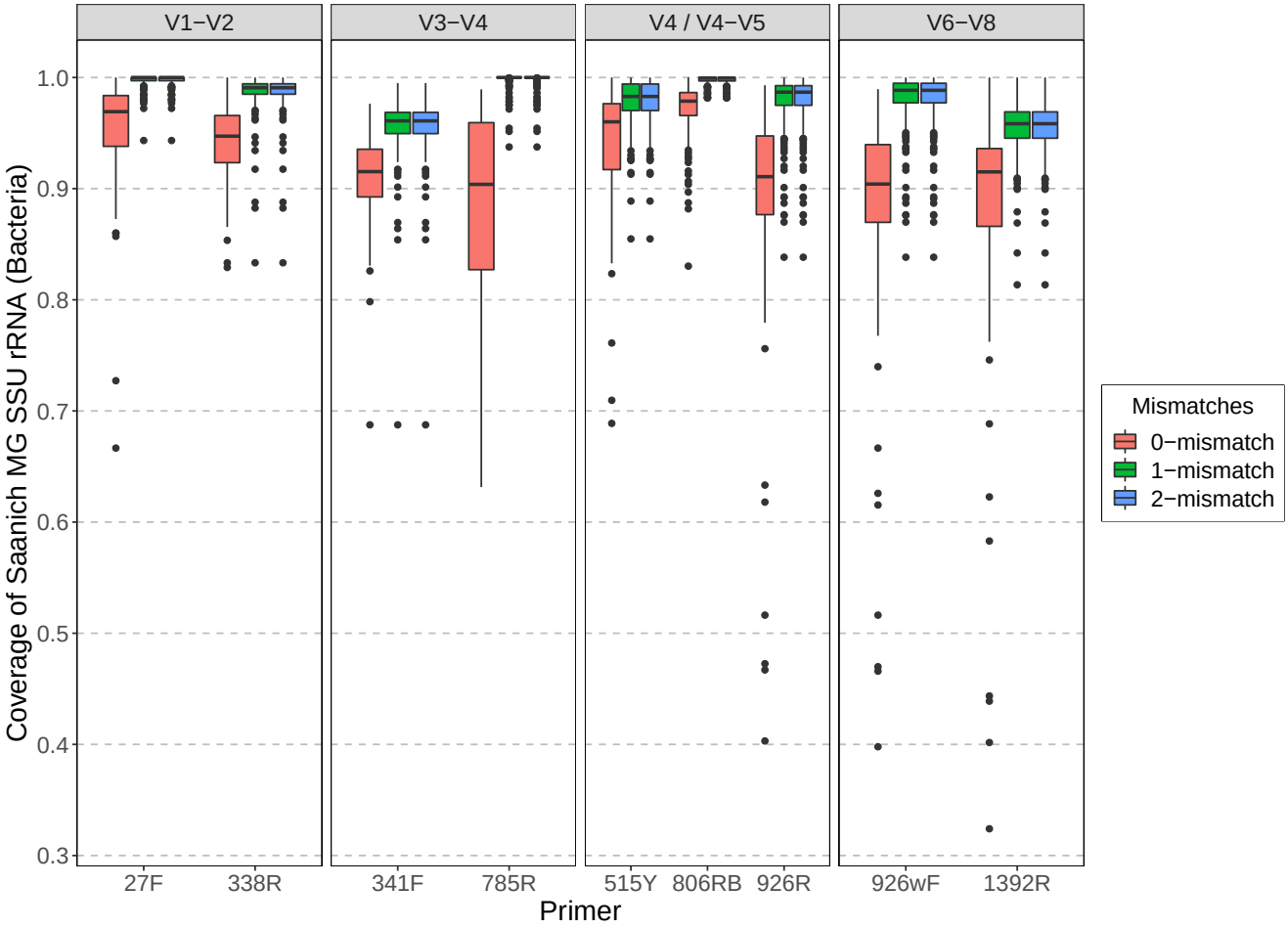

Fig S16: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Bacteria* (not including *Cyanobacteria* and plastids) using metagenomes derived from Saanich Inlet water samples.

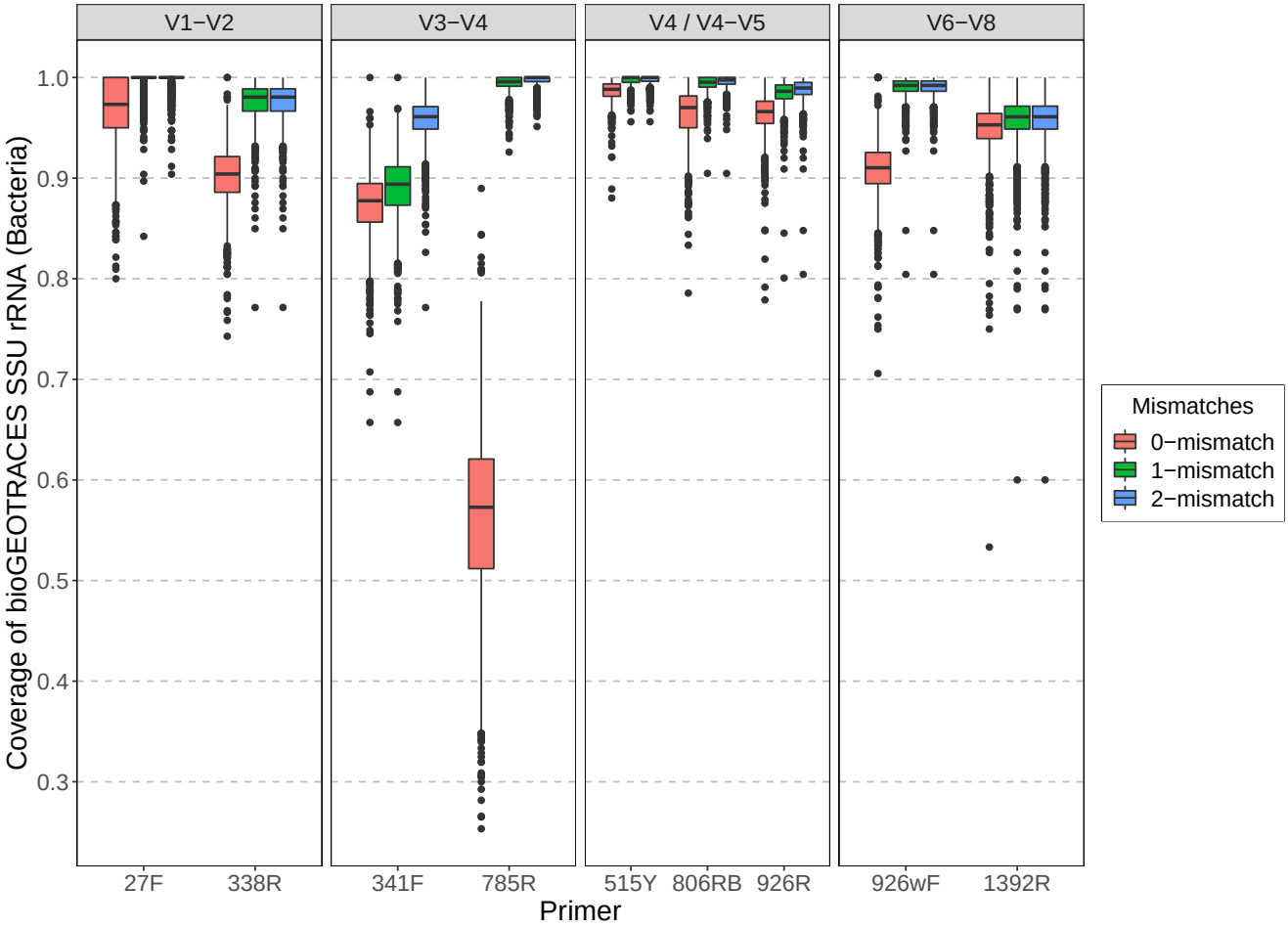

Fig S17: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Bacteria* (not including *Cyanobacteria* and plastids) using metagenomes derived from BioGEOTRACES/BATS/HOT water samples.

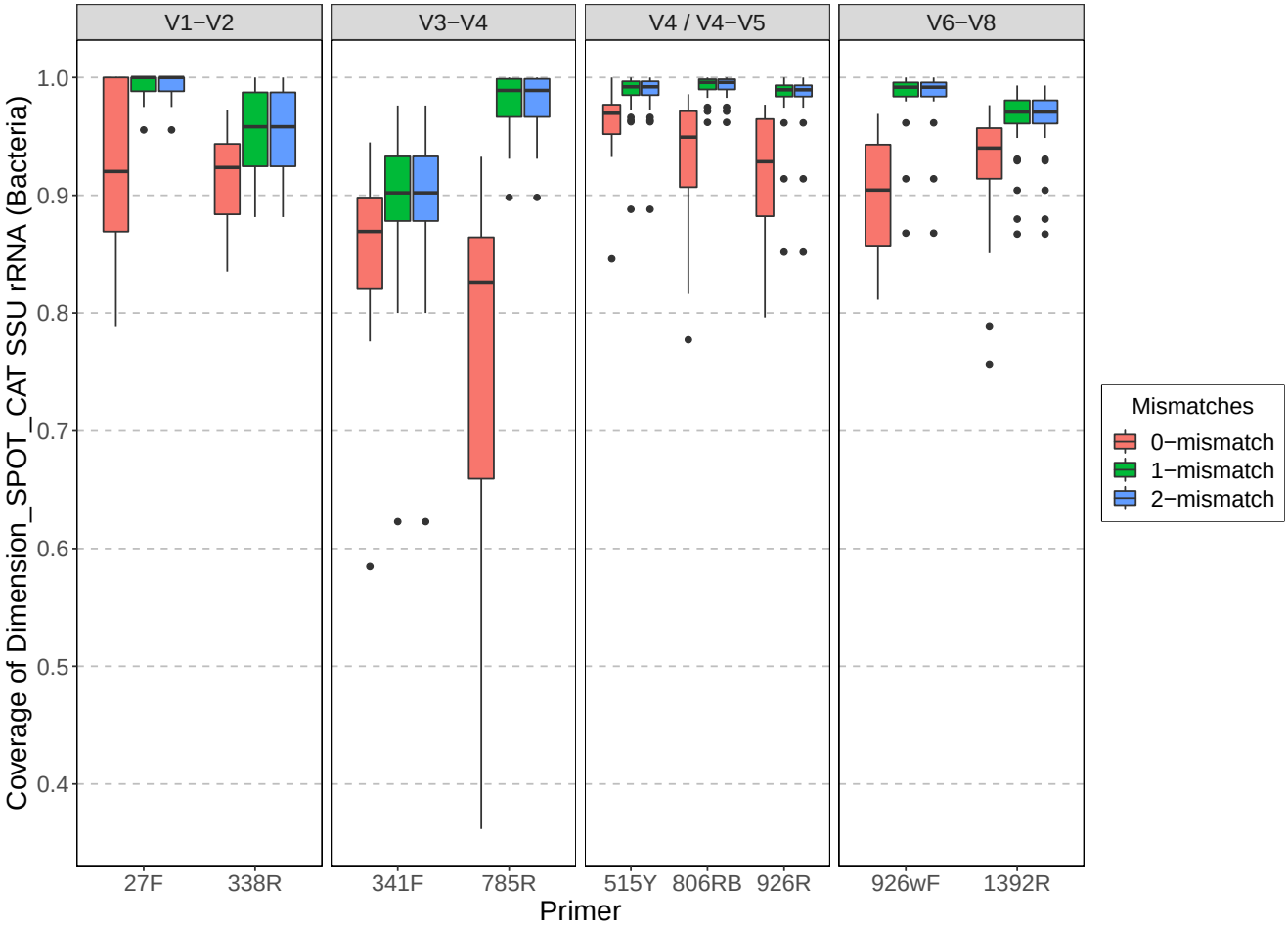

Fig S18: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Bacteria* (not including *Cyanobacteria* and plastids) using metagenomes derived from SPOT/Catalina water samples.

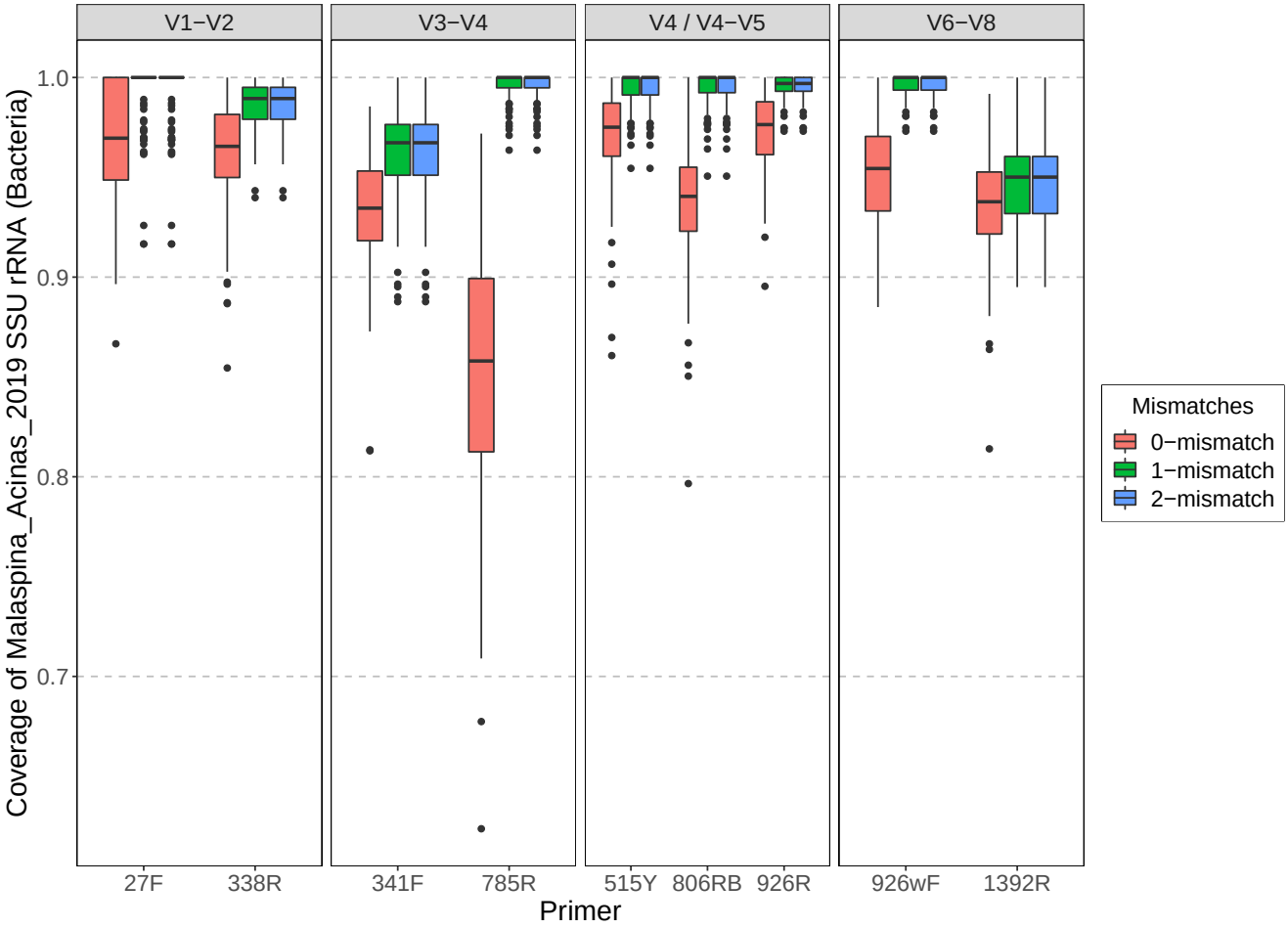

Fig S19: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Bacteria* (not including *Cyanobacteria* and plastids) using metagenomes derived from Malaspina deep-sea water samples.

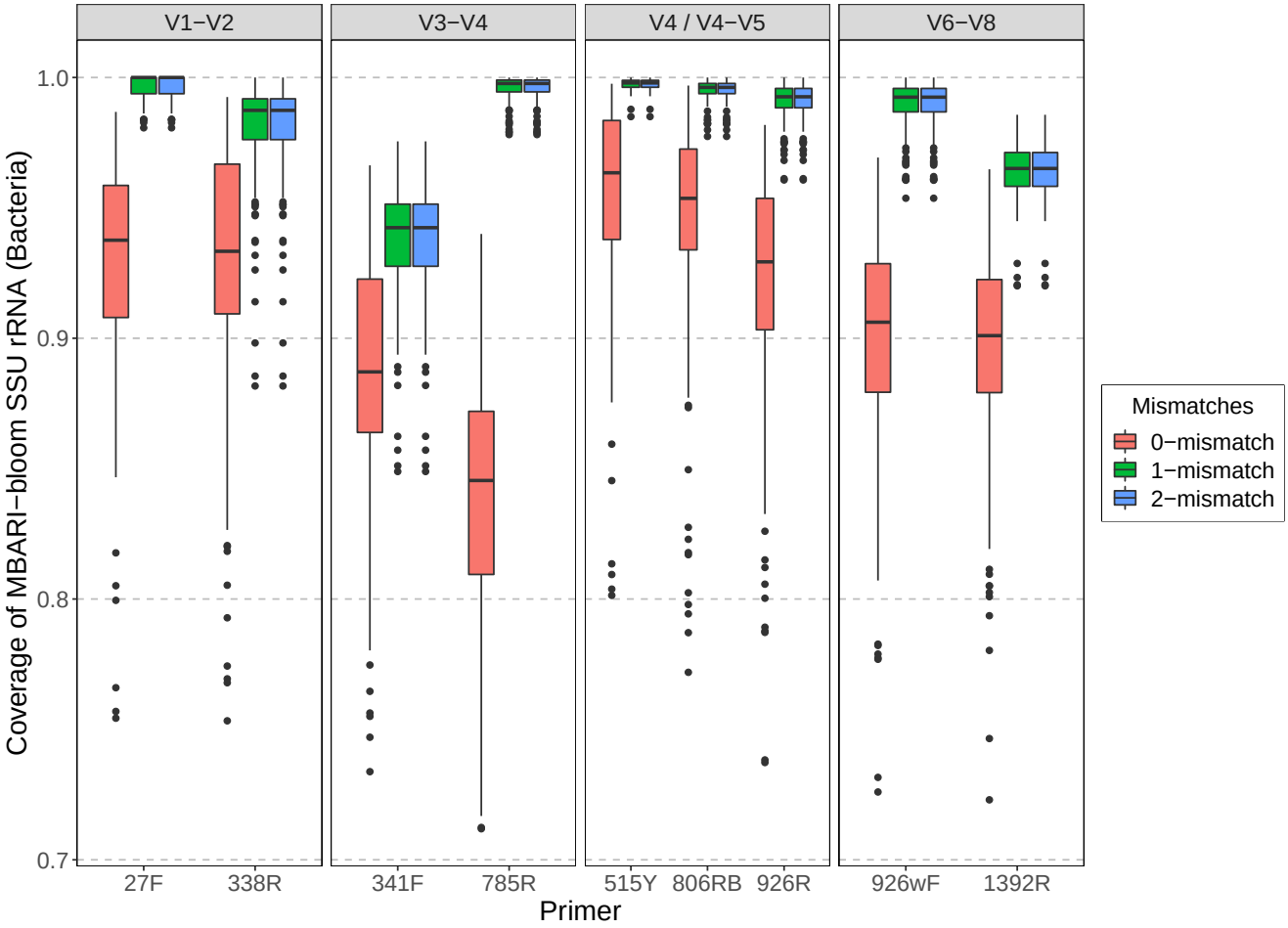

Fig S20: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Bacteria* (not including *Cyanobacteria* and plastids) using metagenomes derived from MBARI bloom water samples.

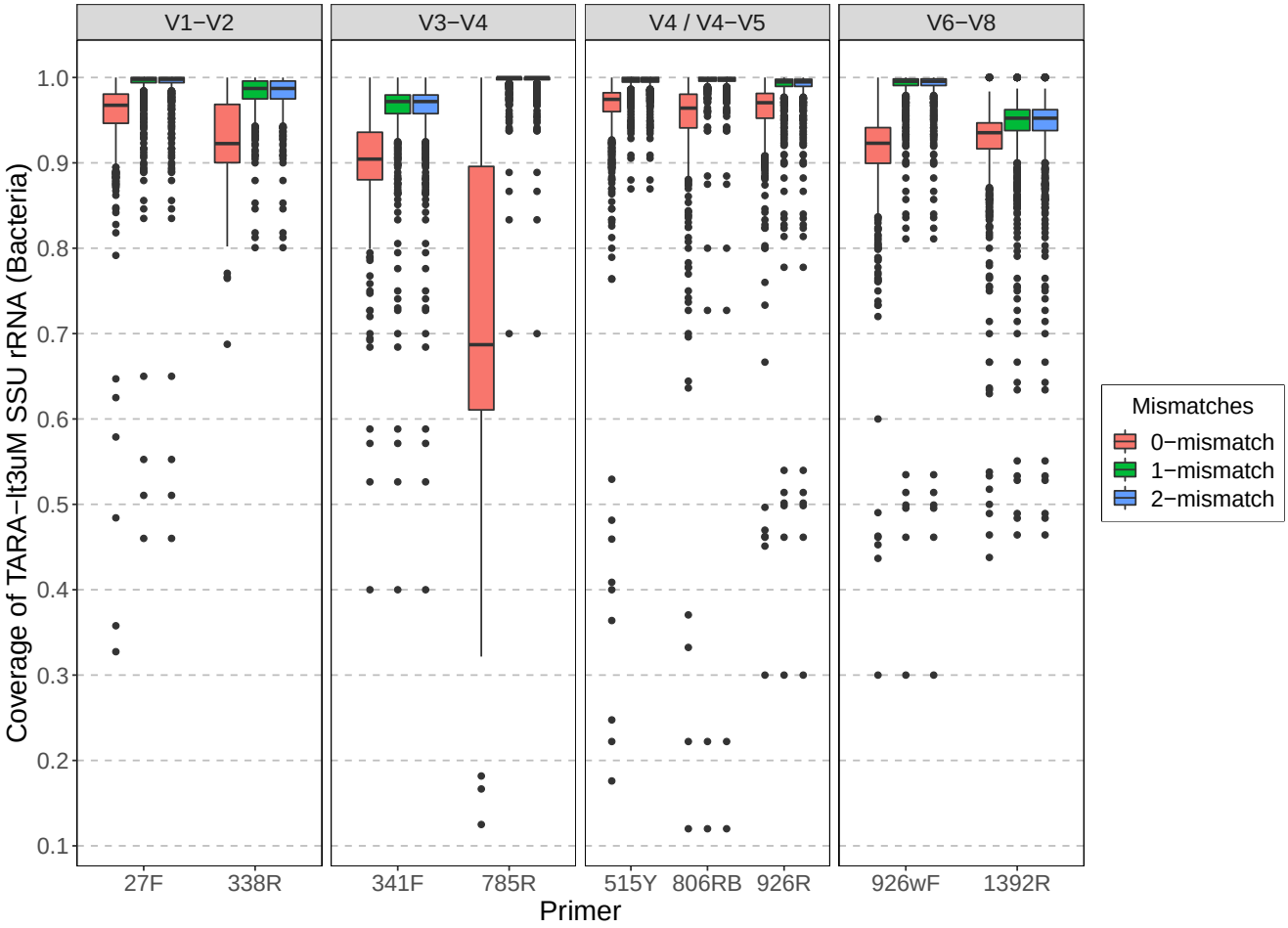

Fig S21: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Bacteria* (not including *Cyanobacteria* and plastids) using metagenomes derived from Tara Oceans water samples (<3µm).

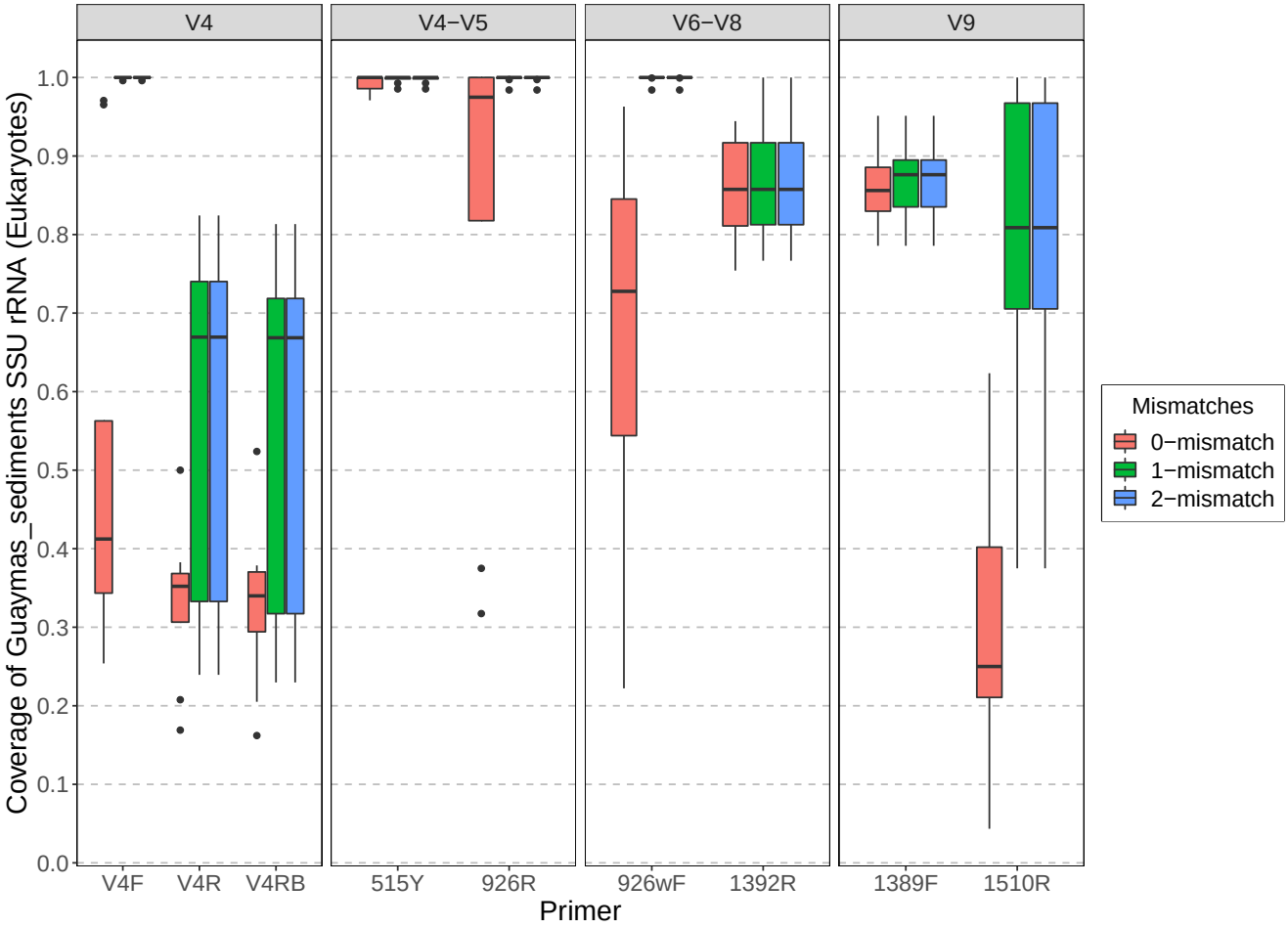

Fig S22: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Eukarya* using metagenomes derived from Guaymas Basin sediment samples.

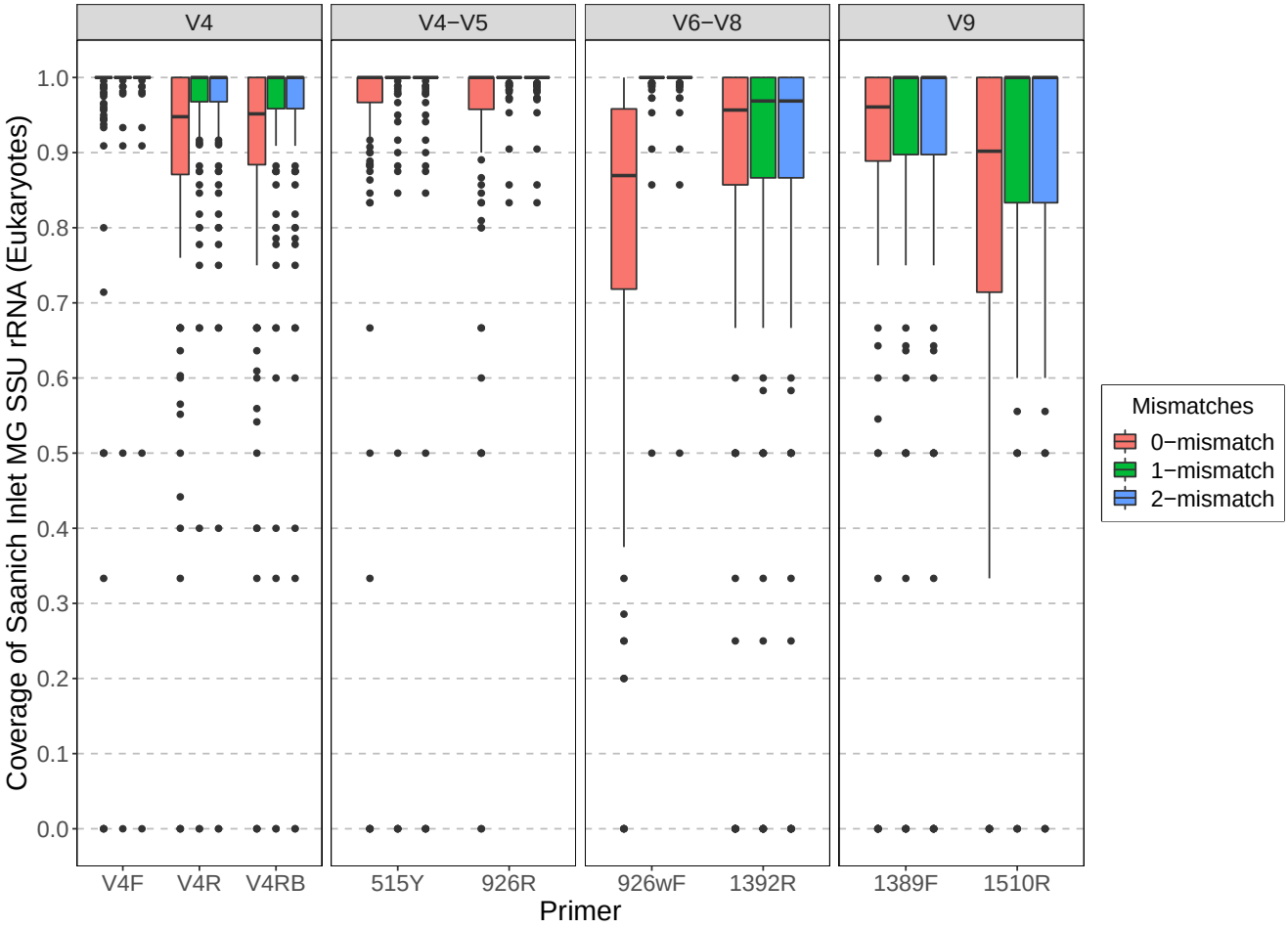

Fig S23: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Eukarya* using metagenomes derived from Saanich Inlet water samples.

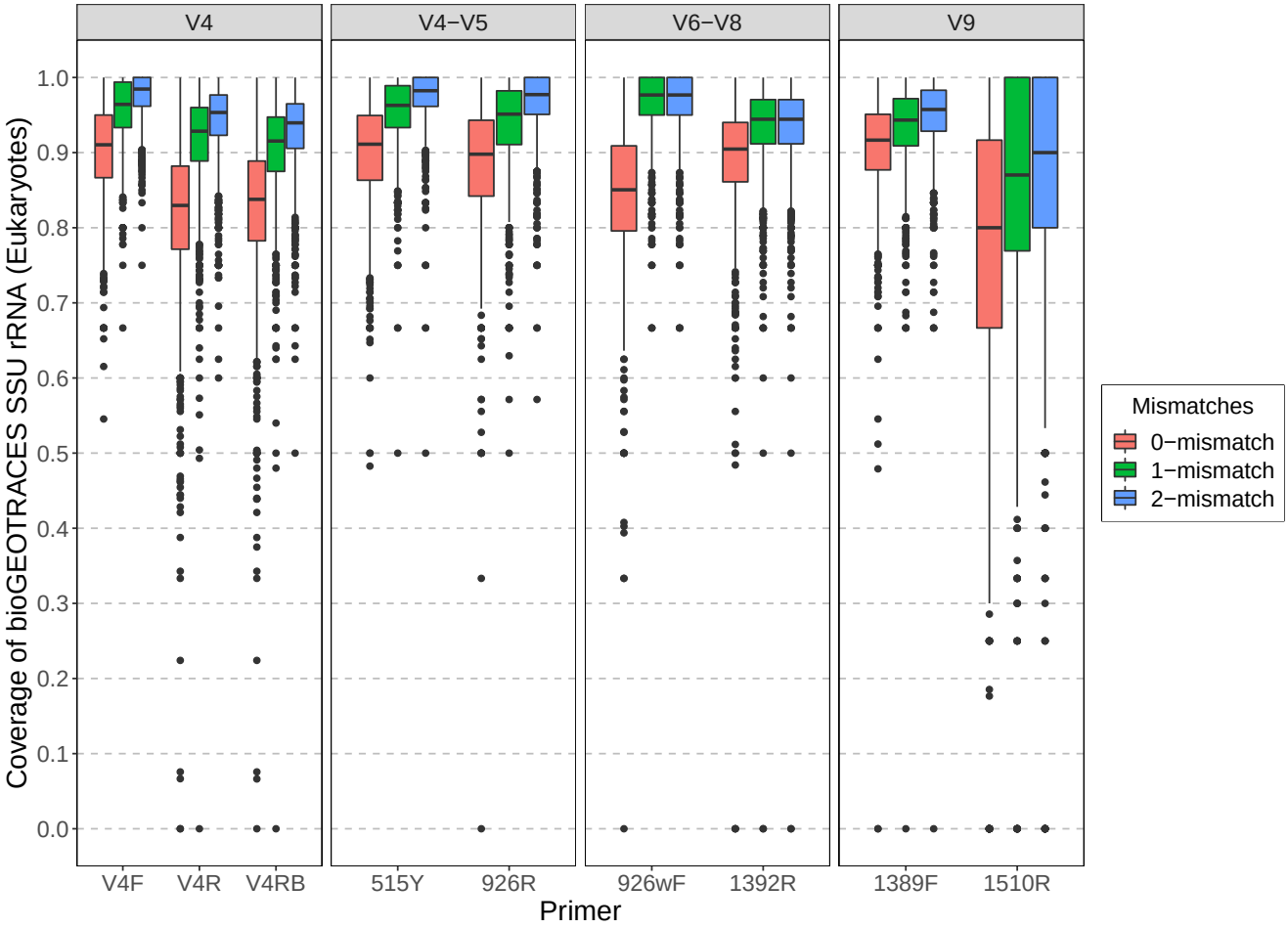

Fig S24: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Eukarya* using metagenomes derived from BioGEOTRACES/BATS/HOT water samples.

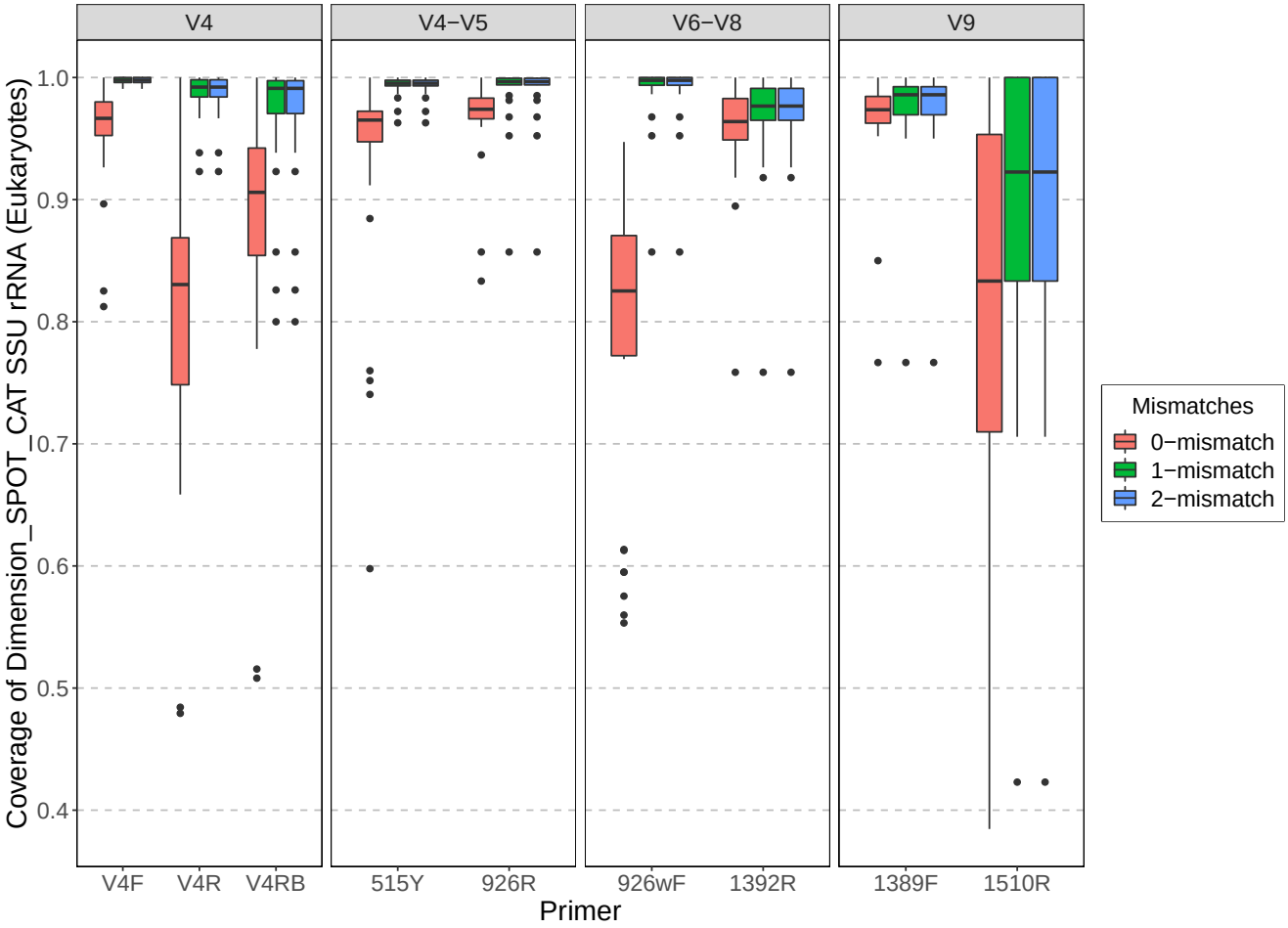

Fig S25: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Eukarya* using metagenomes derived from SPOT/Catalina water samples.

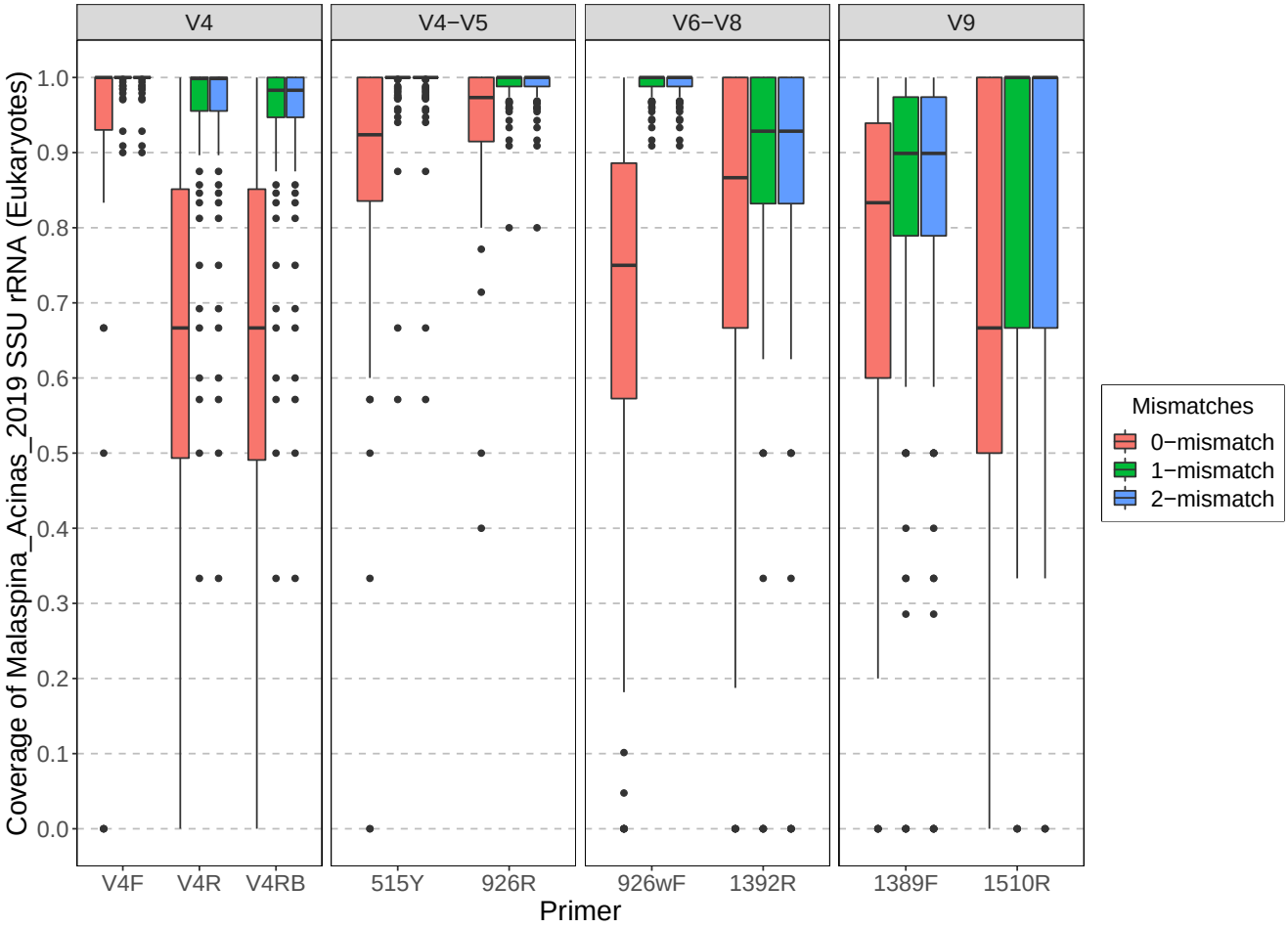

Fig S26: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Eukarya* using metagenomes derived from Malaspina deep-sea water samples.

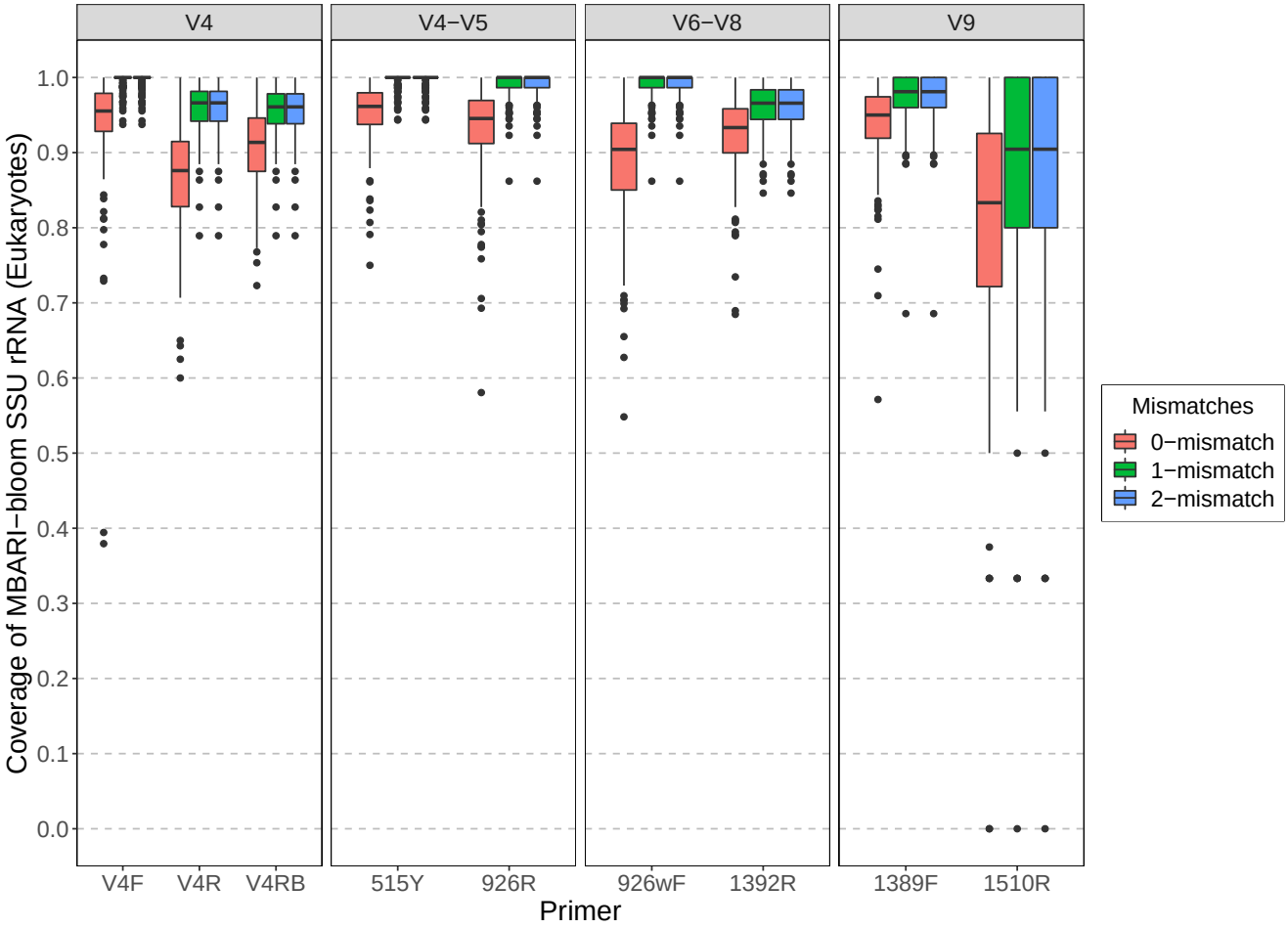

Fig S27: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Eukarya* using metagenomes derived from MBARI bloom water samples.

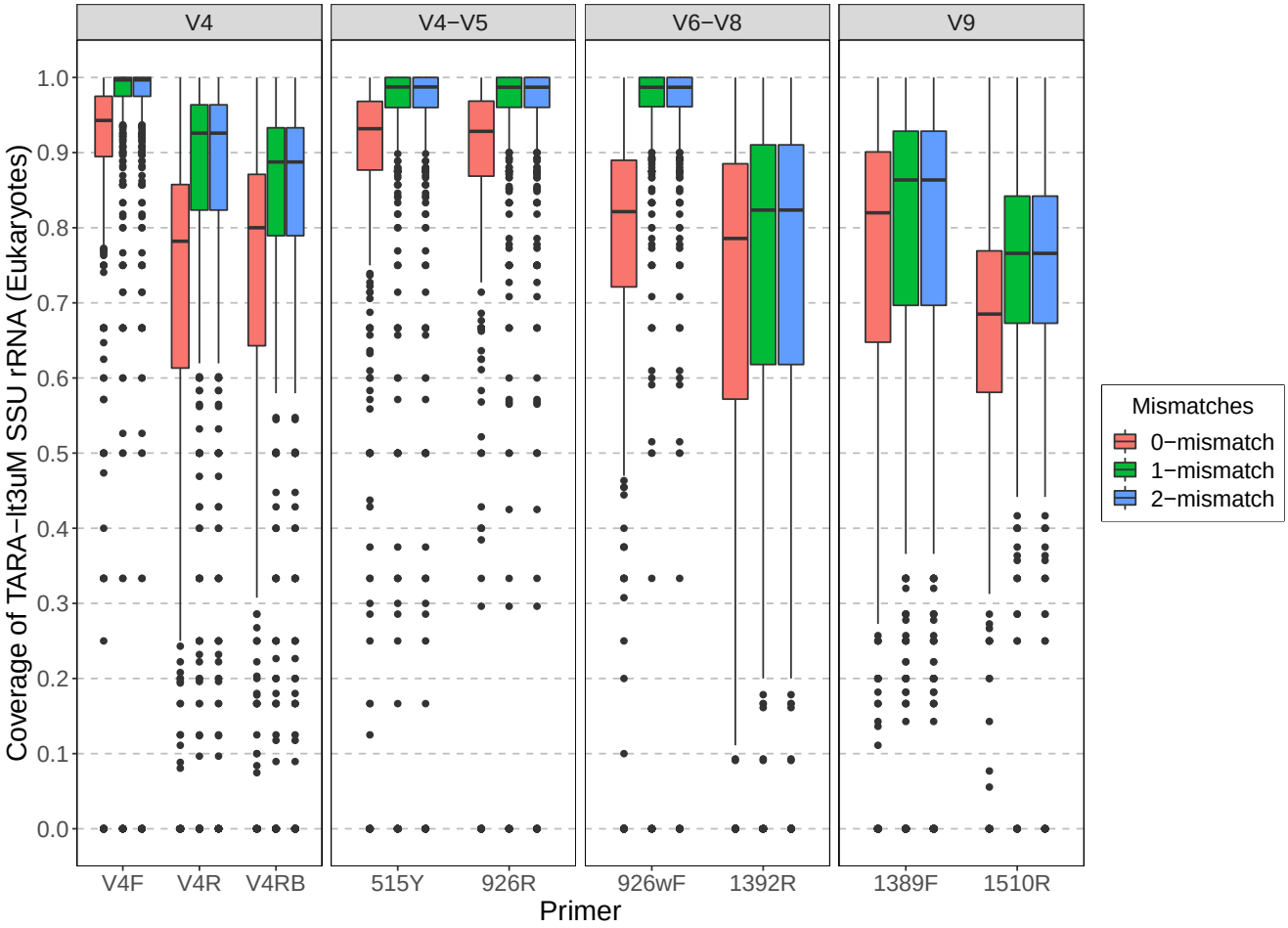

Fig S28: Predicted individual primer coverage allowing 0, 1, and 2-mismatches for *Eukarya* using metagenomes derived from Tara Oceans water samples (<3µm).

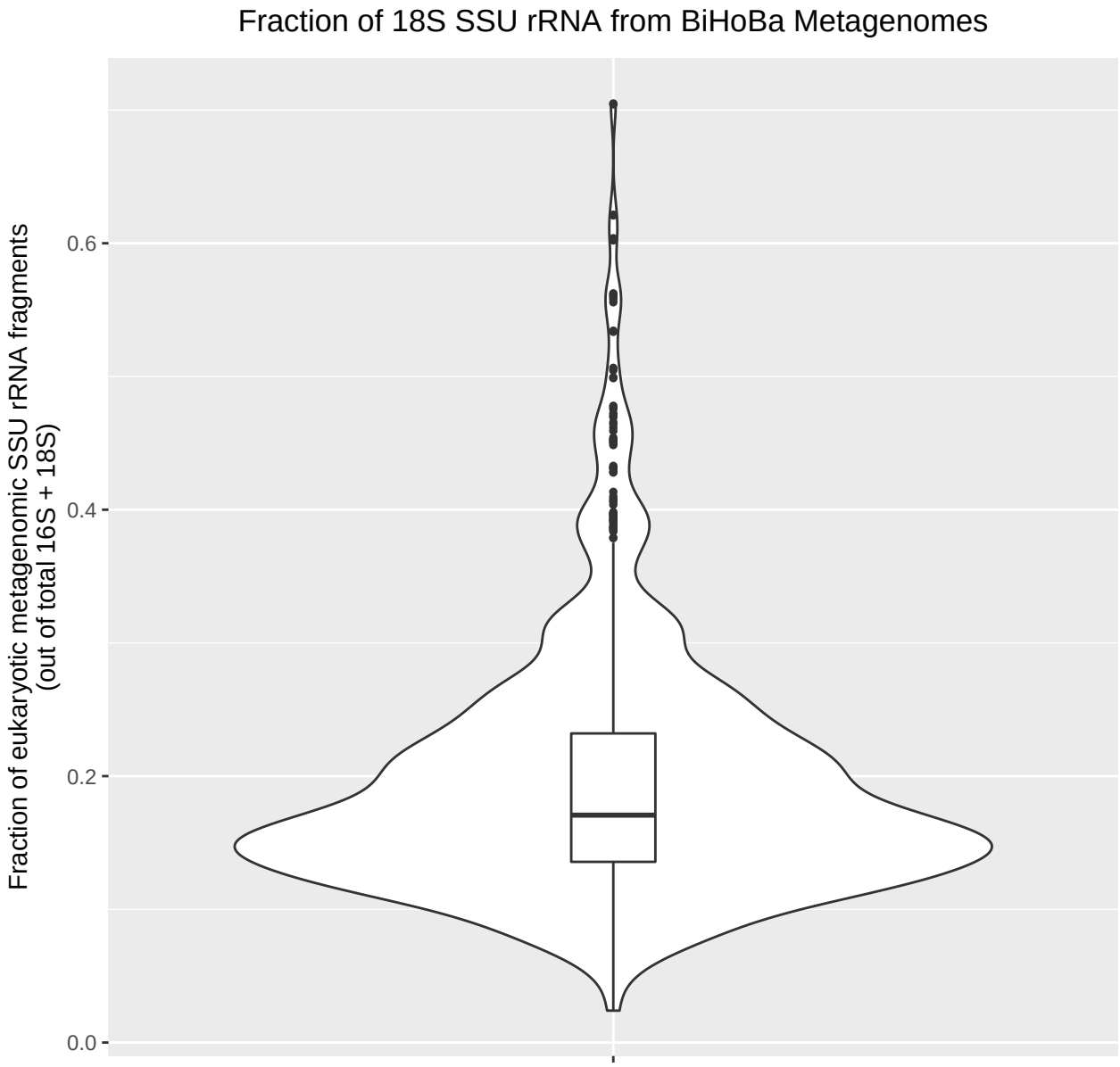

Figure S29: Violin plot with inlaid boxplot showing the fraction of eukaryotic (18S) SSU rRNA as a fraction of total across BATS, HOT, and bioGEOTRACES metagenomic datasets. The violin plot provides a visual description of the distribution of data - i.e. in this case showing that the vast majority of samples had < 30% of SSU rRNA reads accounted for by 18S sequences.

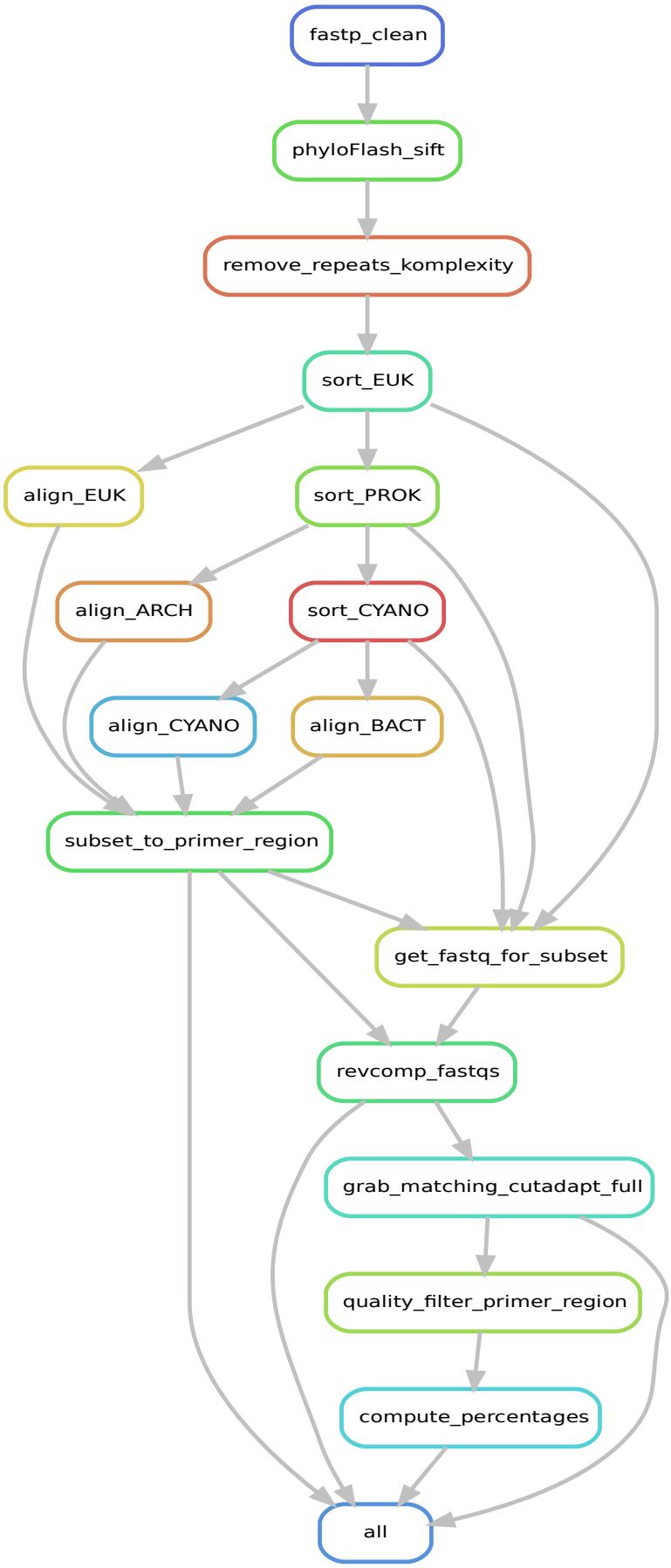

Figure S30: **Rulegraph for "Snakefile-compute.smk" portion of workflow**. Briefly, SSU rRNA is identified using phyloFlash from fastp quality-filtered fastq reads, which are then sorted with bbsplit.sh into 4 categories, and aligned to SSU rRNA references with pyNAST after removing low-complexity sequences. For each group and primer, matches to primer regions are then identified using custom python scripts and cutadapt (at 0, 1, 2, and 6-mismatches), quality filtered to omit sequences where the primer region plus 5 leading and trailing bases has one individual base with a phred score below 30. Matches are then summarized in plain-text output and ggplot2 graphs.

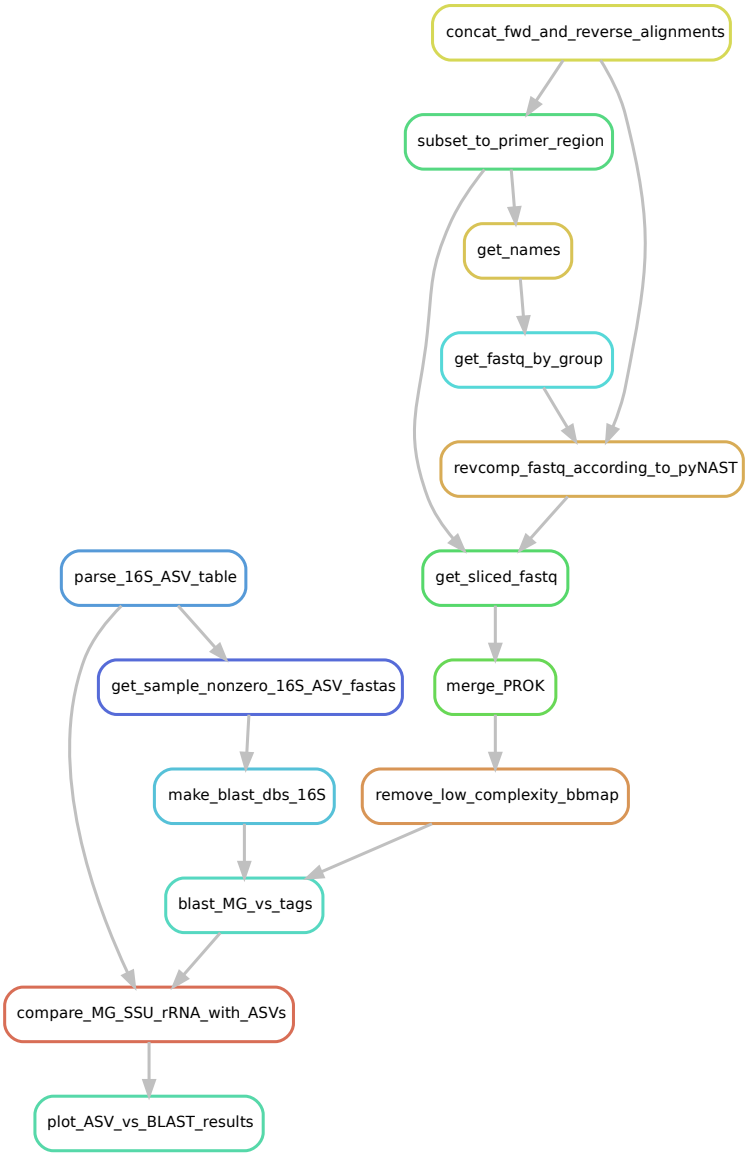

Figure S31: **Rulegraph for "Snakefile-compare.smk" portion of workflow to intercompare metagenome and amplicon data**.