Long-read Sequencing of Eukaryotic rDNA Operon from the Singapore Marine Straits

Loh Dan Hong[1], Christaline George[2] and Adriana Lopes dos Santos[2]
1. School of Biological Sciences, 2. Asian School of Environment

Abstract **–** High-throughput environmental DNA sequencing, also called metabarcoding, has largely contributed to the discovery and description of microbial communities worldwide, from humans to the deepest parts of our oceans. Metabarcoding today is limited to an amplicon sequence of approximately 600 base pairs. Although these short sequences contain enough taxonomic resolution, they have limited phylogenetic signal and can hide cryptic diversity. New long-read sequencing technologies such as the Pacific Biosciences and Nanopore platforms seem to provide a solution for these problems by producing reads that are over 20kb and at a cost cheaper than that of Sanger sequencing. While promising, long read sequencing (as metabarcoding) is highly dependent on the taxon sampling and annotation accuracy of the reference database. For marine microeukaryotes, the focus of this work, both PR2 and SILVA databases are heavily populated with short and long 18S gene sequences and only a few species and groups have their remaining ribosomal genes sequenced (e.g. large submit ribosomal gene 28S and internal transcribed spacers). In order to provide long-read reference sequences from phytoplankton species from Singapore waters, we applied Sanger sequencing technology with a combination of 12 different primers covering the whole 18S, ITS 1 and 2 gene and partial 28S gene region to strains isolated from Singapore waters. In total, 16 strains were sequenced- 10 of which are from Singapore Marine Straits, and 6 from other culture collections. Of these, the full rDNA operon was obtained for 7 strains, 18S and 28S sequences were obtained separately for 3 strains, the 18S sequence only was obtained for 3 strains, and no sequences were obtained for 3 strains. Other methods to obtain the rDNA gene will be needed for successful long-read sequencing- this will increase phylogenetic resolution and capture the extent of diversity of marine microeukaryotes.

Keywords – microeukaryotes, long-reads, Sanger sequencing, rDNA, phylogenetic resolution, reference sequences

1 INTRODUCTION

Marine unicellular eukaryotes, also known as protists, are taxonomically diverse with taxa distributed widely across several branches of the eukaryotic tree of life [1]. The pelagic marine ecosystem host a great diversity of planktonic microeukaryotes which include autotrophic cells that can fix CO_2 into organic matter, heterotrophic cells with predatory behavior such as ciliates, and mixotrophic organisms that can switch between photosynthetic and heterotrophic modes of growth such as dinoflagellates [2]. According to global [3] and local [4] metabarcoding surveys, Singapore waters harbor photosynthetic microeukaryotes that have yet to be described.

Marine microeukaryotes play important roles in the ecosystem; it is essential to the carbon cycle [5], but can also form harmful algal blooms that are deleterious to other marine organisms, and have negative impacts on fisheries and aquaculture industry [6]. It is hence important to study and understand the diversity and functions of marine microeukaryotes.

Sequencing of environmental DNA is a popular method to study the diversity of microbes, including microeukaryotes. These methods target the sequencing of genetic markers – ideally a segment of DNA present in all organisms that is highly variable between individuals of different species but similar among individuals of the same species [7]. Traditionally, the genes within the rDNA operon (or cluster) have been used as genetic markers to identify microbes.

Across all three domains of life, the ribosomal RNA is transcribed from ribosomal DNA (rDNA) operon as a single molecule of RNA. During a post-transcriptional process, the secondary structure assumed by the newly synthesized rRNA guides

cellular enzymes that will release the RNAs that will form the small and large ribosome subunits. For eukaryotes, for example, the rRNA operon contains: the 18S rDNA gene, the internal transcribed spacer (ITS) 1, the 5.8S rDNA gene, ITS 2, and the 28S rDNA genes arranged in tandem. Together with proteins, the 18S and 28S RNAs form the small and large ribosome subunits. 8 (V1-V9, excluding V6, which is relatively conserved in eukaryotes) and 12 (D1-D12) hypervariable regions are found in the sequences of the 18S [8] and 28S [9] RNAs respectively.

Traditionally, metabarcoding (metaB) studies of marine planktonic microeukaryotes have targeted the 18S rDNA gene hypervariable regions 4 (V4) and 9 (V9). Two main reasons have led the use of these variable regions; the taxonomic resolution [10] combined with read length. The read length (in base pairs) of these regions are around 200 for V9 and 450 for V4, which allows the use of high throughput sequencing technologies such as Illumina [11]. Although these short sequences contain enough taxonomic resolution to distinguish different lineages, classes and to a lower extent, genera, they have limited phylogenetic signal and often are unable to distinguish amongst very close species, underestimating the cryptic diversity present in the environment. New long-read sequencing (LRS) technologies such as the Pacific Biosciences and Nanopore platforms seem to overcome these issues by producing reads that are over 20kb and at a cost cheaper than that of Sanger sequencing. By using these new sequence technologies, the entire rDNA operon, including the 28S and ITS regions, can be sequenced at once giving rise to better phylogenetic resolution of the players in a given environment [4].

However, interpretation of such data (metaB of short variable regions or LRS of environmental DNA) is highly dependent on the taxon sampling and annotation accuracy of the reference databases. The taxonomic affiliation of each sequence generated through these methods is obtained by comparing against a reference database. The reference database for eukaryotes, PR2 [12] for example, consists of nuclear-encoded sequences originating from the environment or isolated strains which were carefully annotated by taxonomist experts.

For microeukaryotes, both PR2 and SILVA [13] databases are heavily populated with 18S rDNA gene references sequences, both partial and complete, but few sequences of the 28S rDNA gene and ITS. In order to provide long-read reference sequences from phytoplankton species from Singapore waters, we applied Sanger sequencing technology with a combination of 12 different primers covering the whole 18S, ITS 1 and 2 gene and partial 28S gene region. We used some strains isolated from Singapore waters as well as strains obtained from other locations.

2 MATERIAL AND METHODS

2.1 CULTURING

Cultures were grown in untreated ventilated T-25 flasks with 15 mL of L1 media [14]. All cultures were incubated at 22°C with a 14:10 light–dark cycle and transferred to new medium for at least a week or two, before DNA extraction. Light intensity was approximately 100 μmol photons.$m^{-2}.s^{-1}$. Details of the cultures used are described in Table 1.

2.2 DNA EXTRACTION

Cells were harvested by centrifugation from 2 mL of fresh culture, with the addition of 0.5 μL of Pluronic F-127. The cultures were centrifuged at 11000 g for 1 minute. The supernatant was discarded in 10% bleach, and

Table 1 Strains and species used for PCR optimization and sequencing in this project.

Class	Species	Strain	Location of Isolation
Dinophyceae	*Amphidinium carterae*	CS-21	Halifax, Canada
Raphidophyceae	*Fibrocapsa japonica**	DHI	Singapore
Dinophyceae	*Karenia mikimotoi*	RCC1513	Rance Estuary, France
Bacillariophyceae	*Phaeodactylum tricornutum*	RCC2967	North Atlantic Ocean, UK
Trebouxiophyceae	*Picochlorum costavermella*	RCC4223	Massane River, France
Trebouxiophyceae	*Picochlorum sp**	SENEW3	San Elijo Lagoon, California, USA
Dinophyceae	*Scripsiella sp.^*	-	Singapore

Raphidophyceae	*Fibrocapsa japonica*	RCC1501, SMS1	English Channel, France
Raphidophyceae	*Heterosigma akashiwo*	RCC1502, SMS2	French Coast, France
Dinophyceae	*Symbiodinium sp*	RCC4010, SMS5	Coral Sea, Australia
Coccolithophores	*Emiliania huxleyi*	RCC1731, SMS7	South Pacific Ocean, Ecuador
Chloropicophyceae	*Chloropicon roscoffensis*	RCC2335, SMS14	Sagami Bay, Japan
Prasinophyceae	*Tetraselmis sp*	SMS19	Singapore
Coscinodiscophyceae	*Thalassiosira sp*	SMS39	Singapore
Trebouxiophyceae	*Picochlorum sp*	SMS40	Singapore
Bacillariophyceae	*Cylindrotheca sp*	SMS41	Singapore
Bacillariophyceae	*Nitzschia sp*	SMS45	Singapore

*Extracted DNA provided by mentor Christaline George

^Only used during PCR optimization. Given by Tropical Marine Science Institute at St John Island.

the cells were re-suspended in the remaining media. DNA was then extracted from the pellet with the Nucleospin Plant II Kit (Machery-Nagel), according to protocol from the manual. The concentration of the extracted DNA was determined with an Invitrogen Qubit Fluorometer (Thermo Fisher Scientific).

2.3 PCR OPTIMIZATION

Polymerase Chain reaction (PCR) optimization was done using previously extracted DNA (Table 1). The amplification conditions of the rDNA operon were tested by using 1 μL of 10 μM 63f forward primer [15] and 1 μL of 10 μM nH2R [16], 22R [16], LSUR2 [17], or R1318 [18] reverse primers (Table 2), with or without 0.125 μL of 20mg/ml Bovine Serum Albumin (BSA). The PCR reaction consisted of 5 μL of 5x Long Amp Taq Buffer (New England Biolabs), 1 μL Long Amp Taq Polymerase (New England Biolabs), 0.75 μL of 10μM dNTP (New England Biolabs), 5 μL of 25mM $MgCl_2$. DNA template (10 – 20ng per reaction), and nuclease-free water MilliQ were adjusted accordingly to a final 25 μL reaction volume. Thermal conditions were: initial denaturation for 4 minutes at 94℃; 30 cycles of denaturation at 94°C for 30 seconds, annealing at 58℃ for 30 seconds, and extension at 65°C for 7 minutes; final extension at 65℃ for 6 minutes for LSUR2 or R1318 reverse primers and 10 minutes for 22R or nH2R reverse primers .The final product was held at 8℃ until removed from the thermocycler. The amplified products were visualised by gel electrophoresis using the Invitrogen E-gel Ex 2% agarose gel (Thermo Fisher Scientific). Each well was loaded with 5 μL of amplified products and 15 μL of MilliQ, comparing against 20 μL of E-gel 1kb Plus DNA ladder (Thermo Fisher Scientific). The E-gel was visualized with E-gel Imager (Life Technologies). The ideal PCR condition was determined by the appearance of the bands of the amplicons using the following criteria:

I. Amplicons produced a band of the appropriate size that is clear and thick.
II. No unspecific bands of other sizes observed.

2.4 PURIFICATION AND QUANTIFICATION OF PCR PRODUCTS

Using the optimized PCR conditions, multiple rounds of PCR were performed for the selected 16 species (Table 1) in order to obtain enough mass for sequencing. All amplicons

Table 2 Primers used during PCR optimization and/or for bi-directional primer walking.

Forward Primer	Sequence	Region
63F [15]	5'ACGCTTGTCTCAAAGATTA3'	V1
V4F [11]	5'CAGCAGCCGCGGTAATTCC3'	V4
V5F[a]	5'GCTTAATTTGACTCAACACGGG3'	V5
V9F [19]	5'TTGTACACACCGCCC3'	V9
D1R-28S-F [20]	5'ACCCGCTGAATTTAAGCATA3'	D1
D1R-2CR-F[a]	5'TCTTGAAACACGGACCAAGG3'	D2
Reverse Primer	**Sequence**	**Region**
V4R[a]	5'GGAATTACCGCGGCTGCTG3'	V4
V5R [11]	5'CCCGTGTTGAGTCAAATTAAGC3'	V5
1818R [15]	5'ACGGAAACCTTGTTACGA3'	V9

D1R-R[a]	5'TATGCTTAAATTCAGCGGGT3'	D1
D1R-2CR [20]	5'CCTTGGTCCGTGTTTCAAGA3'	D2
LSUR2 [17]	5'ATTCGGCAGGTGAGTTGTTAC3'	D5
R1318 [18]	5'TCGGCAGGTGAGTTGTTACACAC3'	D5
22R [16]	5'CCATTCATGCRCGTCACWART3'	D9-D12
nH2R [16]	5'GAHHBARCKGTTCCTCTCGTACT3'	D12

[a] These primers were constructed by reverse complementing the original sequence.

successfully obtained for each species were pooled together and purified with QIAquick Purification Kit (QIAgen), according to protocol from the manual. The final concentrations of the purified products were measured with NanoDrop One Microvolume UV-Vis Spectrophotometer (Thermo Fisher Scientific).

2.5 SANGER SEQUENCING

The purified PCR products were sent for sequencing at Macrogen Singapore (https://dna.macrogen-singapore.com/eng/) using the 12 internal primers described in Table 2. The 18S rDNA gene from the different strains were sequenced with primers 63F, V4F, V5F, V9F, V4R, V5R, and 1818R, while the 28S rDNA gene sequences were sequenced with the D1R-F, D1R-2CR-F, D1R, D1R-2CR, and LSUR2 primers.

2.6 ASSEMBLY AND SEQUENCING ANALYSES

De Novo assembly on Geneious Prime (version 2019.2.3) [21] platform was performed with the reads from each species in order to build a consensus sequence of the nearly complete rDNA operon (~3.5 - 4.0Kb). The settings were configured such that Geneious would trim and annotate the 3' and 5' ends, and regions of the sequences with an error probability of >= 0.01. These trimmed regions were not considered in the final consensus sequenced. If the complete fragment was not obtained, *De Novo* Assembly was performed again in order to obtain the 18S and 28S rDNA gene separately.

Partial or complete sequences were compared with those available in GenBank using the *on line* nBLAST tool.

3 RESULTS

3.1 *IN SILICO* BI-DIRECTIONAL PRIMER WALKING

Since the length of rDNA operon is around 3.5 - 4.0kb and therefore longer than the maximum length of individual fragments that can be sequenced at once by Sanger sequencing, suitable primers for bi-directional primer walking were chosen *in silico* using Geneious "Test with saved primers" function. This function uses the sequences of primers to predict which region of the gene sequence the primers will anneal to. An ideal set of 12 primers were chosen producing fragments of 500-700bp that overlapped with one another (Table 2). Having sufficient overlap was crucial to guide the assembly of the sequenced fragments and to produce the complete rDNA sequence. A representation of the annealing position of the selected primers on rDNA operon sequence of *Picochlorum sp.* SENEW3 which was retrieved from NCBI (accession number: JPID01000156.1) is shown in Figure 1.

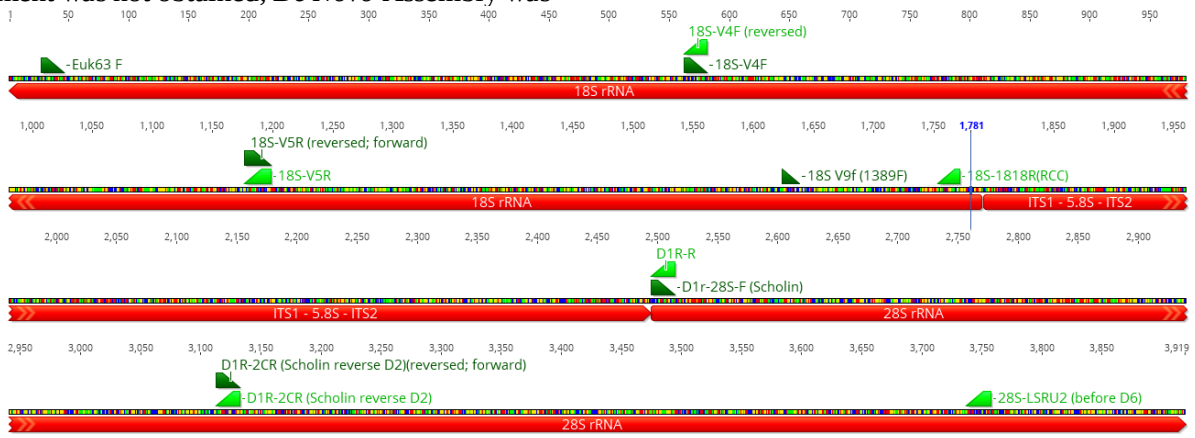

Figure 1 Representation of the 12 selected primers annealing positions on Picochlorum sp. SENEW3 rDNA operon. Dark green annotations are forward primers, while light green ones are reverse primers.

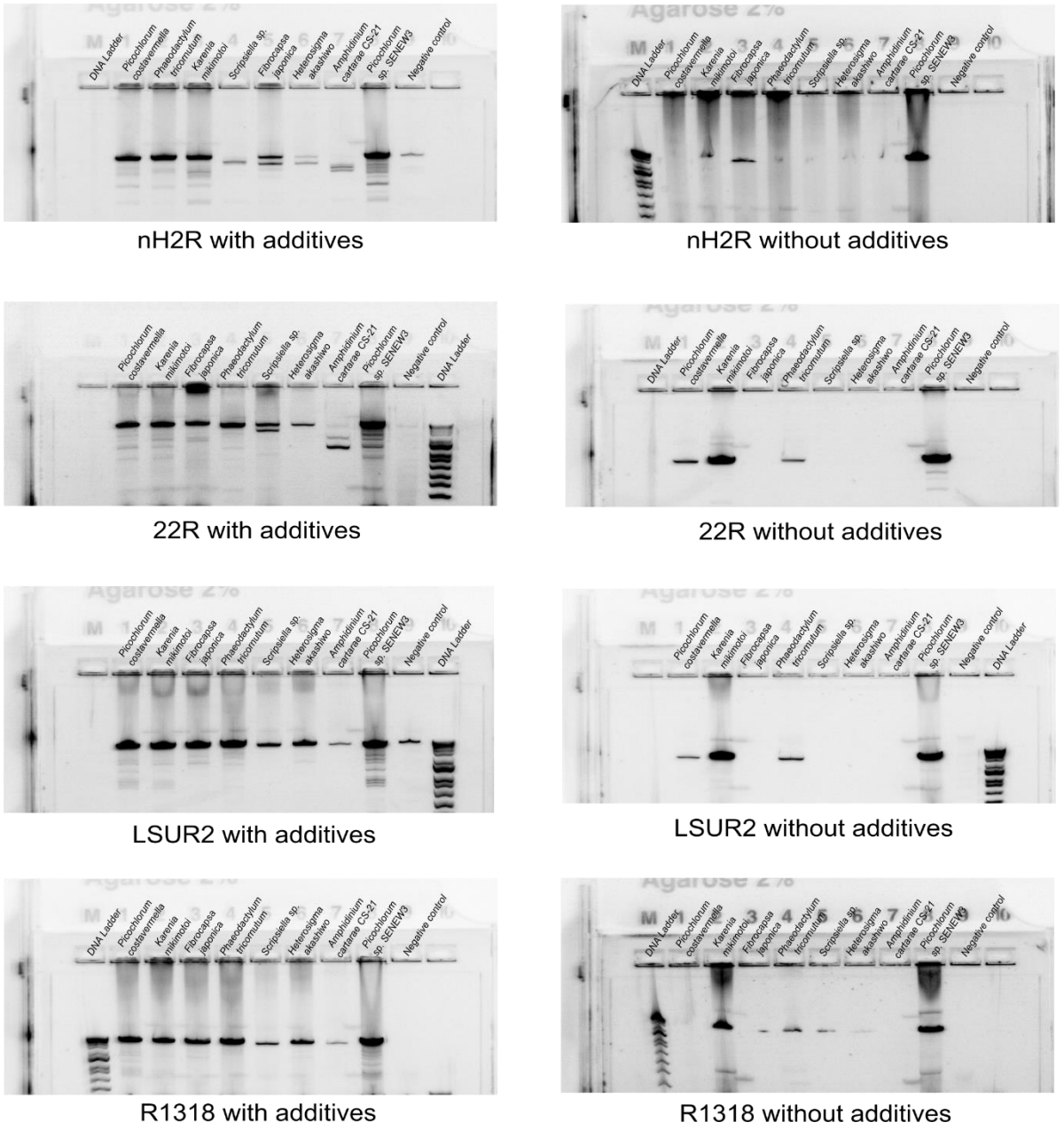

Figure 2 Gel electrophoresis images from PCR optimization. 8 conditions were tested, with the 4 reverse primers, with and without additives.

3.2 PCR OPTIMIZATION AND SEQUENCING QUALITY

During the optimization of the PCR reactions we tested the choice of reverse primers (nH2R, 22R, LSUR2, or 1318 – Table 2), which determines the length of rDNA operon amplified, and the addition of BSA to reduce unspecific amplification.

From the gel electrophoresis images (Figure 2), we observed that the PCR reactions without BSA only produced amplicons, from a few species, particularly those in lane 1 (*Picochlorum costavermella* RCC4223), lane 2 (*Karenia mikimotoi* RCC1513), lane 4 (*Phaeodactylum tricornutum* RCC2967), and lane 8 (*Picochlorum sp*. SENEW3*)*. The addition of BSA to the reactions was critical to produce amplicons for all species tested (Figure 2).

Comparing the different reverse primers with 63F, both nH2R and 22R did produce amplicons of the correct size for some species (~4.5kb). However, the amplification with nH2R generated unspecific bands for the strains *Scripsiella sp*. (lane 4), *Heterosigma akashiwo* RCC1502 (lane 6) and *Amphidinium carterae* CS-21 (lane 7), and multiple faint bands for other species. With 22R, the reactions for *Amphidinium carterae* CS-21 (lane 7) resulted in bands with incorrect size and multiple faint, random bands for the other species. Both LSUR2 and R1318 amplification resulted in bands with correct size, high yield for most species and fewer faint and unspecific bands.

Based on these results, since LSUR2 and R1318 reactions with BSA produced the best PCR results, and these primers amplified the same gene region within the 28S rDNA (D5), we decided to proceed with LSUR2 with BSA to amplify the near full-length rDNA operon of selected strains.

We observed the amplification of contaminants in the negative control of the reactions with nH2R, 22R, and LSUR2 and BSA. These contaminations were likely caused by DNA contaminants in the lab environment. To resolve this problem, we changed our protocol for preparing the PCR reaction mixture in a PCR work station free of DNA and adding DNA templates in the BioSafety Cabinet, instead of doing so on the bench. Contaminants in the negative control were completely absent in subsequent PCR runs.

Sanger sequences of the DNA fragments from Macrogen had varying percentages of untrimmed bases which are of high quality, ranging from 0.0% to 99.7%, with a mean of 72.8%. DNA fragments from the same species generally had similar percentages. *De Novo* assembly of the DNA fragments in Geneious formed contigs of the 18S and 28S regions for most species, with the exception of *Fibrocapsa japonica* SMS 1, *Emiliania huxleyi* SMS 7, and *Karenia mikimotoi* RCC1513, which did not have contigs formed at all, due to the low quality of their sequenced DNA fragments. For SMS2, SMS5, and SMS40, only contigs for their 18S rDNA gene were formed. For SMS 2 and SMS 40, only one directional primer walking was performed with only the 6 reverse primers (Table 2) due to the low yield of their rDNA amplicons. Hence it was less likely for their 28S contig to be formed. Out of the 16 species, only 7 species had contigs formed across the whole length of the rDNA operon- *Chloropicon roscoffensis* SMS 14, *Tetraselmis sp.* SMS 19, *Cylindrotheca sp*. SMS 41, *Amphidinium carterae* CS-21, *Fibrocapsa japonica* DHI, *Phaeodactylum tricornutum* RCC2967, *and Picochlorum sp.* SENEW3.

The summaries of the contigs formed and their BLAST results against the NCBI database is shown in Table 3. Some BLAST searches yielded ambiguous results, with more than 1 genus or species listed, which mostly had high percentage identity and E value of 0, which is the number of expected hits of similar quality that can be found by chance.

4 DISCUSSION

The results of the amplification shown with the gel electrophoresis (Figure 2) that the addition of BSA was essential to amplifying our target genomic region. BSA has been shown to bind to phenols, lipids, anions, and compounds that have endogenous protease activity [22], [23]. These are compounds that could disrupt the activity of DNA polymerase and inhibit PCR, and hence BSA is essential to reduce inhibition and increasing yields of PCR products.

Table 3 BLAST results for Sanger sequences

Species	Region	Primers	Consensus length (bp)	Matches from BLAST	E value	Query cover (%)/ Per. Ident. (%)	Accession number
Fibrocapsa japonica SMS 1				No contigs found			
Heterosigma akashiwo SMS 2	18S	V4R, V5R, 1818R	1523	*Heterosigma akashiwo* 18S ribosomal gene, partial sequence, isolate RCC1502 (acc. no.: LC214052.1)	0.0	100/99.61	MT489358
	28S			No contigs found			
Symbiodinium sp. SMS 5	18S	V4F, V5R, 1818R	469	*Symbiodinium sp.* CBr-I1 18S rRNA gene, partial sequence, strain JCUSG-1 (acc. no.: AB016723.1)	0.0	100/100	MT489386
	28S			No contigs found			
Emiliania huxleyi SMS 7				No contigs found			
Chloropicon roscoffensis SMS 14	18S	all	2404		0.0	95/99.01	
	28S	all	2315	*Chloropicon primus* CCMP1205 chromosome 6 (acc. no.: CP031039.1)	0.0	95/94.96	MT489379
	Whole	all	3505		0.0	97/99.01	
Tetraselmis sp. SMS 19	18S	all	2147	*Tetraselmis sp.* CCAP 66/72 small subunit ribosomal RNA gene, partial sequence (acc. no.: MG022701.1)	0.0	100/99.91	
	28S	All except D1R	1435	*Tetraselmis striata* 18S rRNA gene (partial), 5.8S rRNA gene, 28S rRNA gene (partial), ITS1 and ITS2, strain SAG 41.85 (acc. no.: HE610129.1)	0.0	99/96.50	MT489354 MT489359 MT489380
	Whole	All except D1R	3575	*Tetraselmis sp.* CCAP 66/64 small ribosomal subunit RNA gene, partial sequence (acc. no.: MG022702.1)	0.0	65/99.87	
Thalassiosira sp. SMS 39	18S	all	2256	*Thalassiosira allenii* 18S ribosomal RNA gene, partial sequence (acc. no.: HM991688.1)	0.0	78/99.94	MT489360 MT489355
	28S	all	1525	*Thalassiosira allenii* 28S ribosomal RNA gene, partial sequence (acc. no.: HM991673.1) (results contain a variety of species and genera, like *Thalassiosira*, *Stephanodiscus*, *Discostella*, *Cyclotella* etc.)	0.0	61/99.47	
Picochlorum sp. SMS 40	18S	V4R, V5R, 1818R	1646	*Picochlorum sp.* PACC8946 small subunit ribosomal RNA gene, partial sequence (acc. no.: MN088860.1) (results contain *Picochlorum*, *Nannochloris*, and *Nanochlorum*, many species)	0.0	100/99.57	MT489361 MT489356
	28S			No contigs found			
Cylindrotheca sp. SMS 41	18S	all	2729	*Cylindrotheca closterium* strain MGB0501 18S rRNA gene partial sequence (acc. no.: DQ019446.1)	0.0	63/99.94	
	28S	all	2300	*Cylindrotheca closterium* ITS1, 5.8S ribosomal RNA gene, ITS2, and 28S ribosomal RNA gene, complete sequence	0.0	83/96.25	MT489381
	Whole	all	3630	(acc. no.: AF289049.1)	0.0	52/96.25	
Nitzschia sp. SMS 45	18S	all	2012	*Nitzschia microcephala* strain Som 18S ribosomal RNA gene, partial sequence; IITS 1, 5.8S ribosomal RNA gene, and ITS 2, complete sequence;28S ribosomal RNA gene, partial sequence (acc. no.: KC759159.1) (results contain mostly *Nitzschia* and *Cymbella* genera)	0.0	90/98.85	MT489357 MT489362
	28S	Except D1R-F and D1R	1235	*Nitzschia pusilla* partial 28S rRNA strain L1 (acc. no.: HF679193.1)	0.0	60/99.34	
Amphidinium carterae CS-21	18S	all	2685	*Amphidinium sp.* CCAP 1102/5 genomic DNA containing 18S rRNA gene, ITS1, 5.8S rRNA gene, ITS2, 28S rRNA gene (acc. no.: FR865623.1)	0.0	96/99.85	MT489382
	28S	all	2342		0.0	63/99.93	
	Whole	all	3466		0.0	75/99.88	
Fibrocapsa japonica DHI	18S	all	2516	*Fibrocapsa japonica* strain RCC1501 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S	0.0	100/99.64	
	28S	Except D1R	1400	ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial	0.0	53/99.47	
	Whole	Except D1R	3909	sequence (acc. no.: KP780264.1)	0.0	83/99.64	
Karenia mikimotoi				No contigs found			
Picochlorum costavermella RCC4223	18S	all	2221	*Picochlorum sp.* RCC13 18S ribosomal RNA gene, partial sequence (acc. no.: KT860853.1)	0.0	73/100	MT489383
	28S	all	1576	*Chaetophora pisiformis* isolate XR201704 SSU RNA gene, partial sequence; ITS1, 5.8S ribosomal RNA gene, and ITS 2, complete sequence; and LSU RNA gene, partial sequence (acc. no.: MH002628.1) (ambiguous results, contain *Chlorella*, *Chaetophora*, *Miractidinium* and other genera)	0.0	81/86.77	
Phaeodactylum tricornutum RCC2967	18S	all	2734	*Phaeodactylum tricornutum* strain CCAP 1055/1 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence (acc. no.: EF553458.1)	0.0	100/99.67	MT489384
	28S	all	2336		0.0	100/99.79	
	Whole	all	3873		0.0	100/99.97	
Picochlorum sp. SENEW3	18S	all	2222	*Picochlorum sp.* SENEW3 18S ribosomal RNA gene, partial sequence (acc. no.: KF591594.1)	0.0	77/99.83	MT489385
	28S	all	2278	Pseudochlorella pringsheimii 18S ribosomal RNA gene, ITS 1, 5.8S ribosomal RNA gene, ITS 2, 28S ribosomal RNA gene, complete sequence (acc. no.: KY364701.1) (ambiguous results, contain *Chlorella*, *Chaetophora*, *Micractinium* and other genera)	0.0	78/86.70	
	Whole	all	3708	*Picochlorum sp.* SENEW3 18S ribosomal RNA gene, partial sequence (acc. no.: KF591594.1)	0.0	46/99.83	

The different reverse primers used resulted in different lengths of target gene sequences being amplified; nH2R, which anneals to the end of the 28S region, resulted in an amplicon of ~4500bp; 22R anneals to around 2/3rd of the 28S region, and produced an amplicon of ~4200bp; LSRU2 and R1318 anneal to the end of 28s rDNA D5 region, producing a PCR product of ~3700bp. It should be noted that the actual lengths of the amplicons vary among the different species. Length variation arises from hypervariable regions of 18S and 28S. V2, V4, and V7 vary most in length for 18S, which can vary from 1.5kb to 4.5kb [24]. As for 28S, D3 is the most length-variable region [25].

PCR products of shorter length and random, faint bands corresponding to smaller sizes were produced with the primers nH2R and 22R. Primer degeneracy could be the main factor in the decrease in specificity of these reverse primers. The sequence of nH2R and 22R primers contain 5 and 2 degenerate nucleotides respectively, while those of LR and 1318 have no degenerate nucleotides (Table 2). While primer degeneracy allows for primers to anneal to the DNA of broad number of species, it could also cause primers to be less specific and bind to other regions in the genome [26], contributing to the amplification of more unspecific and fragments. Using a universal primer that is more specific, or designing degenerate primers based on more effective algorithms can overcome this issue [26].

The low quality sequences of *Fibrocapsa japonica* SMS 1, *Emiliania huxleyi* SMS 7, *Karenia mikimotoi*, and DNA fragments of other species could be due to a number of reasons. The culture for *Emiliania huxleyi* SMS 7 was subsequently found to be mixed with another microeukaryote of a completely different lineage, resulting in 2 different rDNA operon being sequenced. The high variation between the 2 sequences would have led to different signals arising during Sanger sequencing at the same nucleotide position, which explains the low quality sequences obtained. However, for other species, different methods may have to be used to improve the quality of sequences.

For DNA extraction of our species, a commercially-available kit was used, with the use of a lysis buffer that is based on the CTAB DNA isolation method to lyse the cell walls. However, efficiency of DNA extraction can be increased by using different methods of cell breakage, as different species vary in their cell covering, from some having no cell wall at all, to some having fortified cell walls, such as silica frustules on diatoms [27]. For example, *Tetraselmis sp*. SMS 19 is known to have a hard thecate cell wall with a unique composition that makes DNA isolation difficult [28], [29]. This is seen in how the DNA extracted from it has a very low concentration of 0.934 ng/μL, as compared to other species which have concentrations that are 4 to 5 times more. Similarly, DNA extracted from diatoms like *Phaeodactylum tricornutum* RCC2967 have low concentrations of 1.09 ng/μL as well. Subsequently, this will affect extracted DNA and PCR product yield. Hence, a preparatory light microscopic analysis of different species could be performed to determine what method of cell breakage can be used for each species for DNA extraction, such as repeat freeze-thaw, liquid nitrogen grinding, and sonication. Most notably, incubation in a lysis buffer for 3 days, before a round of bead-beating to further break any intact cells have been proven to effectively break cells while preserving the integrity of DNA across all species, making it suitable for unidentified species [27]. While the commercial DNA kit used is targeted at plant cells which have stronger cell walls than marine microeukaryotes in general, method of cell lysis should still be considered subsequently when different species of marine microeukaryotes are studied.

For some groups of micro-eukaryotes, identical copies of the rDNA operon exist in the nuclear genome [30]. However, in cases like *Fibrocapsa japonica*, it has been shown that some strains have polymorphism within their ITS region of the rDNA, resulting in copies where the ITS regions vary from other rDNA copies [31]. This fact could explain the low quality of sequences produced for *Fibrocapsa japonica* SMS 1. The polymorphism within their ITS region could give rise to different signals during Sanger sequencing at the same nucleotide position, lowering the quality of the final sequence. Species that have intra-individual polymorphism at certain regions of their rDNA would hence be harder to characterize and have their rDNA sequenced successfully. To overcome this issue, DNA cloning could be used such that each variation of the rDNA can be amplified and sequenced separately. By using PCR to amplify the rDNA of a given species first, the different variations of rDNA fragments can then be inserted into plasmids that are transformed into a species of bacteria such as *Escherichia coli*. Each bacteria cell can then propagate into colonies; with each colony having plasmids with one variation of the rDNA. These rDNA fragments can then be extracted and purified from each colony and subsequently sent for

sequencing. This allows every variation of rDNA to be sequenced separately without contrasting signals using Sanger sequencing technology enabling to capture the polymorphisms along the length of rDNA. However, given that *Fibrocapsa japonica DHI* was sequenced successfully, the low quality sequences of *Fibrocapsa japonica* SMS 1 could probably be due to contamination by another microeukaryote as well, as in the case of *Emiliania huxleyi* SMS 7.

Several BLAST results do not give a clear identity to the different species, especially when BLAST was performed only with the 18S or 28S rDNA gene. The taxonomy is only resolved up to the genus level, with results ranging across species; some results even range across genera. This can be seen in the BLAST results from the 28S consensus of *Thalassiosira sp*. SMS 39, which ranges across genera like *Thalassiosira*, *Stephanodiscus*, *Discostella*, and *Cyclotella*; these 4 genera belong to the same order of *Thalassiosirales* [32] and hence similar regions in the 28S gene sequence could have aligned to those of other genera. As for *Nitzschia sp*. SMS 45, the BLAST results of its 18S consensus ranged across *Nitzschia* and *Cymbella* genera. While both are diatoms, they are of different orders and hence it is unlikely their 18S gene sequence is highly similar; the conflictual results could be due to the sequences being assigned a wrong ID on the NCBI GenBank database.

As sequencing the complete or partial 18S rDNA gene is a common way to characterize microeukaryotes, there is a disproportionate abundance of 18S rDNA gene sequences present in the NCBI database which may result in different BLAST results for the 18S and 28S rDNA gene sequences. For example, the BLAST results for *Picochlorum costavermella* RCC4223 and *Picochlorum sp*. SENEW3 28S rDNA sequence matched several sequences belonging to another phytoplankton called *Chlorella*. Both *Chlorella* and *Picocholorum* belong to the same order Chlorellales [33] and it is likely that there are no 28S sequences present in the database for *Picochlorum* species.

While using the complete rDNA sequence could increase phylogenetic resolution of organisms, there are still obstacles that need to be overcome in order to obtain it. 18S and 28S rDNA of most species were easily assembled in Geneious, but the complete rDNA of only 7 species were able to be assembled from the DNA fragments. Furthermore, DNA fragments that were of low quality and hence unable to be built into contigs are mostly formed by the D1R primer, a reverse primer that synthesizes from the ITS2 region onwards (Figure 1). Though V9F, a forward primer that synthesizes part of the ITS regions as well,- formed higher quality contigs, only about 1/3 of the sequence remained after low quality bases were trimmed away by Geneious. These lend further weight to the fact that ITS regions are often unwieldy and hard to amplify and sequence in the case of marine microeukaryotes, and could be due to intraspecific polymorphism of the ITS region. Other methods of obtaining the ITS regions more effectively when amplifying the rDNA operon will be needed to work around this problem.

5 CONCLUSION AND FUTURE WORK

Contributing near full-length rDNA sequences in databases not only helps with identifying marine microeukaryotes, but can potentially assist in the use of Third Generation Sequencing technologies such as Nanopore. While Nanopore can perform high-throughput real-time sequencing of long sequences unlike Sanger sequencing, which can only sequence DNA of a certain length, the error rates associated with Nanopore are still higher than Sanger sequencing. With a database of full-length rDNA operons sequenced by Sanger sequencing, sequences by Nanopore could be compared against the database to determine its error rate and further refine the assignation method, thus allowing for direct use of Nanopore sequencing in subsequent projects.

However, we were unable to obtain full rDNA sequences for all species. Another method of obtaining rDNA for sequencing may therefore prove to be more effective across all species- by isolating rRNA from harvested total RNA, and creating a cDNA library for the rDNA and sequencing it [34], avoiding possible primer biases, and allowing for the full rDNA to be amplified as well. Nevertheless, using this method will add to the complexity of lab work involved, and require more time as compared to the method used in this project. Further research will be needed to determine what method will be most efficient and effective in long-read sequencing of rDNA operons.

ACKNOWLEDGMENT

I would like to thank my mentor, Christaline George, and my supervisor, Associate Professor Adriana Lopes dos Santos, for their invaluable

guidance over the course of this project, and for teaching me valuable skills needed in research.

I also appreciate the funding and opportunity to take up a research project from the CN Yang Scholars Programme.

I would also like to thank Associate Professor Federico Lauro and the staff at the Singapore Laboratory for Integrative Microbial Ecology for the use of their resources and equipment in the lab.

REFERENCES

[1] S. Ohtsuka, T. Suzaki, T. Horiguchi, N. Suzuki, and F. Not, Eds., *Marine Protists: Diversity and Dynamics*. Springer Japan, 2015.

[2] F. Not, R. Siano, W. H. C. F. Kooistra, N. Simon, D. Vaulot, and I. Probert, "Diversity and Ecology of Eukaryotic Marine Phytoplankton," in *Advances in Botanical Research*, vol. 64, Elsevier, 2012, pp. 1–53.

[3] M. Tragin and D. Vaulot, "Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding," *Sci. Rep.*, vol. 9, no. 1, pp. 1–14, Mar. 2019, doi: 10.1038/s41598-019-41680-6.

[4] M. Jamy *et al.*, "Long metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity," *Mol. Ecol. Resour.*, 2020, doi: 10.1111/1755-0998.13117.

[5] A. Z. Worden, M. J. Follows, S. J. Giovannoni, S. Wilken, A. E. Zimmerman, and P. J. Keeling, "Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes," *Science*, vol. 347, no. 6223, Feb. 2015, doi: 10.1126/science.1257594.

[6] J. H. Landsberg, "The Effects of Harmful Algal Blooms on Aquatic Organisms," *Rev. Fish. Sci.*, vol. 10, no. 2, pp. 113–390, Apr. 2002, doi: 10.1080/20026491051695.

[7] G. J. Olsen, D. J. Lane, S. J. Giovannoni, N. R. Pace, and D. A. Stahl, "Microbial Ecology and Evolution: A Ribosomal RNA Approach," *Annu. Rev. Microbiol.*, vol. 40, no. 1, pp. 337–365, 1986, doi: 10.1146/annurev.mi.40.100186.002005.

[8] J.-S. Ki, "Hypervariable regions (V1–V9) of the dinoflagellate 18S rRNA using a large dataset for marker considerations," *J. Appl. Phycol.*, vol. 24, no. 5, pp. 1035–1043, Oct. 2012, doi: 10.1007/s10811-011-9730-z.

[9] N. Hassouna, B. Mithot, and J.-P. Bachellerie, "The complete nucleotide sequence of mouse 28S rRNA gene. Implications for the process of size increase of the large subunit rRNA In higher eukaryotes," *Nucleic Acids Res.*, vol. 12, no. 8, pp. 3563–3583, Apr. 1984, doi: 10.1093/nar/12.8.3563.

[10] M. Tragin, A. Zingone, and D. Vaulot, "Comparison of coastal phytoplankton composition estimated from the V4 and V9 regions of the 18S rRNA gene with a focus on photosynthetic groups and especially Chlorophyta," *Environ. Microbiol.*, vol. 20, no. 2, pp. 506–520, 2018, doi: 10.1111/1462-2920.13952.

[11] K. Hadziavdic, K. Lekang, A. Lanzen, I. Jonassen, E. M. Thompson, and C. Troedsson, "Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers," *PLOS ONE*, vol. 9, no. 2, p. e87624, Feb. 2014, doi: 10.1371/journal.pone.0087624.

[12] L. Guillou *et al.*, "The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy," *Nucleic Acids Res.*, vol. 41, no. D1, Jan. 2013, doi: 10.1093/nar/gks1160.

[13] C. Quast *et al.*, "The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools," *Nucleic Acids Res.*, vol. 41, Nov. 2012, doi: 10.1093/nar/gks1219.

[14] R. R. L. Guillard and P. E. Hargraves, "Stichochrysis immobilis is a diatom, not a chrysophyte," *Phycologia*, vol. 32, no. 3, pp. 234–236, May 1993, doi: 10.2216/i0031-8884-32-3-234.1.

[15] C. Lepere, M. Demura, M. Kawachi, S. Romac, I. Probert, and D. Vaulot, "Whole-genome amplification (WGA) of marine photosynthetic eukaryote populations," *FEMS Microbiol. Ecol.*, vol. 76, no. 3, pp. 513–523, Jun. 2011, doi: 10.1111/j.1574-6941.2011.01072.x.

[16] A. Schwelm, C. Berney, C. Dixelius, D. Bass, and S. Neuhauser, "The Large Subunit rDNA Sequence of Plasmodiophora brassicae Does not Contain Intra-species Polymorphism," *Protist*, vol. 167, no. 6, pp. 544–554, Dec. 2016, doi: 10.1016/j.protis.2016.08.008.

[17] Y. Takano and T. Horiguchi, "Acquiring Scanning Electron Microscopical, Light Microscopical and Multiple Gene Sequence Data from a Single Dinoflagellate Cell1," *J. Phycol.*, vol. 42, no. 1, pp. 251–256, 2006, doi: 10.1111/j.1529-8817.2006.00177.x.

[18] J.-S. Ki and M.-S. Han, "Molecular analysis of complete ssu to lsu rdna sequence in the harmful dinoflagellatealexandrium tamarense (korean isolate, HY970328M)," *Ocean Sci. J.*, vol. 40, no. 3, pp. 43–54, Sep. 2005, doi: 10.1007/BF03022609.

[19] L. A. Amaral-Zettler, E. A. McCliment, H. W. Ducklow, and S. M. Huse, "A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes," *PLoS ONE*, vol. 4, no. 7, p. e6372, Jul. 2009, doi: 10.1371/journal.pone.0006372.

[20] C. A. Scholin, M. Herzog, M. Sogin, and D. M. Anderson, "IDENTIFICATION OF GROUP- AND STRAIN-SPECIFIC GENETIC MARKERS FOR GLOBALLY DISTRIBUTED ALEXANDRIUM (DINOPHYCEAE). II. SEQUENCE ANALYSIS

OF A FRAGMENT OF THE LSU rRNA GENE1," *J. Phycol.*, vol. 30, no. 6, pp. 999–1011, 1994, doi: 10.1111/j.0022-3646.1994.00999.x.

[21] M. Kearse *et al.*, "Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data," *Bioinforma. Oxf. Engl.*, vol. 28, no. 12, pp. 1647–1649, Jun. 2012, doi: 10.1093/bioinformatics/bts199.

[22] C. A. Kreader, "Relief of amplification inhibition in PCR with bovine serum albumin or T4 gene 32 protein.," *Appl. Environ. Microbiol.*, vol. 62, no. 3, pp. 1102–1106, Mar. 1996.

[23] M. Nagai, A. Yoshida, and N. Sato, "Additive effects of bovine serum albumin, dithiothreitol and glycerolon PCR," *IUBMB Life*, vol. 44, no. 1, pp. 157–163, 1998, doi: 10.1080/15216549800201172.

[24] Q. Xie, J. Lin, Y. Qin, J. Zhou, and W. Bu, "Structural diversity of eukaryotic 18S rRNA and its impact on alignment and phylogenetic reconstruction," *Protein Cell*, vol. 2, no. 2, pp. 161–170, Feb. 2011, doi: 10.1007/s13238-011-1017-2.

[25] S. Yu, Y. Wang, D. Rédei, Q. Xie, and W. Bu, "Secondary structure models of 18S and 28S rRNAs of the true bugs based on complete rDNA sequences of Eurydema maracandica Oshanin, 1871 (Heteroptera, Pentatomidae)," *ZooKeys*, no. 319, pp. 363–377, Jul. 2013, doi: 10.3897/zookeys.319.4178.

[26] J. Iserte, B. Stephan, S. Goñi, C. Borio, P. Ghiringhelli, and M. Lozano, "Family-Specific Degenerate Primer Design: A Tool to Design Consensus Degenerated Oligonucleotides," *Biotechnol. Res. Int.*, vol. 2013, p. 383646, Feb. 2013, doi: 10.1155/2013/383646.

[27] J. Yuan, M. Li, and S. Lin, "An Improved DNA Extraction Method for Efficient and Quantitative Recovery of Phytoplankton Diversity in Natural Assemblages," *PLOS ONE*, vol. 10, no. 7, p. e0133060, Jul. 2015, doi: 10.1371/journal.pone.0133060.

[28] B. Becker, M. Melkonian, and J. P. Kamerling, "The Cell Wall (theca) of Tetraselmis Striata (chlorophyta): Macromolecular Composition and Structural Elements of the Complex Polysaccharides," *J. Phycol.*, vol. 34, no. 5, pp. 779–787, 1998, doi: 10.1046/j.1529-8817.1998.340779.x.

[29] L. Liming, O. Yexin, and H. Hongjun, "Minipreparation of Genomic DNA of a Marine Unicellular Green Algae Tetraselmis," *Wuhan Zhi Wu Xue Yan Jiu Wuhan Bot. Res.*, vol. 21, no. 4, pp. 295–300, Jan. 2003.

[30] J. Gong, J. Dong, X. Liu, and R. Massana, "Extremely High Copy Numbers and Polymorphisms of the rDNA Operon Estimated from Single Cell Analysis of Oligotrich and Peritrich Ciliates," *Protist*, vol. 164, no. 3, pp. 369–379, May 2013, doi: 10.1016/j.protis.2012.11.006.

[31] W. H. C. F. Kooistra, M. K. de Boer, E. G. Vrieling, L. B. Connell, and W. W. C. Gieskes, "Variation along ITS markers across strains of Fibrocapsa japonica (Raphidophyceae) suggests hybridisation events and recent range expansion," *J. Sea Res.*, vol. 46, no. 3, pp. 213–222, Dec. 2001, doi: 10.1016/S1385-1101(01)00086-7.

[32] A. J. Alverson, "Timing marine–freshwater transitions in the diatom order Thalassiosirales," *Paleobiology*, vol. 40, no. 1, pp. 91–101, ed 2014, doi: 10.1666/12055.

[33] W. J. Henley *et al.*, "Phylogenetic analysis of the 'Nannochloris-like' algae and diagnoses of Picochlorum oklahomensis gen. et sp. nov. (Trebouxiophyceae, Chlorophyta)," *Phycologia*, vol. 43, no. 6, pp. 641–652, Nov. 2004, doi: 10.2216/i0031-8884-43-6-641.1.

[34] S. M. Karst, M. S. Dueholm, S. J. McIlroy, R. H. Kirkegaard, P. H. Nielsen, and M. Albertsen, "Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias," *Nat. Biotechnol.*, vol. 36, no. 2, pp. 190–195, Feb. 2018, doi: 10.1038/nbt.4045.