Received: 6 October 2021 | Revised: 17 March 2022 | Accepted: 11 April 2022

DOI: 10.1111/1755-0998.13623

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Short- and long-read metabarcoding of the eukaryotic rRNA operon: Evaluation of primers and comparison to shotgun metagenomics sequencing

Meike A. C. Latz[1,2] | Vesna Grujcic[1] | Sonia Brugel[3] | Jenny Lycken[4] | Uwe John[5,6] | Bengt Karlson[4] | Agneta Andersson[3] | Anders F. Andersson[1]

[1]Department of Gene Technology, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden

[2]Department of Plant and Environmental Sciences, University of Copenhagen, Frederiksberg C, Denmark

[3]Department of Ecology and Environmental Sciences, Umeå University, Umeå, Sweden

[4]Oceanographic Research, Swedish Meteorological and Hydrological Institute, Gothenburg, Sweden

[5]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany

[6]Helmholtz Institute for Functional Marine Biodiversity, Oldenburg, Germany

Correspondence
Meike A. C. Latz and Anders F. Andersson, Department of Gene Technology, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden.
Emails: meike.latz@scilifelab.se (M.A.C.L); anders.andersson@scilifelab.se (A.F.A.)

Funding information
Naturvårdsverket, Grant/Award Number: NV-03728-17; Villum Fonden, Grant/Award Number: 34442

Handling Editor: Naiara Rodriguez-Ezpeleta

Abstract

High-throughput sequencing-based analysis of microbial diversity has evolved vastly over the last decade. Currently, the go-to method for studying microbial eukaryotes is short-read metabarcoding of variable regions of the 18S rRNA gene with <500 bp amplicons. However, there is a growing interest in applying long-read sequencing of amplicons covering the rRNA operon for improving taxonomic resolution. For both methods, the choice of primers is crucial. It determines if community members are covered, if they can be identified at a satisfactory taxonomic level, and if the obtained community profile is representative. Here, we designed new primers targeting 18S and 28S rRNA based on 177,934 and 21,072 database sequences, respectively. The primers were evaluated *in silico* along with published primers on reference sequence databases and marine metagenomics data sets. We further evaluated a subset of the primers for short- and long-read sequencing on environmental samples *in vitro* and compared the obtained community profile with primer-unbiased metagenomic sequencing. Of the short-read pairs, a new V6-V8 pair and the V4_Balzano pair used with a simplified PCR protocol provided good results *in silico* and *in vitro*. Fewer differences were observed between the long-read primer pairs. The long-read amplicons and ITS1 alone provided higher taxonomic resolution than V4. Together, our results represent a reference and guide for selection of robust primers for research on and environmental monitoring of microbial eukaryotes.

KEYWORDS
marine plankton, microbial eukaryotes, PacBio long-read sequencing, primer design, rRNA operon, metabarcoding

1 | INTRODUCTION

High-throughput sequencing of taxonomic marker genes – metabarcoding – is a widely used tool for biodiversity surveys and ecological studies. The method is independent of distinguishable morphological features; consequently, it has transformed the field of microbial ecology and our understanding of environmental microbial communities. Numerous metabarcoding surveys have explored the hidden

diversity of prokaryotes (Bolhuis & Stal, 2011; Yarza et al., 2014). Microbial eukaryotes, however, and eukaryotic plankton in particular, are still far from routinely studied with this approach, despite their fundamental roles in the global ecosystem (Hu et al., 2016; Massana et al., 2015; Pawlowski et al., 2016; Santoferrara et al., 2020; Taib et al., 2013).

Planktonic single-celled eukaryotes play central roles in aquatic ecosystems and in global biogeochemical cycles, through primary production, transfer of organic material to higher trophic levels and sequestration of carbon to the deep ocean (Barton et al., 2013; Carradec et al., 2018; Worden et al., 2015). Some taxa (e.g., *Pseudo-nitzschia* spp. and *Alexandrium* spp.) can, however, also produce toxins and cause harmful algal blooms, threatening marine life and human health (Karlson et al., 2021; Lewitus et al., 2012). The study of eukaryotic plankton is thus central for understanding aquatic ecosystems; for this undertaking metabarcoding offers a time and cost-efficient method for gaining insight into their diversity and biogeography.

The method, however, is not without its limitations and challenges. The 18S ribosomal RNA (rRNA) gene is a commonly used marker, due to its conserved regions separated by highly variable regions. In theory, this allows to both universally target eukaryotes and discriminate between closely related organisms. However, no primers identified to date equally target all eukaryotic organisms, and despite the abundance of studies that have attempted to develop "universal" eukaryotic primers for the 18S rRNA gene (Amaral-Zettler et al., 2009; Hadziavdic et al., 2014; Hugerth, Muller, et al., 2014), there is no consensus in the research community for one primer pair. Within the 18S rRNA gene, different variable regions have been targeted, for example, V1–V2 and V3 (Medinger et al., 2010; Mohrbeck et al., 2015), however, V4 (Balzano et al., 2015; Pagenkopp Lohan et al., 2016; Piredda et al., 2017; Stoeck et al., 2010) and V9 (Bradley et al., 2016; de Vargas et al., 2015; Stoeck et al., 2010) have been most commonly used for eukaryotic plankton.

A few recent studies have compared the suitability of the variable regions for metabarcoding of eukaryotes (Hadziavdic et al., 2014; Hugerth, Muller, et al., 2014) and eukaryotic plankton in specific (Tanabe et al., 2016). However, all limiting factors have not always been taken into consideration, for example a primer pair with broad taxonomic coverage designed by Hugerth, Muller, et al. (2014) generates amplicons too long for merging forward and reverse illumina reads. Short-read metabarcoding using Illumina sequencing poses restrictions on the length of the barcode region. As a result, the taxonomy of amplicon sequences can often not be resolved beyond the genus level (Hugerth, Muller, et al., 2014), since closely related species regularly have identical barcode regions. The advancement of third generation sequencing, or long-read sequencing, through platforms from Pacific Bioscience (PacBio) and Oxford Nanopore technologies, enables sequence recovery of >10 kb (Amarasinghe et al., 2020). Disadvantages include higher costs and typically higher error rates compared to Illumina (~15% compared with ~0.1%). However, these errors tend to be random and by sequencing the same amplicon several times using circular consensus sequencing (CCS) error rates are considerably lowered (Jamy et al., 2020; Larsen et al., 2014; Wenger et al., 2019; Westbrook et al., 2015). Long-read sequencing allows sequencing of amplicons covering a large fraction of the rRNA operon. The inclusion of the fast-evolving internal transcribed spacers (ITS1 and ITS2) and the hypervariable D1D2 region of the 28S rRNA gene allows detailed (species-level) taxonomic identification when closely related taxa are present in the reference database, and higher-level taxonomic placement using the rRNA genes in other cases (Orr et al., 2018; Tedersoo & Anslan, 2019). The long sequences of 18S and 28S rRNA genes also allow reconstruction of well resolved phylogenetic trees which facilitates evolutionary analysis of uncultivated organisms (Jamy et al., 2020). However, to our knowledge, no extensive evaluation of PCR primers for rRNA operon sequencing has been published so far.

The aim of our study was to evaluate existing and newly designed primers for short- and long-read metabarcoding *in silico* and *in vitro* to guide the selection of robust primers for research and environmental monitoring of microbial eukaryotes.

2 | MATERIALS AND METHODS

2.1 | Primer design and tests *in silico*

2.1.1 | Degenerate primer design

Primers were designed as described previously (Hugerth, Wefer, et al., 2014) with small modifications. In short, sequences of the eukaryotic small subunit (SSU) rRNA gene were downloaded from PR2 4.12.0 (Guillou et al., 2013). SSU (138) and Large subunit (LSU) (132) sequences were downloaded from SILVA (Quast et al., 2013). Sequences in each database were sorted by length and clustered at 97% identity using UCLUST tools (Edgar, 2010). Centroid sequences were aligned using MOTHUR ALIGN.SEQ (v.1.42.3) (Schloss, 2020) and reference alignments of SILVA. The reference files were created as previously described (https://mothur.org/blog/2020/SILVA-v138-reference-files/), using eukaryotic sequences only and covering full genes. The alignment was trimmed with TRIMALIGNMENT and degenerate primers generated at every alignment position using DEGEPRIME (Hugerth, Wefer, et al., 2014).

We designed primers targeting highly conserved rRNA gene regions (Tables 1 and 2) (Figure 1a) on nonredundant 18S (Figure 1b) and 28 rRNA gene sequences (Figure 1c). Information on clustering and alignment of the databases is provided in Table S1. Primers were selected to evenly match sequences across taxonomic groups. In addition, general considerations for primer design were applied: length (≥18 bp), GC content (30%–80%), and melting temperature (>45°C).

2.1.2 | Generation of rRNA databases from marine and brackish metagenomics contigs

SSU and LSU rRNA sequences in primary metagenomic contigs from the "protistan" filter size fraction (0.8–5.0 μm) from the Tara Oceans project (Tully et al., 2018), and contigs from samples (size

TABLE 1 Sequences of 18S rRNA primers

Name	Sequence (5'–3')	Length	Degen eracy	GC (%)	Tm[a]	Position[b]	Primer_ID	References
V4_Balzano_F	CCAGCASCYGCGGTAATTCC	20	4	62.5	60.1	565	V4F	Stoeck et al. (2010)
V4_Balzano_R	ACTTTCGTTCTTGATYRR	18	8	36.1	46.4	967	V4RB	Balzano et al. (2015), adapted from Stoeck et al. (2010)
V4_Bråte_F	GGCAAGTCTGGTGCCAG	17	0	64.7	56.5	552	3NDf	Cavalier-Smith et al. (2009)
V4_Bråte_R	ACGGTATCTRATCRTCTTCG	20	4	45	51.4	990	V4_euk_R2	Cavalier-Smith et al. (2009)
V4_Hugerth_F	CGGTAAYTCCAGCTCYAV	18	12	53.7	52	575	574*F	Hugerth, Muller, et al. (2014)
V4_Hugerth_R	CCGTCAATTHCTTYAART	18	12	35.2	45.4	1133	1132R	Hugerth, Muller, et al. (2014)
V4_new_F	GCCAGCAVCYGCGGTAAYT	19	12	61.4	60.3	564	564F	This study, adapted from Hugerth, Muller, et al. (2014)
V4_new_R	GGTATCTRATCVYCTTCG	18	12	48.1	48.2	990	990R	This study
V4_Piredda_F	CCAGCASCYGCGGTAATTCC	20	4	62.5	60.1	565	V418SNextFor	Piredda et al. (2017), adapted from Stoeck et al. (2010)
V4_Piredda_R	ACTTTCGTTCTTGATYRATGA	21	4	33.3	49.7	961	V418SNextRev	Piredda et al. (2017), adapted from Stoeck et al. (2010)
V4_Stoeck_F	CCAGCASCYGCGGTAATTCC	20	4	62.5	60.1	565	TAReuk454FWD1	Stoeck et al. (2010)
V4_Stoeck_R	ACTTTCGTTCTTGATYRA	18	4	33.3	45.9	964	TAReukREV3	Stoeck et al. (2010)
V6-V8_new_F	AATTYGAHTCAACRCGGG	18	12	46.3	51.5	1183	1183F	This study
V6-V8_new_R	CGACRGGMGGTGTGBACA	18	12	64.8	59.4	1625	1625R	This study
V9_Amazet_F	CCCTGCCHTTTGTACACAC	19	3	54.4	54.6	1628	1389F	Amaral-Zettler et al. (2009)
V9_Amazet_R	CCTTCYGCAGGTTCACCTAC	20	2	57.5	56.6	1774	1510R	Amaral-Zettler et al. (2009)
V9_EMP_F	GTACACACCGCCCGTC	16	0	68.8	56.1	1629	Euk_1391f	Based on Stoeck et al. (2010)
V9_EMP_R	TGATCCTTCTGCAGGTTCACCTAC	24	0	50	58.4	1774	EukBr	Based on Amaral-Zettler et al. (2009)
V9_Piredda_F	TTGTACACACCGCCCGTCGC	20	0	65	62.9	1627	V918SNext. For	Stoeck et al. (2010)
V9_Piredda_R	CCTTCYGCAGGTTCACCTAC	20	2	57.5	56.6	1774	V918SNext. Rev	Piredda et al. (2017), adapted from Stoeck et al. (2010)

[a]Melting temperature Tm depicts the average for degenerate primers and was calculated with IDT OligoAnalyser.

[b]Position of the first (5' end) base of the primer on the *S. cerevisiae* 18S rRNA gene (GeneBank: Z75578.1).

TABLE 2 Sequences of 28S rRNA primers

Name	Sequence (5'–3')	Length	Degeneracy	Tm[a]	Target region	Primer_ID	References
D9_2737R	AAHARGGTCTTCTTTCCY	18	12	47.9	D9		This study
D9_2741R	GCTCAAHARGGTCTTCTT	18	8	48.9	D9		This study
D9_2742R	AGCTYAAHAGGGTCTTCT	18	8	49.6	D9		This study
D9_2593R	GAGAGTCATAGTTACBYC	18	8	46.2	D9		This study
D11_3143R	RCCACAAGCYARTTATCC	18	12	50	D11		This study
21R	GACGAGGCATTTGGCTACCTT	21	1	57.6	D9	21R	Schwelm et al. (2016)
22R	CCATTCATGCRCGTCACWART	21	1	55.9	D9	22R	Schwelm et al. (2016)

[a]Melting temperature (Tm) depicts the average for degenerate primers and was calculated with IDT OligoAnalyser.

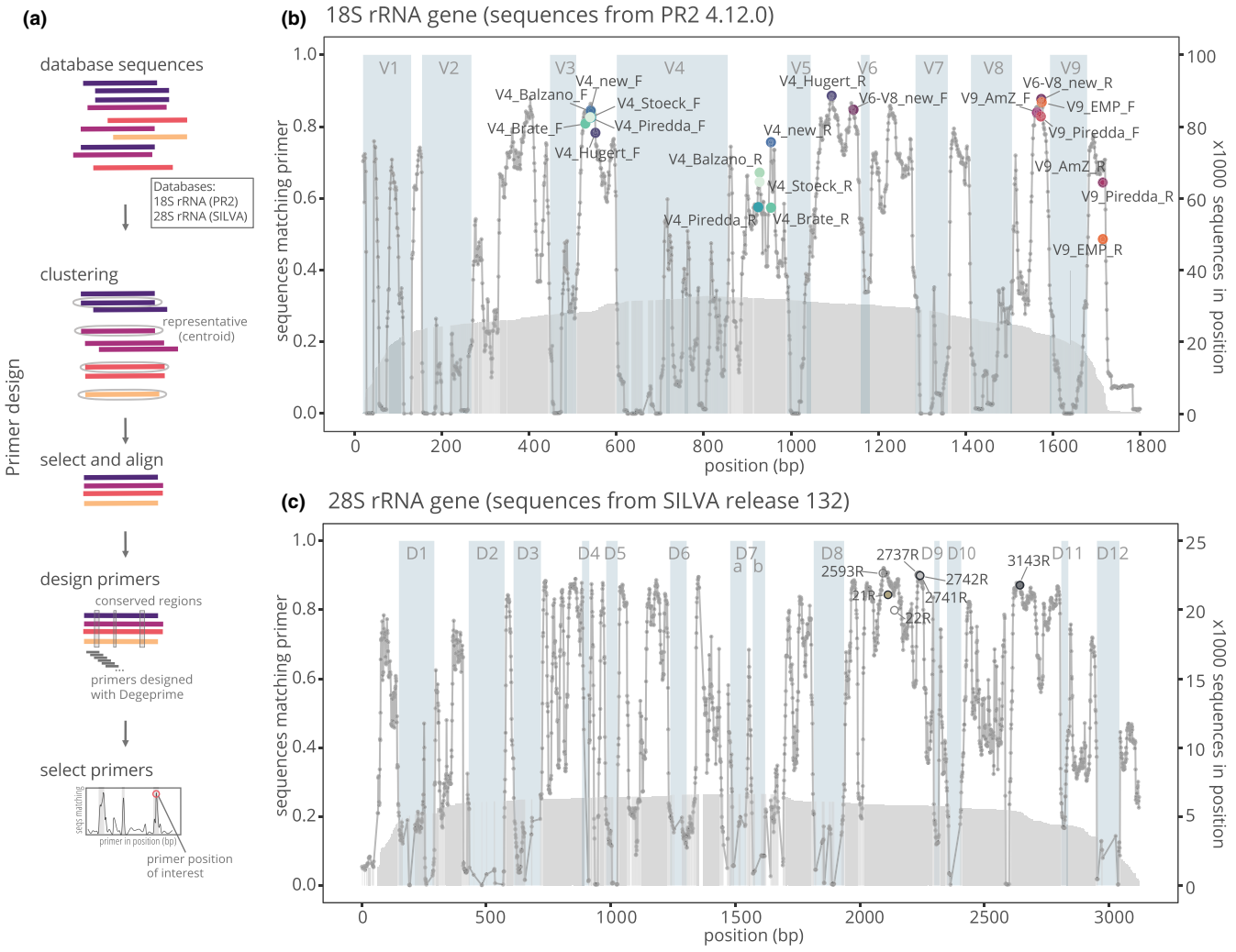

FIGURE 1 Design and evaluation of 18S and 28S rRNA gene primers based on public database sequences. (a) Workflow for designing primers targeting eukaryotic rRNA genes based on public database sequences. Sequences are clustered, centroid sequences aligned, and primers designed in each position of the alignment with DegePrime (Hugerth, Wefer, et al., 2014). (b) Plot of primers designed on 18S rRNA gene sequences (PR2 version 4.12.0) with a primer length of 18 bp and degeneracy of 12. Variable regions are highlighted (pale blue). Grey bars represent the number of sequences spanning each position of the alignment. Connected dots represent primers designed at each position and the proportion of spanning sequences being matched. Primers designed and evaluated in this study are highlighted. (c) Plot of primers designed on 28S rRNA gene sequences (SILVA release 132). The colour markings for the primer pairs are consistent throughout the figures

fraction 0.1–200 μm) across the Baltic Sea (Alneberg et al., 2020), were extracted using METAXA2 (Bengtsson-Palme et al., 2015). From the METAXA2 output, eukaryotic sequences from each sample were combined into a database, and taxonomy files generated for each database. To reduce overrepresentation of abundant sequences, clustering was performed at 97% identity as described in section.

2.1.3 | *In silico* testing of primers

Primers for 18S rRNA (Table 1) and 28S rRNA (Table 2) were tested *in silico* on PR2 and SILVA database sequences, and the marine and brackish databases generated from metagenomic sequencing. For *in silico* PCRs, the alignment files of the databases were trimmed to the region covered by the respective tested individual primer or primer pair; partial sequences were removed. Sequences with 100% match to individual primers/primer pairs were selected with the USEARCH commands search_pcr and search_oligodbs; the search_pcr ampout option generated fasta files with amplicons for further analyses.

2.2 | DNA extraction, library preparation and sequencing

2.2.1 | Sampling and DNA extraction

Sea water samples were collected in duplicates from three sampling station in the Bothnian Bay at the offshore station A13 (64°42'30.0"N 22°04'00.0"E), at the coastal Kattegat station N14 Falkenberg (56°56'24.0"N 12°12'42.1"E), and at the offshore North Sea station HE513_3 (55°47'35.5"N 3°33'50.0"E) (Figure 5a). For the samples from the Bothnian Bay and Kattegat, 500 ml water from 0–10 m depth were filtered onto 0.22 μm membrane filters (Merck Millipore) and DNA extraction from filters performed using the ZymoBIOMICS DNA Miniprep kit (Zymo Research Corp) by following the manufacturer's instructions. For the North Sea samples, 60 l of water were sampled on a P20 filter (>20 μm size fraction) after >200 μm prefiltering, and DNA extracted with the NucleoSpin Soil Kit (Macherey und Nagel) according to the manufacturer's instructions.

2.2.2 | Short-read metabarcoding library preparation and sequencing

The primers for short-read amplification on 18S rRNA (Table 1) were ordered from IDT DNA (IA, USA) with sequence adapters (forward: 5-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3', reverse: 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3) appended in the 5' ends. Amplification conditions for the new primer pairs V4 and V6–V8 were optimised with gradient PCRs to increase yield and to reduce nonspecific amplifications, and the product analysed on a Bioanalyser 2100 (Agilent Technologies). 25 μl PCR reactions were carried out with the KAPA HiFi HotStart ReadyMix PCR Kit (Kapa Biosystems), according to the manufacturer's instructions, with 10 ng template DNA and 0.3 μM final concentration of each primer. For the new V4 and V6–V8 primer pairs, annealing temperatures were set to 52°C. For library preparation with the published primers, the protocol was adapted to the polymerase (Table S2). Cleaning of the product with MagSI-NGS prep plus (MagnaMedics), indexing through a second PCR with Kapa HiFi HotStart ReadyMix and sequencing on a MiSeq (Illumina Inc.) was performed at SciLifeLab/NGI (Solna, Sweden). The PCR conditions for indexing were 95°C for 2 min, 8 cycles of 98°C for 20 s, 55°C for 30 s and 72°C for 30 s, followed by a final elongation step of 72°C for 2 min.

2.2.3 | Long-read metabarcoding library preparation

We evaluated three sets of primers targeting a ~4.5 kbp region of the eukaryotic rRNA operon (Table 2). Three reverse primers were paired with the same forward primer (V4_Balzano_F/V4F): 21R (Schwelm et al., 2016) and the new primers 2742R and 3143R. Amplification conditions were optimised on environmental water samples. 50 μl PCR reactions were carried out using the Takara PrimeStar GXL DNA polymerase by following the manufacturer's instructions for the rapid PCR protocol, using 10 ng template DNA, 0.3 μM final concentration of each primer, and 30 cycles of 98°C for 10 s, 55/60°C for 15 s and 68°C for 90 s, followed by a final elongation step at 68°C for 60 s. The PCR product was analysed on a Bioanalyser 2100 with a High Sensitivity DNA Kit. Libraries for sequencing were prepared using fusion primers with PacBio-specific barcodes. After purifying the PCR product with the QIAquick PCR Purification kit (Qiagen), the libraries were pooled at equimolar concentrations and sequenced at SciLifeLab/NGI (Uppsala, Sweden) with circular consensus sequencing on PacBio Sequel.

2.2.4 | Shotgun metagenomics sequencing

Libraries were prepared at SciLifeLab/NGI (Sweden) using the TAKARA SMARTer ThruPLEX DNA-Seq kit that allows for low DNA input. The libraries were sequenced on Illumina NovaSeq6000 with 2 × 150 bp.

2.3 | Processing and analyses of sequencing data

Sequencing data generated in this study are available at the European Nucleotide Archive (ENA) under the study accession number PRJEB47297. Analyses of sequencing data and plotting of the data was performed in R version 4.0.3 using the packages DADA2 (Callahan et al., 2016), DECIPHER (Wright, 2016), VEGAN (Vegan:

Community Ecology Package), and GGPLOT2 (Valero-Mora, 2010). The 18S rRNA gene database PR2 version 4.12.0 (Guillou et al., 2013) was used as training set for taxonomic classification with assignTaxonomy of DADA2.

2.3.1 | Short-read Illumina data

The median sequencing depth was 0.11 M reads per sample with 88.78% of reads of a quality score ≥30. The DADA2 pipeline was used to infer biological sequences from amplicon reads (Callahan et al., 2016), resulting in 6173 amplicon sequence variants (ASVs) from six samples and nine primer pairs (V4_Hugert primers were excluded since forward and reverse reads did not overlap). ASV abundance was rarefied with the function rarefy from the VEGAN package version 2.5–7 to 50,000 per sample. Off-target bacterial reads were removed from the data set by only keeping amplicons >350 bp for the V4 region and 110–200 bp for the V9 region.

2.3.2 | Long-read PacBio data

Circular consensus sequences (CCS) were generated from raw reads by SMRT link version 8.0.0.79519 with a minimum of three passes and a quality score of 20. A mean of 12,525 reads per sample was generated, with a mean read length of 4283 bp and mean barcode quality of 94.7%. A DADA2 pipeline adapted to PacBio data was used for read processing: after primer removal, reads were filtered and trimmed with "minQ=2", "maxEE=30" and reads outside the length range of 2000–6000 were removed. After dereplicating, roughly 90% of the sequences remained as unique, indicating a high error frequency. Error learning was performed with "errorEstimationFunction =PacBioErrfun", and denoising with the detect singletons option and pool="pseudo", resulting in 500–2500 ASVs per sample and 35,608 ASVs in total. For assigning taxonomy, both the 18S rRNA gene and the V4 region of the ASVs were used. Abundances were rarefied to 6000 per sample. For extraction of rRNA genes and internal transcribed spacers (ITS) from the reads, the tools METAXA2 (Bengtsson-Palme et al., 2015) and ITSX (Bengtsson-Palme et al., 2013) were used.

2.3.3 | Shotgun metagenomics data

Shotgun metagenomics reads were trimmed from remaining adapters with CUTADAPT (Martin, 2011), merged, and METAXA2 (Bengtsson-Palme et al., 2015) used to extract reads of 18S rRNA gene sequences. We created a custom reference database for METAXA2 based on PR2 sequences clustered at 97% identity with the function metaxa2_dbb in conserved mode. On average 0.042% of the total reads were identified as 18S rRNA genes, and their taxonomy classified with the assignTaxonomy function in DADA2 as for the metabarcoding data.

3 | RESULTS

3.1 | Design and evaluation of rRNA primers with broad taxonomic coverage

Two new primer pairs targeting the 18S rRNA gene's V4 and V6-8 regions and five reverse primers targeting the 28S rRNA gene were designed in this study. We evaluated these primers together with nine published primer pairs commonly used for metabarcoding of 18S and 28S rRNA genes (Tables 1 and 2, Figure 1). From *in silico* PCRs on 18S rRNA gene sequences of PR2, amplicons were used for evaluation of amplicon length distribution and thus suitability for Illumina sequencing (Figure 2a), inferable taxonomic information (Figure 2b), and overall- and taxon-specific coverage (Figure 3a). For the reverse primers targeting the 28S rRNA gene, overall coverage (Figure 3b) and taxonomic bias were determined on SILVA. No mismatches to the primer sequences were allowed in the *in silico* PCRs.

3.1.1 | Lengths of amplicons generated by primer pairs targeting 18S rRNA

The choice of amplicon size is a trade-off between obtaining a high taxonomic resolution by longer amplicons and meeting the limitations of current sequencing technologies. Illumina MiSeq sequencing limits the amplicon size to 2 × 300 bp, leaving ~480 bp for the amplicon after deducting primer sequences and >20 bp to merge forward and reverse reads. Variation in amplicon lengths between taxa can also introduce PCR biases; however, a narrow range of amplicon lengths was observed for all pairs tested (Figure 2a). Meanwhile, the size of amplicons generated varied depending on the targeted variable regions (Figure 2a). Primer pairs targeting the V9 region generate amplicon lengths of <200 bp and were used in early metabarcoding studies to meet the technological sequencing limitations (Amaral-Zettler et al., 2009). In recent metabarcoding studies, the V4 region has been more frequently targeted (Bruhn et al., 2021; Egge et al., 2021; Geisen et al., 2019; Hörstmann et al., 2021), generating amplicons >400 bp; however, the >500 bp amplicons generated by V4_Hugerth primers exceed the ~480 bp length limitation for merging of Illumina read pairs. The two new primer pairs "V4_new" and "V6–V8_new" generated *in silico* amplicon lengths within 400–480 bp. The "V6-V8_new" primer pair targets the V6, V7 and V8 variable regions.

3.1.2 | Inferable taxonomic information from *in silico* amplicons

Longer amplicons can potentially provide higher taxonomic resolution, depending on the variability within the amplified region. In Figure 2b the taxonomic resolution of amplicons generated *in silico* was plotted. The plot with absolute values primarily highlights the difference in number of ASVs/unique sequences generated depending

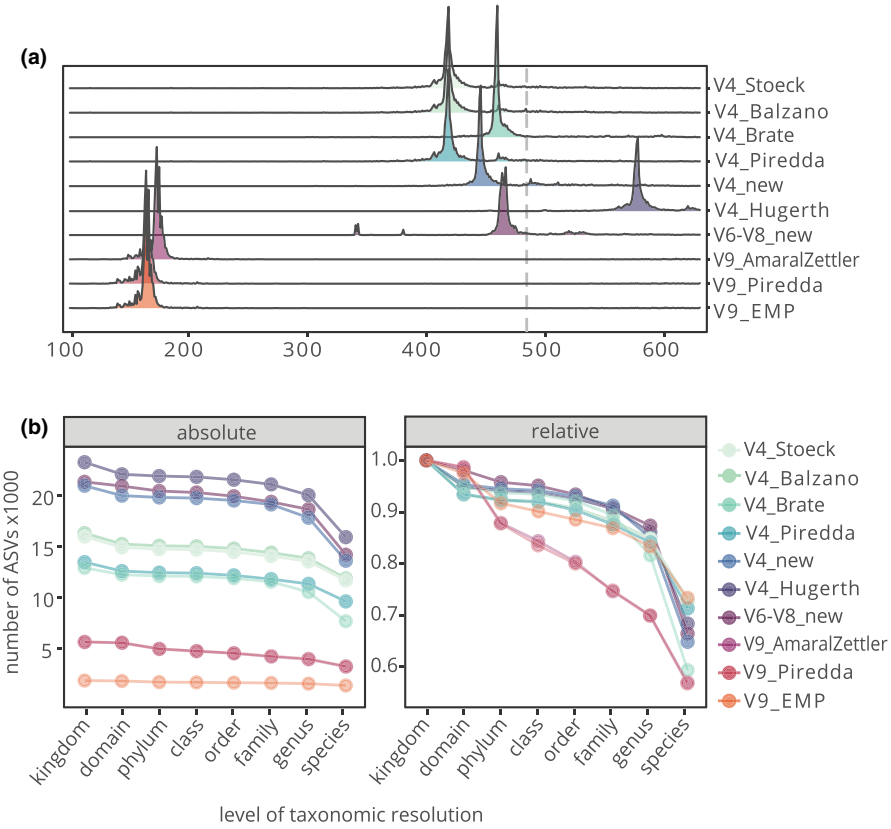

FIGURE 2 Amplicon length distribution and taxonomic resolution of *in silico* PCR on public database sequences of ten 18S rRNA primer pairs. The sequences were simulated from a clustered PR2 database (version 4.12.0). (a) Amplicon length distribution of amplicons generated on the database. (b) Inferable taxonomic information from unique *in silico* amplicon sequences (ASVs). Taxonomy was assigned using the function assignTaxonomy from the DADA2 R package. Note that V9_AmaralZettler and V9_Piredda share similar primer sequences and therefore plot on top of each other

on the primer pair; the two new pairs and the V4_Hugerth pair produced the highest numbers of unique amplicons. The plot with relative values reveals a steady decrease in taxonomic assignment from kingdom down to genus level, and a steeper drop to species level. Around 90% of ASVs generated from the V4 and V6-8 region were classified to genus level, and ~70% to species level. The primers targeting V9 resulted in fewer ASVs and a steeper drop in taxonomic assignment from domain to genus level; ~70% of the V9 ASVs were classified at genus level compared to ~80%–90% of the V4 ASVs. The relative resolution of the V9_EMP primers did not follow the pattern of the other two V9 primers; due to the very low number of ASVs generated for this primer pair the results are probably distorted. Similar patterns of taxonomic resolution were observed *in vitro* when applying the primer pairs on six water samples (Figure S1).

3.1.3 | Coverage of primers on public and constructed environmental databases

The amplicons generated from the *in silico* PCR were used to evaluate the relative coverage of the primers across public database sequences (Figure 3a). Several eukaryotic supergroups are underrepresented, by the reverse primers in particular. The V6–V8_new pair provided the highest and most even coverage across taxonomic groups. Apart from the V6–V8_new_pair and V4_Hugerth_pair, the coverage of the Excavata supergroup was very limited by all primer pairs. The V4_Bråte pair provided low coverage of Stramenopiles, and the V4_Piredda pair of Rhizaria. The V9 targeting primers provided an uneven coverage across groups with several major groups underrepresented. To simulate how well the primers would work on aquatic samples, additional *in silico* PCRs were performed on metagenomes from marine (Tully et al., 2018) and brackish environments (Alneberg et al., 2020). In general, the coverage of the primers on the metagenomes reflected the coverage observed on database sequences, but for example, the V4_Hugerth_pair performed better on the PR2 database than on the environmental ones, and vice versa for the V4_Piredda_pair. For both the V4 and V9 primer pairs, the reverse primer was the limiting factor since it matched considerably fewer sequences than the forward primer. In some cases, primers designed for targeting 18S rRNA can also amplify prokaryotic 16S rRNA genes, but the evaluated primers were in general specific to eukaryotic 18S rRNA. The V9_EMP forward primer matched to bacterial and to a lesser extent archaeal sequences, but combined with the V9_EMP_R primer the pair was specific to eukaryotic sequences.

3.1.4 | *In silico* evaluation of primers targeting 28S rRNA

In addition to two previously published reverse 28S rRNA primers, five primers designed in this study were evaluated (Figure 3b). The coverage of all primers on the SILVA database and the simulated brackish sample was generally >90%. Lower coverage was obtained on the simulated marine sample, where the primers 2593R and 3143R provided the highest coverage of ~60%–70%. As for many of the SSU

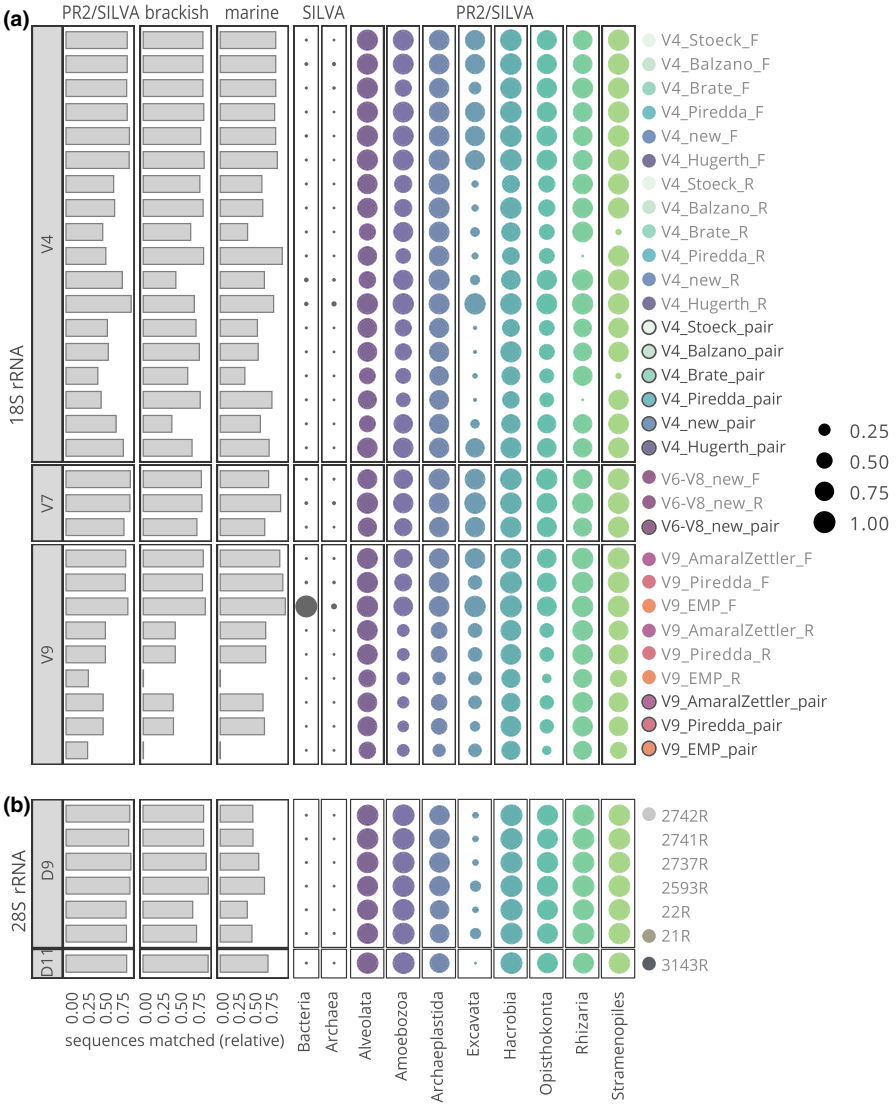

FIGURE 3 Coverage of primer pairs and individual primers across database sequences and taxonomic groups. Database sequences derive from published databases (PR2 for 18S rRNA, SILVA for 28S rRNA, SILVA for prokaryotic 16S and 23S rRNA sequences) and constructed databases of rRNA sequences extracted from shotgun metagenomics data of brackish (Baltic Sea: Alneberg et al., 2020) and marine water (TARA Oceans; Tully et al., 2018). Coverage of primers on eukaryotic supergroups was based on public database sequences (PR2/SILVA) and corresponds to the size of the circle (legend). Matches to prokaryotic sequences are shown as grey circles. The primers are sorted by variable regions amplified within (a) 18S and (b) 28S rRNA genes, and by forward "_F", reverse "_R" and pair "_pair" (marked with black stroke)

primers, the coverage of the Excavata supergroup was limited by all primers, while the other supergroups appeared evenly covered.

3.2 | *In vitro* evaluation of primers targeting 28S rRNA and long-read amplicons

Since a thorough *in silico* evaluation of primer pairs for long-range amplification is not feasible due to the low number of eukaryotic genomes for which complete rRNA operon data is available, analyses of amplicon length distribution and taxonomic resolution were performed on sequencing data generated in this study (Figure 4; and further described in the next section). The length distributions of the amplicons of the three primer pairs tested were much wider than the distributions from the 18S rRNA short-read amplicons, probably reflecting length variation in the ITS sequences, variable regions, and introns. Since the same 18S rRNA targeting forward primer (V4_Balzano_F) was used in the pairs, the shifts in length distribution between the pairs were due to the different reverse primers used (Figure 1c). 3143R produced the longest amplicons with the highest peak at ~4700 bp length, 2742R with a main peak at ~4300 bp and the previously published 21R (which binds closely to 2742R) at ~4200 bp.

The taxonomic resolution of the long-range amplicons was evaluated for the V4 region and the 18S rRNA region of the amplicon (Figure 4b; the different regions of the amplicons are illustrated in Figure 4c). The three primer pairs performed similarly for both regions; for the full 18S rRNA region, 85%–90% of the sequences were classified down to species level, while for the V4 region trimmed to the V4_Balzano primers, <80%.

To investigate the potential of different regions of the eukaryotic rRNA operon for metabarcoding, we evaluated the variability within the regions 18S, 28S, ITS1, ITS2 and V4 (Figure 4c,d). Since we found that denoising with DADA2 artificially suppresses variation in highly variable ITS regions (Schoch et al., 2012) of the long-reads (Figure S2), this analysis was conducted without denoising on unique CCS reads from the primer pair V4_Balzano_F – 2742R. The sequences of the individual regions were clustered at 99% identity to compensate for sequencing errors that would artificially inflate the observed variation. The long regions of the 18S and 28S rRNA genes formed fewer sequence clusters than the

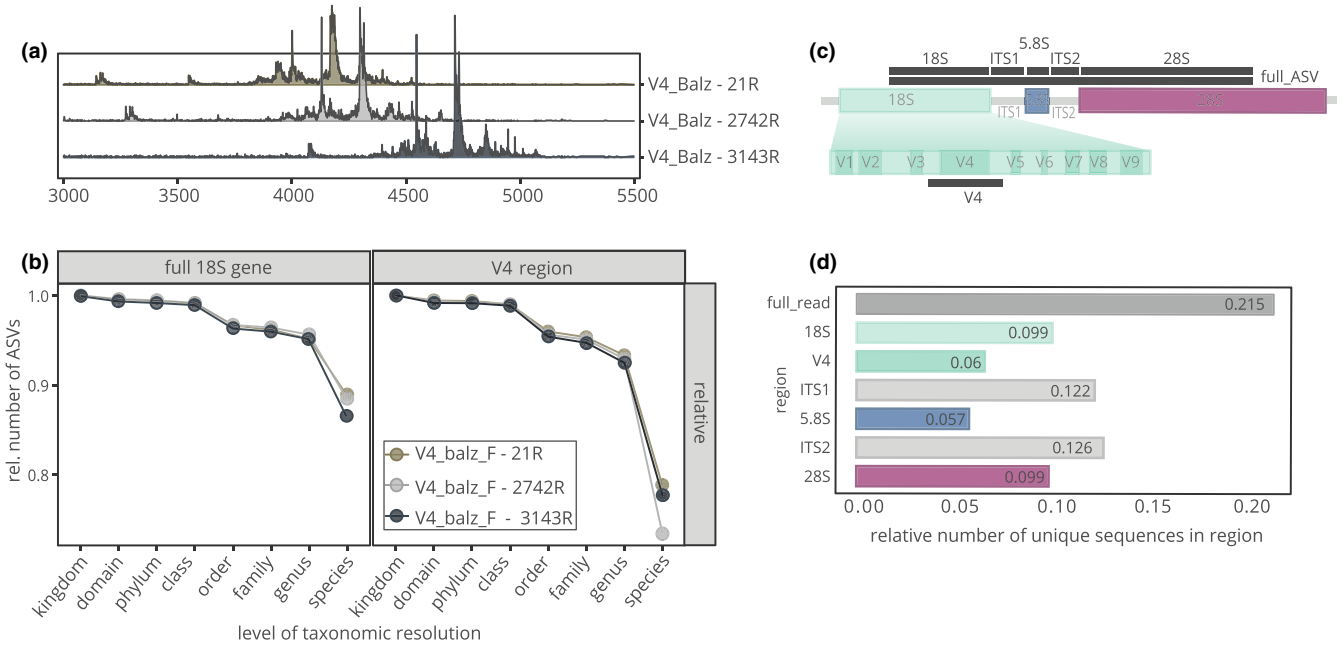

FIGURE 4 Exploring long-read sequencing results. (a) Amplicon length distribution from PacBio reads, (b) taxonomic resolution of rarefied ASVs, based on 18S part and V4 region (trimmed to V4_Balzano primer positions), (c) sketch of the eukaryotic rRNA operon (not to scale) depicting the regions targeted by long-read amplification and the regions selected for (d) bar plot depicting the sequence variability within the targeted regions. X-axis values represent number of sequence clusters for the region divided by the total number of unique full amplicon sequences. The corresponding plots for clustering at 100, 98 and 97% identity, as well as an ASV-based analysis, are provided in Figure S2

full amplicon, with 9.9% compared to 21.5% of sequences from the full amplicons serving as centroid sequences during clustering. The short and variable ITS1 and ITS2 are frequently targeted for species discrimination, for example, dinoflagellates (John et al., 2014), and fungal metabarcoding (Schoch et al., 2012) and 12.2 and 12.6% of sequences, respectively, formed centroids for clusters. The corresponding number for the V4 region was 6%, indicating a lower variability than the ITS regions.

3.3 | Effect of sequencing method and primer pair on inferred community composition in natural samples

In order to evaluate the influence of sequencing method and primer choice on the obtained community composition, we compared Illumina short-read sequencing of 18S rRNA regions and PacBio long-read sequencing of the rRNA operon, using different primer pairs, and Illumina shotgun metagenomic (MG) sequencing, on six environmental samples (Figure 5). The sample duplicates were taken at three locations with diverging temperature and salinity levels: the North Sea, Kattegat and Bothnian Bay (Figure 5a). All primer pairs performed well under the selected PCR conditions (Table S2), and sequencing of amplicons yielded a sufficient number of reads for each sample; for short-read samples >50,000 and for long-read samples >9000 apart from one outlier with 6554 reads. Taxonomic assignment of the short-read sequences identified off-target amplification of bacteria by the primer pairs V4_new (15% of reads), V9_Piredda (3%) and V9_EMP (1.5%); all other primers <0.5%.

A non-metric multidimensional scaling (NMDS) plot at genus level of all samples showed strong clustering by sampling location (Figure 5b); separate plots per location revealed a secondary clustering by sequencing method and primer pair. NMDS plots at the other taxonomic levels can be found in Figure S3. Typically, biological replicates processed with the same method grouped closely. Depending on the taxonomic level chosen for creating the NMDS plots, the clustering shifted but the general patterns remained (Figure S3). Within the short-read samples, those targetting the same variable region grouped closely. For the long-read samples there was a minor separation depending on if the V4 or the full 18S rRNA region was used for taxonomic classification. Generally, Illumina samples clustered closer to the shotgun MG samples than long-read samples, and of those in particular V9 primers, V4_new and V6-V8_new.

When comparing the community composition on supergroup level (Figure 5c), the influence of the sequencing methods and primer pairs on the observed composition became more evident. The differences in community profile depending on the variable region targeted was also visible here, but more notable at lower taxonomic levels (Figure S4). The composition obtained by shotgun MG can be assumed to reflect the true community composition most closely, since it is not affected by primer biases.

To directly compare the metabarcoding data to shotgun MG, we plotted the average Bray-Curtis dissimilarity across samples at each taxonomic level, where a value of 1 represents the highest

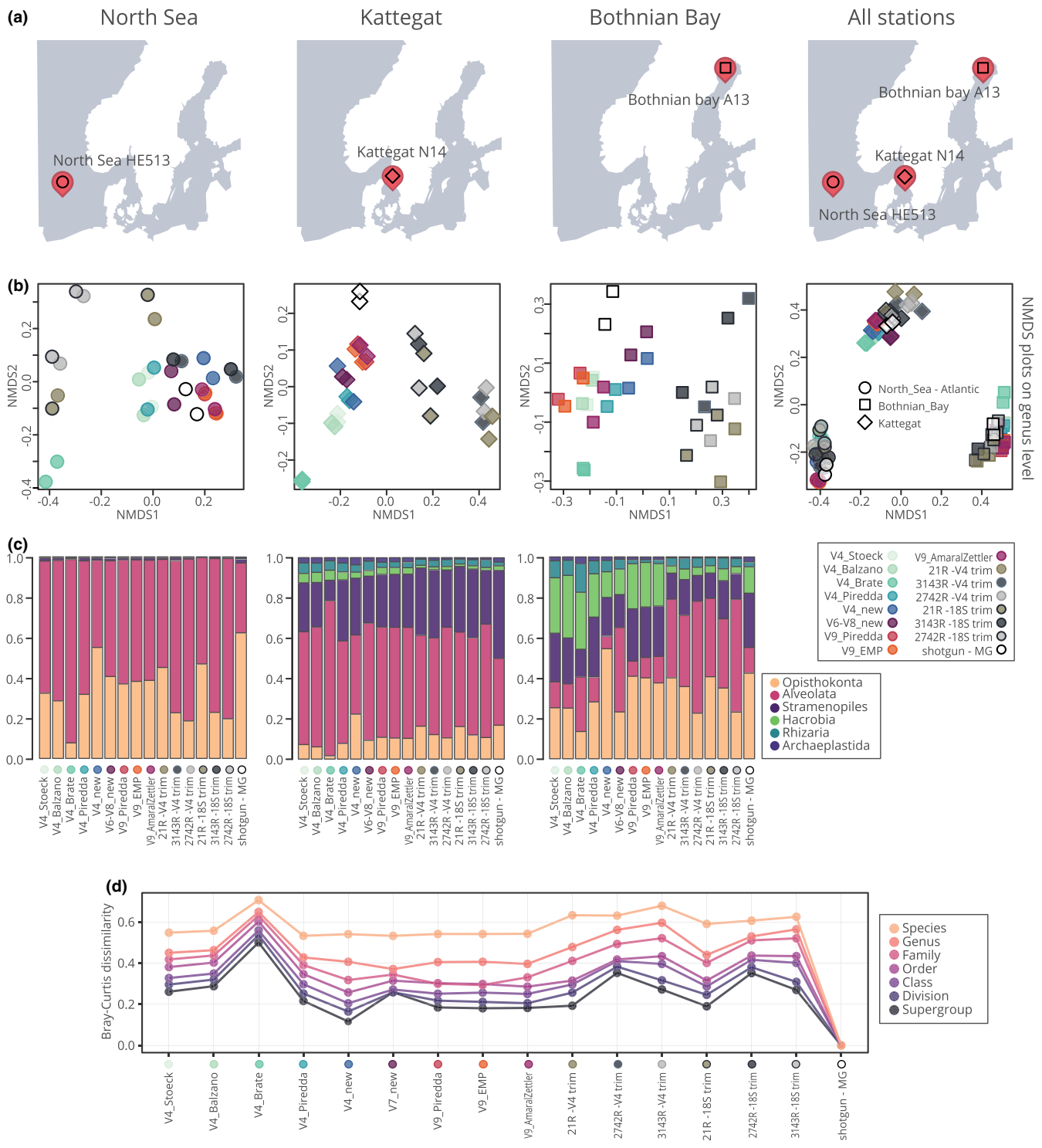

FIGURE 5 Comparison of community composition reconstructed depending on choice of sequencing methods and primer pair. (a) Samples were taken in duplicates from three locations with expectedly distinct microbial eukaryote profile: Atlantic/North Sea station HE513, Kattegat station N14 and Bothnian Bay station A13. PacBio ASVs were trimmed to 18S rRNA region and 18S-V4 region for taxonomic classification. (b) NMDS plots of beta-diversity calculated on genus-level taxonomic assignments for all samples and samples separated by location (as shown in the maps [a]). NMDS plots on other taxonomic levels are available in Figure S3. (c) Bar plots of one biological replicate per location on phylum level (supergroup in PR2 taxonomy), abundance of each group was normalised across samples, bar plots at all taxonomic levels and of both biological replicates in Figure S4 (replicate 1) and Figure S5 (replicate 2). (d) Average Bray-Curtis dissimilarity of primer pairs to shotgun-MG data across six samples on seven taxonomic levels

dissimilarity (Figure 5d). It became apparent that the dissimilarity to the shotgun MG data increased at lower taxonomic levels. Of the short-read primer pairs, V4_new showed the lowest dissimilarity up to family level, while the V7_new primers showed the narrowest range of dissimilarity across taxonomic levels, except species level. At genus and species level little difference could be observed between the pairs, with the exception of the V4_Bråte pair which most strongly diverged from the MG data (Figure 5c,d).

The V4_Bråte pair had a notable bias against the supergroup Opisthokonta (Figure 5c); on closer observation the highly abundant copepod species *Eurytemora affinis* (Bothnian Bay) and *Oithona* sp. (North Sea) were strongly underrepresented by V4_Bråte (Figure S4). Consistent with this, the V4_Bråte primers did not match to sequences of these genera *in silico*, while V4_Balzano and V4_new did. The three V9 primer pairs created very similar community profiles (Figure 5c) and had low dissimilarity to the shotgun MG profile at supergroup level; however, at lower taxonomic levels differences became evident (Figure 5d) connected to that many taxonomic groups were not identified (Figure S4). The community profiles created by the three long-read primer pairs and PacBio sequencing differed based on the reverse primer only (Figure 5c), and generally overrepresented Alveolates and in specific Ciliophora (Figure S4). Based on Bray-Curtis dissimilarity the 21R reverse primer was most successful in mimicking the shotgun MG data (Figure 5d), both when the 18S rRNA and the V4 region was used for taxonomic assignment. The results obtained from the V4 region and full 18S did not differ at the supergroup level but were surprisingly different on lower taxonomic levels (Figure S4). Even though the V4 region from PacBio amplicons was obtained by trimming reads based on the V4_Balzano primers' positions, the community profile did not bear similarity with those created by the V4_Balzano primers, indicating amplification biases was part of the problem. The supergroup Hacrobia is not well covered by any of the PacBio reverse primers tested; the bias cannot be explained by the forward primer V4_Balzano_F since the V4_Balzano pair even overrepresented this taxonomic group, compared to the shotgun MG samples.

4 | DISCUSSION

In this study, we performed *in silico* evaluations of primers for short- and long-read metabarcoding of eukaryotes and compared the community composition obtained *in vitro* from six marine samples with shotgun metagenomic (MG) sequencing. To our knowledge, this is the first comparison of the three sequencing methods for microbial community characterisation. The rationale for repeating the primer design effort by Hugerth, Muller, et al. (2014) was that the expanded sequence databases and clustering of sequences would enable design of broader targeting primers with less bias towards overrepresented taxa. We also included a larger set of previously published primers in the evaluations. We showed that our approach is suitable by confirming the primers performance both *in silico* and *in vitro*. The taxonomic resolution tests only measured what proportion of sequences was classified at the different levels and not the accuracy of the classifications. Although a subset of the classifications was probably wrong, we argue that this is still a valid approach for comparing the taxonomic information of the different primer pairs/regions, since the applied *k*-mer-based algorithm implemented in DADA2 (Callahan et al., 2016) will assign more informative sequences to more detailed taxonomic levels (based on the patterns of *k*-mer sharing). There was an indication that the choice of 18S rRNA region influenced the taxonomic classifications since classification of long-read sequences by the full 18S rRNA gene and the V4 region frequently yielded different taxonomic assignments.

From the tested primer pairs, the pairs V6-V8_new, V4_Hugerth and V4_new provided the highest coverage (Figure 3a) and the largest number of unique amplicons *in silico* (Figure 2b), due to targeting longer regions of the 18S rRNA gene and potentially higher variability within the regions (Figure 2a). The Excavata supergroup received low coverage by all tested pairs, except V6-V8_new and V4_Hugerth (Figure 3a), potentially connected to the many fast-evolving lineages within the group (Burki et al., 2020). However, since the V4_Hugerth pair generated too long amplicons for sequencing on Illumina Miseq (Hugerth, Muller, et al., 2014), and the V4_new pair amplified off-target bacterial sequences, these pairs are less suitable for metabarcoding of eukaryotes. The V9 pairs did not perform as well as the other pairs, probably due to the short amplicons and more limited representation of the region in the PR2 database (Figure 1a). Our analyses indicated the V6–V8_new pair performs the best of the tested pairs, and that it most closely resembled the taxonomic profile from unbiased shotgun MG sequencing *In vitro* (Figure 5b,c). Primers targeting the V6–V8 regions are not widely used, but a pair targeting these regions was previously shown to provide high coverage (Vaulot et al., 2021; Wilkins et al., 2013). However, since it is desirable to be able to compare across studies, a more commonly used primer pair with slightly lower performance might be preferable. In that case, the V4_Balzano pair (Balzano et al., 2015) seems a reasonable choice for short-read 18S rRNA metabarcoding studies, since it had the best performance after V6–V7_new, is used in many studies (Berdjeb et al., 2018; Obiol et al., 2020; Questel et al., 2021) and additionally differs from the most commonly used pair V4_Stoeck (Vaulot et al., 2021) by only the last nucleotide of the reverse primer (Table 1). The simplified PCR protocol we established in this study (Table S2) with only one reaction instead of six, one annealing temperature, and fewer cycles, makes the primers more suitable for high-throughput library preparation. We compared the simplified PCR protocol to the original protocol (Table S2) without observing significant differences between the samples (Figure S7) in neither alpha-diversity (Wilcoxon signed-rank test, p-value = .164) nor beta-diversity (analysis of similarities (ANOSIM), R = −.112, p-value = .995, number of permutations = 9999).

While the advantages of long-read metabarcoding of almost the whole rRNA operon are ample, this relatively new method has its limitations. Next to the high costs per base pair, technical limitations should be taken into consideration. The wide distribution

of amplicon lengths could lead to shorter sequences being over-represented (Tedersoo et al., 2018), and PCRs on such long DNA regions increase the formation of chimeric sequences. A current bottleneck for data analysis is the lack of an established pipeline for denoising of long-read sequences, although attempts have been made for the full rRNA operon (Heeger et al., 2018; Jamy et al., 2020) and for SSU only using DADA2 (Callahan et al., 2019). When we used DADA2 for sequence denoising on the full rRNA operon, the biological variation in the ITS regions appeared to be falsely suppressed, since the numbers of unique ITS sequences were lower than the number of unique 18S V4 sequences, as shown in Figure S2, contrary to what would be expected and what was obtained without applying denoising. This is probably due to the combination of long amplicons and low sequencing depth, as previously noted (Furneaux et al., 2021), rendering too few reads for many of the true biological sequences to be recaptured by the algorithm. Furthermore, there is currently no database or collection of full eukaryotic rRNA operon sequences available, like it is attempted for bacteria (Benítez-Páez & Sanz, 2017; Kinoshita et al., 2021); such an initiative would greatly benefit the ease of application.

As part of this study, we evaluated seven reverse primers for long-read metabarcoding, including two used for long-read metabarcoding in a previous study (Jamy et al., 2020), and sequenced libraries created with three primer pairs with PacBio Sequel. For design of new primers we focussed on the end of the 28S rRNA gene, as this would include additional variable regions in the resulting sequences. All tested primers provided a high *in silico* coverage of 28S rRNA gene sequences (Figure 3b) but did generally not match well to the Excavata supergroup. *In vitro*, the newly designed reverse primers yielded good PCR products (Figure S6). Two of the new reverse primers and one from Jamy et al. (2020) were selected to prepare long-read amplicon libraries from six environmental samples (Figure 5). The V4_Balzano forward primer was chosen for library preparation, since it outperformed the V4_Bråte_F primer used by Jamy et al. (2020) in our tests. The pairs provided somewhat different community profiles, in particular with regard to ratios between Opisthokonta and Alveolata (Figure 5c). Since the performance of the primers did not significantly differ, and since 3143R (designed here) generated the longest amplicon with two additional variable regions (D9 and D10), 3143R might be the preferable choice for long-read rRNA operon metabarcoding studies.

ACKNOWLEDGEMENTS

DNA sequencing was conducted by the Swedish National Genomics Infrastructure (NGI) in Stockholm and Uppsala. Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). We thank Luis Fernando Delgado Zambrano for bioinformatic support. This study was supported by the Swedish Agency for Marine and Water Management and the Swedish Environmental Protection Agency under the grant number NV-03728-17. Meike A. C. Latz was additionally supported by a research grant (34442) from VILLUM FONDEN and Anders F. Andersson by a research grant (2017-00694) from Formas - a Swedish Research Council for SustainableDevelopment.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Meike A. C. Latz and Anders F. Andersson designed the research and Uwe John provided input on the study design. Jenny Lycken and Sonia Brugel collected water samples and extracted DNA, Uwe John provided DNA samples. Meike A. C. Latz performed the laboratory experiments, bioinformatic analyses, and made the figures. Anders F. Andersson gave advice on the bioinformatic analyses. Meike A. C. Latz and Vesna Grujcic wrote the first complete draft of the manuscript, and all authors contributed to the final manuscript.

DATA AVAILABILITY STATEMENT

Raw sequence data have been submitted to the European Nucleotide Archive (ENA) under the study accession number PRJEB47297.

ORCID

Meike A. C. Latz https://orcid.org/0000-0002-6583-9291
Vesna Grujcic https://orcid.org/0000-0002-3322-599X
Sonia Brugel https://orcid.org/0000-0002-1298-3839
Uwe John https://orcid.org/0000-0002-1297-4086
Bengt Karlson https://orcid.org/0000-0002-7524-3504
Agneta Andersson https://orcid.org/0000-0001-7819-9038
Anders F. Andersson https://orcid.org/0000-0002-3627-6899

REFERENCES

Alneberg, J., Bennke, C., Beier, S., Bunse, C., Quince, C., Ininbergs, K., Riemann, L., Ekman, M., Jürgens, K., Labrenz, M., Pinhassi, J., & Andersson, A. F. (2020). Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Communications Biology*, *3*(1), 119. https://doi.org/10.1038/s42003-020-0856-x

Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W., & Huse, S. M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One*, *4*(7), e6372. https://doi.org/10.1371/journal.pone.0006372

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5

Balzano, S., Abs, E., & Leterme, S. C. (2015). Protist diversity along a salinity gradient in a coastal lagoon. *Aquatic Microbial Ecology: International Journal*, *74*(3), 263–277. https://doi.org/10.3354/ame01740

Barton, A. D., Pershing, A. J., Litchman, E., Record, N. R., Edwards, K. F., Finkel, Z. V., Kiørboe, T., & Ward, B. A. (2013). The biogeography of marine plankton traits. *Ecology Letters*, *16*(4), 522–534. https://doi.org/10.1111/ele.12063

Bengtsson-Palme, J., Hartmann, M., Eriksson, K. M., Pal, C., Thorell, K., Larsson, D. G. J., & Nilsson, R. H. (2015). METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular Ecology Resources*, *15*(6), 1403–1414. https://doi.org/10.1111/1755-0998.12399

Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., De Wit, P., Sánchez-García, M., Ebersberger, I., de Sousa, F., Amend, A. S., Jumpponen, A., Unterseher, M., Kristiansson, E., Abarenkov, K., Bertrand, Y. J. K., Sanli, K., Eriksson, K. M., Vik, U., … Nilsson, R. H. (2013). Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution/British Ecological Society*, *4*, 914–919. https://doi.org/10.1111/2041-210x.12073

Benítez-Páez, A., & Sanz, Y. (2017). Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION™ portable nanopore sequencer. *GigaScience*, *6*(7), 1–12. https://doi.org/10.1093/gigascience/gix043

Berdjeb, L., Parada, A., Needham, D. M., & Fuhrman, J. A. (2018). Short-term dynamics and interactions of marine protist communities during the spring–summer transition. *The ISME Journal*, *12*(8), 1907–1917. https://doi.org/10.1038/s41396-018-0097-x

Bolhuis, H., & Stal, L. J. (2011). Analysis of bacterial and archaeal diversity in coastal microbial mats using massive parallel 16S rRNA gene tag sequencing. *The ISME Journal*, *5*(11), 1701–1712. https://doi.org/10.1038/ismej.2011.52

Bradley, I. M., Pinto, A. J., & Guest, J. S. (2016). Design and Evaluation of Illumina MiSeq-Compatible, 18S rRNA Gene-Specific Primers for Improved Characterization of Mixed Phototrophic Communities. *Applied and Environmental Microbiology*, *82*(19), 5878–5891. https://doi.org/10.1128/AEM.01630-16

Bruhn, C. S., Wohlrab, S., Krock, B., Lundholm, N., & John, U. (2021). Seasonal plankton succession is in accordance with phycotoxin occurrence in Disko Bay, West Greenland. *Harmful Algae*, *103*, 101978. https://doi.org/10.1016/j.hal.2021.101978

Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The new tree of eukaryotes. *Trends in Ecology & Evolution*, *35*(1), 43–55. https://doi.org/10.1016/j.tree.2019.08.008

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Callahan, B. J., Wong, J., Heiner, C., Oh, S., Theriot, C. M., Gulati, A. S., McGill, S. K., & Dougherty, M. K. (2019). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Research*, *47*(18), e103. https://doi.org/10.1093/nar/gkz569

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., … Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, *9*(1), 373. https://doi.org/10.1038/s41467-017-02342-1

Cavalier-Smith, T., Lewis, R., Chao, E. E., Oates, B., & Bass, D. (2009). Helkesimastix marina n. sp. (Cercozoa: Sainouroidea superfam. n.) a gliding zooflagellate of novel ultrastructure and unusual ciliary behaviour. *Protist*, *160*(3), 452–479. https://doi.org/10.1016/j.protis.2009.03.003

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., … Velayoudon, D. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1261605. https://doi.org/10.1126/science.1261605

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Egge, E., Elferink, S., Vaulot, D., John, U., Bratbak, G., Larsen, A., & Edvardsen, B. (2021). An 18S V4 rDNA metabarcoding dataset of protist diversity in the Atlantic inflow to the Arctic Ocean, through the year and down to 1000 m depth. *{Earth System Science Data*, *13*(10), 4913–4928.

Furneaux, B., Bahram, M., Rosling, A., Yorou, N. S., & Ryberg, M. (2021). Long- and short-read metabarcoding technologies reveal similar spatiotemporal structures in fungal communities. *Molecular Ecology Resources*, *21*(6), 1833–1849. https://doi.org/10.1111/1755-0998.13387

Geisen, S., Vaulot, D., Mahé, F., Lara, E., de Vargas, C., & Bass, D. (2019). A user guide to environmental protistology: primers, metabarcoding, sequencing, and analyses. *bioRxiv* 850610. https://doi.org/10.1101/850610

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., … Christen, R. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597–D604. https://doi.org/10.1093/nar/gks1160

Hadziavdic, K., Lekang, K., Lanzen, A., Jonassen, I., Thompson, E. M., & Troedsson, C. (2014). Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLoS One*, *9*(2), e87624. https://doi.org/10.1371/journal.pone.0087624

Heeger, F., Bourne, E. C., Baschien, C., Yurkov, A., Bunk, B., Spröer, C., Overmann, J., Mazzoni, C. J., & Monaghan, M. T. (2018). Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. *Molecular Ecology Resources*, *18*(6), 1500–1514. https://doi.org/10.1111/1755-0998.12937

Hörstmann, C., Raes, E. J., Buttigieg, P. L., Lo Monaco, C., John, U., & Waite, A. M. (2021). Hydrographic fronts shape productivity, nitrogen fixation, and microbial community composition in the southern Indian Ocean and the Southern Ocean. *Biogeosciences*, *18*(12), 3733–3749. https://doi.org/10.5194/bg-18-3733-2021

Hu, Y. O. O., Karlson, B., Charvet, S., & Andersson, A. F. (2016). Diversity of Pico- to Mesoplankton along the 2000 km salinity gradient of the baltic sea. *Frontiers in Microbiology*, *7*, 679. https://doi.org/10.3389/fmicb.2016.00679

Hugerth, L. W., Muller, E. E. L., Hu, Y. O. O., Lebrun, L. A. M., Roume, H., Lundin, D., Wilmes, P., & Andersson, A. F. (2014). Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS One*, *9*(4), e95567. https://doi.org/10.1371/journal.pone.0095567

Hugerth, L. W., Wefer, H. A., Lundin, S., Jakobsson, H. E., Lindberg, M., Rodin, S., Engstrand, L., & Andersson, A. F. (2014). DegePrime, a program for degenerate primer design for broad-taxonomic-range PCR in microbial ecology studies. *Applied and Environmental Microbiology*, *80*(16), 5116–5123. https://doi.org/10.1128/AEM.01403-14

Jamy, M., Foster, R., Barbera, P., Czech, L., Kozlov, A., Stamatakis, A., Bending, G., Hilton, S., Bass, D., & Burki, F. (2020). Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular Ecology Resources*, *20*(2), 429–443. https://doi.org/10.1111/1755-0998.13117

John, U., Litaker, R. W., Montresor, M., Murray, S., Brosnahan, M. L., & Anderson, D. M. (2014). Formal revision of the *Alexandrium tamarense* species complex (Dinophyceae) taxonomy: the introduction of five species with emphasis on molecular-based (rDNA) classification. *Protist*, *165*(6), 779–804. https://doi.org/10.1016/j.protis.2014.10.001

Karlson, B., Andersen, P., Arneborg, L., Cembella, A., Eikrem, W., John, U., West, J. J., Klemm, K., Kobos, J., Lehtinen, S., Lundholm, N., Mazur-Marzec, H., Naustvoll, L., Poelman, M., Provoost, P., De Rijcke, M., & Suikkanen, S. (2021). Harmful algal blooms and their effects in coastal seas of Northern Europe. *Harmful Algae*, *102*, 101989. https://doi.org/10.1016/j.hal.2021.101989

Kinoshita, Y., Niwa, H., Uchida-Fujii, E., & Nukada, T. (2021). Establishment and assessment of an amplicon sequencing method targeting the 16S-ITS-23S rRNA operon for analysis of the equine gut microbiome. *Scientific Reports*, *11*(1), 1–12. https://doi.org/10.1038/s41598-021-91425-7

Larsen, P. A., Heilman, A. M., & Yoder, A. D. (2014). The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. *BMC Genomics*, *15*, 720. https://doi.org/10.1186/1471-2164-15-720

Lewitus, A. J., Horner, R. A., Caron, D. A., Garcia-Mendoza, E., Hickey, B. M., Hunter, M., Huppert, D. D., Kudela, R. M., Langlois, G. W., Largier, J. L., Lessard, E. J., RaLonde, R., Jack Rensel, J. E., Strutton, P. G., Trainer, V. L., & Tweddle, J. F. (2012). Harmful algal blooms along the North American west coast region: History, trends, causes, and impacts. *Harmful Algae*, *19*, 133–159. https://doi.org/10.1016/j.hal.2012.06.009

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R., Claverie, J.-M., Decelle, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Forn, I., Forster, D., Guillou, L., Jaillon, O., Kooistra, W. H. C. F., Logares, R., … de Vargas, C. (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology*, *17*(10), 4035–4049. https://doi.org/10.1111/1462-2920.12955

Medinger, R., Nolte, V., Pandey, R. V., Jost, S., Ottenwälder, B., Schlötterer, C., & Boenigk, J. (2010). Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular Ecology*, *19*(Suppl 1), 32–40. https://doi.org/10.1111/j.1365-294X.2009.04478.x

Mohrbeck, I., Raupach, M. J., Martínez Arbizu, P., Knebelsberger, T., & Laakmann, S. (2015). High-throughput sequencing-the key to rapid biodiversity assessment of marine metazoa? *PLoS One*, *10*(10), e0140342. https://doi.org/10.1371/journal.pone.0140342

Obiol, A., Giner, C. R., Sánchez, P., Duarte, C. M., Acinas, S. G., & Massana, R. (2020). A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Molecular Ecology Resources*, *20*(3), 718–731. https://doi.org/10.1111/1755-0998.13147

Orr, R. J. S., Zhao, S., Klaveness, D., Yabuki, A., Ikeda, K., Watanabe, M. M., & Shalchian-Tabrizi, K. (2018). Enigmatic Diphyllatea eukaryotes: culturing and targeted PacBio RS amplicon sequencing reveals a higher order taxonomic diversity and global distribution. *BMC Evolutionary Biology*, *18*(1), 115. https://doi.org/10.1186/s12862-018-1224-z

Pagenkopp Lohan, K. M., Fleischer, R. C., Carney, K. J., Holzer, K. K., & Ruiz, G. M. (2016). Amplicon-based pyrosequencing reveals high diversity of protistan parasites in ships' ballast water: implications for biogeography and infectious diseases. *Microbial Ecology*, *71*(3), 530–542. https://doi.org/10.1007/s00248-015-0684-6

Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J., & Esling, P. (2016). Protist metabarcoding and environmental biomonitoring: Time for change. *European Journal of Protistology*, *55*(Pt A), 12–25. https://doi.org/10.1016/j.ejop.2016.02.003

Piredda, R., Tomasino, M. P., D'Erchia, A. M., Manzari, C., Pesole, G., Montresor, M., Kooistra, W. H. C. F., Sarno, D., & Zingone, A. (2017). Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. *FEMS Microbiology Ecology*, *93*(1), fiw200. https://doi.org/10.1093/femsec/fiw200

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590–D596. https://doi.org/10.1093/nar/gks1219

Questel, J. M., Hopcroft, R. R., DeHart, H. M., Smoot, C. A., Kosobokova, K. N., & Bucklin, A. (2021). Metabarcoding of zooplankton diversity within the Chukchi Borderland, Arctic Ocean: improved resolution from multi-gene markers and region-specific DNA databases. *Marine Biodiversity: A Journal of the Senckenberg Research Institute/Senckenberg Forschungsinstitut Und Naturmudeum*, *51*(1), 1–19. https://doi.org/10.1007/s12526-020-01136-x

Santoferrara, L., Burki, F., Filker, S., Logares, R., Dunthorn, M., & McManus, G. B. (2020). Perspectives from ten years of protist studies by high-throughput metabarcoding. *The Journal of Eukaryotic Microbiology*, *67*(5), 612–622. https://doi.org/10.1111/jeu.12813

Schloss, P. D. (2020). Reintroducing mothur: 10 Years Later. *Applied and Environmental Microbiology*, *86*(2), e02343-19. https://doi.org/10.1128/AEM.02343-19

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., & Chen, W., Fungal Barcoding Consortium, & Fungal Barcoding Consortium Author List (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), 6241–6246. https://doi.org/10.1073/pnas.1117018109

Schwelm, A., Berney, C., Dixelius, C., Bass, D., & Neuhauser, S. (2016). The large subunit rDNA sequence of *Plasmodiophora brassicae* does not contain intra-species polymorphism. *Protist*, *167*(6), 544–554. https://doi.org/10.1016/j.protis.2016.08.008

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H.-W., & Richards, T. A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, *19*(Suppl 1), 21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x

Taib, N., Mangot, J.-F., Domaizon, I., Bronner, G., & Debroas, D. (2013). Phylogenetic affiliation of SSU rRNA genes generated by massively parallel sequencing: new insights into the freshwater protist diversity. *PLoS One*, *8*(3), e58950. https://doi.org/10.1371/journal.pone.0058950

Tanabe, A. S., Nagai, S., Hida, K., Yasuike, M., Fujiwara, A., Nakamura, Y., Takano, Y., & Katakura, S. (2016). Comparative study of the validity of three regions of the 18S-rRNA gene for massively parallel sequencing-based monitoring of the planktonic eukaryote community. *In Molecular Ecology Resources*, *16*(2), 402–414. https://doi.org/10.1111/1755-0998.12459

Tedersoo, L., & Anslan, S. (2019). Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences. *Environmental Microbiology Reports*, *11*(5), 659–668. https://doi.org/10.1111/1758-2229.12776

Tedersoo, L., Tooming-Klunderud, A., & Anslan, S. (2018). PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *The New Phytologist*, *217*(3), 1370–1385. https://doi.org/10.1111/nph.14776

Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, *5*, 170203. https://doi.org/10.1038/sdata.2017.203

Valero-Mora, P. M. (2010). ggplot2: elegant graphics for data analysis. *Journal of Statistical Software, Book Reviews*, *35*(1), 1–3. https://doi.org/10.18637/jss.v035.b01

Vaulot, D., Geisen, S., Mahe, F., & Bass, D. (2021). PR2-primers: an 18S rRNA primer database for protists. *bioRxiv*. https://doi.org/10.1101/2021.01.04.425170v1

Vegan: Community ecology package. (n.d.). Retrieved June 29, 2021, from https://cran.r-project.org/package=vegan

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read

sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Westbrook, C. J., Karl, J. A., Wiseman, R. W., Mate, S., Koroleva, G., Garcia, K., Sanchez-Lockhart, M., O'Connor, D. H., & Palacios, G. (2015). No assembly required: Full-length MHC class I allele discovery by PacBio circular consensus sequencing. *Human Immunology*, *76*(12), 891–896. https://doi.org/10.1016/j.humimm.2015.03.022

Wilkins, D., van Sebille, E., Rintoul, S. R., Lauro, F. M., & Cavicchioli, R. (2013). Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects. *Nature Communications*, *4*, 2457. https://doi.org/10.1038/ncomms3457

Worden, A. Z., Follows, M. J., Giovannoni, S. J., Wilken, S., Zimmerman, A. E., & Keeling, P. J. (2015). Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science*, *347*(6223), 1257594. https://doi.org/10.1126/science.1257594

Wright, E. (2016). Using DECIPHER v2.0 to analyze big biological sequence data in R. *The R Journal*, *8*(1), 352. https://doi.org/10.32614/RJ-2016-025

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., Whitman, W. B., Euzéby, J., Amann, R., & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, *12*(9), 635–645. https://doi.org/10.1038/nrmicro3330

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Latz, M. A. C., Grujcic, V., Brugel, S., Lycken, J., John, U., Karlson, B., Andersson, A., & Andersson, A. F. (2022). Short- and long-read metabarcoding of the eukaryotic rRNA operon: Evaluation of primers and comparison to shotgun metagenomics sequencing. *Molecular Ecology Resources*, *22*, 2304–2318. https://doi.org/10.1111/1755-0998.13623