Catlett Dylan (Orcid ID: 0000-0002-9431-4101)
Matson Paul (Orcid ID: 0000-0003-2105-7308)

Evaluation of Accuracy and Precision in an Amplicon Sequencing Workflow for Marine Protist Communities

Catlett D [1,5*], Matson PG [2,5,6], Carlson CA [2,3], Wilbanks EG [2,3], Siegel, DA [1,4], Iglesias-Rodriguez MD [2,3]

[1] Earth Research Institute, UC Santa Barbara, CA 93106

[2] Marine Science Institute, UC Santa Barbara, CA 93106

[3] Department of Ecology, Evolution, and Marine Biology, UC Santa Barbara, CA 93106

[4] Department of Geography, UC Santa Barbara, CA 93106

[5] These authors contributed equally to this work

[6] Current address: Department of Biological Sciences, Bowling Green State University, OH 43402

[*]Corresponding author:

Dylan Catlett

Earth Research Institute, UC Santa Barbara, CA 93106

Email: dsc@ucsb.edu

Running head: marine protist amplicon method evaluation

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/lom3.10343

Keywords: amplicon sequencing; marine protist; 18S rDNA; DNA metabarcoding; protist diversity

Abstract

Advances in high-throughput DNA sequencing methods reveal the vast diversity of marine protists. Amplicon sequencing of "barcode" genes, such as the 18S small subunit ribosomal RNA gene (henceforth, 18S gene), is a cost-effective and widely used genetic method for assessing the composition of marine protist communities. This method is now being applied from local to global scales to interrogate the causes and consequences of protist community variations. Significant efforts have been made to validate amplicon methods targeting prokaryotes, but the precision, accuracy, and quantitative potential of 18S gene amplicon sequencing methods for marine protists remain unclear. Here, we use artificial (mock) communities and environmental samples collected from the Santa Barbara Channel, CA, to evaluate the precision and accuracy in an amplicon workflow targeting the V9 hypervariable region of the 18S gene for marine protists. Overall, we find that this amplicon workflow has high precision and reasonable accuracy, but the magnitude of analytical uncertainty can increase significantly unless certain procedural issues are avoided. Finally, we demonstrate the value of positive and negative controls in, and the quantitative potential of, amplicon sequencing assessments of marine protist communities.

Introduction

Marine protists (defined here as unicellular eukaryotes excluding fungi and macroalgae) play a critical role in marine ecosystems and global biogeochemical cycles, with marine phytoplankton accounting for almost half of the 105 Pt C yr^{-1} global net primary production (Field et al. 1998) and protist grazers modulating the flow of primary and secondary production through the pelagic food web (Sherr and Sherr 1994; Worden et al. 2015). Recent developments in high-throughput nucleic acid sequencing (HTS) offer unprecedented resolution of the vast diversity of marine protist communities (Amaral-Zettler et al. 2009; De Vargas et al. 2015) and in turn, an improved understanding of the impact of protist diversity and community composition on the structure and function of marine ecosystems (Lima-Mendez et al. 2015; Guidi et al. 2016; Caputi et al. 2018; Wang et al. 2018). The size and complexity of marine protist genomes often precludes metagenomic analysis of environmental samples in which these organisms are abundant (Keeling and del Campo 2017). However, amplicon sequencing of widely conserved "barcode" genes (e.g., the 18S gene in protists) from an environmental DNA sample is a powerful and widely used tool for characterizing protist diversity and community composition.

Standard amplicon sequencing workflows include sampling the community of interest, extracting genomic DNA, and amplifying the barcode gene of interest by Polymerase Chain Reaction (PCR; Hugerth and Andersson 2017). Sequencing adapters and sample-specific index sequences, the latter of which are used for multiplexing multiple samples into a single sequencing run, can be attached during amplification of the barcode gene (a "1-step PCR"

approach; Kozich et al. 2013) or in a second PCR or ligation reaction (a "2-step PCR" approach; Gohl et al. 2016; O'Donnell et al. 2016). Amplicons from each sample are then purified, pooled in equimolar concentrations, and sequenced. This workflow culminates in millions of sequence reads representing tens of thousands of unique sequences. Various bioinformatic algorithms have been developed to classify these sequences into either operational taxonomic units (OTUs) or amplicon sequence variants (ASVs; e.g., Callahan et al. 2016; Callahan et al. 2017). Despite differences in their computational formulations and biological significance, OTUs or ASVs generally serve as proxies for biological species, and proportional sequence counts of each ASV within a sample are frequently used as a proxy for each species' relative abundance. Amongst protists, interspecific variance in 18S gene copy numbers obscures the relationship between sequence counts and cell abundances; however, the high correlation between 18S gene copy number and cell biovolume (Zhu et al. 2005; Godhe et al. 2008) suggests that amplicon sequencing assessments of marine protists may provide ecologically and biogeochemically powerful indices of community composition.

Uncertainties can be introduced at each step of the amplicon workflow (e.g., Bradley et al. 2016; Gohl et al. 2016; see Bálint et al. 2016 and Hugerth and Andersson 2017, for reviews). In addition to the uncertainty inherent in biological sampling, interspecific variability in cell lysis efficiencies caused by differences in the robustness of extracellular structures can introduce taxon-specific biases at the DNA extraction step. Rigorous cell lysis procedures (e.g., bead-beating and enzymatic methods) prior to extraction of genomic DNA reduce this bias (Yuan et

al. 2015; Djurhuus et al. 2017). PCR is considered the most critical step in the amplicon sequencing workflow as the choice of hypervariable region, PCR primers, reaction conditions, and many other factors can alter results of downstream analysis and the detection of specific lineages (Stoeck et al. 2010; Hu et al. 2015; Bradley et al. 2016; Parada et al. 2016). PCR can introduce biases in observed relative sequence abundances as large as 10-fold from those expected for specific OTUs, particularly when primer mismatches are observed (Bradley et al. 2016; Parada et al. 2016; Wear et al. 2018). The choice of bioinformatic pipeline and other data pre-processing procedures can introduce additional uncertainty (Bokulich et al. 2013; McMurdie and Holmes 2014), although recent advances in inferring exact ASVs dramatically reduce the impact of analytical artifacts in amplicon data and are poised to facilitate methodological improvements, cross-study comparisons, and meta-analyses (Callahan et al. 2016; Callahan et al. 2017). Despite these numerous sources of uncertainty, attempts to validate and/or quality-control amplicon methods are often not considered and only rarely published.

For studies of marine protist communities, the V4 and V9 hypervariable regions of the 18S gene offer the most complete picture of marine protist diversity at present, and various workflows and primer sets have been proposed and employed to target these regions (Amaral-Zettler et al. 2009; Stoeck et al. 2010; Hugerth et al. 2014; Hu et al. 2015; Bradley et al. 2016). The major benefit of targeting the V4 rather than the V9 region is the increased phylogenetic and taxonomic resolution enabled by the increased amplicon length (Hu et al. 2015). However, recent work demonstrated reduced accuracy in community composition for the V4 relative to the

(V8-)V9 region for eukaryotic phytoplankton communities, and attributed these differences in part to the longer, lower fidelity sequencing reads required for the V4 region (Bradley et al., 2016). The shorter read lengths required for the V9 region also enable more cost-effective large scale analyses (e.g., De Vargas et al. 2015). Taken together, these factors suggest that the V9 region may be better suited for large-scale investigations (e.g., long time series, global-scale analysis) requiring increased quantitative power. Further investigation of the bias and uncertainty in amplicon workflows targeting the V9 and surrounding regions of the 18S gene are thus needed to constrain the quantitative potential of amplicon sequencing assessments of marine protist communities and identify workflows suitable for linking community variations with environmental, biogeochemical, and other biotic parameters. Such investigations will enable comparisons to well-validated methods for characterizing marine microbial eukaryotic communities (e.g., High Performance Liquid Chromatography phytoplankton pigment determinations; Hooker et al. 2010) and help establish methods best-suited for large-scale quantification of plankton functional types (e.g., Quere et al. 2005).

Here, we use a combination of mock communities and environmental samples to evaluate analytical uncertainty at different stages of an amplicon sequencing workflow using a 1-step PCR approach to target the V9 hypervariable region of the 18S gene for marine protist communities. We provide estimates of precision and accuracy at different steps of our workflow, from sampling and DNA extraction to PCR and sequencing, and examine the effects of procedural modifications on downstream results, including different DNA extraction methods

and PCR protocols. Our analysis provides strong support for the continued use of positive and negative controls in amplicon sequencing analysis. Finally, we show that validations of amplicon workflows employing a single primer set are an often overlooked yet imperative first step in more general comparisons of different primer sets and hypervariable regions, as well as for conducting quantitative downstream analyses of these data.

Materials and Procedures

DNA Extraction Method Comparisons

To evaluate the effects of DNA extraction method on estimates of protist community composition, triplicate samples of natural marine protist communities were collected during Plumes and Blooms (PnB) cruises in the Santa Barbara Channel, CA, USA in May and August, 2017. The first cruise occurred on May 9, 2017, and samples were collected at PnB station 5 during an abnormally large bloom of pennate diatoms (Fig. S1). The second took place on August 17, 2017 and samples were collected at PnB station 3 where low chlorophyll concentrations were observed (Fig. S1). During each cruise, three replicate 2 L surface water samples were collected from Niskin bottles deployed on a CTD rosette and returned to the laboratory. One-liter aliquots were filtered through a 47 mm 1.2 µm mixed cellulose esters membrane filter using positive pressure (~5 psi) and stored frozen in 5 mL cryovials at -80 ºC. These two batches of samples are referred to as “May” and “August” samples. Details regarding PnB station locations and data products are available at

https://seabass.gsfc.nasa.gov/experiment/Plumes_and_Blooms/. Relevant PnB analytical methods are described in Catlett and Siegel (2018), and briefly in Supporting Text S1.

The May and August samples were used to compare two DNA extraction methods, both of which included mechanical (via bead-beating) and enzymatic lysis procedures. Prior to DNA extraction, each filter was cut in half and each half-filter was subjected to one of two DNA extraction procedures. DNA was extracted from one filter-half using the DNeasy PowerWater DNA extraction kit (Qiagen; the PW method) following the manufacturer's instructions, including an additional 10 min incubation at 65 ºC to enhance cell lysis as recommended by the manufacturer. The remaining half-filter was returned to the cryovial, to which 1.8 mL of sucrose lysis buffer (750 mM sucrose, 20 mM EDTA, 400 mM NaCl, 50mM Tris-HCl; pH 8.0) was added. These half-filters were stored at -80 ºC until DNA extractions using a "custom" lysis (see below; Countway et al. 2007) and phenol-chloroform extraction protocol (Giovannoni et al. 1990; the PC method) were conducted.

Lysis for the PC extraction method consisted of three rounds of bead beating with 0.4 mm Zirconium beads (Molecular Biology Grade, OPS Diagnostics, USA) for 1 min followed by incubation at 70 ºC for 5 min. For bead-beating, up to 10 sample tubes were secured horizontally to a Fisherbrand Scientific Vortex Mixer, and bead-beating was carried out at maximum speed (~3200 rpm). Next, 200 µL of SDS (10% w/v) and 20 µL of proteinase K (20 mg/mL) were added and the samples were incubated at 55 ºC with gentle rotational mixing for 1 hour. Genomic DNA was extracted from 1 mL of lysate using an equal volume of phenol-chloroform-

isoamyl alcohol (25:24:1, pH 8.0), followed immediately by two additional extractions with equal volumes of chloroform-isoamyl alcohol (24:1). DNA was precipitated with 7.5M NH_4OAc and 100% isopropanol for three hours in the dark at room temperature, washed with 70% ethanol, and resuspended in 1M Tris HCl (pH 8.0) overnight at 4 ºC before being stored at -20 ºC. Triplicate DNA extraction blanks were included for each method by subjecting unused filters to the procedures described above. Results from additional DNA extraction blanks using the PC method are also considered.

Mock Community Construction

Mock communities were composed of 22 unique, full-length 18S gene amplicons isolated from marine protist species spanning seven major eukaryotic phyla (Table 1). Many of the species included in mock communities are widely distributed and often abundant in the global surface ocean and in the Santa Barbara Channel, CA, USA where our amplicon workflow has been applied toward a time series of surface ocean samples of marine protist communities (Catlett et al. unpublished data). Genetic material for mock communities was generated from marine phytoplankton cultures. Isolates were acquired from several different collections, including UC Santa Barbara, Moss Landing Marine Laboratory, University of Southern California, Monterey Bay Aquarium Research Institute, and the National Center for Marine Algae. Batch cultures were maintained under conditions most suitable for growth, in either

f/2+Si, f/2-Si, or modified f/25-Si culture media (Guillard 1975) under a 12:12 light:dark cycle at either 15 or 22 ºC.

Samples from single phytoplankton cultures were filtered onto either 0.2 µm pore size, 47 mm diameter Supor polyethersulfone filters or 1.2 µm pore size, 47 mm diameter mixed cellulose ester filters under gentle vacuum and stored frozen at -80 ºC. Genomic DNA was extracted from each filter using a PowerSoil DNA extraction kit (MoBio, USA) following the manufacturer's instructions. Full-length 18S genes were amplified from genomic DNA extracts using the primers EukA and EukB (Medlin et al. 1988; see Table S1 for all primer names, sequences, and sources used in the present analysis). These PCRs used 0.8x KAPA2G Robust HotStart ReadyMix (KAPA Biosystems), 0.4 µM each primer, and 2 µL template with thermal cycling as follows: 95 ºC for 2 min; 35 cycles of 95 ºC for 30 s, 55 ºC for 30 s, 72 ºC for 2.5 min; and 72 ºC for 7 min. PCR products were cloned using the pGEM®-T Easy Vector system (Promega, USA) following manufacturer instructions. Unique full-length 18S genes were amplified directly from clonal cultures in 25 µL reactions using 1x KAPA2G Robust HotStart ReadyMix, 0.5 µM each M13F and M13R primers (Table S1), and 1 µL clonal culture as template, with the following thermal cycling settings: 95 ºC for 5 min; 35 cycles of 95 ºC for 15 s, 65 ºC for 15 s, 72 ºC for 2.5 min; and 72 ºC for 7 min. PCR products were purified using the Qiagen QIAquick Gel Extraction kit following manufacturer's instructions, and submitted to the DNA Sequencing Facility at UC Berkeley for Sanger sequencing using the M13F, 563F, and M13R primers (Messing 1983; Hugerth et al. 2014; Bradley et al. 2016; Table S1). Sanger

sequences were merged and initial taxonomic identities were assigned by analysis of BLASTN searches against the Protist Ribosomal Reference database (PR2; v4.11.1; Guillou et al. 2012) using MEGAN's Lowest Common Ancestor Algorithm (LCA; Altschul et al. 1990; Huson et al. 2007; see Table 1 for initial taxonomic assignments).

Amplicon concentrations were quantified in triplicate with the Qubit 3.0 dsDNA High Sensitivity Assay kit (ThermoFisher Scientific, USA) prior to addition to each mock community. Expected relative mass abundances were determined using the average of the three independent quantifications and were converted to molar ratios based on the lengths of the Sanger sequences of each full-length 18S amplicon (EukA to EukB primer binding regions, plus the remaining plasmid sequence from the pGEM-T Easy Vector according to the manufacturer's documentation) and an average molecular weight of 617.96 g mol bp^{-1}. Eight unique mock communities, or amplicon assemblages (AAs), were created with varying community composition. The identities, sequence characteristics, and relative abundances of each amplicon included in each AA sample are summarized in Table 1. Target compositions were an even representation of all 22 amplicons (AA1), an even representation of a subset of 12 amplicons (AA5), dominance by several diatom or dinoflagellate amplicons (AA2 and AA3, respectively), "mixed dominance" by a diverse collection of amplicons (AA4), and dominance by a single haptophyte amplicon (AA6), dinoflagellate amplicon (AA7), or prasinophyte amplicon (AA8).

Assessments of PCR Precision and Accuracy

The eight AAs, along with genomic DNA extracts from natural marine protist communities, were used to assess the precision and accuracy of PCR and sequencing. The AAs allowed for assessments of both accuracy and precision, and the environmental samples allowed for assessments of precision for natural, more diverse protist communities. In addition to the environmental samples collected for comparisons of DNA extraction methods, we considered 94 environmental samples collected on PnB cruises between June 2011 and September 2014 as described in Wear et al. (2018). All samples considered in these analyses and their experimental data are detailed in Supporting File S1. Table 2 provides a quick reference for the abbreviations used to denote sample names and characteristics, procedural variations employed, and other terms with definitions specific to the present analysis.

In all PCR reactions, the V9 hypervariable region of the 18S gene was amplified with a 1-step PCR using custom dual-indexed primers (Kozich et al. 2013) designed from the 1391F and EukB primers (Stoeck et al. 2010; Table S1). Reactions were carried out using 0.4 µM of each primer and 1x KAPA2G Robust HotStart ReadyMix (KAPA Biosystems). Thermal cycling conditions were: 94 ºC for 3 min; 35 cycles of 94 ºC for 45 sec, 65 ºC for 15 sec, 57 ºC for 30 sec, and 72 ºC for 90 sec; 72 ºC for 10 min; 4 ºC until being stored frozen at -20 ºC or continuing with clean-up and normalization. For field and blank samples, 1 µL of genomic DNA template was used in each reaction, regardless of the DNA concentration (all blank samples had DNA concentrations below detection limits of the Qubit 3.0 Fluorometer dsDNA High Sensitivity Kit). AA samples used 20 pg of template DNA in each PCR reaction unless otherwise noted. For each

sequencing library, at least one no-template control using 1 µL PCR-grade water as template per 25 µL reaction was also amplified and included in all subsequent purification and sequencing steps. Unless otherwise noted (see below), samples were amplified in triplicate 25 µL reactions and products were pooled before proceeding with Illumina library preparation and sequencing.

We assessed the impacts of modifications to the above PCR protocol (referred to as the "Standard" protocol) on the precision and accuracy of sequencing results in order to inform the potential for standardization and meta-analyses of amplicon sequencing data employing the same primer set with different PCR protocols. All of the PCR protocols are derived from the Earth Microbiome Project (Thompson et al. 2017), with slight modifications. Under the assumption that the unique 8-nt index sequences included on the dual-indexed 1391F and EukB primers have negligible effects on downstream sequencing results (this was not a valid assumption; see *PCR/Sequencing Method Comparisons* in Results), each procedural modification was tested with at least one AA sample and the Aug1_PC sample (Fig. 1) amplified and sequenced in triplicate. PCR protocols were modified relative to the Standard protocol as follows:

(1) Rather than pooling the products of triplicate PCRs prior to clean-up and normalization, some samples were amplified in a single 25 µL PCR with identical reaction conditions. This method is referred to as the "PCRx1 method".

(2) Some samples were amplified with a modified thermal cycling protocol, with a single annealing step of 57 ºC for 60 sec rather than two annealing steps of 65 ºC for 15 sec, 57 ºC for 30 sec. This method is referred to as the "EMP method".

(3) Some samples were amplified in the presence of a mammal blocking primer as suggested by Vestheim and Jarman (2008; Mammal_block_I-short_1391F; Table S1). For these samples, each 25 µL PCR reaction consisted of 1x KAPA2G Robust HotStart ReadyMix, 0.4 µM forward and reverse primer, and 1.6 µM mammal blocking primer, and thermal cycling was consistent with the Standard method. This method is referred to as the "MB method".

Illumina Library Preparation and Sequencing

For most sequencing runs, PCR products were purified and normalized using the SequalPrep Normalization Plate Kit (Applied Biosystems), pooled into a single library, concentrated with Amicon Ultra-0.5 Centrifugal Filter Devices (Millipore), gel extracted using the Qiagen QIAquick Gel Extraction kit, and concentrated again with an Amicon Ultra-0.5 Centrifugal Filter Device. Library sequencing was performed using a MiSeq PE150 v2 kit (Illumina) at the DNA Technologies Core of the UC Davis Genome Center. An additional MiSeq PE150 Micro sequencing run and MiSeq PE150 Nano run were performed at UC Davis and at the California NanoSystems Institute at UC Santa Barbara, respectively. For these two sequencing runs, PCR products were purified using the Qiagen QIAquick Gel Extraction kit, quantified using the Qubit 3.0 dsDNA High Sensitivity kit, and pooled in equal concentrations before sending to the sequencing facility. The data presented here come from a total of nine

sequencing runs. Each sequencing run included technical PCR/sequencing triplicates of AA1, as well as at least one PCR blank, and often multiple DNA extraction blanks, as negative controls.

Bioinformatics

Demultiplexed Illumina data were processed using the DADA2 pipeline (Callahan et al. 2016). Reads 1 and 2 were trimmed to 140 and 120 nt, respectively, filtered (maxEE=2, truncQ=2), and denoised using the DADA algorithm. The DADA error model was parameterized for each MiSeq run using at least 10^8 bases. Following error correction, paired reads were merged and overhanging sequences (due to the sequencer reading through primer sequences at the 3' end of the amplicon) were trimmed. Chimeras were removed simultaneously from all 9 sequencing libraries using the "consensus" method. Initial taxonomic assignment was performed with a Bayesian classifier algorithm (Wang et al. 2007) with a minimum bootstrap confidence of 80% using the training set file for the Protist Ribosomal Reference (PR2; Guillou et al. 2012) database (v4.11.1) and the Silva SSU reference database (v132; Quast et al. 2012) available for the DADA2 pipeline (https://benjjneb.github.io/dada2/training.html). The Silva database was used to identify prokaryotic sequences while the PR2 database was used for all other assignments due to the Bayesian classifier's propensity to "over-classify" sequences to the Kingdom Eukaryota when used only with the PR2 database.

Secondary Taxonomic Assignments and Generation of a Merged Taxonomy Array

Following initial taxonomic assignment, many ASVs were unable to be unambiguously identified as protists (defined here as unicellular eukaryotes, excluding fungi and macroalgae; (Adl et al. 2012; Guillou et al. 2012). These sequences (lacking classification at the Kingdom level with Silva, or sequences unclassified beyond the Kingdom level or the Supergroups *Opisthokonta* or *Archaeplastida* with PR2) were re-classified via analysis of BLASTN searches against the Silva and PR2 databases using LCA in MEGAN6 (Altschul et al. 1990; Huson et al. 2007). This resulted in four unique taxonomy arrays for our data set. According to the method of taxonomic assignment and the database used, we call these four taxonomy arrays Bayesian-pr2, Bayesian-silva, LCA-pr2, and LCA-silva.

The three taxonomies (PR2, Silva, and LCA, the latter of which is mapped to the NCBI taxonomy) are not consistent in their naming or ranking conventions. We thus sought to map the three taxonomies onto a common taxonomy and generate a single "merged" taxonomy array for our data set, with the goal of preserving the true protist diversity present in our data while adequately controlling for ASVs of non-protist origin. Since the LCA-pr2 and LCA-silva assignments are mapped to a common taxonomy, we first merged the two LCA taxonomy arrays (henceforth, LCA-merged) with sequences assigned to either LCA-pr2 or LCA-silva according to the following rules:

1. If no hits were assigned by both databases, the sequence remained unassigned.
2. If one database provided a taxonomic assignment while the other had no hits assigned, the former was used.

3. If both databases agreed in their assignments but not in their resolution, the more finely resolved taxonomy was used.
4. If the two databases disagreed at the Kingdom rank, the LCA-silva classification was used.
5. In all other disagreements, the LCA-pr2 assignment was used.

Next, we mapped the LCA and Silva taxonomies to the PR2 taxonomy. Since PR2 does not include bacterial or archaeal sequences, sequences assigned to either of these domains with the Silva database retained these assignments at the "Kingdom" rank when mapped onto PR2 (see Supporting Files S2 and S3). We first collected all unique taxonomic assignments observed in our data set from the four highest taxonomic ranks of the Bayesian-silva, Bayesian-pr2, and LCA-merged taxonomies. We then mapped those from the Bayesian-silva and LCA-merged taxonomies to those in the Bayesian-pr2 taxonomy. Mapping files were generated by first manually searching for identical taxonomic names, regardless of rank, between the Bayesian-silva and LCA-merged taxonomies and the Bayesian-pr2 taxonomy. If an identical taxonomic name was found in the Bayesian-pr2 taxonomy, the taxonomy was mapped to the rank at which the identical taxonomic name was found, and the remaining ranks were left unclassified. If an identical taxonomic name was not found in the Bayesian-pr2 taxonomy, we searched for synonyms in Adl et al. (2012). If a synonym with a corresponding entry in the Bayesian-pr2 taxonomy was found, this taxonomy was mapped as above. Otherwise, the taxonomic name was left unmapped.

After executing the taxonomic mapping, we created a merged taxonomy array taking into account assignments from the Bayesian-pr2 taxonomy array and the mapped Bayesian-silva and LCA-merged taxonomy arrays as needed in order to determine whether each sequence was of protist origin. Each sequence was assigned as follows:

1. If the Bayesian-pr2 assignment included an assignment at the Division rank, or if it was assigned to an unambiguous protist Supergroup (not *Opisthokonta* or *Archaeplastida*) it was used.
2. If the Bayesian-pr2 assignment did not satisfy the conditions of (1), but the mapped Bayesian-silva assignment did, the Bayesian-silva annotation was used.
3. If (1) and (2) were not satisfied and the mapped Bayesian-silva assignment was a prokaryote, the Bayesian-silva assignment was used.
4. If (1), (2), and (3) were not satisfied and the mapped LCA-merged taxonomy could delineate sequences of protist origin (e.g., satisfied any of the above conditions), the mapped LCA-merged annotation was assigned.
5. In all other scenarios, the Bayesian-pr2 assignment was retained.

This merged taxonomy array was used to remove non-protist ASVs at the data pre-processing step, and is used in all downstream analyses presented in this study.

Additional Preprocessing

ASVs less than 90 nt or greater than 180 nt in length (target amplicon is 120-130 nt) were discarded. ASVs assigned as *Bacteria*, *Archaea*, *Metazoa*, *Fungi*, *Streptophyta*, *Rhodophyta*, or *Ulvophyceae* were removed. ASVs that remained unassigned at the Kingdom or Supergroup rank, or that were assigned to the Supergroups *Opisthokonta* or *Archaeplastida* but not assigned to a Division, were also discarded. We note that some valid protist ASVs may have been discarded using this strategy, but for all samples considered in the present analysis, we typically discarded less than 30 ASVs composing a summed total of less than 1% of reads in each sample (with some exceptions; Supporting Files S4 and S5). We also note that we assume ASVs unclassified beyond the division *Chlorophyta* are protists, though they may still include unidentified *Ulvophyceae* ASVs. *Ulvophyceae* ASVs were rare in our data set and are expected to be less common at the stations occupied by PnB.

Following removal of non-protist ASVs, ASV abundances in each sample were normalized by the remaining total reads in each sample. Unless otherwise noted, analyses of environmental samples rely on these normalized data. In AA samples, spurious ASVs (those that were not exact matches to the 22 target ASVs) were often detected at low relative abundance (see Table 3 below). For most AA analyses, the 22 target ASVs were computationally isolated from each sample and relative abundances were calculated only considering reads attributable to target ASVs. However, we also investigated the effects of spurious ASVs on downstream analyses and note these in the Results section.

Assessment

DNA Extraction Method Comparisons

Surface ocean samples collected during May and August from the Santa Barbara Channel (CA, USA) were used to assess uncertainty in protist community composition introduced by different DNA extraction methods and compare this uncertainty to that present amongst sampling/biological replicates (Fig. 1). In May, an abnormally large bloom of pennate diatoms was sampled, while August samples were collected during relatively low biomass conditions (Fig. S1). We evaluated differences between a phenol-chloroform DNA extraction method employing mechanical and enzymatic lysis procedures (adapted from Giovannoni et al. 1990 and Countway et al. 2007; the PC method) and the Qiagen DNeasy PowerWater DNA extraction kit (henceforth, the PW method).

Large differences in protist community composition were observed between the May and August samples, regardless of the DNA extraction method employed (Fig. 1). Triplicate DNA extraction blanks (negative controls) were also included for each method (see Table 3 below). While protist ASVs were observed in all DNA extraction blanks using both methods, five or less "contaminant" ASVs (see Table 2 and *Accuracy of ASV Detection in AA and Blank Samples* for definitions) were found in each blank sample. Given that >200 ASVs were detected in all May and August samples and the standard deviations of observed ASVs across sampling replicates were >20, we conclude that contamination was minimized for both methods and is unlikely to influence the results of these comparisons.

The variability introduced by different DNA extraction methods depended on the community sampled (Fig. 1). When considering all protist ASVs, samples clustered according to extraction method in May, and according to sampling replicates in August (Fig. 1A). This clustering pattern remained robust when only considering ASVs found at >1% relative abundance in at least one sample (Fig. 1B), indicating these results were not driven by rare ASVs. In August, the magnitude of Bray-Curtis dissimilarities (BCD) amongst sampling replicates (within-PC and within-PW) was not significantly larger than that observed across extraction methods (PC vs. PW; Fig. 1C; Tukey's HSD test, $p > 0.8$). Conversely, in May, significantly higher BCDs were observed between samples extracted by different DNA extraction methods (Tukey's HSD test, $p < 0.015$), while BCDs amongst sampling replicates remained comparable to those observed in August (Fig. 1C; Tukey's HSD test, $p > 0.7$).

The differences observed between the two DNA extraction methods in the May samples were primarily driven by differences in the relative abundances of dinoflagellates, cercozoans, and diatoms (Classes *Dinophyceae*, *Filosa-thecofilosea*, and *Bacillariophyta*, respectively; Fig. 1D). May samples extracted with the PC method showed higher relative abundances of diatoms, while PW samples showed higher abundances of dinoflagellates and cercozoans (Fig. 1D). The May samples were collected during a large diatom (*Pseudo-nitzschia sp.*) bloom in the Santa Barbara Channel, and biomarker pigment and cell count data indicated that diatoms dominated the community (Fig. S1). In addition to the observed dominance of diatoms in PC samples, more ASVs (mean ± standard deviation) were detected in May PC samples (335 ± 43) than in the May

PW samples (248 ± 29). These differences in ASV richness were not statistically significant (paired *t*-test, $p = 0.13$), though rarefaction analysis (Fig. S2) suggested that differences in sequencing depth did not confound comparisons of ASV richness or alpha diversity. This suggests the more rigorous lysis employed by the PC method yields greater accuracy in observed diversity and community composition.

PCR/Sequencing Method Comparisons

The AAs, along with genomic DNA extracts from natural marine protist communities, were used to assess precision at the PCR and sequencing steps of the workflow. Principal coordinate analyses (PCoA) of all AA1 replicates (39 total across nine MiSeq runs) and Aug1_PC replicates (19 total across three MiSeq runs) showed clear outliers (Fig. 2A, 2C, 2E). No single procedural modification (MB, EMP, or PCRx1) was able to explain these outliers, nor were they exclusive to a single MiSeq run (Figs. 2 and 3). Further analysis revealed that two specific indexed forward primers (SA501 and SA505; Table S1) produced estimates of community composition that were qualitatively and quantitatively dissimilar from the other eight forward indexed primers used (Figs. 2 and 3), demonstrating the importance of mock communities as a diagnostic tool for ground-truthing methods.

Mean BCDs amongst AA1 and Aug1_PC PCR/sequencing replicates are shown in Fig. 3A and B. For these analyses, mean BCDs were calculated across all replicates (including those amplified with modified PCR protocols) amplified and sequenced with or without the SA501 or

SA505 primers. BCDs across other duplicated environmental samples (Fig. 3C) represent the average across all pairs of duplicates partitioned according to the pair of forward indexed primers used. Across all sample types and experimental treatments, BCDs were inflated two to three-fold on average when replicates that were amplified with either SA501 or SA505 were included in the calculation relative to when they were not. This pattern was consistent both within each MiSeq run and across all nine MiSeq runs. We inspected sequence alignments of the forward indexed primers with the corresponding 18S gene regions of the AA ASVs and calculated various biochemical properties of each forward indexed primer (using the Integrated DNA Technologies OligoAnalyzer tool; https://www.idtdna.com/; see Supporting File S6), but were unable to identify systematic patterns to explain these differences.

Ignoring the variance introduced by certain indexed primers, the MB method (including a mammal blocking primer in the PCR) was the most dissimilar to all other experimental treatments (Fig. 2B, 2D, 2F) and consistently failed to detect one of the 22 target ASVs in both AA1 and AA2 (see Tables 3 and S2 below). The PCRx1 and EMP methods produced results indistinguishable from the Standard method (Fig. 2).

Precision of Relative Sequence Abundances

The relative sequence abundances of specific ASVs or OTUs are often analyzed directly to assess the impacts of a biotic or abiotic forcing on the dynamics of a particular population, or to derive correlation networks that are subsequently used to define the dynamics of specific sub-

communities (e.g., Guidi et al. 2016; Needham and Fuhrman 2016; Wang et al. 2018). Such correlative analyses implicitly assume that relative sequence abundances are reasonably precise and accurate, though very few empirical estimates of the precision of relative sequence abundances are available in the literature. The May and August environmental samples described above (Fig. 1) can be used to quantify the precision of relative sequence abundances for complex natural samples of marine protists across biological/sampling and PCR/sequencing replicates. Here we seek to place constraints on the precision that can be expected for the relative abundances of specific protist ASVs or lineages in complex environmental samples.

The coefficient of variation (CV) is a widely used estimate of experimental precision. CV values are calculated here as the ratio of the standard deviation to the mean and are expressed as a percentage. Figure 4 shows the CVs for relative sequence abundances amongst the May_PC (Fig. 4A, 4B) and Aug_PC (Fig. 4C, 4D) sampling replicates, and amongst PCR/sequencing replicates of the Aug1_PC sample (Fig. 4E, 4F). CVs were calculated for the relative abundances of individual ASVs (only the most abundant 250; Fig. 4A, 4C, 4E), and for summed abundances of ASVs within each taxonomic Class (Fig. 4B, 4D, 4F), and were sorted by rank abundance along the x-axis. Notably, the Aug1_PC PCR/sequencing replicates included those amplified with the EMP and PCRx1 protocols and excluded those amplified with the MB protocol and with the SA501 or SA505 primers. Variations in sequencing depth across the Aug1_PC PCR/sequencing replicates considered here were as large as 5-fold (14447 to 73647 reads per replicate) and differences in the number of ASVs observed were greater than 2-fold (293 to 593

ASVs per sample). Thus, the estimates of precision for the PCR/sequencing step provided by this analysis are conservative.

For all three sample sets, the relative abundances of the most abundant ~50 ASVs and ~30 Classes were reasonably precise (CVs < ~50%), with some exceptions. Precision was reduced for less abundant ASVs regardless of whether sampling variability was considered, demonstrating that caution should be applied when interpreting abundances of rare ASVs (Fig. 4A, 4C, 4E). Notable differences in precision were observed between the Aug_PC and May_PC sampling replicates, which may be attributable to the differences in observed community composition for these two sets of samples (Fig. 1D; Fig. 4B and 4D). The precision for ASV relative abundances was worse in the May_PC samples than in the Aug_PC samples, with CV's of the most abundant 50 ASVs often exceeding 20% in the May_PC samples. Relative abundances of the most abundant 50 ASVs and 20 Classes in the Aug_PC sampling replicates showed high precision overall (CVs generally < 20%), while only ~10 Classes were quantified with high precision in the May_PC sampling replicates.

As expected, ASV and Class relative abundances were most precise in the absence of sampling variability (Fig. 4E and 4F). CVs across the Aug1_PC samples were consistently less than 20% for the most abundant 50 ASVs and 25 Classes (Fig. 4E and 4F). The Aug1_PC replicates considered in these analyses incorporate analytical uncertainty associated with variations in indexed primers, PCR protocols, sequencing depth, and ASV richness. Taken together, these results indicate that if sampling/biological variability is ignored or reduced

relative to that observed here, the abundances of the most abundant ~50 ASVs and ~30 Classes within a sample are likely robust and reproducible. If biological/sampling variability is expected to be significant, the abundances of specific ASVs should be interpreted with caution, but grouping ASVs to a coarser taxonomic rank can mask this biological variability and allow for more robust interpretations of the abundances of specific taxonomic groups (e.g., Classes).

Accuracy of ASV Detection in AA and Blank Samples

Overall, the primer set and PCR protocol evaluated here coupled with the DADA2 pipeline were exceptionally effective in detecting all 22 target ASVs included in AA samples, with each ASV perfectly matching Sanger sequences of the cloned 18S amplicons (Tables 3 and S2). Notably, DADA2 was able to accurately distinguish two haptophyte ASVs (ASVs 11 and 12; Table 1) that differed by a single nt. However, spurious ASVs were often detected in AA and blank samples. These are frequently observed in mock community studies and can arise due to PCR and sequencing errors that escape bioinformatic correction, "tag-jumping" or "cross-talk", or contamination (Schnell et al. 2015; Callahan et al. 2016; Edgar 2016; Minich et al. 2019). Spurious ASVs can confound downstream comparisons of diversity, and alter the relative abundance of ASVs of biological interest due to the compositional nature of HTS data.

We attempted to classify spurious ASVs in AA and blank samples to better understand the contributions of the above different sources of error to our workflow (see Table 2 for definitions). Spurious ASVs were labeled "contaminants" if they were detected in any non-

control sample (and for blank samples, any AA sample) amplified and/or sequenced simultaneously with the control sample, "artifacts" if they were detected only in AA samples and were >95% similar to a target ASV sequence in that sample, and "unknown" if neither of the above conditions was satisfied (Tables 3 and S2). Thus, the contaminant classification includes cross-talk in addition to true contaminants. The artifact classification includes spurious ASVs that are derived from a true ASV but escaped bioinformatic correction. The unknown classification includes additional analytical artifacts (e.g., chimeras that escaped bioinformatic correction, highly aberrant PCR/sequencing errors, etc.) and other contaminants that are not sufficiently abundant relative to the target ASVs found in non-control samples to be detected.

With the exception of a single PCR blank that appeared contaminated (PCR_blank_Run8; Table 3), less than 15 spurious ASVs and less than five contaminant ASVs were typically detected in PCR blanks. The contaminated PCR blank was amplified and sequenced alongside an AA sample that also appeared contaminated (MB_AA1A_Run8) with 19 spurious ASVs detected, 14 of which were classified as contaminants (Table 3). These contaminated samples were amplified and sequenced in the presence of a single environmental sample that was clearly the source of cross-sample contamination. Interestingly, the other blank and AA samples amplified and sequenced in this batch did not show signs of unusually high contamination (sample names ending in "Run8"; Tables 3 and S2). Less than 30 ASVs were generally detected in DNA extraction blanks, and less than 10 of these were typically classified as contaminants. When the same DNA extraction blank was amplified and sequenced on

different MiSeq runs, the number of ASVs and the relative contributions of contaminant and unknown ASVs often varied across runs. Systematic increases were not observed in sequential MiSeq runs, suggesting that these variations were introduced by reaction-specific variations in cross-sample contamination, cross-talk, and/or artifact formation rather than contamination of the sample itself.

ASV richness in AA samples was generally robust to false negatives using the standard, EMP, and PCRx1 PCR protocols (see Table 2 for definitions), while one target ASV (ASV 11, within the Division *Haptophyta*) was consistently undetected in AA samples amplified with the MB protocol and with certain indexed primers (SA501 and SA505; Tables 3 and S2) that were determined to significantly reduce accuracy and precision in our workflow (Figs. 2 and 3). Accuracy in ASV richness was more heavily skewed by false positives. Typically, we detected fewer than five spurious ASVs in each AA sample, with summed relative abundances generally less than 1% of sequence reads (Tables 3 and S2). Most spurious ASVs in AA samples were classified as contaminants (e.g., were also present in non-control samples), although occasionally artifact or unknown ASVs were detected. The only instance in which artifact ASVs became highly significant were in samples of AA2 using template concentrations diluted 100-fold from standard AA samples (AA2lowA, B, and C; Tables 3 and S2). In these samples, 24, 53, and 3 artifact ASVs were detected, but these samples also had high sequencing depth (Supporting File S4).

Accuracy of Relative Sequence Abundances in AA Samples

The AA samples can be used to derive estimates of the accuracy of relative sequence abundances in representing the relative abundances of 18S gene copies in a mixed assemblage (Fig. 5). All estimates of accuracy consider only target ASVs in each mock community. A "pseudocount" of 0.0001 was added to all relative abundances for analyses relying on log-transformations to prevent undefined values when taking the logarithm of 0. Most of the results presented are robust to the effects of spurious ASVs, though analyses employing a log-transformation are highly sensitive to the magnitude of the pseudocount used (Fig. S3).

Across the six different AAs that included all 22 target ASVs with varied relative abundances, observed relative abundances of the target ASVs were highly correlated with those expected in both linear and log-transformed space (Fig. 5A and 5B). Coefficients of determination between observed and expected abundances calculated for each AA individually were also generally high ($R^2 > 0.8$) in both linear and log space (Fig. 5C and 5D). In linear space these were reduced in the presence of highly abundant outliers (see AA4, $R^2 < 0.75$ and AA6 $R^2 < 0.5$; Fig. 5C), and in log space decreased when an ASV was undetected (in MB_AA2; Fig. 5D). Values of Spearman's rank correlation coefficient between observed and expected community compositions were also high ($\rho > 0.75$) for these AA samples (Fig. 5E), and were more robust to outliers and variations in community composition and PCR protocols. Finally, root mean square error (RMSE; Fig. 5F) between observed and expected abundances was

consistently < 0.02, but increased for some communities, most notably AA5 and AA6 due to significant bias against highly abundant ASVs.

The magnitude of bias introduced by PCR and sequencing can be quantified as the base-two logarithm of the fold-change in observed abundances relative to those expected (e.g., Parada et al. 2016; Wear et al. 2018; Figs. 6 and S4). In most AA samples, the magnitude of bias was less than two-fold for most ASVs, though in some instances was greater than four-fold (Fig. 6). Many sources of PCR and sequencing bias are known: variations in taxon-specific primer affinities, template sequence characteristics, template abundance, and others (Suzuki and Giovannoni 1996; Polz and Cavanaugh 1998; Aird et al. 2011; Gohl et al. 2016; Parada et al. 2016; Laursen et al. 2017). Primer mismatches are considered the largest source of PCR/sequencing bias (Gohl et al. 2016; Parada et al. 2016; Wear et al. 2018). In the AA samples analyzed here however, amplicon GC content appeared to be the primary source of PCR/sequencing bias (Fig. 6). The V9 primer set evaluated here exhibited single nucleotide mismatches between the forward primer (1391F) and ASVs 11 and 20 at primer positions eight and seven, respectively (Table 1). ASV 11 was under-represented in all AAs, but also had the highest GC content of any target ASV (Figs. 6 and S4). ASV 20 was slightly underrepresented in most, but not all AAs (Figs. 6 and S4). Primer mismatch bias can be more significant when mismatches are observed towards the 3' end of the primer (Gohl et al. 2016), but these results suggest single primer mismatches towards the 5' end of the primer can be tolerated with minimal bias in our workflow.

Discussion

Implications for Meta- and Downstream Analyses

The results above demonstrate which steps of an amplicon workflow, excluding the choice of PCR primers and hypervariable gene region, introduce the largest uncertainties in amplicon data sets and, in turn, which steps require empirical evaluations of accuracy and precision to ensure robust sample analysis and scientific conclusions. In our workflow, the most significant sources of analytical variance were differences in biological/sampling replicates, differences in DNA extraction methods (for some community compositions), and the sample multiplexing strategy. We found that regardless of the community sampled, a baseline BCD of ~0.2 was present amongst sampling replicates (Fig. 1). Although this value is somewhat large (20% of the maximum theoretical dissimilarity), published estimates of dissimilarities inherent in biological/sampling replicates are rare, making it difficult to determine whether modifications to the sampling procedure employed here would reduce this uncertainty or if this is due to inherent environmental heterogeneity of protist communities. Regardless, this value provides a useful threshold for determining the significance of the uncertainties introduced at other steps in our workflow, and in interpreting downstream results in our data, where BCDs of ~0.2 or less should not be interpreted as biological signal distinguishable from analytical noise.

Previous studies demonstrated that both mechanical and chemical lysis procedures are needed prior to DNA extraction to accurately represent protist community composition (Yuan et

al. 2015; Djurhuus et al. 2017). Our comparisons of DNA extraction methods (Fig. 1), both of which employed enzymatic and mechanical (bead-beating) lysis procedures, showed that the composition of the sampled community can also cause discrepancies in diversity and community composition estimates from different DNA extraction methods. Specifically, the PC and PW methods produced comparable estimates of community composition for dinoflagellate-dominated communities, but diverged when diatom-dominated communities were targeted (Figs. 1 and S1). Thus, either conservative interpretations of community dissimilarities (BCDs > ~0.4 in our comparisons), or empirical comparisons of different DNA extraction methods employed across multiple sampled protist communities, are required for robust meta-analyses of amplicon data sets employing different DNA extraction methods. Our analysis also suggests that more rigorous cell lysis procedures prior to DNA extraction improve the accuracy of protist community composition estimates (Figs. 1 and S1).

We found that the choice of multiplexing strategy is a critical procedural consideration that requires rigorous validation for robust downstream analysis of amplicon data. In workflows employing 1-step PCR such as ours, the use of certain dual-indexed primers can inflate average BCDs amongst PCR/sequencing replicates as much as two-fold relative to BCDs amongst sampling replicates (Figs. 1 and 3). Recent studies targeting the 16S small subunit ribosomal DNA in human microbiomes and the mitochondrial 16S DNA in metazoans suggest that 2-step PCR yields greater accuracy and precision in amplicon data (Gohl et al. 2016; O'Donnell et al. 2016). Although we did not evaluate a 2-step PCR protocol for protists, we expect 2-step PCR to

enhance precision and accuracy in amplicon sequencing analysis and recommend validation of this multiplexing strategy for studies of marine protists.

If 1-step PCR is employed, thorough analysis of the uncertainty introduced by different indexed primers is critical. Failure to ensure reproducibility within the workflow should require conservative interpretations of downstream analyses and preclude incorporation of the data into meta-analyses. Clearly, a failure to ensure reproducibility across primer indices will also confound more general methodological comparisons such as empirical comparisons of different 18S primer sets and/or hypervariable regions (Bradley et al. 2016). Fortunately, the other PCR procedure modifications evaluated here, including small variations in thermal cycling routines, the inclusion of a mammal blocking primer, and the pooling of products from replicate PCRs, do not appear to significantly alter the precision or accuracy of results (Figs. 2, 3, and 5). Overall, our findings demonstrate high potential for standardization and meta-analyses of amplicon data sets assuming they meet certain procedural requirements, and provide a starting point for comparing the workflow(s) evaluated here with others.

Quantitative Potential of Relative Sequence Abundances

The high correlation between 18S gene copy numbers and cell biovolume across diverse protist lineages (Zhu et al. 2005; Godhe et al. 2008) justifies the use of relative sequence abundances as an ecologically and geochemically meaningful proxy of protist population and community dynamics. Amplicon sequencing data are thus used in many studies of marine protist

communities to assess the causes and consequences of community or population variations (Guidi et al. 2016; Needham and Fuhrman 2016; Berdjeb et al. 2018; Wang et al. 2018). Such analyses assume that relative sequence abundances are reasonably precise and accurate, despite a paucity of empirical estimates of their precision and accuracy available in the literature. Here we discuss the quantitative potential of relative sequence abundances of specific protist ASVs and lineages in complex environmental samples of marine protist communities.

All HTS data and especially relative sequence abundances are compositional. In other words, both biologically- and analytically-driven variations in the abundance of one ASV alter the relative abundances of all other ASVs in that sample or sequencing run. This, in turn, means that in any single sample, the relative abundance of an ASV may be skewed by the detection of spurious ASVs, a failure to detect biological ASVs, or a significant analytical bias for or against any ASV found within the community. Our analysis of precision in the May_PC and Aug_PC sampling replicates and the Aug1_PC PCR/sequencing replicates (Fig. 4) addressed many of these potential analytical issues. These analyses demonstrate that most of the abundant ASVs within a sample can be quantified with reasonable precision despite significant variations in sequencing depth and the number of ASVs detected. Precision is especially high if sampling/biological variability is ignored or reduced relative to that observed here (e.g., Fig. 4E and 4F). However, sampling/biological uncertainty can dramatically reduce the precision observed for particular ASVs. This was most obvious in the Aug_PC sampling replicates, where an ASV (also the only representative of its Class within those samples; classified as Division

Radiolaria, the blue bar at a rank abundance of 8 in Fig. 4D) was >1% abundant in one sampling replicate but absent from the other two replicates. Grouping ASVs to a coarser taxonomic rank (assuming there are multiple ASVs present within a taxonomic Class) can mask this biological variability and allow for more robust interpretations of the abundances of specific Classes.

The precision observed for protist Classes, particularly if biological/sampling variability is ignored, is comparable to that observed in HPLC determinations of phytoplankton pigment concentrations. HPLC pigment determinations are currently the best standardized and most rigorously validated method for quantitative evaluations of marine phytoplankton communities (Hooker et al. 2010). However, HPLC pigment determinations only resolve the photosynthetic component of the protist community, are associated with much greater taxonomic ambiguity than amplicon data, and only allow for the quantification of a few (~4 or 5) broad taxonomic groups (e.g., Catlett and Siegel 2018). Thus, precise quantification of ~30 protist Classes despite variations in indexed primers and PCR protocols used, sample-specific sequencing depth, and observed ASV richness, offers high quantitative potential for amplicon data if reasonable accuracy is also achieved.

Accuracy is difficult to measure empirically in amplicon analysis of natural marine protist communities. The analyses of accuracy in the AA samples demonstrate that significant analytical bias for or against the dominant ASV(s) in a sample significantly impacts the accuracy of all relative abundances in the sample due to the compositional nature of amplicon data (e.g., AA6). In the absence of such bias high accuracy can be obtained (e.g., AA7; Figs. 5 and 6). The

most likely sources of such bias are mismatches with the 3' end of the primer or unusually high or low amplicon GC content (Fig. 6; Bradley et al. 2016; Parada et al. 2016; Wear et al. 2018). We fit third order polynomials to the relationships between GC content and $\log_2$ fold-change shown in Figs. 6 and S4 to investigate to what degree PCR/sequencing bias could be explained by amplicon GC content. For individual AAs, R^2 values ranged from 0.27 (AA8) to 0.74 (AA1), and for global relationships across all eight AAs, R^2 was 0.46. This range in predictability demonstrates the non-linear nature of PCR/sequencing bias driven by complex interactions amongst multiple sources of bias, and suggests robust corrections of this bias would be difficult for more complex natural communities. Further work should thus be devoted to developing computational approaches to correct PCR/sequencing bias and more generally, to validating accuracy in amplicon methods applied to natural samples of protist communities.

However, our results provide a path for making qualitative assessments of the accuracy of protist community composition in particular samples. In natural samples of marine protist communities, GC content of each ASV can be estimated directly. Primer mismatches for each ASV can be estimated via alignments to reference sequences, despite the potential ambiguity in these alignments (e.g., Edgar 2017). Therefore, if one verifies that abundant (e.g., > ~0.1% abundance) ASVs within a sample do not exhibit high GC content and/or mismatches with the 3' ends of the PCR primers, it can be inferred with reasonable confidence that the relative abundances of the dominant ASVs are accurate in that sample. Overall, our assessments of accuracy and precision in this amplicon workflow suggest strong quantitative potential for

amplicon data if the procedural recommendations, method validation, and quality-control procedures outlined above are employed.

Benefits of Positive and Negative Controls in Amplicon Sequencing Analysis

The use of positive and negative controls has been demonstrated as a valuable means to validate and continuously quality-control amplicon sequencing analysis (Bradley et al. 2016; Gohl et al. 2016; Parada et al. 2016; Wear et al. 2018; Yeh et al. 2018). Despite the significant investment required to account for the many potential sources of uncertainty in amplicon data, the continued use of positive and negative controls is critical to ensure high-quality and reproducible sample analysis, safeguard against invalid scientific conclusions, and move the field toward a consensus on methodological best practices.

The use of negative controls is required for robust analyses of protist diversity and helps control for potential effects of contaminants, cross-talk, and other spurious ASVs in downstream analysis. In our analyses, the negative controls allowed us to identify a sequencing run where some samples appeared contaminated (Table 3). The negative controls also suggested that the largest source(s) of spurious ASVs in our workflow was via cross-sample contamination and/or tag-jumping (Tables 3 and S2). This information can enable systematic improvements in our workflow by modifying our multiplexing strategy and/or taking additional precautions to reduce contamination. Finally, negative controls allowed us to better assess the potential effects of spurious ASVs in downstream analyses, thus making cross-sample comparisons of diversity and

the presence/absence of specific ASVs more robust (see comparisons of ASV richness in *DNA Extraction Method Comparisons* above). While there is no current consensus on the best practices to account for spurious ASVs in amplicon analysis, bioinformatic methods are being developed to address this issue (Edgar 2016; Davis et al. 2018). As bioinformatic tools become more robust and generalized, negative controls will aid in the identification and removal of spurious ASVs (e.g., Davis et al. 2018).

Positive controls helped determine baseline estimates of analytical uncertainty in our workflow (Figs. 1, 3, 4-6) and ensured reproducibility across sequencing runs (Figs. 2, 3, 5), and thus will be useful in distinguishing true biological signal from analytical noise in future analyses of our data. Our continued use of mock community positive controls was also beneficial in our diagnosis of the lack of precision introduced by specific indexed primers in our workflow; this issue came to light after eight MiSeq runs, and the inclusion of multiple positive controls with each run was critical to ruling out aberrant sequencing runs (Figs. 2 and 3; see Yeh et al. 2018). Finally, our results provide baseline uncertainty estimates for other investigators seeking to compare the magnitude of uncertainty in their workflow and/or uncertainties for other 18S gene primer sets and hypervariable regions. This is currently made difficult by the lack of published uncertainty estimates for amplicon workflows, but may prove critical in moving the field toward a consensus on methodological best practices and standardizing amplicon sequencing assessments of marine protist communities.

Conclusions

We compared methodological options and characterized the accuracy and precision in an amplicon sequencing workflow for marine protist communities. Notable biases and uncertainties were introduced by sampling, different DNA extraction methods for certain community states, and by PCR bias for specific ASVs. We found that the choice of multiplexing strategy is critical as certain indexed primers significantly inflated BCDs amongst PCR replicates. Despite these differences, many methodological variations were comparable with one another, which bodes well for standardization and meta-analyses of amplicon data. Our analysis demonstrates the importance of validation and continued quality-control of amplicon methods. Finally, we show that quality-controlled amplicon methods have high quantitative potential for determining the diversity and composition of marine protist communities and the relative abundances of specific ASVs and lineages, though further work is needed to validate the accuracy in these methods.

Acknowledgements

We thank Emma Wear and Anni Djurhuus for advice on sample preparation and analyses, Thomas Lankiewicz for assistance with bench work, Alexis Pasulka for helpful comments regarding mock community composition, Alyson Santoro for providing laboratory space, and the Plumes and Blooms team (especially Nathalie Guillocheau and Sarah Amiri) for assistance with field sampling and supplying ancillary environmental data. We thank Alexandra Worden at MBARI, Avery Tatters at USC, and Holly Bowers at Moss Landing Marine Lab for

providing algal isolates for culturing. This research was supported by the National Aeronautics and Space Administration Biodiversity and Ecological Forecasting program (Grant NNX14AR62A), the Bureau of Ocean and Energy Management Ecosystem Studies program (BOEM award MC15AC00006), and NOAA in support of the Santa Barbara Channel Biodiversity Observation Network. DC is supported on a NASA Earth and Space Science Fellowship (Grant NNX16AO44HS02). Most DNA sequencing was carried out by the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation (Grant 1S10OD010786-01). We acknowledge our general use of, and the assistance of the staff members from, the DNA Technologies Core at the UC Davis Genome Center, and the Biological NanoStructures Lab within the California NanoSystems Institute supported by UC Santa Barbara and the University of California, Office of the President, for all sequencing analysis. Raw DNA sequencing data are publicly available in the National Center for Biotechnology Information's Sequence Read Archive under accession number PRJNA532583.

References

Adl, S. M., A. G. Simpson, C. E. Lane, and others. 2012. The revised classification of eukaryotes. Journal of eukaryotic microbiology **59**: 429–514.

Aird, D., M. G. Ross, W.-S. Chen, and others. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome biology **12**: R18.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. Journal of molecular biology **215**: 403–410.

Amaral-Zettler, L. A., E. A. McCliment, H. W. Ducklow, and S. M. Huse. 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PloS one **4**: e6372.

Bálint, M., M. Bahram, A. M. Eren, and others. 2016. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. FEMS microbiology reviews **40**: 686–700.

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

Bokulich, N. A., S. Subramanian, J. J. Faith, D. Gevers, J. I. Gordon, R. Knight, D. A. Mills, and J. G. Caporaso. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature methods **10**: 57.

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

Callahan, B. J., P. J. McMurdie, and S. P. Holmes. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME journal **11**: 2639.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nature methods **13**: 581.

Caputi, L., Q. Carradec, D. Eveillard, and others. 2018. Community-Level Responses to Iron Availability in Open Ocean Planktonic Ecosystems. Global Biogeochemical Cycles.

Catlett, D., and D. Siegel. 2018. Phytoplankton pigment communities can be modeled using unique relationships with spectral absorption signatures in a dynamic coastal environment. Journal of Geophysical Research: Oceans **123**: 246–264.

Countway, P. D., R. J. Gast, M. R. Dennett, P. Savai, J. M. Rose, and D. A. Caron. 2007. Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). Environmental Microbiology **9**: 1219–1232.

Davis, N. M., D. M. Proctor, S. P. Holmes, D. A. Relman, and B. J. Callahan. 2018. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome **6**: 226.

De Vargas, C., S. Audic, N. Henry, and others. 2015. Eukaryotic plankton diversity in the sunlit ocean. Science **348**: 1261605.

Djurhuus, A., J. Port, C. J. Closek, and others. 2017. Evaluation of filtration and DNA extraction methods for environmental DNA biodiversity assessments across multiple trophic levels. Frontiers in Marine Science **4**: 314.

Edgar, R. C. 2016. UNCROSS: Filtering of high-frequency cross-talk in 16S amplicon reads. bioRxiv 088666.

Edgar, R. C. 2017. UNBIAS: An attempt to correct abundance bias in 16S sequencing, with limited success. bioRxiv 124149.

Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. science **281**: 237–240.

Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field. 1990. Genetic diversity in Sargasso Sea bacterioplankton. Nature **345**: 60.

Godhe, A., M. E. Asplund, K. Härnström, V. Saravanan, A. Tyagi, and I. Karunasagar. 2008. Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. Appl. Environ. Microbiol. **74**: 7174–7182.

Gohl, D. M., P. Vangay, J. Garbe, and others. 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nature biotechnology **34**: 942.

Guidi, L., S. Chaffron, L. Bittner, and others. 2016. Plankton networks driving carbon export in the oligotrophic ocean. Nature **532**: 465.

Guillard, R. R. 1975. Culture of phytoplankton for feeding marine invertebrates, p. 29–60. *In* Culture of marine invertebrate animals. Springer.

Guillou, L., D. Bachar, S. Audic, and others. 2012. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic acids research **41**: D597–D604.

Hooker, S. B., C. S. Thomas, L. Van Heukelem, and others. 2010. The fourth SeaWiFS HPLC analysis round-Robin experiment (SeaHARRE-4).

Hu, S. K., Z. Liu, A. A. Lie, and others. 2015. Estimating protistan diversity using high-throughput sequencing. Journal of Eukaryotic Microbiology **62**: 688–693.

Hugerth, L. W., and A. F. Andersson. 2017. Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. Frontiers in microbiology **8**: 1561.

Hugerth, L. W., E. E. Muller, Y. O. Hu, L. A. Lebrun, H. Roume, D. Lundin, P. Wilmes, and A. F. Andersson. 2014. Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. PLoS One **9**: e95567.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster. 2007. MEGAN analysis of metagenomic data. Genome research **17**: 377–386.

Keeling, P. J., and J. del Campo. 2017. Marine protists are not just big bacteria. Current biology **27**: R541–R549.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss. 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.

Laursen, M. F., M. D. Dalgaard, and M. I. Bahl. 2017. Genomic GC-content affects the accuracy of 16S rRNA gene sequencing based microbial profiling due to PCR bias. Frontiers in microbiology **8**: 1934.

Lima-Mendez, G., K. Faust, N. Henry, and others. 2015. Determinants of community structure in the global plankton interactome. Science **348**: 1262073.

McMurdie, P. J., and S. Holmes. 2014. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS computational biology **10**: e1003531.

Minich, J. J., J. G. Sanders, A. Amir, G. Humphrey, J. A. Gilbert, and R. Knight. 2019. Quantifying and understanding well-to-well contamination in microbiome research. mSystems **4**: e00186-19.

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

O'Donnell, J. L., R. P. Kelly, N. C. Lowell, and J. A. Port. 2016. Indexed PCR primers induce template-specific bias in large-scale DNA sequencing studies. PloS one **11**: e0148698.

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

Polz, M. F., and C. M. Cavanaugh. 1998. Bias in template-to-product ratios in multitemplate PCR. Appl. Environ. Microbiol. **64**: 3724–3730.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research **41**: D590–D596.

Quere, C. L., S. P. Harrison, I. Colin Prentice, and others. 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Global Change Biology **11**: 2016–2040.

Schnell, I. B., K. Bohmann, and M. T. P. Gilbert. 2015. Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. Molecular ecology resources **15**: 1289–1303.

Sherr, E., and B. Sherr. 1994. Bacterivory and herbivory: key roles of phagotrophic protists in pelagic food webs. Microbial Ecology **28**: 223–235.

Stoeck, T., D. Bass, M. Nebel, R. Christen, M. D. Jones, H. BREINER, and T. A. Richards. 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Molecular ecology **19**: 21–31.

Suzuki, M. T., and S. J. Giovannoni. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl. Environ. Microbiol. **62**: 625–630.

Thompson, L. R., J. G. Sanders, D. McDonald, and others. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature **551**: 457.

Vestheim, H., and S. N. Jarman. 2008. Blocking primers to enhance PCR amplification of rare sequences in mixed samples–a case study on prey DNA in Antarctic krill stomachs. Frontiers in zoology **5**: 12.

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

Wang, S., Y. Lin, S. Gifford, R. Eveleth, and N. Cassar. 2018. Linking patterns of net community production and marine microbial community structure in the western North Atlantic. The ISME journal 1.

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

Worden, A. Z., M. J. Follows, S. J. Giovannoni, S. Wilken, A. E. Zimmerman, and P. J. Keeling. 2015. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. Science **347**: 1257594.

Yeh, Y.-C., D. M. Needham, E. T. Sieradzki, and J. A. Fuhrman. 2018. Taxon Disappearance from Microbiome Analysis Reinforces the Value of Mock Communities as a Standard in Every Sequencing Run. MSystems **3**: e00023-18.

Yuan, J., M. Li, and S. Lin. 2015. An improved DNA extraction method for efficient and quantitative recovery of phytoplankton diversity in natural assemblages. PloS one **10**: e0133060.

Zhu, F., R. Massana, F. Not, D. Marie, and D. Vaulot. 2005. Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. FEMS microbiology ecology **52**: 79–92.

Figure Captions

Figure 1. Results of DNA extraction method comparisons. The DNA extraction methods compared were the DNeasy PowerWater DNA extraction kit (PW) and a phenol-chloroform method (PC; see text for details). (A) and (B) show results of hierarchical cluster analyses using Bray-Curtis dissimilarity (BCD) and the average linkage method on the three replicated environmental samples from May and August cruises in the Santa Barbara Channel. Analyses consider (A) all ASVs and (B) ASVs comprising >1% of sequences in at least one sample. (C) shows the magnitude of BCD (mean ± standard deviation) amongst biological replicates extracted with a single method (within-PC and within-PW; n = 3 for each bar), amongst the same samples extracted by different DNA extraction methods (PW vs. PC; n = 3 for each bar), and that present between May and August samples for the PC and PW method. (D) considers ASVs comprising >1% of sequence reads in at least 1 sample from the May cruise, and shows the mean relative abundances of the dominant Classes in May samples extracted with either the PC (red) or PW (blue) method. Individual ASVs within each Class are delineated with a horizontal black line.

Figure 2. Principal coordinate analyses of (A, B, C, D) all PCR/sequencing replicates of the Aug1_PC sample considering (A, B) all ASVs and (C, D) only ASVs that account for more than 1% of sequence reads in at least one PCR/sequencing replicate, and of (E, F) all PCR/sequencing replicates of AA1 considering only the 22 target ASVs included in mock communities with (A, C, E) and without (B, D, F) the SA501 and SA505 indexed forward primers. Each point represents a PCR/sequencing replicate, it's color shows the forward indexed primer used in PCR/sequencing, and it's shape shows the PCR protocol used. The percent variance explained by each axis of the ordination is printed next to the axis name.

Figure 3. Mean Bray-Curtis dissimilarities (BCD) ± standard deviation observed across all PCR/sequencing replicates of (A) AA1 and (B) Aug1_PC in individual sequencing runs, and in all sequencing runs ("All Runs"), when samples amplified with the primers SA501 and SA505 are (red) or are not (blue) considered. In (A), bars are omitted where replicates were not sequenced with one of the primer groups, and error bars are omitted if less than 3 replicates were available. (C) shows mean BCD ± standard deviation for all pairs of environmental samples amplified and sequenced in duplicate across different sequencing runs. Pairs of duplicate samples are grouped according to whether they were sequenced with SA501 and another forward indexed primer (except for SA505; "With SA501"), with SA505 and another forward indexed primer (except for SA501; "With SA505"), or with any combination of forward indexed primers

excluding SA501 and SA505. The number of duplicate pairs used in each category is denoted in the plot.

Figure 4. Coefficients of variation (CV) of relative abundances in the environmental samples used in DNA extraction method comparisons. (A-C) Individual ASVs and (D-F) ASVs agglomerated according to their taxonomic Class are sorted from left to right according to their mean relative abundance in each set of samples (May_PC, Aug_PC, or Aug1_PC). For (A, B, D, E) n = 3, and for (C, F) n = 9. Bars are colored according to taxonomic Division.

Figure 5. Accuracy estimates for AAs. (A) linear and (B) $\log_{10}$-transformed expected vs. observed relative abundances. Solid lines show a 1:1 relationship, and dashed lines are the lines of best fit. All observed abundances are a mean of PCR/sequencing replicates (n = 3). The colors of points correspond to taxonomic Divisions and the shapes correspond to the AA sample. (C-F) show indices of accuracy in observed vs. expected abundances calculated for each AA, including the coefficient of determination in (C) linear and (D) $\log_{10}$-transformed space, (E) Spearman's rank correlation coefficient, and (F) root mean square error. In (C-F), bar colors correspond to the AA sample, and bars show the mean ± standard deviation of the statistic of interest for three PCR/sequencing replicates. Samples amplified and sequenced with the SA501 or SA505 primers were omitted.

Figure 6. ASV-specific bias in AA samples. The base 2 logarithm of observed divided by expected relative abundances is shown for each of the 22 amplicons included in AAs. (A, C, E, G) show the bias for each ASV (ASV numbers on the x-axis were assigned arbitrarily and correspond to those used in Table 1) and (B, D, F, H) show the relationship of bias with amplicon GC content. Bars and points are colored according to taxonomic Division, and represent an average fold-change ± standard deviation calculated across 3 PCR/sequencing replicates. Dashed lines show a 2-fold change. Only samples amplified with the Standard protocol and without the SA501 or SA505 primers are considered.

Table 1. Mock community species compositions.

Clone ID	ASV	Lowest Initial Taxonomic Assignment (Expected Genera)	ASV V9 GC Content	Most Similar V9 ASV (% V9 Similarity)	Expected Relative Abundance							
					AA1	AA2	AA3	AA4	AA5	AA6	AA7	AA8
ChaspCC5	1	*Chaetoceros debilis* (*Chaetoceros*)	0.464	6 (82.0%)	0.046	0.181	0.010	0.020	0.081	0.015	0.005	0.005
OlilutC1	2	*Olisthodiscus luteus* (*Olisthodiscus*)	0.433	3 (85.0%)	0.045	0.010	0.030	0.020	0.089	0.010	0.030	0.010
HetakaC2	3	*Chattonellaceae* (*Heterosigma*)	0.424	2 (85.0%)	0.045	0.030	0.010	0.059	0.000	0.020	0.005	0.005
LepdanC3	4	*Caecitellus* (Unknown**)	0.538	7 (81.1%)	0.048	0.011	0.011	0.011	0.000	0.011	0.011	0.011
PsefraC2	5	*Bacillariophyta* (*Pseudo-nitzschia*)	0.425	6 (95.3%)	0.046	0.200	0.010	0.010	0.000	0.005	0.040	0.020
PsemulC5	6	*Bacillariophyta* (*Pseudo-nitzschia*)	0.433	5 (95.3%)	0.045	0.150	0.020	0.010	0.000	0.040	0.005	0.010
CosspC2	7	*Coscinodiscophyceae* (*Coscinodiscus*)	0.520	6 (85.8%)	0.045	0.030	0.010	0.040	0.000	0.050	0.010	0.005
PelcalC2	8	*Pelagophyceae* (*Pelagomonas*)	0.480	3 (80.5%)	0.045	0.030	0.010	0.069	0.079	0.030	0.010	0.005
ChaspBC7	9	*Chaetoceros debilis* (*Chaetoceros*)	0.438	6 (85.3%)	0.045	0.089	0.040	0.020	0.080	0.010	0.010	0.005
OlilutC3	10	*Cercozoa* (Unknown**)	0.527	12 (75.7%)	0.045	0.020	0.010	0.010	0.000	0.010	0.010	0.118
GepoceC2*	11	*Haptophyceae* (*Gephyrocapsa*)	0.603	12 (99.2%)	0.045	0.010	0.030	0.020	0.080	0.379	0.020	0.020
EmihuxC5	12	*Haptophyceae* (*Emiliania*)	0.595	11 (99.2%)	0.045	0.020	0.010	0.110	0.080	0.050	0.010	0.020
PhagloC2	13	*Haptophyceae* (*Phaeocystis*)	0.573	12 (91.6%)	0.045	0.040	0.010	0.139	0.089	0.040	0.025	0.020
CocbraC3	14	*Haptophyceae* (*Coccolithus*)	0.557	11 (93.9%)	0.045	0.010	0.020	0.010	0.090	0.010	0.010	0.010
RhosalC3	15	*Cryptophyta* (*Rhodomonas*)	0.421	17 (83.2%)	0.046	0.020	0.040	0.101	0.081	0.031	0.202	0.005
MicpusC2	16	*Mamiellaceae* (*Micromonas*)	0.519	17 (86.3%)	0.046	0.040	0.010	0.050	0.081	0.005	0.010	0.422
BatpraC4	17	*Bathycoccus prasinos* (*Bathycoccus*)	0.496	18 (89.1%)	0.046	0.010	0.010	0.050	0.081	0.081	0.005	0.121
OstlucC3	18	*Ostreococcus* (*Ostreococcus*)	0.465	17 (89.1%)	0.046	0.020	0.030	0.030	0.091	0.061	0.071	0.010
GymakaC2	19	*Gymnodiniaceae* (*Gymnodinium*)	0.496	20 (88.1%)	0.045	0.010	0.010	0.040	0.000	0.070	0.030	0.060
PromicC8*	20	*Dinophyceae* (*Prorocentrum*)	0.468	21 (89.7%)	0.045	0.010	0.219	0.070	0.000	0.030	0.030	0.020
ChaspBC10	21	*Alexandrium* (*Alexandrium*)	0.397	20 (89.7%)	0.046	0.020	0.190	0.100	0.000	0.005	0.448	0.010
LinpolC3	22	*Gonyaulacales* (*Lingulodinium*)	0.416	20 (87.3%)	0.045	0.040	0.260	0.010	0.000	0.040	0.005	0.090

*These species each exhibit one mismatch with the forward primer, 1391F. Mismatches were observed at positions (5' to 3') 8 and 7 on the 1391F primer for ASV 11 and 20, respectively.

**Amplicons with unknown expected genera were suspected to arise from contamination of phytoplankton cultures.

Table 2. Terms and abbreviations and their use in the present manuscript.

	Term	Definition
General Terms and Abbreviations for Method and Results Descriptions	AA	Amplicon assemblage, a term used to describe samples of artificial protist communities composed of full-length amplicons isolated from individual protist species. When followed by a number from 1-8, denotes a mock community of a particular composition (Table 1).
	PC	Samples where genomic DNA was extracted using a "custom" phenol-chloroform method.
	PW	Samples where genomic DNA was extracted using the DNeasy PowerWater DNA extraction kit.
	PCRx1	Samples that were amplified in a single PCR. All other samples were amplified in triplicate PCRs which were then pooled prior to PCR product purification and quantitation, sample pooling, and sequencing.
	EMP	Samples that were amplified with a modified thermal cycling routine relative to all other samples (see text for detail).
	MB	Samples that were amplified in the presence of a mammal blocking primer.
	Taxonomy array	An array of ASV sequences and their corresponding taxonomic assignments generated by a single taxonomic assignment algorithm applied to a single reference database, or a combination of assignment algorithms and reference databases employing a single taxonomy.
	Taxonomy	A general term used to describe a taxonomic naming and ranking framework employed by one of the reference databases considered in the present study.
ASVs Detected in AA and Blank Samples	Spurious ASV	Any false positive ASV detected in positive (AAs) or negative (blanks) control samples.
	Contaminant ASV	Spurious ASVs detected in any non-control sample amplified and/or sequenced alongside the control sample. For spurious ASVs detected in blank samples, these also include target ASVs.
	Artifact ASV	Spurious ASVs detected in AA samples that were >95% similar to a target ASV sequence in that AA and not detected in any environmental samples amplified and/or sequenced alongside the AA sample.
	Unknown ASV	Spurious ASVs that could not be classified as contaminant or artifact ASVs.
	Target ASV	One of the 22 ASVs intentionally included in AA samples (see Table 1).

Table 3. Target and spurious ASVs detected in selected positive and negative control samples.

Sample Type	Sample Name	Total ASVs	Target ASVs Detected	Spurious ASVs*	Contaminant ASVs	Unknown ASVs	Artifact ASVs
AA	AA1A_Run1	24	22	2 (8.05×10^{-5})	1 (5.75×10^{-5})	1 (2.30×10^{-5})	0
	AA1A_Run2	28	22	6 (5.31×10^{-4})	6 (5.31×10^{-4})	0	0
	AA1A_Run3	27	22	5 (4.55×10^{-3})	5 (4.55×10^{-3})	0	0
	AA1A_Run4	24	22	2 (2.77×10^{-4})	2 (2.77×10^{-4})	0	0
	AA1A_Run5	26	22	4 (5.10×10^{-4})	4 (5.10×10^{-4})	0	0
	AA1A_Run6	24	22	2 (7.58×10^{-5})	2 (7.58×10^{-5})	0	0
	AA1A_Run7	22	22	0	0	0	0
	AA1A_Run8	22	21	1 (1.56×10^{-4})	0	1 (1.56×10^{-4})	0
	AA1A_Run9	23	22	1 (1.74×10^{-5})	0	1 (1.74×10^{-5})	0
	AA2lowB	76	22	54 (2.82×10^{-2})	0	1 (4.30×10^{-4})	53 (2.77×10^{-2})
	AA5C	16	13**	3 (8.11×10^{-3})	1 (8.01×10^{-3})	0	2 (1.00×10^{-4})
	MB_AA1A_Run8	40	21	19 (6.45×10^{-3})	14 (5.87×10^{-3})	5 (5.75×10^{-4})	0
PCR Blanks	PCR_blank_Run1	NA***	NA	7	3 (0.220)	4 (0.780)	NA
	PCR_blank_Run2	NA	NA	11	5 (0.301)	6 (0.699)	NA
	PCR_blank_Run3	NA	NA	15	6 (0.796)	9 (0.204)	NA
	PCR_blank_Run4	NA	NA	4	3 (0.900)	1 (0.100)	NA
	PCR_blank_Run5	NA	NA	4	4 (1)	0	NA
	PCR_blank_Run6	NA	NA	0	0	0	NA
	PCR_blank_Run7	NA	NA	2	0	2 (1)	NA
	PCR_blank_Run8	NA	NA	35	18 (0.519)	17 (0.481)	NA
	PCR_blank_Run9	NA	NA	1	1 (1)	0	NA
DNA Extraction Blanks	PW_ExtractionBlank1_Run1	NA	NA	6	5 (0.985)	1 (0.015)	NA
	PW_ExtractionBlank2_Run1	NA	NA	6	2 (0.396)	4 (0.604)	NA
	PW_ExtractionBlank3_Run1	NA	NA	15	5 (0.599)	10 (0.401)	NA
	PC_ExtractionBlank1_Run1	NA	NA	11	2 (0.060)	9 (0.940)	NA
	PC_ExtractionBlank2_Run1	NA	NA	10	2 (0.235)	8 (0.765)	NA
	PC_ExtractionBlank3_Run1	NA	NA	22	3 (0.200)	19 (0.800)	NA

*Values in the final 4 columns are the number of ASVs detected in each classification, while values in parentheses are the summed relative abundances (proportions, not percentages) of all ASVs of a given classification found within each sample. Complete descriptions of the ASV classifications are given in *Accuracy of ASV Detection in AA and Blank Samples* and in Table 2.

**Only 12 target ASVs were expected in AA5 samples. All other AA samples included 22 target ASVs. The additional target ASV in AA5C represents a contaminant from the other AA samples, and was found at a relative abundance of 1.48×10^{-4}. Two unexpected/contaminant target ASVs were also detected in AA5B, accounting for relative abundances of 1.73×10^{-4} and 1.13×10^{-4} (see Table S2).

***For all negative controls, Total ASVs are equivalent to Spurious ASVs, and our definition of Artifact ASVs precludes their detection. We thus use NA in place of the Total ASVs and Artifact ASVs.