Resolving the molecular ecology of marine microbial eukaryotes with metatranscriptomes

Ryan D. Groussman

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

E. Virginia Armbrust, Chair

David A.C. Beck

Gabrielle Rocap

Program Authorized to Offer Degree:

Oceanography

University of Washington

Abstract

Resolving the molecular ecology of marine microbial eukaryotes with metatranscriptomes

Ryan D. Groussman

Chair of the Supervisory Committee:

Professor E. Virginia Armbrust

School of Oceanography

The surface waters of the North Pacific ocean host diverse populations of microbial eukaryotes, who collectively mediate the flux of energy and matter in their environment with a significant impact on global biogeochemical cycles. The evolutionary history of these organisms spans billions of years and includes multiple endosymbiotic events that radiated into successful lineages with complex, chimeric genomes. The trophic mode of marine microbial eukaryotes ranges from purely photo-autotrophic to heterotrophic, including a number of flexible mixotrophic strategies. The metabolism, behavior, and physiology of a majority of microbial eukaryote species are not fully understood. The emergence of environmental genomics and transcriptomics has allowed researchers to elucidate the functional capacity and activity of microorganisms *in situ*, including perspectives on species and entire lineages that evade isolation and culture. The influx of massive amounts of high-throughput sequence data necessitates new

approaches to digesting and extracting information from these valuable data. In this dissertation I leverage poly-A+ selected, deeply sequenced metatranscriptomes to resolve the molecular ecology of *in situ* marine microbial eukaryote communities.

In Chapter 1, I studied a diel-resolved metatranscriptome time series in the North Pacific Subtropical Gyre. I developed a method for assembly, annotation, and quantification of these metatranscriptomes to reveal oscillating patterns of gene transcription over diel cycles. I identified differences in the magnitude of diel transcript regulation across different environmental genera and examined the enrichment of diel-regulated genes in key metabolic pathways to show orchestrated metabolic re-arrangement over day and night cycles.

In Chapter 2, I constructed an updated and reproducible protein reference library for marine microbial eukaryotes (MarFERReT), with the aim of improving the taxonomic annotations of environmental metatranscriptomes. I collected reference sequence material from a variety of sources, including recently available sequence data from novel and uncultured taxa, and ingested sequences through a standardized and open-source pipeline. I identify sets of core transcribed genes from these reference species that I use to estimate the coverage of environmental taxa bins and show how the incorporation of new sequence material improves the specificity of annotations.

I combine the methodological approaches of Chapter 1 with the enhanced annotation capacities generated from Chapter 2 to consolidate and standardize metatranscriptome data from four different cruises into a North Pacific Eukaryotic Gene Catalog (Appendix 1). This catalog contains clustered protein sequences from all metatranscriptome assemblies, together with taxonomic annotations from the MarFERReT reference library (Chapter 2), functional annotations, and transcript abundances.

Chapter 3 explores the picoeukaryotes of the dynamic North Pacific Transition Zone, using data from 3 cruises included in the North Pacific Eukaryotic Gene Catalog (Appendix 1). I focus on the small size fraction samples from latitudinal transections of the transition zone, where I observe shifts in the transcript inventory of species across biogeochemical and physical gradients. Perturbing the resource ratio of these communities in on-deck incubations revealed a subset of fast-responding species responsive to different nitrogen and iron ratios, and their differentially transcribed functions indicate a wide assortment of adaptive metabolic strategies.

This dissertation has elucidated numerous strategies that marine microbial eukaryotes use to sustain, survive and thrive in their environments, ranging from the fine-tuned diel transcription of metabolic machinery in the North Pacific Subtropical Gyre, to the rapid response of species with mixed trophic modes under changing nutrient conditions in the North Pacific Transition Zone. In the process, I have developed public resources for enhanced taxonomic annotation and improved reproducibility and accessibility of these valuable metatranscriptome data sets, so they can continue to provide insight and discovery of microbial eukaryotes in the oceans.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation would not have been possible without the wide community of mentors, peers, family and friends that supported my health, happiness and scientific curiosity through the last six years and beyond. Ginger Armbrust has been my greatest advocate and the best advisor I ever could have hoped for; she was always present to encourage my talents and curiosity, constructively challenge the areas where I could grow, and hold space and kindness for me when I suffered difficult periods in my personal life. I will forever be grateful for the mentorship Ginger has provided during my graduate studies, and her support and encouragement as an undergraduate before that.

I am grateful for the enthusiasm and insight of the Armbrust lab group and the collaborative and productive environment we share. Rita Peterson, Micaela Parker and Gwenn Hennon were generous supporters of my early academic ambitions leading up to and into graduate school. Sacha Coesel, Bryn Durham, Zinka Bartolek and Stephen Blaskowski were my brave companions in navigating the challenges that came with the massive influx of metatranscriptomic data. Chris Berthiaume helped guide me into the field of bioinformatics and access the computational power I needed and keep it working when I pushed our compute cluster to its limits. Francois Ribalet and Annette Hynes shared with me their endless curiosity for the world of picophytoplankton.

The UW Oceanography community and the numerous departments that collaborate with it have been a supportive and exciting environment to conduct research and grow as a scientist. My colleagues in the 2016 ocean grad cohort and the greater graduate community have been enormously kind, generous, and creative. I'm deeply grateful to Erik Fredrikson, Katy Christensen, Zho Ragen, and Elisabeth Brasseale for keeping together as a party in facing challenges and celebrating victory. Angie Boysen and Katherine Heal were welcoming and inspirational as a new student. I'm thankful for Gabrielle Rocap, Randie Bundy and David Beck for taking the time to serve on my committee and offer insight and guidance. Gabrielle Rocap, Alex Gagnon and Mikelle Nuwer helped me grow as an educator. It was a pleasure and adventure to work with collaborators on the 2016, 2017 and 2019 Gradients cruises and the SCOPE consortium of brilliant researchers doing exciting science.

The love and support from my family, friends and partners have been essential in keeping me healthy and whole as a graduate student. My parents encouraged my curiosity of the natural world from a young age, and my siblings share my appreciation for it. I am grateful for my dear friend Elling for her love, tenderness, wisdom and care; for my godchild Ursa and new life; for my long time friend Lisa and our shared bond; for Izzy and their un-ending warmth and good vibes; for Alex and the depth of feeling we share; for Mars and his friendship, energy, and joy; for Ev and her emotional support; for Dani in providing solid inspiration; and for all of the other friends and lovers in my life that have crossed through their own personal transition zones in search of the optimal conditions to thrive in life.

INTRODUCTION

The North Pacific Subtropical Gyre (NPSG) is one of the largest biomes on the planet, characterized by warm, stratified and nitrogen-deplete surface water (Karl 1999). Microbial life in the mixed layer of the NPSG is fine-tuned to the daily cycles of light and darkness. Populations of the picocyanobacteria *Prochlorococcus* predictably increase in mean cell diameter over the day as photosynthesis powers carbon fixation and growth, and mean diameter decreases during the night as cells undergo cell division (Vaulot and Marie, 1999; Ribalet et al., 2015), yet the total abundance of *Prochlorococcus* remains fairly level as division is countered by predation, viruses, and other forms of loss (Ribalet et al., 2015). Growth and division in small eukaryotic protists (<10 µm size fraction) is also coupled to diel cycles (Freitas et al., 2020), though distinguishing between different pico- and nano-eukaryote species in this small size class is a challenge with most optical instruments.

The relative stability of the North Pacific Subtropical Gyre is contrasted across its northern border by a region known as the North Pacific Transition Zone (NPTZ); a latitudinal band of strong physical, chemical, and biological gradients and high community productivity where warm, nitrogen-deplete, iron-replete water from the subtropical gyre mixes with cold, nutrient-rich and iron poor water from the northern subpolar gyre. (Roden 1991, Polovina et al., 2001, Juranek et al., 2012, Juranek et al., 2020). Along these gradients of salinity, temperature, nitrogen and iron, the particle size distribution increases with latitude as picocyanobacteria populations fade and eukaryotic picoplankton, nanoplankton and microplankton rise in successive waves (Juranek et al., 2020).

In Chapter 1, I studied a 4-day diel time series of transcriptional abundance profiles for the protist community (0.2–100 µm cell size) in the NPSG near Station ALOHA. *De novo* assembly, annotation and enumeration of poly-A+ selected metatranscriptomes revealed oscillating patterns of gene transcription over diel cycles, with a notably high level of diel regulation in haptophytes and ochrophytes. Metabolic pathways enriched in diel-regulated transcripts showed the tuning of eukaryotic transcription to day and night cycles, as daytime peaks in photosynthesis, carbon fixation, and fatty acid biosynthesis processes gave way at dusk to oxidative phosphorylation and fatty acid degradation.

Underlying our ability to understand environmental metatranscriptome samples is the coverage, quality, and diversity of the reference material that they are annotated against. A significant contribution to marine eukaryote reference sequences came from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP, Keeling et al., 2014), which coordinated a community-wide effort to sequence hundreds of new reference sequences from previously unsequenced cultures. The impact of the MMETSP is still being made, yet the public sphere of known marine eukaryote genetic material continues to increase as new transcriptome and genomes are made public from other sequencing efforts. In Chapter 2, I constructed an updated and reproducible protein reference library for marine microbial eukaryotes (MarFERReT), and incorporated sequence data from 890 genomes and transcriptomes, with many of these not contained in previous marine eukaryote references. I ingested sequences through a standardized pipeline, identifying sets of core transcribed genes that I use to estimate the coverage of environmental taxa bins. Addition of novel open-ocean taxa demonstrates how new sequence material improves the specificity of annotations from similar biomes.

The North Pacific Eukaryotic Gene Catalog (Appendix 1) consolidates eukaryotic metatranscriptome data from three latitudinal transects of the transition zone and one cruise in the subtropical gyre. This catalog contains 182 million clustered protein sequences from 175 separate *de novo* metatranscriptome assemblies, together with taxonomic annotations from the MarFERReT reference library (Chapter 2), functional Pfam annotations, and transcript abundances.

Chapter 3 explores the molecular ecology of picoeukaryotes in the NPTZ using metatranscriptome data from three latitudinal cruise transects in the Spring of 2016, 2017 and 2019, where water column parameters were surveyed with latitudinal resolution (Juranek et al., 2020). Metatranscriptome samples were assembled and annotated as described in Appendix 1. I focus on the small size fraction samples (0.2-3 μm), observing shifts in the transcript inventory of species across biogeochemical and physical gradients. On-deck incubation experiments were conducted on the 2017 cruise to test the community response to altered N:Fe ratios at three sites in the transition zone, and metatranscriptomes from these experiments revealed shifts in the community transcript profile 4 days after nutrient amendment, allowing us to identify a core subset of significant responder species, in particular osmo- and phago-mixotrophic ochrophyte and haptophytes. Differential gene expression in responder species suggests metabolic flexibility in response to changing nutrient conditions; including a pattern of enhanced photosynthesis and energy storage functions under higher N:Fe ratios. The findings from this study add to our understanding of the importance of the N:Fe ratio in structuring picoeukaryote communities, and the dynamic responses of picoeukaryote species adapted to changing nutrient conditions in the North Pacific Transition Zone.

INTRODUCTION REFERENCES

Freitas, F. H., Dugenne, M., Ribalet, F., Hynes, A., Barone, B., Karl, D. M., & White, A. E. (2020). Diel variability of bulk optical properties associated with the growth and division of small phytoplankton in the North Pacific Subtropical Gyre. *Applied Optics*, *59*(22), 6702-6716.

Juranek, L. W., Quay, P. D., Feely, R. A., Lockwood, D., Karl, D. M., & Church, M. J. (2012). Biological production in the NE Pacific and its influence on air-sea CO2 flux: Evidence from dissolved oxygen isotopes and O2/Ar. *Journal of Geophysical Research: Oceans*, *117*(C5).

Juranek, L. W., White, A. E., Dugenne, M., Henderikx Freitas, F., Dutkiewicz, S., Ribalet, F., Ferrón, S. E., Armbrust, E. V., & Karl, D. M. (2020). The importance of the phytoplankton "middle class" to ocean net community production. *Global Biogeochemical Cycles*, *34*(12), e2020GB006702.

Karl, D. M. (1999). A sea of change: biogeochemical variability in the North Pacific Subtropical Gyre. *Ecosystems*, *2*(3), 181-214.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A.., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., Deschamps, P., Dyhrman, S. T., Edvardsen, B., Gates, R. D., Gobler, C. J., Greenwood, S. J., Guida, S. M., Jacobi, J. L., Jakobsen, K. S., James, E. R., Jenkins, B., John, U., Johnson, M. D., Juhl, A. R., Kamp, A., Katz, L. A., Kiene, R., Kudryavtsev, A., Leander, B. S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A. M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M. A., Murray, S., Nadathur, G., Nagai, S., Ngam, P. B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M. C., Rengefors, K., Romano, G., Rumpho, M. E., Rynearson, T., Schilling, K. N., Schroeder, D. C., Simpson, A. G. B., Slamovits, C. H., Smith, D. R., Smith, G. J., Smith, S. R., Sosik, H. M., Stief, P., Theriot, E., Twary, S. N., Umale, P. E., Vaulot, D., Wawrik, B., Wheeler, G. L., Wilson, W. H., Xu, Y., Zingone, A., & Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS biology*, *12*(6), e1001889.

Polovina, J. J., Howell, E., Kobayashi, D. R., & Seki, M. P. (2001). The transition zone chlorophyll front, a dynamic global feature defining migration and forage habitat for marine resources. *Progress in oceanography*, *49*(1-4), 469-483.

Ribalet, F., Swalwell, J., Clayton, S., Jiménez, V., Sudek, S., Lin, Y., Johnson, C. I., Worden, A. Z., & Armbrust, E. V. (2015). Light-driven synchrony of Prochlorococcus growth and mortality in the subtropical Pacific gyre. *Proceedings of the National Academy of Sciences*, *112*(26), 8008-8012.

Roden, G. I. (1991). Subarctic-subtropical transition zone of the North Pacific: large-scale aspects and mesoscale structure. *NOAA Technical Report NMFS*, *105*, 1-38.

Vaulot, D., & Marie, D. (1999). Diel variability of photosynthetic picoplankton in the equatorial Pacific. *Journal of Geophysical Research: Oceans*, *104*(C2), 3297-3310.

CHAPTER 1

Diel-regulated transcriptional cascades of microbial eukaryotes in the North Pacific Subtropical Gyre

Groussman, R. D., Coesel, S. N., Durham, B. P., & Armbrust, E. V. (2021). Diel-Regulated Transcriptional Cascades of Microbial Eukaryotes in the North Pacific Subtropical Gyre. *Frontiers in microbiology*, *12*. doi: 10.3389/fmicb.2021.682651

1.1 Abstract

Open-ocean surface waters host a diverse community of single-celled eukaryotic plankton (protists) consisting of phototrophs, heterotrophs, and mixotrophs. The productivity and biomass of these organisms oscillate over diel cycles, and yet the underlying transcriptional processes are known for few members of the community. Here, we examined a four-day diel time series of transcriptional abundance profiles for the protist community (0.2-100 µm in cell size) in the North Pacific Subtropical Gyre near Station ALOHA. *De novo* assembly of poly-A+ selected metatranscriptomes yielded over 30 million contigs with taxonomic and functional annotations assigned to 54% and 25% of translated contigs, respectively. The completeness of the resulting environmental eukaryotic taxonomic bins was assessed, and 48 genera were further evaluated for diel patterns in transcript abundances. These environmental transcriptome bins maintained reproducible temporal partitioning of total gene family abundances, with haptophyte and ochrophyte genera generally showing the greatest diel partitioning of their transcriptomes. The haptophyte *Phaeocystis* demonstrated the highest proportion of transcript diel periodicity, while most other protists had intermediate levels of periodicity regardless of their trophic status. Dinoflagellates, except for the parasitoid genus *Amoebophrya*, exhibit the fewest diel oscillations of transcript abundances. Diel-regulated gene families were enriched in key metabolic pathways; photosynthesis, carbon fixation, and fatty acid biosynthesis gene families had peak times concentrated around dawn, while gene families involved in protein turnover (proteasome and protein processing) are most active during the high intensity daylight hours. TCA cycle, oxidative phosphorylation and fatty acid degradation predominantly peaked near dusk. We identified temporal pathway enrichments unique to certain taxa, including assimilatory sulfate reduction at dawn in dictyophytes and signaling pathways at early evening in haptophytes,

pointing to possible taxon-specific channels of carbon and nutrients through the microbial community. These results illustrate the synchrony of transcriptional regulation to the diel cycle, and how the protist community of the North Pacific Subtropical Gyre structure their transcriptomes to guide the daily flux of matter and energy through the gyre ecosystem.

1.2 Introduction

The North Pacific Subtropical Gyre (NPSG) is a warm and oligotrophic (nitrogen-limited) ecosystem that hosts a diverse community of phototrophic, heterotrophic, and mixotrophic microbial eukaryotes (protists) spanning over three orders of magnitude in cell sizes. The phylogenetically diverse eukaryotic phytoplankton (phototrophs) represent nearly half of the phytoplankton biomass of the NPSG, and are composed primarily of organisms derived through secondary or tertiary endosymbiosis such as dinoflagellates, haptophytes, and ochrophytes (photosynthetic stramenopiles) (Alexander et al., 2015, Hu et al., 2018). Strictly phototrophic members of the NPSG eukaryotic phytoplankton consist of haptophytes including *Emiliania huxleyi* (Hernández et al., 2020) and diatoms such as *Rhizosolenia* and *Hemiaulus* (Villareal et al., 1993), two genera found in symbioses with nitrogen fixing cyanobacteria that bloom during sporadic injections of nutrients into the surface waters. The strictly heterotrophic (zooplankton) component of the protist community is dominated by the Alveolata supergroup (including dinoflagellates), as well as stramenopiles and Rhizaria (Rii 2016, Hu et al., 2018). Many eukaryotic lineages within the NPSG have mixotrophic life strategies, adjusting their relative balance of photosynthesis and phagotrophy to changing light and nutrient conditions (Mitra et al., 2016, Lambert 2021). In addition, mixotrophs can be distinguished between constitutive (vertical inheritance of plastids) and non-constitutive (kleptoplastic, or acquisition of plastids from prey) (Mitra et al., 2016). Members of haptophyte, ochrophyte and dinoflagellate lineages are constitutive mixotrophs (Faure et al., 2019), with evidence that they can graze on picocyanobacteria in the NPSG (Frias-Lopez et al., 2009). In the well-lit and low nutrient conditions of the NPSG, mixotrophy may be advantageous (Rothhaupt 1996) and the gene family abundance profiles of many environmental protist species suggests widespread mixotrophy (Lambert et al., 2021). Some ciliates, such as *Strombidium*, are non-constitutive mixotrophs that retain the plastid of their consumed prey (Faure et al., 2019, Stoecker et al., 2009).

The daily cycles of light and darkness synchronize cell growth and division across the diverse members of the microbial communities within the sunlit waters of the NPSG. For example, the phototrophic cyanobacteria *Prochlorococcus* displays reproducible increases in mean cell diameter over the day as cells fix carbon into biomass, followed by decreases during the night as cells undergo cell division (Vaulot & Marie 1999, Ribalet et al., 2015). From day to day there is little variation in the total cellular abundance of *Prochlorococcus*, indicating that growth rates roughly equal loss rates in an ecological system finely tuned to the daily solar cycle (Ribalet et al., 2015). The tight coupling of growth and division to the light/dark cycle is also clear within small eukaryotic protists (<10 μm size fraction), although the different species underlying the observed changes in biomass cannot be distinguished based on the optical measures (Freitas et al. 2020). Heterotrophic bacteria also display oscillating waves of species-specific transcriptional patterns over the diel cycle (Ottesen et al., 2014), potentially due to the daily cycle of phytoplankton production of organic matter.

Recent 'omics analyses have enhanced our understanding of the metabolic cascades resulting from synchrony to the daily light cycle in the NPSG. For example, transcript abundances in the haptophyte phototroph *E. huxleyi* underwent daily oscillations, with transcripts associated with carbon fixation and lipid synthesis proteins reaching a peak in abundance near dawn and at mid-day or dusk for those genes encoding respiration and lipid degradation, suggesting a cycle of energy-store biosynthesis and consumption (Hernández et al., 2020). A similar pattern was observed in the more nutrient-rich waters of the California Current upwelling, with oscillations in the abundance of diatom and green algae photosynthesis-related transcripts during the morning and cell division-related transcripts during the night (Kolody et al., 2019). Organisms in the NPSG are equipped with a diversified repertoire of photoreceptors (Coesel et al., 2021), which may allow them to sense and regulate their daily responses to changes in light conditions. Responses to the light cycle include the biosynthesis of energy-rich triacylglycerols during the day and their consumption during the night (Becker et al., 2018) and the replenishment of pigments during the night to compensate for photodegradation during the day (Becker et al, 2020). Moreover, a diel structuring of cross-kingdom interactions was demonstrated by the species-specific exchange of phytoplankton-produced organic sulfonated compounds, produced primarily by haptophytes and diatoms and consumed by heterotrophic bacteria (Durham et al., 2019).

Here, we examined diel transcriptional patterns across the microbial eukaryotes (protists) (0.2-100 μm in size) of the NPSG to evaluate how metabolic pathways may be synchronized across the microbial eukaryote community and the varied trophic states that comprise it. We hypothesized that eukaryotic gene expression would be strongly patterned by evolutionary lineage and trophic level, with phototrophs and mixotrophs expected to demonstrate the highest degree of diel periodicity in their transcriptome and by inference, their metabolism. Analyses centered on the direct annotation of metatranscriptome-assembled contigs allowed us to investigate large-scale transcriptional patterns in abundant eukaryotic taxa and to identify functional metabolic processes operating on diel cycles.

1.3 Materials and Methods

1.3.1 Cruise and sample collection

Duplicate samples for eukaryotic metatranscriptomes were collected from 15 m depth every four hours (06:00, 10:00, 14:00, 18:00, 22:00, and 02:00 HST) over the course of four days on the R/V *Kilo Moana* cruise KM1513 (July and August 2015) approximately 100km NE of Station ALOHA in the North Pacific Subtropical Gyre (see Wilson et al. 2017 for additional cruise details). For each of the 48 samples, seven liters of seawater were pre-filtered through a 100 μm nylon mesh and collected onto a 142 mm 0.2 μm polycarbonate filter using a peristaltic pump. Filters were immediately flash frozen in liquid nitrogen and subsequently stored at -80°C until further processing. Filters were extracted using the ToTALLY RNA Kit (Invitrogen) with some modifications. Briefly, frozen filters were added to 50 mL falcon tubes containing 5 mL of denaturation solution and extraction beads (125 μL 100 μm zirconia beads, 125 μL 500 μm zirconia beads, and 250 μL 425-600 μm silica glass beads). A set of 14 internal mRNA standards were added to the extraction buffer for each sample to generate quantitative transcript inventories; these standards were synthesized as previously described (Satinsky *et al.*, 2013), with the exception that eight standards were synthesized with poly(A) tails to mimic eukaryotic mRNAs. Total, extracted RNA was treated with DNase I (Ambion, New York, USA) and purified with DNase inactivation reagent (Ambion). Eukaryotic mRNAs were poly(A)-selected, sheared to ~225 bp fragments, and used to construct TruSeq cDNA libraries according to the

Illumina TruSeq® RNA Sample Preparation v2 Guide for paired-end (2 X 150 bp) sequencing using the Illumina NextSeq 500 sequencing platform.

1.3.2 DNA sequence quality control and de novo assembly

Raw Illumina sequence reads were quality controlled with trimmomatic v0.36 (Bolger *et al*., 2014, parameters: *MAXINFO:135:0.5, LEADING:3, TRAILING:3, MINLEN:60,* and *AVGQUAL:20)*. A total of 2,426,923,906 merged paired-end sequences were generated with a median length of ~240 bp. Sequences were pooled for each of the 24 sampling times and the 24 pooled samples were assembled using the Trinity *de novo* transcriptome assembler v2.3.2 (Grabherr *et al*., 2011, parameters: --normalize_reads --min_kmer_cov 2 --min_contig_length 300) on the Pittsburgh Supercomputing Center's Bridges Large Memory system. Trinity assembly yielded 52,489,585 total contigs from all 24 assemblies.

1.3.3 Quality control of assemblies, translation and longest frame selection

The raw assemblies were quality controlled with Transrate v1.0.3 (Smith-Unna *et al*., 2016) to check contigs for chimeras, structural errors, and base errors, using their paired-end assembly method (parameters: --assembly $sample.fasta --left $left_reads --right $right_reads). A total of 31,284,431 contigs (59.6% of the raw pool) passed the optimized assembly score threshold and were retained for further analysis. The quality-controlled contigs were translated in six frames with transeq vEMBOSS:6.6.0.059 (Rice *et al*., 2000) using Standard Genetic Code. The longest (or multiple frames if of equal lengths) open reading frame from each contig was retained for downstream analyses. A total of 32,536,410 translated frames that passed these criteria were retained.

1.3.4 Clustering

The 24 peptide assemblies were merged and clustered at the 99% identity threshold level with linclust within the MMseqs2 package (version 31e25cb081a874f225d443eec307a6254f06a291, Steinegger and Söding 2018, --min-seq-id 0.99). A total of 30,015,008 peptide sequences (92% of input sequences) were retained as cluster representatives and used for further analysis.

1.3.5 Annotation

Following clustering at 99% identity, representative contigs were annotated for taxonomy and function. Contigs were annotated against the curated MarineRefII reference database (http://roseobase.org/data/) of 641 marine eukaryotes and prokaryotes, including the MMETSP transcriptomes (Keeling *et al.*, 2014), and supplemented with available marine animal, fungal, choanoflagellate and viral sequences (Coesel et al., 2021). Assembled contigs were aligned to the reference database using DIAMOND v 0.9.18 (Buchfink *et al.*, 2014, parameters: --sensitive -b 65 -c 1 -e 1e-5 --top 10 -f 100). The Lowest Common Ancestor (LCA) was estimated using the LCA algorithm in DIAMOND in conjunction with NCBI taxonomy. The frame with the lowest (best) e-value was retained if multiple frames of a contig received annotations. A total of 15,302,768 contigs were assigned an NCBI tax_id (51.0%). Contigs assigned to the same NCBI taxon or daughter nodes were defined as members of the same environmental taxon "bin." To determine the putative function of each contig, we used hmmsearch, from HMMER 3.1b2 (Eddy 2011, parameters: -T 30 --incT 30), to assign KEGG Orthology IDs (KOs) to contigs using 22,247 hmm profiles from KOfam ver. 2019-03-20 (Aramaki 2020). The profile with the highest bitscore was retained for those contigs that mapped to more than one KOfam profile. If multiple frames of a contig received annotations, the frame with the highest annotation bitscore was retained. A total of 7,707,191 contigs were assigned an KEGG KO (25.7%).

1.3.6 Quantification and normalization of abundance

The clustered contig representatives were quantified by alignment of their nucleotide sequences against the paired-end reads using kallisto v0.43.1 (Bray *et al.*, 2016, parameters: quant --rf-stranded -b 40). We normalized contigs abundance to fragments per kilobase of transcript per million total reads (FPKM), using total read values mapped to each taxonomic bin, rather than the total library size, as the denominator, 'M'. FPKM values for contigs with the same taxonomy and functional annotation combinations (NCBI tax_id and KEGG KO) were summed. Environmental taxa bins were integrated down taxonomic levels, summing abundance values from lower-ranking nodes in the NCBI taxonomy.

1.3.7 NMDS Ordination

Non-metric multidimensional scaling (NMDS) ordination was used to reduce taxonomic, temporal, functional, and abundance information into three-dimensional space. The input was a matrix of observations consisting of each of the 48 genus-level taxa for each of the 24 time points (1,152 total observations). The features for each observation were the *in silico* normalized counts for the 6,925 KOfams present in >5% of the observations. The features within observations were normalized such that the row sums equal 1. The metaMDS function in the R package 'vegan' version 2.5-5 (Oksanen *et al*., 2019, parameters: k=3, trymax=100) was used to compute the Bray-Curtis distance matrix and find a solution between runs. A solution for this ordination with two dimensions was not achievable. The three-dimensional NMDS ordination results were visualized with the R package 'plotly' (Sievert 2018). NMDS ordinations were also individually generated for the 48 genera. The individual ordinations were resolved in two dimensions (parameters: k=2). Mean stress of 48 NMDS = 0.127 ± 0.028 stdev.

1.3.8 Determining significant diel periodicity

Significant periodicity of gene family abundances for each genus were determined with the Rhythmicity Analysis Incorporating Non-parametric Methods (RAIN) package implemented in R (Thaben and Westermark 2014). The p-values from RAIN analyses were ranked and corrected at an FDR < 0.05 using the Benjamini-Hochsberg false-discovery rate method (Benjamini and Hochberg, 1995), as implemented in Ottesen *et al.,* 2014.

1.3.9 Enrichment of significantly diel gene families in KEGG pathways

KEGG Pathways and their associated knums (gene family identifiers representing a KOfam) were parsed using the R package KEGGREST (Tenenbaum 2016) to access the KEGG database. Only pathways with associated knums were used. A contingency matrix for each genus-pathway-time combination (a total of 124,704 combinations of contingency matrices for genus-level analysis, and 106,518 for order-level analysis) was constructed for the test and populated with the appropriate counts: 'knum in pathway and is diel', 'knum is in pathway and is not diel', 'knum is not in pathway and is diel', and 'knum is not in pathway and is not diel'. The 'diel' status of each knum was determined from the RAIN results, above. We removed contingency matrices with no identified knums in the pathway, reducing our number of tests to 108,288 for the genus-level analysis and 93,114 for the order-level analysis. We used Fisher's Exact Test on these matrices to determine enrichment, combined with a Benjamini-Hochberg multiple comparison correction and False Discovery Rate of less than 5% (genera-level adjusted maximum p-value < 3.412673e-05, order-level adjusted maximum p-value 5.792821e-05) Benjamini and Hochberg, 1995), both executed within R.

1.4 Results

1.4.1 Taxonomic and functional composition of environmental transcriptome bins

We examined the transcriptional profiles of eukaryotic microbes (protists) over the diel cycle by collecting size-fractionated (0.2 – 100 µm) RNA samples every four hours (at 06:00, 10:00, 14:00, 18:00, 22:00 and 02:00 HST) over four consecutive days in the oligotrophic North Pacific Subtropical Gyre (NPSG), ~100km NE of Station ALOHA. Local sunrise was at ~06:00 and sunset was at ~18:00 HST. Surface illumination intensities reached over 2000 □mol m^{-2} s^{-1} between 10:00 and 14:00 HST over the four-day sampling period, during which the picoeukaryotic phytoplankton grew and divided on an oscillating daily basis, as estimated from continuous flow cytometry measurements (Coesel et al., 2021). Water column properties during the cruise confirm a warm (26.6° to 27.04°C with some diel variability) and nitrogen-deplete environment (nitrate + nitrite 8±4 nmol L^{-1}) (Wilson et al., 2017). Casts were taken at 15 meters depth, with the research vessel (KM1513) tracking a Lagrangian drogue placed at 15 m to allow repeat sampling of the planktonic community from the same water mass. Illumina deep sequencing of libraries created from poly-A+ selected RNA yielded a total of ~2.4 billion transcript fragments (merged paired-end reads) with an average merged length of ~240 bp (Table 1.1). *De novo* assembly of the metatranscriptome sequences generated about 52 million nucleotide contigs. Subsequent quality control, translation, frame selection, and clustering at 99% amino acid identity (Table 1.1) yielded 30 million amino acid sequences (hereafter referred to as 'contigs'), with an N50 of 423 base pairs (141 amino acid residues, Supplementary Figure S1.1, Table 1.1).

Table 1.1 Sequence data metrics. Volume of sequence data by short reads, assembled contigs, and annotated contigs. Type refers to nucleotide ('nt') or translated amino acid ('aa') sequence. Genus-level taxonomy and functional counts include contigs assigned to nested daughter taxonomies.

Sequence data	**Count**	**Type**
Merged read pairs	2,426,923,906	nt
Raw contigs	52,489,585	nt
QCed contigs (transrate)	31,284,431	nt
Translated, clustered contigs	30,015,008	aa
Contigs w/ any taxonomy	16,061,543	aa
Contigs w/ any KEGG function	7,707,191	aa
Genus-level taxonomy	5,181,384	aa
Genus-level w/KEGG function	1,390,232	aa

The translated contigs were annotated in two ways. First, taxonomic identity and rank within the NCBI taxonomic framework were determined by mapping contigs against a reference database of 18.5 million predicted protein sequences from 554 marine eukaryotes, bacteria, archaea, and viruses (Coesel et al., 2021) and estimating the Lowest Common Ancestor (LCA) of matches. Placement to any taxonomic level was possible for 16 million contigs (51% of total; Figure 1.1A, Supplementary Figure S1.2, Table 1.1); the remaining 49% received no taxonomic annotation. Bacteria or Archaea were assigned to 55,099 and 740 contigs, respectively; these contigs were not considered further in this study. Second, potential function was assigned by mapping the contigs against the Kyoto Encyclopedia of Gene and Genomes (KEGG) database of orthogenes (KOfams) using HMM profiles (Aramaki 2020). A total of 7.7 million contigs (25.7%) were assigned to a putative KOfam and the associated KO term, with 13,765 total KOfams identified among all contigs (Figure 1.1A, Table 1.1). Approximately 5.2 million contigs were annotated to a genus rank or lower (Figure 1.1A, Table 1.1), with a subset of 1.4 million contigs that also received a putative KEGG function. This set of 1.4 million contigs with putative genus-level taxonomy and function were grouped based on their assigned genus-level taxonomy to create environmental taxonomic bins and used for subsequent analyses. Two hundred thirty-one environmental eukaryotic genera bins were represented by at least one functional KOfam assignment (Figure 1.1B). A low proportion of reads aligned to reference sequences of metazoans (opisthokonts) larger in size than 100 μm, potentially reflecting varied biological sources including sloughed cells, gametes, and other life cycle stages.

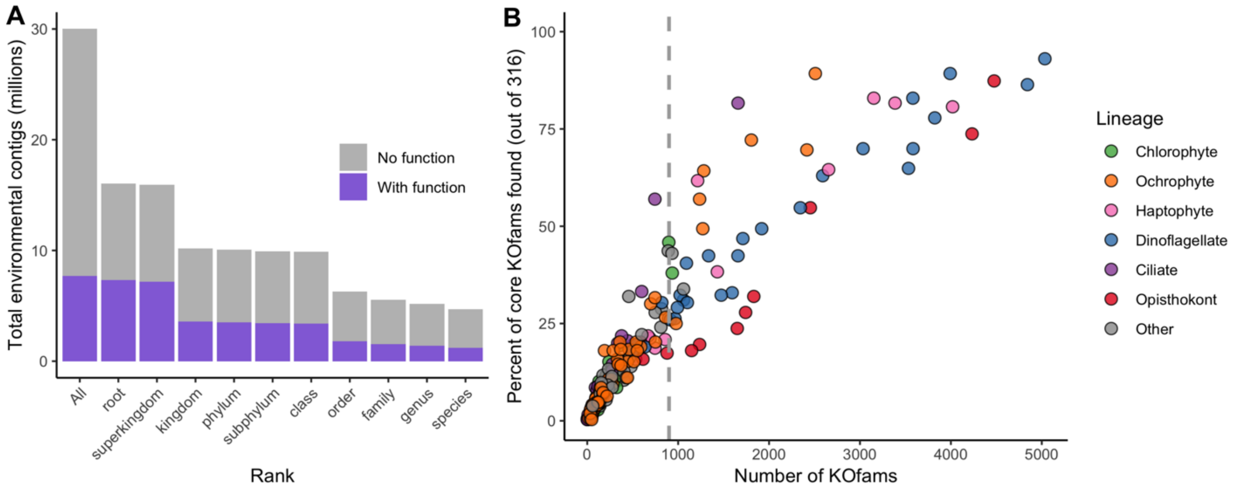


Figure 1.1. Annotation of metatranscriptome assemblies by taxonomy, function, and abundance. A) Cumulative number of contigs placed at each rank or lower. 'All' includes taxonomically-unassigned contigs. **B)** Completeness of genus-level taxonomic bins based on total number of KOfams and percent of the 316 'core KOfams'. Circles represent genus-level taxonomic bins, colored by lineage as in legend. Dashed line represents cutoff criteria of 900 detected KOs for subsequent analyses.

We developed a metric of transcriptome completeness to identify a subset of well-represented environmental genera from this initial set of taxonomic bins for further analyses. We first estimated a minimum number of KOfams expected within marine eukaryotes by *de novo* mapping the proteomes of our 366 reference marine eukaryotes against 22,247 KOfam hmm profiles (ver. 2019-03-20; Aramaki 2020). Three hundred fifty-five of the reference protists (over 96% of eukaryotic reference taxa) each contained 900 or more KOfams; only parasitic protists with a reduced gene content possessed significantly fewer KOfams (Supplementary Figure S1.3, Supplementary Data Sheet S1.1). A set of 316 KOfams were defined as "core KOfams" due to their presence in ~95% of reference marine eukaryotes (Supplementary Data Sheet S1.2). The completeness of a given genus-level environmental taxonomic bin was estimated based on the percentage of core KOfams identified in the environmental bins (Figure 1.1B, Supplementary Figure S1.3). We constrained further analysis to 48 environmental genera bins that each had >900 detected KOfams. These 48 genera were grouped by 8 higher-level lineages (Supplementary Data Sheet S1.3): dinoflagellates (n = 23), opisthokonts (n = 8), ochrophytes (photosynthetic stramenopiles, n = 7), haptophytes (n = 6), ciliates (n = 1), chlorophytes (n = 1), and two "other" lineages of kinetoplastids and oomycetes. These 48 genera corresponded to genera previously detected in the NPSG through metabarcoding (Hu et al., 2018) and collectively had an average of 54% (170 of 316) core KOfams positively identified (Supplementary Data Sheet S1.3). The highest proportion of core KOfams were detected in the taxonomic bin most closely related to the dinoflagellate *Karlodinium* (93%). Confidence values (e-value) for taxonomic assignments were assessed for all 48 genera, with the majority of contigs assigned to each bin receiving the highest-possible confidence value of 0 (Supplementary Figure S1.4). Some genera displayed a shallower distribution of e-value placements, reflecting a more likely entrainment of distantly-related taxa to genus representatives. This is most noticeable in multicellular metazoans with single genus reference representatives (e.g., *Octopus*, *Salmo*, *Orcinus*) and some dinoflagellates (*Amphidinium*, *Kryptoperidinium*, *Lingulodinium*).

1.4.2 Dimensionality reduction of environmental transcriptome bins.

Non-metric multidimensional scaling (NMDS) of Bray-Curtis dissimilarity was used to compare the similarity of transcript abundances for thousands of gene families across the 48 genus-level environmental bins over the 24 time points (Figure 1.2A). The input matrix consisted of each of the 48 genus-level environmental bins for each of the 24 time points (1,152 total observations) and the row-normalized number of transcripts associated with KOfams present in > 5% of observations (6,925 total features). Several patterns emerged from the resulting 3D NMDS ordination. First, the observations clustered together according to genus designation rather than time point, indicating that each taxonomic bin displayed a relatively distinct transcriptional fingerprint irrespective of the time of sampling. Second, phylogenetically-related genera clustered near one another. One notable exception to this pattern was the environmental bin corresponding to the dinoflagellate genus *Amoebophrya*, a highly-derived genus of dinoflagellates with a parasitoid life cycle (Guillou et al., 2008). Third, both the dinoflagellate genera and opisthokont genera (7 metazoa and 1 choanoflagellate) clustered apart from other protists. Among the remaining protists, known phototrophic eukaryotes (diatoms, chlorophytes, and some haptophytes), heterotrophs, and potential mixotrophs formed distinct clusters, with the positioning of putative mixotrophs between the heterotroph and autotroph clusters reflecting their mixed metabolism.

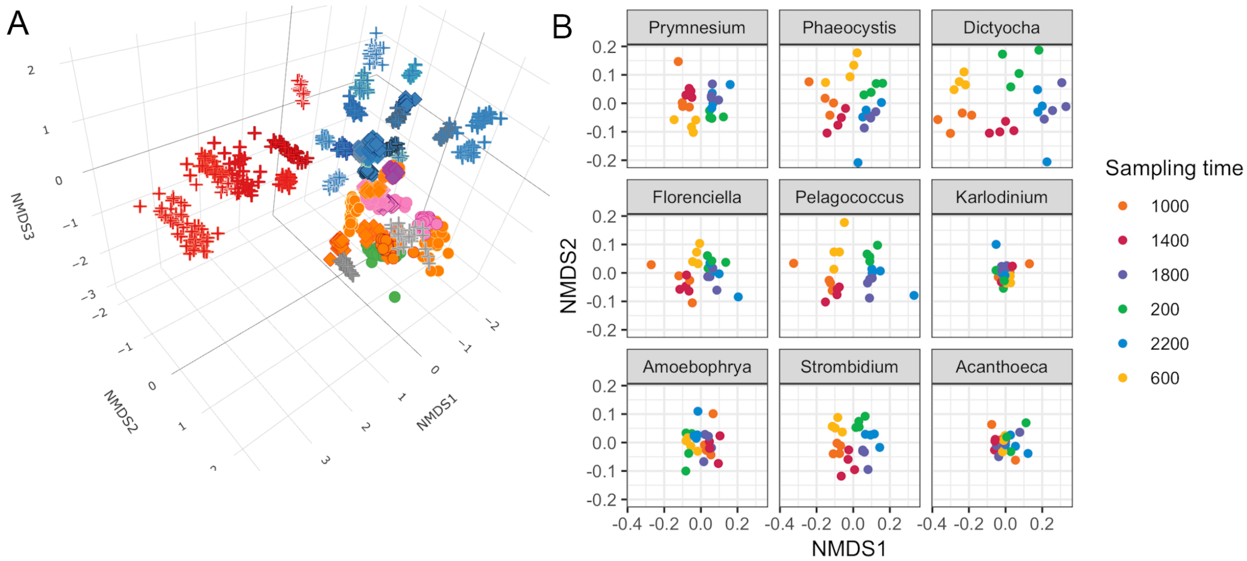


Figure 1.2. NMDS ordinations of transcript abundances from 48 genus-level environmental bins. A) Three-dimensional ordination for 48 genus-level taxonomic bins meeting completeness criteria. Stress = 0.160. Each point is a discrete time point for each genus based on the mean transcript abundances for each KOfam from duplicate samples. Shapes indicate assumed trophic level: circle, phototroph; diamond, mixotroph; cross, heterotroph. Inner color is specific to lineage in legend; outer color is unique to genera within a lineage. **B)** NMDS ordination performed independently on gene families belonging to 9 representative protist genera. Colored dot by genus name corresponds to Lineage from A. Mean stress of 9 independent NMDS = 0.114 ± 0.025 stdev.

We conducted independent NMDS ordinations on each of the 48 environmental genera bins to evaluate whether temporal partitioning of transcript abundances resolved within each environmental bin. Individual NMDS ordinations for the 48 genera were generated (mean stress = 0.114 ± 0.025 stdev) (Supplementary Figure S5). We highlighted a select subset of representative genera as representatives of the NPSG protist community in Figure 1.2B. We focused on nine representative environmental protist genera, based on their high sequence coverage, a high proportion of total and core KOfams, and representation of different trophic states and evolutionary lineages (Table 1.2, Supplementary Data Sheet S1.3). Haptophytes with contrasting trophic modes were represented by the genus *Prymnesium* (order Prymnesiales), which contains known mixotrophic species (Faure et al., 2018) and the genus *Phaeocystis* (order Phaeocystales), which is considered strictly photosynthetic and can exist as a free-living flagellate or in a colonial form (Rousseau et al., 2007). Ochrophytes were represented by the silicoflagellate genus *Dictyocha* (order Dictyochales), the genus *Florenciella* (order Florenciellales), and the pico-eukaryotic (< 2 µm cell size diameter) genus *Pelagococcus* (order Pelagomonadales). Both *Dictyocha* and *Florenciella* include known mixotrophic species (Quéguiner 2016, Li et al., 2021), whereas *Pelagococcus* is thought to be strictly photosynthetic

(Lewin et al., 1977). Dinoflagellates were represented by the genus *Karlodinium* (order Gymnodiniales), which contains mixotrophic species, similar to many of the dinoflagellate genera observed in this study (Faure et al., 2018). Isotopically-labeled grazing experiments on *Prochlorococcus* and *Synechococcus* in the NPSG identified prymnesiophytes (Prymnesiophyceae), dictyophytes (Dictyochophyceae), and dinoflagellates (Dinoflagellata) as grazers of picocyanobacteria (Frias-Lopez et al., 2009). We also highlighted *Amoebophrya* (order Syndiniales), a parasitic dinoflagellate that infects eukaryotic host cells, because of its distinct lifestyle and because its notable abundance in this environment as evidenced by the rRNA/rDNA libraries from the NPSG (Hu et al., 2018). The ciliate genus *Strombidium* (order Oligotrichida) can live heterotrophically as well as by non-constitutive mixotrophy through retention of plastids from engulfed prey (Stoecker et al., 2009). The choanoflagellate genus *Acanthoeca* (order Acanthoecida) is thought to be an obligate heterotroph.

Table 1.2. General Features of Nine Representative Genera. Taxonomic levels are from the NCBI taxonomic framework.

Genus	Order	Class	Putative trophic mode	Reference
Prymnesium	Prymnesiales	Haptophyta	mixotrophic	Faure et al., 2019
Phaeocystis	Phaeocystales	Haptophyta	phototrophic	Rousseau et al., 2007
Florenciella	Florenciellales	Ochrophyta	mixotrophic	Quéguiner 2016, Li et al., 2021
Dictyocha	Dictyochales	Ochrophyta	photo- or mixotrophic	Quéguiner 2016
Pelagococcus	Pelagomonadales	Ochrophyta	phototrophic	Lewin et al., 1977
Karlodinium	Gymnodiniales	Dinophyceae	mixotrophic	Faure et al., 2019
Amoebophrya	Syndiniales	Dinophyceae	parasitic	Guillou et al., 2008
Acanthoeca	Acanthoecida	Choanoflagellata	heterotrophic	
Strombidium	Oligotrichida	Spirotrichea	hetero- or mixotrophic	Faure et al., 2019

The temporal partitioning of transcript abundances in the individual NMDS ordinations (Figure 1.2B, Supplementary Figure S1.5) showed that sampling time was an important driver for the transcript ordination for a majority of genera, with samples clustered by collection time and organized in a clock-like fashion. Whereas most genera showed some form of diel partitioning, the transcriptional profiles of dinoflagellates such as *Karlodinium* and *Amoebophrya*, the choanoflagellate *Acanthoeca* (Figure 1.2B, Supplementary Figure S1.5), and the heterotrophic protists kinetoplasid *Neobodo* and the stramenopile *Phytopthora*

(Supplementary Figure S1.5) were not distinguished by sampling time, indicating that not all organisms entrain their transcriptional activity to the diel cycle.

1.4.3 Diel signatures of environmental transcriptome bins

We determined the proportion of oscillating transcripts within each of the 48 environmental genera. The diel periodicity of transcript abundances across the 48 environmental genera was tested for a combined total of 103,904 KOfam gene families (RAIN analyses, maximum p-value = 0.0044, FDR < 0.05) (Figure 1.3A, Supplementary Data Sheet S1.4). Statistically significant diel periodicity in transcript abundance was detected for 9,153 gene families, with peaks in abundance assigned to one of the six sampling times: 06:00, 10:00, 14:00, 18:00, 22:00 and 02:00 HST (Figure S1.6, Supplementary Data Sheet S1.4). Haptophytes displayed the highest proportions of diel-oscillating transcript abundances for different gene families (Figure 1.3A, Supplementary Data Sheet S1.4). About 34% of the *Phaeocystis* and about 18% of *Prymnesium* gene families underwent significant oscillations in transcript abundance. Most ochrophyte environmental genera displayed diel periodicity in transcript abundance in at least 15% of their gene families, with the highest proportion (21.5%) observed in the environmental silicoflagellate *Dictyocha* (Figure 1.3A, Supplementary Data Sheet S1.3). The dinoflagellates displayed a low proportion (average of less than 3%) of gene families with diel transcript abundance patterns, as did the purely heterotrophic opisthokonts (Figure 1.3A). The heterotrophic environmental stramenopile *Phytophthora* displayed a comparably low (3.2%) proportion of diel gene families. Three other heterotrophic organisms stood out in this analysis. The environmental genera of *Lepeophtheirus* (copepod), *Acanthoeca* (choanoflagellate), and *Strombidium* (ciliate) each displayed relatively high proportions of diel oscillations in transcript abundance across gene families (23%, 16%, 20% respectively), comparable to the haptophytes and non-diatom ochrophytes (Figure 1.3A, Supplementary Data Sheet S1.3). Thus, the extent of diel periodicity was not directly correlated with trophic mode and appeared instead to be taxa specific.

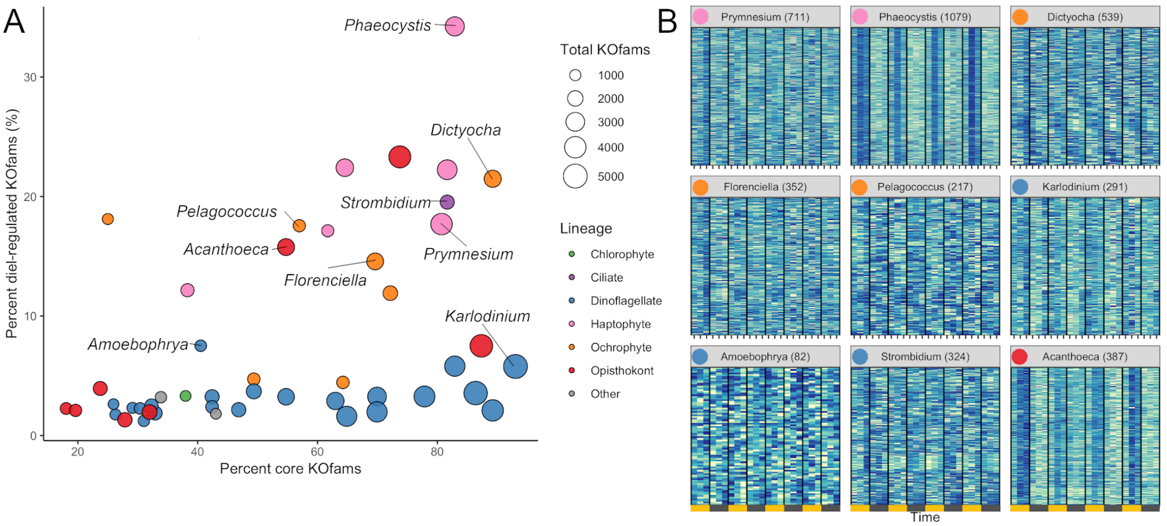


Figure 1.3: Diel periodicity in environmental genus bins. A) Estimated completeness and diel periodicity of gene family transcript abundance in 48 genera from Fig 1B. Each circle represents one environmental genus bin. The area is proportional to the total number of KOfams identified in the genus. The nine representative genera (Fig. 2B) are labelled. **B)** Normalized abundance heat map of significantly periodic gene families from 9 representative protist genera. Yellow and gray bars denote light (06:00, 10:00, 14:00 HST) and dark (18:00, 22:00, 02:00 HST) periods, respectively. Numbers in parentheses indicate the number of significantly period gene families (FDR < 0.05). Each row corresponds to a gene family, ordered by hierarchical clustering of abundance patterns. Color corresponds to row-normalized abundance values for each gene family. Colored dot by genus name corresponds to Lineage from A. The order and presence of gene families is not maintained between facets.

The transcription profiles of diel oscillating gene families (KOfams) were evaluated for the nine representative genera (Figure 1.3A, B) along with the remaining 41 environmental taxa (Figure S1.6, S1.7). The representative environmental haptophyte *Phaeocystis* showed a transcriptional minimum at 10:00 (Figure 1.3B), coinciding with environmental light intensities reaching 2000 µmol/m/s at the sea surface. Similar mid-morning minima are seen in *Acanthoeca* (Figure 1.3B) and other opisthokonts (Figures S1.6, S1.7). Transcriptional patterns of the ochrophytes *Florenciella*, *Dictyocha*, and *Pelagococcus*, and the ciliate environmental genus *Strombidium* are more equally distributed across the 24-hour diel cycle (Figure 1.3B). The dinoflagellate *Karlodinium* and most (21 of 23) other dinoflagellate environmental genera (Figures S1.6, S1.7) displayed relatively minimal distinction of transcript abundances by time (Figure 1.3B) whereas the parasitic dinoflagellate *Amoebophrya* (Figure 1.3B) maintained a relatively high proportion (7.5%) of diel oscillating transcript abundances (Data Sheet S1.3). We retrieved sufficient sequences with similarity to representative multicellular metazoan (animal) genera in our reference database (*Capitella*, *Orcinus*, *Octopus*, *Salmo*, *Lepeophtheirus*, *Nematostella*, and *Oikopleura)* to pass our cut-off criteria. Each of these environmental "genera" displayed some temporal partitioning (Figure 1.3A, Figs, S1.5, S1.6, S1.7), particularly those identified as *Capitella* (annelid) and *Octopus* (mollusk). We constrained further analysis to protistan genera, recognizing that many metazoan taxa undergo substantial vertical migration over diurnal cycles. Diel vertical migration has also been observed in some species of motile phytoplankton (Shikata et al., 2015), but we assume that the relatively minor swimming speed of migrating protists is not a significant factor in the well-mixed surface layer of the NPSG. Overall, haptophyte and ochrophyte genera tended to display the highest proportions of diel oscillating gene families, regardless of whether the examined genera were primarily mixotrophic or photoautotrophic.

1.4.4 KEGG pathways with diel-oscillating transcript levels

The observed variation in diel transcriptional patterns between environmental protistan taxa suggested targeted allocation of transcriptional resources to different functional processes over the diel cycle. We sought to identify specific pathways with strong temporal partitioning by using Fisher's Exact Test. We determined whether particular KEGG pathways were enriched in diel-oscillating gene families at each of the six time points by conducting tests on all unique taxa-time-pathway combinations. A total of 110,754 genus-time-pathway tests identified 78 significant taxa-time-pathway combination enrichments (BH < 0.05; Table 1.3, Supplementary Data Sheet S1.5); these enrichments represent 28 total enriched KEGG pathways out of 430 pathways tested. These pathways encompass varied metabolic pathways including central carbon metabolism, lipid biosynthesis and degradation, protein biosynthesis and turnover, organellar processes, and signaling. We focused on KEGG pathways enriched in at least two of the 9 representative environmental genera: 'Photosynthesis'; 'Carbon fixation in photosynthetic organisms'; 'Porphyrin and chlorophyll metabolism'; 'Proteasome'; 'Protein processing in endoplasmic reticulum'; 'TCA cycle'; 'Circadian entrainment'; 'Oxidative phosphorylation'; and 'Ribosome' (Figure 1.4A).

Table 1.3. KEGG pathway enrichment analysis at the Genus level.
Italics denote representative pathways plotted in Figure 1.4A. Human Disease pathways not shown.

Class / *Genus*	**peak (HST)**	**KEGG pathway**
Haptophyceae		
Prymnesium	06:00	*Carbon fixation*, Glycolysis / Gluconeogenesis*, Photosynthesis,* Pentose phosphate pathway
	14:00	*Proteasome, Protein processing in ER,* Antigen processing and presentation
	18:00	*TCA cycle, Oxidative phosphorylation,* Thermogenesis
	22:00	Circadian entrainment
	02:00	*Ribosome*
Phaeocystis	06:00	*Photosynthesis, Carbon fixation*
	14:00	*Protein processing in ER*
	18:00	*TCA cycle, Proteasome,* Thermogenesis
	02:00	Lysosome
Dictyochophyceae		
Dictyocha	06:00	*Photosynthesis, Carbon fixation,* Glycolysis / Gluconeogenesis
	18:00	Thermogenesis
	22:00	*Ribosome*
Florenciella	-	*No enriched pathways at FDR < 0.05*
Pelagococcus	18:00	Fatty acid degradation
	02:00	*Ribosome*
Dinophyceae		
Karlodinium	-	*No enriched pathways at FDR < 0.05*
Amoebophrya	14:00	*TCA cycle, Oxidative phosphorylation*
Spirotrichea		
Strombidium	06:00	*Ribosome*
	10:00	Biosynthesis of unsaturated fatty acids
	14:00	*Carbon fixation, TCA cycle,* Glyoxylate and dicarboxylate metabolism
Choanoflagellata		
Acanthoeca	02:00	Amoebiasis

As expected, morning was characterized by enrichments in photosynthesis-related pathways. At dawn (06:00 HST), ‘Photosynthesis’ and ‘Carbon fixation’ pathways were enriched for diel oscillating gene families in phototrophs, including those with mixotrophic capabilities, with ‘Photosynthesis’ the most frequently enriched pathway. A majority of gene families in the ‘Photosynthesis’ pathway displayed peak transcript abundances at dawn (Figure 1.4A). At 10:00, ‘Porphyrin and chlorophyll metabolism’ was the only pathway enriched in photosynthetic and mixotrophic lineages. The paucity of enrichments at 10:00 was consistent with the observed minimum in transcript peak across most environmental genera (Figures S1.6, S1.7). One of the few whole pathway enrichments observed in dinoflagellate lineages was ‘Photosynthesis’ in *Alexandrium* (Data Sheet S1.5) at 10:00. Within the phagocytic taxa, *Strombidium* displayed significant enrichments in the ‘Ribosome’ pathway at dawn (Figure 1.4A), similar to that seen in the sea squirt *Oikopleura* (Data Sheet S1.5).

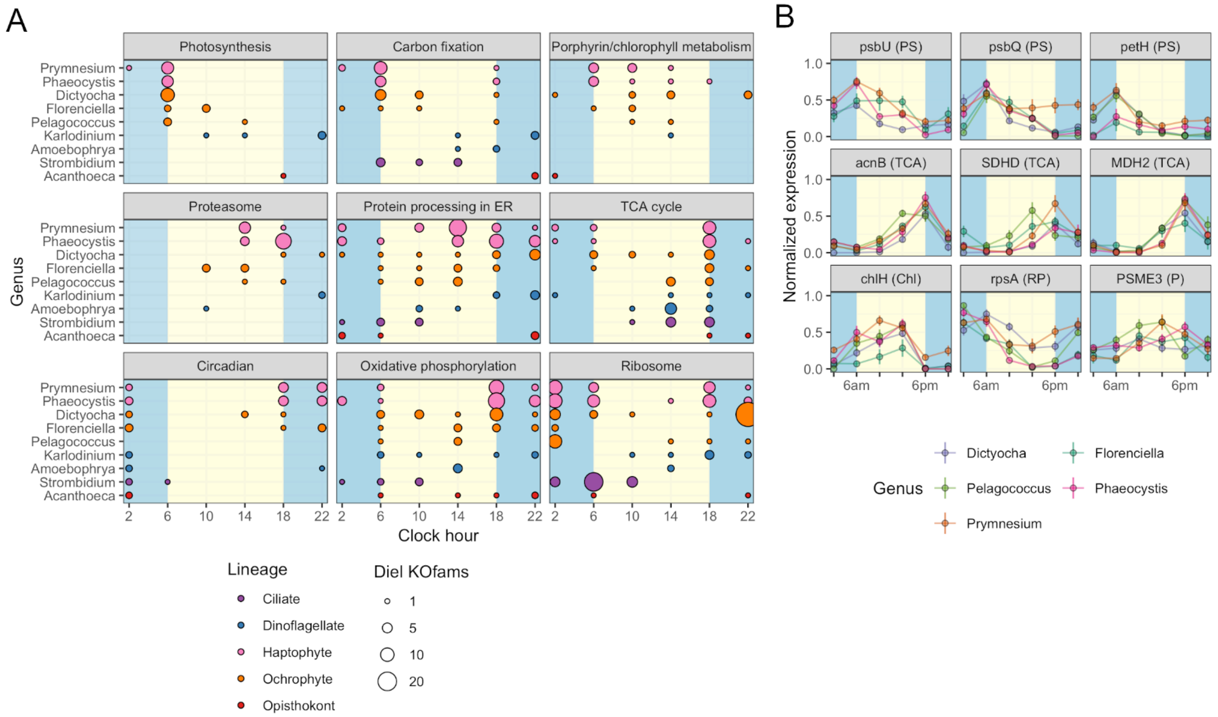


Figure 1.4. Diel periodicity of metabolic pathways and gene families in select genera. A) Peak times for gene families in select KEGG pathways for 9 genera. Circles indicate peak timing for KOfams with significant periodicity (FDR < 0.05), with circle size scaled by the number of KOfams peaking at a given time. Blue and yellow shading denote night and day periods, respectively. **B)** Temporal abundance of nine gene families from pathways in A. Haptophytes and ochrophytes from A are shown. Transcript abundances are min-max normalized. Lines connect the mean expression level across all identical sampling times (two replicates and four days). Vertical bars are standard error (n = 8). Blue and yellow shading denote night and day periods, respectively. Full protein names are psbU, photosystem II PsbU protein; psbQ, photosystem II oxygen-evolving enhancer protein 3; petH, ferredoxin-NADP+ reductase; acnB, aconitate hydratase 2; SDHD, succinate dehydrogenase membrane anchor subunit; MDH2, malate dehydrogenase; chlH, magnesium chelatase subunit H; rpsA, small subunit ribosomal protein S1; PSME3, proteasome activator subunit 3. Letters in parentheses next to the protein symbol indicate KEGG pathways: PS, Photosynthesis metabolism; TCA, TCA Cycle; N, Chl, Porphyrin and chlorophyll metabolism; RP, Ribosomal protein; P, Proteasome.

Later in the day, organisms shifted to protein processing and respiration-based pathways. The ‘Proteasome’, ‘Protein processing in endoplasmic reticulum’, and ‘TCA cycle’ pathways were enriched at 14:00; with the latter pathway enriched in five haptophyte genera (Figure 1.4A, Data Sheet S1.5). The haptophyte and ochrophyte genera had transcript abundance peaks in ‘Proteasome’-associated gene families in the afternoon (Figure 1.4A) whereas *Amoebophyra* and *Strombidium* displayed transcript abundance peaks in ‘TCA cycle’-associated gene families at this time point. By dusk (18:00), pathways involved in energy-yielding processes became transcriptionally prominent and included enrichments in ‘Oxidative phosphorylation’ and additional enrichments in the TCA cycle. At this time the greatest number of pathway enrichments were detected, although the timing of transcript abundance peaks for individual gene families varied across genera. The ochrophytes displayed peaks in transcript abundance for individual TCA-associated gene families throughout the day, while the haptophytes also displayed peaks for TCA-associated genes before and at dawn (Figure 1.4A). ‘Oxidative phosphorylation’ peak times also occurred across other time points in ochrophytes, but primarily between dusk to dawn in *Prymnesium* and *Phaeocystis*, afternoon in *Amoebophrya* and before noon in *Strombidium* (Figure 1.4A).

The night timepoints (22:00 and 02:00) were characterized by enrichments in ‘Circadian entrainment’, ‘Oxidative phosphorylation’, and ‘Ribosome’ pathways in most of the nine genera. *Pelagococcus* did not display enrichment in Circadian entrainment at any time point, was enriched in ‘Oxidative phosphorylation’ during the day rather than at night and was enriched in the ‘Ribosome’ pathway at the 02:00 time point.

Additional pathways were enriched in specific genera (Data Sheet S1.5). The greatest number of pathway enrichments at the genus level was found in the haptophyte *Prymnesium* (16), followed by other haptophytes and ochrophyte genera (38 and 17 pathway-time enrichments in total, respectively). ‘Fatty acid biosynthesis’ and ‘Fatty acid degradation’ were enriched in a subset of the genus-level analyses. For example, biosynthesis of unsaturated fatty acids was enriched in the ciliate *Strombidium* at 10:00 (Data Sheet S1.5), alluding to a build-up of energy storage reserves during the day. Enrichment of the ‘Thermogenesis’ pathway in several genera (*Phaeocystis*, *Prymnesium*, and *Dictyocha*) was driven by peaks in transcripts encoding mitochondrial-targeted proteins. In addition, a subset of enriched pathways at this time point are characterized as Human Diseases in KEGG; these pathways contain ubiquitous gene families such as calmodulin, calcium channels, cytochrome oxidase, ATPase, and some components of the TCA cycle (Data Sheet S1.5).

Five pathway enrichments were detected in Dinoflagellate genera with the most in *Amoebophrya* (‘TCA cycle’ and ‘Oxidative phosphorylation’ were both enriched at 14:00); the other dinoflagellates had either one or no enrichments. The choanoflagellate genus *Acanthoeca* had few significant peak times in the most commonly enriched diel pathways (Figure 1.4A), even though nearly 16% of *Acanthoeca* gene families displayed diel oscillations in transcript abundances, a value comparable to *Prymnesium* (18%) (Figure 1.3A, Supplementary Data Sheets S1.3, S1.5). The only significant enrichment attributed to *Acanthoeca* was the disease pathway 'Amoebiasis', at 02:00.

The striking similarity in the overall patterns of the representative ochrophyte and haptophyte genera (Figure 1.4A) prompted an examination of transcription abundances for select gene families across the diel cycle, including those involved in 'Photosynthesis' and 'Chlorophyll metabolism', the 'TCA cycle', 'Ribosome', and 'Proteasome' pathways (Figure 1.4B). We focused on those gene families with transcript abundances that oscillated over the diel cycle and were detected in at least three genera from the representative ochrophytes (*Dictyocha*, *Florenciella*, and *Pelagococcus*) or haptophytes (*Phaeocystis* and *Prymnesium*). Each of the photosynthesis gene family transcript abundance patterns were remarkably consistent across genera. In general, transcript abundance was highest at dawn with a decline through the day culminating in a dusk minimum. Gene family transcripts in the TCA cycle pathway were also tightly correlated with an inverse transcriptional pattern to photosynthesis-related transcripts, with sharp peaks at the dusk time point and 02:00 or 06:00 minima. Two gene families involved in either protein synthesis (small subunit ribosomal protein S1, rpsA) or protein degradation (proteasome activator subunit 3, PSME3) had generally opposing phases, with rpsA transcripts at higher abundances in the dark until dusk and PSME3 transcripts with higher day-time abundances. These results suggested that the ochrophyte and haptophyte phytoplankton lineages maintained similar diel regulation of genes within these pathways.

To identify potentially distinguishing features of the major phytoplankton lineages, we further examined pathway enrichments at the class level corresponding to the Haptophyceae, Dictyophyceae, Pelagophyceae, and Dinophyceae (Table 1.4). These four classes are inclusive of 6, 23, 1 and 3 of the 48 genera meeting cut-off criteria, respectively. A total of 45 KEGG pathway enrichments over the diel cycle were detected for the Haptophyceae; the Dictyophyceae and Pelagophyceae displayed both lower overall numbers of KOfams and sequence coverage with detection of 21 and 13 enrichments, respectively. Few diel enrichments were detected within the Dinophyceae.

Table 1.4. KEGG pathway enrichment analysis at the Class level.
Bold, pathways specific to each Class. Italics denote representative pathways plotted in Figure 4A. Human Disease pathways not shown. Pathways for each time per taxa are listed by decreasing significance order (FDR < 0.05).

Class	peak (HST)	KEGG pathway
Haptophyceae		
	06:00	*Carbon fixation, Photosynthesis*, Fatty acid biosynthesis, Glycolysis / Gluconeogenesis, Pentose phosphate pathway, **Phenylalanine tyrosine and tryptophan biosynthesis**, **Cysteine and methionine metabolism**, **Alanine aspartate and glutamate metabolism, Biotin metabolism, Tropane piperidine and pyridine alkaloid biosynthesis**, **Fructose and mannose metabolism**
	10:00	*Porphyrin and chlorophyll metabolism*
	14:00	*Protein processing in ER, Proteasome*, **RNA transport**, Spliceosome
	18:00	Spliceosome, **DNA replication,** *TCA cycle, Oxidative phosphorylation*, **Cell cycle,** Synaptic vesicle cycle, **Nucleotide excision repair**, *Ribosome*, **Meiosis**, Fatty acid degradation, **Mismatch repair,** Thermogenesis, Dopaminergic synapse
	22:00	**Calcium signaling pathway, cAMP signaling pathway**, Thermogenesis, **Fc epsilon RI signaling pathway**, *Circadian entrainment*, **Natural killer cell mediated cytotoxicity**, Oxytocin signaling pathway, **Renin secretion**, **Aldosterone synthesis and secretion**, **Phospholipase D signaling pathway**, **MAPK signaling pathway**
	02:00	**Valine leucine and isoleucine biosynthesis**, **Protein export**
Dictyochophyceae		
	06:00	*Carbon fixatio*n, Glycolysis / Gluconeogenesis, *Photosynthesis*, **Glycine serine and threonine metabolism**, Pentose phosphate pathway, **Sulfur metabolism**, **Pyruvate metabolism**, **Bacterial secretion system**
	10:00	*Porphyrin and chlorophyll metabolism, Proteasome*, *Photosynthesis*, **Photosynthesis - antenna proteins**
	14:00	*Protein processing in ER,* **Antigen processing and presentation**
	18:00	*TCA cycle*, **Cardiac muscle contraction**, Thermogenesis, *Oxidative phosphorylation*
	22:00	*Ribosome*, **Salivary secretion**, **NOD-like receptor signaling pathway**
	02:00	*No enriched pathways at FDR < 0.05*
Pelagophyceae		
	06:00	*Carbon fixation*, *Photosynthesis*, Fatty acid biosynthesis,
	10:00	*Porphyrin and chlorophyll metabolism*
	14:00	*Protein processing in ER*
	18:00	*Ribosome*, *Proteasome, TCA cycle*
	22:00	Dopaminergic synapse, **Olfactory transduction**, Oxytocin signaling pathway, **Melanogenesis**, *Circadian entrainment*
	02:00	*No enriched pathways at FDR < 0.05*
Dinophyceae		
	06:00	*No enriched pathways at FDR < 0.05*
	10:00	*No enriched pathways at FDR < 0.05*
	14:00	**Drug metabolism - other enzymes**
	18:00	*Carbon fixation*, **Lysosome**
	22:00	Synaptic vesicle cycle, **Collecting duct acid secretion**, *Oxidative phosphorylation, Photosynthesis*
	02:00	*Carbon fixation*

These analyses uncovered a number of distinguishing features of specific groups (Table 1.4). Specific enrichment of 'Fructose and mannose metabolism' pathways in Haptophyceae may reflect an enhanced sensitivity for detecting enrichments in this well-covered class of organisms as this pathway is directly linked to glycolysis. The specific enrichment in 'Pyruvate metabolism' in Dictyophyceae may reflect a similar routing of fixed carbon towards lipids. Haptophyceae were also specifically enriched at dawn in multiple amino acid metabolic pathways ('Alanine aspartate and glutamate metabolism', 'Phenylalanine, tyrosine and tryptophan biosynthesis', and 'Cysteine and methionine metabolism') as well as 'Biotin metabolism'. Other pathways with significant temporal enrichment only in Haptophyceae include 'Aldosterone synthesis and secretion', 'Cellular senescence pathway', 'Valine, leucine, and isoleucine biosynthesis', and 'Protein export' pathways.

The Dictyophyceae specifically displayed dawn enrichments in both 'Sulfur metabolism' and 'Glycine, serine and threonine metabolism', two pathways linked via the metabolite homoserine. (Figures S1.8, S1.9). Dictyophyceae also had a unique enrichment in the 'Photosynthesis - antenna proteins' pathway. The morning-peaking genes families in the sulfur metabolism pathway form the majority of complete sulfate assimilation pathways in Dictyophyceae. Specific enrichment of 'Antigen processing and presentation' and 'Cardiac muscle contraction' pathways in the afternoon and at dusk, respectively may reflect enrichments in general ATPase and transport functions. Enrichments identified only in Pelagophyceae were limited to 'Olfactory transduction' and 'Melanogenesis', both at 22:00. Surprisingly, the Dinophyceae were enriched in 'Carbon fixation in photosynthetic organisms' and 'Photosynthesis' pathways during the night whereas all other classes were enriched in these pathways in the early morning. The Dinophyceae also specifically displayed enrichment in the 'Lysosome' pathway at dusk (18:00), suggesting a dominant signal from heterotrophic or mixotrophic dinoflagellates, with nighttime partitioning of gene transcription likely associated with digestion of engulfed prey. (Figure S1.10). We note that aside from the 'Drug metabolism' pathway, all enrichments identified in Dinophyceae are organelle localized (plastid or lysosome).

1.5 Discussion

Microbial eukaryotes perform vital functions in the NPSG ecosystem, including phototrophy, heterotrophy, mixotrophy, and parasitism. The generation of time-, function-, and taxonomy-resolved environmental transcriptome bins in this study produced insights about community composition and abundance, the connection between transcriptome composition and trophic state, the degree of diel regulation utilized by these environmental taxa, and the timing of functional processes throughout the diel cycle. Because of the high amount of functional, temporal and taxonomic resolution provided by the annotated metatranscriptomes, we've utilized an analytical hierarchy in this study: we begin with broad survey of all environmental bins, then narrow our global functional analysis to the most complete environmental genera, and further focus our examination of metabolism to abundant representative genera and prominent pathways. This includes the smaller sized genera (< 7 µm diameter) that make significant contributions to daily productivity in the mixed layer of the NPSG (Freitas et al., 2020). Although there may be subtle variations in environmental factors contributing to the transcriptional differences between genera, our assumption is that diel cycles are the critical driver of microbial life in the NPSG, and we've constrained the majority of our statistical analysis to diel periodicity of gene families and the timing of pathways.

The environmental genera in this study can be roughly categorized as having either high, intermediate, or low levels of diel transcriptional regulation. The haptophyte *Phaeocystis* has the singular distinction of having the highest proportion of diel gene families. Although *Phaeocystis* was not the only pure phototroph captured by our study, it was the most abundant obligate phototroph among the complete genera bins, as such it is difficult to conclude from this study alone whether similarly high levels of diel regulation are a common phototroph strategy. Furthermore, though we chose to constrain our analysis to the genus-level or higher, there may be strain-level differences in diel regulation magnitude throughout the cosmopolitan *Phaeocystis*.

Organisms of comparable sequencing depth and completeness show similar levels of diel regulation in the 'intermediate' range (~12-22% of gene families being diel-oscillating) despite differences in their trophic state and evolutionary lineage. This includes non-dinoflagellate mixotrophic genera (including the haptophytes *Prymnesium* and *Chrysochromulina*, the dictyophyte *Florenciella,* and the non-constitutive mixotrophic ciliate *Strombidium),* the representative copepod genus *Lepeophtheirus* and the choanoflagellate genus *Acanthoeca*. Diel transcriptome structuring in these genera may reflect physiological attunement to the diel cycle; in copepods the high diel periodicity could be linked to diel vertical migration, but also reflect confounding issues from sub-group migration in and out of the mixed layer over diel cycles. The repeated observation of an 'intermediate' level of diel regulation across disparate trophic modes and evolutionary lineages suggests that transcript synchronization to diel cycles is a common and advantageous strategy in the NPSG.

The 'low' range of diel transcriptomes bin regulation (~8% or less of gene families) is predominantly occupied by dinoflagellates, including mixotrophic and heterotrophic genera. The low degree of regulation in in dinoflagellates is consistent with studies showing a relatively dampened transcriptional response to environmental stimuli (Lin 2011). Other strictly heterotrophic genera also had a low fraction of diel-regulated gene families, including the metazoan (animal) groups aside from *Lepeophtheirus* and heterotrophic protists other than *Acanthoeca*.

The coordination of protists to the diel period is also apparent in the distribution of diel transcript peak times. The highest proportions of transcript peak times occur at dusk, followed by dawn, underscoring the considerable metabolic re-arrangement of cells between light and dark periods. This has been observed previously in laboratory cultures: in the diatom *Thalassiosira pseudonana*, more gene transcript abundances peaked at dusk rather than dawn (Ashworth et al., 2013). The relative lack of gene family peak times at 10:00 HST, in particular, could be attributed to photo-protective purposes, minimizing transcription before the noon irradiation peak. This mid-morning minimum is also seen in prokaryotes: an earlier metatranscriptome study of the NPSG noted that the transcriptional minimum occurred closer to noon in contrast to MED4 culture transcriptomes (Ottesen 2014). This 'mid-day depression' could be a response to the detrimental effect of high UV radiation at mid-day in the surface layer and has been attributed to reduced growth, reduced DNA synthesis, and photochemical quenching in picoplankton of the equatorial Pacific (Vaulot and Marie 1999). The mid-day depression in transcription could also play a role in anti-viral defense, in that it restricts the replication of viral transcripts that are themselves tightly coordinated to diel cycles (Waldbauer et al., 2012).

Functional analysis of diel-regulated gene families allowed us to infer how protist lineages allocate transcriptional resources to functional processes over the diel cycle. Some metabolic features of phototrophs appear to be common strategies, such as dawn peaks in photosynthesis and energy storage pathways, and up-regulation of DNA synthesis and cell cycle elements at dusk. This strategy has been observed in diatoms and haptophytes previously. In the diatom *Thalassiosira pseudonana*, genes encoding cell division, DNA replication and repair, carbon metabolism, and oxidative phosphorylation enzymes are highly expressed at dusk, while at dawn transcripts involved in Photosynthesis, Carbon fixation, and Ribosomes were higher (Ashworth et al 2013); this general phototrophic strategy was also observed in metatranscripts from the calcifying haptophyte *Emiliania huxleyi* (Hernández et al., 2020). Transcriptional evidence for the cycling of carbon is consistent with *in situ* field measurements of triacylglycerol in the NPSG that show increasing concentrations in cellular biomass through the day and decreases after nightfall (Becker et al., 2018). In particular, the striking similarity in transcriptional patterns between haptophytes and ochrophytes, both members of the Chromealveolate supergroup that separated an estimated 1 billion years ago (Yoon et al., 2004), implies selective pressure to conserve diel regulation of key cellular processes.

Taxa-specific metabolic features may be important in understanding the fate of carbon in the surface layer as a function of community composition. At the class-level, the haptophytes (Haptophyceae) constitute a 'best case scenario' for diel pathway enrichments, as a combination of high diel regulation and population abundance contribute to sufficiently deep sequencing of haptophyte transcripts. The coordination of protein biosynthesis and turnover pathways to the diel cycle in haptophytes highlights the importance of these processes in diel cycles. The focused peaks of these pathway peaks in the afternoon suggests a large-scale proteomic remodeling prior to the evening metabolic switch, possibly involving the tagging and degradation of photosynthesis-related proteins in the dark hours when they are no longer useful. Intensive recycling of the proteome in the N-limited gyre could be an advantageous strategy to alleviate nitrogen stress, allowing nitrogen to be recycled between alternating proteomic regimes in a manner similar to the 'hot-bunking' of iron atoms between anti-phase metalloenzymes in *Crocosphaera* (Saito et al., 2011). The functional pathway enrichments specific to dictyophytes, such as 'Sulfur metabolism', 'Pyruvate metabolism', and 'Glycine, serine and threonine metabolism' are evidence of additional lineage-specific temporal partitioning of metabolic processes. The 'Sulfur metabolism' pathway includes sulfate assimilation, which is dominated by morning transcriptional peaks in this group. The connection of the 'Sulfur metabolism' pathway to 'Glycine, Serine, and Threonine Metabolism' through the metabolite homoserine highlights at important connections between diel-enriched metabolic features. Many protists participate in diel cycling of sulfonated compounds (Durham et al 2019), and these results hint at a possibly unexplored role of dictyophytes in sulfur cycling in the oceans. Future studies into the apparent metabolic differences between major protist lineages would benefit from metabolite- or protein-level data and fine-scale targeted investigation of specific metabolic pathways.

This study lends new perspective into dinoflagellate genetic regulation, which has significantly diverged from the transcriptional-level control utilized by most other eukaryotic lineages. The low proportion of diel gene families in dinoflagellate taxonomic bins is consistent with other studies showing a loss of transcriptional regulation in dinoflagellates (Lin 2011, Kojima et al., 2011). The presence of 5' spliced leaders, a transcript modification observed extensively in all major dinoflagellate orders, is a dinoflagellate adaptation that has been invoked as an alternative mechanism of gene regulation (Zhang *et al.,* 2007). As expected from their low level of diel regulation, the dinoflagellates did not return many significant pathway enrichments at any phylogenetic level of investigation. The few enriched dinoflagellate pathways are intriguing, as they contrast the generally minimum transcriptional regulation in dinoflagellates. At the class level, we observed organelle-targeted enrichments in three pathways ('Photosynthesis', 'Carbon fixation', and 'Lysosome'), and it would be interesting to investigate whether the gene families involved use 5' spliced leaders in a similar manner as cytosolic transcripts. Along with the enrichments, we found the timing of the Photosynthesis and Carbon fixation pathways (evening) to be surprising, as all other photosynthetic groups maintained these pathways with predominantly morning (06:00) peaks. Proteomic studies have been helpful in describing the altered function of environmental dinoflagellates across spatial gradients (Cohen et al., 2021) and a future proteomic studies of dinoflagellates with diel temporal resolution would be useful in determining the translational offset of dinoflagellate proteins from their transcript peaks over these cycles. Within the dinoflagellates, the notable outlier was the Syndiniales (represented here by genus *Amoebophyra*), an order of obligate parasitoids that infect dinoflagellates and other marine organisms (Guillou *et al*., 2008). This lineage branched off from other dinoflagellates early in their evolution, before the transcriptional adaptations that characterize more-derived dinoflagellates. Conservation of diel periodicity in *Amoebophyra* may provide hints into the ancestral transcriptional regulation of dinoflagellates that appears to have been lost the course of their evolutionary history in most other genera.

The proportion of taxonomy and function-calling for the assembled contigs in our dataset is comparable to the Tara Oceans Initiative eukaryotic gene catalog, which assigned about 25% of putative coding frames a functional annotation (Pfam) and about ~50% of the contigs a taxonomy at any level (Carradec et al, 2018). Despite some differences in our reference databases and separate functional databases, we see nearly identical annotation results in our data, with 54% and 25.7% of contigs receiving confident taxonomical and functional (KEGG Orthology) annotations, respectively. Much remains to be discovered in the "microbial dark matter" of sequence data that has no match to extant taxonomic or functional databases. As novel organisms continue to be sequenced and functional annotation databases are expanded in the future, the annotation of the raw transcript data generated by this study can continue to be improved and exploited to gain further scientific insight. Regardless, the thousands of currently catalogued gene family profiles provide us rich detail into the presence and timing of known metabolic pathways.

This study reveals the diel transcriptional dynamics of eukaryotic protists in the surface layers of the NPSG, elucidating common patterns and striking differences in the transcriptional phenotypes of the most abundant protists in the surface community. Taking all of these results in aggregate, a picture emerges of a eukaryotic community tightly orchestrated to the daily rhythms of sunlight, with phototrophic organisms structuring their transcriptomes around the clock to harness and store solar energy during the day, and to replicate, divide and possibly exchange signaling molecules at night. Mixotrophs are abundant and vital members of the protist community, and employ most of the core metabolic strategies as phototrophs. Understanding the most prevalent metabolic strategies employed by microbial eukaryotes in conjunction with the differences that distinguish them is critical to furthering our understanding of how carbon, nutrients and energy flow through the surface ocean ecosystem.

1.6 Author Contributions

RDG, SC, BPD, and EVA conceived and designed the research project. RDG and SC conducted the metatranscriptomic analyses. All authors contributed to data interpretation and the writing of the manuscript.

1.7 Funding

This work was supported by a grant from the Simons Foundation (SCOPE Award 329108 to E.V.A) and XSEDE Grant Allocation (OCE160019 to R.D.G).

1.8 Acknowledgements

We thank the crew and science party of the R/V *Kilo Moana* on cruise KM1513.

1.9 Data Availability Statement

KM1513 cruise information, plots, and associated environmental data for the HOE Legacy II cruise can be found online at http://hahana.soest.hawaii.edu/hoelegacy/hoelegacy.html. Raw metatranscriptome short-read sequence data is available in the NCBI Sequence Read Archive under BioProject ID PRJNA492142. Assembled contigs are deposited in Zenodo (https://doi.org/10.5281/zenodo.5009803). Code associated with this project is available on Github (https://github.com/armbrustlab/diel_eukaryotes).

1.10 References

Alexander, H., Rouco, M., Haley, S. T., Wilson, S. T., Karl, D. M., & Dyhrman, S. T. (2015). Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. *Proceedings of the National Academy of Sciences*, *112*(44), E5972-E5979.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020). KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, *36*(7), 2251-2252.

Ashworth, J., Coesel, S., Lee, A., Armbrust, E. V., Orellana, M. V., & Baliga, N. S. (2013). Genome-wide diel growth state transitions in the diatom Thalassiosira pseudonana. *Proceedings of the National Academy of Sciences*, *110*(18), 7518-7523.

Becker, K. W., Collins, J. R., Durham, B. P., Groussman, R. D., White, A. E., Fredricks, H. F., Ossolinski, J. E., Repeta, D. J., Carini, P., Armbrust, E. V., & Van Mooy, B. A. (2018). Daily changes in phytoplankton lipidomes reveal mechanisms of energy storage in the open ocean. *Nature communications*, *9*(1), 5179.

Becker, K. W., Harke, M. J., Mende, D. R., Muratore, D., Weitz, J. S., DeLong, E. F., Dyhrman, S. T., & Van Mooy, B. A. (2020). Combined pigment and metatranscriptomic analysis reveals highly synchronized diel patterns of phenotypic light response across domains in the open oligotrophic ocean. *The ISME Journal*, 1-14.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. **30**, 2114-2120 *Bioinformatics* (2014).

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, *12*(1), 59.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, *34*(5), 525.

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., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Tara Oceans Coordinators, Jaillon, O., Aury, J., Karsenti, E., Sullivan, M. B, Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., de Vargas, C., Iudicone, D., Bowler C., & Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature communications*, *9*(1), 373.

Coesel, S. N., Durham, B. P., Groussman, R. D., Hu, S. K., Caron, D. A., Morales, R. L., Ribalet, F., & Armbrust, E. V. (2021). Diel transcriptional oscillations of light-sensitive regulatory elements in open-ocean eukaryotic plankton communities. *Proceedings of the National Academy of Sciences*, *118*(6).

Durham, B. P., Boysen, A. K., Carlson, L. T., Groussman, R. D., Heal, K. R., Cain, K. R., Morales, R. L., Coesel, S. N., Morris, R. M., Ingalls, A. E., & Armbrust, E. (2019). Sulfonate-based networks between eukaryotic phytoplankton and heterotrophic bacteria in the surface ocean. *Nature microbiology*, *4*(10), 1706-1715.

Cohen, N. R., McIlvin, M. R., Moran, D. M., Held, N. A., Saunders, J. K., Hawco, N. J., Brosnahan, M., DiTullio, G. R., Lamborg, C., McCrow, J. P., Dupont, C. L., Allen, A. E., & Saito, M. A. (2021). Dinoflagellates alter their carbon and nutrient metabolic strategies across environmental gradients in the central Pacific Ocean. *Nature Microbiology*, 1-14.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS computational biology*, *7*(10), e1002195.

Faure, E., Not, F., Benoiston, A. S., Labadie, K., Bittner, L., & Ayata, S. D. (2019). Mixotrophic protists display contrasted biogeographies in the global ocean. *The ISME journal*, *13*(4), 1072-1083.

Freitas, F. H., Dugenne, M., Ribalet, F., Hynes, A., Barone, B., Karl, D. M., & White, A. E. (2020). Diel variability of bulk optical properties associated with the growth and division of small phytoplankton in the North Pacific Subtropical Gyre. *Applied Optics*, *59*(22), 6702-6716.

Frias-Lopez, J., Thompson, A., Waldbauer, J., & Chisholm, S. W. (2009). Use of stable isotope-labelled cells to identify active grazers of picocyanobacteria in ocean surface waters. *Environmental microbiology*, *11*(2), 512-525.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman., N., & Regev, A.. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, *29*(7), 644.

Guillou L., Viprey M., Chambouvet A., Welsh R. M., Kirkham A. R., Massana R., Scanlan D. J., Worden A. Z. (2008). Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environmental Microbiology*, *10*(12), 3349-3365.

Hernández, M. L., Hennon, G. M. M., Harke, M. J., Frischkorn, K. R., Haley, S. T., & Dyhrman, S. T. (2019). Transcriptional patterns of Emiliania huxleyi in the North Pacific Subtropical Gyre reveal the daily rhythms of its metabolic potential. *Environmental microbiology*.

Hu, S. K., Connell, P. E., Mesrop, L. Y., & Caron, D. A. (2018). A hard day's night: Diel shifts in microbial eukaryotic activity in the North Pacific Subtropical Gyre. *Frontiers in Marine Science*, *5*, 351.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A.., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., Deschamps, P., Dyhrman, S. T., Edvardsen, B., Gates, R. D., Gobler, C. J., Greenwood, S. J., Guida, S. M., Jacobi, J. L., Jakobsen, K. S., James, E. R., Jenkins, B., John, U., Johnson, M. D., Juhl, A. R., Kamp, A., Katz, L. A., Kiene, R., Kudryavtsev, A., Leander, B. S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A. M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M. A., Murray, S., Nadathur, G., Nagai, S., Ngam, P. B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M. C., Rengefors, K., Romano, G., Rumpho, M. E., Rynearson, T., Schilling, K. N., Schroeder, D. C., Simpson, A. G. B., Slamovits, C. H., Smith, D. R., Smith, G. J., Smith, S. R., Sosik, H. M., Stief, P., Theriot, E., Twary, S. N., Umale, P. E., Vaulot, D., Wawrik, B., Wheeler, G. L., Wilson, W. H., Xu, Y., Zingone, A., & Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS biology*, *12*(6), e1001889.

Kojima, S., Shingle, D. L., & Green, C. B. (2011). Post-transcriptional control of circadian rhythms. *Journal of cell science*, *124*(3), 311-320.

Kolody, B. C., McCrow, J. P., Allen, L. Z., Aylward, F. O., Fontanez, K. M., Moustafa, A., Moniruzzaman, M., Chavez, F. P., Scholin, C. A., Allen, E. E., Worden, A. Z., DeLong, E. F., & Allen, A. E.. (2019). Diel transcriptional response of a California Current plankton microbiome to light, low iron, and enduring viral infection. *The ISME journal*, *13*(11), 2817-2833.

Lambert, B. S., Groussman, R. D., Schatz, M. J., Coesel, S. N., Durham, B. P., Alverson, A. J., White, A. E., & Armbrust, E. V. (2022). The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. *Proceedings of the National Academy of Sciences*, *119*(7), e2100916119.

Lewin, J., Norris, R. E., Jeffrey, S. W., & Pearson, B. E. (1977). AN ABERRANT CHRYSOPHYCEAN ALGA PELAGOCOCCUS SUBVIRIDIS GEN. NOV. ET SP. NOV. FROM THE NORTH PACIFIC OCEAN 1, 2. *Journal of Phycology*, *13*(3), 259-266.

Li, Q., Edwards, K. F., Schvarcz, C. R., Selph, K. E., & Steward, G. F. (2021). Plasticity in the grazing ecophysiology of Florenciella (Dichtyochophyceae), a mixotrophic nanoflagellate that consumes Prochlorococcus and other bacteria. *Limnology and Oceanography*, *66*(1), 47-60.

Lin, S. (2011). Genomic understanding of dinoflagellates. *Research in microbiology*, *162*(6), 551-569.

Mitra, A., Flynn, K. J., Tillmann, U., Raven, J. A., Caron, D., Stoecker, D. K., Not, F., Hansen, P. J., Hallegraeff, G., Sanders, R. Wilken, S., McManus, G., Johnson, M., Pitta, P., Våge, , B., Calbet, A., Thingstad, F., Jeong, H. J., Burkholder, J., Glibert, P. M., Granéli, E., & Lundgren, V. (2016). Defining planktonic protist functional groups on mechanisms for energy and nutrient acquisition: incorporation of diverse mixotrophic strategies. *Protist*, *167*(2), 106-120.

Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2019). vegan: Community Ecology Package. R package version 2.5-5. https://CRAN.R-project.org/package=vegan

Ottesen, E. A., Young, C. R., Gifford, S. M., Eppley, J. M., Marin III, R., Schuster, S. C., Scholin, C. A., & DeLong, E. F. (2014). Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science*, *345*(6193), 207-212.

Quéguiner, B. (2016). *The Biogeochemical Cycle of Silicon in the Ocean*. John Wiley & Sons.

Ribalet, F., Swalwell, J., Clayton, S., Jiménez, V., Sudek, S., Lin, Y., Johnson, C. I., Worden, A. Z., & Armbrust, E. V. (2015). Light-driven synchrony of Prochlorococcus growth and mortality in the subtropical Pacific gyre. *Proceedings of the National Academy of Sciences*, *112*(26), 8008-8012.

Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 16, 276–277 (2000).

Rii, Y. M. (2016). *Ecology of photosynthetic picoeukaryotes in the oligotrophic ocean: Diversity, activity, and dynamics* (Doctoral dissertation, University of Hawai'i at Manoa).

Rothhaupt, K. O. (1996). Laboratory experiments with a mixotrophic chrysophyte and obligately phagotrophic and photographic competitors. *Ecology*, *77*(3), 716-724.

Rousseau, V., Chrétiennot-Dinet, M. J., Jacobsen, A., Verity, P., & Whipple, S. (2007). The life cycle of Phaeocystis: state of knowledge and presumptive role in ecology. *Biogeochemistry*, *83*(1-3), 29-47.

Saito, M. A., Bertrand, E. M., Dutkiewicz, S., Bulygin, V. V., Moran, D. M., Monteiro, F. M., Follows, M. J., Valois, F. W. & Waterbury, J. B. (2011). Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph Crocosphaera watsonii. *Proceedings of the National Academy of Sciences*, *108*(6), 2184-2189.

Satinsky, B. M., Gifford, S. M., Crump, B. C. & Moran, M. A. in *Methods in Enzymology* (ed. DeLong, E. F.) **531**, 237–250 (Elsevier Inc., 2013).

Sievert, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., & Despouy, P. (2018). plotly for R.

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome research*.

Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature communications*, *9*(1), 2542.

Stoecker, D. K., Johnson, M. D., de Vargas, C., & Not, F. (2009). Acquired phototrophy in aquatic protists. *Aquatic Microbial Ecology*, *57*(3), 279-310.

Tenenbaum, D. (2016). KEGGREST: Client-side REST access to KEGG. *R package version*, *1*(1).

Thaben, P. F., & Westermark, P. O. (2014). Detecting rhythms in time series with RAIN. *Journal of biological rhythms*, *29*(6), 391-400.

Vaulot, D., & Marie, D. (1999). Diel variability of photosynthetic picoplankton in the equatorial Pacific. *Journal of Geophysical Research: Oceans*, *104*(C2), 3297-3310.

Villareal, T. A., Altabet, M. A., & Culver-Rymsza, K. (1993). Nitrogen transport by vertically migrating diatom mats in the North Pacific Ocean. *Nature*, *363*(6431), 709-712.

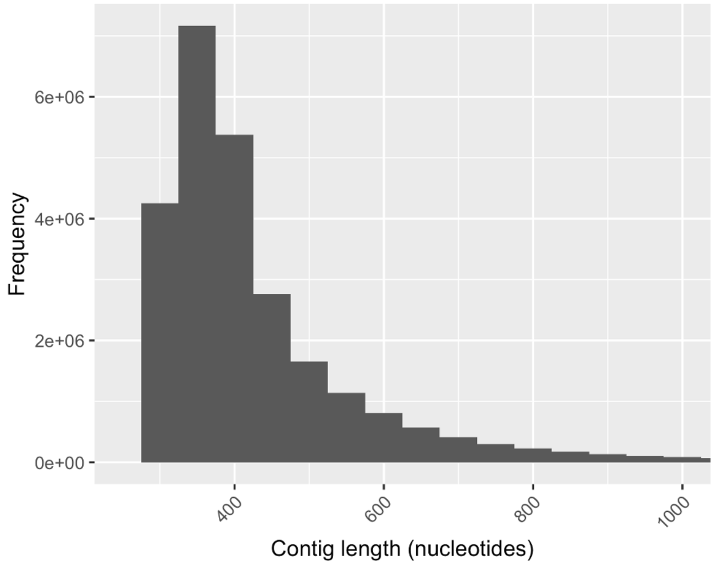
Waldbauer, J. R., Rodrigue, S., Coleman, M. L., & Chisholm, S. W. (2012). Transcriptome and proteome dynamics of a light-dark synchronized bacterial cell cycle. *PloS one*, *7*(8), e43432.

Wilson, S. T., Aylward, F. O., Ribalet, F., Barone, B., Casey, J. R., Connell, P. E., Eppley, J. M., Ferrón, S. E., Fitzsimmons, J. N., Hayes, C. T., Romano, A. E., Turk-Kubo, K. A., Vislova, A., Armbrust, E. V., Caron, D. A., Church, M. J., Zehr, J. P., Karl, D. M., & DeLong, E. F. (2017). Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium Crocosphaera. *Nature microbiology*, *2*(9), 1-9.

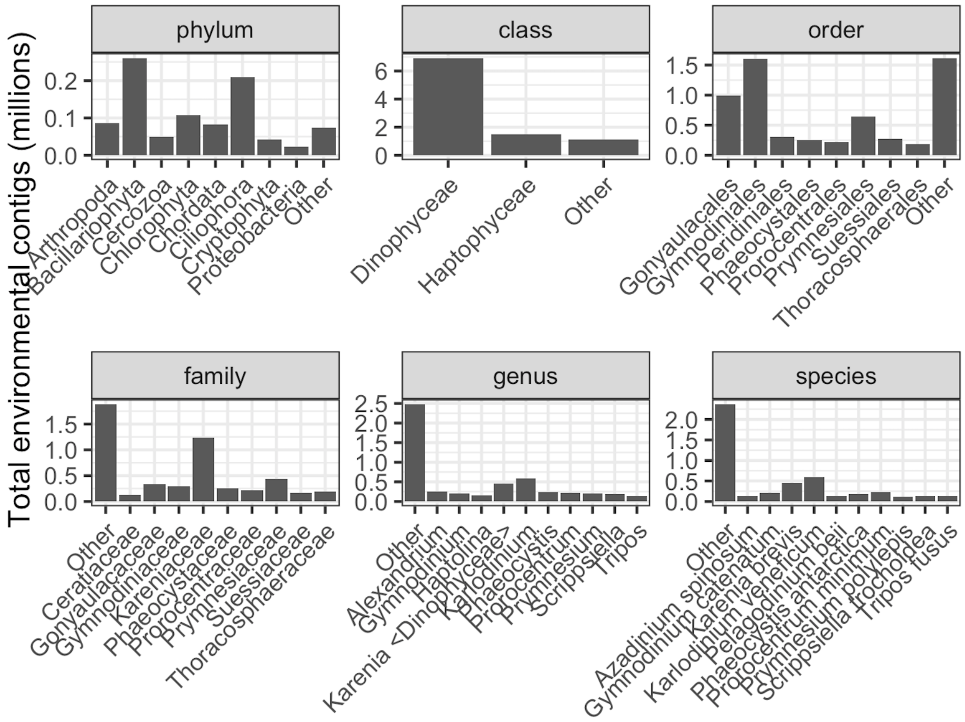
Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G., & Bhattacharya, D. (2004). A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular biology and evolution*, *21*(5), 809-818.

Zhang, H., Hou, Y., Miranda, L., Campbell, D. A., Sturm, N. R., Gaasterland, T., & Lin, S. (2007). Spliced leader RNA trans-splicing in dinoflagellates. *Proceedings of the National Academy of Sciences*, *104*(11), 4618-4623.

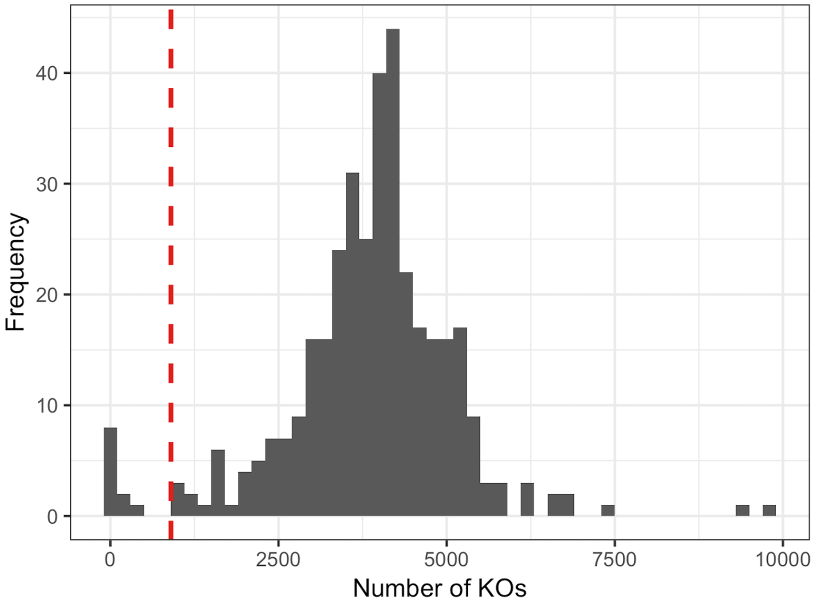
1.11 Supplementary Figures



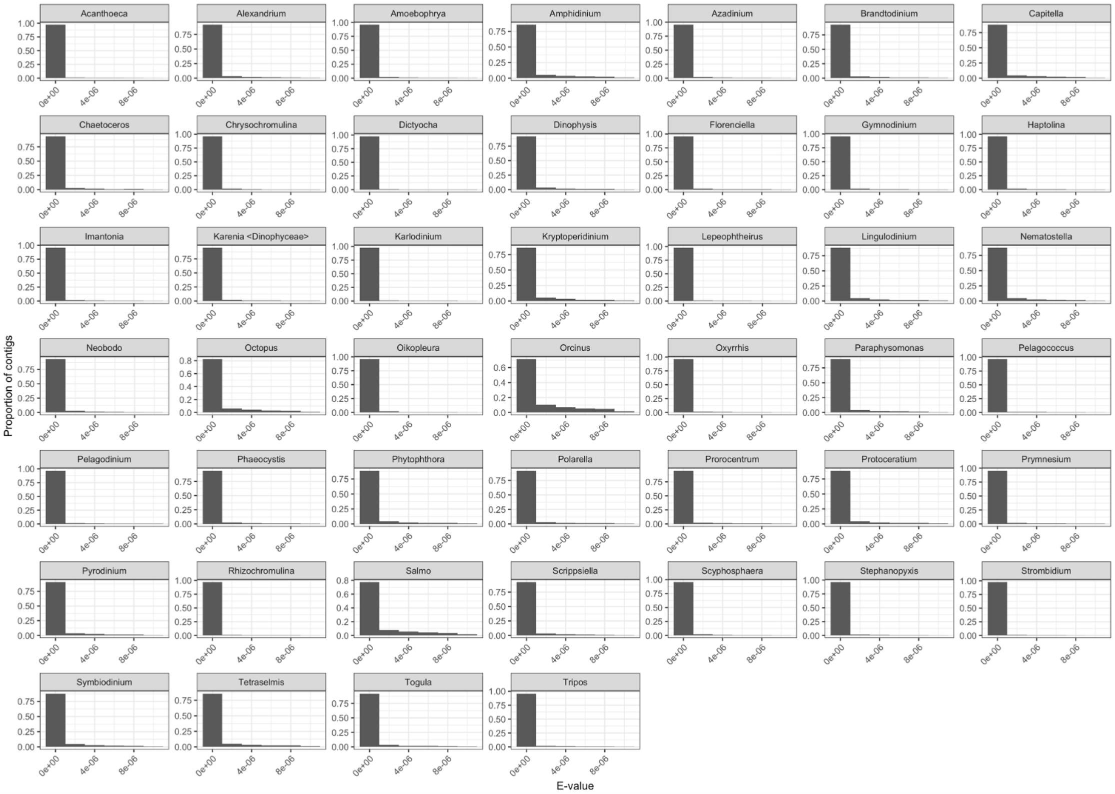
Supplementary Figure S1.1. Sequence length distribution of ~25 million quality-controlled Trinity-assembled contigs. Contigs less than 1,000 nucleotides in length are shown; approximately 482,000 contigs (1.9% of total) are longer than 1,000 nucleotides and range from 1,001 to 18,833 nucleotides.



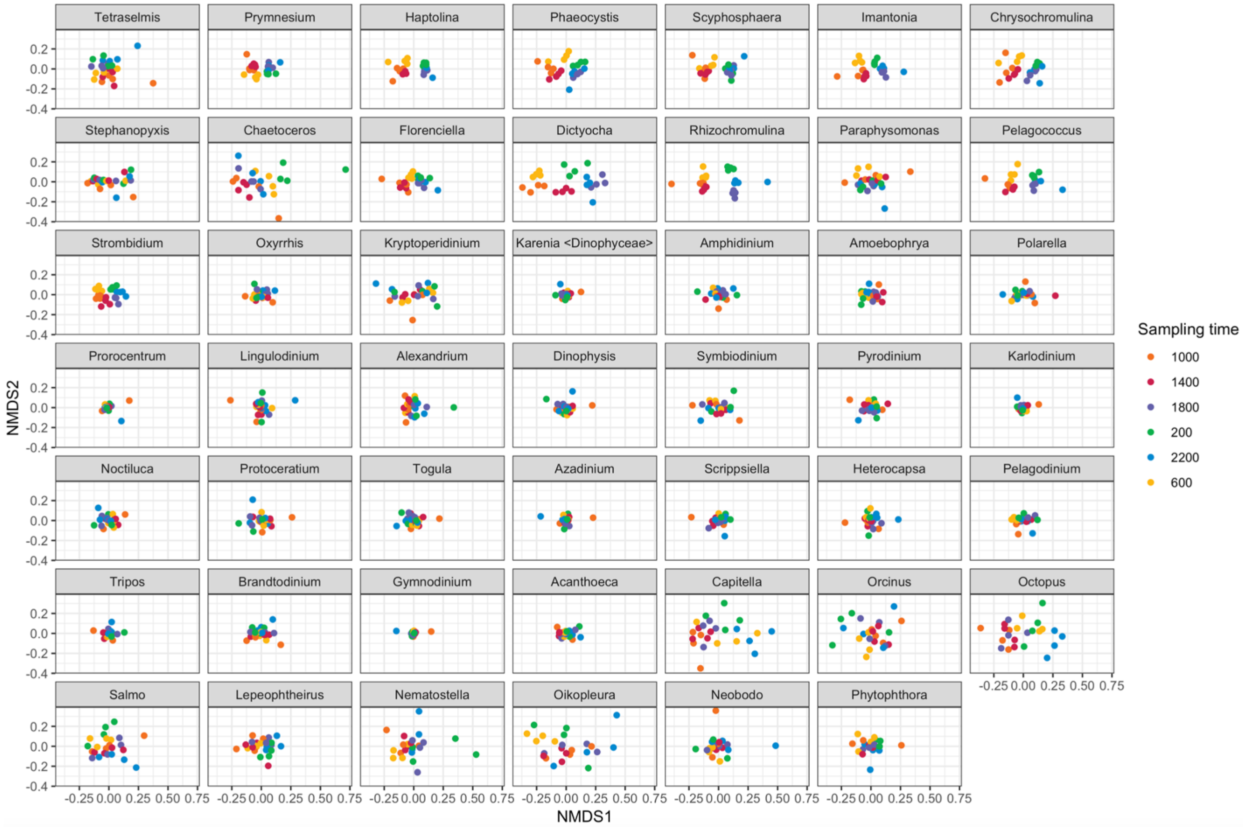
Supplementary Figure S1.2. Taxonomic assignments by primary Linnean ranks. Assignments totaling less than 2.5% of total placements for each rank are aggregated into “Other”. Not all taxa have higher-level rank assignments in the NCBI taxonomy; Phylum and Class rank assignments do not exist for many lineages and were not manually assigned.



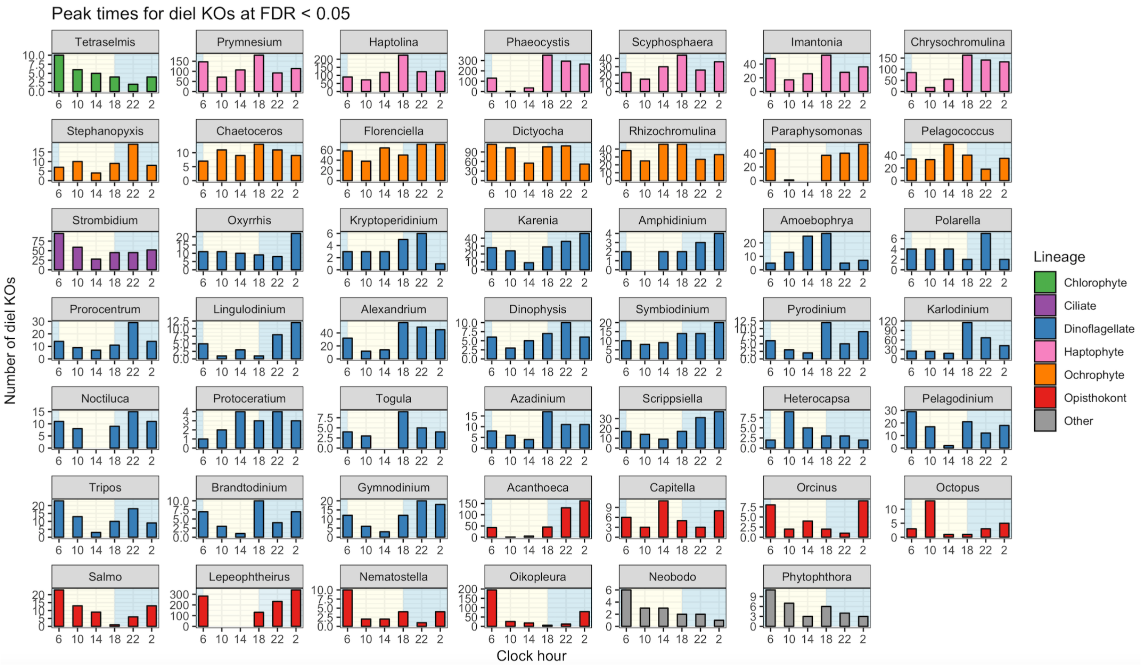
Supplementary Figure S1.3. Frequency of unique KEGG Ontologies (KOs) in eukaryotic MarineRefII reference taxa. X-axis: number of unique KOs found in each taxa after mapping to KEGG's KOfam library of HMMer profiles. Y-axis: frequency of occurrence. Dashed vertical line: Minimum threshold of 900 KOs required for determination of 'core KOs' present in >95% of reference taxa.



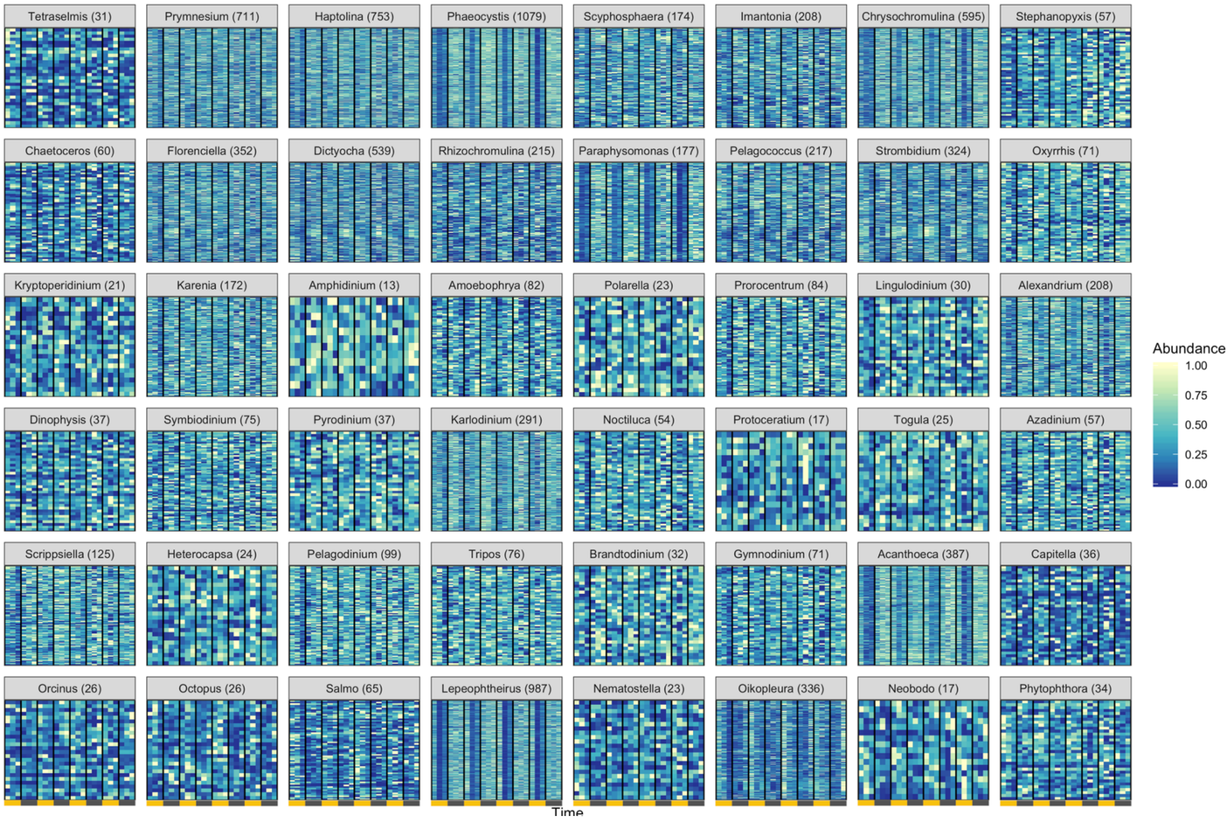
Supplementary Figure S1.4: E-value distributions for DIAMOND taxonomy assignments for environmental genera bins. Counts include assignments directly at the genus level and lower nodes (e.g., species under a genus).



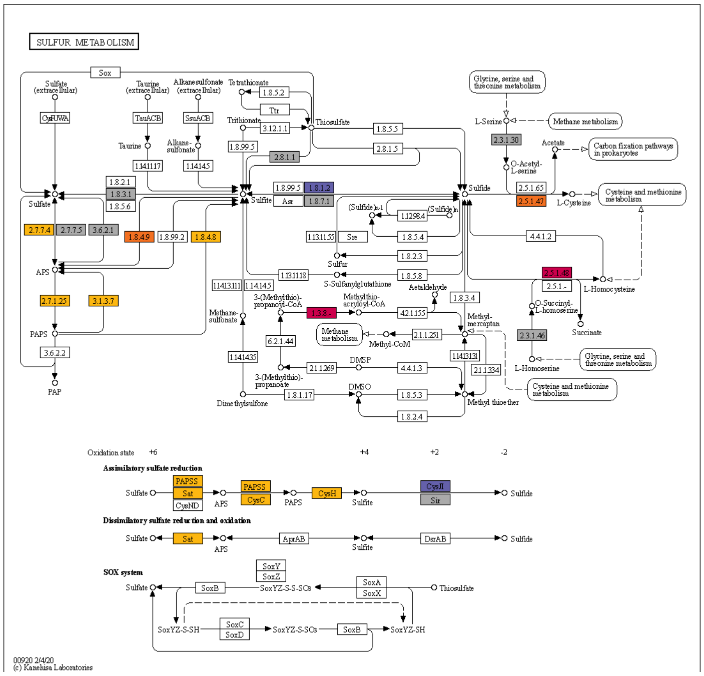
Supplementary Figure S1.5: NMDS ordination of Bray-Curtis from row-normalized KOfam counts. NMDS ordination performed independently on gene families belonging to each of 48 environmental genera that met completeness criteria. Mean stress of 48 NMDS = 0.127 ± 0.028 stdev.



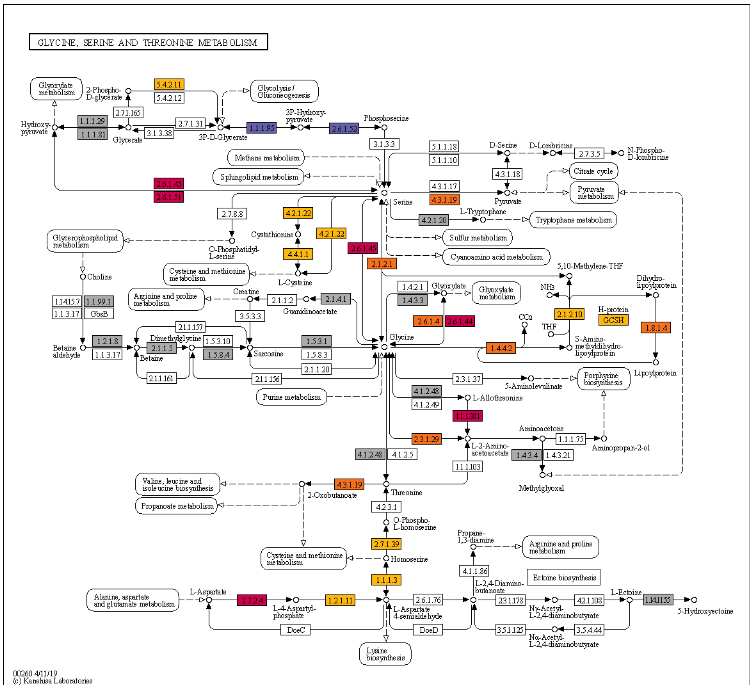
Supplementary Figure S1.6: Gene family transcript peak times for 48 protist genera across the diel cycle. Gene families are significantly periodic with an FDR < 0.05. X-axis: Clock hour in 24-hour day. Y-axis: percentage of significantly diel KOs with a peak at this time.



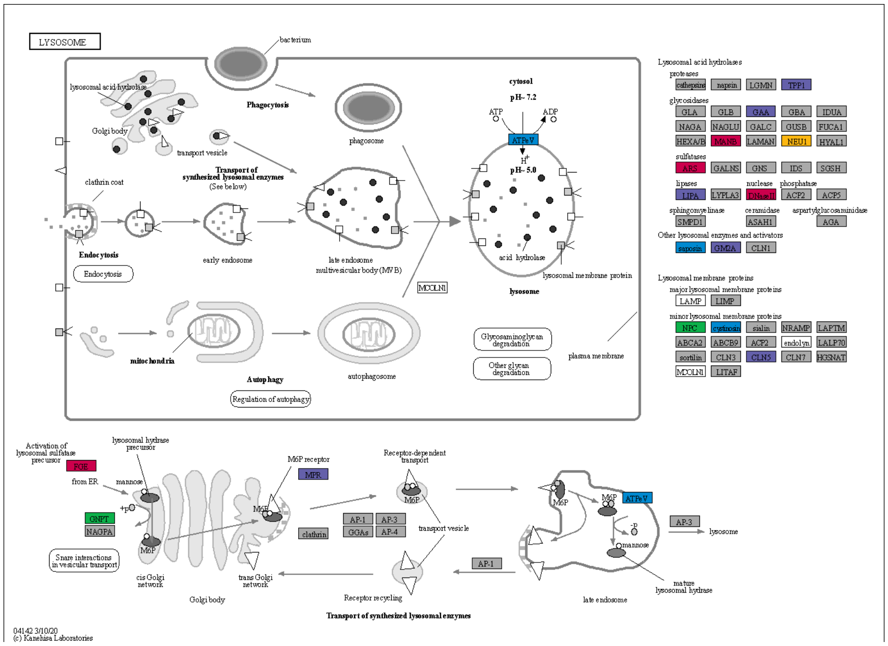
Supplementary Figure S1.7. Normalized abundance heat map of significantly periodic gene families from 48 protist genera. Yellow and gray bars denote light (06:00, 10:00, 14:00 HST) and dark (18:00, 22:00, 02:00 HST) periods, respectively. Each row corresponds to a gene family, ordered by hierarchical clustering of abundance patterns. Color corresponds to row-normalized abundance values for each gene family. Colored dot by genus name corresponds to Lineage from 3A. The order and presence of gene families is not maintained between facets.



Supplementary Figure S1.8. Peak times for Sulfur metabolism transcripts in Dictyophytes. Most enzymes in this pathway peak at dawn, including the sulfate to sulfite component of assimilatory sulfate reduction. Metabolic maps were generated using KEGG Color Mapper (https://www.kegg.jp/kegg/mapper/color.html). Enzymes with significant periodicity are colored according to their peak time as determined through RAIN analysis of transcript abundance. Warm colors indicate peaks in dawn/daylight hours: yellow, 06:00 HST; orange, 10:00; red, 14:00. Cool colors indicate peaks in dusk/night hours: purple, 18:00 HST; blue, 22:00; green, 02:00.



Supplementary Figure S1.9. Peak times for Glycine, serine and threonine metabolism transcripts in Dictyophytes. Dictyophytes maintain peak times for most of the enzymes in this pathway throughout daylight hours. Metabolic maps were generated using KEGG Color Mapper (https://www.kegg.jp/kegg/mapper/color.html). Enzymes with significant periodicity are colored according to their peak time as determined through RAIN analysis of transcript abundance. Warm colors indicate peaks in dawn/daylight hours: yellow, 06:00 HST; orange, 10:00; red, 14:00. Cool colors indicate peaks in dusk/night hours: purple, 18:00 HST; blue, 22:00; green, 02:00.



04142 3/10/20
(c) Kanehisa Laboratories

Supplementary Figure S1.10. Peak times for Lysosome transcripts in Dinoflagellates. Most periodic lysosomal enzymes have peaks in the afternoon (14:00), dusk (18:00) and early evening (22:00). Metabolic maps were generated using KEGG Color Mapper (https://www.kegg.jp/kegg/mapper/color.html). Enzymes with significant periodicity are colored according to their peak time as determined through RAIN analysis of transcript abundance. Warm colors indicate peaks in dawn/daylight hours: yellow, 06:00 HST; orange, 10:00; red, 14:00. Cool colors indicate peaks in dusk/night hours: purple, 18:00 HST; blue, 22:00; green, 02:00.

CHAPTER 2

MarFERReT: an open-source, version-controlled reference library of marine microbial eukaryote functional genes

Groussman, R.D., Blaskowski, S., Coesel, S., and Armbrust, E.V.

2.1 Abstract

Metatranscriptomics generates large volumes of sequence data about transcribed genes in natural environments. Taxonomic annotation of these datasets depends on availability of curated reference sequences. For marine microbial eukaryotes, current reference libraries are limited by gaps in sequenced organism diversity and barriers to updating libraries with new sequence data resulting in taxonomic annotation of only about half of eukaryotic environmental transcripts. Here, we introduce Marine Functional EukaRyotic Reference Taxa (MarFERReT), an updated marine microbial eukaryotic sequence library with version-controlled contents designed for taxonomic annotation of eukaryotic metatranscriptomes. MarFERReT contains over 32 million protein sequences from 890 marine eukaryotic genomes and transcriptomes, covering 504 species and 323 genera. We highlight regions of the marine eukaryotic tree of life currently lacking reference coverage and identify core sets of transcribed genes in major protistan lineages. Addition of open-ocean references refined taxonomic placements and increased annotation of previously unknown reads. Continued expansion of MarFERReT as new reference sequences become available will enable up-to-date taxonomic annotations into the future.

2.2 Introduction

Microbial eukaryotes perform essential ecological functions in marine ecosystems as phototrophs, predators, and parasites (Caron et al., 2017). This evolutionarily diverse group of organisms collectively possesses hundreds of millions of taxonomically distinct genes encoding metabolic processes that shape global biogeochemical cycles (Carradec et al., 2019). Eukaryotic metatranscriptomes are a nucleotide representation of community-wide transcriptional patterns and provide a window into how different members of the community function *in situ*. Identifying the taxonomic origin of these sequences depends on the quality and depth of the library of reference sequences used for taxonomic annotation. Eukaryotic metatranscriptomes became a component of environmental studies beginning around 2010, and yet annotation of these early metatranscriptomes was hampered by a sparsity of reference sequences among marine protists, with many lineages lacking any sequenced representatives. The sequence landscape improved dramatically in 2014 with development of the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP, Keeling et al., 2014), a community-wide undertaking that resulted in public availability of over 650 assembled transcriptomes derived from over 340 marine eukaryotic isolates. The MMETSP increased the number of sequenced marine protist species by approximately one order of magnitude (Fig. S2.1), greatly enhancing the ability to annotate environmental sequences.

In the time since release of the MMETSP, a variety of new reference libraries have arisen that build upon MMETSP by aggregating additional marine eukaryote genomes and transcriptomes from different sources. The MarineRefII sequence library (http://roseobase.org/data/) combines additional curated eukaryotic and bacterial genomes and transcriptomes to the MMETSP transcriptomes, with the last public update in 2014. A total of 1.2 million additional metazoan, fungi and viral sequences were added to MarineRefII to increase the diversity of coverage, with the new contents described in Coesel et al., 2021. The PhyloDB reference library (https://github.com/allenlab/PhyloDB) was last updated in 2015 and contains microbial eukaryotes, bacteria, archaea and viruses isolated from both marine and non-marine environments. The EukZoo protein database of aquatic microbial eukaryotes (https://zenodo.org/record/1476236) was released and updated in 2018 and added 48 genomes and transcriptomes to the MMETSP dataset. In 2019, the METdb repository (Niang et al., 2020) included the MMETSP re-assemblies (Johnson et al., 2018) in addition to 34 marine protist transcriptomes generated from the Roscoff Culture Collection (RCC). MARMICRODB is a prokaryote-focused database that includes reference transcriptomes, genomes, SAGs and metagenome-assembled genomes (MAGs) from marine microbes, as well as some eukaryotic reference sequences from the MMETSP (Becker et al., 2019). In addition to these different public repositories, 'custom reference sequence libraries' are constructed independently by different research groups to meet specific research needs. The composition, sources, and documentation for these custom sequence libraries are often not readily available, which may impact reproducibility of the analyses.

Sequence data from reference organisms maintained in different culture collections is now supplemented by sequences generated from uncultured single cells, so-called Single-cell Amplified Genomes (SAGs) (Roy et al., 2014). Taxonomic identification of a SAG is assigned, when possible, from analysis of 18S rDNA generated during the amplification step, although as with other annotations, 18S rDNA-based taxonomy is also dependent on reference sequences. SAGs from MAST-3 and MAST-4 clades of marine stramenopiles and the Chrysophyte H1 and H2 clades are available to the public (Seeleuthner et al., 2018). The availability of SAGs from uncultured organisms in combination with the continued isolation and sequencing of marine eukaryote strains from diverse environments such as Antarctic coastal waters (Guajardo et al., 2021), deep waters (Cooney et al., 2020), and the oligotrophic open ocean (Lambert et al., 2021) improves the ability to provide taxonomic affiliation to previously unknown sequences. Although expansive, remaining gaps in sequence representation persist within current reference libraries, and approximately half of assembled transcripts from metatranscriptomes currently have no match to those from any known organism (Carradec et al., 2018, Groussman et al., 2021).

Here, we introduce Marine Functional EukaRyotic Reference Taxa (MarFERReT), an updated open-source marine eukaryote reference protein sequence library with a reproducible framework allowing for community-supported expansion over time. MarFERReT was designed to be a comprehensive marine microbial eukaryote reference library for the taxonomic annotation of environmental metatranscriptome data. Our collation of publicly available reference sequences highlights sections of the marine eukaryotic tree of life that lack reference coverage and can guide future targeted sequencing projects. We use MarFERReT to identify the core transcribed genes of key marine eukaryote lineages that can serve as a metric for estimating the completeness of environmental taxonomic transcript bins. Two use case studies, with accompanying code, are provided to illustrate how MarFERReT can be used either by itself or in conjunction with other protein sequence libraries to assign taxonomic identity to environmental sequences, and to approximate sequencing coverage within environmental taxonomic bins. As more references are integrated over time, future releases of MarFERReT will provide version-controlled library updates with documented changes. We anticipate this framework to collect and integrate new sequences into a growing library that will increase the accessibility of marine eukaryote references and further our understanding of the diverse functional potential of marine protists.

2.3 Materials & Methods

2.3.1 Collation of sequence data

Genomic, transcriptomic, and single-cell-amplified (SAG) data were collected from publicly available web resources. The repository/database origin and web link to the original data source along with additional metadata for MarFERReT is available on the Zenodo repository: (link). Four projects were major sources of sequence information for MarFERReT. The Marine Microbial Eukaryote Sequence Project (MMETSP, Keeling et al., 2014) is the largest contributor to MarFERReT with 666 entries. The MMETSP sequences were collected as peptide translations from Version 2 of the MMETSP re-assemblies (Johnson et al., 2018, https://doi.org/10.5281/zenodo.740440). A total of 117 entries were collected from JGI Phycocosm (Grigoriev et al., 2021) as translated protein predictions of gene models from assembled genomes. A total of 41 transcriptomes were collected from the Roscoff Culture Collection (RCC) as nucleotide transcriptome assemblies from the METdb (http://metdb.sb-roscoff.fr/metdb/). To provide classification breadth to metazoan sequences consistently captured in protist-focused metatranscriptomes (Carradec et al., 2018), sequences from 36 Metazoan (animal) species were included from NCBI GenBank, including nucleotide transcriptome assemblies from 14 copepod transcriptomes (Maas 2018) and translated gene models from 22 marine species selected for broad coverage of 12 metazoan phyla (Bucklin et al., 2011). The remaining 23 entries were collected from other sequencing projects including translated gene models from 8 single-cell amplified genomes of uncultured stramenopiles (Seeleuthner et al., 2018), 10 assembled transcriptomes of diatom isolates from Antarctic and North Atlantic coastal waters (Guajardo et al., 2021), two single-cell amplified assembled transcriptomes from early-branching dinoflagellates (Cooney et al., 2020), and translated transcriptome assemblies from open-ocean isolates of haptophytes and diatoms of the North Pacific (Lambert et al., 2021, link).

2.3.2 Curation of sequence metadata

Gaps in entry source metadata were manually curated to ensure that each sequence entry has an organismal name and an associated NCBI taxonomy ID (tax_id). With few exceptions, the sequence data ingested into the library had an associated NCBI taxonomy IDs (tax_id), genus and species names. For entries that had an NCBI tax_id but no genus and species (e.g., some of the unclassified MMETSP samples), organismal names were updated when possible, from the NCBI taxonomy database (Federhen 2012, https://www.ncbi.nlm.nih.gov/taxonomy). Outdated NCBI tax_ids were replaced with their current version (as of October 11 2022). Entries without a provided NCBI tax_id were manually assigned an NCBI taxonomy at the most specific possible rank. NCBI Taxonomy hierarchical relationships were gathered for the 517 unique taxIDs in MarFERReT using the 'taxit' package (v0.9.2, https://github.com/fhcrc/taxtastic).

2.3.3 Incorporation of NCBI Taxonomy relationships and generation of cladograms

The taxonomic relationships of the 517 MarFERReT taxIDs to internal taxonomic ranks was gathered from the NCBI Taxonomy database (Federhen 2012) using the version of NCBI Taxonomy published on October 1st 2022 (link). This was used to generate a cladogram of MarFERReT species from the NCBI Taxonomy CommonTree tool (link). For the cladogram of MarFERReT coverage across marine protists~~,~~ taxonomic relationships were retrieved from NCBI for all Eukarya and the Family-rank taxIDs were filtered down to include only Families with catalogued marine species in the World Register of Marine Species (Vanhoorne et al., 2008, link). To maintain focus on unicellular protists, animals (metazoa), plants (streptophyta), fungi and rhodophytes were not included.

2.3.4 Six-frame translation and frame selection of nucleotide sequences

Nucleotide sequences were translated in six frames with transeq vEMBOSS:6.6.0.059 (Rice *et al*., 2000) using Standard Genetic Code, to bring all MarFERReT reference material into translated amino acid sequence. The longest coding frame (longest uninterrupted stretch of amino acid residues) was retained for downstream analysis.

2.3.5 Functional annotation of protein sequences

The MarFERReT protein sequences were annotated against the Pfam 34.0 collection of 19,179 protein family Hidden Markov Models (HMMs) (Mistry et al., 2021) using HMMER 3.3 (Eddy 2011). The highest-stringency cutoff score ('trusted cutoff') assigned by Pfam to each hmm profile was used as a minimum score threshold. The best scoring Pfam annotation (highest bitscore) was used if the protein received more than one Pfam match. A total of 12,918,785 MarFERReT sequences received annotation with 12,620 of the 19,179 total possible profiles in Pfam 34.0.

2.3.6 Identification and analysis of Core Transcribed Genes

Core transcribed genes (CTGs) were identified based on the Pfam 34.0 annotation of proteins translated from the 742 transcriptome and SAT entries (the 148 genomic and SAG-sourced entries were not included). We identified 11,247 Pfam profiles in the transcriptome-derived MarFERReT protein reference sequences, or 59% of the 19,179 total possible profiles in Pfam 34.0. A presence-absence matrix of Pfam functions was generated from functional annotation for the 368 species that had at least 1000 Pfams (pre-filtered taxonomic species bins). Core transcribed genes were operationally defined as the set of Pfam IDs assigned to annotated transcripts within ≥95% of pre-filtered species bins within a higher-order lineage. CTGs were defined for all eukaryotes as a whole group, and from pre-filtered species in 9 lineages: Bacillariophyta, Chlorophyta, Dinophyceae, Ochrophyta, Haptophyta, Ciliophora, Opisthokonta, Rhizaria, and Amoebozoa.

2.3.7 Species-level protein clustering

To reduce sequence redundancy from multiple sequence entries for a single organism, the protein sequences for NCBI taxIDs with more than one entry were combined and clustered at the 99% amino acid sequence identity threshold with MMseqs2 (Steinegger and Söding 2018). TaxIDs with a single entry were not clustered.

2.3.8 Case Study 1: Taxonomic annotation with DIAMOND protein-alignment

MarFERReT protein sequence data and the accompanying taxonomic information were used to generate a database for use with DIAMOND (Buchfink et al., 2015). DIAMOND databases were generated for both MarFERReT alone and for a combined database of MarFERReT and the prokaryotic MARMICRODB (Becker et al, 2019). Prior to database construction, the MARMICRODB sequence file was filtered to remove metagenome-assembled genomes, eukaryotes, and taxa not represented with an NCBI tax_id, retaining 7,921 prokaryotic genomes, transcriptomes and single-cell amplified genomes. These data were merged with MarFERReT protein sequences and used to construct a combined DIAMOND database.

We used the DIAMOND protein alignment tool to assign a putative lowest-common-ancestor (LCA) to assembled transcripts from an environmental metatranscriptome (more information about these metatranscriptomes and their generation in Appendix 1). The results from MarFERReT-only and combined MarFERReT-MARMICRODB databases were compared. To compare MarFERReT against an older marine-eukaryote focused sequence library, DIAMOND annotation results were compared to annotations against the same set of environmental transcripts from DIAMOND annotation using the enhanced MarineRefII sequence library (Coesel et al., 2021). Case Study 1 code can be found here.

2.3.9 Case Study 2: Estimating completeness of metatranscriptome taxonomic bins

We estimated the completeness of eukaryotic environmental transcriptome bins derived from MarFERReT annotation in Case Study 1, using the Core Transcribed Genes (CTGs) derived from Pfam annotation of MarFERReT protein sequences. Case Study 2 code can be found here.

2.4 Results

2.4.1 Composition and construction workflow of MarFERReT

MarFERReT (v1.0) contains 32,825,421 translated and clustered protein sequences, derived from 504 species and 323 unique genera gathered from 890 public-access large-scale and smaller-scale marine sequencing projects released between 2002 and 2021, primarily focusing on marine microbial eukaryotes (Figures 2.1, S2.1, Table 2.1). The reference sequences were gathered from 741 assembled transcriptomes (including 2 single-cell amplified transcriptomes), and 148 genomes (including 8 single-cell amplified genomes) (Figure 2.1). A common framework was employed for inclusion of all sequences into MarFERReT (Figure 2.2). Sequences were collected from primary sources as either nucleotides or translated amino acids (protein). Nucleotide sequences were translated into predicted protein sequences, and the longest coding frame was retained. Protein sequences were functionally annotated with Pfam version 34.0 (Mistry et al., 2021) to provide uniform functional prediction. Sequence redundancy in NCBI taxIDs with multiple entries was reduced by clustering at the 99% protein sequence identity threshold; single-entry taxIDs were not clustered.

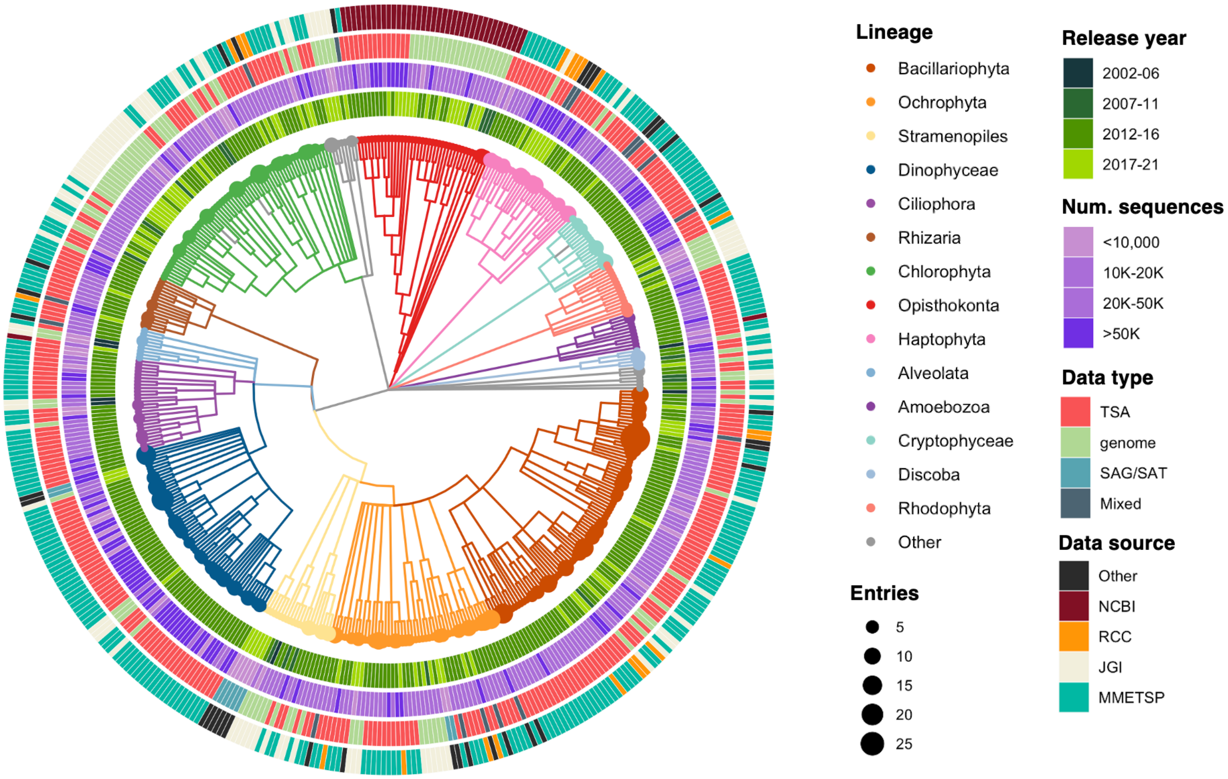


Figure 2.1. Taxonomy of species in MarFERReT. Cladogram of hierarchical taxonomic ranks of marine eukaryotes based on the NCBI Taxonomy framework (Federhen 2012, link) using their CommonTree tool (link). Branches are colored by lineage with size of closed circle at each tip proportional to the number of independent sequence datasets in each species. Concentric rings describe metadata and statistics for each species. From innermost out: (a) year of publication or data release for sequence data (average year of release for multiple entries), (b) number of final protein sequences in MarFERReT species, (c) initial format of sequence data: transcriptome shotgun assembly, TSA; single-cell amplified genome, SAG; single-cell amplified transcriptome, SAT; genome-derived gene models; or a combination of types (mixed), and (d) data source: National Center for Biotechnology Information, NCBI; Roscoff Culture Collection, RCC; DOE Joint Genome Institute, JGI; Marine Microbial Eukaryote Transcriptome Sequencing Project, MMETSP.

Table 2.1. Summary of MarFERReT taxonomic and sequence statistics.

Number of entries:	890
Number of NCBI taxIDs:	517
Number of species:	504
Number of genera:	323
Total sequences from entries:	39,127,537
as nucleotide:	5,712,383
as amino acid:	33,415,154
Total MarFERReT proteins:	32,825,421
clustered multi-entry:	16,485,085
unclustered single-entry:	16,340,336
Total Pfam annotations	12,918,785

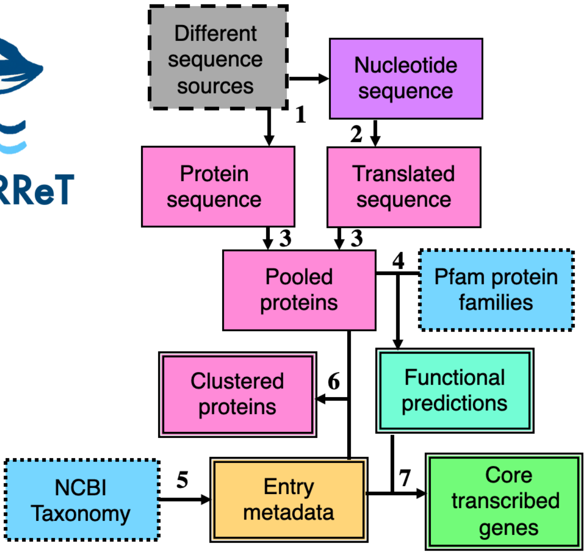


Figure 2.2. Diagrammatic overview of MarFERReT build processes. Boxes indicate data sets; box borders indicate external sequence inputs (dashed line), external taxonomic and functional annotation resources (dotted lines), internal data products (solid line) and output MarFERReT data products (double lines). Arrows indicate processes: 1) collection of nucleotide and protein reference sequences from primary data sources; 2) six-frame translation and frame-selection of nucleotide sequences into amino acid sequence; 3) compilation of protein sequences from all sources; 4) functional annotation of protein sequences; 5) curation of NCBI taxID for all MarFERReT entries; 6) clustering of protein sequences within taxIDs with multiple entries at 99% identity; 7) identification of core transcribed genes (CTGs) from Pfam functional annotations.

‘

2.4.2 Representation of marine protist diversity in MarFERReT

We assessed the distribution of MarFERReT taxa across the 850 families of marine protists that are both cataloged in NCBI Taxonomy (Figure 2.3) and listed in the World Register of Marine Species (https://www.marinespecies.org/). MarFERReT contains taxa from 160 (19%) marine protist families that are represented by at least one species with transcriptomic or genomic functional sequences. Representation across families is biased by the number of NCBI-identified families across eukaryotic lineages, likely reflecting historical legacies of species descriptions. For example, Ciliophora is composed of 154 marine families whereas Cryptophyceae is composed of 8 families. Moreover, available sequences within each lineage differ dramatically due to biases in the availability of marine reference isolates maintained in culture collections or the relatively limited diversity of available single-cell genomes and transcriptomes (Supp Figs S2.2-S2.13). Together, these factors result in dramatic differences in family-level representation across lineages within MarFERReT. In descending order, coverage consists of 75% of cryptophyte families (6 of 8), 56% of haptophytes families (9 out of 16), 39% of diatom families (31 of 80), 13% of Ciliophora (20 of 154), 8% of Rhodophyta (9 of 110), and 4% of Rhizaria families (6 of 137 families). Significant gaps remain in classified marine lineages at finer taxonomic ranks, notably within the Ciliophora, Rhizaria and Ochrophyta. Even within relatively well-characterized lineages such as diatoms, dinoflagellates and haptophytes, the distribution of sequenced representatives across the family levels is uneven, with several unrepresented clades remaining (Figure 2.3, Supp Figs S2.2-S2.14).

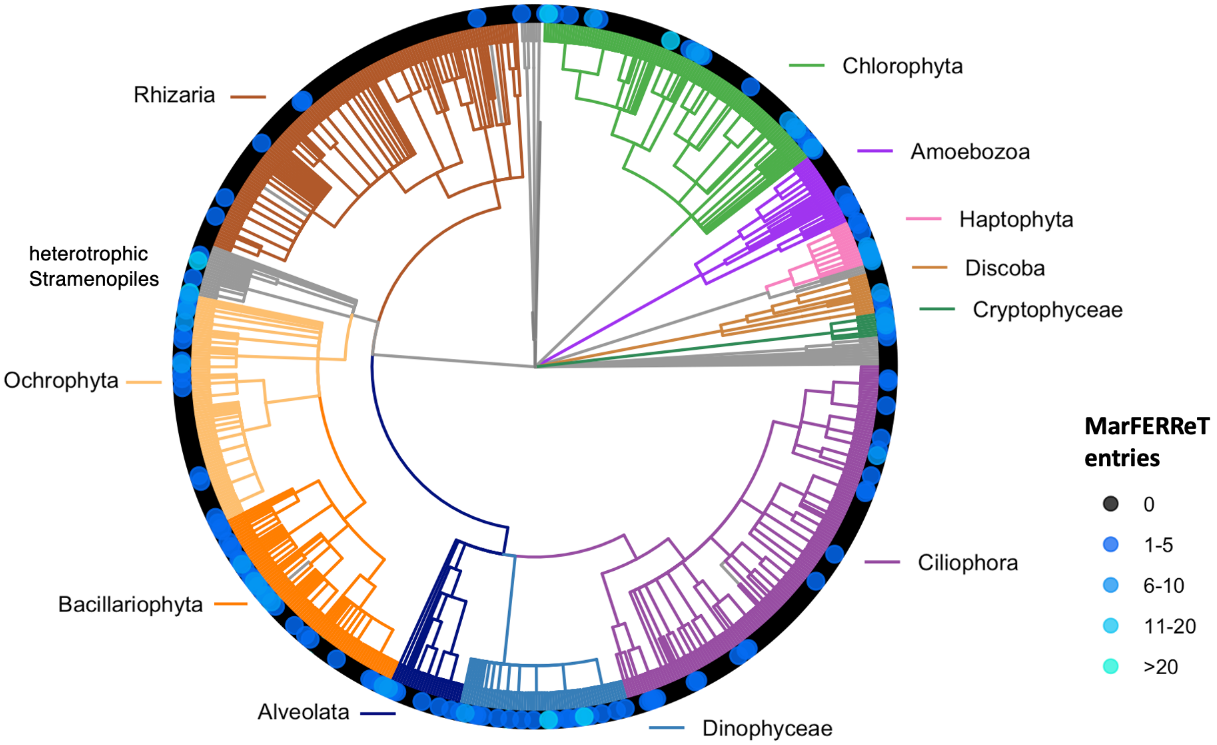


Figure 2.3. Family-level coverage of marine eukaryotes in MarFERReT. Cladogram of hierarchical taxonomic ranks of marine eukaryotes based on the NCBI Taxonomy framework (Federhen 2012, link) using their CommonTree tool (link). Tips were pruned down to the Family level and filtered by taxonomic Families that include marine species as listed in the World Register of Marine Species (Vanhoorne et al., 2008, link). Animals (Metazoa), plants (Streptophyta), fungi and rhodophytes are not shown here. Branches are colored by lineage (ranging from superkingdom to phylum level; lineage shown next to clade), and the color of each tip indicates the number of species in the family represented in MarFERReT v1.0.

2.4.3 Identification of Core Transcribed Genes for assessing environmental transcriptome bin coverage

One of the key purposes of MarFERReT is to annotate metatranscriptome sequence data. Environmental metatranscriptomes can be deconvolved into individual taxonomic transcriptomes by combining sequences (assembled transcript contigs or short sequence reads) with shared taxonomic identity into taxonomic bins. The size of the taxonomic bins (in transcripts or reads) for each species population is a factor of the population organismal abundance, the number of transcribed genes in the population, and the sample sequencing depth, among other possible factors. The total number of reads mapped to a taxonomic bin provides a rough approximation of sequencing depth, but transcriptome coverage offers a richer assessment. To estimate metatranscriptome bin completeness of environmental eukaryotic species bins, we used eukaryote species in MarFERReT to develop a transcript-based metric analogous to the BUSCO-based (Simão et al., 2015, Manni et al., 2021) estimate of genome assembly completeness.

Here we derive sets of core transcribed genes (CTGs; link) for major marine protistan lineages, leveraging the 748 transcriptomes derived from a total of 386 species within MarFERReT (Figure 2.4). Reference protein sequences were annotated with predicted protein families from Pfam 34.0 (Mistry et al., 2021). Catalogs of transcribed genes based on Pfam protein family annotations were generated for each species (Supp Fig S2.14); a median of 3080 Pfams was detected across species. The break in the lower end of the normal gene distribution was used to identify 368 species with a minimum threshold of 1000 Pfam functions, removing possibly incomplete transcriptomes. The core transcribed genes for the 368 eukaryotic species as a whole are defined here as those transcripts associated with Pfam protein annotations identified in at least 95% of the species with at least 1000 assigned Pfams (Supp Fig S2.15). We also identified CTGs using the same criteria for nine key lineages with transcriptomes from at least 10 separate species (Supp Figs S2.16-S2.24): Haptophyta (haptophytes), Dinophyceae (dinoflagellates), Bacillariophyta (diatoms), Ochrophyta (ochrophytes not including diatoms), Chlorophyta (chlorophytes), Opisthokonta (includes fungi and animals), Ciliophora (ciliates), Rhizaria and Amoebozoa. Each lineage was defined by ~600-1400 CTGs with a mean of 1042 CTGs (Table 2.2), and 884 CTGs detected in least 95% of all MarFERReT species with transcriptomes (Supp Fig S2.15).

Table 2.2. Number of species and core transcribed genes (CTGs) for major marine eukaryote lineages. The ‘Eukaryota’ group contains all MarFERReT species with transcriptomes.

Lineage	Species	CTGs
Bacillariophyta	86	1231
Dinophyceae	47	922
Eukaryota	46	884
Ochrophyta	46	1112
Chlorophyta	43	1421
Haptophyta	32	1435
Ciliophora	21	603
Opisthokonta	15	906
Rhizaria	15	1186
Amoebozoa	11	720
Alveolata	6	683

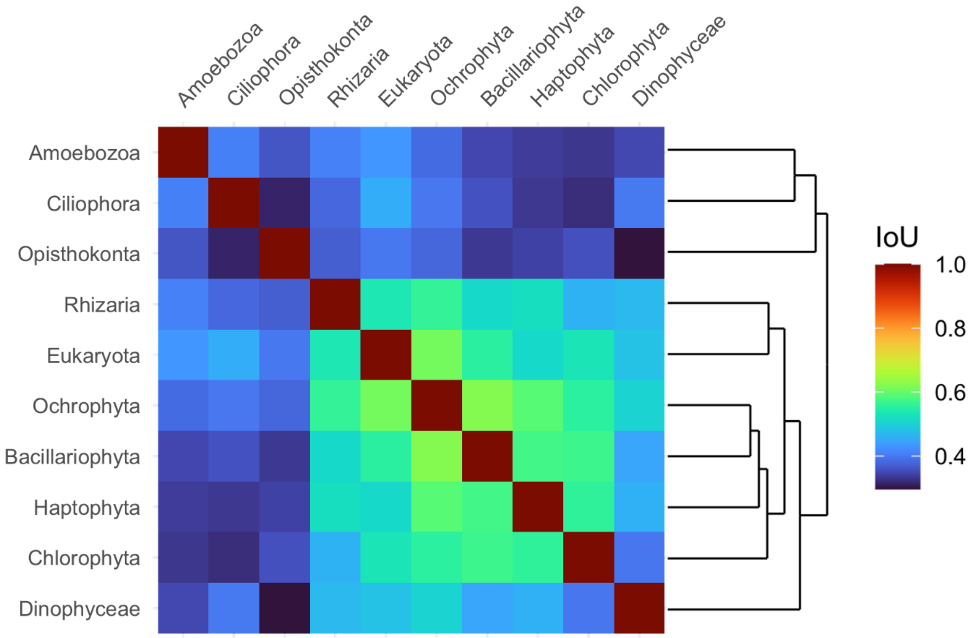


Figure 2.4. Overlap in core transcribed gene sets between major eukaryotic lineages. The heatmap colors represents the overlap in CTG sets between two lineages through the ratio of the size of the intersecting and union sets (Intersect over Union, IoU). 'Eukaryotes' contains MarFERReT species that do not belong to the other lineages. Dendrogram on right generated from hierarchical clustering of a distance matrix of IoU scores.

The proportion of overlapping CTG sets between these eukaryotic lineages recapitulates their shared evolutionary history and general trophic mode (Figure 2.4). The four predominantly phototrophic lineages, bacillariophyta, ochrophyta, haptophyte, and chlorophyta are the most similar to each other and share over half of their core transcribed gene families. In contrast, the essentially heterotrophic clades of amoebozoa, ciliophora, and opisthokonta have roughly one-third of their CTGs in common with any other group. The 'eukaryota' set of CTGs, derived from the full set of eukaryotic transcriptomes in MarFERReT, represent the core suite of transcribed protein gene families found in nearly all species and shares the most overlap with the ochrophyte and diatom groups; this is a probable artifact of the present bias towards phototrophic organisms in reference taxa. In general, the greatest number of CTGs were found in predominantly phototrophic lineages (diatoms, ochrophytes and haptophytes) whereas the fewest CTGs were found in heterotrophic lineages, with a weak positive correlation between the number of species and number of CTGs in a lineage from linear regression ($r^2 = 0.24$). The full CTG catalog is available online here: (https://zenodo.org/record/7055912). An example of use of the CTG metric to estimate the completeness of environmental taxa bins from metatranscriptomes is given in Case Study 2, below.

2.4.4 Case Study 1: Taxonomic annotation using DIAMOND protein-alignment

This case study illustrates use of the DIAMOND fast protein-alignment tool (Buchfink et al., 2015) in combination with MarFERReT to infer the phylogenetic identity of unknown environmental transcripts based on a set of reference proteins and demonstrates the impact of novel reference species on annotation results (Figure 2.5). We use DIAMOND directly with MarFERReT, and also show how to customize MarFERReT with novel reference sequences by merging in a bacterial reference database. Documentation for database preparation and DIAMOND protein alignment is available online: (Case Study 1 on github). The unknown environmental sequences used for annotation in this Case Study were 31.9 million transcript contigs from 19 poly-A+ selected size-fractionated metatranscriptome assemblies, with large (3 – 200 µm pore size) and small (0.2 – 3 µm pore size) size-fractions (referred to hereafter as 'environmental sequences) collected from surface seawater at eight sites on a latitudinal transect of in the North Pacific Ocean (Lambert et al., 2022, Appendix 1).

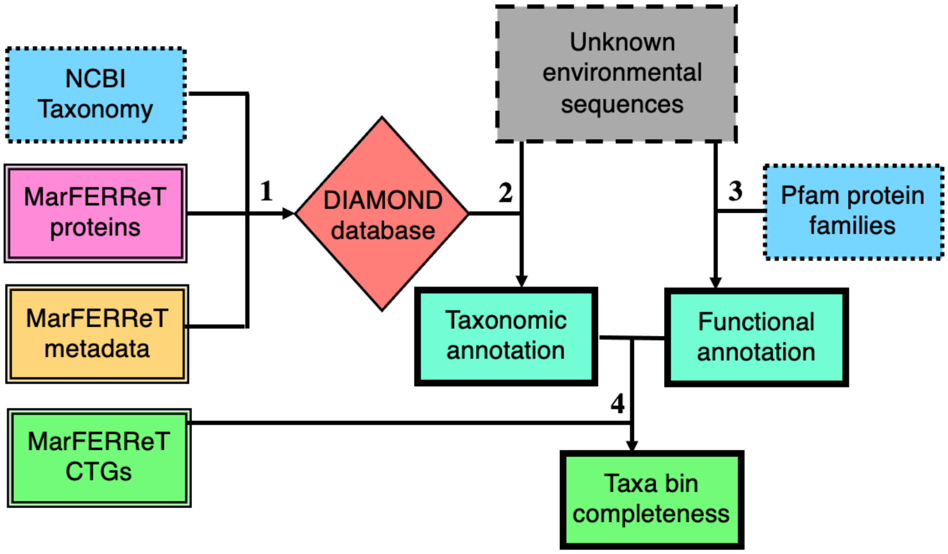


Figure 2.5. Schematic of use of MarFERReT for annotation of environmental metatranscriptomes. Diagram illustrates how MarFERReT data products are used to annotate unknown sequences and assess taxonomic bins, as described in Case Studies 1 and 2. Boxes indicate datasets; box borders indicate environmental sequence input (dashed line), external taxonomic and functional annotation resources (dotted lines), MarFERReT library resources (double lines) and taxonomic and functional annotation

results (bold lines). Red diamond indicates use of user-constructed DIAMOND database for lowest common ancestor determination. Arrows indicate processes: 1) construction of DIAMOND database using MarFERReT proteins and data, and taxonomy files from NCBI Taxonomy; 2) Taxonomic annotation of environmental sequences against the database from (1) using DIAMOND database; 3) Functional prediction of environmental sequences with hmmsearch with Pfam protein family hmm profiles; 4) Completeness assessment of taxonomically- and functionally-annotated metatranscriptome bins using MarFERReT core transcribed genes (CTGs).

We annotated the environmental sequences with DIAMOND using three reference libraries (Table 2.3): the first using MarFERReT, the second using a combined database of MarFERReT and MARMICRODB (MarFERReT+), and the third using an older combined eukaryotic and prokaryotic database (MarRef2+). The MarFERReT+ library combines MarFERReT's eukaryotes with bacterial and archaeal reference sequence references from the publicly available MARMICRODB reference. MARMICRODB is a prokaryote-focused database that includes reference transcriptomes, genomes, SAGs and metagenome-assembled genomes (MAGs) from marine microbes (Becker et al., 2019). We removed the MAGs and the limited set of eukaryotic reference sequences from MARMICRODB for this analysis, as the MAGs were of uncertain levels of taxonomic confidence and the eukaryote sequences were redundant with those in MarFERReT. This resulted in a join of 27,890,788 bacterial and archaeal proteins to MarFERReT's 32,825,421 eukaryotic proteins. The third library (MarRef2+), is a modified version of an older cross-kingdom marine reference library (MarRefII, http://roseobase.org/data/), amended with additional protein sequences from marine prokaryotic, eukaryotic and viral genomes and transcriptomes to expand reference diversity, but notably does not contain sequences that became available after 2015 (Coesel et al., 2021). We compared high-level annotation metrics of the environmental sequences against these three references (Table 2.3). The cross-kingdom MarFERReT+ reference resulted in annotation of more total transcripts (60.9%) than MarFERReT alone (57.6%) or the older cross-kingdom MarRef2+ (55.9%). The greater total number of Eukaryota-annotated reads from MarFERReT+ compared to MarFERReT could be a result of increased statistical confidence in DIAMOND assignments to Eukaryota taxa with the addition of non-Eukaryotic references. The difference between MarFERReT+ and MarRef2+ also shows the influence of the new eukaryotic reference material (Table 3); assignments to eukaryotic taxa increased by 8.8% (1.5 million transcript annotations). Assignments to species or sub-species eukaryotic taxa similarly increased by 8.2%, indicating an improvement in annotation specificity (Fig 2.6a).

Table 2.3: Comparison of annotation metrics from three reference databases. All databases were run against the same set of 31.9 million environmental transcripts. Labeled columns indicate the reference database used: MarFERReT+ (combined MarFERReT and MarMicroDB); MarFERReT eukaryotes only, or the MarRef2+ reference. The first row is the total number of query environmental transcripts (see 'Gradients 1' in Appendix 1 for more information on environmental transcripts). Subsequent rows indicate the total number of transcripts and the percent of all transcripts assigned to a eukaryotic taxID (Total Eukaryota), the total number and percent of transcripts assigned to a bacterial taxID (Total Bacteria), and the total and percent of transcripts assigned a eukaryotic taxID at the species or subspecies level.

	MarFERReT+	**MarFERReT**	**MarRef2+**
Total transcripts	31,905,677	31,905,677	31,905,677
Total with annotations	19,421,265	18,372,500	17,840,953
% of all transcripts	60.9%	57.6%	55.9%
Total Eukaryota	19,082,003	18,372,500	17,541,361
% of all transcripts	59.8%	57.6%	55.0%
Total Bacteria	58,458	0	35,413
% of all transcripts	0.18%	0.00%	0.11%
Total Eukaryota species	6,844,519	6,618,085	6,323,260
% of all transcripts	21.5%	20.7%	19.8%

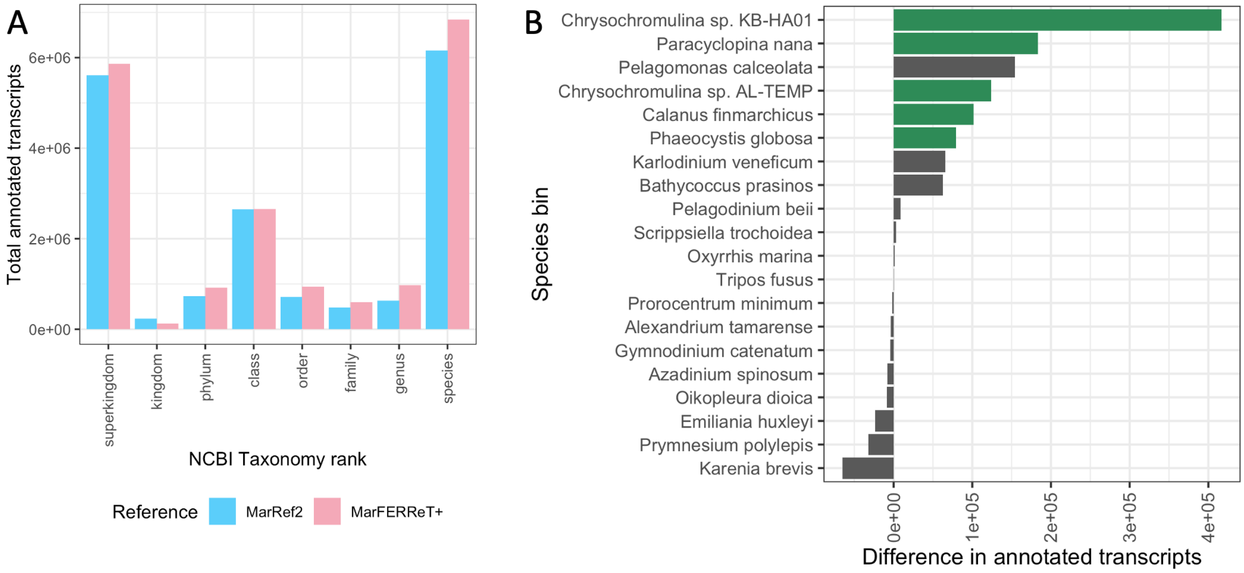


Figure 2.6. Annotation of environmental eukaryotic metatranscriptomes. A. Comparison of lowest-common-ancestor rank placements using the MarFERReT+ combined reference library (pink fill) against another mixed-domain reference based on MarineRefII (MarRef2, blue fill, Coesel et al., 2021). Placements to NCBI minor ranks have been collapsed (e.g., 'phylum' contains sums for 'subphylum'). **B.** Comparison of total transcript assignments to environmental species bins from MarFERReT+ and MarRef2. Horizontal axis shows the difference in the number of transcripts assigned to species bin from new (MarFERReT+) minus old (MarRef2) annotations. Green bars indicate new species bins in MarFERReT not present in MarRef2, with zero previous annotations. Only species with at least 1% of total eukaryotic species-level assignment using either reference are shown here.

A common first step when analyzing eukaryotic metatranscriptomes is to assess the efficiency of poly-A+ selection of eukaryotic messenger RNA by identifying the proportion of bacterial-annotated transcripts. As expected with the eukaryotic-only reference library, there were no assignments to bacterial taxa using MarFERReT alone (Table 2.3). With the combined library, the sensitivity of adding bacterial references is illustrated by detection of 0.18% bacterial-annotated transcripts (58,458 bacterial transcripts within 31,905,677 total annotated transcripts); indicating the effectiveness of poly-A+ selection prior to sequencing. The proportion of bacterial transcripts with MarRef2+ annotation was similarly low (0.11%). The remaining ~400,000 non-Eukaryotic annotations are predominantly assigned to the basal taxID 'cellular organisms', which DIAMOND cannot classify to the kingdom level with confidence.

A prime motivation for developing a new reference library specific to marine eukaryotic plankton that incorporates sequences from novel and uncultured organisms is to increase the specificity of taxonomic bins by incorporating advances in sequencing the diversity in marine ecosystems. To understand the impact of new reference sequences on annotation results, we compared environmental sequence annotations between the newer MarFERReT+ library and the older MarRef2+ (Fig 2.6). We looked at the distribution of transcripts along internal and external taxonomic ranks (Fig 2.6a) and note shifts in new and extant environmental species-level bins (Fig 2.6b).

DIAMOND's lowest common ancestor (LCA) algorithm infers the most-likely taxonomic identity of a transcript along ranks within NCBI's Taxonomy hierarchy, ranging from broad high-level ranks (e.g., superkingdom, phylum) down to the species level or below. We compared the distribution of rank assignments to eukaryotic taxa between the newer MarFERReT+ and older MarRef2+ reference libraries and found that eukaryotic rank placement shifted towards lower (more specific) ranks using MarFERReT+, highlighting the increase in the proportion of transcripts that can be classified with confidence at lower taxonomic levels using updated expanded eukaryotic references (Fig 2.6a, Supp Fig S2.25).

The increased placements at the species-level are evidenced in the difference in total transcript annotations between the new and old annotations (Table 2.3, Fig 2.6b, Supp Figs S2.26-S2.30). Inclusion of novel taxonomic species bins increased the taxonomic resolution of environmental transcripts previously annotated at coarse high-level ranks, refined species-level annotations, and provided new annotations for transcripts with no previous assignment at all (Fig

2.6, Supp Fig S2.31). In particular, the haptophyte, ochrophyte, dinoflagellate and metazoan lineages were populated with new species bins that had no previous categorization (Supp Figs S2.26-S2.30, respectively). The haptophyte *Chrysochromulina* sp KB-HA01, isolated from the North Pacific and new to MarFERReT, was the environmental species bin with the largest total change (Fig 2.6b). Nearly 40% of the transcripts now assigned to this bin had no previous annotation, and approximately another third was placed at higher ranks above the species (Supp Fig S2.31). Around 7% of transcripts were previously assigned to the closely related species *Haptolina brevifila*. Similarly, the haptophyte *Phaeocystis globosa*, another new species bin (Fig 2.5b), consisted of mostly previously unannotated or high-level transcripts and ~17% of transcripts previously assigned to another *Phaeocystis* species (Supp Fig S2.31).

Some of the species contained in earlier references have been expanded with the addition of newly sequenced subspecies. For example, the cosmopolitan pelagophyte *Pelagomonas calceolata* was included in the MMETSP (and in MarRef2+), and MarFERReT adds four new *P. calceolata* strains not present in MMETSP or MarRef2+. These strain additions resulted in recruitment of new annotated environmental sequences to the ochrophyte *Pelagomonas calceolata* environmental species bin (Fig 2.6b, S2.27). Several new animal species were added to MarFERReT (Fig 2.6b, S2.29) such as the copepods *Paracyclopina nana* and *Calanus finmarchicus*. These copepod bins recruited previously unannotated environmental transcripts as well as transcripts previously assigned to the lone *L. salmonis* copepod reference in MarRefII showing that representative inclusion of animal sequences serves to increase the overall annotation efficiency of marine eukaryotic metatranscriptomes.

The addition of single-cell amplified genomes (SAGs) and transcriptomes (SATs) from uncultured lineages, where publicly available, allowed for annotation of new species-level bins. Five of the uncultured marine stramenopile SAGs each now comprise over 0.1% of total species-level eukaryotic transcripts (Supp Fig S2.30). The majority of transcripts assigned to the most abundant SAG-derived bin, Stramenopiles sp TOSAG23-2, were previously assigned to 'Eukaryota' or were not annotated at all.

In this Case Study, we examined the annotation results drawn from the DIAMOND protein alignment with and without prokaryotes and compared MarFERReT against an older eukaryote-focused reference with limited taxonomic coverage. Compared to the older reference, MarFERReT received more placements at the species-level and other low-level classifications

(Fig 2.6a). Addition of new genera sequences allows for the recognition of new abundant environmental transcript bins (Fig 2.6b), providing new classifications to unknown sequences and refining earlier classifications. Case Study 1 tutorial and code can be found here: (Case Study 1 on github). Overall, with the older reference library, 55.9% of contigs received an annotation at any level, comparable to previous studies. With MarFERReT, 60.9% of contigs received an annotation at any level, an increase of 8.9% that could be taxonomically annotated (Table 2.3).

2.4.5 Case Study 2: Estimating completeness of metatranscriptome taxonomic bins

MarFERReT taxonomic annotations of the metatranscriptome used for Case Study 1 initially generated 501 species-level environmental eukaryotic taxonomic bins. The core transcribed genes (CTG) generated for the different major taxonomic lineages (Table 2.2) can be used to assess the functional coverage of each environmental taxonomic bin by incorporating a metric of functional species metatranscriptome bin completeness similar to a BUSCO completeness analysis (Simão et al., 2015). To assess the coverage of our Case Study 1 environmental taxonomic bins, we assigned functional predictions to the G1 transcripts with the Pfam 34.0 protein family database (Mistry et al., 2021), in a similar fashion to the functional annotation of MarFERReT proteins (Figure 2.5). The species bins were grouped according to lineage and the functional transcript annotations within a species were queried for the percentage of lineage-specific core transcribed genes (%CTGs) developed from MarFERReT. The %CTGs was calculated for all species-level environmental bins across the entire cruise (Fig. 2.7a), highlighting the relative coverage of each bin. As expected, species bins with smaller numbers of transcripts have a smaller fraction of the core transcripts that would be expected in a complete transcriptome; bins with less than 1,000 transcripts have less than a quarter of core transcripts identified. Most species bins are in this rare tail; 388 of the 501 bins have less than 25% CTGs found. Bins recruiting an intermediate level of transcripts (~1,000 to ~20,000) were accompanied by a steep rise in the %CTG identified, and 52 species had at least 50% CTGs in this range. Above ~20,000 transcripts the %CTG began to saturate, indicating species bins inferred to have relatively complete coverage of their transcriptome. A total of 28 species bins had over 75% CTGs found, indicating the species bins within the metatranscriptome with the best overall coverage. We compared the %CTGs coverage estimates to the coverage of 174 BUSCO Protist

Pfams (Simão et al., 2015) (Supp Fig S2.32). Despite the greater number of CTGs per lineage used as a completeness metric, %CTG recovery generally provides a higher estimate of completeness than those derived from the BUSCO set of Protistan genome markers for the identical sets of taxonomic bins (Supp Fig S2.32). We note some lineage-dependent differences in the relationship between %CTG and %BUSCO 'Protist' markers, with metazoans and chlorophytes falling closer to the 1:1 line between estimates, while dinoflagellates were generally higher in their %CTG estimate than the BUSCO counterpart. We also calculated the %CTGs of species in every discrete metatranscriptome sample, allowing us assess coverage independently across samples (Figure 2.7b). In the case of the environmental sequences used in this case study, this can be used to compare species coverage between different sample sites and size-fractions (Figure 2.7b). In this Case Study, we leveraged the combined taxonomic and functional annotation of environmental query sequences to assess the functional transcript inventory of each species bin and estimate completeness of these bins from the percentage of core transcribed genes we expect to find in complete transcriptomes of the same taxonomic lineage. This allowed us to distinguish rare and low-coverage species bins from those approaching a nearly complete transcriptome. We compared coverage estimates against the proportion of Pfam protein families found from a genome-based equivalent and showed an example of how independent assessment of coverage within samples allows us to compare the coverage of recruited species bins across sample sites and size classes. Case Study 2 tutorial and code can be found here: (Case Study 2 on github).

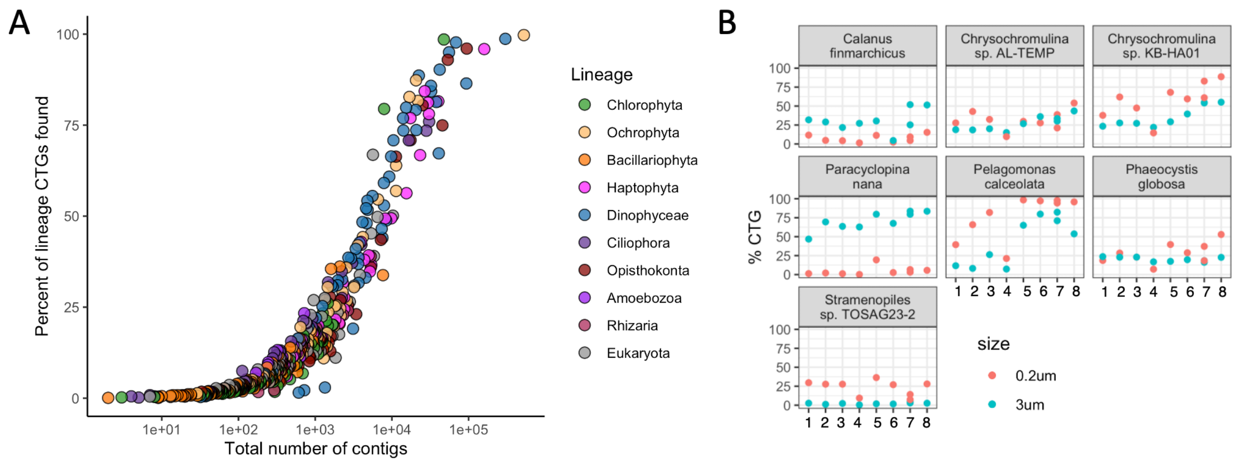


Figure 2.7. **Completeness estimates of environmental eukaryote species bins.** Each circle is an environmental species from the sample metatranscriptome, colored by eukaryotic lineage as in the legend. **A**. Percent of transcribed genes (CTGs) identified in species bins against the $\log_{10}$-transformed number of total transcripts assigned to each species using MarFERReT+ as a reference. **B.** Bin completeness of select species bins within individual sampling sites along 8 environmental stations from an open-ocean oceanographic transect (Lambert et al., 2021). Sample sites are labeled by number on the horizontal axis, increasing with latitude. Points are colored by size class; small (0.2 – 3 µm) and large (3 – 200 µm) size fractions.

2.4.6 Improving MarFERReT with future updates

MarFERReT was designed to be updated as new microbial eukaryote functional reference sequences are publicly released, with releases identified either through literature reviews, the JGI Genomes On Line Database (GOLD), or via user nominations through the 'Issues' request function in the MarFERReT github repository (link). New sequences included in MarFERReT will need to fulfill four requirements: 1) the organism is a marine eukaryote and preferably, a protist, 2) the sequences have been quality-controlled and assembled as transcripts derived from transcriptomes and SATs, or as gene models derived from genomes and SAGs, 3) all sequences are publicly available on a stable repository with an accessible URL, and 4) the organisms should have an associated NCBI taxID, ideally at the species or subspecies level. As new species are incorporated into MarFERReT, the core transcribed gene sets will be refined. New releases of the Pfam-based protein family annotations will be updated as new Pfam releases become available. The structure is flexible enough that additional functional annotations such as KEGG gene ontologies (Kanehisa et al., 2020) can be added for future versions. New versions of MarFERReT will be described in a changelog on the github repository (link), describing any additions or changes to the library composition. The changelog will also detail updates to the MarFERReT code on github and MarFERReT files hosted on Zenodo, including any revisions to the scripts, metadata files, protein sequence library, binary DIAMOND database, and Core Transcribed Gene inventories.

2.5 Discussion

MarFERReT was created out of the need for a reference sequence library that 1) focuses on marine microbial eukaryotes, 2) is represented by a stable and accessible publication and/or DOI, 3) captures recent advances in published sequence data, 4) has transparent and replicable code, and 5) can expand over time with documented releases. Collating eukaryotic reference sequences into a single reference library has made it possible to compare species coverage in MarFERReT with known organisms within the NCBI taxonomic structure and to generate a new transcriptome-based metric for assessing coverage of environmental bins derived from metatranscriptomes.

We find that currently less than 20% of known families of marine eukaryotic microbes have at least one representative species with genome-wide sequence coverage, with the greatest coverage within photosynthetic lineages. And yet, even within those photosynthetic lineages such as diatoms or haptophytes with a relatively high number of sequenced species, the coverage is often biased towards regions with cultured organism from near-shore environments while representation from open ocean and oligotrophic environments remains under-sampled. MarFERReT-based identification of phylogenetic gaps in sequence coverage can help guide future targeted sequencing efforts of cultured isolates, such as the recently completed transcriptome sequencing of 34 isolates from the Roscoff Culture Collection (Niang et al., 2020). In addition, concerted efforts to either bring new isolates from diverse environments into culture, or to further expand single cell sequencing will also generate more diverse coverage. For example, the addition of sequences from new isolates, such as two *Chrysochromulina* strains from the open-ocean North Pacific Ocean, improves the annotation efficiency of environmental reads from the same area and allows a more specific identity for these transcripts.

The nature of trying to match relatively under-described marine taxa to an established taxonomic rank in the NCBI Taxonomy architecture comes with challenges. Accurate phylogenetic placement of sequences using approaches such as DIAMOND's LCA often require public bioinformatic resources like NCBI Taxonomy, but the inclusion of novel species, strains and other taxonomic ranks can take time to make their way into such resources. As a result, there are some inconsistencies in the underlying taxonomic architecture that users should be mindful of: newly described taxa may have no associated species, genus, or other major taxonomic rank; the listed names or taxIDs and taxa may change over time as phylogenetic relationships are revised and reflected in NCBI, and a taxID may not reflect the best or most accurate taxID available for a given organism over time. Curation of MarFERReT sequency entry data ensures the most up-to-date versions of taxID are incorporated, documented and reviewed upon release of new versions of MarFERReT, and will continue to refine associations to the NCBI Taxonomy database as they both expand and change over time.

A second outcome of the creation of MarFERReT is the identification of core gene families transcribed by >95% of species within a lineage. We restricted this analysis to those lineages represented by at least 10 species with transcriptomes to create a new metric of CTGs that can be used to assess coverage within taxonomically binned environmental transcripts and thus address the sparsity issues inherent to metatranscriptome analyses. A common metric of environmental genome coverage is BUSCO (Simão et al., 2015), a compilation of single-copy orthologs for different phyla, including protists. The species used to identify protist single-copy genes are biased towards medically relevant species and thus do not accurately reflect gene distributions in marine protists nor contain ortholog sets tailored to key marine lineages. We propose that our lineage-specific marine CTG compilation provides an important metric of transcriptome coverage of environmental samples. As additional reference sequences are added to MarFERReT, CTG sets will be refined over time and new CTGs can be developed for additional lineages.

In Case Study 1 we demonstrated the practicality of combining MarFERReT with other reference libraries for modular cross-kingdom coverage and compared the results of eukaryotic annotation with MarFERReT against an older library with references from 2015 or earlier. The addition of novel reference species increased the number of eukaryote-annotated transcripts, reflecting the influence of recent reference sequencing efforts on our improved ability to annotate environmental sequence data. This increase also propagated to species-level taxonomic identity, which is the most valuable in terms of taxonomic specificity. Several of the new references in MarFERReT are new species or subspecies from open ocean isolates, and the importance of their addition was seen in their impact on the annotations of the open-ocean environmental sequences used as examples.

We use MARMICRODB (Hogle 2019) an example of sourcing references sequences from complementary taxonomic kingdoms and encourage users to survey publicly-available curated reference libraries to find the resource most appropriate to their dataset. We intend to maintain the focus of MarFERReT on marine microbial eukaryotes through subsequent versions as the parallel field of curated prokaryote-focused reference libraries also continues to develop over time.

Reproducible method and code used for developing MarFERReT are available in a public github repository along with documentation, allowing users to get started with primary MarFERReT data products or recreate the development pipeline described here. The version-controlled nature of the code and data contents can be forked by others if desired and provides users with a venue to nominate new sequences to be included in future versions. The URLs are provided for the sources of the individual sequences, and the compiled, translated, and clustered sequences are available through the public repository, Zenodo (link). We provide code for two Case Studies that allow users to deploy MarFERReT in a command-line environment for common analyses. The MarFERReT structure is specifically designed to expand as needed in future versions with the ability to add new sequences, update functional and/or taxonomic annotations as new versions become available, and update or expand case studies to incorporate additional downstream analyses. Our goal is to create a resource that can keep pace with future marine eukaryote sequencing efforts and further our understanding of the molecular world of marine eukaryotes.

2.6 Data and Code Availability

The aggregated and processed protein data, Pfam functional predictions, CTG catalog, MarFERReT and MarFERReT+ DIAMOND database and other data products are available on Zenodo (link). The source URL, file names and references for raw entry data are listed on this metadata table (link). Code, documentation, and tutorials for this project are available on github: https://github.com/armbrustlab/marine_eukaryote_sequence_database.

2.7 References

Becker, J. W., Hogle, S. L., Rosendo, K., & Chisholm, S. W. (2019). Co-culture and biogeography of *Prochlorococcus* and SAR11. *The ISME journal*, *13*(6), 1506-1519.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, *12*(1), 59-60.

Bucklin, A., Steinke, D., & Blanco-Bercial, L. (2011). DNA barcoding of marine metazoa. *Annual review of marine science*, *3*, 471-508.

Caron, D.A., Alexander, H., Allen, A.E., Archibald, J.M., Armbrust, E., Bachy, C., Bell, C.J., Bharti, A., Dyhrman, S.T., Guida, S.M. and Heidelberg, K.B., Kaye, J. Z., Metzner, J., Smith, S. R., & Worden, A. Z. (2017). Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology*, *15*(1), 6-20.

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., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Tara Oceans Coordinators, Jaillon, O., Aury, J., Karsenti, E., Sullivan, M. B, Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., de Vargas, C., Iudicone, D., Bowler C., & Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature communications*, *9*(1), 373.

Coesel, S. N., Durham, B. P., Groussman, R. D., Hu, S. K., Caron, D. A., Morales, R. L., Ribalet, F., & Armbrust, E. V. (2021). Diel transcriptional oscillations of light-sensitive regulatory elements in open-ocean eukaryotic plankton communities. *Proceedings of the National Academy of Sciences*, *118*(6).

Cooney, E. C., Okamoto, N., Cho, A., Hehenberger, E., Richards, T. A., Santoro, A. E., Worden, A. Z., Leander, B. S. & Keeling, P. J. (2020). Single-cell transcriptomics of Abedinium reveals a new early-branching dinoflagellate lineage. *Genome biology and evolution*, *12*(12), 2417-2428.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS computational biology*, *7*(10), e1002195.

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic acids research*, *40*(D1), D136-D143.

Groussman, R. D., Coesel, S. N., Durham, B. P., & Armbrust, E. V. (2021). Diel-Regulated Transcriptional Cascades of Microbial Eukaryotes in the North Pacific Subtropical Gyre. *Frontiers in microbiology*, *12*.

Guajardo, M., Jimenez, V., Vaulot, D., & Trefault, N. (Assemblies) Transcriptomes from Thalassiosira and Minidiscus diatoms from English Channel and Antarctic coastal waters (Version 1) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.4591037

Grigoriev, I.V., Hayes, R.D., Calhoun, S., Kamel, B., Wang, A., Ahrendt, S., Dusheyko, S., Nikitin, R., Mondo, S.J., Salamov, A. and Shabalov, I., & Kuo, A. (2021). PhycoCosm, a comparative algal genomics resource. *Nucleic acids research*, *49*(D1), D1004-D1011.

Johnson, L. K., Alexander, H., & Brown, C. T. (2019). Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *Gigascience*, *8*(4), giy158.

Kanehisa, M., & Sato, Y. (2020). KEGG Mapper for inferring cellular functions from protein sequences. *Protein science*, *29*(1), 28-35.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A.., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., Deschamps, P., Dyhrman, S. T., Edvardsen, B., Gates, R. D., Gobler, C. J., Greenwood, S. J., Guida, S. M., Jacobi, J. L., Jakobsen, K. S., James, E. R., Jenkins, B., John, U., Johnson, M. D., Juhl, A. R., Kamp, A., Katz, L. A., Kiene, R., Kudryavtsev, A., Leander, B. S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A. M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M. A., Murray, S., Nadathur, G., Nagai, S., Ngam, P. B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M. C., Rengefors, K., Romano, G., Rumpho, M. E., Rynearson, T., Schilling, K. N., Schroeder, D. C., Simpson, A. G. B., Slamovits, C. H., Smith, D. R., Smith, G. J., Smith, S. R., Sosik, H. M., Stief, P., Theriot, E., Twary, S. N., Umale, P. E., Vaulot, D., Wawrik, B., Wheeler, G. L., Wilson, W. H., Xu, Y., Zingone, A., & Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS biology*, *12*(6), e1001889.

Lambert, B. S., Groussman, R. D., Schatz, M. J., Coesel, S. N., Durham, B. P., Alverson, A. J., White, A. E., & Armbrust, E. V. (2022). The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. *Proceedings of the National Academy of Sciences*, *119*(7), e2100916119.

Maas, A. E., Blanco-Bercial, L., Lo, A., Tarrant, A. M., & Timmins-Schiffman, E. (2018). Variations in copepod proteome and respiration rate in association with diel vertical migration and circadian cycle. *The Biological Bulletin*, *235*(1), 30-42.

Mangot J. F., Logares R., Sánchez P., Latorre F., Seeleuthner Y., Mondy S., Sieracki M. E. , Jaillon O., Wincker P., de Vargas C., Massana R. (2017). Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Scientific Reports*, *7*(1), 1-12.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, *38*(10), 4647-4654.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic acids research*, *49*(D1), D412-D419.

Niang, G., Hoebeke, M., Meng, A., Liu, X., Scheremetjew, M., Finn, R., Pelletier, E., & Corre, E. (2020). METdb: A genomic reference database for marine species. *F1000Research*, *9*.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends in genetics*, *16*(6), 276-277.

Roy, R. S., Price, D. C., Schliep, A., Cai, G., Korobeynikov, A., Yoon, H. S., Yang, E. C., & Bhattacharya, D. (2014). Single cell genome analysis of an uncultured heterotrophic stramenopile. *Scientific reports*, *4*(1), 1-8.

Sunagawa, S., de Berardinis, V., Salanoubat, M., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Tara Oceans Coordinators, Pesant, S., Poulton, N., Stepanauskas, R., Bork, P., Bowler, C., Hingamp, P., Sullivan, M. B., Iudicone, D., Massana, R., Aury, J., Henrissat, B., Karsenti, E., Jaillon, O., Sieracki, M., de Vargas, C., & Wincker, P. (2018). Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nature communications*, *9*(1), 1-10.

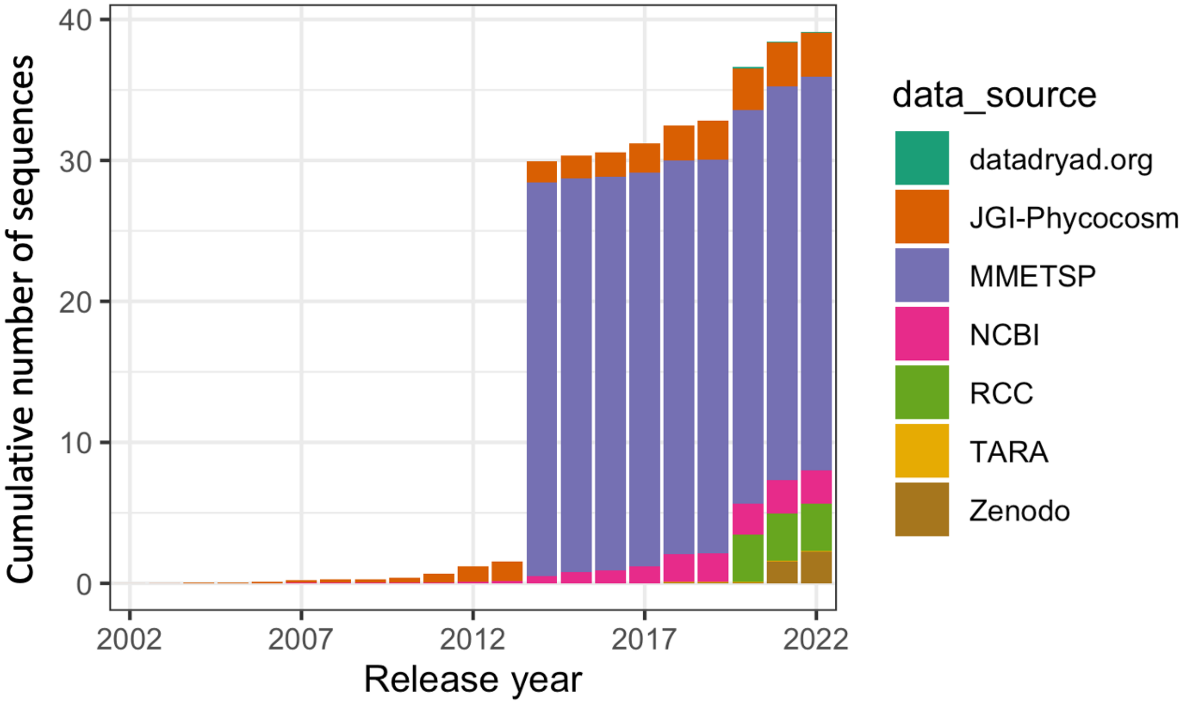
Sieracki, M. E., Poulton, N. J., Jaillon, O., Wincker, P., De Vargas, C., Rubinat-Ripoll, L., Stepanauskas, R., Logares, R., & Massana, R. (2019). Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Scientific reports*, *9*(1), 1-11.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212.

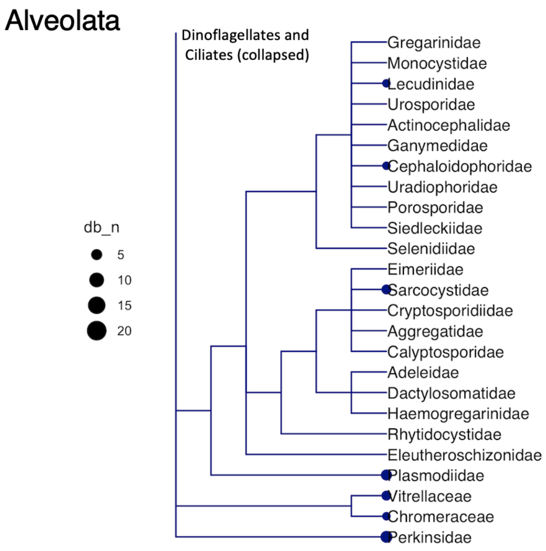
Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature communications*, *9*(1), 1-8.

Vanhoorne, B., Deneudt, K., Appeltans, W., Hernandez, F., & Mees, J. (2008). The aphia taxon match tool, an online quality control tool for checking your biological data against the World and European Registers of Marine Species. In *2008 International Marine Data and Information Systems conference (IMDIS 2008)* (pp. 145-145). Hellenic Centre for Marine Research (HCMR).

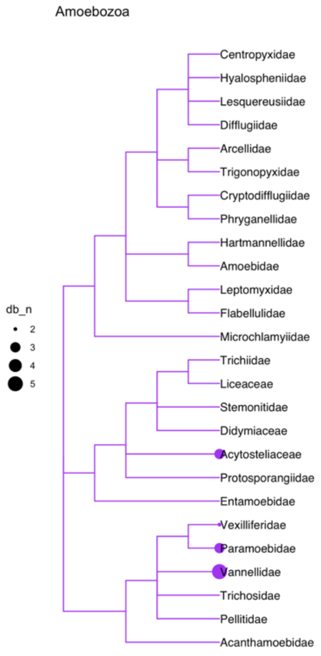
2.8 Supplementary Figures



Supplementary Figure S2.1. Cumulative growth in marine microbial eukaryote reference sequences over time. From MarFERReT entries based on the year of associated publication or public release of material. Counts include number of translated protein sequences.

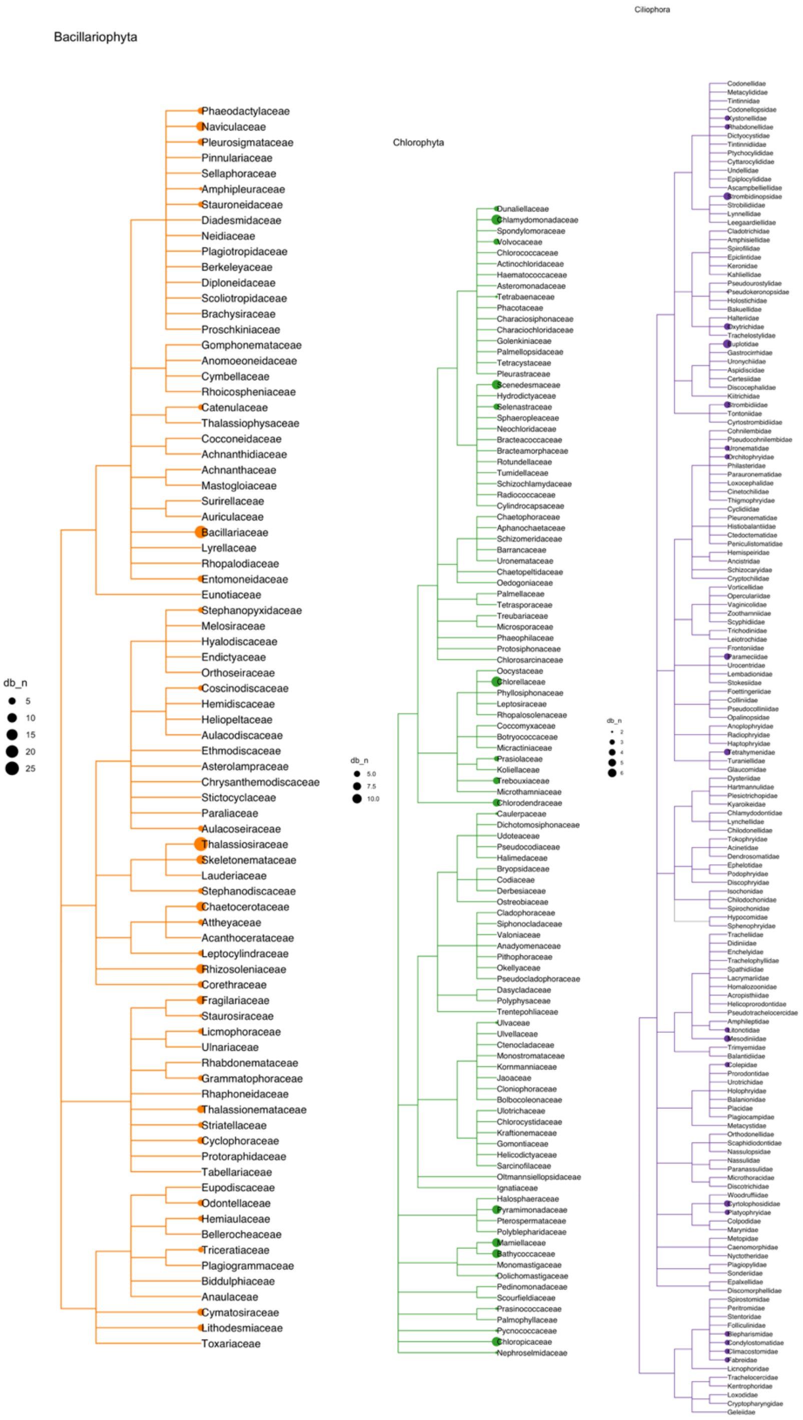


Supplementary Figure S2.2. Cladogram of Alveolata families. Cladogram of hierarchical taxonomic ranks of marine eukaryotes based on the NCBI Taxonomy framework (Federhen 2012, link) using their CommonTree tool (link). Cladogram is subsetted by lineage shown in title. Tips were pruned down to the Family level and filtered by taxonomic Families that include marine species as listed in the World Register of Marine Species (Vanhoorne et al., 2008, link). The size of the dot on each tip indicates the number of species in the family represented in MarFERReT v1.0.

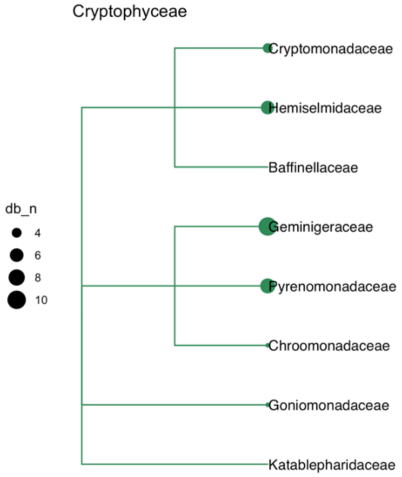


Supplementary Figure S2.3. Cladogram of Amoebozoa families. Cladogram of hierarchical taxonomic ranks of marine eukaryotes based on the NCBI Taxonomy framework (Federhen 2012, link) using their CommonTree tool (link). Cladogram is subsetted by lineage shown in title. Tips were pruned down to the Family level and filtered by taxonomic Families that include marine species as listed in the World

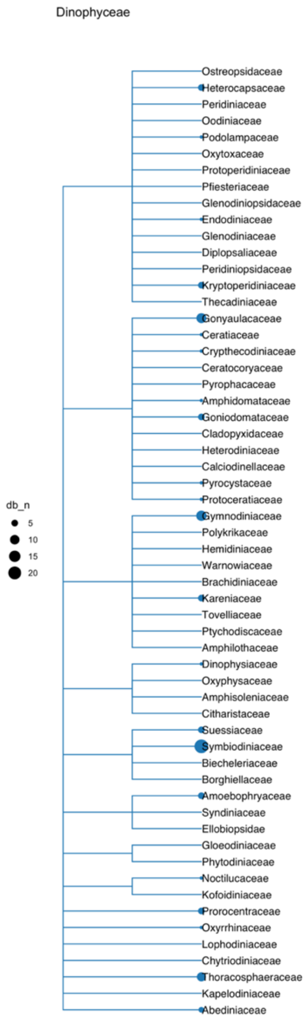
Register of Marine Species (Vanhoorne et al., 2008, link). The size of the dot on each tip indicates the number of species in the family represented in MarFERReT v1.0.



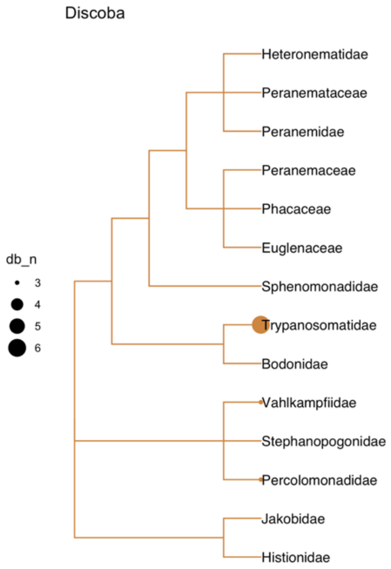
Supplementary Figure S2.4, S2.5, S2.6. Cladogram of Bacillariophyta, Chlorophyta, and Ciliophora families. Cladogram of hierarchical taxonomic ranks of marine eukaryotes based on the NCBI Taxonomy framework (Federhen 2012, link) using their CommonTree tool (link). Cladogram is subsetted by lineage shown in title. Tips were pruned down to the Family level and filtered by taxonomic Families that include marine species as listed in the World Register of Marine Species (Vanhoorne et al., 2008, link). The size of the dot on each tip indicates the number of species in the family represented in MarFERReT v1.0.



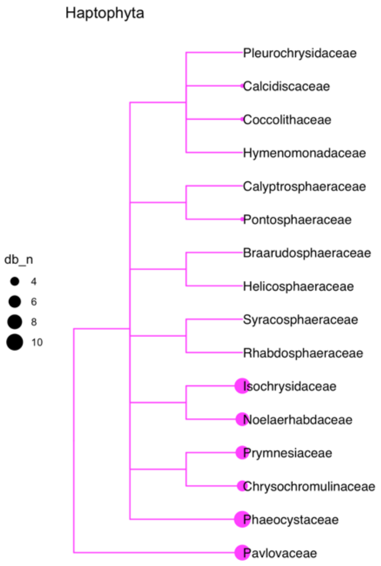
Supplementary Figure S2.7. Cladogram of Cryptophyceae families. Cladogram of hierarchical taxonomic ranks of marine eukaryotes based on the NCBI Taxonomy framework (Federhen 2012, link) using their CommonTree tool (link). Cladogram is subsetted by lineage shown in title. Tips were pruned down to the Family level and filtered by taxonomic Families that include marine species as listed in the World Register of Marine Species (Vanhoorne et al., 2008, link). The size of the dot on each tip indicates the number of species in the family represented in MarFERReT v1.0.



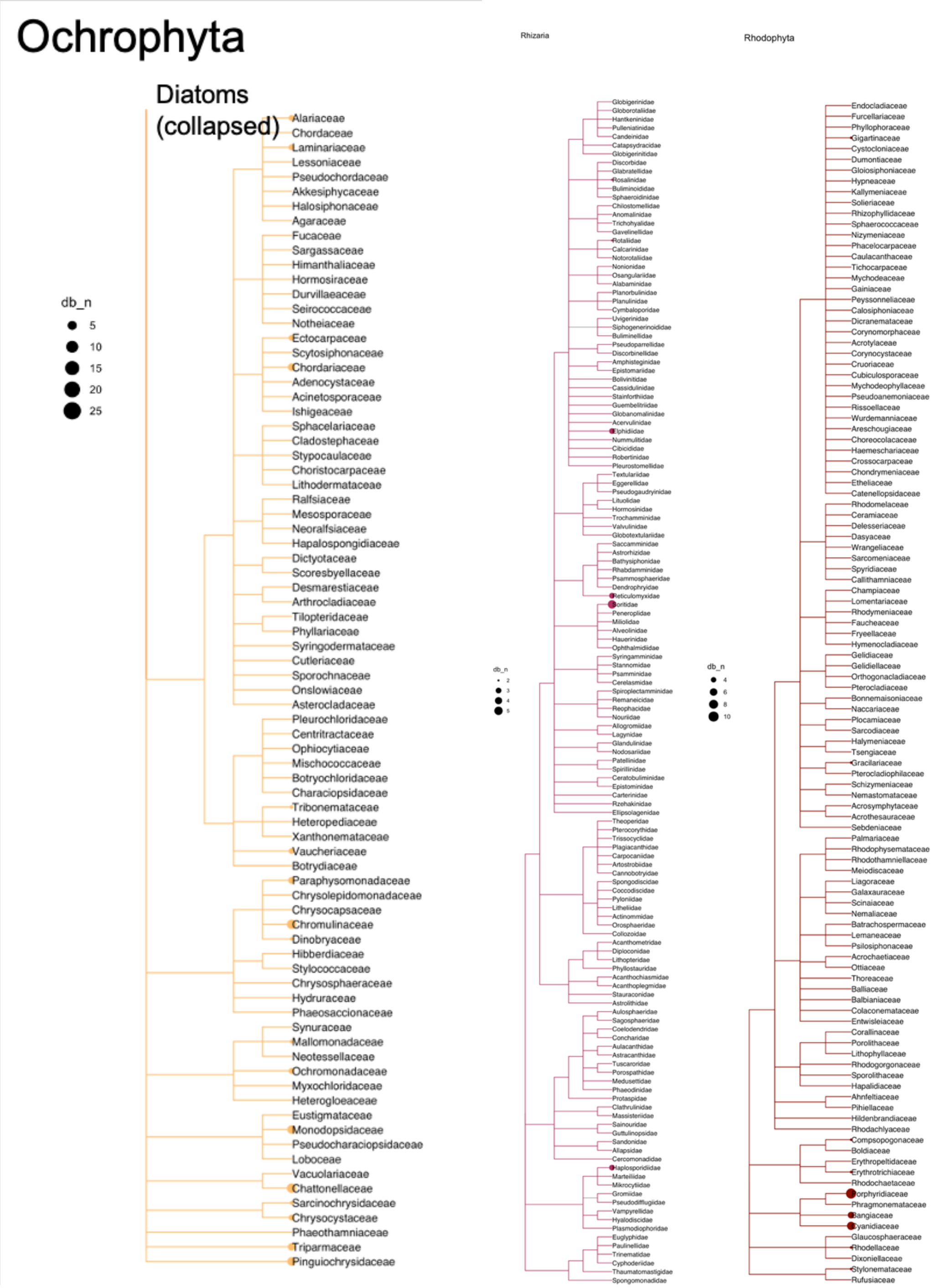
Supplementary Figure S2.8. Cladogram of Dinophyceae families. Cladogram of hierarchical taxonomic ranks of marine eukaryotes based on the NCBI Taxonomy framework (Federhen 2012, link) using their CommonTree tool (link). Cladogram is subsetted by lineage shown in title. Tips were pruned down to the Family level and filtered by taxonomic Families that include marine species as listed in the World Register of Marine Species (Vanhoorne et al., 2008, link). The size of the dot on each tip indicates the number of species in the family represented in MarFERReT v1.0.



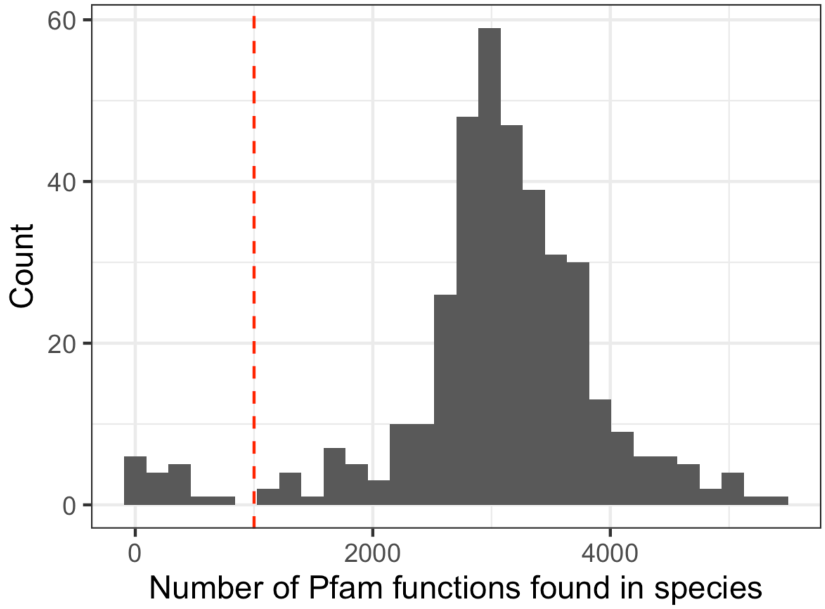
Supplementary Figure S2.9. Cladogram of Discoba families. Cladogram of hierarchical taxonomic ranks of marine eukaryotes based on the NCBI Taxonomy framework (Federhen 2012, link) using their CommonTree tool (link). Cladogram is subsetted by lineage shown in title. Tips were pruned down to the Family level and filtered by taxonomic Families that include marine species as listed in the World Register of Marine Species (Vanhoorne et al., 2008, link). The size of the dot on each tip indicates the number of species in the family represented in MarFERReT v1.0.



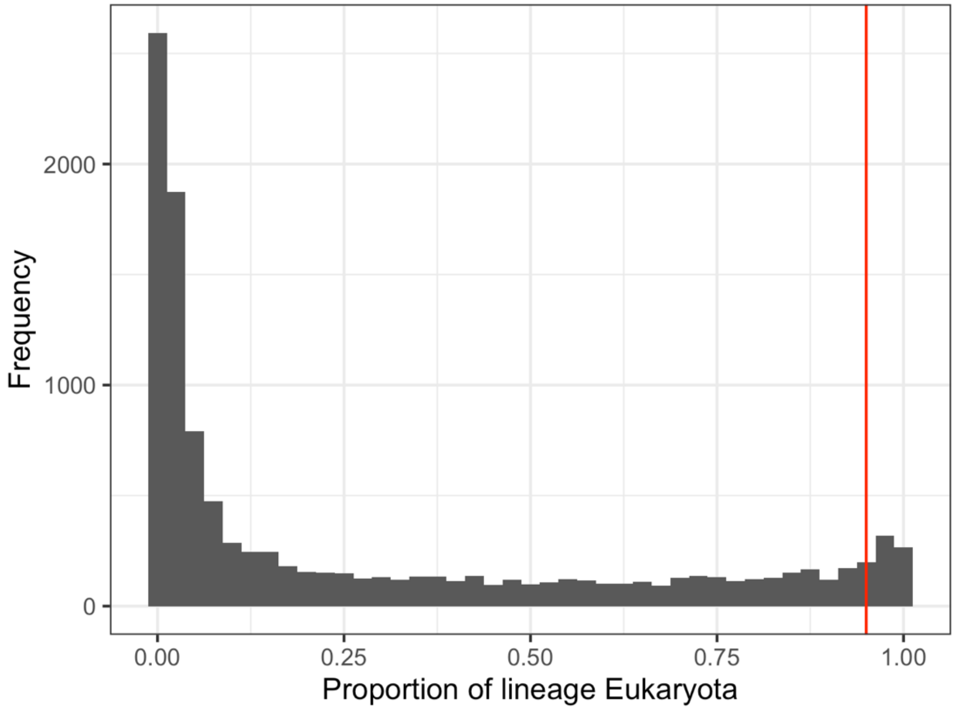
Supplementary Figure S2.10. Cladogram of Haptophyta families. Cladogram of hierarchical taxonomic ranks of marine eukaryotes based on the NCBI Taxonomy framework (Federhen 2012, link) using their CommonTree tool (link). Cladogram is subsetted by lineage shown in title. Tips were pruned down to the Family level and filtered by taxonomic Families that include marine species as listed in the World Register of Marine Species (Vanhoorne et al., 2008, link). The size of the dot on each tip indicates the number of species in the family represented in MarFERReT v1.0.



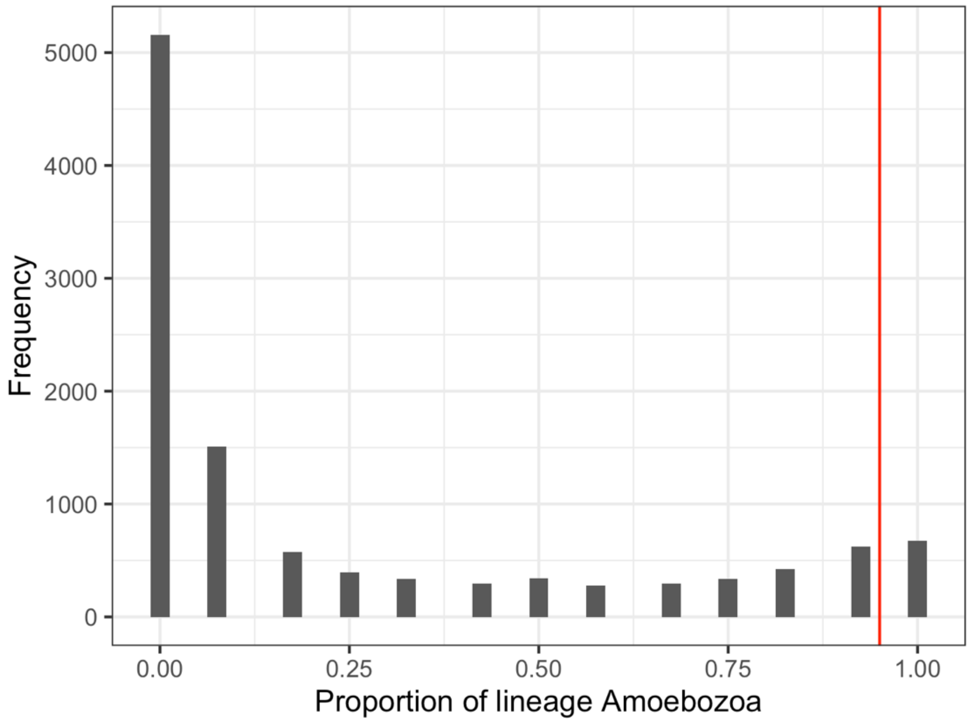
Supplementary Figure S2.11, S2.12, S2.13. Cladogram of Ochrophyta, Rhizaria, and Rhodophyta families. Cladogram of hierarchical taxonomic ranks of marine eukaryotes based on the NCBI Taxonomy framework (Federhen 2012, link) using their CommonTree tool (link). Cladogram is subsetted by lineage shown in title. Tips were pruned down to the Family level and filtered by taxonomic Families that include marine species as listed in the World Register of Marine Species (Vanhoorne et al., 2008, link). Diatoms 'Bacillariophyta' are collapsed in the Ochrophyta cladogram. The size of the dot on each tip indicates the number of species in the family represented in MarFERReT v1.0.



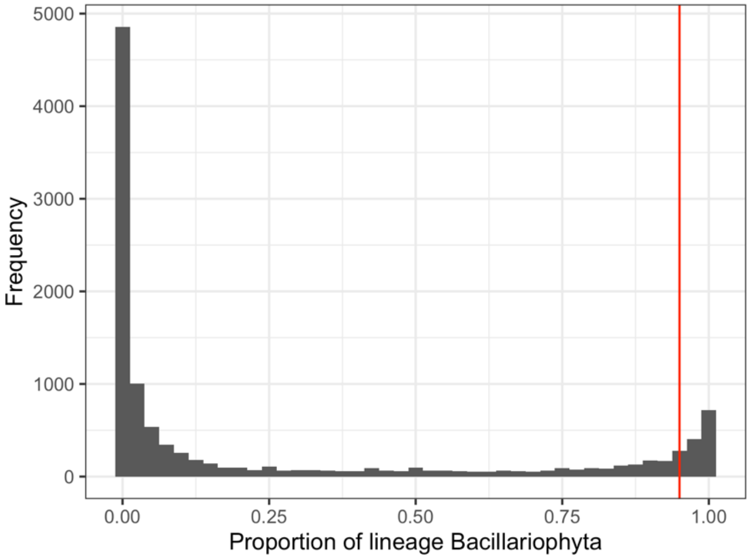
Supplementary Figure S2.14. Histogram of the Pfam function count in MarFERReT species. Horizontal axis is the number of identified Pfams, vertical axis is the count of species with this number in the histogram bin. Dotted vertical line at 1000 Pfams indicates the cut-off for inclusion in core transcribed genes identification.



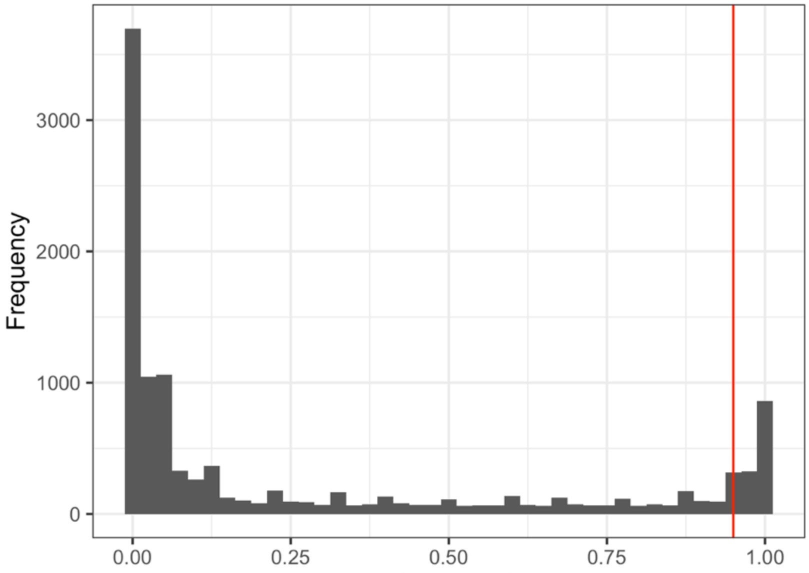
Supplementary Figure S2.15. Histogram of Pfam occurrence in all Eukaryota species. Histogram of the frequency of appearance of Pfam 34.0 protein families in MarFERReT species, within Eukarya. Horizontal axis shows the fraction of species, vertical axis shows the frequency of detection in samples of this lineage. The vertical red line marks the 95% frequency cutoff; Pfams to the right of this line are present in at least 95% of species within this group and considered core transcribed genes.



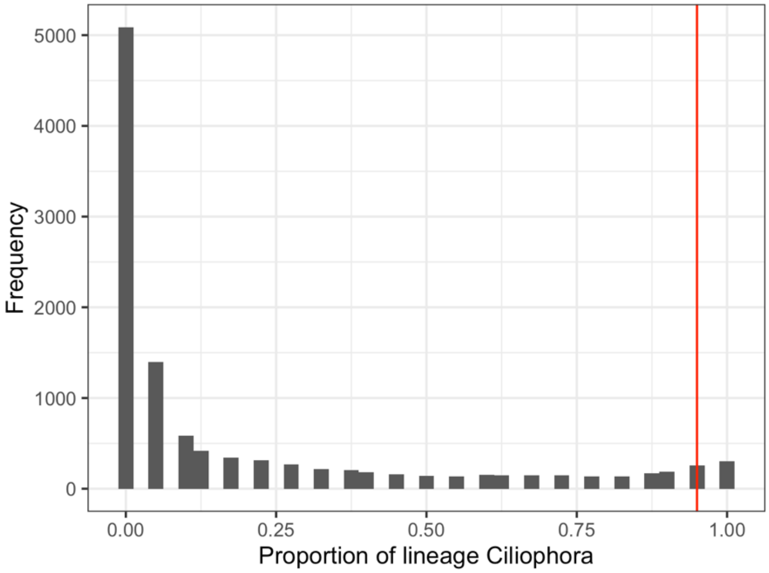
Supplementary Figure S2.16. Histogram of Pfam occurrence in Amoebozoa species. Histogram of the frequency of appearance of Pfam 34.0 protein families in MarFERReT species, within the stated lineage. Horizontal axis shows the fraction of species in the group indicated in the title, vertical axes values show the frequency of detection in samples of this species. The vertical red line marks the 95% frequency cutoff; Pfams to the right of this line are present in at least 95% of species within this group and considered core transcribed genes.



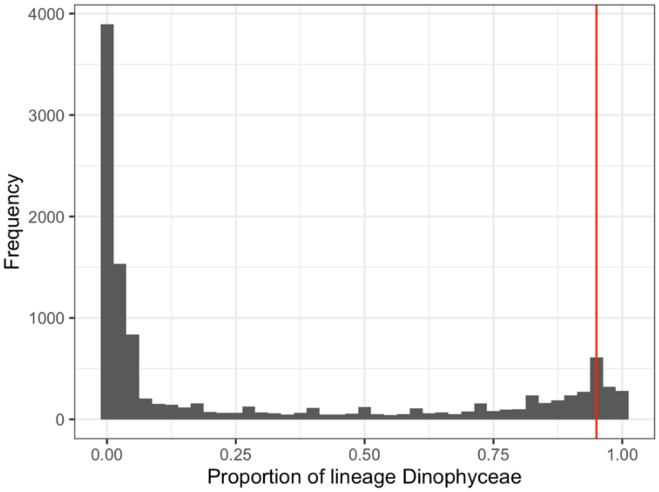
Supplementary Figure S2.17. Histogram of Pfam occurrence in Bacillariophyta species. Histogram of the frequency of appearance of Pfam 34.0 protein families in MarFERReT species, within the stated lineage. Horizontal axis shows the fraction of species in the group indicated in the title, vertical axes values show the frequency of detection in samples of this species. The vertical red line marks the 95% frequency cutoff; Pfams to the right of this line are present in at least 95% of species within this group and considered core transcribed genes.



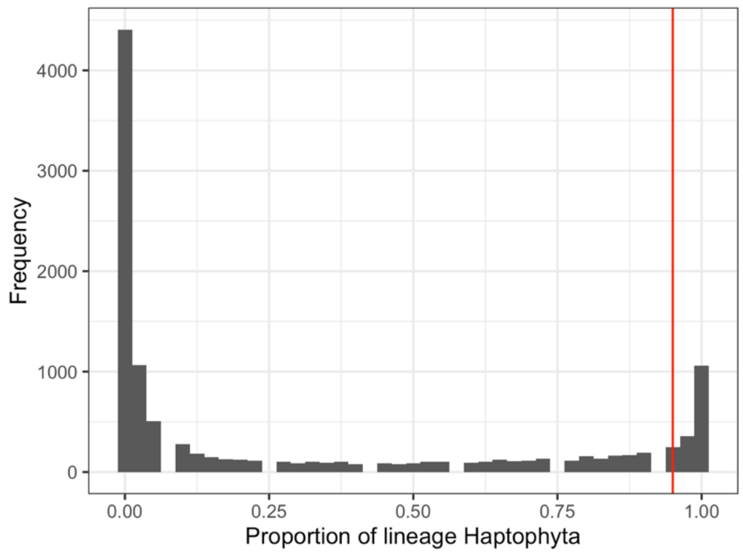
Supplementary Figure S2.18. Histogram of Pfam occurrence in Chlorophyta species. Histogram of the frequency of appearance of Pfam 34.0 protein families in MarFERReT species, within the stated lineage. Horizontal axis shows the fraction of species in the group indicated in the title, vertical axes values show the frequency of detection in samples of this species. The vertical red line marks the 95% frequency cutoff; Pfams to the right of this line are present in at least 95% of species within this group



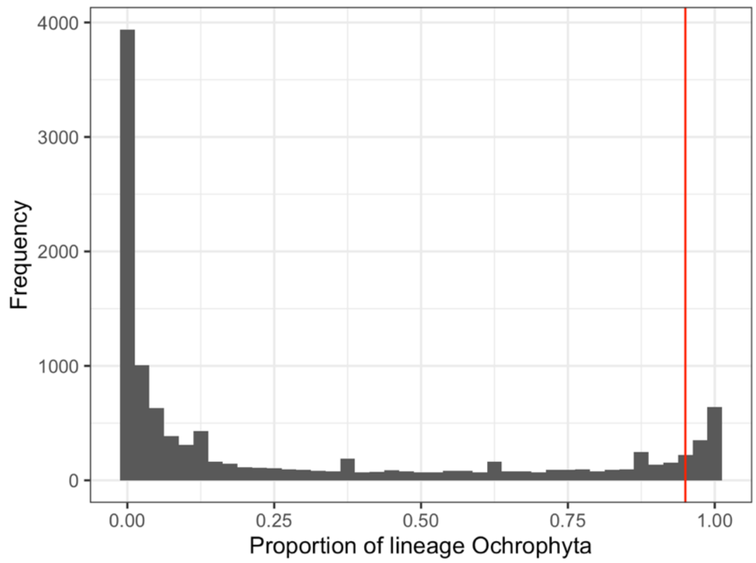
Supplementary Figure S2.19. Histogram of Pfam occurrence in Ciliophora species. Histogram of the frequency of appearance of Pfam 34.0 protein families in MarFERReT species, within the stated lineage. Horizontal axis shows the fraction of species in the group indicated in the title, vertical axes values show the frequency of detection in samples of this species. The vertical red line marks the 95% frequency cutoff; Pfams to the right of this line are present in at least 95% of species within this group.



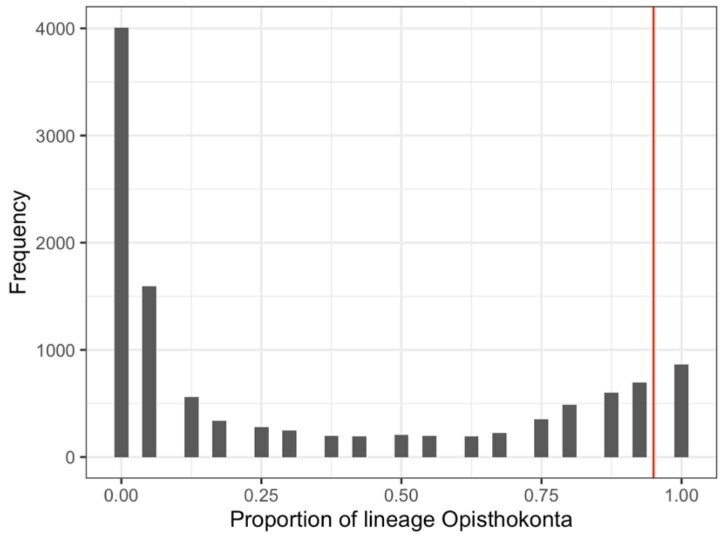
Supplementary Figure S2.20. Histogram of Pfam occurrence in Dinophyceae species. Histogram of the frequency of appearance of Pfam 34.0 protein families in MarFERReT species, within the stated lineage. Horizontal axis shows the fraction of species in the group indicated in the title, vertical axes values show the frequency of detection in samples of this species. The vertical red line marks the 95% frequency cutoff; Pfams to the right of this line are present in at least 95% of species within this group and considered core transcribed genes.



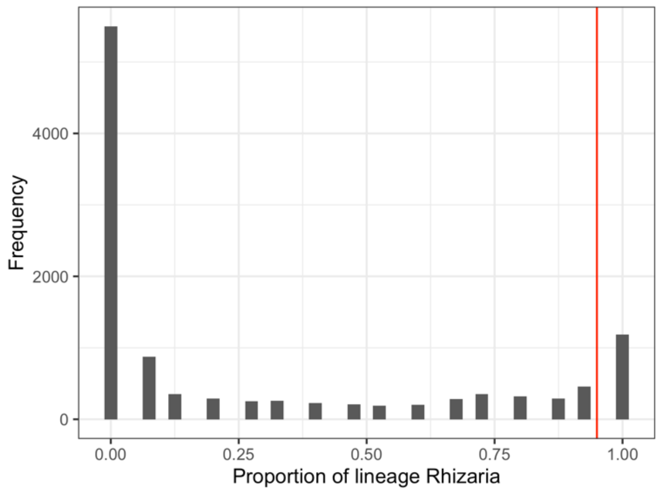
Supplementary Figure S2.21. Histogram of Pfam occurrence in Haptophyta species. Histogram of the frequency of appearance of Pfam 34.0 protein families in MarFERReT species, within the stated lineage. Horizontal axis shows the fraction of species in the group indicated in the title, vertical axes values show the frequency of detection in samples of this species. The vertical red line marks the 95% frequency cutoff; Pfams to the right of this line are present in at least 95% of species within this group and considered core transcribed genes.



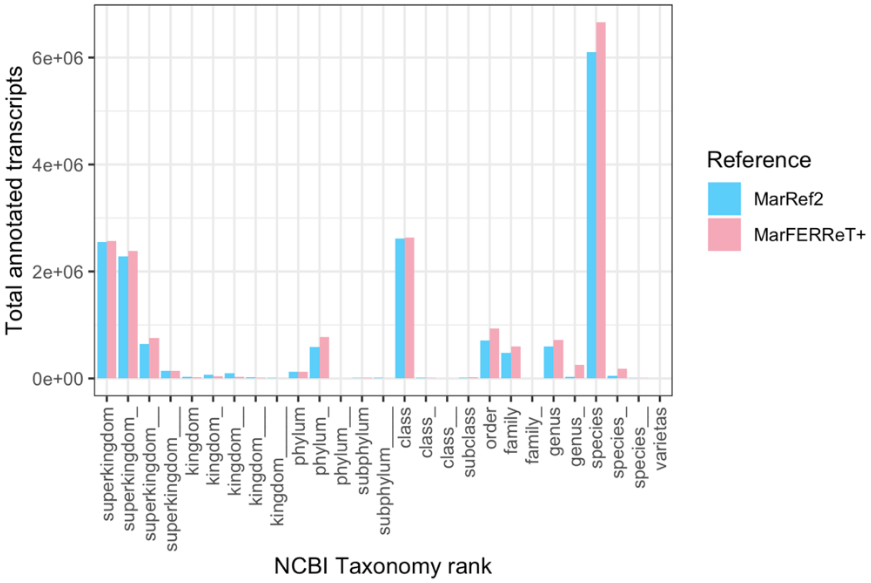
Supplementary Figure S2.22. Histogram of Pfam occurrence in Ochrophyta species. Histogram of the frequency of appearance of Pfam 34.0 protein families in MarFERReT species, within the stated lineage. Horizontal axis shows the fraction of species in the group indicated in the title, vertical axes values show the frequency of detection in samples of this species. The vertical red line marks the 95% frequency cutoff; Pfams to the right of this line are present in at least 95% of species within this group.



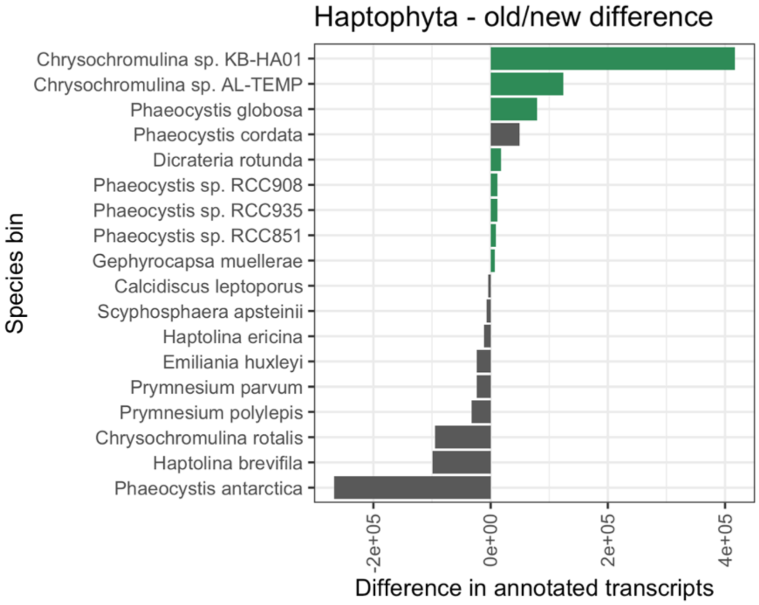
Supplementary Figure S2.23. Histogram of Pfam occurrence in Opisthokonta species. Histogram of the frequency of appearance of Pfam 34.0 protein families in MarFERReT species, within the stated lineage. Horizontal axis shows the fraction of species in the group indicated in the title, vertical axes values show the frequency of detection in samples of this species. The vertical red line marks the 95% frequency cutoff; Pfams to the right of this line are present in at least 95% of species within this group.



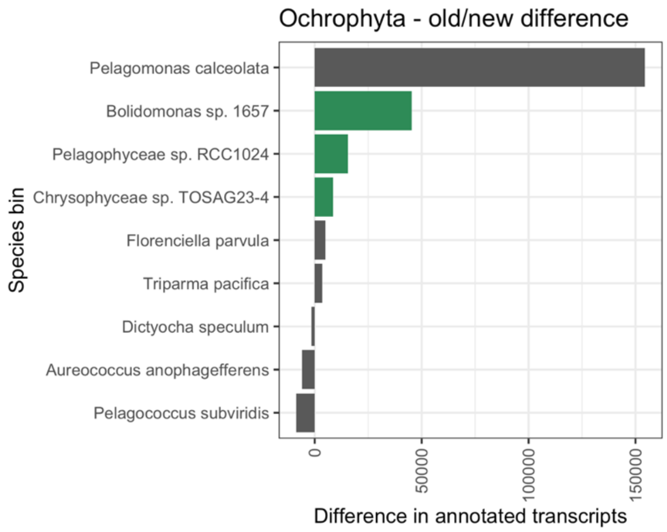
Supplementary Figure S2.24. Histogram of Pfam occurrence in Rhizaria species. Histogram of the frequency of appearance of Pfam 34.0 protein families in MarFERReT species, within the stated lineage. Horizontal axis shows the fraction of species in the group indicated in the title, vertical axes values show the frequency of detection in samples of this species. The vertical red line marks the 95% frequency cutoff; Pfams to the right of this line are present in at least 95% of species within this group and considered core transcribed genes.



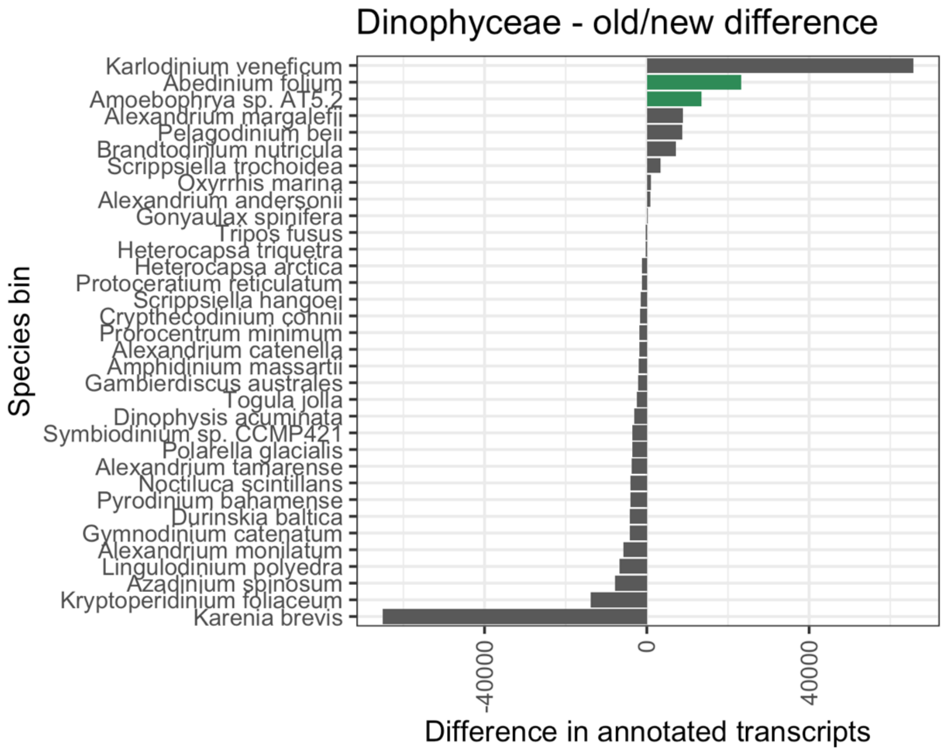
Supplementary Figure S2.25. Comparison of lowest-common-ancestor rank placements. The MarFERReT + MarMicroDB combined reference library (MarFERReT+, pink fill) is shown against another mixed-domain reference based on MarineRefII (MarRef2+, blue fill, Coesel et al., 2021).



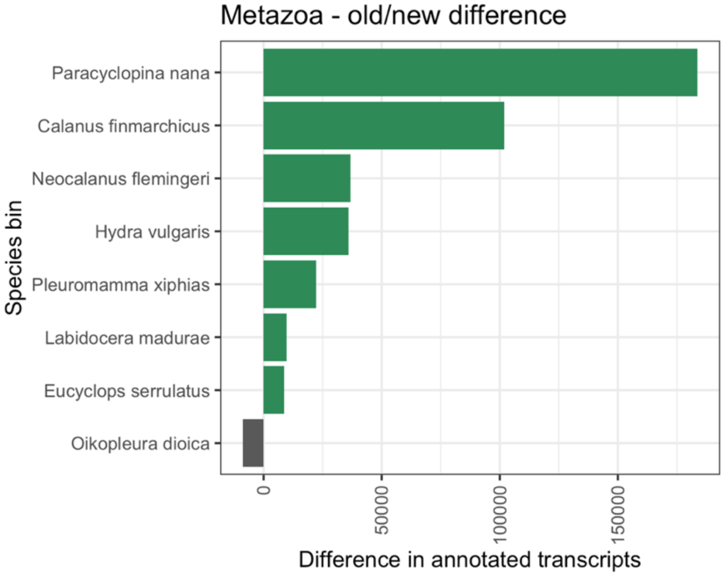
Supplementary Figure S2.26. Difference in transcript annotations by reference for Haptophyte species bins. Horizonal axis is the difference in the number of annotated transcripts between references (MarFERReT+ minus MarRef+), for haptophyte species bins with at least 0.1% of total transcript annotations in eukaryotic species. Green color indicates new species bins in MarFERReT v1 that were not present in the older reference.



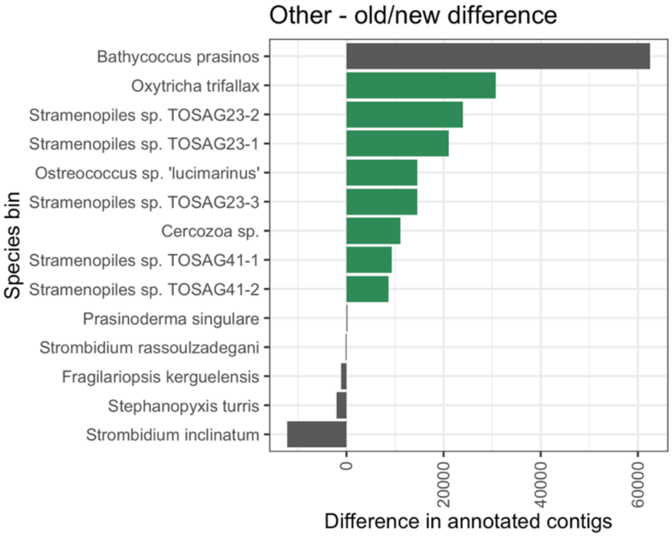
Supplementary Figure S2.27. Difference in transcripts annotations by reference for Ochrophyte species bins. Horizonal axis is the difference in the number of annotated transcripts between references (MarFERReT+ minus MarRef+), for ochrophyte species bins with at least 0.1% of total transcript annotations in eukaryotic species.



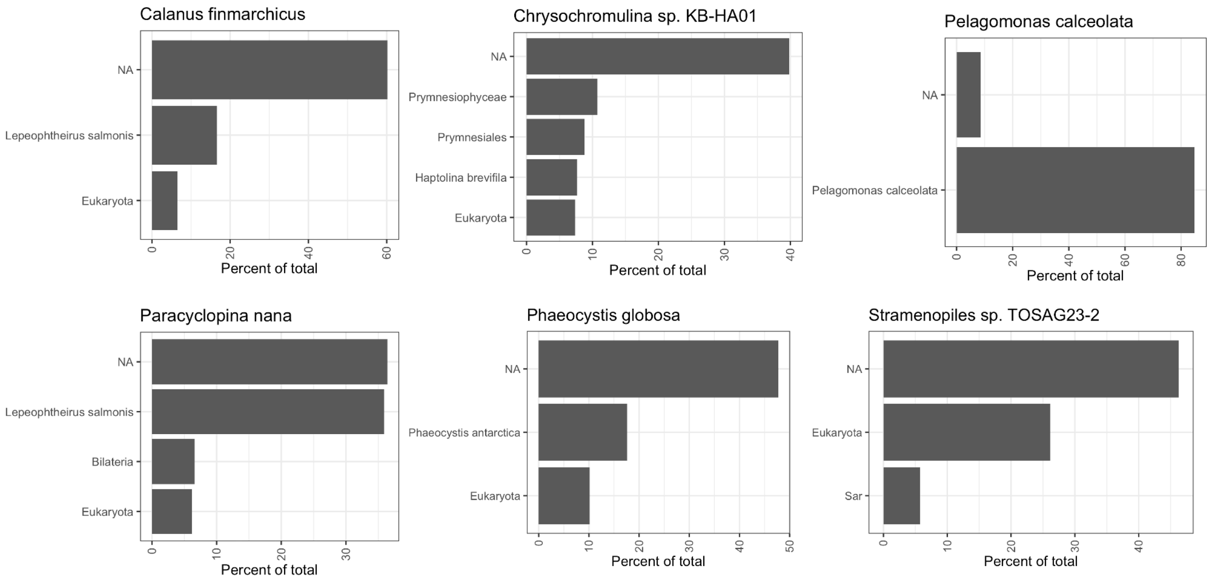
Supplementary Figure S2.28. Difference in transcripts annotations by reference for Dinoflagellate species bins. Horizonal axis is the difference in the number of annotated transcripts between references (MarFERReT+ minus MarRef+), for dinoflagellate species bins with at least 0.1% of total transcript annotations in eukaryotic species. Green color indicates new species bins in MarFERReT v1 that were not present in the older reference.



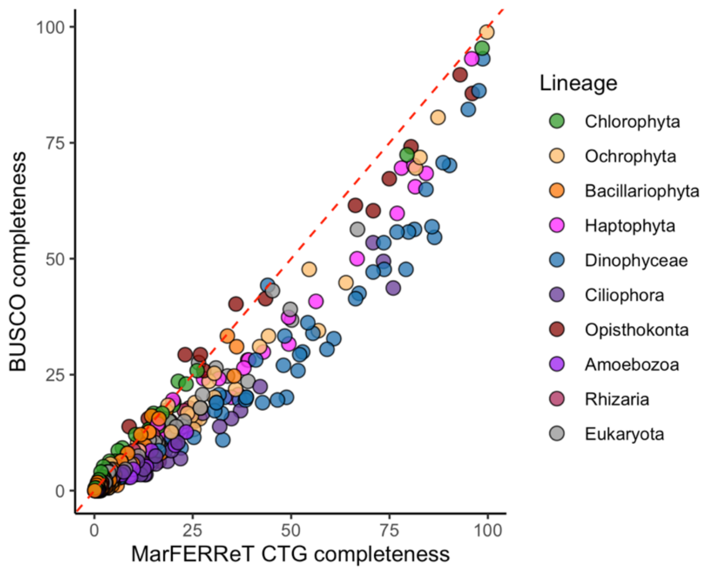
Supplementary Figure S2.29. Difference in transcripts annotations by reference for Metazoa (animal) species bins. Horizonal axis is the difference in the number of annotated transcripts between references (MarFERReT+ minus MarRef+), for metazoan species bins with at least 0.1% of total transcript annotations in eukaryotic species. Green color indicates new species bins in MarFERReT v1 that were not present in the older reference.



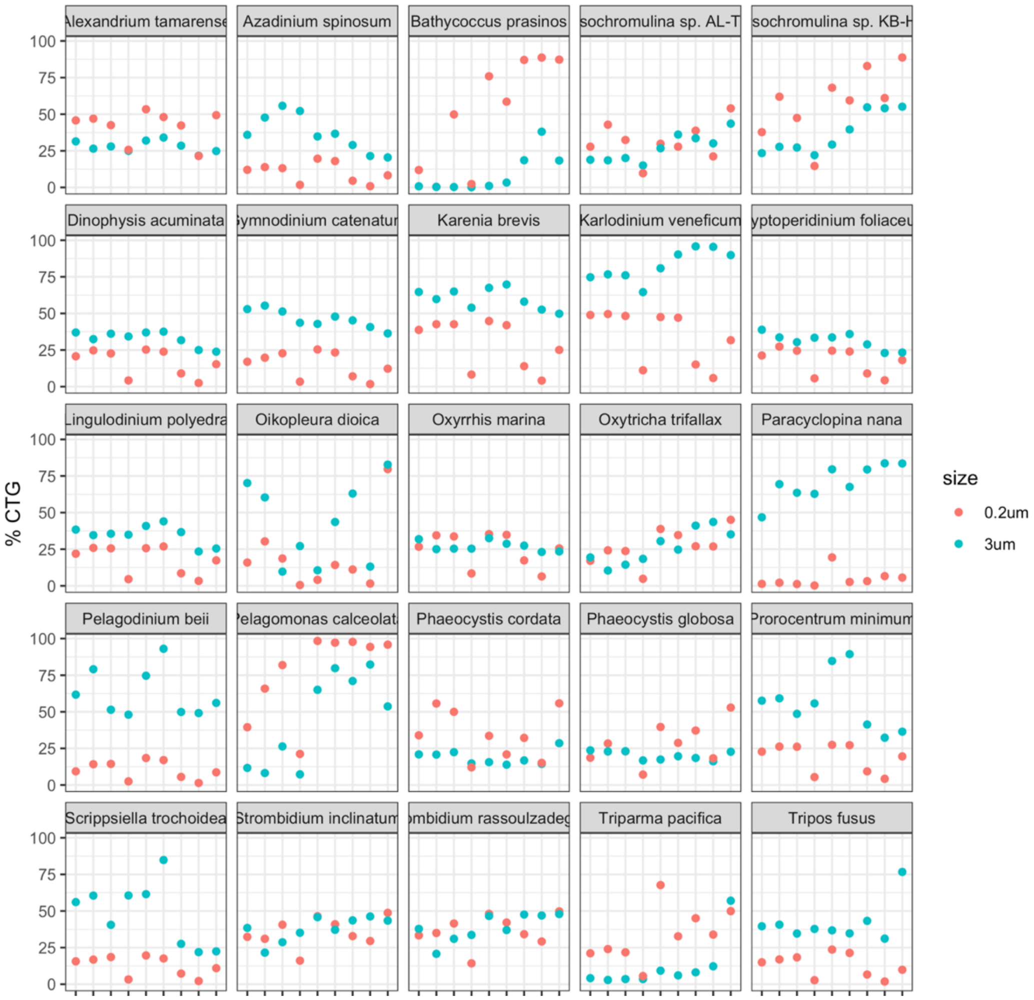
Supplementary Figure S2.30. Difference in transcripts annotations by reference for other species bins. Horizonal axis is the difference in the number of annotated transcripts between references (MarFERReT+ minus MarRef+), for eukaryotic species bins with at least 0.1% of total transcript annotations except for haptophyte, ochrophyte, dinoflagellate and metazoan species (Shown in previous Supplemental Figures). Green color indicates new species bins in MarFERReT v1 that were not present in the older reference.



Supplementary Figure S2.31. A selection of MarFERReT species bins and the composition of transcript annotations from an earlier eukaryotic reference, MarRef. The title of each subplot is a species bin derived from annotations using MarFERReT v1. The horizontal axis is the percent of transcripts in this bin found in previous annotation assignments to all ranks; only bins with at least 5% of the previous composition are shown. A previous assignment of 'NA' reflects transcripts with no annotation using the earlier reference.



Supplemental Figure S2.32. MarFERReT %CTG identification vs BUSCO Protistan genome markers for species-level bins. Each dot is an environmental genus from the Gradients 1 metatranscriptome, colored by eukaryotic lineage in the legend. Comparison of completeness estimates using the percentage of Pfams (out of 174 total) associated with the BUSCO set of Protistan genome markers, against the percentage of MarFERReT lineage-specific or general Eukaryotic set of CTG transcriptome markers.



Supplemental Figure S2.33. Species bin completeness at separate sampling sites and size fractions. Bin completeness of the 25 species bins with the highest average completeness across individual sampling sites along 8 environmental stations from a latitudinal transect along 158°W. Sample sites are labeled by number on the horizontal axis, increasing with latitude. Points are colored by size class; 0.2um (small, 0.2 – 3 µm) and 3um (large, 3 – 200 µm) size fractions.

CHAPTER 3

Molecular ecology of picoeukaryotes in the North Pacific Transition Zone

Ryan D. Groussman

3.1 Abstract

The North Pacific Transition Zone is a basin-scale oceanographic feature with dynamic physical biogeochemical gradients, separating the warm and nitrogen-deplete North Pacific Subtropical Gyre to the south and the cold, iron-limited North Pacific Subpolar Gyre to the north. Northward across the transition zone, net community productivity increases as a diverse community of photosynthetic picoeukaryotes increase in biomass and abundance. The ecological functions that picoeukaryote species perform in this region are not well understood. The Gradients cruises of 2016, 2017 and 2019 conducted high-resolution transects across the North Pacific Transition Zone to study this productive and dynamic region. Size-fractionated metatranscriptomes taken in conjunction with other measurements indicate that the picoeukaryote community is more diverse than larger size class community with a higher proportion of haptophyte, chlorophytes, ochrophytes in terms of community transcript composition. We conducted on-deck incubation experiments test the community transcriptional response to amendments with different ratios of N, P and Fe at three sites in the transition zone. Treatments induced significant changes in community transcript composition 4 days after the amendments. A subset of responder species significantly increased their transcript pools under the treatments. These responders are primarily osmo-mixotrophic and phago-mixotrophic ochrophytes and haptophytes, and the differentially transcribed transcript functions of these species illustrate their ability to shift metabolic modes to adapt to changing environmental conditions. This is advantageous in the dynamic NPTZ, where the seasonal dynamics of the chlorophyll front and mixing of different water masses expose communities to gradients in biochemical and physical conditions.

3.2 Introduction

The North Pacific Ocean is structured by large-scale physical processes into provinces with distinct biogeochemical regimes. The North Pacific Subtropical Gyre is a basin-spanning feature characterized by warm and nutrient-deplete water, and the North Pacific Subpolar Gyre situated in higher latitudes is characterized by cold, nutrient-rich and iron-poor waters. The region between these gyres is known as the North Pacific Transition Zone (NPTZ), a latitudinal band spanning ~32°-42°N of strong physical, chemical, and biological gradients and high net community productivity (Roden 1991, Polovina et al., 2001, Juranek et al., 2012, Juranek et al., 2020). The NPTZ is defined by sharp gradients in water column parameters. The southern edge of the NPTZ is separated from the norther boundary of the NPSG by a salinity front (Roden 1991), operationally defined as the 34.82 isohalocline (commonly at ~32°N, Follett et al., 2021). A secondary feature of the NPTZ is Transition Zone Chlorophyll Front (TZCF), operationally defined by chlorophyll *a* concentrations of 0.2 mg chl m^{-3} that shifts latitudinally over seasonal cycles (Polovina et al., 2001, Follett et al., 2021), extending as far south as ~33°N in February and retreating to maximum northern extent of ~40°N in August. The TZCF is associated with regional maxima in net community production (Juranek et al., 2020) and is used to separate northern and southern subregions (Juranek et al., 2012, Follett et al., 2021). This region has high productivity and net CO_2 flux: net and gross oxygen production here are 2 to 4 times that of the surrounding areas (Juranek et al., 2012), with a relatively high net flux of CO_2 out of the atmosphere (Takahashi et al., 2002). The NPTZ is a known habitat of albacore tuna, loggerhead turtles, and other macrofauna sustained by this productive ecosystem (Polovina et al., 2017). The particle size distribution shifts across the NPTZ (Sheldon et al., 1972), with the shift towards larger particles in northern latitudes corresponding to increases in the pico- and nano-eukaryote population (Juranek et al., 2012, 2020). The southern salinity front acts as a natural barrier between the southern NPSG communities and the northern NPTZ communities. Nitrogen fixation by UCYN-A and other diazotrophs via the *nifH* gene is common in the NPSG, and *nifH* decreases below detection thresholds north of the salinity front (Gradoville et al., 2020), concomitant with decreases in *Prochlorococcus* and increases in *Synechococcus* abundance.

Transition zones between biomes, or ecotones, are hotspots of species diversity that often host communities with distinct features from their neighboring biomes (Kark et al., 2007, Kark

2013). One explanation for the shift in community composition over the natural biogeochemical gradients is the Resource Ratio hypothesis (Tilman 1985), which assumes that a given species will outcompete other species under different ratios of available nutrients. The N to Fe ratio, for instance, has been used to explain the biogeography of nitrogen-fixing organisms in the ocean (Ward et al., 2013). In aquatic environments, water column parameters have also been invoked to explain shifts in community composition. Margalef (1978) developed a conceptual model of phytoplankton community structure as a function of nutrients and water column turbulence ('Margalef's mandala'), which has since been expanded to also consider adaptations to light levels, nutrient reduction state, motility, and other functional traits (Glibert 2016).

The ecological fitness of microbial eukaryotes to the conditions of the NPTZ reflects individual species adaptations as well as the legacy of a long evolutionary history that shapes their macromolecular stoichiometry and trace metal requirements. These eukaryote lineages are separated by hundreds of millions of years of evolution that have created chimeric genomes resulting from a series of endosymbiotic events (Bhattacharya et al., 2004). This genomic diversity is apparent in the metabolic diversity of microbial eukaryotes that span trophic modes from photoautotrophy to heterotrophy. Many eukaryotes, including picoeukaryote species in the haptophyte, dinoflagellate, and ochrophyte lineages, are mixotrophs able to both conduct photosynthesis and consume other organisms (Caron et al., 2017, Stoecker et al., 2017). Eukaryote groups also differ in their macromolecular ratios of proteins, lipids, and nucleic acids carbohydrates (Finkel et al, 2016, Fiset et al., 2019), resulting in deviations in C:N:P ratios from the canonical 106:16:1 Redfield ratio (Geider & La Roche 2002) and latitudinal patterns in C:N:P ratios (Martiny et al., 2013). Eukaryote lineages also vary in trace metal utilization; the red plastid lineage (including ochrophytes, haptophytes, and some dinoflagellates) has a lower ratio of Fe, Zn and Cu to P than observed in the green plastid lineage (chlorophytes and other dinoflagellates) (Quigg et al., 2010).

Dinoflagellates, haptophytes, ochrophytes and chlorophytes are well-recognized lineages with notable contributions to the picoeukaryote size class in eukaryotic metatranscriptomes from the NPSG (Groussman et al., 2021) and globally (Carradec et al., 2019), and in global eukaryote metabarcoding surveys (de Vargas et al., 2015). Their contributions to the picoeukaryote community of the NPTZ has not been investigated in high resolution, and the genetic adaptations underpinning community shifts across the NPTZ are not well understood. Here, we use metatranscriptome data from three latitudinal cruise transects across this transition zone along 158°W in the Spring of 2016, 2017 and 2019 to resolve the molecular ecology of picoeukaryotes. Use of on-deck nutrient amendment incubation experiment were used to identify top responding species to the nutrient amendments and evaluate changes to their transcription of functional genes. Our investigations into the molecular mechanisms underlying the dominant populations of the transition zone provides insight into the niche specializations and flexible metabolisms granting ecological fitness to picoeukaryotes species of the dynamic North Pacific Transition Zone.

3.3 Results

3.3.1 Conditions of the North Pacific Transition Zone and general taxonomic trends

A total of 166 metatranscriptomes were collected from mixed-layer surface samples (7-15 m depth) on 3 latitudinal transects (Gradients 1 (2016), Gradients 2 (2017), Gradients 3 (2019) along 158°W that spanned the North Pacific Transition Zone (NPTZ, Fig 3.1a). Along these transects, temperature and salinity decrease with latitude (Fig 3.1b) while nutrient concentrations increase north of the ~32°N salinity front (Fig 3.1c) concomitant with biomass increases in the ~picophytoplankton size class (Fig 3.1d). Metatranscriptome samples were serially fractionated into small (0.2-3 μm) and large (3-100 or 200 μm) size fractions (Fig 3.2a, Supp Figs S3.1, S3.2, S3.3), to create a small picoplankton-dominated size class and a larger nanoplankton and microplankton size class. The metatranscriptomes were poly-A selected to select for transcripts derived from eukaryotes and transcriptional patterns were determined by mapping the resulting short sequence reads against the North Pacific Eukaryotic Gene Catalog (NPEGC, Appendix 1), which consists of 182 million assembled and annotated transcripts.

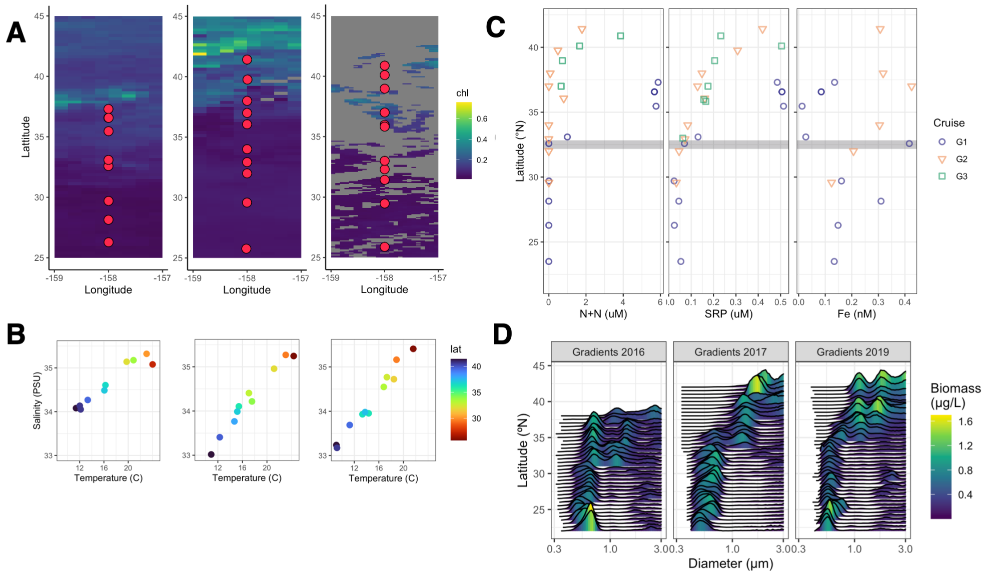


Figure 3.1. Environmental conditions of the Transition Zone Chlorophyll Front. A. Satellite estimates of average chlorophyll-*a* concentrations during cruise dates for the 2016 (left), 2017 (middle) and 2019 (right) Gradients cruises. 2016 and 2017 plots use CMEMES estimates (link), 2019 cruise uses MODIS estimates (link). The points on the maps indicate the locations of sample sites where metatranscriptomes were collected. **B.** Temperature and salinity plots of the three cruises (ordering the same as Fig 1a); the points are metatranscriptome sample sites colored by latitude. **C.** Surface concentrations of key nutrients. Left, nitrate + nitrite (N+N, left); middle, soluble reactive phosphate (SRP); right, iron (Fe). The gray line represents the approximate location of the salinity front (34.82 PSU) at ~32°N on all three cruises. For 2019, N+N and SRP measurements below 0.5 µM detection threshold not shown. **D.** Latitudinal biomass estimates in the 0.3-3.0 µm size class from the SeaFlow underway cytometer for the 2016 (left), 2017 (middle) and 2019 (right) cruises (adapted from Juranek et al., 2020).

Across all cruise samples and both size classes, five major marine microbial eukaryote groups accounted for more than half (56%) of the short reads mapped to taxonomically-classified environmental transcripts: dinoflagellates (27%, class 'Dinophyceae'), haptophytes (10%, phylum 'Haptophyta), ochrophytes (9.2%, clade 'Ochrophyta'), alveolates (6.8%, clade 'Alveolata'), and chlorophytes (3.2%, phylum 'Chlorophyta') (Supp Figs S3.1-S3.3). In the picoeukaryote-containing fraction (0.2-3 µm), these four groups accounted for 59% of mapped reads and the fractional composition was more evenly distributed across the dinoflagellates (16%), ochrophytes (16%), haptophytes (14%), alveolates (6.3%) and chlorophytes (5.9%). The proportion of sequences derived from animals, primarily copepods, was 8.9% and 5.2% of the total in the large and picoeukaryote fractions, respectively. Bacterial reads accounted for 0.4% and 0.6% of the total in the large and picoeukaryote fractions, respectively .

The changing latitudinal composition of the five most-abundant eukaryotic groups across the NPTZ is evident in the proportions of mapped metatranscriptome reads to dinoflagellates, ochrophytes, haptophytes and chlorophytes (Fig 3.2a, Supp Figs S3.1-3) relative to the total number of reads that could be mapped. In all three cruises, the southernmost stations (southward of 32°N) resemble the eukaryotic community of the North Pacific Subtropical Gyre (Pasulka et al., 2013, Hu et al., 2018, Groussman et al., 2021) with relatively high transcript abundances of dinoflagellates and mixotrophic haptophytes. Ochrophytes increased in the transcript proportion of the picoeukaryote metatranscriptomes in the southern region of the TZ (Fig 3.2a), while chlorophyte transcript inventory increases further north along the transects.

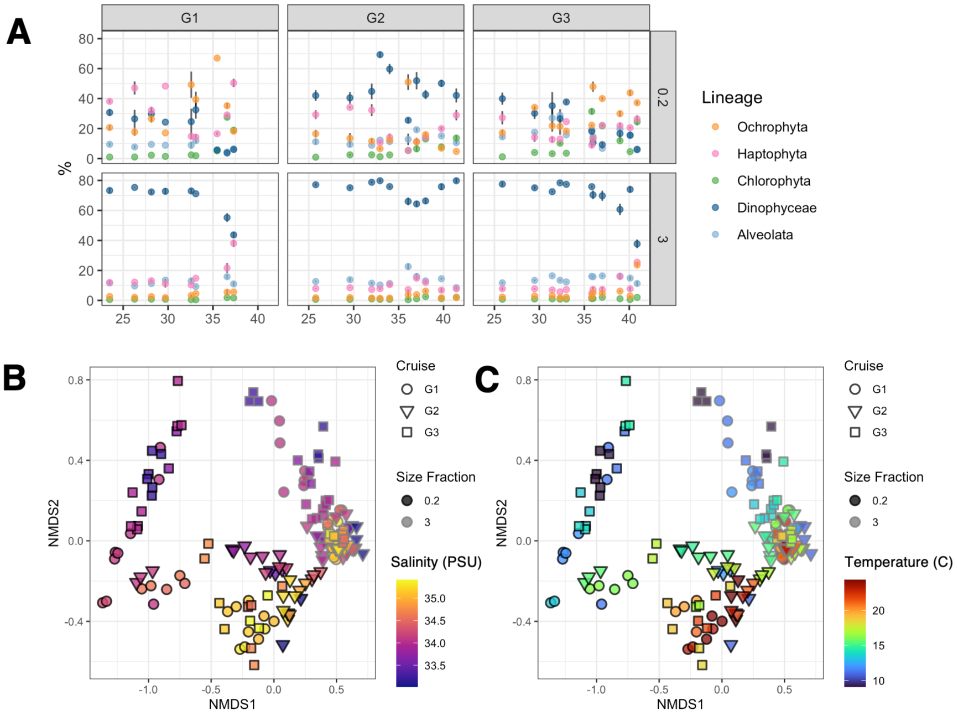


Figure 3.2. Broad taxonomy of large and small fraction eukaryotic metatranscriptomes. A. Proportions of total mapped environmental reads to dinoflagellate, ochrophyte, haptophyte, alveolate and chlorophyte taxa. Top panel labels indicate the cruise; G1, Gradients 1 (Apr 20-May 2, 2016); G2, Gradients 2 (May 27-Jun 11, 2017) and G3, Gradients 3 (Apr 8-28, 2019). Side panel labels indicate the sample size fraction (top row 0.2 – 3 µm, bottom row 3-100 µm (G2, G3) or 3-200 µm (G1)) **B-C.** Similarity of species transcript abundance in 166 metatranscriptome samples from Gradients 1, 2 and 3 from non-metric multidimensional scaling (stress = 0.11) of the fractional composition of reads mapped to species-level environmental bins. Each point indicates one metatranscriptome sample. Border color indicates size fraction; black, small fraction (0.2 – 3 µm); grey, large fraction (3-100 µm for G2 and G3, 3-200 µm for G1), shape indicates the cruise, fill color indicates *in situ* salinity (B) or temperature (C).

We constrained our analysis to the species-level to analyze the portion of the community transcript pool with the best-resolved phylogenetic annotation. Approximately one-third of the annotated assembled transcripts from the North Pacific Eukaryotic Gene Catalog are confidently assigned to species-level ranks, with the rest apportioned to higher ranks (Supp Fig S3.4). The fractional composition of short reads mapped to these species-level transcripts were used as a measure of similarity of eukaryotic communities across the three Gradients cruises (Fig 3.2b,c). Sixty-six eukaryotic species bins represented at least 0.2% of mapped reads in all samples. Non-metric multidimensional scaling (nMDS) was conducted on the fractional read abundances of these species. We excluded metazoa (animal) species from this analysis to focus on picoeukaryote species and reduce the effect of variability from metazoan transcripts. Samples were distinguished by size fraction, emphasizing the differences in community transcript composition between the picoeukaryotes and the larger plankton communities. The picoeukaryote-size fraction (0.2-3 μm) samples grouped along contours of similar temperature and salinity (Fig 3.2b,c) whereas the large-size fraction samples displayed less variability between samples.

3.3.2 Estimating coverage of species-level environmental metatranscriptome bins

We assessed the coverage of all 502 species bins within the 81 picoeukaryote-fraction samples to identify a core subset of picoeukaryotes with well-covered transcriptome bins. To estimate completeness of species-level bins between cruises, we identified the proportion of core transcribed genes (CTGs) in all 502 species bins (Fig 3.2a), using sets of core transcribed genes from references transcriptomes of the same lineage (Groussman et al., 2022, Scientific Data. *in prep*). Most species bins (464 species, 92% of total) had less than 50% coverage across the expected transcriptome, with only 38 species having 50% or more coverage (Fig 3.3a). On the high end of coverage, a small number species bins containing thousands of protein functions were sequenced deeply enough to detect transcripts associate with all expected gene functions. The similarity in coverage estimates between three cruises suggests the presence of a re-occurring community with interannual and seasonal variability.

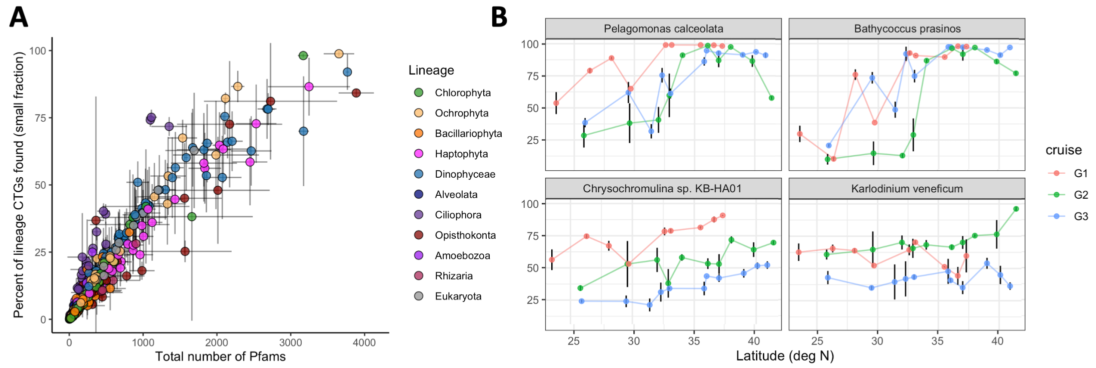


Figure 3.3. Sequencing depth of small fraction species-level metatranscriptome bins. A. Each dot indicates an environmental eukaryotic species bin from the small size-fraction (0.2 – 3 µm) Gradients metatranscriptomes, color indicates the eukaryotic lineage. X-axis, number of total non-redundant Pfam protein families identified across all cruise samples (average of three cruises). Y-axis, percent of core transcribed genes (CTGs) found (from Pfam annotations, average of three cruises). Error bars indicate standard deviation across three cruises. **B**. Completeness estimates within individual metatranscriptome samples in the small size fraction (0.2 – 3 µm) for a representative selection of four genera across the three cruises (see legend) based on the percent of CTGs detected in each sample.

To assess picoeukaryote species bin coverage within individual samples, we identified the percent of CTGs in small-fraction replicates across the three cruises (Fig 3.3b, Supp Fig S3.5). We focused on picoeukaryotes (protists in the small size-fraction) with a high completeness north of the salinity front at ~32°N that traditionally demarcates the northern boundary of the NPSG (Fig 3.1c, defined as the 34.82 salinity isohaline (Roden, 1971). The four microbial eukaryote species with the highest average completeness across all small-fraction cruise samples were the ochrophyte *Pelagomonas calceolata* (88.7%), the chlorophyte *Bathycoccus prasinos* (88.6%), the haptophyte *Chrysochromulina* sp. KB-HA01 (60.2%) and the dinoflagellate *Karlodinium veneficum* (58.8%) (Figure 3.3b). All three of these named species and the genus *Chrysochromulina* are recognized as cosmopolitan in literature (Guérin et al., 2021, Moreau et al., 2012, Li et al., 2000) and this widespread distribution is apparent in these North Pacific samples. Coverage patterns differ between species: some have dramatic shifts in coverage on either side of the ~32°N salinity front like *Pelagomonas calceolata* and *Bathycoccus prasinos,* while others like *Chrysochromulina* sp. KB-HA01 and *Karlodinium veneficum* have relatively even coverage on either side of the front (Fig 3.3b, Supp Fig S3.5). The differences in coverage between species on either side of the salinity front and across the transition zone reflect the difference in population abundance in these areas, and indicate varying levels of competitive fitness under the environmental conditions of these areas.

3.3.3 On-deck nutrient amendment incubation experiments

To identify potential mechanisms underlying the observed differences in picoeukaryote metatranscriptome composition and species coverage levels across the NPTZ, we conducted three on-deck nutrient amendment incubation experiments (REXP1-3) at 3 different latitudes on Gradients 2 (Fig 3.4a). At the northernmost REXP1 station at 41.42°N, *in situ* dissolved inorganic nitrogen (DIN) was 2 µM and iron (Fe) was 0.3 nM resulting in a molar N:Fe of 6.6 × 10^3 and $\log_{10}$N:Fe of 3.8. At the transition zone REXP2 station at 37.00°N, *in situ* nutrient concentrations were 0.06 µM DIN and 0.51 nM Fe resulting in a $\log_{10}$N:Fe of 2.1 At the southernmost REXP3 station at 32.93°N, *in situ* concentrations were 0.01 µM DIN and 0.22 nM Fe, resulting in a $\log_{10}$N:Fe of 1.7 (Fig 3.4 a, b). For REXP1, the *in situ* $\log_{10}$N:Fe of 3.8 was shifted to 2.9 (HiFe), 3.5 (LoFe) or 3.7 (NPFe) through amendments with different ratios of N, P, and Fe. In REXP2, the *in situ* $\log_{10}$N:Fe of 2.1 was shifted to 1.6 (Fe), 3.5 (NPFe) or 4.0 (NP). In REXP3, the *in situ* $\log_{10}$N:Fe of 1.7 was shifted to 3.4 (NP), 3.8 (NPFe), or 4.4 (NP) (Fig 3.4b). Triplicate metatranscriptome samples were collected from each treatment at T= 4 days and compared to T=0 (control) or the station sample.

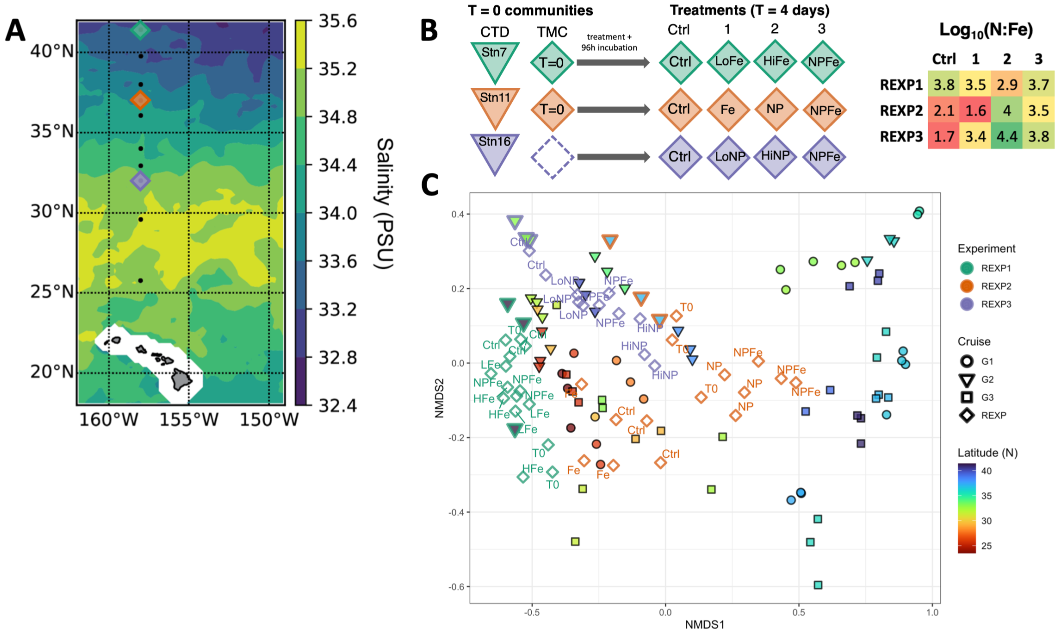


Figure 3.4. Gradients 2 resource ratio incubation experiments. A) Salinity contour map of the Gradients 2 transect. Black dots mark station sites where surface samples were collected; diamonds indicate stations where seawater water for incubation experiments was also collected. Cruise map adapted from Lambert et al., 2021. **B)** Schematic of the resource ratio incubations. The *in situ* communities

corresponding to the experiments (T=0 communities) were sampled from the CTD rosette, and for REXP1 and REXP2 additional T=0 samples were collected with trace metal clean (TMC) pumps immersed at 15m depth. The water for the experiments were collected from the TMC pumps and spike with nutrient amendments (treatments 1-3) or left as unamended controls (Ctrl), and maintained in temperature-controlled incubators for 96 hours. The ambient/adjusted $\log_{10}$(N:Fe) is shown for each of the experimental treatments. **C)** nMDS ordination of species transcript abundance in the small-fraction metatranscriptomes of Gradients 1, 2 and 3 the Gradients 2 on-deck resource ratio incubation experiments (REXP) (stress = 0.12). Each point is one metatranscriptome sample. Black-bordered points are surface samples, with inner color showing sample site latitude and shape indicating cruise. REXP samples appear as hollow diamonds colored by their experiment; and the corresponding station samples share the border color with their corresponding treatments.

The resulting short sequence reads from each REXP metatranscriptome were mapped to the annotated Gradients 2-derived contigs within the North Pacific Gene Catalog and aggregated into eukaryotic species bins. As observed with the size fractionated metatranscriptomes from surface samples (Fig. 3.2b, c), the 82 large- and picoeukaryote-size fractionated metatranscriptome samples from the three REXP experiments are distinct from each other (Supp Fig S3.6), with large-size fraction REXP samples closely grouped together, and the picoeukaryote sample communities displaying more dispersion. In particular, the picoeukaryote samples from the REXP2 experiment samples were more dispersed than the control samples, indicating a strong community transcriptional response. The three picoeukaryote communities between the three experiments are statistically distinct (PERMANOVA; $p < 0.001$) from each other (Table 3.1). Within the REXP2 and REXP3, the different nutrient amendments resulted in statistically significant differences in the species composition of mapped metatranscriptome reads (REXP2: $r^2 = 0.84$, Pr(>F) = <0.001; REXP3: $r^2 = 0.79$, Pr(>F) = <0.001). In contrast, the different nutrient amendments in REXP1did not result in statistically different community transcript inventories ($r^2 = 0.45$, Pr(>F) = 0.06).

Table 3.1. Permutational multivariate analysis of variance (PERMANOVA). Tests were conducted between the three REXP experiments as a whole, and for each of the treatment conditions in REXP1, REXP2, and REXP3.

Test: Differences between REXP1, REXP2, and REXP3 transcript composition

Source	Df	SumsOfSqs	MeanSqs	F.Model	r^2	Pr(>F)
Experiment	2	1.6838	0.84192	27.754	0.59362	0.001***
Residuals	38	1.1527	0.03033		0.40638	
Total	40	2.8365			1	

Test: Differences between REXP1 treatments

	Df	SumsOfSqs	MeanSqs	F.Model	r^2	Pr(>F)
Treatment	3	0.065313	0.0217711	2.2593	0.45865	0.06
Residuals	8	0.077089	0.0096362		0.54135	
Total	11	0.142402			1	

Test: Differences between REXP2 treatments

	Df	SumsOfSqs	MeanSqs	F.Model	r^2	Pr(>F)
Treatment	3	0.38077	0.126924	14.354	0.84333	0.001***
Residuals	8	0.07074	0.008842		0.15667	
Total	11	0.45151			1	

Test: Differences between REXP3 treatments

	Df	SumsOfSqs	MeanSqs	F.Model	r^2	Pr(>F)
Treatment	3	0.175778	0.058593	8.7972	0.79037	0.001***
Residuals	7	0.046623	0.00666		0.20963	
Total	10	0.222401			1	

To place the resulting REXP picoeukaryotic community transcriptional profiles within the context of surface picoeukaryotic community profiles across the three gradients communities, we created an nMDS ordination of Bray-Curtis dissimilarity of species transcript fractions in all 122 picoeukaryote samples: the 41 picoeukaryote metatranscriptomes derived from the REXP treatments and the 81 picoeukaryote metatranscriptomes derived from the Gradients 1, 2, and 3 surface samples (Figure 3.4c). The Gradients 1 and 3 picoeukaryote community transcriptomes were separated by their geography along the first ordination axis (NMDS1); low latitude samples with low N:Fe ratios (roughly corresponding to the ~32°N salinity front and southward of it), and high latitude samples with higher N:Fe ratios (~37°N and north, closer to the chlorophyll front) were closer together along NMDS1 with more distance across the second axes (NMDS2). Samples from intermediate latitudes between the salinity front and the chlorophyll front were closer to either the low-latitude or high-latitude samples, suggesting a sharp gradient in the transition between northern and southern communities from Gradients 1 and 3. Unlike the two earlier-season (April/May) cruises, the later-season (June) Gradients 2 samples from all latitudes were less dissimilar and grouped low along NMDS1 and high on NMDS2 (Figure 3.4c) and nearer to low-latitude samples from Gradients 1 and 3. Manipulating the N:Fe ratio of the origin communities in the REXP2 and REXP3 experiments shifted the balance of species transcript inventories (Figure 3.4c); higher N:Fe treatments resulted in a species transcript composition ordinated in the intermediate space between high-latitude and low-latitude communities, and treatments that lowered the N:Fe ratio resulted in lower values on the NMDS2 axis. Samples from REXP1 did not notably shift from their origin communities, consistent with the non-significance of results from PERMANOVA analysis (Table 3.1).

We assessed the relative proportion of species transcript inventories apportioned to dinoflagellate, haptophyte, ochrophyte and chlorophyte lineages in each sample (Supp Fig S3.7) to infer how community transcript inventories vary across the ordination space, and the influence of altered N:Fe ratios on this structure. Specific regions of the ordination space are associated with high proportions of transcript inventories in these four lineages. The Gradients 2 samples and southern-latitude Gradients 1 and 3 samples were proportionally higher in dinoflagellate-mapped transcripts (Fig 3.2a, Supp Fig S3.7a). The higher-latitude samples had a lower dinoflagellate proportion and are distinguished along the NMDS2 axis by either higher proportions of ochrophytes (Supp Fig S3.7c) or chlorophytes (Supp Fig S3.7d). Haptophytes displayed a different pattern of proportional distribution; samples with relatively high haptophyte proportions bridged the dinoflagellate, ochrophyte and chlorophyte-dominated communities, and had sample-specific hotspots of higher transcript abundance (Supp Fig S3.7d). Resource ratio experiments with an increased N:Fe treatment above *in situ* conditions were driven closer towards the ochrophyte- and chlorophyte-rich communities (Supp Figs S3.7c,d), and lowered N:Fe treatments resulted in a sample-specific high haptophyte proportions (Supp Fig S3.7b).

To understand how environmental parameters explain the species transcript fractions of these surface and experimental metatranscriptome, we fit environmental vectors onto the combined picoeukaryote sample ordination (Supp Fig S3.8, Table 3.2). Temperature, nitrate, phosphate and the N:Fe ratio had strong significant correlations to the structure (Pr(>r) < 0.001), iron had a correlation at a weaker level of significance (Pr(>r) = 0.029), and salinity and latitude had no significant correlation. The significantly-correlated environmental parameters align with the communities characterized by relatively high proportions of dinoflagellates, ochrophytes or chlorophytes (Supp Fig S3.7); dinoflagellates are more prominent in high temperatures and lower N:Fe samples (log10(N:Fe) < 3), ochrophytes with higher N:Fe (log10(N:Fe) > 3), and chlorophytes with high nitrate and phosphate concentrations and log10(N:Fe) > ~2 (Supp Fig S3.8). Haptophyte populations appear less sensitive to N:Fe as relative haptophyte abundance spans large gradients in N:Fe across the transition zone.

Table 3.2. Correlation of environmental parameters to species ordination. Latitude is from sample site, CTDTMP is temperature (°C) from the CTD rosette, CTDSAL is salinity (PSU), NO3_uM are nitrate concentrations (µM), PO4_uM is phosphate concentration (µM), Fe_nM is iron concentration (nM), and NtFe is the N to Fe ratio (N:Fe).

	r^2	**Pr(>r)**	
LATITUDE	0.1164	0.085	
CTDTMP	0.3332	0.001	***
CTDSAL	0.0486	0.366	
NO3_uM	0.6087	0.001	***
PO4_uM	0.406	0.001	***
Fe_nM	0.163	0.029	*
NtFe	0.3842	0.001	***

3.3.4 Responder species in the REXP incubations

To identify which specific species displayed significant changes in transcript inventories in the incubation experiments, we conducted two-tailed t-tests of the relative species transcript inventory in the triplicate treatment bottles to the un-amended control bottles, followed by a Benjamini-Hochberg correction for multiple comparisons. A total of 108 significant species-treatment pairs were discovered; 23 species-treatment pairs with significant increases in community read fraction and 85 species-treatment pairs with significant decreases (Figure 3.5, Table 3.3). Twelve species accounted for the significant increases in transcript pools, and seven of these twelve species had significant transcript fraction increases in more than one treatment, accounting for 18 of the 23 significant increases. *Pelagomonas calceolata* was the most frequent significant responder with increased transcript fractions in three NP treatments and two NPFe treatments. Dinoflagellate species diminished in fractional abundance *en masse*, with twice as many significantly decreased species in the NP treatments as in Fe and no significant increases in any treatment. Alveolates similarly decreased without no significant increases in any treatment.

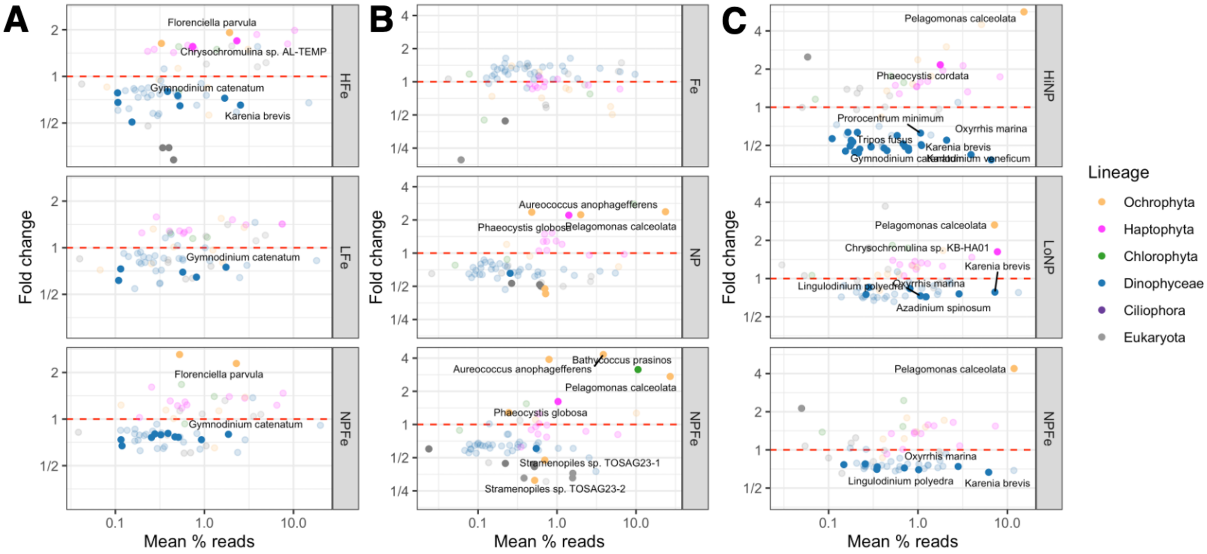


Figure 3.5. Changes in picoeukaryote species abundance following nutrient amendments. Subpanels are labeled by treatment. Horizontal axis is $\log_{10}$-transformed average of species read fraction in small-fraction triplicate bottles, vertical axis is $\log_2$-transformed fold change of read fraction vs control bottles; the vertical axes between A, B, and C are scaled separately. Species are indicated by dots, colored by their lineage (the 'Eukaryota' category includes species from lineages other than the ones shown). Species with statistically different changes are plotted in darker colors; those with non-significant changes are paler. Significantly different species with an average read percent greater than 1% are labeled. **A)** REXP1, **B)** REXP2, **C)** REXP3.

Table 3.3. Summary of significant responder species to REXP treatment by lineage. For each lineage, values in the table indicate the number of significantly-different species in the treatments. Rows are labeled by directionality; 'up' for increasing abundance and 'down' for decreasing.

			REXP1			**REXP2**			**REXP3**	
Lineage	**Direction**	**LFe**	**HFe**	**NPFe**	**Fe**	**NP**	**NPFe**	**LoNP**	**HiNP**	**NPFe**
Ochrophyta	up		2	2		3	4	1	1	1
Haptophyta	up		2			1	1	1	1	
Chlorophyta	up						1			
Other	up								1	1
Dinophyta	down	5	8	10		1	1	7	26	8
Alveolata	down		3		1	3	3			
Rhizaria	down						1			
Ochrophyta	down					2	2			
Other	down				1		1			

While the overall REXP1 community did not show significant differences following treatment (Fig 3.4c, Table 3.1), two ochrophytes species (*Florenciella parvula* and *Florenciella* sp. RC1007) and two haptophyte species (*Chrysochromulina* sp. AL-TEMP and *Chrysochromulina rotalis*) had significant increases in their respective read fractions compared to the control bottles (Fig 3.5a); the *Florenciella* species increased in the high Fe and NPFe treatments and the *Chrysochromulina* species under the high Fe condition only. The low Fe treatment did not have any significantly increased species.

The REXP2 experiment stimulated significant increases in the community transcript proportion of ochrophyte species in the NP and NPFe treatments (Fig 3.5b): *Pelagomonas calceolata*, *Pelagococcus subviridis*, and *Aureococcus anophagefferens* increased transcript inventories in both treatments, along with the haptophyte *Phaeocystis globosa*, while *Florenciella* sp. RC1007 and the chlorophyte *Bathycoccus prasinos* increased transcripts under NPFe only. There were no positive responses to the Fe treatment.

In REXP3, *Pelagomonas calceolata* increased in transcript proportion under all three treatment conditions (Fig 3.5c). *Phaeocystis globosa* and *Chrysochromulina* sp. KB-HA01 increased under high NP and low NP, respectively, and the prasinodermophyte green alga *Prasinoderma singulare* increased in NPFe and high NP conditions.

3.3.5 Functional transcriptome response to nutrient amendment in responder species

The REXP incubation experiments revealed a set of species bins that were responsive to manipulated N:Fe ratios, in some instances doubling their transcript inventory over the four experiments (Fig 3.5, Table 3.3). We examined the individual transcriptional response of the 12 significant responder species to the nutrient amendments (Fig 3.6, Table 3.4), and determined statistically significant fold changes in the transcript abundance of Pfam protein families (Mistry et al., 2021) from the metatranscriptome functional annotations included in the North Pacific Eukaryotic Gene Catalog (Appendix 1). Two-tailed *t*-tests with a multiple comparison p-value correction identified a total of 279 statistically significant protein families with differential transcript abundance in the 12 species, with a false discovery rate of 10%. Species in the REXP2 (Fig 3.6a-f) and the REXP3 (Fig 3.6g-i) experiments had the greatest number of significantly different functions, and the number of significantly different functions was higher in some species than others.

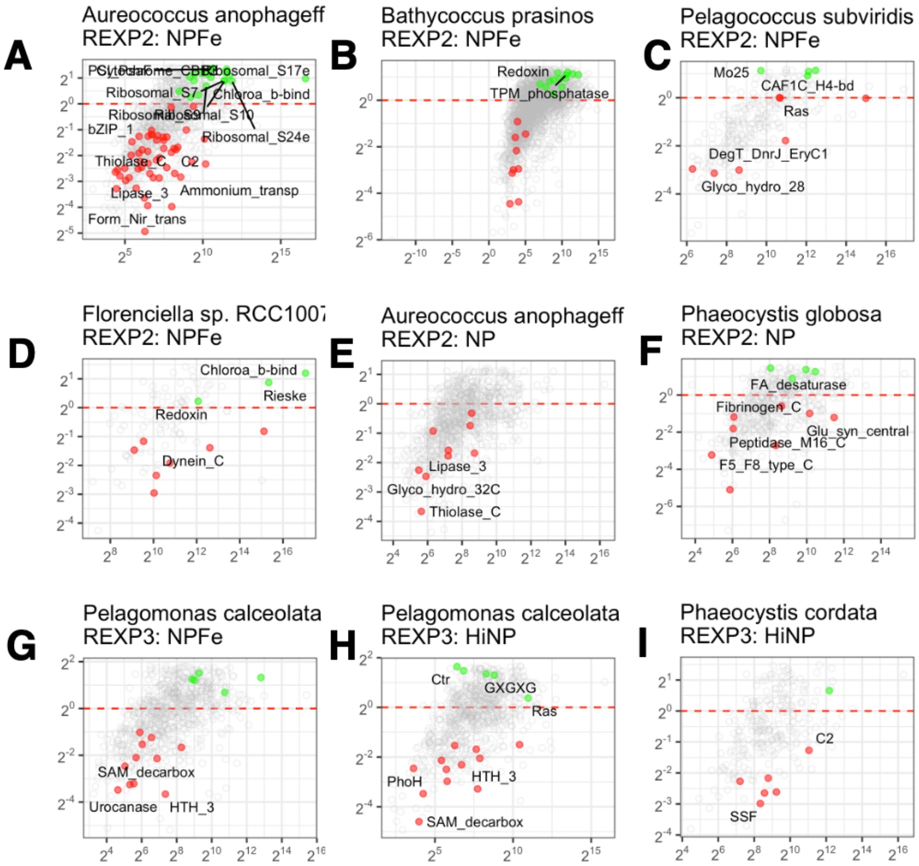


Figure 3.6. Changes in functional transcript abundance in responder species following nutrient amendment in the resource ratio experiments. Seven responder species with the greatest number of significantly-different Pfam protein families are shown here by experiment and treatment. Horizontal axis is $\log_2$-transformed mean transcripts per million species reads (TPM) in the treatment, vertical axis is $\log_2$-transformed fold change of treatment against the control. Significantly different Pfam functions are shown in green (increasing) and red (decreasing), and Pfams with non-significant changes are hollow grey circles. Pfam protein families discussed in the text are labeled.

Table 3.4. Summary of differentially-transcribed Pfam protein functions in 12 responder species. Species, lineage, NCBI, and directionality (dir) of significant Pfam protein families under all Fe, NP, and NPFe treatments, and the total up-regulated or down-regulated Pfams.

Species	Lineage	taxID	dir	Fe	NP	NPFe	total
Aureococcus anophagefferens	Ochrophyta	44056	up	0	5	38	43
Aureococcus anophagefferens	Ochrophyta	44056	down	0	9	49	58
Florenciella parvula	Ochrophyta	236787	up	0	0	1	1
***Florenciella* sp. RCC1007**	Ochrophyta	464225	up	0	5	5	10
***Florenciella* sp. RCC1007**	Ochrophyta	464225	down	0	3	10	13
Pelagococcus subviridis	Ochrophyta	35679	up	3	0	7	10
Pelagococcus subviridis	Ochrophyta	35679	down	0	2	4	6
Pelagomonas calceolata	Ochrophyta	35677	up	0	7	7	14
Pelagomonas calceolata	Ochrophyta	35677	down	0	14	11	25

Chrysochromulina rotalis	Haptophyta	412157	up	0	2	3	5
Chrysochromulina rotalis	Haptophyta	412157	down	1	3	0	4
***Chrysochromulina* sp. AL-TEMP**	Haptophyta	2802064	up	0	2	1	3
***Chrysochromulina sp.* AL-TEMP**	Haptophyta	2802064	down	0	1	1	2
***Chrysochromulina sp.* KB-HA01**	Haptophyta	2802065	up	0	1	0	1
***Chrysochromulina sp.* KB-HA01**	Haptophyta	2802065	down	0	1	1	2
Phaeocystis cordata	Haptophyta	118079	up	2	2	1	5
Phaeocystis cordata	Haptophyta	118079	down	1	6	3	10
Phaeocystis globosa	Haptophyta	33658	up	1	9	2	12
Phaeocystis globosa	Haptophyta	33658	down	0	8	4	12
Prasinoderma singulare	Eukaryota	676789	up	1	0	0	1
Prasinoderma singulare	Eukaryota	676789	down	0	1	1	2
Bathycoccus prasinos	Chlorophyta	41875	up	1	0	24	25
Bathycoccus prasinos	Chlorophyta	41875	down	1	3	11	15

The NPFe treatment in REXP2 elicited significant increases in species transcript inventory and the greatest number of differentially expressed protein families (Table 3.4, Fig 3.5, Fig 3.6). Under the REXP2 NPFe treatment, *Aureococcus anophagefferens* was the top responder with over a 4-fold increase in transcript inventory (Fig 3.5a) and 85 differentially expressed protein families (Fig 3.6a), the greatest number in any species under any treatment (Table 3.4). *A. anophagefferens* significantly increased transcription of Pfam protein families involved in photosystem proteins and chlorophyll biosynthesis (PSI_PsaF, Chloroa_b-bind), ribosomes (Ribosomal_S17e, Ribosomal_S24e, Ribosomal_S10, Ribosomal_S7, Ribosomal_S9), and redox processes and oxidative stress (Cytochrome_CBB3, Thioredoxin_6, Redoxin), and significantly decreased transcripts involved in fatty acid degradation (Lipase_3), nitrogen transport (Form_Nir_trans) and ammonium transport (Ammonium_transp). In the same treatment, *Bathycoccus prasinos (*Fig 6b) increased transcription of Pfam protein families involved in photosystem proteins (TPM_phosphatase), chlorophyll biosynthesis (ALAD, Porphobil_deam), the TCA cycle (Ldh_2), the ribulose phosphate pathway (Rib_5-P_isom_A), redox control (Redoxin), and fatty acid biosynthesis (ACC_central). In *Pelagococcus subviridis* (Fig 3.6c), notable increased transcript functions are associated with cell growth and cytokinesis (Mo25, Ras) and histone binding (CAF1C_H4-bd). Decreased functions include glycoside hydrolases (Glyco_hydro_28), aminotransferases (DegT_DnrJ_EryC1) and aconitase (Aconitase_C). *Florenciella* sp. RCC1007 had increases in functions related to electron-transfer chains (Rieske), photosystem complexes (MSP) and chlorophyll biosynthesis (Chloroa_b-bind) and decreases in flagella-associated proteins (Dynein_C) (Fig 3.6d).

The importance of the N:Fe ratio in the REXP2 community is underscored by the relative lack of significantly different protein families in the NP-only treatment. *Aureococcus anophagefferens* had smaller set of differentially transcribed functions under the NP-only treatment, down-regulating transcripts involved in fatty acid degradation (Thiolase_C, Lipase_3) and glycoside hydrolases (Glyco_hydro_32C) (Fig 3.6e). *Phaeocystis globosa* had significantly increased transcription of genes encoding ribosomal proteins (S1, Ribosom_S12_S23, Ribosomal_L1), DNA binding (bZIP_1), fatty acid biosynthesis (FA_desaturase, PP-binding) along with decreases in NH_4^+ assimilation (Glu_syn_central), and extracellular glycoprotein and phospholipid processes (Peptidase_M16_C, Fibrinogen_C, F5_F8_type_C) (Fig 3.6f).

In the southern experiment proximal close to the salinity front (REXP3), *Pelagomonas calceolata* was the responder with the highest fold change increase in community read fraction of the HiNP and NPFe treatments, both of which raised the N:Fe conditions from an ambient $\log_{10}$(N:Fe) of 1.7 to 4.4 and 3.8, respectively (Fig 3.5b,c) though they did not share sets of up-regulated functions. In the higher N:Fe treatment (hiNP, $\log_{10}$(N:Fe) = 4.4), *P. calceolata* increased transcripts involved in copper uptake (Ctr), ammonia assimilation (GXGXG), and a peroxisome-associated function (Mpv17_PMP22) in the NPFe treatment ($\log_{10}$(N:Fe) = 3.8). *P. calceolata* shared sets of down-regulated processes under these increased N:Fe conditions, including transcriptional regulators (HTH_3) and spermidine biosynthesis (SAM_decarbox), while decreased protein families in HiNP only included Pfam functions associated with phosphate starvation (PhoH), and decreased functions in NPFe included histidine degradation (Urocanase). In this experiment, *Phaeocystis cordata* was a significant responder to the HiNP but not the NPFe treatment. In the former, this species upregulated protein functions involved in porphyrin biosynthesis (URO-D) and decreased functions involved in membrane trafficking and transporter families (C2, SSF).

3.4 Discussion

The dynamic and productive North Pacific Transition Zone region hosts picoeukaryote communities that are distinct from the subtropical gyre to the south and the subarctic gyre to the north. We used a combination of survey studies and nutrient amendment incubations to examine the molecular ecology of these organisms . We relied on changes in transcript abundances within species bins to infer how communities respond to different conditions and nutrient amendments, recognizing that identification of a species bin depends on availability of diverse reference sequences for taxonomic annotation. We therefore developed and used for this study an updated reference library that includes reference species isolated from the study region (Groussman et al., 2022). In addition, our analyses rely on compositional changes, which are sensitive to cellular mRNA pool sizes differ between eukaryote lineages. For example, dinoflagellates appear to transcribe most of their genes at all times, relying on post-transcriptional regulation to modulate their proteome (Wisecaver and Hackett 2011, Bowazolo et al., 2022), which may decreases the proportional signal of other lineages. In contrast, the streamlined genomes and reduced gene content of small picoeukaryotes such as *Bathycoccus prasinos* (Moreau et al., 2012) result in a relatively small fraction of the eukaryotic metatranscriptome pool. We focused our analyses on only those species bins that displayed significant increases or decreases in the community transcript inventories, noting that most species within a given treatment maintained relatively constant. Using these approaches, we identified 12 picoeukaryote species with significantly increased transcript inventories after manipulating the dissolved N:Fe. A majority of these responder species are known osmo-mixotrophic and phago-mixotrophic ochrophytes and haptophytes that can adjust their metabolic mode to adapt to changing environmental conditions, which is expected to be advantageous in the dynamic NPTZ where the seasonal dynamics of the chlorophyll front and mixing of different water masses expose communities to gradients in biochemical and physical conditions.

On-deck nutrient incubation experiments tested the community response to altered N:Fe ratios with at three sites in different regions of the NPTZ (Fig 3.3a). Across treatments, the responder species fall into four general categories. Five of the responder species appear to increase in relative transcript abundance (as measured by an increase in %CTG coverage) across the three Gradients transects: *Bathycoccus prasinos*, *Pelagomonas calceolata*, *Aureococcus anophagefferens*, *Prasinoderma singulare* and *Chrysochromulina* sp strain KB-HA01 (Fig. 3.3b, Fig. S3.5). Three responder species appear to maintain relatively constant proportions across the Gradients transects: *Phaeocystis globosa*, *Pelagomonas subviridis*, and *Chrysochromulina* sp AL-TEMP (Fig. S3.5), one species decreased in abundance across the transects (*Phaeocystis cordata*) and three species that had less than 50% average %CTG coverage on the Gradients transects and yet significantly responded to nutrient amendments: *Florenciella parvula, Florenciella* sp. RCC1007 and *Chrysochromulina rotalis* (Fig. 3.5). Station 11 had the highest *in situ* surface Fe levels measured on both G1 and G2 transects (0.43 nM), and the Fe addition treatment elicited no significant response in species. There is evidence that aeolian dust deposition from Asia contributed to the high iron levels in the NPTZ in 2017, forming a local Fe maximum around 35°N and stimulating production (Pinedo-González et al., 2020), explaining the high iron levels in the REXP2 source water. Interestingly, isotopic analysis suggested that ~20-60% of this Fe dust deposition is attributed to anthropogenic emissions from East Asia, providing an intriguing connection between human activity and productivity in the NPTZ. The high iron concentrations may have primed the community for the response to NP amendment here after alleviation of N-stress.

Across all experiments, ochrophyte species showed significant transcriptional responses to incubation experiments with important distinctions between the pelagophyte and dictyophyte subclades. The pelagophyte species *Pelagomonas calceolata*, *Pelagococcus subviridis,* and *Aureococcus anophagefferens* were responders had among the highest fold change in treatments that increased N:Fe above ambient levels, while the phago-mixotrophic dictyophytes (another clade of ochrophyta) *Florenciella* sp. RCC1007 and *Florenciella parvula* responded to treatments that lowered N:Fe from ambient conditions. The distinction between pelagophyte and dictyophyte responses to NP and Fe amendments underlines some of the importance differences between these ochrophyte subclades.

Pelagomonas calceolata is a cosmopolitan picoeukaryote species in global oceans (Worden et al., 2012) and a higher relative abundance in high-temperature, low-light and iron-poor conditions attributed to genomic adaptations to low-iron environments (Guérin et al., 2022). Our results confirm the niche adaptation of *Pelagomonas calceolata* to low-iron environments and highlight it as an important responder species to nutrients pulses that result in increased N:Fe. The transcriptional responses of *P. calceolata* to these treatments shows adjustment to their metabolic strategies. In the NP-containing amendments, *P. calceolata* increased transcripts involved in copper uptake, which could be related to low-iron adaptations in ferroprotein equivalents. In reports from another study, this species up-regulates genes under iron poor conditions involved in iron uptake (phytotransferrin, *ISIP2A,* Zn/Fe permease), storage (*ISIP3*), and non-ferrous substitutions for iron-requiring proteins (flavodoxin) (Guérin et al., 2022). Along with most ochrophytes outside of the diatom lineage (Groussman et al., 2015), *P. calceolata* lacks the iron-storage protein ferritin. Additionally, decreases in phosphate-starvation biomarkers indicate alleviation of P stress. The increase in ammonia assimilation protein families would allow *P. calceolata* to take advantage of the enriched nitrogen conditions. This is consistent with the observation by Dupont et al., (2014) that *P. calceolata* had highest proportion of nitrogen-assimilation genes among eukaryotes in the subsurface chlorophyll maximum in the eastern subtropical Pacific, leading the authors to suggest a major role in nitrate assimilation and new production. This species upregulates genes for inorganic nitrogen uptake under low-N conditions and has a higher number of enzymes involved in NH_4^+ assimilation (GS/GOGAT pathway) than other species (Guérin et al., 2022). Organic nitrogen sources may be an important source of nitrogen for *P. calceolata*, which appears to switch between reliance on organic nitrogen sources and inorganic nitrogen depending on N conditions (Guérin et al., 2022, Kang et al., 2021). *P. calceolata* had the highest average coverage of all species north of the salinity front in the G1-3 transects and notable responses to the nutrient amendment experiments, indicating the importance of this species across the transition zone across and its ability to thrive under higher N:Fe.

We identified *Aureococcus anophagefferens* as another pelagophyte responder species with some of the highest fold changes under increased N:Fe treatments in REXP2, and the greatest number of significantly different protein families. *A. anophagefferens* is a phototrophic picoeukaryote (2–3 μm) known for forming massive blooms (‘brown tides’) in response to high inorganic and organic nitrogen concentrations (Pustizzi et al., 2004). They appear to prefer reduced forms of nitrogen, in particular ammonia and dissolved free amino acids (Mulholland et al., 2004) and also uptake dissolved organic carbon and phosphorus (Gobler et al., 2004). It would appear that this blooming behavior is also a response to higher nitrogen and N:Fe ratios in the NPTZ. Here, *A. anophagefferens* significantly increased transcription of protein families involved in photosystem proteins and chlorophyll biosynthesis, ribosomal proteins, redox processes and oxidative stress, and significantly decreased transcripts involved in fatty acid degradation, hydratases and nitrogen and ammonia uptake. This enhancement of their photosynthetic machinery and support functions are most likely essential in powering their bloom behavior, while they de-emphasized catabolic functions and nutrient uptake. Coverage of *A. anophagefferens* transcript inventories is lower in the southern NPTZ and peaks in the north at the approximate location of the chlorophyll front. The patterns of surface community distribution and response to high N:Fe treatments suggests a role for this bloom-forming species in the uptake of nitrogen under increased N concentrations supported by increased photosynthetic capacity.

Unlike the 'phototrophic' pelagophytes, *Florenciella* sp. RCC1007 and *Florenciella parvula* are mixotrophic ochrophyte species from the dictophyte subclade. Along with prymnesiophytes (haptophytes), species of the genus *Florenciella* are grazers of picocyanobacteria and heterotrophic bacteria (Frias-Lopez et al., 2009, Li et al., 2021); prey ingestion alleviates nutrient limitation and allows for faster growth, and they exhibit a faster grazing rate under nutrient limitation (Li et al., 2021). *Florenciella parvula* had the highest fold change among species under the lowered N:Fe treatments in the already iron-rich REXP1 seawater, and their fitness here is likely explained by their ability to supplement macronutrients through grazing. *Florenciella* sp. RCC1007 was a significant responder under the REXP2 NPFe treatment that increased N:Fe, and their transcriptional response demonstrates that they shift along the mixotrophy axis towards phototrophy when the N:Fe ratio is higher; indicated by significant increases in functions related to photosystem complexes and chlorophyll biosynthesis. The decrease in a flagella-associated protein family could mean that they decrease their motility under higher NP conditions as well. *Florenciella* is recognized as an important species in the NPSG (Li et al., 2021, Groussman et al., 2021), but neither *Florenciella* species has over 50% average coverage north of the salinity front in the G-3 surface samples (through %CTGs). Despite their relatively low abundance in the transition zone, the positive response of *Florenciella* species to raised or lowered N:Fe could reflect a niche specialization of these species to exploit changing nutrient conditions through rapid metabolic remodeling.

The pelagophyte and dictyophyte responder species discussed above are mixed-growth strategists capable of balancing photoautotrophy and separate types of mixotrophy. Mixotrophy is defined as the combination of phototrophy with the phagocytosis and catabolism of prey (Mitra et al., 2016, Flynn et al., 2019), while osmo-mixotrophy refers to the uptake and assimilation of endogenous organics. The pelagophyte species *Pelagomonas calceolata* and *Aureococcus anaphagefferens* are capable of thriving on dissolved organic N and P (Guérin et al., 2022, Mulholland et al., 2004, Gobler et al., 2004), suggesting the possibility of osmo-mixotrophic metabolism. Both strategies would contribute to the flexibility of these ochrophytes in responding rapidly to changing nutrient conditions, through alleviation of N and P stress through direct uptake of organic nutrients or prey phagocytosis in low inorganic nutrient conditions and enhancement of photosynthetic machinery under high inorganic nutrients.

The other notable group of positive responder species were haptophytes; and they are also distinguished by different trophic strategies. There were three species in the genus *Chrysochromulina* (*C. rotalis*, AL-TEMP, and KB-HA01) and two in the genus *Phaeocystis* (*P. globosa* and *P. cordata*). *Chrysochromulina* species are known to be constitutive phago-mixotrophs and have been shown to increase ingestion rate under P-starvation (Jones et al., 1993) and used as model mixotrophs to explore optimal foraging theory (Stibor and Sommer 2003). Metatranscriptome analysis from a subset of the data used here proposed that *Chrysochromulina* species shift gene transcription over the Gradients 1 transect between these two metabolisms to optimize environmental fitness (Lambert et al., 2021); with a more-heterotrophic metabolism in the nutrient-limited subtropical gyre and a more-photoautotrophic metabolism northward across the NPTZ where direct uptake of abundant nutrients supports the shift towards autotrophy. In the northern REXP1 experiment, both *Chrysochromulina* species AL-TEMP and *C. rotalis* had significant positive increases under the iron fertilization treatment. The only other significant responders in REXP1 were the two *Florenciella* species. These responses by phago-mixotrophic species under the low N:Fe conditions on REXP1 would indicate that phago-mixotrophy allows both genera to thrive through prey-derived macronutrient supplementation. Unlike the other two *Chrysochromulina* species, *Chrysochromulina* sp. KB-HA01 was a significant responder to only the REXP3 LoNP amendment, although abundance increased under all treatments compared to the control, and it's possible this shifted KB-HA01 towards a more-phototrophic metabolism. Coverage of *Chrysochromulina* sp. KB-HA01 transcript functions was high in the NPSG and across the NPTZ, and the metabolic flexibility of this species likely contributes to its abundance across different nutrient conditions.

The other haptophyte responder genus, *Phaeocystis*, is characterized by flagellated cells that can form colonies held together by mucilaginous secretions. *Phaeocystis antarctica* shifts from flagellated cells to the colonial form under higher iron conditions (Bender et al., 2018). Brisbin and Mitarai (2019) observe transcriptional shifts in *P. globosa* during colony formation such that most processes are down-regulated except those involved in colony formation and DMSP production, suggesting a defensive role for colony formation. *P. globosa* was a significant responder in increased N:Fe treatments in REXP2, and significantly increased transcription of genes encoding ribosomal proteins and fatty acid biosynthesis, along with decreases in NH_4^+ assimilation and extracellular matrix processes. This could reflect an increased proportion of free-living cells relative to the colonial form following increased N:Fe conditions.

The two other species with significant community transcript increases in these experiments were the chlorophyte *Bathycoccus prasinos* and the prasinodermophyte *Prasinoderma singulare*. *Prasinoderma* is the only genus in the prasinodermophyte lineage of green alga. This marine picoeukaryote has recently been placed in a novel third phylum in Viridiplantae, having diverged from the common ancestor of chlorophytes and land plants (Li et al., 2020). *Bathycoccus* is a small picochlorophyte in the order Mamiellales; this group diverged early from other chlorophytes in evolutionary history and are among the smallest free-living eukaryotic autotrophs (around 1 to 2 μm) (Moreau et al., 2012). In the REXP experiments, *Bathycoccus* was only significantly increased under the REXP2 NPFe treatment, showing increased transcriptions of gene families involved in photosystems, chlorophyll biosynthesis, the ribulose phosphate pathway and fatty acid biosynthesis. *Bathycoccus* species are abundant in nutrient rich waters (Vannier et al., 2016), and this treatment likely provided them a more suitable growth environment.

Dinoflagellates in this study are contrasted by their high community transcript proportion and lack of significant response to the REXP treatments. Dinoflagellates were higher in picoeukaryote transcript composition on the 2017 Gradients 2 transect than the other two years, and they were the dominant population by transcript composition in the resource ratio experiments. The station samples for the southern experiment, at the edge of the subtropical gyre, had the highest proportion of dinoflagellate transcripts among all small fraction metatranscriptomes in this study. Collected in mid-June, these samples are the most separated from winter's wind-driven delivery of northern nutrients (Ayers and Lozier, 2010). Many dinoflagellates are also mixotrophic (Stoecker et al., 2017, Caron et al., 2017), and can shift the balance of their trophic modes under different N:Fe conditions (Cohen et al., 2021). The competitive advantage of phago-mixotrophic dinoflagellates in low N:Fe conditions is apparent from their high transcript abundance in low N:Fe samples. The absence of any significant increases in species transcript proportions under the REXP treatments is likely due to the substantially lower growth rates of dinoflagellates compared than other taxa (Tang 1996). Conversely, the higher N:Fe amendments appear to have shifted the balance of the community away from a dinoflagellate-dominated equilibrium.

3.5.2 Conclusions

Our findings illuminate a subset of the southern NPTZ picoeukaryote community that is poised to respond quickly to sporadic nutrient pulses that create perturbations in the ratio of key nutrients. These responders are primarily osmo-mixotrophic and phago-mixotrophic ochrophytes and haptophytes with the ability to shift their metabolic mode to adapt to changing environmental conditions. This is advantageous in the dynamic NPTZ, where the seasonal dynamics of the chlorophyll front and mixing of different water masses expose communities to gradients in biochemical and physical conditions. These responder species occupy different niches in the NPTZ community. *Pelagomonas calceolata, Aureococcus anophagefferens* and other osmo-mixotrophs are situated to take advantage of dissolved N and P inputs, and *P. calceolata* is particularly suited to high N:Fe ratios. Phago-mixotrophs like *Florenciella* and *Chrysochromulina* are more tolerant of lower N:Fe ratios, likely through alleviation of nutrient deficiencies by predation. Distinct from the core communities of the subtropical gyre and subpolar gyre that bound it, the North Pacific transition zone hosts a community of metabolically flexible picoeukaryotes ready to respond within days to changing conditions in this dynamic ecotone.

3.5 Methods

3.5.1 Metatranscriptomes from the North Pacific Eukaryotic Gene Catalog

Assembled eukaryotic metatranscriptomes and associated taxonomic and functional annotations and quantifications were derived from the North Pacific Eukaryotic Gene Catalog. A description of the Catalog and assembly, annotation and quantification methods are available in Appendix 1. Methods, data and code from the NPEGC will be publicly available in GitHub and Zenodo repositories prior to publication in Scientific Data. A subset of that data, the 166 surface metatranscriptomes from the Gradients 1, 2 and 3 cruises, is used in this study.

3.5.2 On-deck incubation experiments

We conducted trace metal clean, temperature-controlled incubation experiments on the 2017 Gradients 2 cruise with seawater from three sites as previously described (Boysen 2020, Lambert et al., 2021). Triplicate T=0 communities were filtered shortly collection from the trace metal pump. Twenty liters of seawater were collected using a trace-metal clean pump at 15m depth into triplicate polycarbonate carboys and incubated at *in situ* temperature for 96 hours in on-deck, temperature-controlled incubators screened with 1/8-inch light blue acrylic panels to approximate *in situ* light levels at 15 m. Triplicate carboys were amended with respective treatments for each experiment at $t = 0$, and triplicate carboys with no amendment served as a control. Resource ratio experiment 1 (REXP1) used water collected from Station 7, the northernmost station on Gradients 2 (latitude 41.42°N, ambient water temperature 10.9°C). Two iron amendments were given: a low-iron pulse (LoFe; 0.3 nM Fe), a high-iron pulse (HiFe; 2 nM Fe), and a combined N+P+Fe pulse (NPFe; 2 nM Fe + 10 µM NO_3^- + 1 µM PO_4^{3-} added). Resource ratio experiment 2 (REXP2) used water collected from Station 11 (latitude 37.00°N, ambient water temperature 15.2°C) and were given one iron amendment (Fe; 1 nM Fe), an N+P amendment (NP; 0.5 µM P + 5 µM NO_3^-) and an N+P+Fe amendment (NPFe; 1 nM Fe + 5 µM NO_3^- + 0.5 µM PO_4^{3-} added). Resource ratio experiment 3 (REXP3) was collected proximal to the Gradients 2 salinity front (Latitude 32.93°N, ambient water temperature 17.1°C), and amended with a low N+P pulse (LoNP, 0.5 µM NO_3^- + 0.05 µM PO_4^{3-}), a high N+P pulse (HiNP, 5 µM NO_3^- + 0.5 µM PO_4^{3-}) and a combined N+P+Fe pulse (NPFe, 5 µM NO_3^- + 0.5 µM PO_4^{3-} 0.5 nM + Fe added). After 96 hours, samples were serially filtered onto a 3 µM polycarbonate and 0.2 µM polycarbonate filter and flash-frozen. RNA extraction and sequencing were carried out as detailed for the surface metatranscriptomes (Appendix 1). The nutrient conditions of the three sites were determined from discrete seawater samples analyzed after the 2017 cruise and indicate the initial total dissolved nutrient concentrations of the treatment communities. In REXP1, the surface concentration of nitrate + nitrite (N+N) was 1.78 µM, phosphate was and 0.42 µM, and iron was 0.31 nM. In REXP2 *in situ* N+N and phosphate concentrations were lower than REXP1 (0.006 µM and 0.13 µM respectively), and the highest iron concentrations of all three experiments (0.43 nM). REXP3 had the lowest initial concentrations of N+N and phosphorus between the three experiments, with nitrate at 0.002 µM and phosphate 0.07 µM. Discrete iron measurements were not taken for REXP3 at Station 16, but are likely near the values measured at adjacent stations; 0.30 nM Fe at 34.00°N and 0.20 nM Fe at 32.00°N.

3.5.3 Mapping of resource ratio experiment metatranscriptome reads to G2PA assemblies

Quality-controlled short reads from the resource ratio experiments were mapped back onto the G2PA assemblies using kallisto, using the parameters for kallisto described in Appendix 1.

3.5.4 Estimates of species bin coverage

Species bin coverage completeness estimates were performed using the Pfam annotations associated with REXP-aligned assembled transcripts, and the positive detection of Pfam protein families was compared with a set of core transcribed genes defined for major phytoplankton lineages as described in Groussman et al., (Sci Data 2022, in prep) to identify the percent of transcribed protein families expected in a completely sequenced transcriptome for cruise-wide comparisons as well as in each separate sample.

3.5.5 Non-metric multidimensional scaling (nMDS) of metatranscriptomes

Non-metric multidimensional scaling (nMDS) was conducted to reduce the dimensionality of metatranscriptome communities for comparison. We performed nMDS on the fractional composition of total reads aligned to eukaryotic species-bin transcripts, with animal counts omitted. This ordination was performed for the 166 Gradients large and small surface metatranscriptome samples together (stress = 0.1123908), for the 82 large and small REXP metatranscriptome samples together (stress = 0.08888119), and for the 122 small samples from the surface and REXP metatranscriptomes (stress = 0.1165293). For the first two ordinations, species were filtered to those with an average of at least 0.2% of the total read fraction and a non-zero presence in all samples, resulting in 66 species for the surface ordination and 65 species for the REXP ordination. For the combined small-fraction, we combined these counts for the small surface and REXP samples and filtered the species again to the subset present in both sets; a total of 62 species were retained for this analysis. A matrix of samples and species relative abundance was created, and samples row-normalized so row sums equal 1. The metaMDS function in the R package “vegan” version 2.5-5 (Oksanen et al., 2019, parameters: $k = 2$) was used to compute the Bray–Curtis distance matrix and find a solution.

3.5.6 PERMANOVA to identify differences in REXP treatments

We conducted PERMANOVA (Permutational Multivariate Analysis of Variance Using Distance Matrices)on the Bray-Curtis dissimilarity of small-fraction REXP samples to determine if the REXP communities were statistically different following treatment using the ‘adonis’ function in R package “vegan” (Oksanen et al., 2019,, parameters: permutations = 999, method = "bray"), with the null hypothesis that groups were not different. PERMANOVA was conducted for the three REXP experiments and independently for the treatments in each experiment. Significance values are shown in Table 1.

3.5.7 Environmental fit

Environmental parameters and species abundances were fitted to the combined surface & REXP small-fraction ordination to assess their fitted correlation to the ordination structure using the ‘envfit’ function from R package “vegan” (Oksanen et al., 2019, parameters: permutations = 999, na.rm = TRUE). Significance values are shown in Table 2.

3.5.8 Two-tailed t-test for significantly different species abundance by read %

We conducted two-tailed t-tests on the fractional read abundance of species in the REXP experiments, with the null hypothesis that treatment communities are not significantly different than the control at the 95% confidence level. Tests were conducted with the ‘t.test’ function in R (parameters: alternative = “two.sided”, na.rm = TRUE, var.equal = TRUE, conf.level = 0.95). P-values were adjusted for multiple comparison corrections with a 10% false discovery rate using the Benjamini-Hochberg method with R function ‘p.adjust’ (parameters: “BH”).

3.5.9 Differentially Expressed Genes

Two-tailed t-tests were conducted on the TPM-normalized Pfam protein abundance of 12 significant responder species comparing the REXP treatments to the control, with the null hypothesis that relative Pfam abundance in a species under the treatment is not significantly different than the control at the 95% confidence level. Tests were conducted with the ‘t.test’ function in R (parameters: alternative = “two.sided”, na.rm = TRUE, var.equal = TRUE, conf.level = 0.95). P-values were adjusted for multiple comparison corrections with a 10% false discovery rate using the Benjamini-Hochberg method with R function ‘p.adjust’ (parameters: “BH”).

3.6 References

Ayers, J. M., & Lozier, M. S. (2010). Physical controls on the seasonal migration of the North Pacific transition zone chlorophyll front. *Journal of Geophysical Research: Oceans*, *115*(C5).

Bender, S. J., Moran, D. M., McIlvin, M. R., Zheng, H., McCrow, J. P., Badger, J., DiTullio, G. R., Allen, A. E. & Saito, M. A. (2018). Colony formation in Phaeocystis antarctica: connecting molecular mechanisms with iron biogeochemistry. *Biogeosciences*, *15*(16), 4923-4942.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Bhattacharya, D., Yoon, H. S., & Hackett, J. D. (2004). Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *Bioessays*, *26*(1), 50-60.

Mars Brisbin, M., & Mitarai, S. (2019). Differential gene expression supports a resource-intensive, defensive role for colony production in the bloom-forming haptophyte, Phaeocystis globosa. *Journal of Eukaryotic Microbiology*, *66*(5), 788-801.

Bowazolo, C., Song, B., Dorion, S., Beauchemin, M., Chevrier, S., Rivoal, J., & Morse, D. (2022). Orchestrated translation specializes dinoflagellate metabolism three times per day. *Proceedings of the National Academy of Sciences*, *119*(30), e2122335119.

Boysen, A. (2020). *Marine microbial metabolomics: a journey through time, space, and metabolism*. University of Washington.

Browning, T. J., Achterberg, E. P., Rapp, I., Engel, A., Bertrand, E. M., Tagliabue, A., & Moore, C. M. (2017). Nutrient co-limitation at the boundary of an oceanic gyre. *Nature*, *551*(7679), 242-246.

Caron, D. A. (2016). Mixotrophy stirs up our understanding of marine food webs. *Proceedings of the National Academy of Sciences*, *113*(11), 2806-2808.

Caron, D.A., Alexander, H., Allen, A.E., Archibald, J.M., Armbrust, E., Bachy, C., Bell, C.J., Bharti, A., Dyhrman, S.T., Guida, S.M. and Heidelberg, K.B., Kaye, J. Z., Metzner, J., Smith, S. R., & Worden, A. Z. (2017). Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology*, *15*(1), 6-20.

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., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Tara Oceans Coordinators, Jaillon, O., Aury, J., Karsenti, E., Sullivan, M. B, Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., de Vargas, C., Iudicone, D., Bowler C., & Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature communications*, *9*(1), 373.

Cohen, N. R., McIlvin, M. R., Moran, D. M., Held, N. A., Saunders, J. K., Hawco, N. J., Brosnahan, M., DiTullio, G. R., Lamborg, C., McCrow, J. P., Dupont, C. L., Allen, A. E., & Saito, M. A. (2021). Dinoflagellates alter their carbon and nutrient metabolic strategies across environmental gradients in the central Pacific Ocean. *Nature Microbiology*, *6*(2), 173-186.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., ... & Karsenti, E. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1261605.

Fernández-Reiriz, M. J., Perez-Camacho, A., Ferreiro, M. J., Blanco, J., Planas, M., Campos, M. J., & Labarta, U. (1989). Biomass production and variation in the biochemical profile (total protein, carbohydrates, RNA, lipids and fatty acids) of seven species of marine microalgae. *Aquaculture*, *83*(1-2), 17-37

Finkel, Z. V., Follows, M. J., Liefer, J. D., Brown, C. M., Benner, I., & Irwin, A. J. (2016). Phylogenetic diversity in the macromolecular composition of microalgae. *PLoS One*, *11*(5), e0155977.

Fiset, C., Irwin, A. J., & Finkel, Z. V. (2019). The macromolecular composition of noncalcified marine macroalgae. *Journal of phycology*.

Flynn, K. J., Mitra, A., Anestis, K., Anschütz, A. A., Calbet, A., Ferreira, G. D., Gypens, N., Hansen, P. J., John, U., Martin, J. L. and Mansour, J. S., Maselli, M., Nikola Medić, A. N. Norlin, A., Not, F. Pitta, P., Romano, F., Saiz, E., Schneidler, L. K., Stolte, W., & Traboni, W. (2019). Mixotrophic protists and a new paradigm for marine ecology: where does plankton research go now?. *Journal of Plankton Research*, *41*(4), 375-391.

Follett, C. L., Dutkiewicz, S., Forget, G., Cael, B. B., & Follows, M. J. (2021). Moving ecological and biogeochemical transitions across the North Pacific. Limnology and Oceanography, 66(6), 2442-2454.
Frias-Lopez, J., Thompson, A., Waldbauer, J., & Chisholm, S. W. (2009). Use of stable isotope-labelled cells to identify active grazers of picocyanobacteria in ocean surface waters. *Environmental microbiology*, *11*(2), 512-525.

Geider, R. J., & La Roche, J. (2002). Redfield revisited: variability of C [ratio] N [ratio] P in marine microalgae and its biochemical basis. *European Journal of Phycology*, *37*(1), 1-17.

Glibert, P. M. (2016). Margalef revisited: a new phytoplankton mandala incorporating twelve dimensions, including nutritional physiology. *Harmful Algae*, *55*, 25-30.

Gobler, C. J., Boneillo, G. E., Debenham, C. J., & Caron, D. A. (2004). Nutrient limitation, organic matter cycling, and plankton dynamics during an Aureococcus anophagefferens bloom. *Aquatic Microbial Ecology*, *35*(1), 31-43.

Gradoville, M. R., Farnelid, H., White, A. E., Turk-Kubo, K. A., Stewart, B., Ribalet, F., Ferrón, S., Pinedo-Gonzalez, P., Armbrust, E. V., Karl, D. M., John, S., & Zehr, J. P. (2020). Latitudinal constraints on the abundance and activity of the cyanobacterium UCYN-A and other marine diazotrophs in the North Pacific. *Limnology and Oceanography*, *65*(8), 1858-1875.

Groussman, R. D., Parker, M. S., & Armbrust, E. V. (2015). Diversity and evolutionary history of iron metabolism genes in diatoms. *PLoS One*, *10*(6), e0129081.

Groussman, R. D., Blaskowski, S., Coesel, S., & Armbrust, E. V. (2022). MarFERReT: an open-source, version-controlled reference library of marine microbial eukaryote functional genes. *SciData (*in prep)

Guérin, N., Ciccarella, M., Flamant, E., Frémont, P., Mangenot, S., Istace, B., Noel, B., Belser, C., Bertrand, L., Labadie, K. Cruaud, C., Romac, S., Bachy, C., Gachenot, M., Pelletier, E., Alberti, A., Jaillon, O., Wincker, P., Aury, J. M., & Carradec, Q. (2022). Genomic adaptation of the picoeukaryote Pelagomonas calceolata to iron-poor oceans revealed by a chromosome-scale genome sequence. *Communications biology*, *5*(1), 1-14.

Hu, S. K., Connell, P. E., Mesrop, L. Y., & Caron, D. A. (2018). A Hard Day's Night: Diel Shifts in Microbial Eukaryotic Activity in the North Pacific Subtropical Gyre. Front. *Mar. Sci*, *5*, 351.

Jones, H. L., Leadbeater, B. S. C., & Green, J. C. (1993). Mixotrophy in marine species of Chrysochromulina (Prymnesiophyceae): ingestion and digestion of a small green flagellate. *Journal of the Marine Biological Association of the United Kingdom*, *73*(2), 283-296

Juranek, L. W., Quay, P. D., Feely, R. A., Lockwood, D., Karl, D. M., & Church, M. J. (2012). Biological production in the NE Pacific and its influence on air-sea CO2 flux: Evidence from dissolved oxygen isotopes and O2/Ar. *Journal of Geophysical Research: Oceans*, *117*(C5).

Juranek, L. W., White, A. E., Dugenne, M., Henderikx Freitas, F., Dutkiewicz, S., Ribalet, F., Ferrón, S. E., Armbrust, E. V., & Karl, D. M. (2020). The importance of the phytoplankton "middle class" to ocean net community production. *Global Biogeochemical Cycles*, *34*(12), e2020GB006702.

Kamykowski, D., Milligan, E. J., & Reed, R. E. (1998). Biochemical relationships with the orientation of the autotrophic dinoflagellate Gymnodinium breve under nutrient replete conditions. *Marine Ecology Progress Series*, *167*, 105-117.

Kark, S., Allnutt, T. F., Levin, N., Manne, L. L., & Williams, P. H. (2007). The role of transitional areas as avian biodiversity centres. *Global Ecology and Biogeography*, *16*(2), 187-196.

Kark, S. (2013). Ecotones and ecological gradients. In *Ecological systems* (pp. 147-160). Springer, New York, NY.

Koppelle, S., López-Escardó, D., Brussaard, C. P., Huisman, J., Philippart, C. J., Massana, R., & Wilken, S. (2022). Mixotrophy in the bloom-forming genus Phaeocystis and other haptophytes. *Harmful Algae*, *117*, 102292.

Lambert, B. S., Groussman, R. D., Schatz, M. J., Coesel, S. N., Durham, B. P., Alverson, A. J., White, A. E., & Armbrust, E. V. (2022). The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. *Proceedings of the National Academy of Sciences*, *119*(7), e2100916119.

Li, A., Stoecker, D. K., & Coats, D. W. (2000). Spatial and temporal aspects of Gyrodinium galatheanum in Chesapeake Bay: distribution and mixotrophy. *Journal of Plankton Research*, *22*(11), 2105-2124.

Li, L., Wang, S., Wang, H., Sahu, S. K., Marin, B., Li, H., Xu, Y., Liang, H., Li, Z., Cheng, S., Reder, T., & Liu, H. (2020). The genome of Prasinoderma coloniale unveils the existence of a third phylum within green plants. *Nature ecology & evolution*, *4*(9), 1220-1231.

Li, Q., Edwards, K. F., Schvarcz, C. R., Selph, K. E., & Steward, G. F. (2021). Plasticity in the grazing ecophysiology of Florenciella (Dichtyochophyceae), a mixotrophic nanoflagellate that consumes Prochlorococcus and other bacteria. *Limnology and Oceanography*, *66*(1), 47-60.

Marchetti, A., Schruth, D. M., Durkin, C. A., Parker, M. S., Kodner, R. B., Berthiaume, C. T., Morales, R., Allen, A. E. & Armbrust, E. V. (2012). Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences*, *109*(6), E317-E325.

Margalef, R. (1978). Life-forms of phytoplankton as survival alternatives in an unstable environment. *Oceanologica acta*, *1*(4), 493-509.

Martiny, A. C., Pham, C. T., Primeau, F. W., Vrugt, J. A., Moore, J. K., Levin, S. A., & Lomas, M. W. (2013). Strong latitudinal patterns in the elemental ratios of marine plankton and organic matter. *Nature Geoscience*, *6*(4), 279-283.

Mitra, A., Flynn, K. J., Tillmann, U., Raven, J. A., Caron, D., Stoecker, D. K., Not, F., Hansen, P. J., Hallegraeff, G., Sanders, R. Wilken, S., McManus, G., Johnson, M., Pitta, P., Våge, , B., Calbet, A., Thingstad, F., Jeong, H. J., Burkholder, J., Glibert, P. M., Granéli, E., & Lundgren, V. (2016). Defining planktonic protist functional groups on mechanisms for energy and nutrient acquisition: incorporation of diverse mixotrophic strategies. *Protist*, *167*(2), 106-120.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic acids research*, *49*(D1), D412-D419.

Moore, C. M., Mills, M. M., Arrigo, K. R., Berman-Frank, I., Bopp, L., Boyd, P. W., Galbraith, E.D., Geider, R.J., Guieu, C., Jaccard, S.L., & Jickells, T. D. (2013). Processes and patterns of oceanic nutrient limitation. *Nature geoscience*, *6*(9), 701-710.

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., Van Bel, M., Poulain, J., Katinka, M., Hohmann-Marriott, M. F., Piganeau, G., & Vandepoele, K. (2012). Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome biology*, *13*(8), 1-16.

Mulholland, M. R., Boneillo, G., & Minor, E. C. (2004). A comparison of N and C uptake during brown tide (Aureococcus anophagefferens) blooms from two coastal bays on the east coast of the USA. *Harmful Algae*, *3*(4), 361-376.

Oksanen, J. (2007). Vegan: community ecology package. R package version 1.8-5. *http://www. cran. r-project. org*.

Pasulka, A. L., Landry, M. R., Taniguchi, D. A., Taylor, A. G., & Church, M. J. (2013). Temporal dynamics of phytoplankton and heterotrophic protists at station ALOHA. *Deep Sea Research Part II: Topical Studies in Oceanography*, *93*, 44-57.

Pinedo-González, P., Hawco, N. J., Bundy, R. M., Armbrust, E., Follows, M. J., Cael, B. B., White, A. E., Ferrón, S., Karl, D. M. & John, S. G. (2020). Anthropogenic Asian aerosols provide Fe to the North Pacific Ocean. *Proceedings of the National Academy of Sciences*, *117*(45), 27862-27868.

Polovina, J. J., Howell, E., Kobayashi, D. R., & Seki, M. P. (2001). The transition zone chlorophyll front, a dynamic global feature defining migration and forage habitat for marine resources. *Progress in oceanography*, *49*(1-4), 469-483.

Polovina, J. J., Howell, E. A., Kobayashi, D. R., & Seki, M. P. (2017). The Transition Zone Chlorophyll Front updated: Advances from a decade of research. *Progress in Oceanography*, *150*, 79-85.

Pustizzi, F., MacIntyre, H., Warner, M. E., & Hutchins, D. A. (2004). Interaction of nitrogen source and light intensity on the growth and photosynthesis of the brown tide alga Aureococcus anophagefferens. *Harmful Algae*, *3*(4), 343-360.

Quigg, A., Irwin, A. J., & Finkel, Z. V. (2010). Evolutionary inheritance of elemental stoichiometry in phytoplankton. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1705), 526-534.

Roden, G. I. (1991). Subarctic-subtropical transition zone of the North Pacific: large-scale aspects and mesoscale structure. *NOAA Technical Report NMFS*, *105*, 1-38.

Sheldon, R. W., Prakash, A., & Sutcliffe Jr, W. (1972). THE SIZE DISTRIBUTION OF PARTICLES IN THE OCEAN 1. *Limnology and oceanography*, *17*(3), 327-340.

Sieburth, J. M., Smetacek, V., & Lenz, J. (1978). Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions 1. *Limnology and oceanography*, *23*(6), 1256-1263.

Stibor, H., & Sommer, U. (2003). Mixotrophy of a photosynthetic flagellate viewed from an optimal foraging perspective. *Protist*, *154*(1), 91-98.

Stoecker, D. K., Hansen, P. J., Caron, D. A., & Mitra, A. (2017). Mixotrophy in the marine plankton. *Annu. Rev. Mar. Sci*, *9*(1), 311-335.

Takahashi, T., Sutherland, S. C., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., Bates, N., Wanninkhof, R., Feely, R. A., Sabine, C. & Olafsson, J. (2002). Global sea–air CO2 flux based on climatological surface ocean pCO2, and seasonal biological and temperature effects. *Deep Sea Research Part II: Topical Studies in Oceanography*, *49*(9-10), 1601-1622.

Tang, E. P. (1996). Why do dinoflagellates have lower growth rates? *Journal of Phycology*, *32*(1), 80-84.

Tilman, D. (1985). The resource-ratio hypothesis of plant succession. *The American Naturalist*, *125*(6), 827-852.

Vannier, T., Leconte, J., Seeleuthner, Y., Mondy, S., Pelletier, E., Aury, J. M., De Vargas, C., Sieracki, M., Iudicone, D., Vaulot, D., Wincker, P., & Jaillon, O. (2016). Survey of the green picoalga Bathycoccus genomes in the global ocean. *Scientific reports*, *6*(1), 1-11.

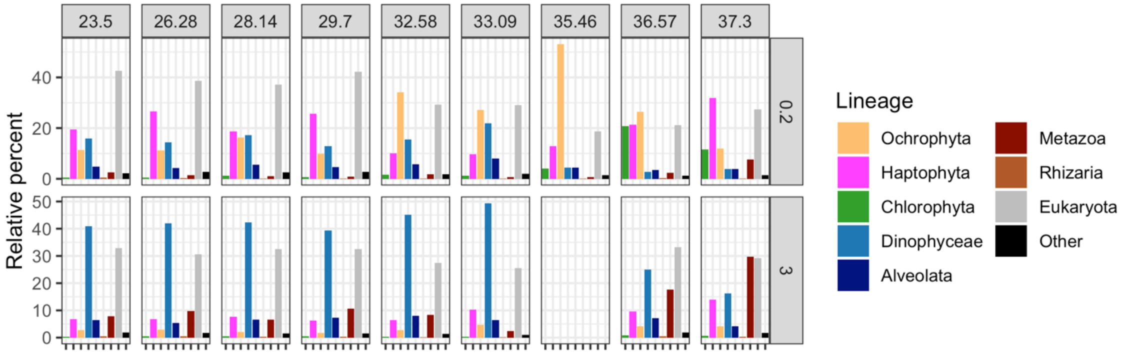
Ward, B. A., Dutkiewicz, S., Moore, C. M., & Follows, M. J. (2013). Iron, phosphorus, and nitrogen supply ratios define the biogeography of nitrogen fixation. *Limnology and Oceanography*, *58*(6), 2059-2075.

Warner, J. R. (1999). The economics of ribosome biosynthesis in yeast. *Trends in biochemical sciences*, *24*(11), 437-440.

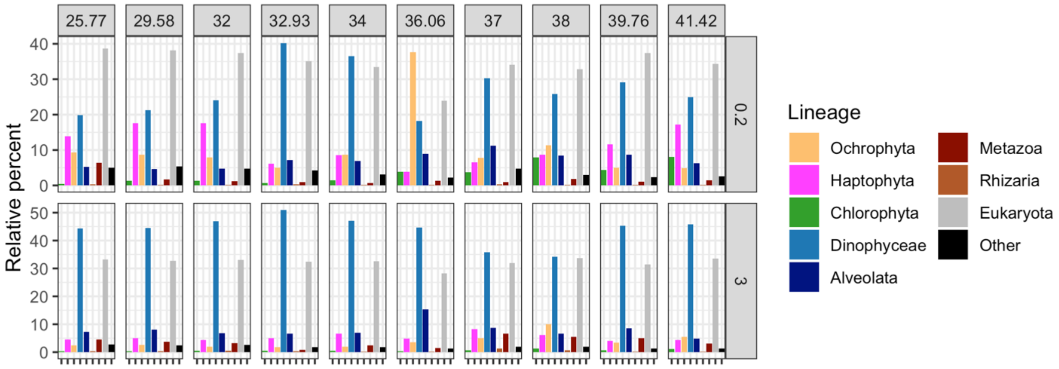
Wisecaver, J. H., & Hackett, J. D. (2011). Dinoflagellate genome evolution. *Annual review of microbiology*, *65*, 369-387.

Worden, A. Z., Janouskovec, J., McRose, D., Engman, A., Welsh, R. M., Malfatti, S., Tringe, S. G., & Keeling, P. J. (2012). Global distribution of a wild alga revealed by targeted metagenomics. *Current Biology*, *22*(17), R675-R677.

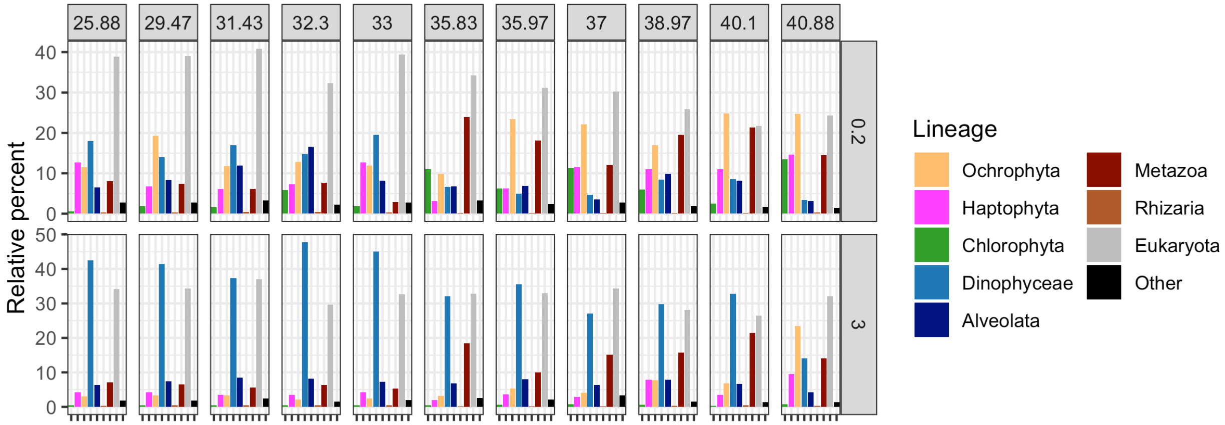
3.7 Supplementary Figures



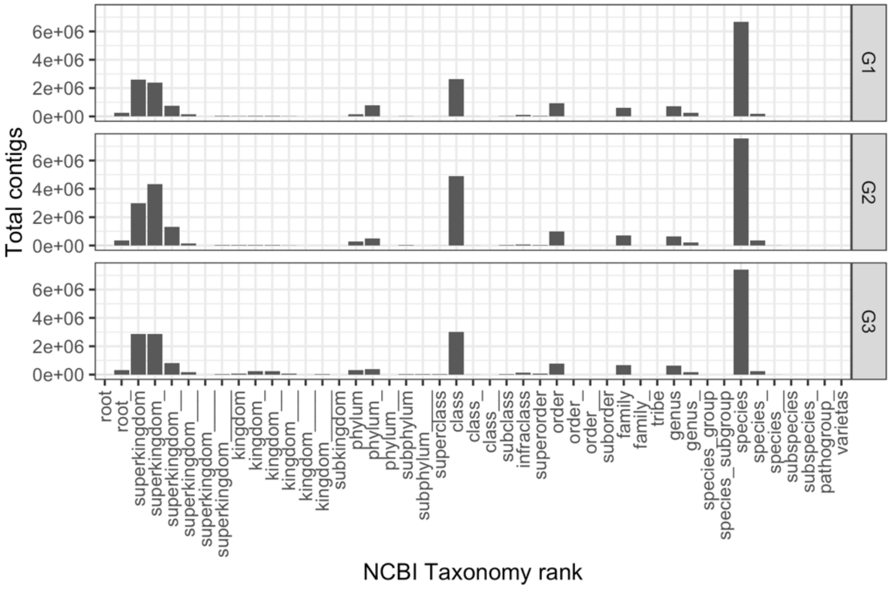
Supplementary Figure S3.1. Gradients 1 taxonomy overview. Taxonomic annotations of size-fractionated polyA+-selected metatranscriptomes from Gradients 1 cruise transect along ~158°W (Apr 20-May 2, 2016). Top panel labels indicate the sampling site latitude (°N). Side panel labels indicate the sample size fraction (top row 0.2 – 3 µm, bottom row 3-200 µm). For each panel, the height of the bars represents the fractional composition (relative percent) of total reads mapped to contigs assigned to major microplankton lineages. 'Other' includes all non-eukaryotic assignments.



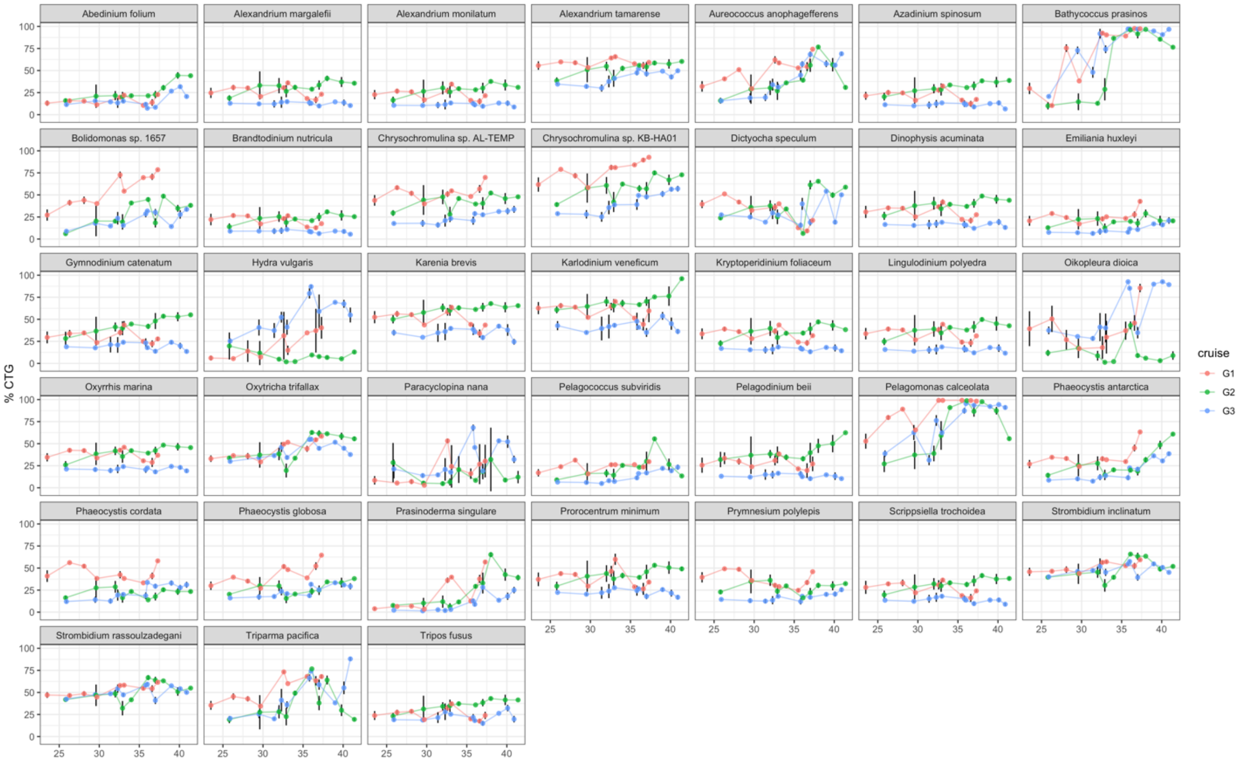
Supplementary Figure S3.2. Gradients 1 taxonomy overview. Taxonomic annotations of size-fractionated polyA+-selected metatranscriptomes from Gradients 2 cruise transect along ~158°W (May 27-Jun 11, 2017). Top panel labels indicate the sampling site latitude (°N). Side panel labels indicate the sample size fraction (top row 0.2 – 3 µm, bottom row 3-100 µm). For each panel, the height of the bars represents the fractional composition (relative percent) of total reads mapped to contigs assigned to major microplankton lineages. 'Other' includes all non-eukaryotic assignments.



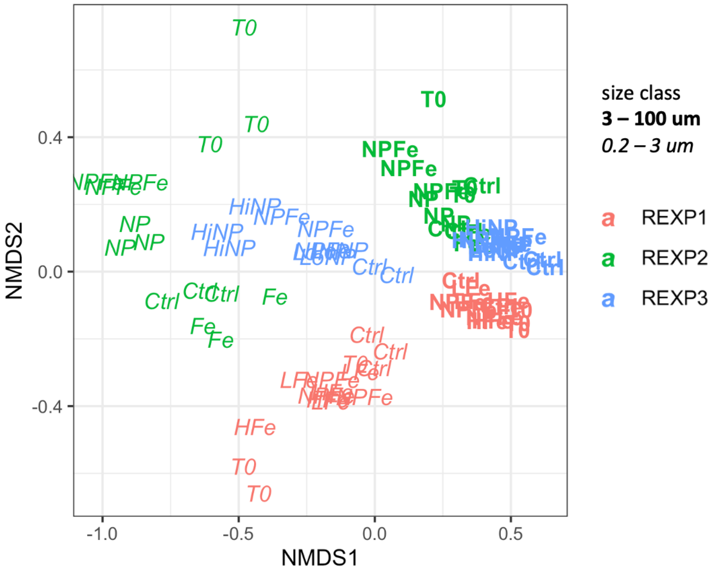
Supplementary Figure S3.3. Gradients 1 taxonomy overview. Taxonomic annotations of size-fractionated polyA+-selected metatranscriptomes from Gradients 3 cruise transect along ~158°W (Apr 8-28, 2019). Top panel labels indicate the sampling site latitude (°N). Side panel labels indicate the sample size fraction (top row 0.2 – 3 µm, bottom row 3-100 µm). For each panel, the height of the bars represents the fractional composition (relative percent) of total reads mapped to contigs assigned to major microplankton lineages. 'Other' includes all non-eukaryotic assignments.



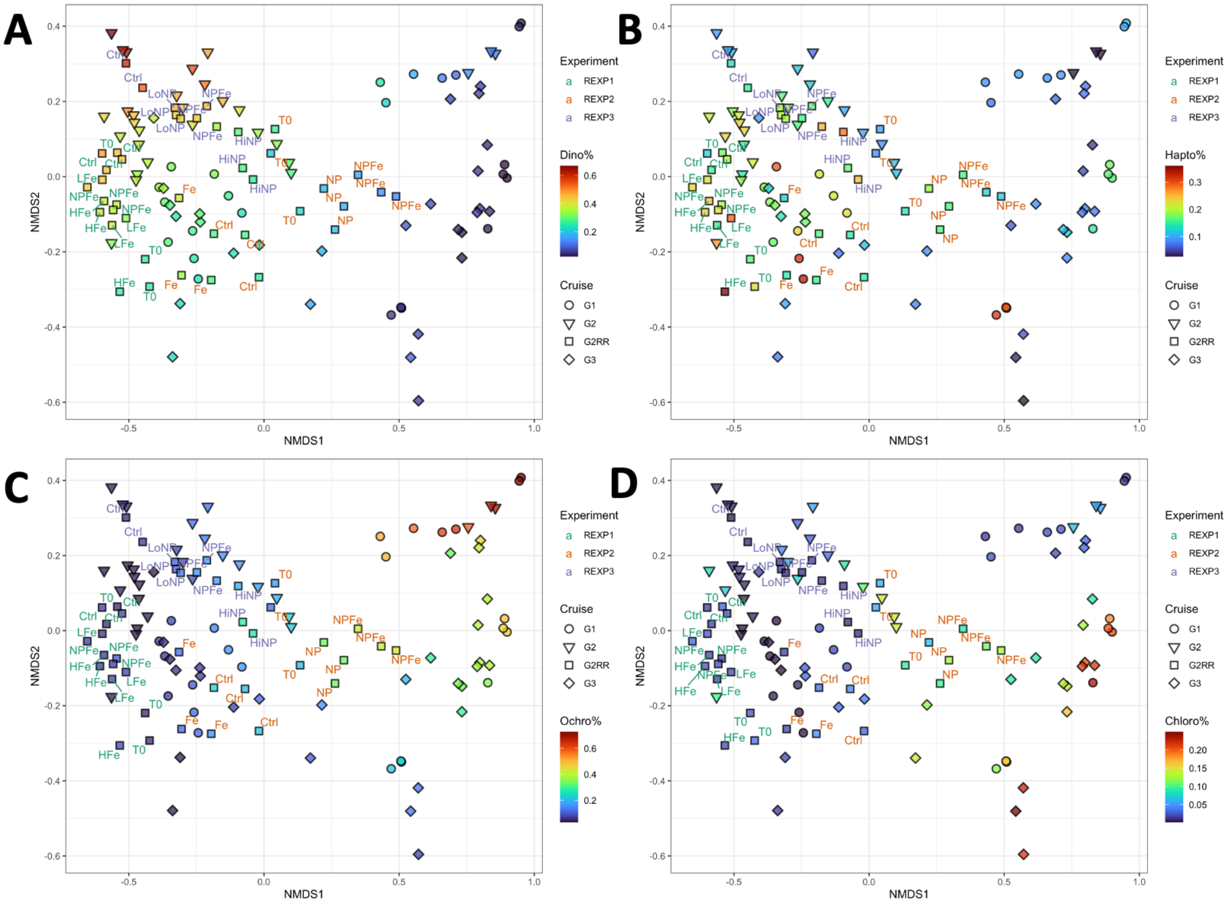
Supplementary Figure S3.4. Metatranscriptome contig taxonomic assignments by rank. Each panel shows the sum of taxonomic annotations to NCBI Taxonomy rank levels, with Gradients 1, 2 and 3 sums shown as separate panels.



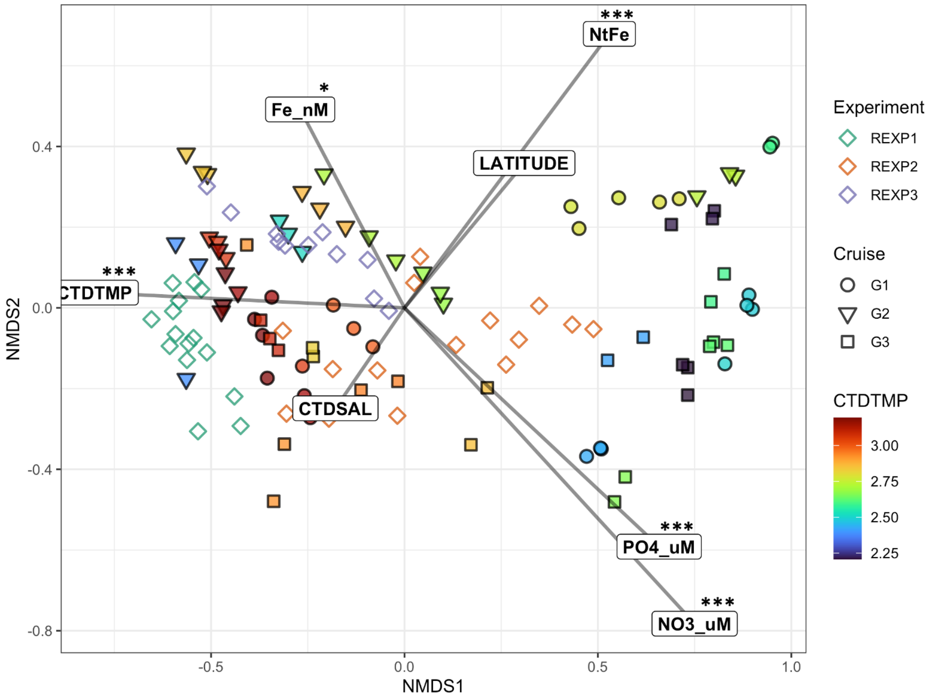
Supplementary Figure S3.5. Small-size fraction completeness for environmental species bins. Species bins with over 50% of total CTGs found in cruise-wide coverage estimates are shown here. Each point is a sample site, with error bars showing the standard deviation between triplicate samples. Horizontal axis shows sampling site latitude (°N). Vertical axis shows the percent of expected transcribed protein functions found. Samples sites are colored by Gradients cruise G1 through G3. Error bars are the standard deviation between replicates.



Supplementary Figure S3.6. nMDS ordination of 82 large and small metatranscriptome samples from the Gradients 2 on-deck resource ratio incubation experiments. (stress = 0.089). Samples are colored by experiment and labeled with their nutrient treatment (see Fig 3b), or as control bottles (Ctrl) and T=0 community samples (T0). Bold text indicates large samples, italicized text, small samples.



Supplementary Figure S3.7. nMDS ordination of species transcript abundance in the small-fraction metatranscriptomes of Gradients 1, 2 and 3 the Gradients 2 on-deck resource ratio incubation experiments (REXP). Ordination results are as described in Fig 3c, with points colored by the total percent of community metatranscriptome reads mapped to species in the **a)** dinoflagellate, **b)** haptophyte, **c)** ochrophyte and **d)** chlorophyte lineages.



Supplementary Figure S3.8. nMDS ordination of species transcript abundance in the small-fraction metatranscriptomes of Gradients 1, 2 and 3 the Gradients 2 on-deck resource ratio incubation experiments (REXP). Ordination results are as described in Fig 3c, with points colored by CTD temperature. Environmental variables tested are shown as labeled vectors, vector lengths indicate the strength of the strength of the correlation, asterisks above labels show significance levels at Pr(>F) <0.001 (***) and 0.05 (*).

APPENDIX 1

The North Pacific Eukaryotic Gene Catalog of metatranscriptome assemblies with taxonomic, function and abundance annotations

Ryan Groussman

A1.1 Abstract

The North Pacific Ocean is structured by large-scale physical processes into oceanographic provinces with different biogeochemical characteristics. The North Pacific Subtropical Gyre is bordered on its northern side by a region known as the North Pacific Transition Zone; a latitudinal band of strong physical, chemical, and biological gradients and high productivity where warm, nutrient-deplete water from the subtropical gyre mixes with cold, nutrient-rich water from the north. The North Pacific Eukaryotic Gene Catalog consolidates eukaryotic metatranscriptome data from three latitudinal transects of the transition zone and one cruise in the subtropical gyre. Metatranscriptomes were gathered from latitudinally-resolved surface samples and diel-resolved temporal studies, with samples taken in triplicate or duplicate and collected on 0.2-100 µm, 0.2-3 µm, and 3 µm-100 or 200 µm size fractions. These metatranscriptome data were de novo assembled into 175 independent assemblies, totaling 182 million clustered transcript contigs. Assemblies were annotated by taxonomy and function and enumerated by short read alignment. This catalog provides assembled environmental contigs, their translated peptide sequences, taxonomic and functional annotations and read counts with the aim of facilitating continued discoveries about the molecular ecology of microbial eukaryotes in the North Pacific.

A1.2 Background & Summary

The North Pacific Ocean is structured by large-scale physical processes into large provinces with differing biogeochemical regimes. The North Pacific Subtropical Gyre (NPSG) is a basin-spanning feature characterized by warm and nutrient-deplete water, bounded on its northern side by the cold, nutrient-rich and iron-poor water of the North Pacific Subpolar Gyre (Karl 1999). The region between these gyres is known as the North Pacific Transition Zone (NPTZ), a latitudinal band spanning ~32°-42°N of strong physical, chemical, and biological gradients and high net community productivity (Juranek et al., 2020). Here we incorporate eukaryotic metatranscriptome data products from four oceanographic cruises in these areas of the North Pacific (Figure A1.1, Table A1.1).

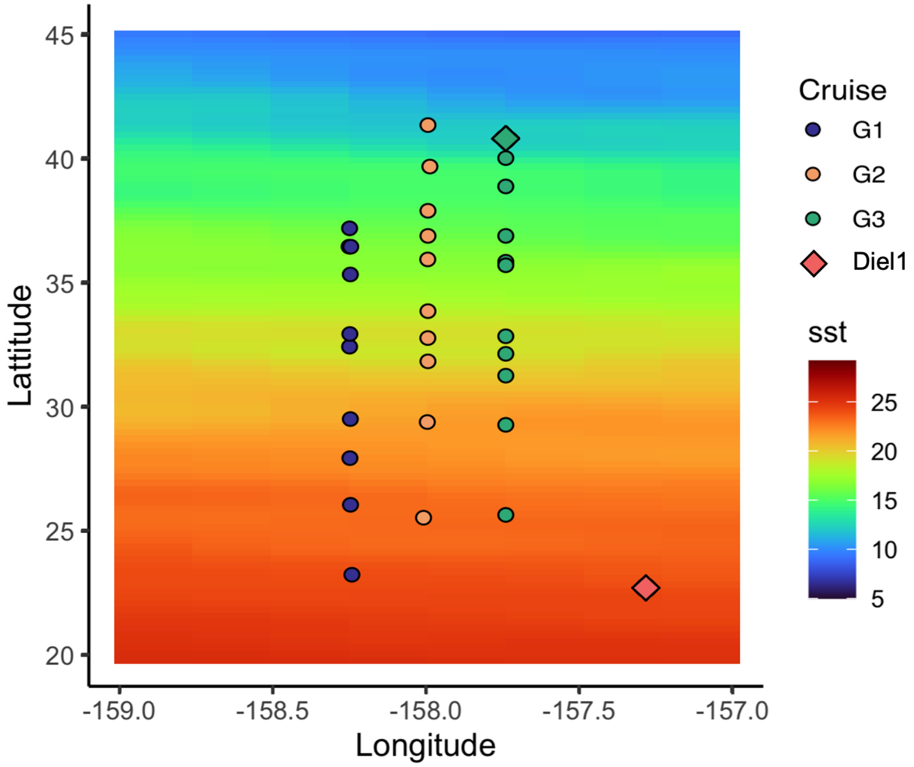


Figure A1.1. Sample sites of metatranscriptomes in the North Pacific Eukaryotic Gene Catalog. Background gradient is average sea surface temperature (sst) in °C from 2016 to 2022 (GHRSST Global Blended Sea Surface Temperature from NCEI). SST data was retrieved from Simons CMAP (https://simonscmap.com/, Ashkezari et al., 2021). Cruises are indicated by color; G1, Gradients 1; G2, Gradients 2; G3, Gradients 3, Diel1. G1 and G3 cruise tracks are offset from 158 °N by -0.25° and +0.25° longitude for visibility. Circles indicate locations of dawn metatranscriptome samples; diamonds indicate the stations where multi-day diel observations were conducted (Diel1 and G3).

The first set of metatranscriptomes came from R/V *Kilo Moana* cruise KM1513 (referred to here with the shorthand name ‘Diel1’), which sought to understand biological variability over diel timescales and followed a semi-Lagrangian drifter near Station ALOHA in the North Pacific Subtropical Gyre (Wilson et al., 2017). Surface seawater samples for metatranscriptomes were taken every four hours in duplicate over a four-day sampling period (Figure A1.1) for the 0.2-100 μm size fraction. These metatranscriptomes have contributed to studies of investigating diel regulation of light-sensitive regulatory elements in eukaryotic phytoplankton (Coesel et al., 2021), assessments of the trophic modes of species across diel cycles (Lambert et al., 2022), and the diel regulation of flavodoxin transcripts in diatoms (van Creveld et al., 2022).

Table A1.1. Summary of cruises in the North Pacific Eukaryotic Gene Catalog and the metatranscriptomes collected from these cruises. The name for metatranscriptome series in this catalog is followed by the cruise ID, cruise dates, and the filter pore size for metatranscriptome samples in μm. The ‘Sum/Mean%’ column reports the total sum of values across rows, or the average of percentages (% with tax, % with Pfam). Total samples is the number of sequenced metatranscriptome samples, total assemblies is the number of *de novo* Trinity assemblies generated from metatranscriptome samples (triplicate samples were pooled for assembly in Diel1, Gradients1, and a portion of Gradients2). Raw transcripts indicates the total number of transcript contigs output by Trinity. Clustered aa is the number of translated, frame-selected proteins following clustering at the 99% amino identity threshold. Clustered nt is the number of unique nucleotide transcripts from clustered aa. Number w/ tax is the number of clustered nt transcripts with an assigned NCBI Taxonomy taxID, and % with tax is the percent of clustered nt transcripts with a taxID. Number w/ Pfam is the number of clustered nt transcripts with an assigned Pfam protein family ID, and % with tax is the percent of clustered nt transcripts with a Pfam ID.

Name	**Diel1**	**Gradients1**	**Gradients2**	**Gradients3**	**G3 diel**	**Sum/Avg%**
Cruise ID	KM1513	KOK1606	MGL1704	KM1906	KM1906	
Cruise dates	July-Aug 2015	April-May 2016	May-June 2017	April 2019	April 2019	
Size (μm)	0.2-100	0.2-3, 3-200	0.2-3, 3-100	0.2-3, 3-100	0.2-100	
Total samples	48	47	59	60	40	**254**
Total assemblies	24	19	32	60	40	**175**
Raw transcripts	52,489,585	40,892,391	57,666,908	47,569,922	35,610,144	**234,228,950**
Clustered aa	50,461,546	37,950,644	49,691,936	38,103,601	23,272,764	**199,480,491**
Clustered nt	48,907,619	31,905,677	44,353,029	34,792,290	22,293,879	**182,252,494**
number w/ tax	26,692,293	19,420,897	25,572,811	21,569,823	14,374,579	**107,630,403**
% with tax	54.6%	60.9%	57.7%	62.0%	64.5%	**59.9%**
number w/ Pfam	17,693,251	11,796,125	14,945,957	13,648,989	8,276,146	**66,360,468**
% with Pfam	36.2%	37.0%	33.7%	39.2%	37.1%	**36.6%**

A series of three cruises transected the North Pacific Transition Zone in the Spring of 2016, 2017 and 2019 along 158W longitude and studied biogeochemical and physical gradients with latitudinal resolution (referred to here as ‘Gradients 1’, ‘Gradients 2’ and ‘Gradients 3’). Surface seawater samples were collected in triplicate for metatranscriptomes on the three cruises at dawn to minimize diel transcript variability. Gradients 1 samples were collected between 19 April to 3 May 2016 aboard the R/V *Ka'imikai-O-Kanaloa* cruise KOK1606, and samples were collected in the 0.2-3 and 3-200 μm size fractions at 15 m depth. The Gradients 1 metatranscriptomes have been used to estimate changing trophic modes of mixotrophic species across the transition zone (Lambert et al., 2022). Gradients 2 samples were collected between 27 May to 13 June 2017 aboard the R/V *Marcus G. Langseth* cruise MGL1704, and samples were collected along this transect in the 0.2-3 and 3-100 μm size fractions from dawn CTD casts at 15m depth. Gradients 3 samples were collected between 10 April to 29 April 2019 aboard the R/V *Kilo Moana* cruise KM1906, and metatranscriptome samples in the 0.2-3 and 3-100 μm size fractions were collected at dawn from the ship's seawater intake at ~7m depth. A diel-resolved study of *in situ* conditions was also conducted on the 2019 Gradients 3 cruise at the northernmost station at ~41.5°N, to capture diel biological variability in the northern area of the over a 72-hour diel observation period. These metatranscriptome samples were collected approximately every 4 hours from CTD casts at 15m depth over ~72 hours with a 0.2-100 μm size fraction.

Sequenced metatranscriptomes were processed in a standardized pipeline to generate assembled transcripts, taxonomic and functional annotations, and mapped-read abundances (Figure A1.2, Table A1.1). Paired-end short reads from these metatranscriptome projects were assembled into longer transcript contigs, translated, and the longest coding frame(s) retained for downstream analysis. All assemblies were clustered together at the 99% amino acid identity level to reduce redundancy between samples. These clustered protein sequences were then annotated by taxonomy against the MarFERReT eukaryotic reference sequence library (Chapter 2, Groussman et al. 2022) and by function against the Pfam 34.0 protein family database (Mistry et al., 2021). Short reads from the metatranscriptomes were mapped back to their assemblies to generate transcript abundances. The North Pacific Eukaryotic Gene Catalog consolidates the metatranscriptome sequence data and associated annotation products together to facilitate the accessibility and continued use of this dataset.

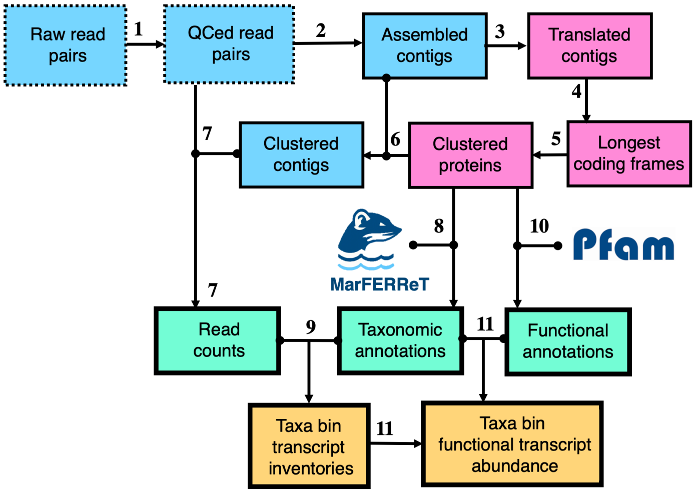


Figure A1.2. Schematic of bioinformatics processing of metatranscriptomes in the North Pacific Eukaryotic Gene Catalog. Diagram illustrates the different steps taken to develop metatranscriptome assemblies and their annotations. Boxes indicate datasets and box borders indicate the data type: short read sequences (dashed lines), assembled transcript sequences (solid line) and annotation products (thick solid line). Box color indicates the sequence type: nucleotide sequences (blue), amino acid sequences (pink), primary annotations products (teal) and derived data using multiple annotations (ochre). Arrows indicate processes: 1) quality-control of short paired-end reads; 2) *de novo* assembly of metatranscriptomes; 3) six-frame translation of assembled nucleotide transcripts; 4) retention of the longest translated coding frame (or multiple, if ties); 5) clustering at the 99% amino acid identity threshold to reduce redundancy; 6) retrieval of the set of nucleotide transcripts associated with the clustered proteins in (5); 7) quantification of assembled transcript abundances through alignment of short reads to clustered nucleotide sequences; 8) taxonomic annotation of protein sequences with the MarFERReT marine eukaryote reference sequence library (Groussman et al., 2022); 9) aggregation of taxonomic annotations and transcript read counts to derive taxa bin transcript inventories; 10) functional annotation of protein sequences with the Pfam 34.0 protein family database (Mistry et al., 2021); 11) aggregation of taxa bin transcript inventories and functional annotations to produce functional transcript inventories of Pfam-annotated transcripts in the taxa bins.

A1.3 Methods

A1.3.1 Cruise and sample collection

Diel1 eukaryotic metatranscriptome samples were collected during R/V *Kilo Moana* cruise KM1513 (SCOPE HOE-Legacy 2, July 2015) near Station ALOHA in the North Pacific Subtropical Gyre as previously described (Durham et al., 2019, Coesel et al., 2021). Briefly, samples were collected from ~15 m depth every 4 hours over a 4-day period. Seawater samples were recovered from Niskin bottles attached to a CTD rosette. Seawater was pre-filtered using 100 μm Nitex mesh and collected on a 0.2 μm polycarbonate filter, and immediately flash-frozen in liquid nitrogen. Total RNA was extracted using the ToTALLY RNA kit (Invitrogen) with the addition of 14 internal mRNA standards (Durham et al., 2019) and poly(A)-selected mRNAs were used for Illumina NextSeq 500 sequencing. Gradients 1 samples were collected during 19 April to 3 May 2016 aboard the R/V *Ka'imikai-O-Kanaloa*, cruise KOK1606 as previously described (Lambert et al., 2021). Seawater samples were recovered from Niskin bottles attached to a CTD rosette, and approximately 6-10 L seawater samples were pre-filtered through a 200 µm nylon mesh and collected by sequential filtration through a 3 µm and a 0.2 µm polycarbonate filters using a peristaltic pump. Gradients 2 samples were collected on the R/V *Marcus G. Langseth*, cruise MGL1704, from the 27 May to 13 June 2017. Seawater samples were recovered from Niskin bottles attached to a CTD rosette, and approximately 6-10 L seawater samples were pre-filtered through a 100 µm nylon mesh and collected by sequential filtration through a 3 µm and a 0.2 µm polycarbonate filters using a peristaltic pump. Gradients 3 samples were collected onboard the 10 April to 29 April 2019 aboard the R/V *Kilo Moana*. Seawater samples for metatranscriptomes were collected at dawn from the ship's seawater intake at ~7m depth and pre-filtered through a 100 µm nylon mesh before sequential filtration through a 3 µm and a 0.2 µm polycarbonate filters using a peristaltic pump. The G3 diel samples were also from the 2019 Gradients 3 cruise at the northernmost station at ~41.5°N over a 72-hour diel observation period. Seawater samples were collected approximately every 4 hours from CTD casts at 15m depth, pre-filtered through a 100 µm nylon mesh and collected on a 0.2 μm polycarbonate filter, and immediately flash-frozen in liquid nitrogen.

A1.3.2 Extraction and sequencing

Metatranscriptome samples were extracted and sequenced as previously described (Durham et al., 2019, Coesel et al., 2019, Lambert et al., 2021). Total RNA for Diel1 and Gradients1 was extracted using the ToTALLY RNA kit (Invitrogen) with the addition of 14 internal mRNA standards (Durham et al., 2019) and poly(A)-selected to retain eukaryotic mRNAs, and were sequenced on an Illumina NextSeq 500 platform. Gradients 2, Gradient 3, and G3 diel metatranscriptomes were extracted using the same methods, except total RNA was extracted using the Direct-zol RNA MiniPrep Plus kit (Zymo Research) and poly(A)-selected mRNAs were sequenced on the Illumina NovaSeq platform. For all samples, extracted RNA was quantified using a Qubit fluorometer (Thermofisher) and quality controlled using a Bioanalyzer (Agilent) prior to sequencing. Samples were randomized across sequencing runs to reduce potential biases.

A1.3.3 Quality control and trimming

Raw Illumina sequence reads were quality controlled with trimmomatic v0.36 (Bolger et al., 2014, parameters: MAXINFO:135:0.5, LEADING:3, TRAILING:3, MINLEN:60, and AVGQUAL:20), as described previously (Chapter 1, Groussman et al., 2021).

A1.3.4 Metatranscriptome assembly

Quality controlled read pairs from the metatranscriptomes were assembled using the Trinity *de novo* assembler (Grabherr et al., 2011, parameters: --normalize_reads --min_kmer_cov 2 --min_contig_length 300). For Diel1, Gradients 1, and Gradients 2, metatranscriptomes were assembled on the Pittsburgh Supercomputing Center's Bridges Large Memory system the using Trinity version v2.3.2, as previously described for Diel1 (Groussman et al., 2021). For Diel1 and Gradients 1, assemblies were conducted on the short reads from combined replicates, Gradients 2 assemblies were conducted on combined replicates (16 of 32 assemblies) and 18 individually assembled replicate samples (18 of 32 assemblies). The Gradients 3 and G3 diel metatranscriptomes were assembled individually for each replicate using Trinity v2.12.0 on a local high-memory cluster. In total, these assemblies generated 234,228,950 nucleotide transcript contigs (Table 1).

A1.3.5 Six-frame translation and frame selection of nucleotide sequences

Nucleotide sequences were translated in six frames with transeq vEMBOSS:6.6.0.059 (Rice *et al*., 2000) using Standard Genetic Code, to bring all MarFERReT reference material into translated amino acid sequence. The longest coding frame(s) (longest uninterrupted stretch of amino acid residues) were retained for downstream analysis.

A1.3.6 Clustering on amino acid identity of translated protein sequences

Translated protein sequences all assemblies were clustered together at the 99% amino acid sequence identity threshold with MMseqs2 (Steinegger and Söding 2018) to reduce sequence redundancy in identical or nearly identical transcript contigs across metatranscriptome assemblies. A total of 199,480,491 protein sequences were retained from the cluster centroid representatives, and the corresponding set of nucleotide transcripts (182,252,494) that these protein sequences were generated from were retained as clustered nucleotide transcripts (Table 1). The difference in totals between the clustered protein and nucleotide transcripts is from retention of multiple longest translated coding frames of equal lengths during frame selection. Clustering on the 99% amino acid identity level resulted in retention of 78% of the 234,228,950 original assembled nucleotide transcripts.

A1.3.7 Alignment of short metatranscriptome reads to clustered, assembled transcripts

To quantify assembled transcript abundance, quality-controlled short reads were aligned back to the clustered nucleotide transcripts using the *kallisto* aligner v0.46.1 (Bray et al., 2016, parameters: quant --rf-stranded).

A1.3.8 Taxonomic annotation

Clustered protein sequences were assigned taxonomic identity through protein alignment to the MarFERReT marine microbial eukaryote reference library (Chapter 2, Groussman *et al*., in prep). We used MarFERReT in conjunction with the DIAMOND protein alignment algorithm (Buchfink et al., 2015, v2.0.5.143, parameters: -e 1e-5, --top10 -f 102) to estimate the lowest common ancestor of translated sequences from matches to MarFERReT and assign a taxonomic identifier (taxID) from the NCBI Taxonomy database (Federhen 2012, https://www.ncbi.nlm.nih.gov/taxonomy). The best scoring taxID annotation (highest bitscore) was used for the associated nucleotide transcript (Table 1).

A1.3.9 Functional annotation of protein sequences

Clustered protein sequences were annotated against the Pfam 34.0 collection of 19,179 protein family Hidden Markov Models (HMMs) (Mistry et al., 2021) using HMMER 3.3 (Eddy 2011). The highest-stringency cutoff score ('trusted cutoff') assigned by Pfam to each hmm profile was used as a minimum score threshold. The best scoring Pfam annotation (highest bitscore) was used for the associated nucleotide transcript (Table 1).

A1.3.10 Data aggregation and generation of taxa bin transcript inventories

The primary annotation products (raw read counts, taxonomy, and function) were combined together to produce taxa bin transcript inventories and conduct inter-taxa normalization of transcript abundances. For transcripts with a taxonomic annotation (taxID), raw transcript counts for each nucleotide transcript were adjusted by the transcript length to fragments-per-kilobase (FPK) and normalized to transcripts-per-million (TPM) by the total sum of FPK for all transcripts with the same taxID.

A1.4 Data Records and Code Availability

Raw short reads from all metatranscriptomes in this catalog have been deposited into NCBI Short Read Archive for the Diel1 under BioProject PRJNA492142; Gradients 1, BioProject PRJNA690573. Gradients 2, Gradients 3 and G3 diel are in submission to SRA. The Zenodo repository associated with this catalog includes complete raw metatranscriptome assemblies, translated and clustered protein sequences, taxonomic annotations against MarFERReT, function annotations against Pfam 34.0, and aligned short read counts. Zenodo repository link: https://zenodo.org/record/7332796. This catalog has a github repository for the code used in these methods: https://github.com/armbrustlab/NPac_euk_gene_catalog

A1.5 Technical Validation

For all samples, extracted RNA was quantified using a Qubit fluorometer (Thermofisher) and quality controlled using a Bioanalyzer (Agilent) prior to sequencing. Samples were randomized across sequencing runs to reduce potential biases. Synthetic spike-in mRNA standards were added to samples during the RNA extraction process as previously described (Durham et al., 2019, Coesel et al., 2021). A FASTA-format file of mRNA standards is available on the Zenodo repository. Taxonomic assessment to determine efficiency of poly-A+ selection and sample size fractionation was performed following DIAMOND alignment of protein sequences against the MarFERReT reference library (Chapter 2, Chapter 3, Groussman et al., 2022 in prep). Bacterial contamination was minimal (<1% of annotated transcripts), and the broad taxonomic profile of the transcript community aligns with expectations according to size class and environmental conditions (Chapter 2, Chapter 3).

A1.6 References

Ashkezari, M. D., Hagen, N. R., Denholtz, M., Neang, A., Burns, T. C., Morales, R. L., Lee, C. P., Hill, C. N & Armbrust, E. V. (2021). Simons collaborative marine atlas project (Simons CMAP): an open-source portal to share, visualize, and analyze ocean data. *Limnology and Oceanography: Methods*, *19*(7), 488-496.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, *34*(5), 525-527.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, *12*(1), 59-60.

Coesel, S. N., Durham, B. P., Groussman, R. D., Hu, S. K., Caron, D. A., Morales, R. L., Ribalet, F., & Armbrust, E. V. (2021). Diel transcriptional oscillations of light-sensitive regulatory elements in open-ocean eukaryotic plankton communities. *Proceedings of the National Academy of Sciences*, *118*(6), e2011038118.

Durham, B. P., Boysen, A. K., Carlson, L. T., Groussman, R. D., Heal, K. R., Cain, K. R., Morales, R. L., Coesel, S. N., Morris, R. M., Ingalls, A. E., & Armbrust, E. (2019). Sulfonate-based networks between eukaryotic phytoplankton and heterotrophic bacteria in the surface ocean. *Nature microbiology*, *4*(10), 1706-1715.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS computational biology*, *7*(10), e1002195.

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic acids research*, *40*(D1), D136-D143.

Groussman, R. D., Coesel, S. N., Durham, B. P., & Armbrust, E. V. (2021). Diel-Regulated Transcriptional Cascades of Microbial Eukaryotes in the North Pacific Subtropical Gyre. *Frontiers in microbiology*, *12*.

Juranek, L. W., White, A. E., Dugenne, M., Henderikx Freitas, F., Dutkiewicz, S., Ribalet, F., Ferrón, S. E., Armbrust, E. V., & Karl, D. M. (2020). The importance of the phytoplankton "middle class" to ocean net community production. *Global Biogeochemical Cycles*, *34*(12), e2020GB006702.

Karl, D. M. (1999). A sea of change: biogeochemical variability in the North Pacific Subtropical Gyre. *Ecosystems*, *2*(3), 181-214.

Lambert, B. S., Groussman, R. D., Schatz, M. J., Coesel, S. N., Durham, B. P., Alverson, A. J., White, A. E., & Armbrust, E. V. (2022). The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. *Proceedings of the National Academy of Sciences*, *119*(7), e2100916119.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic acids research*, *49*(D1), D412-D419.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends in genetics*, *16*(6), 276-277.

Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature communications*, *9*(1), 1-8.

Wilson, S. T., Aylward, F. O., Ribalet, F., Barone, B., Casey, J. R., Connell, P. E., Eppley, J. M., Ferrón, S. E., Fitzsimmons, J. N., Hayes, C. T., Romano, A. E., Turk-Kubo, K. A., Vislova, A., Armbrust, E. V., Caron, D. A., Church, M. J., Zehr, J. P., Karl, D. M., & DeLong, E. F. (2017). Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium Crocosphaera. *Nature microbiology*, *2*(9), 1-9.

CONCLUSIONS

Microbial eukaryotes are a major constituent of marine ecosystems and carry out important functions in global biogeochemical cycling, but the complexities of their metabolism, physiology and behavior makes comprehensive understanding of their ecological functions difficult (Worden et al., 2015). Transcriptomic and metatranscriptomic methodologies have emerged as valuable approaches to understanding the functional complexities of these organisms (Caron et al., 2017). In this dissertation, I have leveraged marine eukaryote metatranscriptomes to present new findings on the diel transcript regulation of *in situ* communities in the North Pacific Subtropical Gyre, and the adaptations of picoeukaryotes to biogeochemical gradients and resource ratios in the North Pacific Transition Zone.

In the subtropical gyre, I characterized the diel transcriptional patterns of microbial eukaryotes and found notable differences in the magnitude of regulation between eukaryotic lineages (Groussman et al., 2021). Phototrophs and photomixotrophs orchestrate transcription of metabolic functions in sync with the photocycle, with dawn peaks in functions associated with photosynthesis, light harvesting, and biosynthetic processes, and a dusk transition to the TCA cycle, oxidative phosphorylation, and catabolic processes. Afternoon peaks related to protein biosynthesis and turnover indicates a coordinated proteome rearrangement in advance of the metabolic transition at dusk. Haptophytes and ochrophytes had the highest proportions of diel-regulated gene families, while dinoflagellate transcription varied little over diel cycles with the notable exception of putative plastid-associated functions. The observation of a diel-regulated assimilatory sulfur pathway and other sulfur metabolism transcripts only in dictyophytes suggests an under looked role for this lineage in the sulfur cycle.

The metatranscriptome data from this study has been analyzed with paired lipid and metabolite samples to link taxonomy and function from transcripts to biochemical measurements of notable compounds in the environment. Diel lipidome data together with transcripts involved in triacylglycerol synthesis showed the role of dinoflagellates, haptophytes, and other eukaryotic plankton in the daily flux of fixed carbon stored in energy-rich triacylglycerols (Becker et al., 2018). The transcriptomes were combined with metabolite and whole-community (prokaryotic) metatranscriptomes to uncover the exchange of organic sulfur compounds between eukaryotic phytoplankton and heterotrophic bacteria over diel scales in the North Pacific Subtropical Gyre

(Durham et al., 2019), and the transcriptional profile of microbial eukaryote genes related to metabolites with diel rhythms in concentration (Boysen et al., 2021).

Detailed examination of light-sensitive regulatory elements in marine eukaryotic phytoplankton leveraged the diel metatranscriptome assemblies and transcript abundances to resolve the diel synchrony of photoreceptors tuned to different wavelengths of light and uncovered a diversification of light-responsive photoreceptor families in haptophytes and photosynthetic stramenopiles (Coesel et al., 2021). The separation of taxa along trophic gradients observed in transcriptional profile ordination inspired the development of a machine-learning method to estimate the trophic mode of species in transcriptomes and metatranscriptomes (Lambert et al., 2022), and the diel regulation of flavodoxin transcripts in diatoms (van Creveld et al., 2022).

To improve taxonomic annotations of metatranscriptomes, I constructed a new marine microbial eukaryote reference sequence library (MarFERReT), incorporating advances in sequencing new reference material in a reproducible and versioned framework (Groussman et al., in submission). In building MarFERReT, I generated inventories of core transcribed genes for use in assessing the coverage of environmental metatranscriptome bins and showed the effect of novel sequence inclusions on improving annotation efficiency and specificity. I provide public access to the MarFERReT resources on stable repositories, with freely available code, metadata, and Case Study tutorials to advance the accessibility and reproducibility of this library.

The MarFERReT reference library has contributed to taxonomic annotations in North Pacific Eukaryotic Gene Catalog (NPEGC, Appendix 1) and several projects in development (Coesel et al., in prep, Groussman et al., in prep.). The metatranscriptomes described in NPEGC were the foundation of the findings in Chapter 3; where I characterized the picoeukaryote communities of the North Pacific Transition Zone by their transcriptional inventory across three cruise transects. Resource ratio incubation experiments allowed me to identify significantly responsive species to altered nutrient ratios and highlighted the role of osmo- and phago-mixotrophs in responding dynamically to changing nutrient conditions. Differentially regulated transcript functions in these responders suggests that flexibility along the trophic spectrum is a common and advantageous strategy.

This dissertation has elucidated numerous strategies that marine microbial eukaryotes use to sustain, survive and thrive in their environments; ranging from the fine-tuned diel transcription of metabolic machinery in the North Pacific Subtropical Gyre, to the rapid response of species with mixed trophic modes under changing nutrient conditions in the North Pacific Transition Zone. In the process, I have developed public resources to enhance the taxonomic annotation and improve the reproducibility and accessibility of these valuable metatranscriptome data sets, so they can continue to provide insight into the molecular ecology of microbial eukaryotes in the oceans.

CONCLUSION REFERENCES

Becker, K. W., Collins, J. R., Durham, B. P., Groussman, R. D., White, A. E., Fredricks, H. F., ... & Van Mooy, B. A. (2018). Daily changes in phytoplankton lipidomes reveal mechanisms of energy storage in the open ocean. *Nature communications*, *9*(1), 1-9.

Boysen, A. K., Carlson, L. T., Durham, B. P., Groussman, R. D., Aylward, F. O., Ribalet, F., ... & Ingalls, A. E. (2021). Particulate metabolites and transcripts reflect diel oscillations of microbial activity in the surface ocean. *Msystems*, *6*(3), e00896-20.

Boysen, A. K., Durham, B. P., Kumler, W., Key, R. S., Heal, K. R., Carlson, L. T., ... & Ingalls, A. E. (2022). Glycine betaine uptake and metabolism in marine microbial communities. *Environmental microbiology*, *24*(5), 2380-2403.

Caron, D. A., Alexander, H., Allen, A. E., Archibald, J. M., Armbrust, E., Bachy, C., ... & Worden, A. Z. (2017). Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology*, *15*(1), 6-20.

Coesel, S. N., Durham, B. P., Groussman, R. D., Hu, S. K., Caron, D. A., Morales, R. L., Ribalet, F., & Armbrust, E. V. (2021). Diel transcriptional oscillations of light-sensitive regulatory elements in open-ocean eukaryotic plankton communities. *Proceedings of the National Academy of Sciences*, *118*(6).

Durham, B. P., Boysen, A. K., Carlson, L. T., Groussman, R. D., Heal, K. R., Cain, K. R., ... & Armbrust, E. (2019). Sulfonate-based networks between eukaryotic phytoplankton and heterotrophic bacteria in the surface ocean. *Nature microbiology*, *4*(10), 1706-1715.

Groussman, R. D., Coesel, S. N., Durham, B. P., & Armbrust, E. V. (2021). Diel-Regulated Transcriptional Cascades of Microbial Eukaryotes in the North Pacific Subtropical Gyre. *Frontiers in microbiology*, *12*.

Lambert, B. S., Groussman, R. D., Schatz, M. J., Coesel, S. N., Durham, B. P., Alverson, A. J., White, A. E., & Armbrust, E. V. (2022). The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. *Proceedings of the National Academy of Sciences*, *119*(7), e2100916119.

van Creveld, S. G., Coesel, S. N., Blaskowski, S., Groussman, R. D., Schatz, M. J., & Armbrust, E. V. (2022). Divergent functions of two clades of flavodoxin in diatoms mitigate oxidative stress and iron limitation. *bioRxiv*.

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.