5-2023

Ecology, Evolution, and Gene Transfer Between Diatoms and Bacteria

Cory B. Gargas
University of Arkansas-Fayetteville

Ecology, Evolution, and Gene Transfer Between Diatoms and Bacteria

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Biology

by

Cory B. Gargas
Kent State University
Bachelor of Science in Conservation Biology, 2013
John Carroll University
Master of Science in Biology, 2018

May 2023
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

_____	_____
Andrew Alverson, Ph.D. Dissertation Director	Jeremy Beaulieu, Ph. D. Committee Member
_____	_____
Jeffery Lewis, Ph.D. Committee Member	Shady Amin, Ph.D. Committee Member

Abstract

Although photosynthetic macro-eukaryotes (i.e., plants) make up the majority of organic biomass on earth, bacteria are the second largest taxonomic group, by biomass. Bacteria are ubiquitous in our environment, living on, and within, man-made surfaces, natural environments, and eukaryotes themselves. The relationship between bacteria and eukaryotes has existed from the very beginning of eukaryotic life in the form of bacterial endosymbioses that resulted in mitochondria and plastids. Other eukaryote–bacteria relationships have evolved since then, ranging from the beneficial (e.g., mutualistic) to harmful (e.g., parasitic or pathogenic). Understanding these eukaryote–bacteria relationships is key to understanding both the evolution of important ecosystem processes and how these interactions affect human endeavors such as agriculture. To better understand how bacterial communities affect and interact with their eukaryotic partners, we have utilized the genomes and transcriptomes of the ubiquitous micro-eukaryotes known as diatoms to analyze their co-occurring bacteria.

This dissertation explores the composition, dynamics, and interactions between diatoms and their bacterial partners. We first sequenced the genome and transcriptome of the araphid pennate diatom *Psammoneis japonica* and examined its associated bacterial metagenome. Repetitive element content in *P. japonica,* and other existing diatom genomes, were found to have a positive relationship with genome size. The partial metagenome of P. japonica revealed a diverse microbial community of at least 25 associated bacterial taxa, including four near-complete genomes for novel species of Planctomycetota, α-proteobacteria, and Bacteroidota. The *P. japonica* genome was found to contain genes and intergenic open reading frame sequences which were transferred to the *P. japonica* lineage from members of the lineages of several cohabiting bacteria. Several of these HGT candidate proteins are located in regions with transposon densities higher than the average for the genic and intergenic regions of the *P. japonica* genome.

Subsequently, we mapped and extracted bacterial 16S sequences from existing transcriptome reads of diatoms. Transcriptomes were sourced from the Alverson Lab and the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) to investigate bacterial diversity, community phylogenetics, and cophylogenetic concordance between diatom-bacteria associations across cultured diatom strains. There was a high degree of dissimilarity in phylogenetic beta-diversity between diatom bacterial communities at all taxonomic levels of the diatom tree of life. Ordination analysis of phylogenetic beta-diversity demonstrated distinct groupings of diatom microbiomes by salinity. Significant cophylogenetic concordance was found between diatoms of the genus Chaetoceros and their bacterial partners. These results support that diatom phycosphere communities are more similar within salinity levels, while still maintaining high diversity within and across genera. Lastly, this research demonstrates that incidentally collected sequence data can be utilized to investigate microbiomes.

These experiments highlight that incidentally collected sequence data can be utilized to investigate the algal phycosphere by using bioinformatics methods to extract bacterial sequences from xenic algal cultures, as well as how normally discarded data can be used to examine community dynamics that would otherwise be overlooked. These findings also suggest that diatom–bacteria relationships are stable over evolutionary timescales and can lead to recurrent horizontal gene transfer events from symbiont to host, as well as cophylogenetic concordance between diatoms and their bacterial partners.

Table of Contents

List of Tables

Chapter 2

Chapter 3

List of Figures

Chapter 2

Figure 3.6. NMDS analysis of unweighted UniFrac distances between (Alverson *et al.*, 2023) transcriptomes, the best solution was found with a stress value of 0.24 after 20 permutations. A) Bacterial beta-diversity between transcriptomes, colored by salinity with individual points depicting pairwise beta-diversity comparisons between transcriptomes. B) OTUs of the ten most abundant phyla, colored and faceted by phylum. Ellipses generated using stat = “norm.”

Figure 3.7. Cophylogentic concordance tanglegram of bacterial OTUs for the diatom genus *Chaetoceros.* Individual links in the tanglegram are colored according to their level of cophylogenetic signal, will lower values (yellow) indicating more phylogenetic concordance (higher cophylogenetic signal) and greater values indicating less phylogenetic concordance (lower cophylogenetic signal).

Figure 3.8. Density plot of squared residual values from our cophylogenetic concordance analyses for the *Chaetoceros*–bacteria subset. Individual points represent global best fit values (m^2_{xy}) for individual links between *Chaetoceros* species and their bacteria. Values closer to zero indicate a greater contribution to cophylogenetic signal.

Chapter 1 Introduction

1.1 Diatoms in microbiome research.

Diatoms are a group of unicellular stramenopile algae with unique silica-based cell walls and a near ubiquitous distribution wherever water is present. Diatoms are responsible for ≥ 40% of marine primary production and 25% of global net primary production (Nelson et al. 1995). As with all eukaryotes, the most intimate association diatoms have with bacteria is in the form of endosymbioses with their mitochondria and secondarily derived plastids (*The Molecular Life of Diatoms* 2022). Beyond these endosymbioses, almost all eukaryotes form intimate ectosymbioses with bacteria. These symbioses range from parasitic interactions of bacteria upon their hosts to mutualistic interactions in which both the host and symbiont benefit from the interaction (Shady A. Amin, Parker, and Armbrust 2012; Kazamia et al. 2016; Seymour et al. 2017). Whereas multicellular eukaryotes may have multiple niches or specialized organs in which their bacterial symbionts dwell, single-celled eukaryotes such as diatoms have evolved complex interactions and signaling mechanisms to maintain and attract their symbionts over, relatively, large scales.

The interactions between diatoms and their bacterial partners are of particular interest to aquatic ecologists, algal bloom researchers, and researchers in industrial applications such as biofuels and pharmaceuticals. Until relatively recently, one of the major issues with past research on microbial ecology was the inability to culture, and sequence, the majority of bacterial strains present in the environment. The advent of high-throughput, short- and long-read sequencing has enabled the direct sequencing of environmental samples and the construction of metagenomes from which the composition and abundance of bacterial consortia can be investigated without disregarding unculturable bacteria (Kirubakaran et al. 2020).

Investigations into diatom–bacteria relationships have found interactions between bacteria and phytoplankton support both biological and geochemical interactions (Cole 1982; Azam and Malfatti 2007; Seymour et al. 2017). At their most basic, these interactions consist of

phytoplankton dependent on bacteria for the remineralization of organic matter back to inorganic forms that support their growth (Field et al. 1998; Falkowski, Fenchel, and Delong 2008). More intimate interactions exist, such as in diatoms that are auxotrophic for B vitamins acquiring them from bacteria (Durham et al. 2015, 2017). Other diatoms actively promote the attachment and growth of beneficial bacteria, while also suppressing the attachment and growth of non-beneficial bacteria, via the rosmarinic and azelaic acid secretion (Shibl et al. 2020). Likewise, bacteria can promote the growth of some diatoms by converting algal-secreted tryptophan into indole-3-acetic acid (IAA), which enhances cell division and may potentially increase its carbon output to bacteria (Amin et al. 2015; Segev et al. 2016). A great diversity of interactions and relationships have been recorded from research such as the aforementioned, but broader studies that span the entirety of the diatom tree have been lacking. As such I focused on utilizing existing and de novo transcriptomic and genomic resources to examine global diatom–bacteria relationships and how they change according to salinity and diatom systematics. I also employed more focused techniques to examine a single diatom species and its metagenome, with special consideration for whether these intimate relationships have allowed for any horizontal gene transfer from bacterial partners to their diatom host.

1.2 Dissertation Outline

This dissertation presents unique research focusing on using data scavenging techniques to characterize and analyze the bacterial consortia of diatoms using existing genomic and transcriptomic sequence resources. Chapter 2 characterizes the bacterial consortia of existing diatom cultures by extracting 16S sequences from ≥ 200 existing transcriptomes. This allowed for the exploration of changes in community composition across different groups of diatoms and their habitats, as well as an exploration of co-evolutionary dynamics between diatoms and their bacterial communities. Chapter 3 explores and characterizes the genome and partial metagenome of the araphid pennate diatom *Psammoneis*

japonica. In doing so, I retrieved four complete bacterial genomes and several 16S sequences for additional bacteria. Strong evidence for horizontal gene transfer between the lineages of these co-occurring bacteria and the *P. japonica* lineage was also found.

1.3 References

Amin, S. A., L. R. Hmelo, H. M. van Tol, B. P. Durham, L. T. Carlson, K. R. Heal, R. L. Morales, et al. 2015. “Interaction and Signalling between a Cosmopolitan Phytoplankton and Associated Bacteria.” *Nature* 522 (7554): 98–101.

Amin, Shady A., Micaela S. Parker, and E. Virginia Armbrust. 2012. “Interactions between Diatoms and Bacteria.” *Microbiology and Molecular Biology Reviews: MMBR* 76 (3): 667–84.

Azam, Farooq, and Francesca Malfatti. 2007. “Microbial Structuring of Marine Ecosystems.” *Nature Reviews. Microbiology* 5 (10): 782–91.

Cole, Jonathan J. 1982. “Interactions Between Bacteria and Algae in Aquatic Ecosystems.” *Annual Review of Ecology and Systematics* 13 (1): 291–314.

Durham, Bryndan P., Stephen P. Dearth, Shalabh Sharma, Shady A. Amin, Christa B. Smith, Shawn R. Campagna, E. Virginia Armbrust, and Mary Ann Moran. 2017. “Recognition Cascade and Metabolite Transfer in a Marine Bacteria-Phytoplankton Model System.” *Environmental Microbiology* 19 (9): 3500–3513.

Durham, Bryndan P., Shalabh Sharma, Haiwei Luo, Christa B. Smith, Shady A. Amin, Sara J. Bender, Stephen P. Dearth, et al. 2015. “Cryptic Carbon and Sulfur Cycling between Surface Ocean Plankton.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (2): 453–57.

Falkowski, Paul G., Tom Fenchel, and Edward F. Delong. 2008. “The Microbial Engines That Drive Earth’s Biogeochemical Cycles.” *Science* 320 (5879): 1034–39.

Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski. 1998. “Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components.” *Science* 281 (5374): 237–40.

Kazamia, Elena, Katherine Emma Helliwell, Saul Purton, and Alison Gail Smith. 2016. “How Mutualisms Arise in Phytoplankton Communities: Building Eco-Evolutionary Principles for Aquatic Microbes.” *Ecology Letters*. https://doi.org/10.1111/ele.12615.

Kirubakaran, Rangasamy, K. N. ArulJothi, Sundaravadivel Revathi, Nowsheen Shameem, and Javid A. Parray. 2020. “Emerging Priorities for Microbial Metagenome Research.” *Bioresource Technology Reports* 11 (September): 100485.

Nelson, David M., Paul Tréguer, Mark A. Brzezinski, Aude Leynaert, and Bernard Quéguiner. 1995. “Production and Dissolution of Biogenic Silica in the Ocean: Revised Global Estimates, Comparison with Regional Data and Relationship to Biogenic Sedimentation.” *Global Biogeochemical Cycles* 9 (3): 359–72.

Segev, Einat, Thomas P. Wyche, Ki Hyun Kim, Jörn Petersen, Claire Ellebrandt, Hera Vlamakis, Natasha Barteneva, et al. 2016. “Dynamic Metabolic Exchange Governs a Marine Algal-Bacterial Interaction.” *eLife* 5 (November). https://doi.org/10.7554/eLife.17473.

Seymour, Justin R., Shady A. Amin, Jean-Baptiste Raina, and Roman Stocker. 2017. “Zooming in on the Phycosphere: The Ecological Interface for Phytoplankton–bacteria Relationships.”

Nature Microbiology 2 (May): 17065.

Shibl, Ahmed A., Ashley Isaac, Michael A. Ochsenkühn, Anny Cárdenas, Cong Fei, Gregory Behringer, Marc Arnoux, et al. 2020. “Diatom Modulation of Microbial Consortia Through Use of Two Unique Secondary Metabolites.” *bioRxiv*. Microbiology. https://doi.org/10.1101/2020.06.11.144840.

The Molecular Life of Diatoms. 2022. Springer International Publishing.

Signatures of a long-term diatom–bacterial association in the genome and metagenome of of the diatom *Psammoneis japonica*

Cory B. Gargas[4], Matthew Parks[1], Elias Spiliotopoulos[2], Marissa Ashner[3], Matthew P. Ashworth, Eveline Pinseel[4], Wade R. Roberts[4], Elizabeth C. Ruck[4], Nina Denne[5], Anni Wang[6], Sarah Schaak[2], Shady Amin[8], Norman J. Wickett[7], Andrew J. Alverson[4]

[1]Department of Biology, University of Central Oklahoma, Edmond, Oklahoma

[4]Department of Biological Sciences, University of Arkansas, Fayetteville, Arkansas

[2]Biology Department, Reed College, Portland, Oregon

[3]Illinois Institute of Technology, Chicago, Illinois

[5]Carleton College, Northfield Minnesota

[6]Florida State University, Tallahassee, Florida

[7]Clemson University, Clemson, South Carolina

[8]New York University - Abu Dhabi, Abu Dhabi, United Arab Emirates

Author Contributions: I performed final eukaryotic and bacterial genome analyses, code production, figures and tables, and wrote the manuscript. Matthew parks assembled genomes. Marissa Ashner, Nina Denne, and Anni Wang assisted with genome analysis. Elias Spiliotopoulos, Sarah Schaak, and Shady Amin assisted with bacterial genome analysis. Elizabeth Ruck performed RNA and DNA extraction. Eveline Pinseel performed additional analysis and provided comments and edits to this manuscript. Wade Roberts performed, and assisted, with additional analyses, produced figure 2, and provided comments and edits to this manuscript. Andrew Alverson and Norman Wickett provided project conception and funding. Andrew Alverson also assisted with analysis and provided comments and edits to this manuscript.

2.1 Abstract

Diatoms (Bacillariophyta) are a diverse lineage of photosynthetic algae with important roles in global carbon and nutrient cycling, and share critical metabolic interactions with bacteria inhabiting the diatom phycosphere. I sequenced the nuclear genome and transcriptome of *Psammoneis japonica*, a chain-forming, benthic pennate diatom. The nuclear genome of *P. japonica* is 91.4 Mbp in length, with 15,170 predicted genes making up 27% of the total genome, repetitive elements accounting for 33% of the genome, and other non-coding elements comprising the remaining 40% of the genome. Repetitive elements were found to have a positive relationship with genome size. The partial metagenome of *P. japonica* revealed a diverse microbial community of at least 25 associated bacterial taxa, including four near-complete genomes for novel species of Planctomycetota, α-proteobacteria, and Bacteroidota. The *P. japonica* genome contains genes and potential pseudogenes which were transferred from several cohabiting bacteria. A total of 17 genic and 40 intergenic HGT candidate proteins were found. Three intergenic ORFs were found to form a potential pseudogene in the intergenic regions of the *P. japonica* genome. Several of these HGT candidate proteins are located in regions with transposon densities higher than the average for the genic and intergenic regions of the *P. japonica* genome. These findings suggest that diatom–bacteria relationships are stable over evolutionary timescales and can lead to recurrent horizontal gene transfer events from symbiont to host.

2.2 Introduction

Since the evolution of eukaryotic cells, prokaryotes and eukaryotes have formed intimate associations.Although the most intimate and well known of these associations would be the endosymbiotic events that first gave rise to eukaryotic structures such as plastids and mitochondria (Archibald, 2015), prokaryotic-eukaryotic interactions span a wide array of beneficial, and detrimental, relationships. Bacteria in the microbiome of ruminant mammals aid in digestion (Cholewińska *et al.*, 2020), some bacteria associating with cicadas have evolved to become essential intracellular symbiotes (McCutcheon *et al.*, 2009), and the bacteria in the rhizosphere of plants are essential for plant health and metabolism (Berendsen *et al.*, 2012). While the associations of multicellular eukaryotes and prokaryotes are more thoroughly studied (Dubilier *et al.*, 2008; Müller *et al.*, 2016), the associations between prokaryotes and microbial eukaryotes (protists) are often overlooked (Keeling & Burki, 08/2019; Burki *et al.*, 2020). Protist–prokaryote interactions have been characterized across the majority of well-studied eukaryotic supergroups (Husnik *et al.*, 2021). Some of the best documented protist-prokaryote associations occur within ciliates and amoebozoans (Horn & Wagner, 2004; Görtz, 2006). In isolates of the amoebozoan genus *Acanthamoeba,* 25% of isolates have been found to harbor obligate intracellular bacteria that are unculturable outside of their hosts (Fritsche *et al.*, 1993). Social amoeba of the genus *Dictyostelium* utilize bacteria as food by ‘farming’ some and using others as defensive symbionts to protect their food source (Brock *et al.*, 2013). Likewise, ciliates in the genus *Paramecium* play host to a wealth of bacterial symbionts (Görtz, 2006). For example, the giant ciliate *Zoothamnium niveum* plays host to a seemingly mutualistic relationship with its obligate ectosymbiont *Candidatus Thiobios zoothamnicoli,* a chemoautotrophic bacteria (Bright *et al.*, 2014). Bacterial associations between photosynthetic protists (algae), such as dinoflagellates and diatoms (Foster & Zehr, 2019), are far less studied, but recent research is ameliorating this.

The area around algal cells in which algae-microbe interactions take place is termed the phycosphere and operates similarly to the rhizosphere in plants (Bell & Mitchell, 1972; Seymour *et al.*, 2017). The phycosphere is a driving force in the structure and dynamics of planktonic ecosystems, especially in regards to large, blooom-forming eukaryotic microalgae, such as diatoms (Amin *et al.*, 2012). Diatoms are some of the most abundant, diverse, and ecologically important lineages of photosynthetic protists (Armbrust, 2009; Mann & Vanormelingen, 2013). Diatom–bacteria interactions in the phycosphere range from mutualistic to parasitic (Amin *et al.*, 2012), and can influence biogeochemical cycles and ecosystems. For example, cobalamin (B_{12}) is an essential vitamin for growth in diatoms, but many are unable to synthesize it (Croft *et al.*, 2005). Since diatoms are unable to synthesize B_{12}, they must acquire it exogenously from bacterial partners (Durham *et al.*, 2015; Bertrand *et al.*, 2015). Iron is another growth limiting micronutrient, and diatom-associated bacteria of the genus *Marinobacter* have been demonstrated to promote iron assimilation in diatoms, and other algae, by facilitating photochemical redox cycling of iron via the production of the siderophore vibrioferrin (Amin *et al.*, 2009). Another highly refined and entrenched diatom-bacteria interaction involves the exchange of tryptophan. The diatom *Pseudo-nitzschia multiseries* secretes tryptophan, which is taken up by the bacterium *Sulfitobacter* sp. SA11 and converted into indole-3-acetic acid (Amin *et al.*, 2015; Segev *et al.*, 2016). *Pseudo-nitzschia multiseries* then absorbs the indole-3-acetic acid, enhancing diatom cell division and potentially increasing carbon output to the bacteria (Amin *et al.*, 2015; Segev *et al.*, 2016).

Currently, diatom–bacteria associations are only known to take place with Proteobacteria (alpha- & gamma-), Bacteroidetes, and Cyanobacteria (Amin *et al.*, 2012; Helliwell *et al.*, 2022). Instances of high host-specificity (Mönnich *et al.*, 2020), as well as diatom (Shibl *et al.*, 2020a) and bacteria controlled community modulation (Majzoub *et al.*, 2019), have been observed for

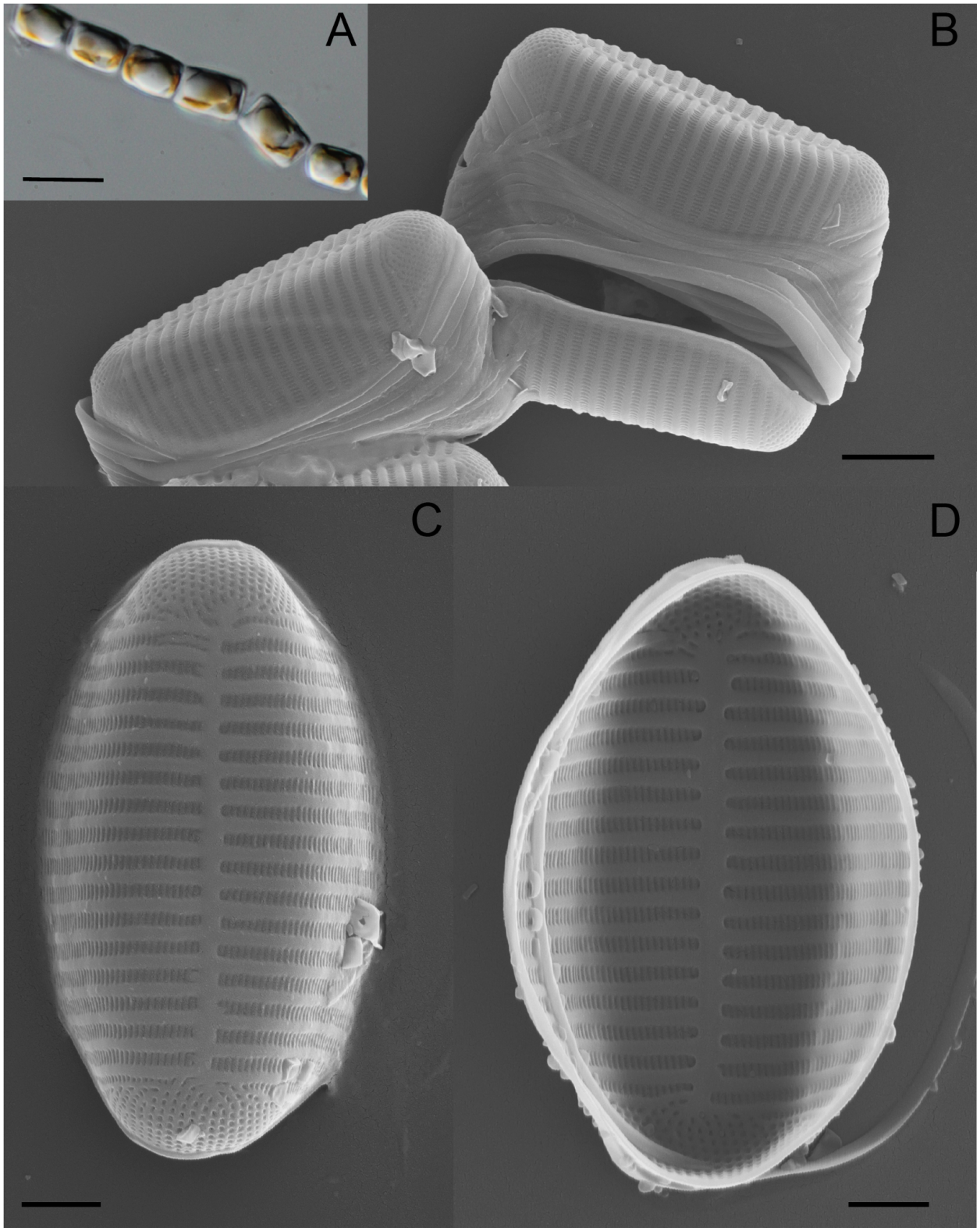

Figure 2.1. Light and scanning electron micrographs (SEMs) of *Psammoneis japonica*. Image A) girdle view of live cells in culture, B–D) SEMs of frustules in B) girdle view, C) valve view, and D) interior view of valve.

diatoms. This suggests that some bacterial lineages are adapted to specific interactions with diatoms and that these relationships might have evolved over long timescales or been enabled via horizontal gene transfer (HGT) events. It has been demonstrated that some diatom-bacterial communities are stable in laboratory culture for >1 year (Behringer *et al.*, 2018). Evidence for these associations persisting over evolutionary times-scales comes from genomic studies in which several instances of HGT between bacteria and diatoms have been documented (Bowler *et al.*, 2008; Vancaester *et al.*, 2020). Understanding the longevity of these relationships is key to understanding how microbial interactions develop and affect the co-evolution of diatoms and their bacterial partners (Brodie *et al.*, 2017).

I expand the ecological and phylogenetic diversity of sequenced diatom genomes by sequencing the genome of the colonial, marine, araphid diatom, *Psammoneis japonica* Shin.Sato, Kooistra & Medlin. I also recovered and characterized the partial metagenome of *P. japonica.* I was able to recover the genomes of four co-occurring bacterial strains and 16S rDNA sequences from many other species, providing a broader view of the phycosphere. Through comparison of bacterial and diatom genomes, I detected HGT between all four bacterial genomes and the *P. japonica* genome. These HGT candidate genes were found in both genic and intergenic regions of the *P. japonica* genome. These findings support the long-term stability of diatom-bacteria relationships.

2.3 Methods

2.3.1 Diatom culturing and RNA extraction and sequencing

The ECT2AJA-110 strain of *Psammoneis japonica* was grown from a sample collected in 2014 by Christopher Lobban at Outhouse Beach, Guam (13.464200, 144.655000). This *P. japonica* strain was maintained under 12h:12h light:dark conditions in L1 medium at 23° C.

Daily growth rates were estimated based on chlorophyll-*a* fluorescence with a Trilogy Laboratory Fluorometer (Turner Designs, Sunnyvale, CA, USA). Cells were harvested during exponential growth and concentrated by centrifugation at 3,000 rpm for 10 minutes at 4° C. RNA was isolated from bead-disrupted cells with a Qiagen RNeasy Kit ® (Qiagen, Venlo, the Netherlands). DNase-treated RNA was quantified with a Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA, USA), and the RNA quality was assessed with a TapeStation 2200 (Agilent Technologies, Santa Clara, CA, USA). An RNA library was constructed using 1 ug of RNA using the TruSeq RNA Sample Preparation Kit v2 (set A adapters) and was sequenced at the Beijing Genomics Institute on the Illumina HiSeq2000 platform.

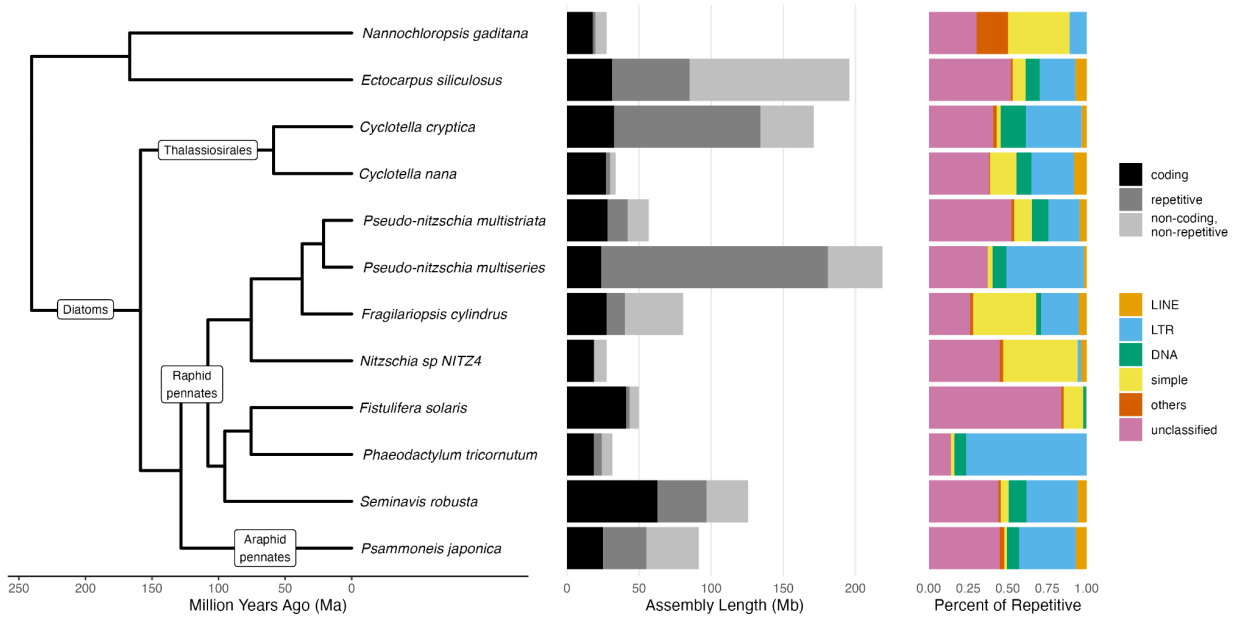

Figure 2.2. Phylogenetic position of *Psammoneis japonica* inferred from single-copy orthologs identified by OrthoFinder. From left to right, bar plots display assembly length megabases (Mb), colored by content, and percents of repetitive sequence content.

DNA extraction and sequencing

Cell pellets were bead-disrupted and DNA was extracted using the Qiagen Plant Mini Kit (Qiagen, Venlo, the Netherlands). Library construction and sequencing was performed by the University of Delaware DNA Sequencing and Genotyping Center using the Illumina HiSeq2500 and the PacBio RSII platforms.

2.3.2 Illumina DNA sequence filtering and assembly

Raw Illumina reads were error corrected using ACE (Sheikhizadeh & de Ridder, 2015) using default settings and with an estimated genome size of 90Mb (based on preliminary kmer-based estimates using Jellyfish ver. 2.1.3 (Marçais & Kingsford, 2011) and GenomeScope (http://qb.cshl.edu/genomescope/). Resulting error-corrected reads were quality-trimmed and cleared of remaining adapter sequence using Trimmomatic ver. 0.32 (Bolger *et al.*, 2014)with the following settings and using Illumina TruSeq adapter sequences: ILLUMINACLIP:TruSeq_adapters.fa:2:30:10 LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:80. Error correction, and quality and adapter filtering resulted in a decrease of 14.4% of read pool size.

A preliminary assembly of error-corrected and trimmed reads was completed using Ray ver. 2.3.1 (Boisvert et al. 2012) with kmer value of 31. Overall assembly quality and contamination levels were estimated numerically and using taxon-annotated GC-coverage (TAGC) plots as implemented in Blobtools (Laetsch & Blaxter, 2017; https://github.com/DRL/blobtools). Blast searches used for generating TAGC plots here and below were completed using BLAST ver. 2.2.9 (Altschul *et al.*, 1990) and the NCBI nt database (ftp://ftp.ncbi.nlm.nih.gov/blast/; downloaded 20 November 2015). Assembly N50 was 8.5 Kbp and the longest assembled sequence was over 190 Kbp, however the TAGC plot for this assembly revealed a significant number of contigs from exogenous sources. From these results, reads in the original error-corrected and quality-trimmed read pool were filtered to exclude all reads mapping to contigs meeting any of the following criteria: 1) GC content less than 0.35 or greater than 0.65, 2) coverage depth <15, 3) GC content >0.52 and coverage depth <25, 4) only non-eukaryotic BLAST hits, at least one of which had E-value $<10^{-5}$ and alignment length >200 bp. This filtering resulted in a further decrease in read pool size of 9.6%, with a final read pair count of 40 Mbp.

Contaminant-filtered reads were then re-assembled using Ray but with a higher kmer value (kmer = 67), resulting in a final Illumina assembly with fewer assembled contigs, smaller total assembly size and increased N50. Contigs from this assembly were subsequently used in PacBio assembly efforts as described below, while the error-corrected, quality, adapter, and contaminant filtered read pool was used further in contaminant filtering and to generate final coverage estimates and TAGC plot, as described below.

2.3.3 Illumina RNA sequence filtering and assembly

Raw Illumina RNA sequence data was cleaned and filtered using the Perl script rnaseq_clean_filter.pl (https://github.com/andrewalverson/RNAseq/blob/master/rnaseq_clean_filter.md), which uses Trimmomatic and Bowtie2 (Langmead & Salzberg, 2012) to quality-trim reads and filter Illumina adapter sequences, and filters reads with BLAST homology to databases of common sequencing vectors, diatom organellar genomes, and diatom rRNA sequences. In addition, BBNorm and BBMerge of BBMap ver. 0.35 (Bushnell *et al.*, 2017) are used for kmer normalization of identified nuclear RNA reads and merging of these reads prior to subsequent assembly. rnaseq_clean_filter.pl uses default settings in Bowtie2 with the local (--local) alignment strategy, and the following Trimmomatic parameters: ILLUMINACLIP:TruSeq_adapters.fa:2:40:15 HEADCROP:10 LEADING:5 TRAILING:5 AVGQUAL:32 MINLEN:72. The following parameters were used to run rnaseq_clean_filter.pl: --min_kmer_rna=1 --min_kmer_org=1 --min_kmer_nuc=1 --window=4:2 --min_len=30 --reference=phaeo.

Filtered and merged paired-end RNA sequence data was assembled with Trinity ver. 2.0.2 (Grabherr *et al.*, 2011), and transcript abundance for the resulting nuclear transcriptome assembly was estimated using the RSEM module of Trinity ver. 2.0.2. Assembly quality was assessed using TransRate ver. 1.0.1 (Smith-Unna *et al.*, 2016). Open reading frames were predicted from the Trinity assembly using TranDecoder.LongOrfs from the TransDecoder ver.

2.0.1 module of Trinity. Identified reading frames were then searched for protein homology by BLAST search against the UniProt Swiss-Prot database (www.uniprot.org, downloaded 11 February 2016) with an E-value cutoff of 10^{-3}, and with hidden Markov modeling using the hmmscan module of TransDecoder against the Pfam database (Finn *et al.*, 2016; downloaded 14 April 2016). Resulting homology information from these searches was applied to final translations using the TransDecoder.Predict module. Redundant and overlapping amino acid sequences were removed using Cd-hit ver. 4.6.5 (Li & Godzik, 2006) with parameter settings –c 0.99 –n 5 to form the final representative pool of translated protein sequences from the transcriptome assembly.

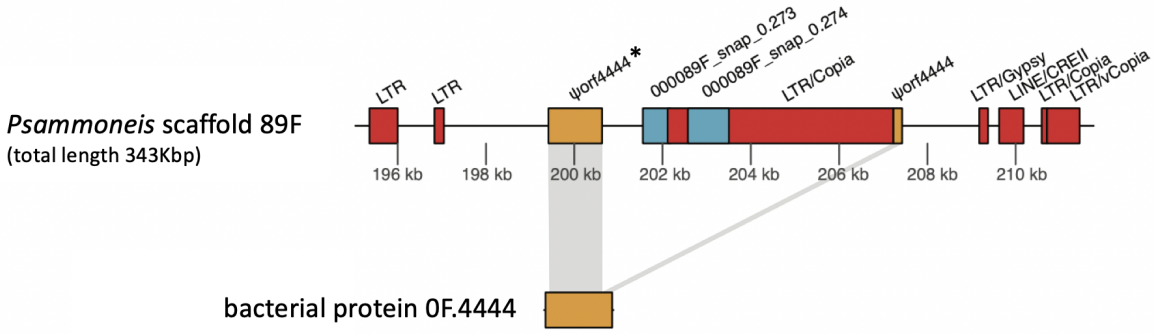

Figure 2.3. Location of ORF loci with BLASTP homology to single bacterial protein. Asterisk (*) indicates that two ORF loci overlap at this location.

PacBio DNA sequence filtering and assembly

Reads from PacBio sequencing were initially assembled using Falcon ver. 0.4.0 (https://github.com/PacificBiosciences/FALCON) with length cutoff of 7000, minimum coverage of 3 and max coverage depth and difference of 100. This resulted in assembly of 900 contigs with N50 of 333 Kbp and total assembly length of 118 Mbp. Similar to the initial assembly with Illumina DNA sequence data, a TAGC plot of this assembly suggested substantial amounts of exogenous contamination. Included in this putative contamination were four large contigs with lengths ranging from ca. 3.5 – 5.5 Mbp and extensive homology to bacterial genomes, suggesting assembly of complete or nearly complete exogenous bacterial genomes. Based on these results, the original PacBio read pool was filtered to remove reads with primary mapping

to contigs meeting any of the following criteria: 1) GC content <0.40 and strongest blast hit to either chloroplast or mitochondrial sequence, 2) any of the identified four large bacterial contigs, 3) any contig with bacterial blast hit with at least 98% identity and GC content not between 0.46 and 0.52. In total, there were 36 contigs falling under these constraints, resulting in a ca. 67.6% decrease in the pool of PacBio reads.

The resulting contaminant-filtered read pool was re-assembled using Falcon with identical settings, resulting in fewer assembled contigs, smaller total assembly size, and a slight decrease in N50. To test whether assembled Illumina contigs might further improve the PacBio assembly, all 24,197 Illumina contigs from the final Ray assembly of error-corrected and quality, adapter, and contaminant-filtered Illumina reads were added to the contaminant-filtered PacBio read pool in order to 'masquerade' as PacBio reads. This combined pool was assembled again in Falcon using the same settings, and resulted in a decrease in the number of contigs assembled, slight increase in total assembly length, and an increase in N50. This combined PacBio/Illumina contig assembly served as the penultimate assembly, from which remaining contaminant contigs were removed in order to capture the final assembly, as described below.

2.3.4 Final contaminant filtering strategy

In addition to GC content and BLAST-identified homology, two other primary strategies were used in identification and removal of remaining contaminant contigs from the penultimate assembly. The first of these strategies involved discriminating between putative endogenous and putative exogenous contigs based on coverage depth. Reads from the error-corrected and quality, adapter, and contaminant-filtered Illumina read pool were mapped to assembled contigs using BWA-MEM ver. 0.7.12 with default settings (Li, 2013). Based on these results, a cutoff of 3x coverage depth was used initially to discriminate putative endogenous contigs (>3 depth, 'high coverage') from putative exogenous contigs (<3× depth, 'low coverage'). Similarly, filtered nuclear Illumina RNA reads were also mapped to the assembly, and again coverage depth was used as an indicator of likely endogenous or exogenous origin for a contig. For RNA read data,

a cutoff of 1×was used to discriminate ‘low coverage’ from ‘high coverage’ contigs, based on the coverage depth distribution of all contigs.

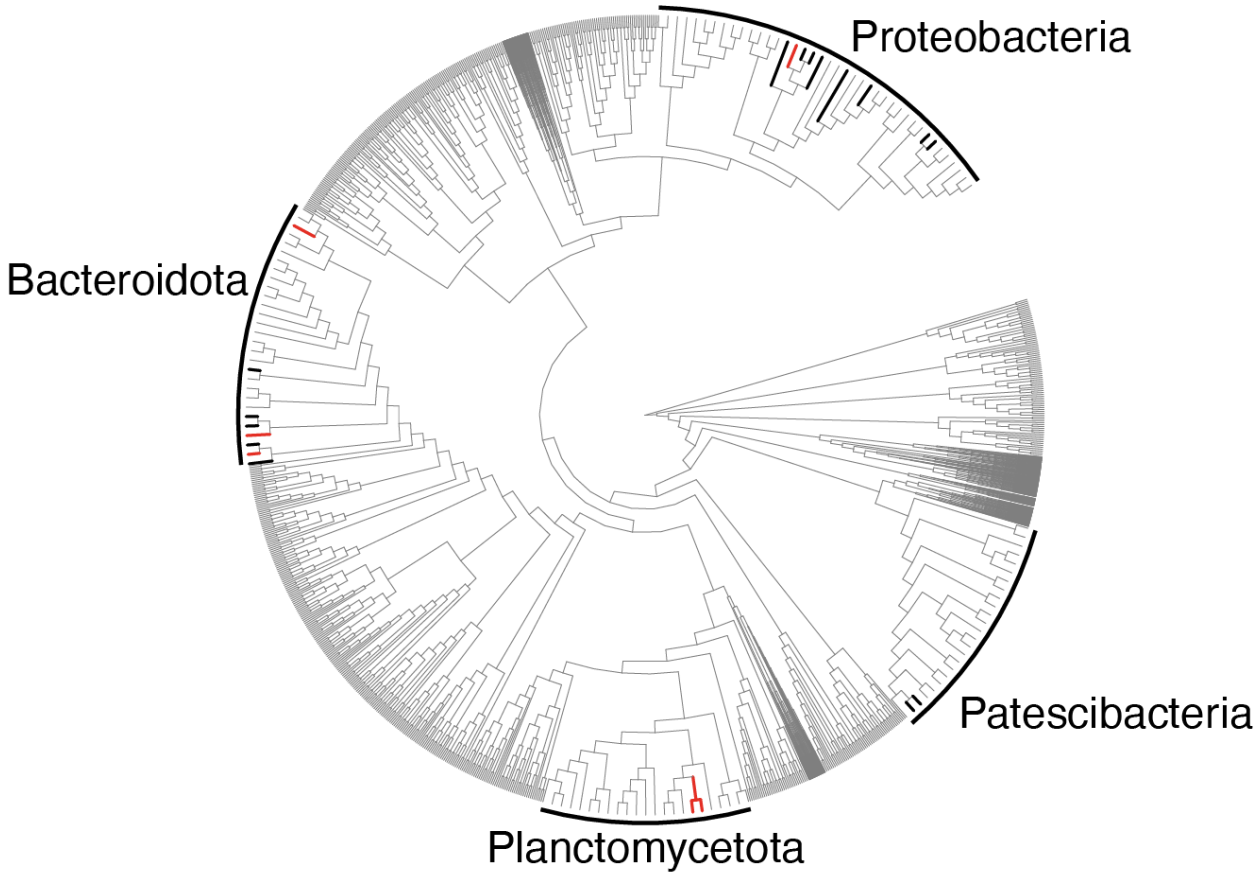

Figure 2.4. Phylogenetic placements of 21 bacterial SSU rDNA sequences (thick branches) from the metagenome assembly of *Psammoneis japonica*. Red branches show the placements of four bacterial species with completely sequenced genomes. The planctomycete genome had two distinct SSU sequences.

A second strategy involved estimation of gene density for all contigs of the assembly, since contaminant contigs of prokaryotic origin would be expected to exhibit a higher gene density than would endogenous eukaryotic contigs (Mira *et al.*, 2001). For this effort, contigs were translated in all six reading frames using EMBOSS (Rice *et al.*, 2000) with a minimum size cutoff of 30 amino acids, and the resulting polypeptide sequences were searched for protein homology against the UniProt Swiss-Prot database using an E-value cutoff of 10^{-3}. Translated polypeptides with qualifying homology to the Swiss-Prot database were considered proxies for

genes. Gene density was subsequently calculated and normalized for each contig as the number of unique translated polypeptides with one or more hits to the Swiss-Prot database per 1 Mbp of contig length.

Contig origin (i.e., endogenous or exogenous) was ultimately judged based on a combination of the above information, including GC content, size, coverage depth of mapped filtered Illumina DNA and RNA reads, BLAST homology, and gene density. For example, when contigs were grouped by BLAST result and coverage depth of mapped error-corrected and quality, adapter, and contaminant-filtered Illumina DNA reads, strong and consistent relationships were found for gene density between low and high coverage contigs. Low-coverage contigs had average gene densities of 639 – 686 polypeptides per Mbp, respectively, while high-coverage contigs ranged from 167 – 261 genes per Mbp.

From the 710 total contigs in the assembly, 490 contigs, totaling 88.8 Mbp, had both high Illumina DNA and RNA coverage. The majority of these contigs had no BLAST homology (301 contigs, 25.6 Mbp), while smaller numbers had a strongest BLAST hit to either Bacillariophyta specifically (72 contigs, 25.7 Mbp) or the broader Eukaryota (110 contigs, 35.8 Mbp). Seven of these contigs (1.7 Mbp) had a strongest BLAST hit to bacterial taxa. Based on Illumina DNA and RNA coverage, gene densities and GC contents, these contigs were considered endogenous in origin. Conversely, a total of 107 contigs totaling 2.3 Mbp in length were identified with high Illumina DNA coverage depth but low RNA coverage depth. Only one contig of this group had any BLAST-identified homology (12.7 Kbp in length, homology to Bacillariophyta); however, all gene density values for these contigs suggested eukaryotic origin. These contigs were considered of endogenous origin and kept in the assembly. 112 contigs totaling 5.7 Mbp in length were identified with both low Illumina DNA and RNA coverage. All of these contigs either had no BLAST-based homology or had the strongest homology to Bacteria. In addition, all gene density values suggested bacterial origin. As a result, these contigs were flagged as likely exogenous in origin. A single contig of 82 Kbp had low Illumina DNA coverage,

but high RNA coverage (4.2× depth). Gene density for this contig was intermediate between putative endogenous and exogenous values, while the strongest BLAST homology was to Bacteria (bitscore = 3219) and GC content was relatively high (0.57%). Based on blast homology, low Illumina DNA coverage and high GC content, this contig was also flagged as exogenous. The final *P. japonica* genome assembly was completed by removing the 113 putative contaminant contigs from the penultimate assembly. This resulted in a total of 597 contigs with N50 of 390 Kbp and totaling 91.2 Mbp in total assembled length. Genome assembly for the four bacterial scaffolds (labeled 0F, 1F, 2F, 3F) followed the same general methods as for the nuclear genome assembly. Sequences were checked for completeness and contamination levels using CheckM (Parks *et al.*, 2015).

2.3.5 Error correction of final assembly

The final contaminant-filtered Psammoneis japonica assembly was first corrected for potential assembly errors using two error-correcting softwares in sequence, Quiver ver. 2.1.0 (Chin *et al.*, 2013) and Pilon ver. 1.2.1 (Walker *et al.*, 2014). These software packages are complementary, as Quiver utilizes re-mapping of PacBio reads for error correction, while Pilon utilizes short-read (Illumina) sequence data for error correction. PacBio bax.h5 files were first converted to BAM formatting using bax2bam ver. 0.0.8 (https://github.com/PacificBiosciences/bax2bam). PBalign ver. 0.3.0 (https://github.com/PacificBiosciences/pbalign) was used to align all original PacBio reads to the final contaminant-filtered Psammoneis japonica alignment, and Quiver ver. 2.1.0 (Chin *et al.*, 2013) was used to create a new consensus sequence for the full Psammoneis japonica genome assembly. Final contaminant-filtered Illumina sequence reads were then aligned to the Quiver consensus sequence and a sorted BAM file representing these mappings was produced using BWA-MEM ver. 0.7.15 (Li, 2013) and SAMtools ver. 1.2 (Li *et al.*, 2009) under default settings. A final, error-corrected consensus sequence was produced from this BAM file using Pilon ver.

1.21 (Walker *et al.*, 2014). Quiver reported 312,600 error corrections, including 177,281 insertions, 50,926 deletions and 84,393 substitutions.

The Quiver- and Pilon-corrected consensus sequence was then checked for remaining contig mis-assemblies using REAPR ver. 1.0.18 (Hunt *et al.*, 2013) with default settings. Identified mis-assembly gaps were masked and subsequently filled where possible using the final contaminant-filtered Illumina sequence reads in GapCloser ver. 1.12 (Luo *et al.*, 2012) with default settings, except the average insert size was set to 310bp and the minimum aligned length to contigs for a reliable read location was set to 50bp. After REAPR processing and prior to GapCloser, the assembly included 1.7 Mbp of gap positions at 2,934 loci throughout the assembly (579.6 masked position per gap locus). After GapCloser, the assembly included 473,998 gap positions at 1096 loci throughout the assembly (average 432.5 masked positions per final gap locus). This resulted in a final assembly length of 91.4 Mbp with an N50 of 377.7 Kbp.

This final assembly was further validated through mapping and summary of the final contaminant-filtered Illumina sequence reads and trimmed and filtered Illumina RNA-seq reads using BWA-MEM ver. 0.7.15 (Li, 2013) and SAMtools ver. 1.2 (Li *et al.*, 2009) under default settings. For these read sets, mapped and properly paired read pairs represented 99.8% (Illumina DNA sequence reads) and 98.9% (Illumina RNA-seq sequence reads) of all read pairs with both mates mapped to the final and repeat-masked assembly (not considering reads with mate pairs mapped to a different chromosomes), indicating a highly accurate assembly.

2.3.6 Gene model annotation of final assembly

The final assembly was annotated through an iterative process using Maker2 ver. 2.31.8 (Holt & Yandell, 2011). Maker was first run under default settings with the following exceptions: 1) EST evidence was supplied as the assembled nuclear transcriptome for *P. japonica,* 2) protein homology was supplied as translated protein-coding sequences from Phaeodactylum tricornutum (NCBI BioProject PRJNA13152), *Cyclotella nana* (NCBI BioProjects PRJNA34119,

PRJNA191), and *Fragilariopsis cylindrus* (JGI Project ID:16035, filtered protein models 1), 3) repeat_protein reference was supplied as the 'te_proteins.fasta' file supplied with the maker distribution, 4) augustus_species was supplied as *Cyclotella nana* translated protein-coding sequences, 5) est2genome was set at 1, 6) proteine2gnenome was set at 1, 7) max_dna_len was set at 240000, 8) min_contig length was set at 200, 9) min_protein length was set at 15, 10) single_exon value was set at 1, 11) single_length was set at 200. After the first run, GFF formatted output was concatenated and re-formatted using the Maker function 'maker2zff,' and an HMM profile was constructed using SNAP (Korf, 2004). The annotation and HMM-building process was repeated three more times, with identical settings as above but with the previous round's HMM supplied for snaphmm and est2genome set to 0 for Gene Prediction settings. Annotated gene counts between the third and fourth rounds of annotation were essentially identical (15,188 versus 15,170 annotated genes, respectively), thus annotations resulting from the fourth round of maker were considered final. Metabolic pathways for *P. japonica* and other assessed diatoms were searched against the KAAS-KEGG (Moriya *et al.*, 2007) database with GHOSTX (Suzuki *et al.*, 2014). The *P. japonica* genome was analyzed for interspersed repeats and low complexity DNA sequences via RepeatMasker and the integrated genome viewer (IGV) (Tarailo-Graovac & Chen, 2009; Robinson *et al.*, 2011). Transposable element density within 10 Kbp up- and downstream of genic and intergenic elements calculated using the bedtools functions slop and intersect functions.

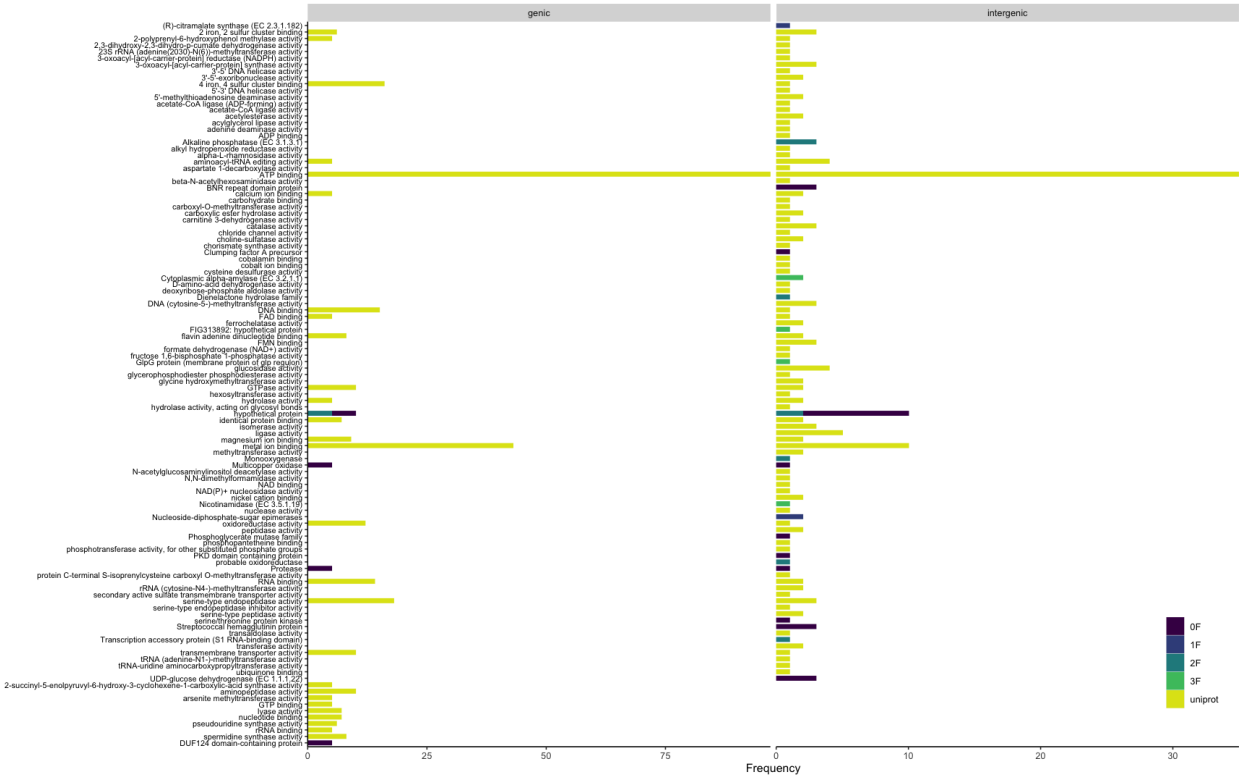

Figure 2.5. Gene annotations of genic and intergenic HGT hits to the *Psammoneis*–associated genomes. Genic hits were subset to only include metabolic functions with ≥ 5 occurrences. HGT hits that had no annotation are excluded from this figure. Colors indicate the source for the sseqid hit. Intergenic hits were subset to only include annotations with > 0 occurrences.

2.3.7 Orthologous clustering and testing for HGT elements in coding loci

Orthologous clustering was performed using OrthoFinder (ver. 2.5.2) with default settings (Emms and Kelly 2015). A total of 18 input proteomes were used from the following diatom genomes: *P. japonica, Pseudo-nitzschia multiseries, Pseudo-nitzschia multistriata, Fragilariopsis cylindrus, Seminavis robusta, Fistulifera solaris, Phaeodactylum tricornutum, Nitzschia sp.* Nitz4*, Cyclotella cryptica, Cyclotella nana*. The heterokonts *Nannochloropsis gaditana* and *Ectocarpus siliculosus* were included as outgroups. To generate the phylogeny used in Figure 2.2, I used the species tree inferred by OrthoFinder, using the STAG algorithm (Emms & Kelly, 2018). I then dated the species tree using TreePL (Smith & O'Meara, 2012). The min/max calibration times for the MRCA node of the ROOT, DIATOMS, and NITZSCHIA were obtained from (Nakov *et al.*, 2018). Gene ontology (GO) enrichment was performed using (Alexa & Rahnenfuhrer, 2022) on

the subset of orthogroups unique to pennate diatoms (*P. japonica, P. multiseries, P. multistriata, F. cylindrus, S. robusta, F. solaris, P. tricornutum, N. sp.* Nitz4) and the subset of orthogroups unique to raphid pennate diatoms (*P. multiseries, P. multistriata, F. cylindrus, S. robusta, F. solaris, P. tricornutum, N. sp.* Nitz4), these were enriched against the entirety of GO terms associated with all of our orthofinder orthogroups. Orthofinder output used in the HGT index was filtered to include only those orthogroups with genes from *P. japonica*. The HGT index was calculated as described above, using orthogroups instead of ORFs. Orthogroups were screened for HGT candidates by calculating the *h*-value for each gene individually as well as the mean *h*-value for each orthogroup *(h_{orth})* and applying the HGT class A, B, and C designations (Crisp *et al.*, 2015). To designate HGT classes, a hit must be classified as coming from the specified donor group (prokaryotes in this case). The three HGT class designations were calculated as follows: Class C candidates having $h \geq 30$ and max donor bitscore ≥ 100, Class B candidates having $h_{orth} \geq 30$, and Class A candidates having $h_{orth} \geq 30$ and a maximum recipient bitscore < 100 for the orthogroup.

For final validation of HGT candidate genes, all Class A candidates were subjected to a final search against the larger NCBI non-redundant (nr) and taxonomy (taxdb) databases (Downloaded: 2022-10-09) using NCBI BlastP (Altschul *et al.*, 1990). I performed this additional search to confirm the origin of these sequences using a larger database and to ensure the use of -max_target_seqs in this Diamond search did not lead to false positive hits to our donor database (Shah *et al.*, 2019). I performed this BLAST analysis using the following command: blastp -db nr -num_threads 64 -max_target_seqs 10 -outfmt '6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore sacc staxids sscinames' -query <query file> -out <output file>. I then downloaded the prot.accession2taxid files and categories.dmp file (Downloaded from NCBI: 2022-10-09) and converted the prot.accession2taxid files into a parquet file for input to the R package ‘arrow’ (Richardson *et al.*, 2022). The BLASTP results, prot.accession2taxid, and categories.dmp files were joined so that I

could determine the taxonomy of each BLASTP hit to our query sequences. If the top hit, based on bitscore, was also to a bacterial or archaeal sequence, the Class A candidate was retained. Sequences with top hits to eukaryotes were discarded, regardless of whether a majority of hits included bacteria as well.

Validated Class A HGT sequences were submitted to the PHYRE2 protein modeling web server for functional assignment (Kelley *et al.*, 2015). These proteins were also submitted to additional analysis to predict subcellular localization using DeepLOC ver. 2.0 (Thumuluri *et al.*, 2022), GPI-anchoring using the NetGPI (v1.1) (Gíslason *et al.*, 2021), and signal peptide category using HECTAR (v1.3), SignalP (6.0) and TargetP (2.0) (Gschloessl *et al.*, 2008; Armenteros *et al.*, 2019; Teufel *et al.*, 2022). Lastly, since diatoms have secondary red-algal derived plastids with 4 membranes, I used ASAFind v1.1.7 (Gruber *et al.*, 2015) to evaluate whether any of the HGT candidate proteins are targeted to the plastid. ASAFind uses the input from SignalP v3.0 (Bendtsen *et al.*, 2004). The ASAFind.py script had to be modified to work with Python v3.9 due to several deprecated functions.

2.3.8 Bacterial HGT in P. japonica intergenic regions

Pseudogenes are disabled copies of genes that do not produce full length protein chains due to the introduction of premature stops, disruptive frameshift mutations, and other mutations. This is most common in duplicated genes that are not the focus of DNA repair mechanisms. It is common for HGT elements to become pseudogenized over time because they are often under similar evolutionary pressures as duplicated genes (Liu *et al.*, 2004; Bock, 2010; Dunning Hotopp, 2011). As such, in addition to annotated genes, I searched the intergenic regions of the *P. japonica* genome for evidence of HGT with the aim to similarly identify pseudogenized intergenic elements with evidence of origin from one or more of the *Psammoneis*-associated bacteria. To do this, I extracted the intergenic regions of the *P. japonica* genome using bedtools (ver. 2.25) (Quinlan & Hall, 2010). I then identified ca. 2.74 million open reading frames (ORFs) of at least 90 base pairs in length from the intergenic regions of the *P. japonica*

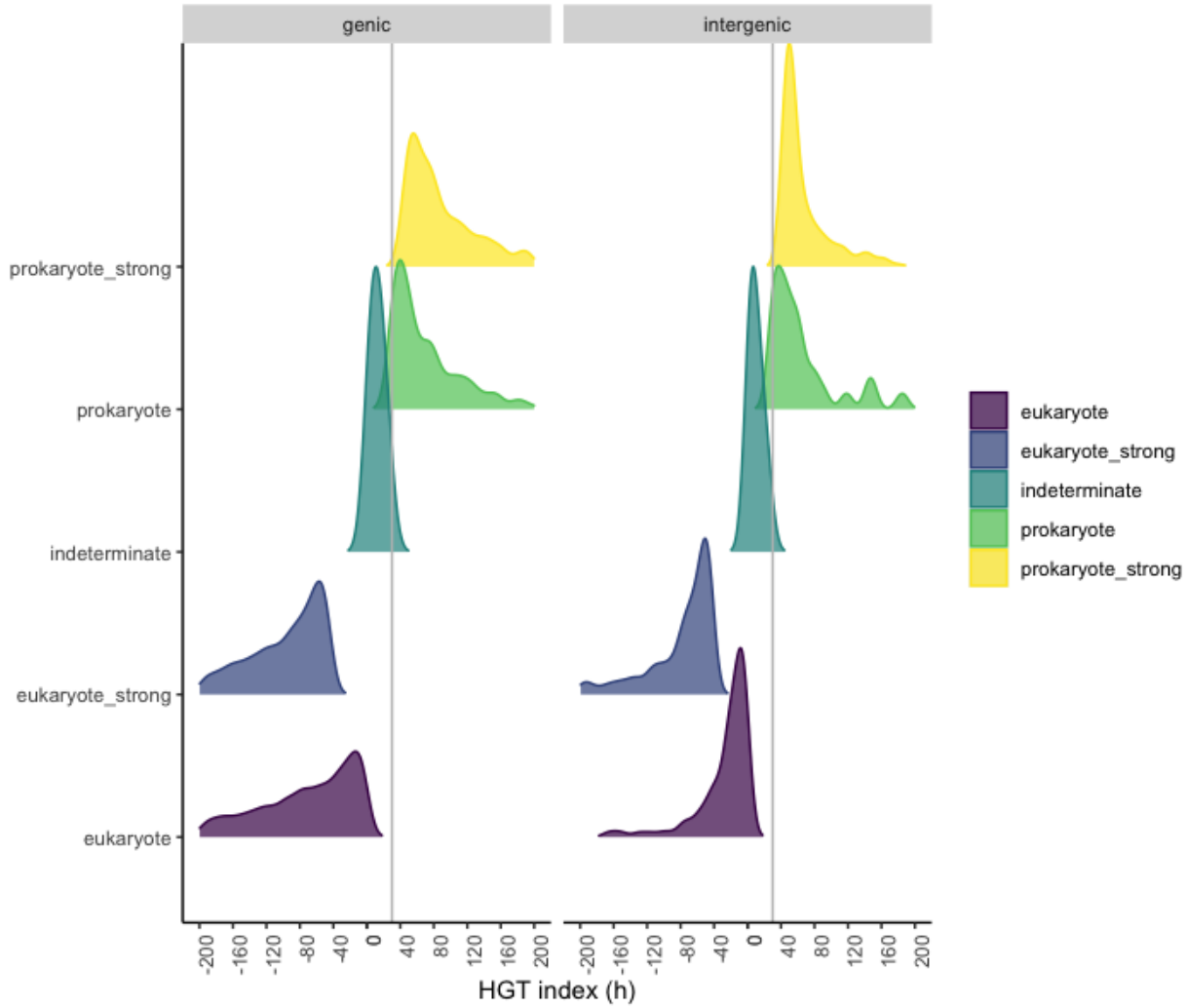

Figure 2.6. Density plots of *h* values for genic and intergenic datasets. Colors indicate the respective domain classification for each Diamond BlastP hit. Vertical gray lines indicate the cutoff *h* value for determining domain (*h* = 30). For clarity, plots have been subset to only include values of -200 ≤ *h* ≤ 200.

genome using the program orfipy (ver. 0.0.4) (Singh & Wurtele, 2021) with the following command line code: orfipy [input fasta file] --pep [output peptide fasta file name] --dna [output nucleotide fasta file name] --strand b --between-stops --include-stop --min 30. Intergenic regions of the genome were converted into ORFs sequences to identify sequences within intergenic regions with the potential to be translated into proteins, and thus the potential to have been horizontally transferred. For a breakdown of ORF lengths by HGT classification, see Figure 2.8. I next conducted Diamond (ver. 2.0.1; Buchfink *et al.*, 2021) BLASTP searches on

the translated versions of these intergenic ORFs against prokaryotic and eukaryotic Diamond databases. The prokaryotic Diamond database combined peptide sequences from the bacterial and archaeal representatives of the UniprotKB-SwissProt (UniProt Consortium, 2021) database (downloaded March 2022) and translated peptide sequences of annotated genes in the bacterial genomes. The eukaryotic Diamond database consisted of peptide sequences from eukaryotic representatives of the UniprotKB-SwissProt database. After completing these BlastP searches, I calculated the HGT index (*h*) for each intergenic query sequence as described in (Boschetti *et al.*, 2012) and determined a cutoff of $h = 30$ (Fig. 10) to be sufficient for categorization of HGT domain. The HGT index (*h*) is calculated by performing BLAST searches on each query sequence against donor (prokaryote) and recipient (eukaryote) databases, then subtracting the bitscore of the top eukaryote hit from the bitscore of the best prokaryote hit. Negative *h* values are indicative of the query sequence being eukaryotic in origin, while positive bitscores are indicative of prokaryotic origin. Following Boschetti et al. (2012), genes were categorized as prokaryotic, eukaryotic, or indeterminate in origin. As in Boschetti *et al.* (2012), eukaryote sequences were defined as having $h \leq 0$, indeterminate sequences having *h*-values between 0 and 30, and prokaryotic sequences having $h \geq 30$, and updated with two additional categories (e.g., prokaryote_strong and eukaryote_strong). For query sequences that only had hits to one database or the other, I applied a bitscore of zero to the missing domain and assigned that query to the eukaryotic_strong or prokaryotic_strong category, depending on which domain the top BLAST hit was. Query sequences that were identified as being of prokaryotic origin were submitted to the PHYRE2 protein modeling web server for functional assignment (Kelley *et al.*, 2015). Finally, three ORFs from the *P. japonica* genome (contig 89F) were submitted to additional analysis to predict subcellular localization using DeepLOC ver. 2.0 (Thumuluri *et al.*, 2022).

For final validation of HGT candidate ORFs, all prokaryotic HGT candidates were subjected to a final search against the larger NCBI non-redundant (nr) and taxonomy (taxdb)

databases (Downloaded: 2022-10-09) using NCBI BlastP (Altschul *et al.*, 1990). I performed this additional search to confirm the origin of these sequences using a larger database and to ensure the use of -max_target_seqs in this Diamond search did not lead to false positive hits to our donor database (Shah *et al.*, 2019). I performed this BLAST analysis using the following command: blastp -db nr -num_threads 64 -max_target_seqs 10 -outfmt '6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore sacc staxids sscinames' -query <query file> -out <output file>. I then downloaded the prot.accession2taxid files and categories.dmp file (Downloaded from NCBI: 2022-10-09) and converted the prot.accession2taxid files into a parquet file for input to the R package ‘arrow’ (Richardson *et al.*, 2022). The BLASTP results, prot.accession2taxid, and categories.dmp files were joined so that I could determine the taxonomy of each BLASTP hit to our query sequences. If the top hit, based on bitscore, was also to a bacterial or archaeal sequence, the HGT candidate ORF was retained. Sequences with top hits to eukaryotes were discarded from final analysis, regardless of whether a majority of hits included bacteria as well.

I used the metabolic-g function of METABOLIC (ver. 4.0) (Zhou *et al.*, 2022) to profile the genomes of four *Psammoneis*-associated bacteria. Using the Metabolic-g function of the program, I classified the metabolic capabilities of four bacterial genomes and generate biogeochemical cycling diagrams highlighting the portions of these cycles that each bacterial genome can perform.

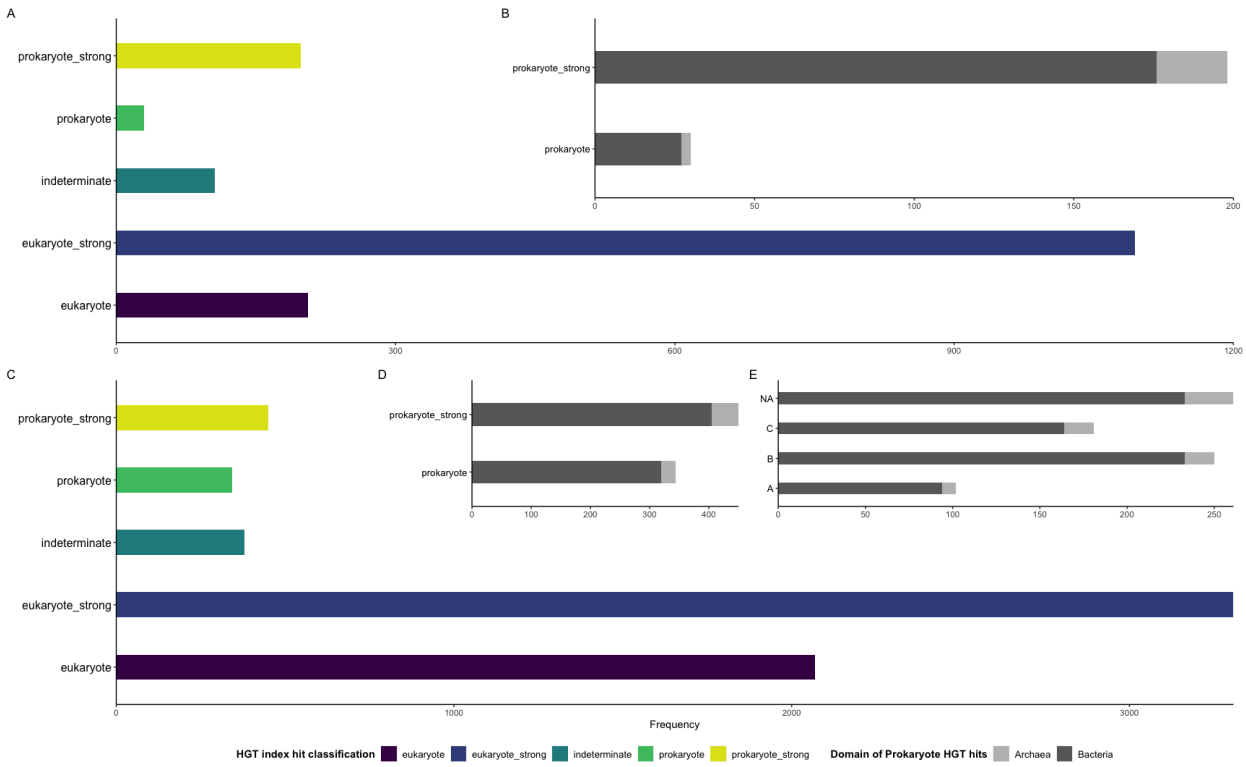

Figure 2.7. HGT index classifications for intergenic and genic analyses. Intergenic analyses: A) Frequencies of prokaryotic, indeterminate, and eukaryotic HGT classifications (Boschetti *et al.*, 2012) and B) domain level classifications for prokaryotic HGT hits. Genic analyses: C) Frequencies of prokaryotic, indeterminate, and eukaryotic HGT classifications (Boschetti *et al.*, 2012), D) domain level classifications for prokaryotic HGT hits, and E) h_{orth} classifications (Crisp *et al.*, 2015).

2.3.9 Phylogenetic placement of bacterial SSU sequences and genomes

Bacterial small-subunit (16S, SSU) ribosomal DNA sequences were extracted from the metagenome assembly with METAXA2 (Bengtsson-Palme *et al.*, 2015) and aligned using SSU-ALIGN (ver 0.1) (Nawrocki, 2009) against the SSU portion of release 202 of the Genome Taxonomy Database (GTDB; release 202) (Parks *et al.*, 2018, 2020). I chose GTDB because it is frequently updated and incorporates uncultured and undescribed bacterial taxa. The multiple sequence alignment, together with the GTDB reference phylogeny and taxonomy file (Release 202), were used as input for phylogenetic placement with EPA-NG v0.3.5, using the SYM+I+G4 nucleotide substitution model (Barbera *et al.*, 2018). Rather than relying on arbitrary percent similarity cutoffs to identify a sequence, EPA-NG takes into account the phylogenetic placement

of each sequence. Taxonomic assignments were extracted from the EPA-NG jplace file with Gappa v0.5.0 (Czech & Stamatakis, 2018), using the functions ‘examine assign’ and ‘examine graft.’ The phylogeny was rendered with the R package ggtree (Yu *et al.*, 01/2017) and polished using Adobe Illustrator.

The four bacterial genomes were taxonomically placed using the GTDB-tk toolkit v1.4.1 (reference data ver. r95; Chaumeil *et al.*, 2019). I chose GTDB-tk because it incorporates a large amount of described and undescribed bacterial diversity and makes use of a genomic reference multiple sequence alignment to make phylogenetic placements and taxonomic identifications.

2.3.10 Gene family evolution

I analyzed gene family evolution using the software Computational Analysis of gene Family Evolution (CAFE) (Han *et al.*, 2013), using the Orthofinder orthogroups from the following taxa as input: *E. siliculosus, N. gaditana, T. pseudonana, C. cryptica, S. robusta, F. solaris, P. tricornutum, F. cylindrus, P. multistriata, P. multiseries, N.* sp. Nitz4, and *P. japonica.* The resulting orthogroups that were found to be rapidly evolving were submitted to the KofamKOALA (ver. 2021-04-01; KEGG release 98.0) web tool (Aramaki *et al.*, 2020) for annotation via KOfam, a customize HMM database of KEGG Orthologs, with an e-value parameter set to 0.01. Finally, InterProScan (Jones *et al.*, 2014) was used to obtain Gene Ontology (GO) terms for each of the rapidly evolving orthogroups.

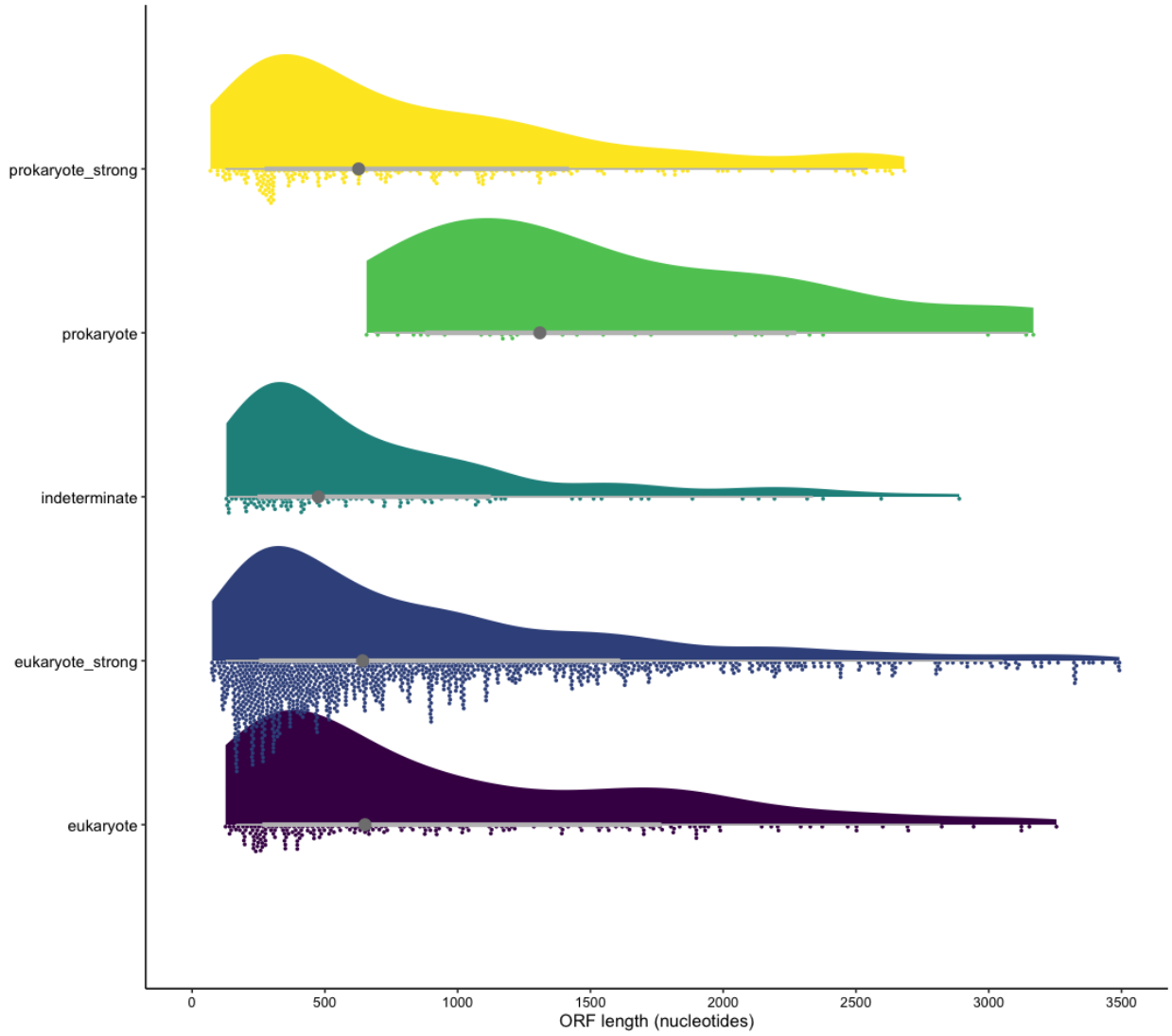

Figure 2.8. Density plot of length ranges in nucleotides for intergenic ORFs by HGT classification.

2.4 Results

2.4.1 Nuclear genome characteristics of *P. japonica*

The nuclear genome assembly of *P. japonica* totaled 91.4 Mbp in length (scaffold N50 = 378 Kbp), which is roughly three times larger than the smallest sequenced diatom genomes (Fig. 2). As can be seen in Figure 2.2, genome composition of *P. japonica* is similar to that of other diatoms. The 15,170 predicted protein-coding genes in *P. japonica* comprise 27% of the total genome, repetitive elements account for 33% of the genome, and other non-coding

elements comprise the remaining 40% of the genome (Fig. 2). Of the repetitive elements of the genome that could be classified, 12% were long terminal repeats (LTR), 2.3% were long interspersed nuclear elements (LINE), 0.5% were simple repeats, 2.65 were DNA. The remaining repeat content of the genome consisted of 1% other and 15% unclassified repeats. The composition of repetitive regions in the *P. japonica* genome is also similar to the other diatom genomes depicted in figure 2.2. Phylogenetically independent contrasts were performed to determine whether there was a relationship between genome size, repeat content, or gene content. Phylogenetically independent contrasts revealed a positive relationship between genome size and repeat content, and no relationship between gene content and genome size among diatoms (Fig. 2). The genome contains six of the seven known diatom-specific Ty1/Copia-like elements (Maumus *et al.*, 2009), but these accounted for a small portion of the genome overall (ca. 24 Kbp).

2.4.2 Gene family evolution of *P. japonica*

Orthofinder recovered a total of 21,959 orthogroups, with 8729 containing genes from *P. japonica* and 396 orthogroups containing genes only from *P. japonica.* Of the orthogroups used in our cafe analysis, only a single orthogroup (OG10091) contained genes from all five raphid diatoms to the exclusion of all other groups sampled , 3,457 orthogroups contained genes from any raphid diatom(s), 3,226 contained genes from all eight pennate diatom proteomes, and 6,023 contained genes from any pennate diatom(s). The small number of orthogroups exclusive to all rapid diatoms may indicate that most of the genes that make rapid pennate diatoms unique comes from the modification of existing gene families in araphid pennate diatoms. The CAFE analysis of gene family evolution in *P. japonica* found five gene families on the *P. japonica* branch that are rapidly evolving, specifically: OG38, OG42, OG66, OG102, & OG509. Of the orthogroups that were rapidly evolving, only OG38 was found to be rapidly contracting, all others were found to be rapidly expanding. Orthogroup 38 was identified as H32_USTMA Histone H3.2, a core component of the nucleosome. Orthogroup 42 was identified as

PEPR2_ARATH Leucine-rich repeat receptor-like protein kinase PEPR2, which acts as a receptor for PEP defense proteins and are kinases that act as enzymes on other proteins. Orthogroup 66 was identified as PNS1_ASHGO Protein PNS1, which is likely involved in transport through the plasma membrane. Orthogroup 102 was identified as NPC1_PIG NPC intracellular cholesterol transporter 1, which is an intracellular cholesterol transporter whose function in diatoms is currently unknown. Finally, orthogroup 509 was identified as DTX3L_MOUSE E3 ubiquitin-protein ligase DTX3L. None of these orthogroups were found to be HGT candidates by our analyses. Likewise, none of these gene families have functions that are obviously related to diatom-bacteria interactions and thus the expansion and contraction of these gene families does not provide any deeper insights at the time of writing. . Additional rapidly evolving gene families were identified in deeper nodes, but are not presented here. For the full results of these analyses, see Zenodo repository.

2.4.3 Community composition of the *P. japonica* phycosphere

I recovered, and phylogenetically placed, 16 bacterial taxa based on their 16S *SSU*-rRNA sequences and four complete bacterial genomes from the *P. japonica* metagenome (Figure 2.4). Bacterial taxa from identified via 16S sequences displayed some overlap with the lineages of the four bacterial genomes. Those that were from similar lineages were placed in the Family Rhodovibrionaceae, and other clades of the Alphaproteobacteria, as well as in the genus *Ekhidna* of the phylum Bacteroidota. The remaining 16S sequences were placed in the Family Alteromonadaceae of the Gammaproteobacteria, and an undescribed order of the Patescibacteria. See the associated Zenodo repository for phylogenetic placements and associated support values of 16S sequences.

Table 2.1. Summary of four bacterial genomes from combined Illumina and PacBio sequencing data. Determination of most closely related species is phylogenetic placement of 16s rRNA sequence alignment using EPA-NG. LWR (Likelihood Weight Ratio) is a measure of placement certainty used by EPA-NG, ranging from 0-100 with higher values indicating greater certainty. Genome completeness estimated using CheckM (Parks, Imelfort et al. 2015).

Scaffold ID	Lowest predicted taxonomy (phylum-level taxonomy)	Likelihood Weight ratio (LWR)	Genome size (Mb)	GC content (%)	Protein-coding gene count	% estimated complete
0F	*Phycisphaerae* (Planctomycetota)	88.7	5.5	57.3	4522	98.9
1F	*Rhodovibrionaceae* (Proteobacteria)	99.9	4.6	69.8	4361	99.6
2F	*Ekhidna* (Bacteroidota)	91.9	4.2	40.1	3869	98.8
3F	*Balneola* (Balneolaeota)	99.9	3.5	40.5	3092	99.7

Phylogenetic placement of the four bacterial genomes was conducted using both 16S sequences from each genome, as well as the genomes themselves, via EPA-NG and GTDB-tk, respectively. Phylogenetic placement of 16S sequences from the bacterial genomes found four distinct lineages: sister to the marine Phycisphaerae clade of Planctomycete bacteria (0F), within the Rhodovibrionaceae clade of the Alphaproteobacteria (1F), within the genus Ekhidna (Family Cyclobacteriaceae) of the Bacteroidota (2F), and within the Family Balneolaceae of the Alphaproteobacteria (3F). Bacterial genomes were also phylogenetically placed using the GTDB-tk toolkit (Chaumeil *et al.*, 2019). The closest hits for each genome are as follows: 000000F did not have any other close hits to the reference dataset, 000001F was related to the bacterium *HHTR118 sp003572205* of the Rhodovibrionaceae family (Average Nucleotide Identity = 80.34%), 000002F was related to *Ekhidna lutea* (Average Nucleotide Identity = 78.15), and 000003F was related to *Balneola vulgaris* (Average Nucleotide Identity = 77.52). Since the next closest hits from GTDB-tk and the taxonomic placement of EPA-NG show a high degree of overlap, I am confident that these taxonomic placements are accurate. Additionally, the four bacterial genomes have already been incorporated and phylogenetically placed by the Genome Taxonomy Database (GTDB) reference, which shows phylogenetic placements nearly identical to their 16S placements. The overlap in taxonomy between the placements made by GTDB for

their reference dataset and our results provides another line of evidence in favor of these placements. Each of the genomes contained genes for auxin synthesis, B vitamin synthesis, iron, and nitrogen metabolism. Identified secondary metabolite clusters among these genomes included polyketide synthesis (Types 1, 3, Other), Bacteriocin, Ectoine, non-ribosomal peptide synthetase and Terpene.

Table 2.2. Functional assignments and genome annotations for intergenic HGT candidate pseduogenic sequences. Functional assignments completed using Phyre2. Full details for 40 ORF candidate loci are found in the Zenodo repository associated with this manuscript.

***Psammoneis* ORF (length in amino acids)**	**Contig coordinates**	**ORF length**	**Top BLASTP hit to *Psammoneis*-associated bacteria**	**bitscore**	**subcellular localization prediction**	**Phyre2 Top functional assignment**
89F.ORF.132	194733-201554	390	0F.4444	363.2	Cytoplasm	haloperoxidase
89F.ORF.31	194733-201554	214	0F.4444	97.1	Extracellular	haloperoxidase
89F.ORF.38	203504-212023	93	0F.4444	87.8	Extracellular	oxidoreductase

2.4.4 Metabolism and biogeochemistry analysis of bacterial genomes

Analysis of the four bacterial genomes found genes related to 16 of the 23 major metabolic categories examined in the METABOLIC workflow (Fig. 9A-C). The majority of genes related to these metabolic categories were only present in 1 genome, the 1F bacterial genome. The only metabolic categories found in ≥3 genomes were arsenic cycling, complex carbon degradation, fatty acid degradation, and oxidative phosphorylation (Fig. 9B). None of the examined genomes possessed genes that contribute to all aspects of the biogeochemical cycles that were examined. A total of 40 unique metabolic functions were found within the 16 metabolic categories (Fig. 10D-F). None of these functions were present more than once in a genome.

2.4.5 Bacterial HGT in *P. japonica* intergenic regions

From the HGT BLASTP search, a total of 1634 ORFs hit either the prokaryotic or eukaryotic Diamond databases. Of these ORFs, 1300 had *h*-values indicative of eukaryotic origin (1094 eukaryote_strong and 206 eukaryote), 228 were indicated as prokaryotic in origin (198 prokaryote_strong and 30 prokaryote), and 106 were indicated as indeterminate in origin. A total of 203 prokaryotic hits were from the domain Bacteria and 25 from domain Archaea. Of those ORFs identified as prokaryotic in origin, 40 were most similar to proteins from the 0F, 1F, 2F, and 3F bacterial genome assemblies (23, 3, 9, and 5, respectively), the remaining 188 were hits to prokaryotic representatives of the UniprotKB-SwissProt database (203 bacterial and 25 archaeal hits). Of these 228 ORFs, only 15 were incomplete, 3'-partial ORFs (lacking a stop codon) with the remainder being complete (start and stop codon) . Density plots of *h*-values for intergenic data can be found in figure 2.6. The final NCBI BLASTP validated 113 HGT candidates as being prokaryotic in origin. Twenty of the forty hits to the 4 bacterial genomes and 93 of the hits to prokaryotic representatives of the UniprotKB-SwissProt database were maintained. These sequences had percent similarities of 24–82% (mean = 49.7%) to their top hits and ranged in length from 25–1302 bp (mean = 268.8 bp). I also examined the molecular function annotation of prokaryotic hits (Fig. 5). Average TE density around all intergenic ORFs was found to be 1.7 TEs per 10 Kbp up- and downstream. Transposable element density around the 40 ORFs most similar to cohabitating bacterial genomes ranged from 0–5 TEs per 10 Kbp up- and downstream, with TE densities above 1.7 indicating that they are in regions TE-rich. The majority of intergenic prokaryotic hits to the ORFs were to UniProt proteins with no protein annotation (60), followed by ATP binding (35) and metal ion binding (10, Fig. 5). The majority of ORF hits to the bacterial genomes were annotated as hypothetical proteins (10, Fig. 5), this was also the only category which hit to multiple bacterial genomes.

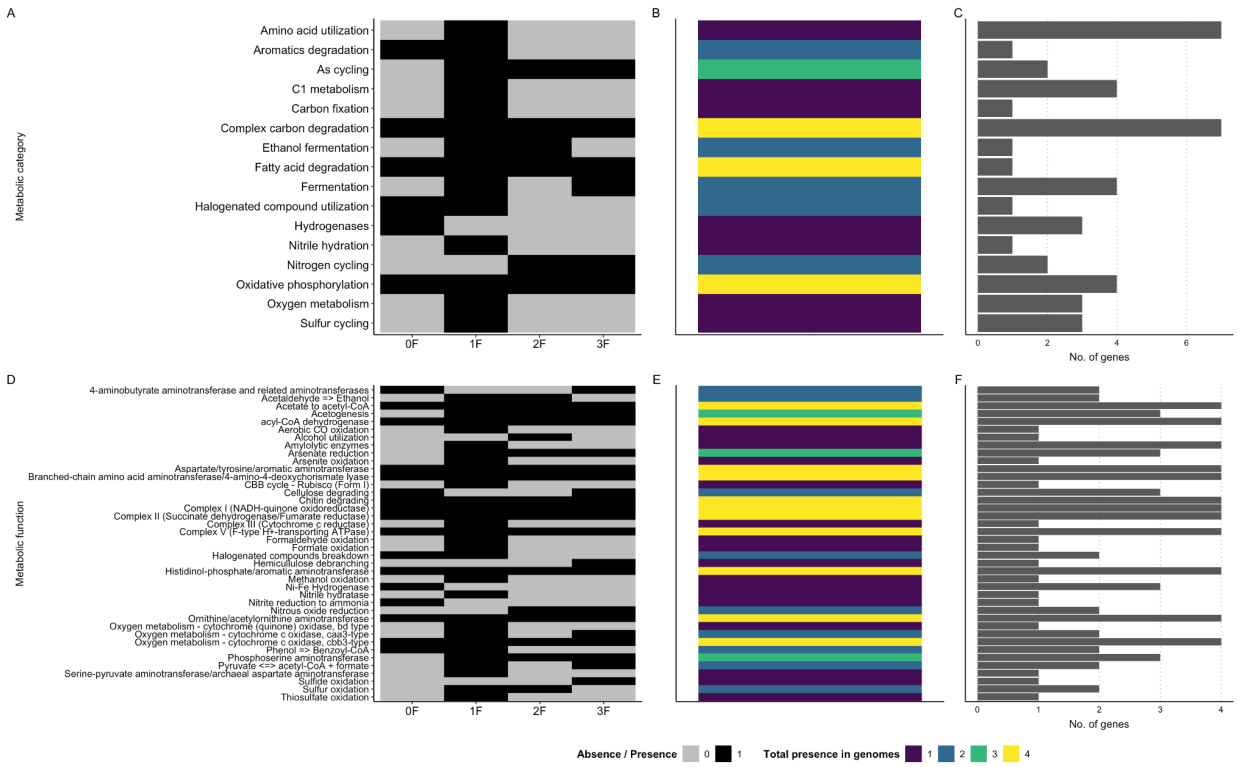

Figure 2.9. METABOLIC category (A-C) and function (D-F) output plots for bacterial genomes. A) Absence / Presence of genes in different metabolic categories for each bacterial genome, B) total number of genomes which possess a gene in a metabolic category, and C) total number of genes in each metabolic category for all genomes. D) Absence / Presence of genes in different metabolic functions for each bacterial genome, E) total number of genomes which possess a gene for a metabolic function, and F) total number of genes in each metabolic function for all genomes.

Nine ORFs were found to be hits within the same contigs of *P. japonica* to the same annotated genes. In particular, the following hit associations between *P. japonica* ORFs and bacterial protein encoding genes were found: *P. japonica* contig 2F had two hits to 0F.1049, *P. japonica* contig 89F had three hits to 0F.4444, *P. japonica* contig 103F had three hits to 0F.1684, *P. japonica* contig 532F had three hits to 0F.3038. Of these ORFs, only the three between *P. japonica* 89F:0F.4444 displayed potential evidence of being a pseudogenized HGT element from a *Psammoneis*-associated bacteria. These three ORFs reside within a 7.8 Kbp segment near the middle of a 343 Kbp contig (Figure 2.3). Diamond BlastP results and functional assignments for these three ORFs are presented in Table 2.2. All three ORF sequences had Diamond BlastP hits to different parts of the same 0F bacterial protein (0F.4444), resulting in

alignment to 464/504 positions of this protein and with an average identity of 60%. The two larger ORFs (89F.ORF.132 and 89F.ORF.31) overlap in the final alignment, with the smaller ORF (89F.ORF.38) separated from these by ca. 6 Kbp. These ORFs are also bordered by LTRs and LINEs within 10 Kbp up- and downstream, with four repeat regions near the two larger ORFs (89F.ORF.132 and 89F.ORF.31) and five near the smaller ORF (89F.ORF.38). All three of these ORFs had singular Diamond BlastP hits to the 0F.4444 bacterial locus. The average TE density around these three loci, 4.7 TEs per 10 Kbp up- and downstream, was also significantly higher than the average for the other 37 ORFs, indicating that the 89F loci are in TE-rich regions. Taken together, these results support that these three ORF loci represent three separate coding sections of a single gene from a Planctomycete 0F bacterium or a closely related species. All three ORF loci were predicted as haloperoxidases/oxidoreductases by Phyre2 with DeepLOC predicting subcellular localization outside of the cell and within the cytoplasm (Table 2.2).

2.4.6 Orthologous clustering and testing for HGT elements in coding loci

Orthofinder recovered a total of 21,959 orthogroups, with 10,126 containing genes from *P. japonica* and 398 orthogroups containing genes only from *P. japonica.* An additional 25,023 genes could not be assigned to an orthogroup, with 1,397 unassigned genes belonging to *P. japonica.* Of the genes that could be assigned to an orthogroup, 7,780 were assigned to orthogroups containing genes only from rapid pennate diatoms, with 3,226 and 4,554 belonging to multi-taxon and single-taxon orthogroups, respectively. For pennate diatoms (both araphid and raphid proteomes), there were a total of 1,701orthogroups, with 1,305 and 396 multi-taxon and single-taxon orthogroups, respectively. Gene ontology enrichment analysis found that the pennate diatom orthogroups had no significantly enriched GO terms associated with them. Enrichment analysis of raphid diatom orthogroups found two GO terms that were significantly enriched for molecular function, GO:0005509 (8 orthogroups) and GO:0016747 (2 orthogroups). The GO term GO:0005509 is annotated for calcium ion binding and GO:0016747 is annotated

for transferase activity, transferring acyl groups other than amino-acyl groups, both of which are very general in function.

HGT analysis of the *P. japonica* orthogroup subset found a total of 935 sequences (731 unique orthogroups) with *h*-values indicative of prokaryotic origin, 5472 as eukaryotic (4113 unique orthogroups), and 416 as indeterminate (351 unique orthogroups)(FIG. 7C). A total of 866 prokaryotic hits were to the domain Bacteria and 69 from domain Archaea. Density plots of *h*-values for genic data can be found in figure 2.6**.** Of the prokaryotic hits, 499 were categorized as Class A, with 459 from the domain Bacteria and 40 from domain Archaea. The 857 *P. japonica* proteins with strong evidence for bacterial origin, based on the class A–C designations (Crisp *et al.*, 2015), are similar in number to previously reported counts for diatom genomes (Fan *et al.*, 2020; Vancaester *et al.*, 2020). Out of the classed HGT candidates, 148 had top hits to proteins from *Psammoneis*–associated bacterial genomes. Of these 148 proteins, 106 were identified as class A HGT candidates, with 15 from orthogroups that contained only *P. japonica* proteins. A final BLASTP validation of these class A HGT candidate proteins found that only 17 of these sequences maintained top bitscore hits to bacterial proteins. Average TE density for genes for *P. japonica* was found to be 1.1 TEs per 10 Kbp up- and downstream. For the 17 HGT candidate proteins, nine were found to have TE densities of ≥ 2 per 10 Kbp up- and downstream.

I also examined the molecular function annotation of prokaryotic hits (Fig. 5). Genic prokaryotic hits to *P. japonica* coding loci were to UniProt sequences with ATP binding (101), followed by no protein annotation (70) and metal ion binding (47, Fig. 5). Phyre2 recovered a wide array of classifications, including cello adhesion, signaling proteins, hydrolases, and others. Additional analyses to predict subcellular localization, GPI-anchoring, and signal peptide category found the following. DeepLOC predicted that the majority of these proteins were located in the cytoplasm or outside of the cell. NetGPI predicted that none of these proteins were GPI-anchored. Finally, consensus between results from DeepLOC, HECTAR, SignalP, and

TargetP predicted eight proteins as possessing signal peptides, with DeepLOC predicting signals for nuclear localization, nuclear export, and transmembrane domain. Finally, ASAFind predicted a single protein to go to the plastid. Please see the associated Zenodo repository for full results of Phyre2, subcellular localization predictions, and signal peptide predictions.

2.5 Discussion

I sequenced and analyzed the nuclear genome for the araphid pennate diatom *P. japonica,* adding to the genomic resources available for diatoms. With a genome size of >91Mbp, the genome of *P. japonica* is among the largest diatom genomes sequenced to date. The genome characteristics of *P. japonica* are largely on trend with other sequenced diatom nuclear genomes in terms of gene number, proportion of coding vs. non-coding regions, and repetitive content (Figure 2.2). The *P. japonica* genome also contains genes associated with a number of common metabolic pathways in diatoms, such as urea metabolism.

Although this report is not the first to characterize complete bacterial genomes associated with a unialgal diatom culture (Shibl *et al.*, 2020b), the recovery of bacterial genomes has enabled us to investigate HGT events between diatoms and their phycosphere symbionts. I characterized 20 distinct taxa in the partial metagenome of *P. japonica,* 16 from only 16S rRNA data and 4 from near-complete genomes and their 16S data (Tables 2.1). Although all of the bacteria found in association with *P. japonica* are from bacterial lineages that are known to associate with diatoms, all represent the first record of these specific taxa associating with a diatom (Grossart *et al.*, 2005; Morris *et al.*, 2006; Bennke *et al.*, 2013; Klindworth *et al.*, 2014; Shibl *et al.*, 2020b) and of the four bacterial genomes recovered, three are the first bacterial genomes of their phyla to be recovered from diatom cultures (0F, 2F, & 3F; Table 2.1). Notably, these genomes possess pathways for auxin and B vitamin synthesis both of which have been shown to be critical components of diatom-bacterial associations (Amin *et al.*, 2012, 2015).

With the large body of evidence for HGT between bacteria and eukaryotes (Bock, 2010; Syvanen, 2012; Boschetti *et al.*, 2012), including diatoms (Bowler *et al.*, 2008; Vancaester *et al.*, 2020), I wanted to investigate the genome of *P. japonica* for evidence of bacterial HGT. Additionally, any instances of HGT between *P. japonica* and members of its metagenome would be strong evidence for the long-term association between a diatom and its metagenome. I have found evidence that numerous coding loci in the *P. japonica* genome are the result of bacteria-to-diatom HGT events, based on the HGT classification system of Crisp et al. (2015). These results are consistent with previous reports on diatom genomes (Bowler *et al.*, 2008; Osuna-Cruz *et al.*, 2020; Vancaester *et al.*, 2020). While I observed many potential genic HGT candidates, only 17 loci passed the final NCBI BLASTP analysis. Fourteen of the loci that did not pass were found to have a majority of hits to bacterial sequences with a top hit to a diatom sequence, potentially indicating very ancient instances of HGT before pennate diatom diversification. Ten of these HGT candidates originate from orthogroups that are composed of proteins strictly from *P. japonica,* indicating that these sequences are more recent examples of HGT to an earlier member of the *P. japonica* lineage or that these genes were lost in all other lineages examined. All other HGT candidate proteins were from orthogroups composed of raphid diatom-only, diatom-only, or include outgroup taxa, indicating either more ancient instances of HGT or multiple instances of HGT between these lineages and similar bacterial lineages. Upon investigating the contigs in which these HGT candidates are located, I found that nine of the 17 genic HGT candidates were located in contigs with one to four transposable elements (transposons) within 10 Kbp up- or downstream of each gene. This further supports that some of these loci may be the result of HGT, as transposons are well known to facilitate HGT (Ochman *et al.*, 2000; Frost *et al.*, 2005; Casacuberta & González, 2013). Regarding the function and localization of these HGT candidate proteins, I found that the majority of these proteins are localized to extracellular secretory pathways and include sugar binding proteins to hydrolases. One protein in particular was identified as a TIM beta/alpha-barrel structural motif

that is transcribed and targeted to the cytoplasm of the cell. The TIM beta/alpha-barrel structural motif is common and occurs across the majority of major enzyme functional classes (Copley & Bork, 2000; Wierenga, 2001; Nagano *et al.*, 2002). While these results indicate that several of these HGT candidates are actively secreted from the cells of *P. japonica*, I was not able to further define a functional role for these HGT candidates.

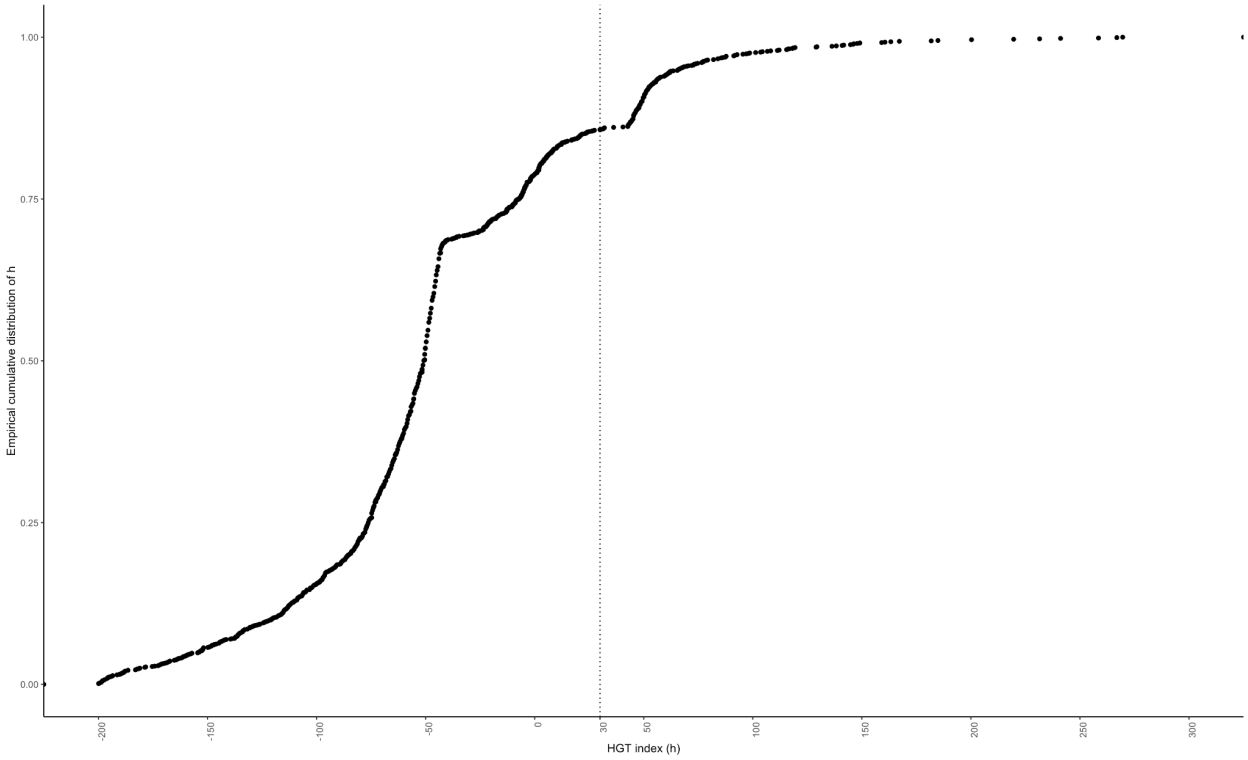

Figure 2.10. Empirical cumulative distribution function (ECDF) plot of *h* values for all blast hits. ECDF plots report the percentage of values that are below a given threshold in a collection of samples. For example, ~85% of *h* values are lower than *h* = 30. The gray, dotted line indicates the value (*h* = 30) I chose as the cutoff for making HGT index classifications.

Through exploration and testing of the intergenic portions of the *P. japonica* genome, I identified multiple ORFs with evidence of bacterial origin. I identified 110 ORFs with strong similarity to proteins from genomes 0F–3F and UniProt reference sequences, with three in particular that were of interest to us. Within a short segment of the 89F contig of the *P. japonica* nuclear genome assembly I identified three ORFs that appear to be derived from a single HGT

from the *Phycisphaerae sp.* bacterial genome, or a member of its lineage, into the *P. japonica* genome. This is supported by all three loci returned BlastP hits to different regions of the *Psammoneis*-associated bacterial protein (0F.4444; Table 2.2) resulting in alignment to 90% of this bacterial protein. Two of these loci had single BLASTP hits to protein 0F.4444, the third locus had a strongest hit to a Planctomycete bacterial protein, followed by a single, second hit, to the 0F.4444 protein. I recovered no evidence that these three ORFs are actively expressed in the *P. japonica* transcriptome, supporting that these ORF loci are currently nonfunctional. Nonetheless, our protein homology and subcellular localization analyses suggest that these three ORFs represent former oxidoreductase enzymes localized to the cytoplasm and outside of the cell. The three ORFs are surrounded by six LTR elements (Figure 2.3), which have been implicated as important mediators in the evolution of genes and their expression (Frost *et al.*, 2005; Franke *et al.*, 2017), suggesting the possibility that these ORFs may be relatively recent examples of HGT in the *P. japonica* genome. The remaining 107 HGT candidate ORFs may represent remnants of pseudogenes or otherwise unannotated genes.

Although the top matches are to bacteria, the low percent similarities for the intergenic (56-65%; Table 2.2) and genic (24-59%) HGT candidate genes suggest transfer events from relatives to representatives in current sequence databases or ancient transfer events. Molecular phylogenetic dating supports an age of 140–165 mya for pennate diatoms (Bacillariophyceae) and 25–35 mya for *P. japonica (Nakov et al., 2018)*. For HGT candidate genes in this analysis, it can then be extrapolated that some may be especially ancient, if they are from orthogroups composed of proteins from only pennate diatoms. For HGT candidates from *Psammoneis*-only orthogroups, it can also be extrapolated that these would be relatively less ancient than HGT elements from more inclusive orthogroups. For HGT candidate sequences with higher similarity to *P. japonica,* such as the intergenic ORFs (Table 2.2), it may be reasonable to assume a younger age than those with significantly lower percent identities because pseudogenes often experience increased rates of sequence divergence compared to coding sequences (Graur *et*

al., 1989; Ophir & Graur, 1997; Zhang *et al.*, 2004). Due to the high divergence in sequence identity between the corresponding diatom and bacterial sequences we are left with two explanations: time and degree of relatedness to what has been sequenced. These HGT candidate proteins could be examples of HGT between ancient members of the *P. japonica* and 0F bacteria lineages. Alternatively, since the majority of life remains unsequenced, these HGT candidate proteins could be a much more recent example of HGT between *P. japonica* and an extant bacterium that is unsequenced, leading to the inference that these genes were transferred from genome 0F instead of the more similar, but unsequenced bacterium. Thus, if the ancient HGT is correct, this would support an ancient and long-term relationship between the lineages of *P. japonica* and its bacterial partners. Although contemporary research on the longevity of diatom-bacterial associations indicate strong conservation, reproducibility, and stability over long periods (Behringer *et al.*, 2018; Mönnich *et al.*, 2020; Barreto Filho *et al.*, 2021), this would be the first demonstration of diatom-bacteria associations over evolutionary timescales.

These results reinforce the importance of diatom-bacteria interactions and support that some observations of co-occurrence can represent long-term associations over evolutionary time-scales between specific diatom and bacterial lineages. The relationship between *P. japonica* and its bacteria, particularly the strain represented by genome 0F, requires further scrutiny as an example of inter-domain co-evolution. This can be accomplished via further exploration of the metagenome of other *Psammoneis* species (Sato *et al.*, 2008), and the Plagiogrammaceae family as a whole. By investigating the functional roles of genes from bacterial-derived HGT, this might also contribute to understanding what metabolic interactions between diatoms and their phycosphere-associated bacteria are most critical. Incorporating metagenomic strategies will not only allow for more precise investigation in metabolic interactions in the diatom phycosphere, but also will allow for investigation into the presence of HGT between diatoms and their bacterial partners. Any evidence of HGT between diatoms and

their bacterial metagenome will further elucidate the longevity, importance, and evolution of these relationships. Considering the diversity of diatom species, and their critical roles in both aquatic ecosystems and global carbon cycles, improving our understanding of diatom-bacteria interactions has the potential to significantly improve our understanding of the evolution of host-symbiote co-evolution and its effect on ecosystem processes.

2.6 Data Availability

Sequencing data used for assembly of the genomes of *P. japonica* and its co-cultured bacterial taxa was deposited on NCBI under the Bioproject PRJNA476996. The full data analysis workflow, including rationale, R code, and commands used in analyses, and raw analysis output files can be found in the Quarto document 'psammoneis_project_workflow.qmd' in Zenodo repository for this manuscript (10.5281/zenodo.7464459).

2.7 Acknowledgements

The authors would like to thank the AHPCC for its services and infrastructure. Matt Ashworth for providing the *P. japonica* culture.

2.8 References

Alexa A, Rahnenfuhrer J. **2022**. *topGO: enrichment analysis for gene ontology*.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. **1990**. Basic local alignment search tool. *Journal of molecular biology* **215**: 403–410.

Amin SA, Green DH, Hart MC, Kupper FC, Sunda WG, Carrano CJ. **2009**. Photolysis of iron-siderophore chelates promotes bacterial-algal mutualism. *Proceedings of the National Academy of Sciences* **106**: 17071–17076.

Amin SA, Hmelo LR, van Tol HM, Durham BP, Carlson LT, Heal KR, Morales RL, Berthiaume CT, Parker MS, Djunaedi B, *et al.* 2015. Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* **522**: 98–101.

Amin SA, Parker MS, Armbrust EV. **2012**. Interactions between Diatoms and Bacteria. *Microbiology and molecular biology reviews: MMBR* **76**: 667–684.

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**: 2251–2252.

Archibald JM. **2015**. Endosymbiosis and Eukaryotic Cell Evolution. *Current biology: CB* **25**: R911–21.

Armbrust EV. **2009**. The life of diatoms in the world's oceans. *Nature* **459**: 185–192.

Armenteros JJA, Salvatore M, Emanuelsson O, Winther O, Von Heijne G, Elofsson A, Nielsen H. **2019**. Detecting sequence signals in targeting peptides using deep learning. *Life science alliance* **2**.

Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, Stamatakis A. **2018**. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Systematic biology* **68**: 365–369.

Barreto Filho MM, Walker M, Ashworth MP, Morris JJ. **2021**. Structure and Long-Term Stability of the Microbiome in Diverse Diatom Cultures. *Microbiology spectrum* **9**: e0026921.

Behringer G, Ochsenkühn MA, Fei C, Fanning J, Koester JA, Amin SA. **2018**. Bacterial Communities of Diatoms Display Strong Conservation Across Strains and Time. *Frontiers in microbiology* **9**: 659.

Bell W, Mitchell R. **1972**. Chemotactic and growth responses of marine bacteria to algal extracellular products. *The Biological bulletin* **143**: 265–277.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. **2004**. Improved prediction of signal peptides: SignalP 3.0. *Journal of molecular biology* **340**: 783–795.

Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson RH. **2015**. METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular ecology resources* **15**: 1403–1414.

Bennke CM, Neu TR, Fuchs BM, Amann R. **2013**. Mapping glycoconjugate-mediated

interactions of marine Bacteroidetes with diatoms. *Systematic and applied microbiology* **36**: 417–425.

Berendsen RL, Pieterse CMJ, Bakker PAHM. **2012**. The rhizosphere microbiome and plant health. *Trends in plant science* **17**: 478–486.

Bertrand EM, McCrow JP, Moustafa A, Zheng H, McQuaid JB, Delmont TO, Post AF, Sipler RE, Spackeen JL, Xu K, *et al.* 2015. Phytoplankton-bacterial interactions mediate micronutrient colimitation at the coastal Antarctic sea ice edge. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 9938–9943.

Bock R. **2010**. The give-and-take of DNA: horizontal gene transfer in plants. *Trends in plant science* **15**: 11–22.

Bolger AM, Lohse M, Usadel B. **2014**. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Boschetti C, Carr A, Crisp A, Eyres I, Wang-Koh Y, Lubzens E, Barraclough TG, Micklem G, Tunnacliffe A. **2012**. Biochemical Diversification through Foreign Gene Expression in Bdelloid Rotifers (J Zhang, Ed.). *PLoS genetics* **8**: e1003035.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, *et al.* 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239–244.

Bright M, Espada-Hinojosa S, Lagkouvardos I, Volland J-M. **2014**. The giant ciliate Zoothamnium niveum and its thiotrophic epibiont Candidatus Thiobios zoothamnicoli: a model system to study interspecies cooperation. *Frontiers in microbiology* **5**: 145.

Brock DA, Read S, Bozhchenko A, Queller DC, Strassmann JE. **2013**. Social amoeba farmers carry defensive symbionts to protect and privatize their crops. *Nature communications* **4**: 2385.

Brodie J, Ball SG, Bouget F-Y, Chan CX, De Clerck O, Cock JM, Gachon C, Grossman AR, Mock T, Raven JA, *et al.* 2017. Biotic interactions as drivers of algal origin and evolution. *The New phytologist* **216**: 670–681.

Buchfink B, Reuter K, Drost H-G. **2021**. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature methods* **18**: 366–368.

Burki F, Roger AJ, Brown MW, Simpson AGB. **2020**. The new tree of eukaryotes. *Trends in ecology & evolution* **35**: 43–55.

Bushnell B, Rood J, Singer E. **2017**. BBMerge – Accurate paired shotgun read merging via overlap. *PloS one* **12**: e0185056.

Casacuberta E, González J. **2013**. The impact of transposable elements in environmental adaptation. *Molecular ecology* **22**: 1503–1517.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. **2019**. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* .

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A,

Huddleston J, Eichler EE, *et al.* 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* **10**: 563–569.

Cholewińska P, Czyż K, Nowakowski P, Wyrostek A. **2020**. The microbiome of the digestive system of ruminants--a review. *Animal health research reviews / Conference of Research Workers in Animal Diseases* **21**: 3–14.

Copley RR, Bork P. **2000**. Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *Journal of molecular biology* **303**: 627–641.

Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. **2015**. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome biology* **16**: 50.

Croft MT, Lawrence AD, Raux-Deery E, Warren MJ, Smith AG. **2005**. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. *Nature* **438**: 90–93.

Czech L, Stamatakis A. **2018**. Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples. *bioRxiv*: 346353.

Dubilier N, Bergin C, Lott C. **2008**. Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nature reviews. Microbiology* **6**: 725–740.

Dunning Hotopp JC. **2011**. Horizontal gene transfer between bacteria and animals. *Trends in genetics: TIG* **27**: 157–163.

Durham BP, Sharma S, Luo H, Smith CB, Amin SA, Bender SJ, Dearth SP, Van Mooy BAS, Campagna SR, Kujawinski EB, *et al.* 2015. Cryptic carbon and sulfur cycling between surface ocean plankton. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 453–457.

Emms DM, Kelly S. **2018**. STAG: Species Tree Inference from All Genes. *bioRxiv*: 267914.

Fan X, Qiu H, Han W, Wang Y, Xu D, Zhang X, Bhattacharya D, Ye N. **2020**. Phytoplankton pangenome reveals extensive prokaryotic horizontal gene transfer of diverse functions. *Science Advances* **6**: eaba0111.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, *et al.* 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic acids research* **44**: D279–85.

Foster RA, Zehr JP. **2019**. Diversity, genomics, and distribution of phytoplankton-Cyanobacterium single-cell symbiotic associations. *Annual review of microbiology* **73**: 435–456.

Franke V, Ganesh S, Karlic R, Malik R, Pasulka J, Horvat F, Kuzman M, Fulka H, Cernohorska M, Urbanova J, *et al.* 2017. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome research* **27**: 1384–1394.

Fritsche TR, Gautom RK, Seyedirashti S, Bergeron DL, Lindquist TD. **1993**. Occurrence of bacterial endosymbionts in Acanthamoeba spp. isolated from corneal and environmental specimens and contact lenses. *Journal of clinical microbiology* **31**: 1122–1126.

Frost LS, Leplae R, Summers AO, Toussaint A. **2005**. Mobile genetic elements: the agents of open source evolution. *Nature reviews. Microbiology* **3**: 722–732.

Gíslason MH, Nielsen H, Almagro Armenteros JJ, Johansen AR. **2021**. Prediction of GPI-anchored proteins with pointer neural networks. *Current Research in Biotechnology* **3**: 6–13.

Görtz H-D. **2006**. Symbiotic Associations Between Ciliates and Prokaryotes. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E, eds. The Prokaryotes: Volume 1: Symbiotic associations, Biotechnology, Applied Microbiology. New York, NY: Springer New York, 364–402.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**: 644–652.

Graur D, Shuali Y, Li WH. **1989**. Deletions in processed pseudogenes accumulate faster in rodents than in humans. *Journal of molecular evolution* **28**: 279–285.

Grossart H-P, Levold F, Allgaier M, Simon M, Brinkhoff T. **2005**. Marine diatom species harbour distinct bacterial communities. *Environmental microbiology* **7**: 860–873.

Gruber A, Rocap G, Kroth PG, Armbrust EV, Mock T. **2015**. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *The Plant journal: for cell and molecular biology* **81**: 519–528.

Gschloessl B, Guermeur Y, Cock JM. **2008**. HECTAR: a method to predict subcellular targeting in heterokonts. *BMC bioinformatics* **9**: 393.

Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. **2013**. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular biology and evolution* **30**: 1987–1997.

Helliwell KE, Shibl AA, Amin SA. **2022**. The Diatom Microbiome: New Perspectives for Diatom-Bacteria Symbioses. In: Falciatore A, Mock T, eds. The Molecular Life of Diatoms. Cham: Springer International Publishing, 679–712.

Holt C, Yandell M. **2011**. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics* **12**: 491.

Horn M, Wagner M. **2004**. Bacterial endosymbionts of free-living amoebae. *The Journal of eukaryotic microbiology* **51**: 509–514.

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. **2013**. REAPR: a universal tool for genome assembly evaluation. *Genome biology* **14**: R47.

Husnik F, Tashyreva D, Boscaro V, George EE, Lukeš J, Keeling PJ. **2021**. Bacterial and archaeal symbioses with protists. *Current biology: CB* **31**: R862–R877.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, *et al.* 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236–1240.

Keeling PJ, Burki F. **08/2019**. Progress towards the Tree of Eukaryotes. *Current biology: CB* **29**: R808–R817.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. **2015**. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols* **10**: 845–858.

Klindworth A, Mann AJ, Huang S, Wichels A, Quast C, Waldmann J, Teeling H, Glöckner FO. **2014**. Diversity and activity of marine bacterioplankton during a diatom bloom in the North Sea assessed by total RNA and pyrotag sequencing. *Marine genomics* **18 Pt B**: 185–192.

Korf I. **2004**. Gene finding in novel genomes. *BMC bioinformatics* **5**: 59.

Laetsch DR, Blaxter ML. **2017**. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**: 1287.

Langmead B, Salzberg SL. **2012**. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**: 357–359.

Li H. **2013**. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. **2009**. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liu Y, Harrison PM, Kunin V, Gerstein M. **2004**. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome biology* **5**: R64.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, *et al.* 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**.

Majzoub ME, Beyersmann PG, Simon M, Thomas T, Brinkhoff T, Egan S. **2019**. Phaeobacter inhibens controls bacterial community assembly on a marine diatom. *FEMS microbiology ecology* **95**: fiz060.

Mann DG, Vanormelingen P. **2013**. An Inordinate Fondness? The Number, Distributions, and Origins of Diatom Species. *The Journal of eukaryotic microbiology* **60**: 414–420.

Marçais G, Kingsford C. **2011**. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764–770.

Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A, Grandbastien M-A, Bowler C. **2009**. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC genomics* **10**: 624.

McCutcheon JP, McDonald BR, Moran NA. **2009**. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 15394–15399.

Mira A, Ochman H, Moran NA. **2001**. Deletional bias and the evolution of bacterial genomes. *Trends in genetics: TIG* **17**: 589–596.

Mönnich J, Tebben J, Bergemann J, Case R, Wohlrab S, Harder T. **2020**. Niche-based assembly of bacterial consortia on the diatom *Thalassiosira rotula* is stable and reproducible. *The ISME journal*.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. **2007**. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research* **35**: W182–5.

Morris RM, Longnecker K, Giovannoni SJ. **2006**. Pirellula and OM43 are among the dominant lineages identified in an Oregon coast diatom bloom. *Environmental microbiology* **8**: 1361–1370.

Müller DB, Vogel C, Bai Y, Vorholt JA. **2016**. The plant Microbiota: Systems-level insights and perspectives. *Annual review of genetics* **50**: 211–234.

Nagano N, Orengo CA, Thornton JM. **2002**. One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions. *Journal of Molecular Biology* **321**: 741–765.

Nakov T, Beaulieu JM, Alverson AJ. **2018**. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *The New phytologist* **219**: 462–473.

Nawrocki E. **2009**. Structural RNA homology search and alignment using covariance models.

Ochman H, Lawrence JG, Groisman EA. **2000**. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299–304.

Ophir R, Graur D. **1997**. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* **205**: 191–202.

Osuna-Cruz CM, Bilcke G, Vancaester E, De Decker S, Bones AM, Winge P, Poulsen N, Bulankova P, Verhelst B, Audoor S, *et al.* 2020. The Seminavis robusta genome provides insights into the evolutionary adaptations of benthic diatoms. *Nature communications* **11**: 3320.

Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. **2020**. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature biotechnology* **38**: 1079–1086.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. **2018**. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology* **36**: 996–1004.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. **2015**. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**: 1043–1055.

Quinlan AR, Hall IM. **2010**. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Rice P, Longden I, Bleasby A. **2000**. EMBOSS: the European Molecular Biology Open

Software Suite. *Trends in genetics: TIG* **16**: 276–277.

Richardson N, Cook I, Crane N, Dunningto D, François R, Keane J, Moldovan-Grünfeld D, Ooms J, Apache Arrow. **2022**. arrow: Integration to 'Apache' 'Arrow'.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. **2011**. Integrative genomics viewer. *Nature biotechnology* **29**: 24–26.

Sato S, Kooistra WHCF, Watanabe T, Matsumoto S, Medlin LK. **2008**. A new araphid diatom genus Psammoneis gen. nov. (Plagiogrammaceae, Bacillariophyta) with three new species based on SSU and LSU rDNA sequence data and morphology. *Phycologia* **47**: 510–528.

Segev E, Wyche TP, Kim KH, Petersen J, Ellebrandt C, Vlamakis H, Barteneva N, Paulson JN, Chai L, Clardy J, *et al.* 2016. Dynamic metabolic exchange governs a marine algal-bacterial interaction. *eLife* **5**.

Seymour JR, Amin SA, Raina J-B, Stocker R. **2017**. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nature Microbiology* **2**: 17065.

Shah N, Nute MG, Warnow T, Pop M. **2019**. Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics* **35**: 1613–1614.

Sheikhizadeh S, de Ridder D. **2015**. ACE: accurate correction of errors using K-mer tries. *Bioinformatics* **31**: 3216–3218.

Shibl AA, Isaac A, Ochsenkühn MA, Cárdenas A, Fei C, Behringer G, Arnoux M, Drou N, Santos MP, Gunsalus KC, *et al.* 2020a. Diatom Modulation of Microbial Consortia Through Use of Two Unique Secondary Metabolites. *bioRxiv*: 2020.06.11.144840.

Shibl AA, Isaac A, Ochsenkühn MA, Cárdenas A, Fei C, Behringer G, Arnoux M, Drou N, Santos MP, Gunsalus KC, *et al.* 2020b. Diatom modulation of select bacteria through use of two unique secondary metabolites. *Proceedings of the National Academy of Sciences of the United States of America* **117**: 27445–27455.

Singh U, Wurtele ES. **2021**. orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics* .

Smith SA, O'Meara BC. **2012**. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**: 2689–2690.

Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. **2016**. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome research* **26**: 1134–1144.

Suzuki S, Kakuta M, Ishida T, Akiyama Y. **2014**. GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PloS one* **9**: e103833.

Syvanen M. **2012**. Evolutionary implications of horizontal gene transfer. *Annual review of genetics* **46**: 341–358.

Tarailo-Graovac M, Chen N. **2009**. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 4**: Unit 4.10.

Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD,

Winther O, Brunak S, von Heijne G, Nielsen H. **2022**. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature biotechnology* **40**: 1023–1025.

Thumuluri V, Almagro Armenteros JJ, Johansen AR, Nielsen H, Winther O. **2022**. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic acids research* **50**: W228–34.

UniProt Consortium. **2021**. UniProt: the universal protein knowledgebase in 2021. *Nucleic acids research* **49**: D480–D489.

Vancaester E, Depuydt T, Osuna-Cruz CM, Vandepoele K. **2020**. Comprehensive and Functional Analysis of Horizontal Gene Transfer Events in Diatoms. *Molecular biology and evolution* **37**: 3243–3257.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, *et al.* 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**: e112963.

Wierenga RK. **2001**. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS letters* **492**: 193–198.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. **01/2017**. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data (G McInerny, Ed.). *Methods in ecology and evolution / British Ecological Society* **8**: 28–36.

Zhang Z, Carriero N, Gerstein M. **2004**. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends in genetics: TIG* **20**: 62–67.

Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, Karaoz U, Anantharaman K. **2022**. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* **10**: 1–22.

Microbiome data scavenged from diatom (Bacillariophyta) transcriptomes reveals high diversity and cophylogentic congruence between diatoms and their bacterial consortia.

Cory B. Gargas[1], Rachel A. Ungar[2], Shady A. Amin[3], Andrew J. Alverson[1]

[1] Alverson Lab, Biology department, University of Arkansas, Fayetteville, AR 72701, USA

[2] Stanford University, Stanford, CA 94305, USA

[3] New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

Author Contributions: I performed and designed the final data processing and all analyses, code production, figures and tables, and wrote the manuscript. Rachel Ungar performed initial exploratory analysis of the dataset. Shady Amin assisted with interpretation of some results and literature recommendations. Andrew Alverson provided project conception, funding, and transcriptomes, as well as assisting with analysis and providing comments and edits to this manuscript.

3.1 Abstract

Diatoms are one of the most abundant groups of microalgae. A key factor in this dominance is their close partnerships with heterotrophic bacteria. Diatoms secrete dissolved organic matter that attracts bacteria, which in turn provide the diatom with essential nutrients and cofactors. These associations have been coined as the 'phycosphere,' an algal analog to the rhizosphere in plants. Genomic projects involving diatoms frequently sequence both diatoms and their co-occurring bacteria, the latter of which aren't usually analyzed. Here, I assembled bacterial 16S sequences from 274 diatom transcriptomes to investigate bacterial diversity and community phylogenetics across a broad diversity of diatoms. From these transcriptomes, I recovered 26,000 prospective prokaryotic 16S sequences, culminating in 4,033 OTUs. Of the 4,033 OTUs found, 3,252 were found to be unique to individual diatom cultures. I found a high degree of dissimilarity in phylogenetic beta-diversity between diatom bacterial communities, even in closely related species of diatoms. Ordination analysis of phylogenetic beta-diversity demonstrated distinct groupings of diatom microbiomes by salinity. PERMANOVA analysis confirmed that beta-diversity was significantly different between these groups. Our results support that diatom phycosphere communities are more similar within salinity levels, while still maintaining high diversity within and across genera. I also investigated the cophylogenetic concordance of diatom-bacteria associations and analyzed these associations in relation to phylogeny, salinity, and other factors. Significant cophylogenetic concordance was found between diatoms of the genus *Chaetoceros* and their bacterial partners, suggesting that some diatom–bacteria relationships are maintained over evolutionary time scales. Off-target sequences that are incidentally collected from genome and transcription data can provide powerful new insights into diatom biology and evolution.

3.2 Introduction

The interactions between bacteria and phytoplankton support a plethora of biological interactions (Cole, 1982; Azam & Malfatti, 2007; Seymour *et al.*, 2017). At their most basic, eukaryotic phytoplankton rely on bacteria for the remineralization of organic matter back to inorganic forms to support their growth (Field *et al.*, 1998; Falkowski *et al.*, 2008). In turn, bacteria depend use the organic carbon produced by phytoplankton to support their own growth (Cho & Azam, 1988; Worden *et al.*, 2015). The majority of these interactions are thought to occur in the organically rich area surrounding algal cells, termed the phycosphere (Bell & Mitchell, 1972; Amin *et al.*, 2012; Seymour *et al.*, 2017). The phycosphere is characterized as the region of relatively high concentrations of organic molecules released by algal cells relative to that found in seawater (Biddanda & Benner, 1997). This concept is analogous to that of the rhizosphere in plants (Mendes *et al.*, 2011). Bacteria are able to colonize the phycosphere actively (e.g., through chemotaxis) or passively (e.g., through random encounters), and may also be passed on to daughter cells following cell division (Seymour *et al.*, 2017). Given the importance of marine phytoplankton in ocean ecosystems, it is increasingly clear that the interactions occurring at the level of single cells in the phycosphere have globally important implications. A complete understanding of these interactions is therefore essential to understanding basic aspects of diatom biology and ecosystem function.

For phytoplankton, these interactions range from beneficial (e.g., commensalistic or mutualistic) to detrimental (e.g., parasitic or predatory; Amin *et al.*, 2012; Kazamia *et al.*, 2016; Seymour *et al.*, 2017). In one mutually beneficial interaction, bacteria synthesize essential vitamins that are taken up by the diatom in exchange for photosynthates (Durham *et al.*, 2015, 2017). More than half of all microalgal species are auxotrophic for the vitamins B_1, B_7, and B_{12} and have to acquire them from outside sources (Croft *et al.*, 2005; Tang *et al.*, 2010). Iron, a limiting element in the pelagic zone (Martin & Fitzwater, 1988), can be increased in bioavailability to phytoplankton via the production of iron-binding chelates by bacteria (Amin *et*

al., 2009). phytoplankton-bacteria relationships go so far as to affect cell-cycles in both bacteria and diatoms. The diatom *Asterionellopsis glacialis* excretes rosmarinic and azelaic acids to promote the attachment and growth of beneficial bacteria, while also suppressing the attachment and growth of non-beneficial bacteria (Shibl *et al.*, 2020). An opposite interaction, in which bacteria affect their diatom host, occurs with the *Roseobacter* clade bacteria, which convert algal-secreted tryptophan into indole-3-acetic acid (IAA). The hormone IAA enhances algal cell division and may potentially increase its carbon output to bacteria (Amin *et al.*, 2015; Segev *et al.*, 2016).

Understanding interactions in the phycosphere requires a full determination of the identity and relative abundance of bacteria that interact with phytoplankton. Most studies to date have identified bacteria from clonal algal cultures or through metagenomic studies of algal blooms. One difficulty with the former approach is that the majority of bacteria are difficult, or impossible, to isolate and maintain in culture in the lab, which may cause some relationships to break down *in situ* (Rappé & Giovannoni, 2003). For example, bacteria isolated from marine diatom cultures and blooms belong to only a fraction of the total number of genera found in the surrounding seawater (Sapp *et al.*, 2007; Amin *et al.*, 2012; Baker & Kemp, 2014). Similar differences between bacterial consortia and surrounding seawater have been observed for other phytoplankton clades (Biegala *et al.*, 2002; Green *et al.*, 2004; Hasegawa *et al.*, 2007; Eigemann *et al.*, 2013). While this highlights the small number of phycosphere associated bacteria that are currently culturable, some of this research is indicative of a core microbiome in the diatom *Thalassiosira rotula* (Mönnich *et al.*, 2020) and the cyanobacterium *Trichodesmium (Frischkorn et al., 2017)*. Other evidence suggests that some microalgae do not possess a core microbiome. For example, across 13 cultures of the green alga *Ostreococcus tauri*, no core microbiome was apparent at the genus level (Abby *et al.*, 2014). Another question is whether co-evolution is occurring between these diatoms and their bacterial consortia.

Historically, co-evolution is defined as the reciprocal changes in two organisms as a result of the selection pressures they impose on one another, which promotes co-adaptation over time (Janzen, 1980), which can lead to co-speciation. Phylosymbiosis is the observation that closely related hosts tend to have microbiomes that are more similar in composition than those of distantly related hosts. Unless treated with antibiotics, diatoms maintain some or all of their bacterial associations in culture. Most diatoms cannot be maintained in axenic conditions for extended periods of time, so it is typical for algal genome projects to contain some amount of non-target DNA reads from resident bacteria (Gargas *et al.*, 2020). I mined the transcriptomes from 274 diverse diatoms for bacterial 16S rDNA reads and used these scavenged sequences to peer into the global diatom phycosphere. I used these data to characterize patterns of diatom–bacterial associations across major diatom clades, habitats, and to test whether some of the more densely sampled diatom lineages show evidence of co-speciation with their associated bacteria.

3.3 Materials & Methods

3.3.1 Programs, packages, and data sources

The following R packages were used in our analyses: ape (Paradis *et al.*, 2004; Paradis & Schliep, 2018), dbConnect (James, 2012), dbplyr, dplyr, ggplot2 (Wickham *et al.*, 2008), ggridges, gt, here, PACo (Balbuena *et al.*, 2013; Hutchinson *et al.*, 2017), patchwork, PhyloMeasures (Tsirogiannis & Sandel, 2016), phyloseq (McMurdie & Holmes, 2013), picante (Kembel *et al.*, 2010), reticulate (Allaire *et al.*, 2017), reshape2 (Wickham & Others, 2007), RSQLite, tidyverse (Wickham *et al.*, 2019), viridis (Simon Garnier, Noam Ross, Bob Rudis, Marco Sciaini, Cédric Scherer, 2018). The following command line (CLI) programs were also utilized in our analyses: BLAST+ (Camacho *et al.*, 2009) using the National Center for Biotechnology Information (NCBI) database (DB; dbV5); CD-HIT (Li & Godzik, 2006; Fu *et al.*, 2012), EPA-NG (Barbera *et al.*, 2019) & Gappa (Czech & Stamatakis, 2019), IQ-TREE2 (Minh *et al.*, 2019).

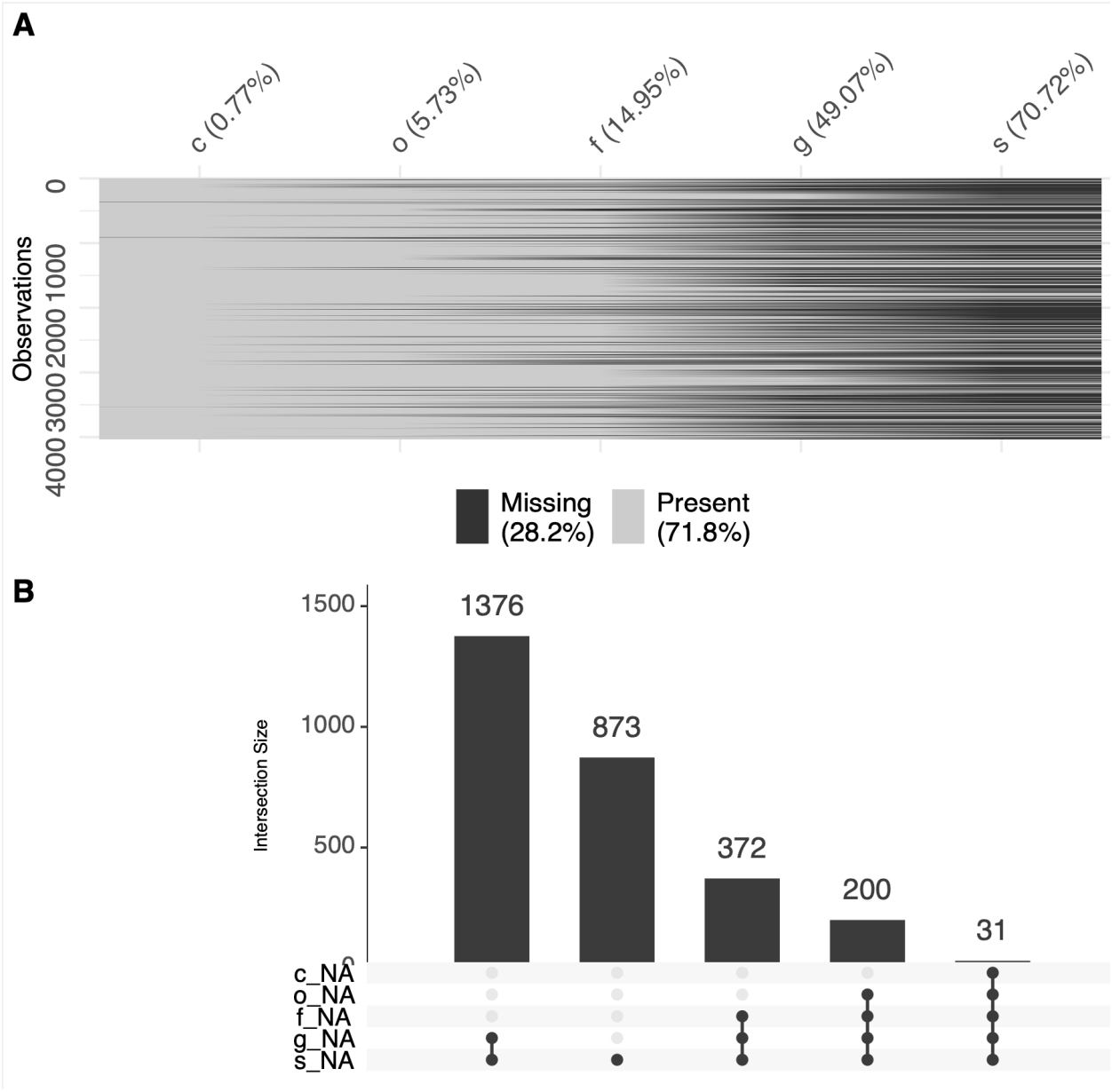

Figure 3.1. Missing data visualizations of taxonomic assignments for Bacteria (A-B) & Archaea (C-D)(C-D). Missingness by variable plots (A & C) depict missing data for each row in each variable. Variables c, o , f, g, & s indicate the taxonomic ranks class, order, family, genus & species, respectively. Parenthetical values next to variable names indicate percent missingness for each variable. Upset plots (B & D) depict intersections and counts of missing data for the variables in A & C.

Diatom transcriptomes were obtained from two sources: the Marine Microbial Eukaryote Transcriptome Sequencing Project (Keeling *et al.*, 2014) and Alverson *et al.*, (2023). A total of

274 transcriptomes were obtained from both sources, 185 from AJA and 77 from MMETSP. Reference 16S sequences and phylogenies for Archaea (ar122) and Bacteria (bac120) were obtained from release 202 of the Genome Taxonomy Database (GTDB; Parks *et al.*, 2018, 2020).

3.3.2 Data procurement, filtering, and subsetting

Trimmomatic (Bolger *et al.*, 2014) was used to trim and clean Illumina short reads. Bowtie2 (Langmead *et al.*, 2009, 2019; Langmead & Salzberg, 2012) was used to map trimmed Illumina reads to a reference database of 16S ribosomal rRNA sequences (Hug *et al.*, 2016). Mapped reads were merged with BBMerge (Bushnell *et al.*, 2017) and assembled with Trinity (Grabherr *et al.*, 2011).

I used the following workflow to generate operational taxonomic units (OTUs) from our query sequences. I first used a custom Python script to concatenate, measure, and assign unique IDs to each sequence. I then removed sequences < 250bp in length from our dataset using dplyr::filter(). These sequences were exported into a FASTA file that was used as input to CD-HIT using cd-hit-est with options: -c 0.97 -n 10 -d 0 -M 8000 -T 4. These sequence clusters were used as OTUs in downstream analyses.

Next I assigned taxonomic groupings at the level of domain to our OTUs via NCBI-BLASTN to perform taxonomy-restricted searches of each OTU sequence against archaeal, bacterial, and eukaryotic sequences from the NCBI nt database (ftp://ftp.ncbi.nlm.nih.gov/blast/; downloaded 20 November 2021). This resulted in bitscores for each query sequence for each domain. I then used a modified version of the HGT index (Boschetti *et al.*, 2012) to assign domain-level taxonomy to each query sequence. The original HGT index only indicated whether sequences are metazoan or non-metazoan in origin. I modified the HGT index into the archaeal (HGT_A) and bacterial (HGT_B) HGT indices to assess whether OTU sequences were archaeal, bacterial, or eukaryotic in origin. The formulas for the HGT_A and HGT_B are as follows: HGT_A = (bit-score for top archaeal hit) - (bit-score for top

eukaryotic hit) and HGT_B = (bit-score for top bacterial hit) - (bit-score for top eukaryotic hit). I then used the criteria depicted in Table 1 to determine which domain to place each OTU, based on the scores for HGT_A and HGT_B.

Table 3.1. Criteria used to assign 16S OTUs to domains after calculating the HGT index for archaea and bacteria for each OTU.

Table 1. Formulae used to assign 16S OTUs to domains after calculating the HGT index for archaea and bacteria for each OTU.

Archaea domain assignment[1,2,3,4,5]	
	(HGTa > 0) > (HGTb > 0)
	(HGTa > 0) & (HGTb = NULL)
	[(HGTa > 0) > (HGTb > 0)] & bitscore_e = NULL
Bacteria domain assignment[1,2,3,4,6]	
	(HGTb > 0) > (HGTa > 0)
	(HGTb > 0) & (HGTa = NULL)
	(bitscore_a < bitscore_b) & (bitscore_e = NULL)
Eukarya domain assignment[1,2,3,4,7]	
	(HGTa ≤ 0) & (HGTb ≤ 0)
	(HGTa = NULL) & (HGTb ≤ 0)
	(HGTa ≤ 0) & (HGTb = NULL)
	(bitscore_e > 0) & (HGTa = NULL & HGTb = NULL)
	(bitscore_e > bitscore_a) & (bitscore_b = NULL)
	(bitscore_e > bitscore_b) & (bitscore_a = NULL)
	HGTa < 0 & HGTb = 0
	HGTb < 0 & HGTa = 0
Indeterminate domain assignment[1,2,3,4,8]	
	HGTa & HGTb < 0
	HGTa & HGTb = NULL
	(bitscore_e = bitscore_a) & (bitscore_b = NULL)
	(bitscore_e = bitscore_b) & (bitscore_a = NULL)

[1] HGTa = (bitscore for top archaeal hit) - (bitscore for top eukaryotic hit)
[2] HGTb = (bitscore for top bacterial hit) - (bitscore for top eukaryotic hit)
[3] The letter suffix after bitscore indicates the domain of that bitscore. these bitscores are for the top hit in that domain.
[4] NULL indicates that that value could not be recovered, or calculated due to a lack of data, by NCBI BLAST.
[5] Formulae used to assign OTUs to Domain Archaea
[6] Formulae used to assign OTUs to Domain Bacteria
[7] Formulae used to assign OTUs to Domain Eukarya
[8] Formulae used to assign OTUs to Domain 'Indeterminate.' This indicates that scores were too close or uncertain to confidently assign an OTU to a domain.

3.3.3. Reference & query sequence alignment and model finding

SSU-ALIGN (v0.1.1; Nawrocki, 2009) was used to construct separate 16S OTU alignments for bacteria and archaea by aligning our OTUs against NCBI nt 16S reference

sequences for each domain against the respective 16S query sequences using SSU-ALIGN and SSU-MASK. Individual searches were restricted to covariance models for each particular domain, such as bacteria or archaea. The aligned reference and query sequences were used as input for phylogenetic placement.

3.3.4 Phylogenetic placement analyses and taxonomic assignment of query sequences

In order to perform phylogenetic placements on OTUs, I needed to determine the best nucleotide substitution model for our sequences. IQ-tree was used to determine the nucleotide substitution model to use for our phylogenetic placement analyses using the option: -m TESTONLY (Kalyaanamoorthy *et al.*, 2017). The ModelFinder function of IQ-tree searches tree space for the best matching model of sequence evolution for the input sequence alignment. Model finding was performed on both the archaeal and bacterial alignments, IQ-tree found that SYM+I+G_4 was found to be the best model of substitution for both. To assign taxonomy to our archaeal and bacterial query sequences I used the programs EPA-NG & Gappa for phylogenetic placement & taxonomic assignment, respectively. EPA-NG was run using the respective query and reference data for each domain & the options: --model SYM+I+G4 --preserve-rooting on. Gappa was used with options ‘examine assign --best-hit --resolve-missing-paths’ and ‘--examine graft’ to generate taxonomic assignments for query sequences and the reference phylogeny with placed query sequences, respectively. Support values for placed sequences are given as likelihood weight ratios ranging from 0 – 1 in value. The likelihood weight ratios indicate the probability that a placed query sequence is located somewhere along a specific branch, , with lower values indicating lower probabilities. I then filtered any sequences that had taxonomic assignments only to the level of domain. I also filtered any sequences that had taxonomic assignments with likelihood weight ratio values <0.5. Finally, I removed OTUs matching bacterial genera that are common lab and reagent contaminants (Kulakov *et al.*, 2002; Salter *et al.*, 2014; Glassing *et al.*, 2016), a complete list of which is available in the accompanying data archive.

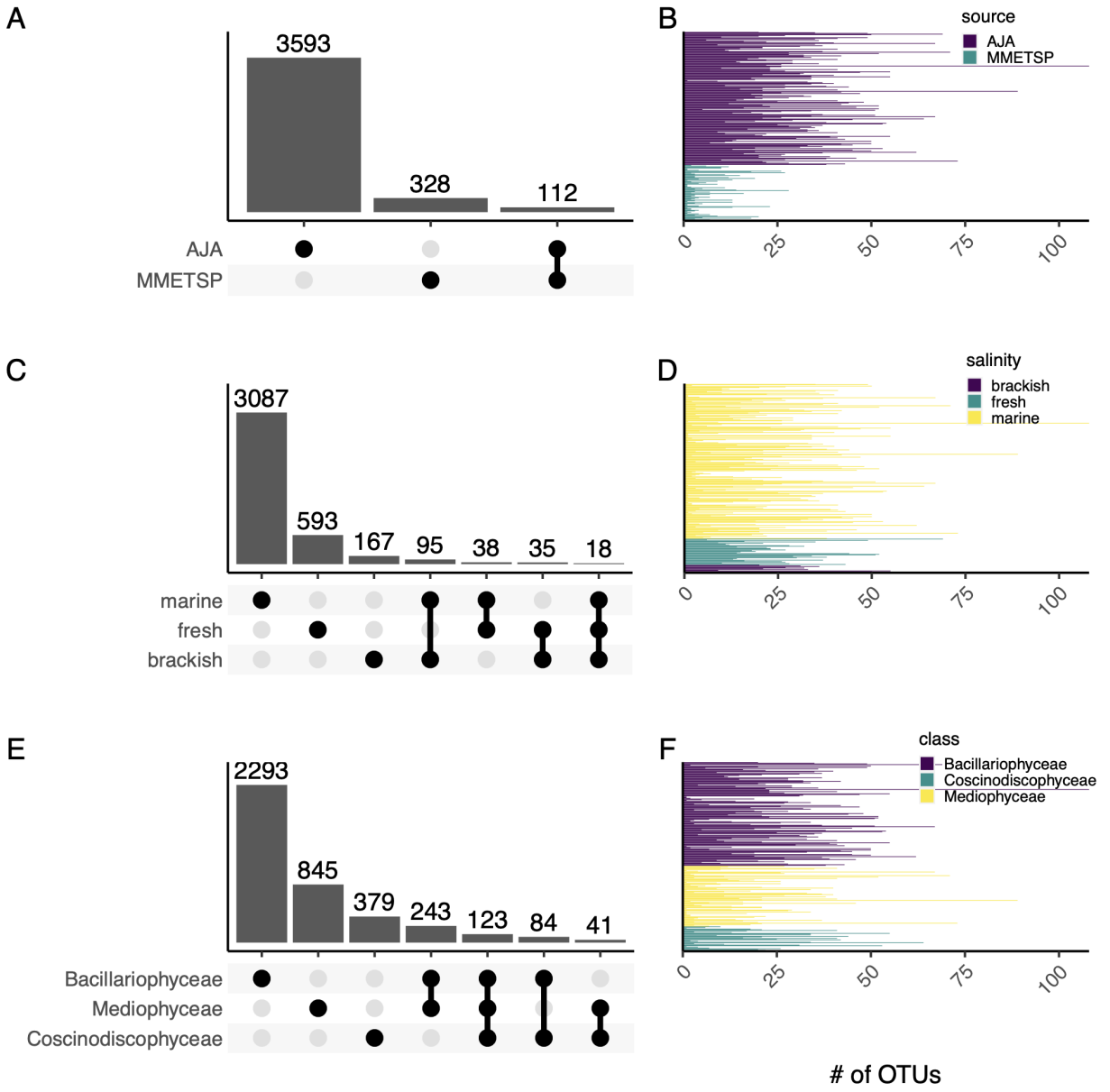

Figure 3.2. OTU occupancy plots for bacteria filtered by lab (A and B), salinity (C and D), and diatom class (E and F). Upset plots (A, C, and E) depict OTU counts by intersection of levels in each variable of interest. Bar plots (B, D, and F) indicate the number of OTUs (x-axis) for each transcriptome (y-axis), colored by levels in each variable and grouped from highest to lowest occupancy.

3.3.5 Community phylogenetic, ordination, phylogenetic diversity, and phylogenetic signal analyses

To examine the alpha- and beta-diversity of diatom bacterial consortia, I used the R package Phylomeasures and phyloseq (McMurdie & Holmes, 2013), respectively. For phylogenetic alpha-diversity, I calculated the standardized and non-standardized versions of the following measures of: core ancestor cost (CAC; Tsirogiannis *et al.*, 2014), Faith's phylogenetic distance (FI; Faith, 1992), mean nearest taxon distance (MNTD; Webb, 2000; Vellend et al., 2011), and mean pairwise distance (MPD; Webb *et al.*, 2003). The CAC is a single-sample measure defined as the distance of the most recent common ancestor (MRCA) node from the tree root of at least chi (X) proportion of the species in a sample, where X is defined as a value between 0.5 and 1. The FI measure is defined as the total branch length in the minimum spanning subtree for a sample. The definition of MNTD is defined as the mean distance from each taxon to its nearest neighbor in a sample. MPD is defined as the mean pairwise distance of branches along the tree between all species in a sample. The measures CAC, FI, and MPD were standardized according to the species richness of each sample (or pair of samples) using the option standardize=TRUE and are indicated with a suffixed ‘s’ (e.g., CAC_s, FI_s, and MPD_s). Standardized measures range from negative to positive in value, with negative values suggesting a bias towards taxa that are more distantly related to each other (i.e., phylogenetic overdispersion) and positive values indicating a bias towards species that are more closely related to each other (phylogenetic clustering). Each measure used 1000 Monte-Carlo random repetitions and the option chi=0.7 was used for our CAC analysis.

I then calculated the phylogenetic beta-diversity between each sample (i.e., the 16S sequences from a given diatom transcriptome) using the unweighted version of phylogenetic

beta-diversity metric UniFrac (uwUF; Lozupone & Knight, 2005; Lozupone *et al.*, 2006). UniFrac

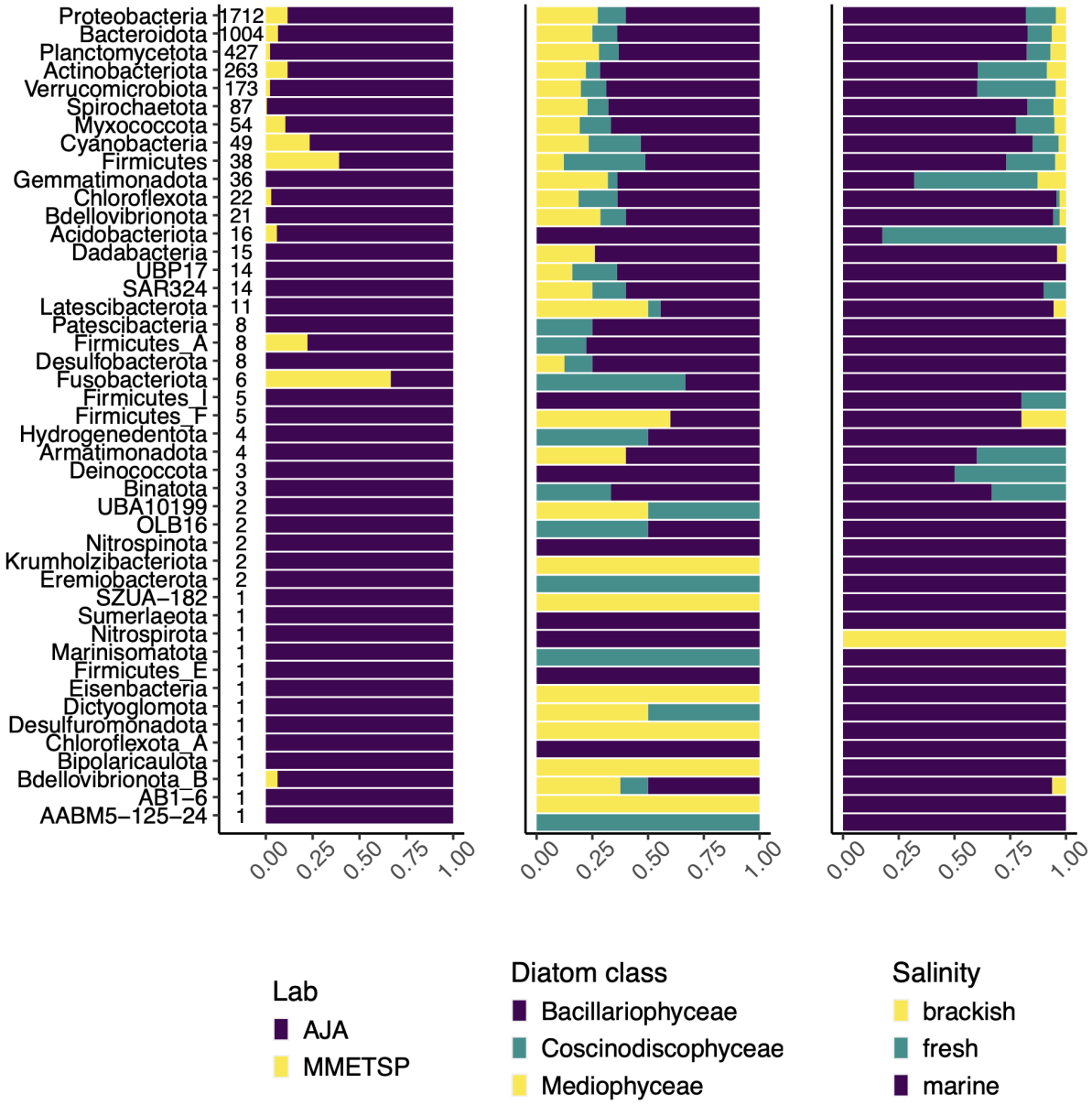

Figure 3.3. Stacked bar plots of bacterial phyla (y-axis) by percent (x-axis) colored by lab (A), salinity (B), and diatom class (C). Values to the left of each bar indicate the number of OTUs assigned to each phylum. bar plots are sorted from greatest to least by number of OTUs.

is a phylogenetic beta-diversity metric of the unique fraction of total phylogenetic diversity between individual communities or samples. The result of this analysis is a matrix that contains dissimilarity values for each pairwise comparison between communities, with greater similarity

indicated by lower values (i.e., a value of 0 meaning no difference in communities and a value of 1 meaning completely different). I used the function UniFrac (options: weighted = FALSE, normalized = TRUE, parallel = TRUE, fast = TRUE) from the R package phyloseq to perform this analysis. I also tested whether our uwUF values were different from a set of simulated uwUF values. I accomplished this using the function phyloseq_randomize (options: null_model = "independentswap", verbose = TRUE) from the package metagMisc to randomize our OTU association matrix. I ran uwUF on the output from phyloseq_randomize and used the base R function ‘replicate()’ to repeat this 100 times and then calculated the average value for those 100 replicates. Finally, I performed a t-test on the observed vs. mean simulated values to determine if the observed values were significantly different from the simulated values. If the observed and simulated values were found to be significantly different, observed uwUF values were then considered to not be from stochasticity alone. If no difference was observed, then the observed values would be considered to be the result of stochasticity in our data.

I also used the ordinate() function from the R package phyloseq to perform a non-metric multidimensional scaling (NMDS) on our Alverson *et al.*, (2023) bacterial dataset. The uwUF distances allow for the incorporation of phylogenetic information and other metadata variables into the NMDS ordinations. The uwUF distance was chosen because our data is presence-absence only, rendering weighted UniFrac unnecessary.

I calculated the phylogenetic signal K statistic for our bacterial query phylogenies using the function multiphylosignal() from the R package picante. I used our transcriptome metadata table to assign traits to each OTUs in our query-only phylogenies. The K statistic describes the tendency for related organisms to have similar phenotypes, with K<1 indicating relatives resemble each other less than expected and K>1 indicating that relatives resemble each other more than expected under brownian motion (BM) evolution. If the value of K is close, or equal, to one, this indicates that trait is evolving according to BM. I would expect to see K<1 in instances of homoplasy and K>1 would be expected when a trait is a synapomorphy.

Initially, analysis of variance (ANOVA) was performed on OTU abundances for each transcriptome to ascertain whether there were significant differences in abundance between transcriptomes from MMETSP and Alverson *et al.*, (2023). I also performed ANOVA on OTU abundances for salinity levels for the AJA transcriptomes. ANOVA was also performed on the results of our phylogenetic alpha-diversity metrics and beta-diversity metric (uwUF). I then computed Tukey honest significant differences for the results of each of our ANOVAs. Lastly, note that because uwUF consists of pairwise comparisons between communities Tukey HSD results will consist of pairwise comparisons of those pairwise variables, thus will have Tukey results consisting of marine-marine mean distances against freshwater-marine mean distances because the original pairwise comparisons of communities resulted in distances for marine-marine communities and freshwater-marine communities. Linear regression models (LMs) were fitted on OTU abundance data and uwUF distances against years in culture for each diatom transcriptome. Because years in culture were not known for all diatom cultures, this subset our data to a total of 166 transcriptomes, 120 and 34 belonging to AJA and MMETSP, respectively.

Non-metric multidimensional scaling (NMDS) was conducted on combined and AJA transcriptome data using the function ordinate() from the phyloseq R package. I chose to use our uwUF dissimilarity matrices as the input distance measure for our NMDS so I could incorporate our community phylogenetic data into the analysis. I then performed a phylogenetically informed, nonparametric multivariate test for differences to obtain a rigorous probabilistic statement about our metadata variables (Anderson, 2001). Using our uwUF distance matrices as input and our transcriptome metadata variables as factors. Specifically, I performed a permutational multivariate analysis of variance (PERMANOVA) using the function adonis.

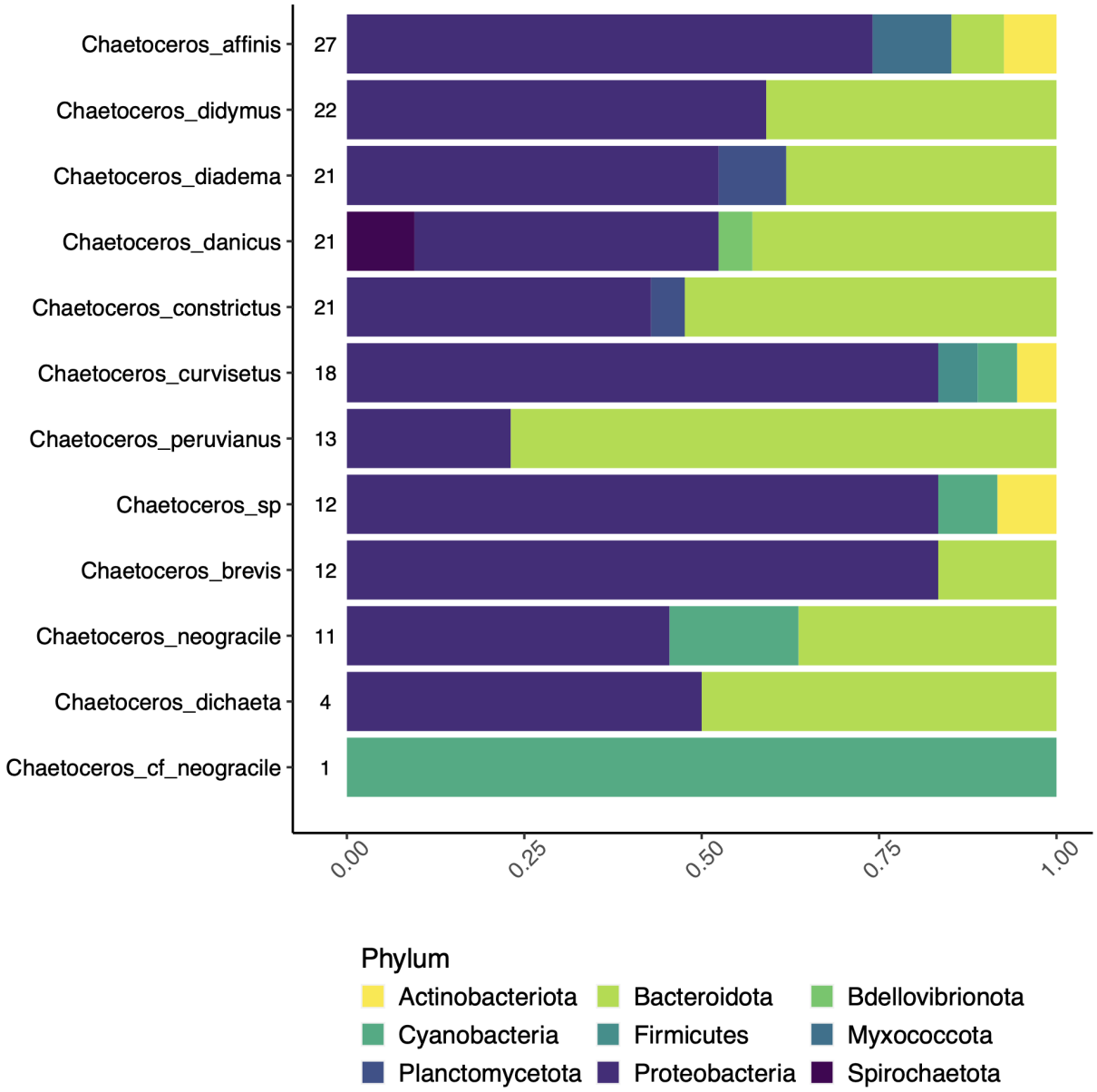

Figure 3.4. Stacked bar plots for bacterial OTUs of *Chaetoceros* transcriptomes (A, y-axis) and archaeal OTUs (B, y-axis). Bars are colored by phylum. Values to the left of each bar indicate the number of OTUs assigned to each phylum. bar plots are sorted from greatest to least by number of OTUs.

3.3.6 Cophylogenetic concordance analysis

Cophylogeny studies have been used to investigate relationships between symbiont(s) (mutualistic, commensalistic, or parasitic) and hosts. I used the R package PACo to determine whether the topology of the diatom phylogeny was concordant with the topology of our archaeal

or bacterial phylogenies with the following options: nperm=1000, symmetric=TRUE, method="quasiswap", shuffled=TRUE. PACo calculates cophylogenetic concordance between two phylogenies, such as two gene trees or host–parasite phylogenies. PACo accomplishes this by converting each phylogeny into a distance matrix, converting those into extended principal coordinate matrices using a host–parasite link matrix, and then scaling and rotating the parasite configuration to fit the host configuration using procrustean superimposition analysis. This then provided us with a global sum of squared residuals (m^2_{xy}) and residuals showing the contribution of individual links. I also subset phylogenies to focus on individual diatom clades and their bacteria. Most cophylogentic concordance studies look at 1:1 associations between hosts and parasites. Since the bacterial communities of diatoms represent 1:many associations, similar to those seen in pollination networks, I chose the 'quasiswap' method and symmetric procrustes statistic. The 'quasiswap' option conserves the number of interactions for each species and the symmetric procrustes statistic makes the assumption that one group tracks the other (Hutchinson *et al.*, 2017). In these analyses, m^2_{xy} is used to determine the amount of congruence (similarity) between the host and bacterial community phylogenies. Residual values of individual links (connections between tips in host and symbiont phylogenies) can be examined to see which links contribute the most to cophylogeny, the lower the value of an individual link, the more congruent it is. Perfect cophylogenetic congruence (identical topology & branch lengths) would be indicated by $m^2_{xy} = 0$, partial congruence would be indicated by values of $0 < m^2_{xy} < 1$, and lower congruence as m^2_{xy} values increase beyond 1. For large datasets with many links (e.g., microbial communities, pollination networks, etc.) and use the asymmetric procrustes statistic (option: symmetric = FALSE), you can end up with quite large m^2_{xy} values (> 10-100) that can still be indicative of significant phylogenetic congruence if the observed arrangement of links shows substantially greater congruence than a null model for link arrangement. When the symmetric procrustes statistic (option: symmetric = TRUE) is chosen, both phylogenies are standardized prior to procrustes superimposition, resulting in the best-fit of

superimposition being independent of both phylogenies and an m^2_{xy} value that is constrained between 0 and 1 (Hutchinson *et al.*, 2017). To determine which diatom clades to investigate for cophylogentic concordance I examined species diversity per genus and number of bacteria per species, this resulted in us choosing the genus *Chaetoceros* as it was the most well sampled, in terms of transcriptome number, and number of bacterial OTUs recovered per species. The phylogeny of bacteria associated with *Chaetoceros* was then analyzed against the respective diatom phylogeny for the genus *Chaetoceros.* To create our diatom phylogeny for *Chaetoceros,* I pruned the transcriptome-based phylogeny from Alverson *et al.*, (2023) to only include *Chaetoceros* species from our microbiome analyses. I then ran PACo on the best phylogeny from these and all other bootstrapped trees to determine whether our PACo results are significantly different from a random distribution of PACo values from a random set of bootstrap trees. Bacterial phylogenies were generated using EPA-NG & Gappa as described above in section 3.3.4.

To test whether our observed PACo results were real and not a result of stochasticity, I performed randomization simulations using our input data. PACo already includes an element of randomization via random allocation of hosts to parasites, since PACo specifically tests whether the parasite phylogeny depends on the host phylogeny. Because of this I randomizaed individual bacterial associations with each host. I utilized the randomize_phyloseq function from the package metagMisc (github.com/vmikk/metagMisc) to randomize the bacterial associations of our *Chaetoceros* data to maintain sample species richness and species occurrence frequency using the options “model = richness” and “model = frequency” options, respectively. I performed 1 000 randomizations for each of these models to determine if our observed result fell within the distribution of results from our randomizations. PACo was run on each of the 1 000 randomizations using the same settings and options as the analysis that produced our observed results. If our observed result was within the distribution of simulated results, then I could surmise that our observed results from the empirical data were no different than those pulled

from a random distribution. If our observed result was outside this distribution, then it would indicate that the observed result from our empirical data could not be derived from a random distribution.

I then generated simulated host-parasite phylogenies and association matrices to determine how varying levels of cophylogeny would affect the results of PACo. Simulated host-parasite data was generated via the R package treeducken (Dismukes & Heath, 2021). I generated six sets of 1 000 simulated host-parasite data sets. Of the six simulated datasets, three had different cospeciation rates but the same host shift rates and three had different cospeciation rates and host shift rates. The following options were used with the following parameters in all analyses: hbr = 1, hdr = 0.05, sbr = 2, sdr = 0.1. Three cospeciation rates (cosp_rate) were used for each set of analyses: 0.1, 0.5, and 1.0. In the simulations with different host shift rates (host_exp_rate), the following values were used in the simulations with the matching cosp_rate value: 0.1, 0.5, and 1.0. So, I end up with the following values for cospeciation rate/host shift rate for each simulation: 0.1/0, 0.5/0, 1.0/0, 0.1/0.1, 0.5/0.5, and 1.0/1.0. PACo analyses were performed on each of the simulated data sets using the same settings and options as the analysis that produced our observed results.

To determine whether there were significant statistical differences between each simulation and randomization, Student's t-Test, ANOVA, and one-sample Z-tests were performed on our simulated data sets. All of these were performed in the R programming environment using the stats (for Student's t-Test and ANOVA) and BSDA (Z-test) packages. Student's t-Test was used to test whether the means of our richness and frequency randomizations were significantly different. ANOVA was used to test if our treeducken cospeciation simulations (cosp= 0.1, 0.5, 1) had significantly different means. Finally, Z-tests were used to determine if the mean of each simulated distribution is significantly different from our observed PACo result.

Finally, I used the parafit function, of the R package ape, to test the hypothesis of coevolution between our host and parasite phylogenies. The null hypothesis (H0) of the global test statistic (ParaFitGlobal) is that the evolution of the two groups, as inferred from the two phylogenetic trees and the set of host-symbiote association links, has been independent. (Legendre *et al.*, 2002; Paradis & Schliep, 2018). If the H0 is rejected, then evolution of the two groups is inferred to be dependent on one another (e.g., coevolution has occurred). Unlike PACo, there is no option for phylogeny standardization in Parafit and thus the global test statistic of parafit scales very high as the number of associations increases. Additionally, Parafit allows for the testing of individual links and reports support values for them, unlike PACo. Parafit was run using the following options: nperm = 1 000, test.links = TRUE, correction = “lingoes”, and seed = 42.

3.4 Results

3.4.1 Data processing & summaries

I initially recovered a total of 26 806 16S sequences from 261 of the 274 diatom transcriptomes. Removing sequences < 250bp in length reduced the number of sequences to 17 727. Sequence clustering using CD-hit further reduced our dataset into a total of 8 698 sequence clusters at the 97% similarity threshold. After removing eukaryotic and indeterminate sequences, 5 864 of the bacterial OTUs were able to be placed using EPA-NG & Gappa, none of the archaeal OTUs were able to be placed and were thus discarded from further analysis. Next, after discarding OTUs that were only placed to the level of domain, had likelihood weight ratio values < 0.5, or had been assigned to taxa in our list of known bacterial contaminants, I were left with a total of 4 033 bacterial OTUs.

Of the bacterial OTUs, the majority (3 593; Fig. 2) were found in transcriptomes from (Alverson *et al.*, 2023), suggesting that the two sample sources handled transcriptomes differently, either before or after sequencing. In reference to salinity, the majority were recovered

from the transcriptomes of marine diatoms (3 087; Fig. 2). Finally, the majority (2293; Fig. 2) of OTUs were recovered from pennate diatom (Bacillariophyceae) transcriptomes. The majority of

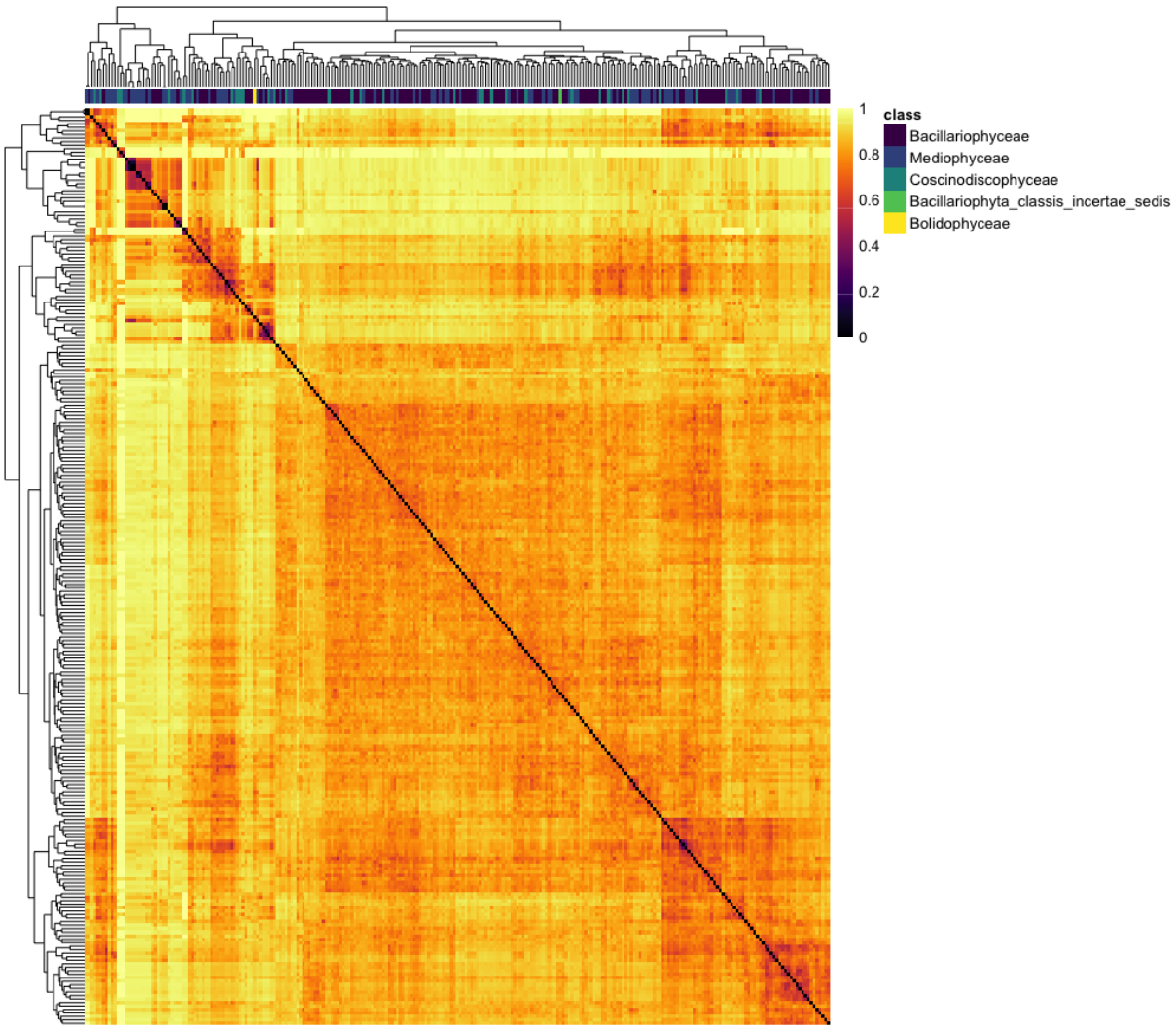

Figure 3.5. Heat plot of pairwise unweighted UniFrac dissimilarity values by transcriptome for Bacterial OTUs. Transcriptome IDs are not shown to preserve readability. Hotter colors (Higher values) indicate higher dissimilarity values between communities of each pairwise comparison. Tips of clustering dendrograms are colored according to the diatom class each transcriptome originates from.

bacterial sequences from both sources were found to be unique to their host (3 252). Bacterial sequences that were recovered from multiple transcriptomes were primarily present in only a ≤

10 transcriptomes (735). Only a fraction of the observed bacterial sequences were found in > 10 transcriptomes (37) and > 20 transcriptomes (9).

3.4.2 Phylogenetic placement and taxonomic identity of 16S query sequences

While the majority of sequences (72%) were able to be placed at the level of class or lower, most of those sequences were not identifiable at the species level (Fig. 1). The vast majority of bacterial sequences were assigned to the phyla Proteobacteria (1 712) and Bacteroidota (1 004), followed by the phyla Planctomycetota, Actinobacteriota, Verrucomicrobiota, and Spirochaetota (Fig. 3). All of these phyla are known from previous studies to associate with diatoms. [Sentence about number of bacteria from aquatic environments]. Of the bacterial OTUs, the majority were The remaining bacterial OTUs belonged to a variety of both described and undescribed bacterial phyla and, other than the phylum Cyanobacteria (Janson, 2002), are not known to associate with diatoms, (Fig. 3).

3.3 *Phylogenetic Diversity and Phylogenetic signal analyses*

To determine if there was a significant lab effect between sample sources, I performed an ANOVA to determine if mean OTU abundance per sample was significantly different between sample sources. Our ANOVA found a significant difference in mean OTU abundance between levels in source ($F = 172.2$, $P = <0.05$). Because of this significant source effect, I continued our analyses only on the Alverson *et al.*, (2023) subset of samples. I then analyzed both subsets using NMDS, phylogenetic signal analyses, & other statistical analyses where noted. Lastly, I fit a LM of our OTU abundance data against the total number of years those diatom strains were in culture. The LM of the Alverson *et al.*, (2023) data set found no significant correlations between OTU abundance and years in culture.

Table 3.2. OTU abundances between sample sources.

	Alverson et al. (2023)	MMETSP
Total OTU abundance	5863	535
Average OTUs per sample	31.9	6.95
Std. dev. of OTUs per sample	16.1	6.72

Using Tukey's HSD, I found that marine diatoms harbored greater bacterial diversity compared to freshwater diatoms, based on the phylogenetic alpha-diversity metrics FI, CAC, and CAC_s. Our phylogenetic beta-diversity found that each transcriptome harbored bacterial assemblages that were mostly unique to individual transcriptomes (Fig. 5), with a range of 0% – 100% dissimilarity and a mean of 85% dissimilarity in beta-diversity. These large ranges of dissimilarities are primarily the result of comparisons between transcriptomes which both have a single OTU that is the same (or very closely related), which results in dissimilarities of zero or near-zero. If I ignore those comparisons, our range becomes 20–100%. The majority of transcriptomes displayed high levels of dissimilarity from each other (Fig. 6). The diatom transcriptomes with the greatest similarity to each other are actually quite distantly related, which indicates they share the same, high occupancy OTUs. Our T-test on the observed vs. mean simulated uwUF distances found a significant difference in means between the observed and simulated uwUF distances ($T = 80.1$, $DF = 33929$, $P = <0.05$), meaning that our observed distances could not come from a random distribution, i.e., this pattern did not arise by chance. Beta-diversity was found to be significantly different between all salinity groupings except for the mean beta-diversity of brackish-fresh – brackish-brackish. Additionally, I examined whether time in culture (years) affected beta-diversity using linear regression. I found that there was no significant effect on beta-diversity from time in culture.

Our NMDS analysis of the Alverson *et al.*, (2023) data set found the best solution with a stress value of 0.22 after 20 permutations. Beta-diversities for freshwater and marine

transcriptomes formed distinct, but slightly overlapping, groups in the NMDS analysis, with brackish transcriptome beta-diversities interspersed in the freshwater and marine groupings (Fig. 6). To quantify the observed differences in the NMDS ordination, I performed a PERMANOVA analysis of the Alverson *et al.*, (2023) transcriptomes , which found significant differences in beta-diversity between transcriptomes of different levels of salinity.

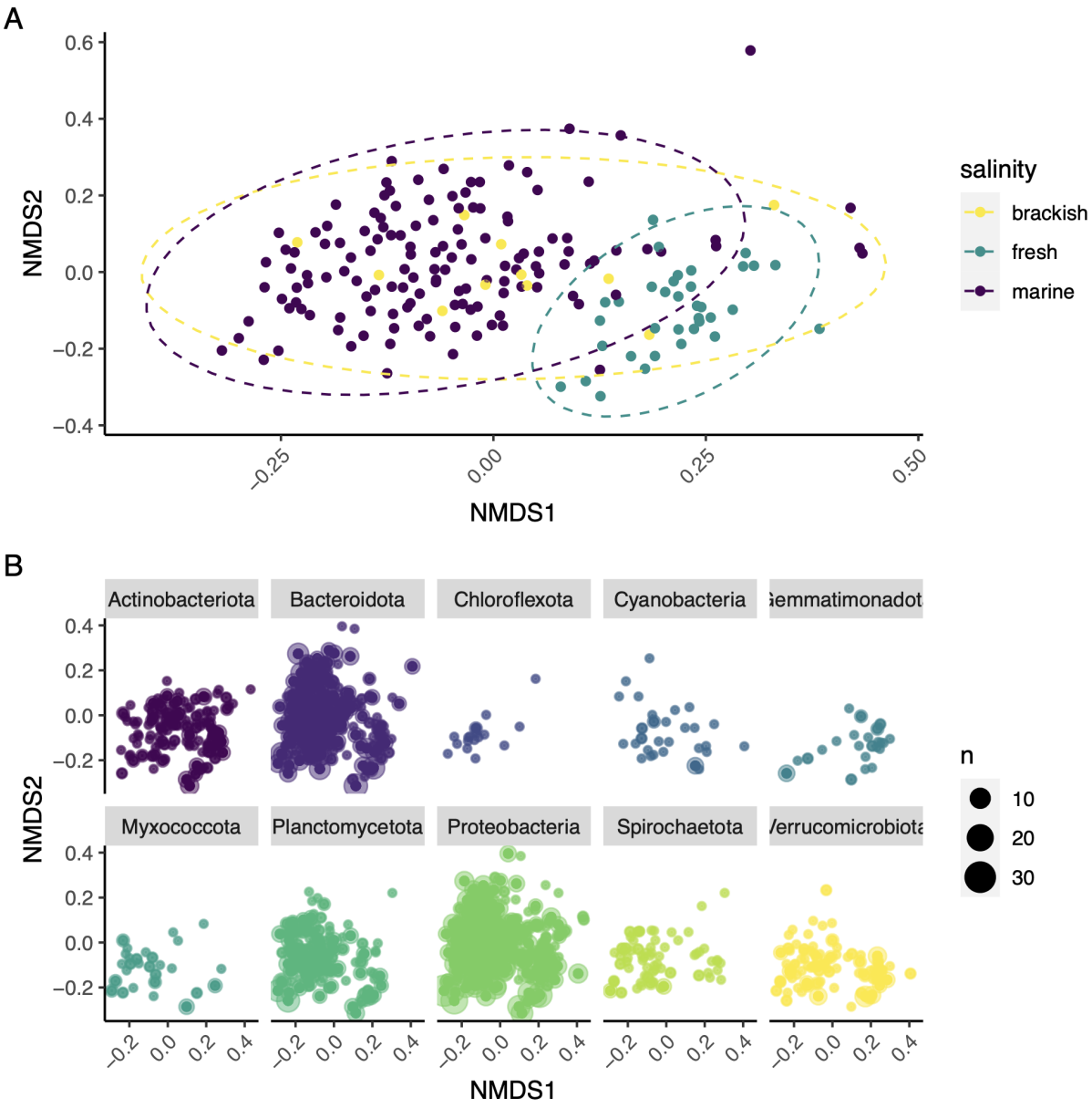

Figure 3.6. NMDS analysis of unweighted UniFrac distances between (Alverson *et al.*, 2023)

transcriptomes, the best solution was found with a stress value of 0.24 after 20 permutations. A) Bacterial beta-diversity between transcriptomes, colored by salinity with individual points depicting pairwise beta-diversity comparisons between transcriptomes. B) OTUs of the ten most abundant phyla, colored and faceted by phylum. Ellipses generated using stat = “norm.”

To determine if any metadata variables for our transcriptomes were homoplasies or synapomorphies, I performed phylogenetic signal analysis using the transcriptome metadata variables in comparison to the bacterial OTU phylogeny for Alverson *et al.*, (2023). I found that all variables were homoplasious in nature ($K < 0.03$), and only four out of 11 variables were found to be statistically significant. I also performed a phylogenetic signal analysis on the subset of bacterial OTUs associated with the diatom genus *Chaetoceros,* since it was the most diversely sampled diatom genus in our data set. The *Chaetoceros* subset analysis was conducted on the metadata variables of transcriptome ID and diatom species, with *K* values of 0.2 and 0.12, respectively, and both were found significant (P = 0.001).

3.4.4 Cophylogenetic concordance analysis

The genus *Chaetoceros* was found to have 133 bacterial sequences associated with it, from 37 families and 48 genera with an average abundance of 12 OTUs per sample . The taxonomic composition of *Chaetoceros* bacterial OTUs used in this analysis consisted primarily of taxa from the phyla Proteobacteria and Bacteroidetes (Fig. 4). The taxonomic composition of bacterial consortia for *Chaetoceros* spp. is similar to the overall taxonomic composition of the full dataset, but the average OTU abundance is less than the average of 16 OTUs per sample. PACo analysis of *Chaetoceros* found significant concordance between the bacterial and diatom phylogenies (P=0, m^2_{xy}=0.90, Fig. 3.7), suggesting that interactions between *Chaetoceros* and their bacterial symbionts demonstrated significantly more cophylogenetic signal than expected by chance alone. Another widely used global-fit test for cophylogeny, parafit, was used to assess cophylogeny in the *Chaetoceros* data set. Parafit differs from PACo in that it uses the 4th-corner method to examine non-independence between phylogenies based on their interactions, compared to PACo which uses procrustean superimposition. The results of our

Parafit analysis of *Chaetoceros* and their bacterial consortia rejected the null hypothesis of the test and found that both phylogenies were dependent on one another (ParaFitGlobal = 244.6131 , p-value = 0.001), indicating coevolution between *Chaetoceros* and its bacterial consortia. Based on this, I can say that our parafit results demonstrate significant evidence of coevolution. Although both PACo and parafit found significant evidence of coevolution between *Chaetoceros* and its bacterial consortia, these analyses disagreed on which individual links contributed the most to these concordances.

Our PACo frequency and richness randomization tests found that our observed m^2_{xy} fell outside of the distribution of m^2_{xy} values for the frequency and richness randomizations. Student's t-Test found that the true difference in means between each randomized data set is not equal to zero (t = 19.1, df = 1968.6, p-value < 2.2e-16). Our one-sample Z-test found that the true mean for both our frequency (z = 409.9, p-value < 2.2e-16) and richness (z = 492.27, p-value < 2.2e-16) randomizations were not equal to our observed value of m^2_{xy} = 0.9. The mean m^2_{xy} for frequency and richness were 0.97 and 0.98, respectively, and mean p-values for both randomizations were > 0.05. The m^2_{xy} values for frequency and richness, ranged from 0.95 to 0.98 and 0.96 to 0.99, respectively. These results indicate that our observed result does not originate from a random distribution.

To test whether our observed results for PACo were similar to the simulated results under six different sets of cospeciation and host-shift rates, I compared our observed results with these simulated results using ANOVA and Z-tests. Our treeducken simulations found that our observed value fell within the distribution of m^2_{xy} of each of the six simulations. Specifically, I found that the distribution of m^2_{xy} values for the cospeciation rate = 0.1 simulation were much higher, which may indicate that our empirical data originated from a similar distribution. Our ANOVA analysis found that the means of each simulation were significantly different from one another (DF = 2, F-value = 844.8, p-value < 2e-16). Our one-sample Z-tests found that the true mean for our low (z = -14.974, p-value < 2.2e-1), medium (z = -39.526, p-value < 2.2e-16), and

high (z = -65.601, p-value < 2.2e-16) cospeciation rate simulations were not equal to our observed value of m^2_{xy} = 0.9. The true m^2_{xy} means for our low, medium, and high cospeciation rate simulations were 0.83, 0.67, and 0.51, respectively. The range of m^2_{xy} values observed for our low, medium, and high cospeciation rate simulations were 0.05 - 1, 0.01 - 0.99, and 0.09 - 0.98, respectively.

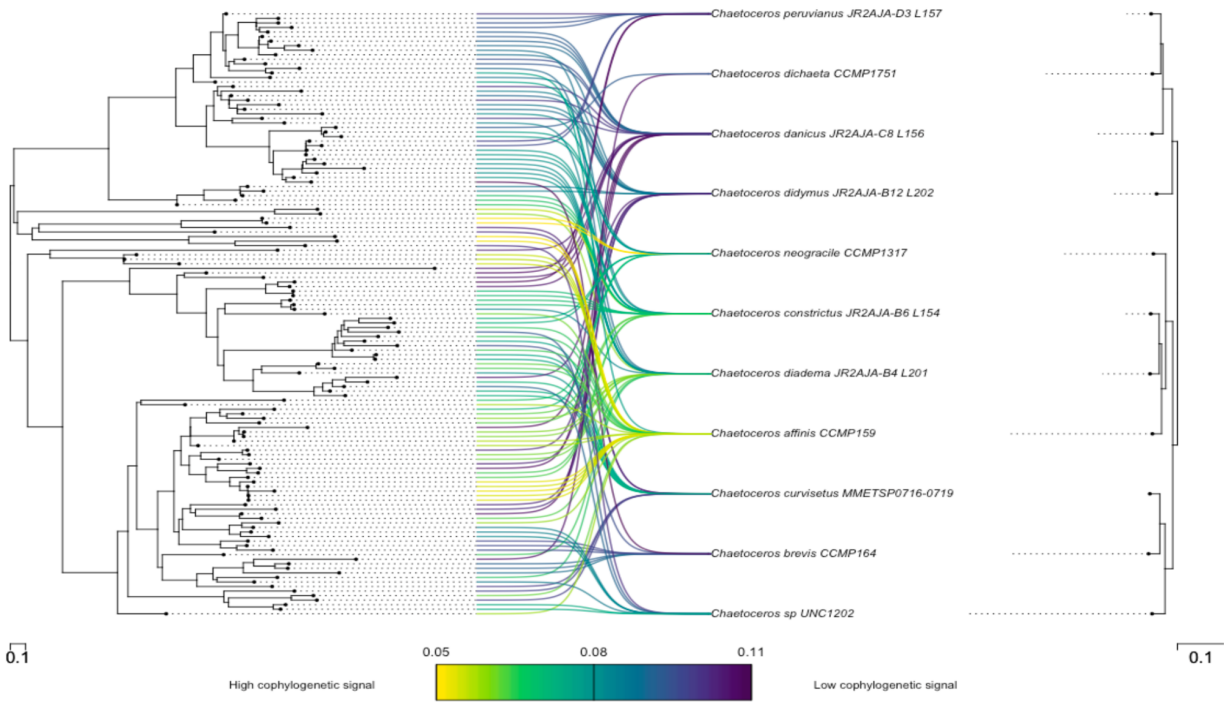

Figure 3.7. Cophylogentic concordance tanglegram of bacterial OTUs for the diatom genus *Chaetoceros.* Individual links in the tanglegram are colored according to their level of cophylogenetic signal, will lower values (yellow) indicating more phylogenetic concordance (higher cophylogenetic signal) and greater values indicating less phylogenetic concordance (lower cophylogenetic signal).

3.5 Discussion

Diatom–bacteria interactions are responsible not only for maintaining the health of diatoms via interactions that supply essential nutrients, but also for promoting the assemblage of beneficial bacteria, or increasing growth rates or cell death in diatom blooms. Understanding how these interactions are formed, and maintained, is key to understanding these interactions and how they operate in nature. Our results demonstrate that non-traditional methods of

obtaining bacterial 16S sequences from transcriptome sequencing can be a useful method for surveying microbiome data from existing transcriptomic datasets. I have also found that diatom cultures maintain similar bacterial communities to those found in previous studies, as well as previously unknown bacterial associations. I also provide additional evidence that some bacteria-diatom associations are preserved over time via cophylogenetic concordance analysis.

Previous studies on bacterial consortia of diatoms have found a group of core bacterial phyla that associate with them. In a study of bacterial community assembly in synthetic phycospheres, several OTUs were recovered that share taxonomy with those that I identified (Fu *et al.*, 2020). Those OTUs include the genera A*lgoriphagus, Celeribacter, Marinobacterium, Oceanospirillum, Roseivirga, Ruegeria, Terasakiella, Thalassospira,* and the families Alteromonadaceae & Flavobacteriaceae. The most abundant OTU found in association with diatom exudates by Fu *et al.*, (2020) was classified as *Ruegeria pomeroyi.* This bacteria was also found in our dataset, but only in transcriptomes belonging to unidentified raphid taxa in the genera *Sternimirus* & *Haslea,* respectively. Fu *et al.*, (2020) also show small numbers of Cyanobacterial and Archaeal OTUs that persisted in their reciprocal transfers. While Cyanobacterial OTUs were recovered from the dataset, none of the prospective archaeal OTUs in the dataset made it past the filtering stage.

There are potential concerns that must be addressed with data. The OTUs recovered from this dataset most likely present an incomplete picture of the existing bacterial diversity associated with diatoms due to the data scavenging methodology I employed. The concern that some of these OTUs may represent prokaryotic diversity that was accrued after isolation and replaced the original communities present in these cultures is also unlikely. While I cannot entirely rule out the accrual of new bacterial associates from the surrounding lab environment, based on our results and those of previous studies it seems unlikely. Our results demonstrate that bacterial communities in diatom cultures are highly diverse and culture-specific, as well as distinct between marine and freshwater diatoms. Our data also demonstrates that the vast

majority of our OTUs were assigned to bacterial groups that are known to associate with diatoms. If cross-culture contamination was occurring in these cultures, I would expect to see low diversity in bacterial communities and highly similar community composition between diatoms, as well as bacterial groups that aren't typically known to associate with diatoms. From the literature, the occurrence of specific strains of bacteria across strains of the same microalgae species, which would seem to indicate that these associations are intimate and likely to be skewed toward conservation by both parties. Only a single study has been conducted on the conservation of bacterial strains in microalgae cultures. Across multiple strains of the diatoms *Asterionellopsis glacialis* and *Nitzschia longissima,* bacterial communities displayed little change in composition at the genus level after one year under laboratory culture conditions (Behringer *et al.*, 2018). Finally, I chose to focus on the Alverson *et al.*, (2023) dataset because the MMETSP dataset either had more aggressive depletion of rRNA during extraction or library preparation, or through filtering of rRNA reads. The supplemental methods described in Keeling *et al.* (2014; supporting information, Text S1), did not describe RNA extraction methods, due to differences in procedures between contributing labs, or how data was filtered and prepared for downstream analysis.

The results of our UniFrac analysis indicate that the bacterial consortia of diatoms are highly strain-specific. Pairwise comparisons of uwUF distances found that bacterial communities were, on average, 85% dissimilar. Despite this high overall dissimilarity, our analyses did find that salinity groupings were significantly different from one another. This might initially be assumed to result from some differences between the diatom hosts' bacteria, but current research suggests that these may be a result of differences in freshwater and marine bacterial communities and not their diatom hosts. Rodríguez-Gijón *et al.*, (2022) found that genome size in archaea and bacteria is highly influenced by ecosystem type and trophic strategy. Based on the results of our NMDS analysis of UniFrac beta-diversity metric, and subsequent PERMANOVA analysis, the separation in our NMDS seems to be driven by differences in

bacterial assemblages between marine and freshwater diatoms (Fig,. 6). However, this does not mean that the diatom bacterial communities are unaffected by their hosts, as our UniFrac results demonstrate that within group beta-diversity for marine and freshwater diatoms are highly dissimilar in composition.

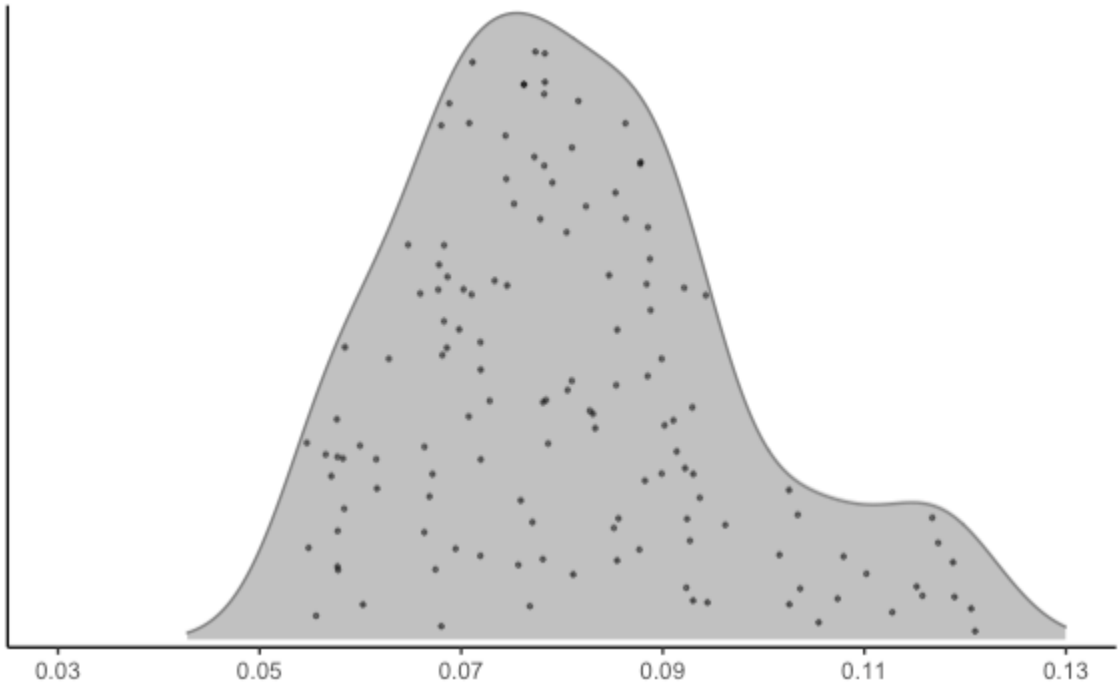

Figure 3.8. Density plot of squared residual values from our cophylogenetic concordance analyses for the *Chaetoceros*–bacteria subset. Individual points represent global best fit values (m^2_{xy}) for individual links between *Chaetoceros* species and their bacteria. Values closer to zero indicate a greater contribution to cophylogenetic signal.

Viewing these results through the lens of community assembly the possible origin of this high diversity and dissimilarity in diatom microbial communities. Historically, community assembly in phycosphere communities was thought to occur via the 'lottery hypothesis,' in that bacteria are competitively equivalent and whichever is able to get to a niche first (in this case a diatom) will persist and out-compete late-comers (Munday, 2004).The application of the lottery hypothesis has been confounded by further research in which it was discovered that chemical signaling occurs between diatoms and their bacterial partners (Amin *et al.*, 2015). Additionally, motile bacteria are able to encounter far more algal cells than their non-motile counterparts

(Seymour *et al.*, 2017), and indeed some possess chemotaxis towards algal cells when present (Miller *et al.*, 2004). Gralka *et al.* (2020) have thoroughly reviewed how trophic interactions drive the assembly and diversity of bacterial communities, mainly via primary consumer bacteria breaking down complex resources into less complex resources that are consumed by secondary consumers. Shibl *et al.* (2020) have found that secondary metabolites (rosmarinic and azelaic acid) that are released by the diatom *Asterionellopsis glacialis* are capable of modulating community assembly by promoting the growth and attachment of beneficial bacterial partners while suppressing opportunistic bacteria. Additionally, selective filtering related to the host microhabitat can greatly restrain lottery assembly in the diatom *Cylindrotheca closterium (Stock et al., 2022)*. Based on this research, I hypothesize that the high diversity and dissimilarity of diatom microbiomes in our research is due to a mixture of microhabitat modulation through the host diatoms as well as trophic interactions within the bacterial communities themselves.

Chaetoceros makes for an interesting candidate for cophylogeny analysis as it is one of the most abundant members of the phytoplankton and the most species-rich genus of marine planktonic diatoms, with over 400 species described (Rines & Hargraves, 1988). In particular, *Chaetoceros* was chosen as the focus of our cophylogeny analyses due to it having the most thorough sampling of any diatom genus in our dataset.. The results of our analyses indicate that a subset of bacteria that associate with the diatom genus *Chaetoceros* form longstanding relationships with these diatoms (Fig. 3.7). In our analysis of *Chaetoceros,* I found that specific *Chaetoceros* spp. and their bacterial associations are driving this strong cophylogenetic signal, although these differ between PACo & Parafit as a result of differing methodologies between the two analyses. The results of our randomized and simulated PACo analyses indicate that our result are not consistent with from a random distribution and that our results are most likely the results of low cospeciation rates, although the similarity with low cospeciation rate simulations may be due to incomplete taxon sampling and a stronger cophylogenetic signal could result with

more complete sampling. These results are supported by previous research in which diatom bacterial communities were found to be stable over long periods of time after sampling and culturing (Mönnich *et al.*, 2020; Barreto Filho *et al.*, 2021).

Significant cophylogeny results can be interpreted under the lens of trophically-driven community assembly. PACo analysis revealed that a third of OTUs (45 links) associated with *C. affinis, C. diadema, C. neogracile,* and *C. constrictus* are responsible for the strong cophylogenetic signal between our bacterial and diatom phylogenies (Figs. 3.7 and 3.8). These stronger links may indicate that these particular bacteria are utilizing primary metabolites from their algal partners, whereas the bacteria displaying weaker links may represent bacteria that are utilizing secondary metabolites. This primary and secondary metabolite tracking by bacteria could influence coevolution in that primary metabolite consuming bacteria would potentially coevolve to maintain these associations with their primary source of metabolites, whereas secondary metabolite consumers would only be tracking the evolution of the primary metabolite consuming bacteria. Gralka *et al.* (2020) found that a subset of primary metabolite processing bacteria directed the assembly of their community through the release of secondary metabolites and antibiotics they produce. This would also explain the high diversity seen in some of our cultures in that the primary consumer bacteria of the phycosphere would enable secondary consumer bacteria to persist in these xenic diatom cultures. Further research will be necessary to determine what the core, functional genes are in these interactions and what the trophic networks of the phycosphere are like and whether they are correlated with those bacteria driving these significant cophylogeny results.

One concern is that the lack of phylogenetic signal in our bacterial community contradicts the results of our PACo and uwUF analyses. This is not necessarily the case, as the “traits” defined in our phylogenetic signal analyses are subjective and broad, e.g. habitat of diatom and salinity level. If I were able to obtain genomes for all of these bacterial taxa and resolve metabolic pathways, I would perhaps find some phylogenetic signal for these better

resolved and finer scale traits. The PACo and uwUF analyses are more objective than the phylogenetic signal analyses in that they are based on objective data and traits (e.g., host-symbiote associations and genetic data).

In conclusion, our results indicate that marine and freshwater diatoms have bacterial community compositions that are significantly distinct from each other and that individual diatoms harbor highly diverse and unique bacterial communities. Cophylogeny analysis indicated that the genus *Chaetoceros* has some cophylogenetic concordance with the bacterial communities associated with its species. Through the addition of a greater range of diatom taxa from across the globe and multiple habitats, this research further expands our knowledge of diatom bacterial communities. An improved understanding of bacterial diversity across diatoms may allow future researchers to observe which members of diatom-bacterial consortia are members of the core microbiome and which are opportunistic secondary consumers. A better empirical understanding of the preferred bacterial partners for diatoms will allow for better approximations of these relationships *in situ* and aid industrial microalgal cultures in increasing yields while protecting them from infection. Future research efforts should focus on developing a detailed understanding of the mechanisms influencing the assembly of both the primary and secondary bacterial communities that make up the phycosphere.

3.6 Data Availability

The full data analysis workflow, including rationale, R code, and commands used in analyses, and raw analysis output files can be found in the Quarto document ‘project_workflow.qmd’ in Zenodo repository for this manuscript (10.5281/zenodo.3929776).

3.7 Acknowledgements

This research is supported by the Arkansas High Performance Computing Center which is funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission. The authors would also like to thank Dr. Matthew Hutchinson for his helpful advice on interpretation of results from the R package PACo and parafit analyses.

3.8 References

Abby SS, Touchon M, De Jode A, Grimsley N, Piganeau G. **2014**. Bacteria in *Ostreococcus tauri* cultures - friends, foes or hitchhikers? *Frontiers in microbiology* **5**: 505.

Allaire JJ, Ushey K, Tang Y, Eddelbuettel D. **2017**. reticulate: R Interface to Python.

Alverson A, Nakov T, Roberts W, Ruck E, Gargas C. **2023**. A global phylogenomic analysis of diatoms.

Amin SA, Green DH, Hart MC, Kupper FC, Sunda WG, Carrano CJ. **2009**. Photolysis of iron-siderophore chelates promotes bacterial-algal mutualism. *Proceedings of the National Academy of Sciences* **106**: 17071–17076.

Amin SA, Hmelo LR, van Tol HM, Durham BP, Carlson LT, Heal KR, Morales RL, Berthiaume CT, Parker MS, Djunaedi B, *et al.* 2015. Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* **522**: 98–101.

Amin SA, Parker MS, Armbrust EV. **2012**. Interactions between Diatoms and Bacteria. *Microbiology and molecular biology reviews: MMBR* **76**: 667–684.

Anderson MJ. **2001**. A new method for non-parametric multivariate analysis of variance. *Austral ecology* **26**: 32–46.

Azam F, Malfatti F. **2007**. Microbial structuring of marine ecosystems. *Nature reviews. Microbiology* **5**: 782–791.

Baker LJ, Kemp PF. **2014**. Exploring bacteria–diatom associations using single-cell whole genome amplification. *Aquatic Microbial Ecology* **72**: 73–88.

Balbuena JA, Míguez-Lozano R, Blasco-Costa I. **2013**. PACo: A novel procrustes application to cophylogenetic analysis. *PloS one* **8**: e61048.

Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, Stamatakis A. **2019**. EPA-ng: Massively parallel evolutionary placement of genetic sequences. *Systematic biology* **68**: 365–369.

Barreto Filho MM, Walker M, Ashworth MP, Morris JJ. **2021**. Structure and Long-Term Stability of the Microbiome in Diverse Diatom Cultures. *Microbiology spectrum*: e0026921.

Behringer G, Ochsenkühn MA, Fei C, Fanning J, Koester JA, Amin SA. **2018**. Bacterial Communities of Diatoms Display Strong Conservation Across Strains and Time. *Frontiers in microbiology* **9**: 659.

Bell W, Mitchell R. **1972**. Chemotactic and growth responses of marine bacteria to algal extracellular products. *The Biological bulletin* **143**: 265–277.

Biddanda B, Benner R. **1997**. Carbon, nitrogen, and carbohydrate fluxes during the production of particulate and dissolved organic matter by marine phytoplankton. *Limnology and oceanography* **42**: 506–518.

Biegala IC, Kennaway G, Alverca E, Lennon J-F, Vaulot D, Simon N. **2002**. Identification of

bacteria associated with Dinoflagellates (Dinophyceae) *Alexandrium* spp. using tyramide signal amplification -- fluorescent In situ hybridization and confocal microscopy. *Journal of phycology* **38**: 404–411.

Bolger AM, Lohse M, Usadel B. **2014**. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Boschetti C, Carr A, Crisp A, Eyres I, Wang-Koh Y, Lubzens E, Barraclough TG, Micklem G, Tunnacliffe A. **2012**. Biochemical diversification through foreign gene expression in bdelloid rotifers. *PLoS genetics* **8**: e1003035.

Bushnell B, Rood J, Singer E. **2017**. BBMerge – Accurate paired shotgun read merging via overlap. *PloS one* **12**: e0185056.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. **2009**. BLAST+: Architecture and applications. *BMC bioinformatics* **10**: 421.

Cho BC, Azam F. **1988**. Major role of bacteria in biogeochemical fluxes in the ocean's interior. *Nature* **332**: 441–443.

Cole JJ. **1982**. Interactions between Bacteria and algae in aquatic ecosystems. *Annual review of ecology and systematics* **13**: 291–314.

Croft MT, Lawrence AD, Raux-Deery E, Warren MJ, Smith AG. **2005**. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. *Nature* **438**: 90–93.

Czech L, Stamatakis A. **2019**. Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples. *bioRxiv*: 346353.

Dismukes W, Heath TA. **2021**. treeducken: An R package for simulating cophylogenetic systems. *Methods in ecology and evolution / British Ecological Society*.

Durham BP, Dearth SP, Sharma S, Amin SA, Smith CB, Campagna SR, Armbrust EV, Moran MA. **2017**. Recognition cascade and metabolite transfer in a marine bacteria-phytoplankton model system. *Environmental microbiology* **19**: 3500–3513.

Durham BP, Sharma S, Luo H, Smith CB, Amin SA, Bender SJ, Dearth SP, Van Mooy BAS, Campagna SR, Kujawinski EB, *et al.* 2015. Cryptic carbon and sulfur cycling between surface ocean plankton. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 453–457.

Eigemann F, Hilt S, Salka I, Grossart H-P. **2013**. Bacterial community composition associated with freshwater algae: Species specificity vs. dependency on environmental conditions and source community. *FEMS microbiology ecology* **83**: 650–663.

Faith DP. **1992**. Conservation evaluation and phylogenetic diversity. *Biological conservation* **61**: 1–10.

Falkowski PG, Fenchel T, Delong EF. **2008**. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**: 1034–1039.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. **1998**. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**: 237–240.

Frischkorn KR, Rouco M, Van Mooy BAS, Dyhrman ST. **2017**. Epibionts dominate metabolic functional potential of *Trichodesmium* colonies from the oligotrophic ocean. *The ISME journal* **11**: 2090–2101.

Fu L, Niu B, Zhu Z, Wu S, Li W. **2012**. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150–3152.

Fu H, Uchimiya M, Gore J, Moran MA. **2020**. Ecological drivers of bacterial community assembly in synthetic phycospheres. *Proceedings of the National Academy of Sciences of the United States of America* **117**: 3656–3662.

Gargas CB, Roberts WR, Alverson AJ. **2020**. Genome sequences of bacteria associated with the diatom *Cyclotella cryptica* strain CCMP332. *Microbiology Resource Announcements* **9**: e01030–20.

Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. **2016**. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut pathogens* **8**: 24.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**: 644–652.

Gralka M, Szabo R, Stocker R, Cordero OX. **2020**. Trophic Interactions and the Drivers of Microbial Community Assembly. *Current biology: CB* **30**: R1176–R1188.

Green DH, Llewellyn LE, Negri AP, Blackburn SI, Bolch CJS. **2004**. Phylogenetic and functional diversity of the cultivable bacterial community associated with the paralytic shellfish poisoning dinoflagellate *Gymnodinium catenatum*. *FEMS microbiology ecology* **47**: 345–357.

Hasegawa Y, Martin JL, Giewat MW, Rooney-Varga JN. **2007**. Microbial community diversity in the phycosphere of natural populations of the toxic alga, *Alexandrium fundyense*. *Environmental microbiology* **9**: 3108–3121.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, *et al.* 2016. A new view of the tree of life. *Nature Microbiology* **1**.

Hutchinson MC, Cagua EF, Balbuena JA, Stouffer DB, Poisot T. **2017**. paco: Implementing Procrustean Approach to Cophylogeny in R (R Fitzjohn, Ed.). *Methods in ecology and evolution / British Ecological Society* **8**: 932–940.

James DA. **2012**. DBI: R Database Interface (2009). R package version 0.2. 5.

Janson S. **2002**. Cyanobacteria in Symbiosis with Diatoms. In: Rai AN, Bergman B, Rasmussen U, eds. Cyanobacteria in Symbiosis. Springer, Dordrecht, 1–10.

Janzen DH. **1980**. When is it coevolution? *Evolution; international journal of organic evolution* **34**: 611–612.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. **2017**. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* **14**: 587–589.

Kazamia E, Helliwell KE, Purton S, Smith AG. **2016**. How mutualisms arise in phytoplankton communities: Building eco-evolutionary principles for aquatic microbes. *Ecology Letters* **19**: 810–822.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, *et al.* 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**: e1001889.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. **2010**. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**: 1463–1464.

Kulakov LA, McAlister MB, Ogden KL, Larkin MJ, O'Hanlon JF. **2002**. Analysis of bacteria contaminating ultrapure water in industrial systems. *Applied and environmental microbiology* **68**: 1548–1555.

Langmead B, Salzberg SL. **2012**. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**: 357–359.

Langmead B, Trapnell C, Pop M, Salzberg SL. **2009**. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**: R25.

Langmead B, Wilks C, Antonescu V, Charles R. **2019**. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**: 421–432.

Legendre P, Desdevises Y, Bazin E. **2002**. A statistical test for host-parasite coevolution. *Systematic biology* **51**: 217–234.

Li W, Godzik A. **2006**. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Lozupone C, Hamady M, Knight R. **2006**. UniFrac--An online tool for comparing microbial community diversity in a phylogenetic context. *BMC bioinformatics* **7**: 371.

Lozupone C, Knight R. **2005**. UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* **71**: 8228–8235.

Martin JH, Fitzwater SE. **1988**. Iron deficiency limits phytoplankton growth in the north-east Pacific subarctic. *Nature* **331**: 341–343.

McMurdie PJ, Holmes S. **2013**. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one* **8**: e61217.

Mendes R, Kruijt M, de Bruijn I, Dekkers E, van der Voort M, Schneider JHM, Piceno YM, DeSantis TZ, Andersen GL, Bakker PAHM, *et al.* 2011. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332**: 1097–1100.

Miller TR, Hnilicka K, Dziedzic A, Desplats P, Belas R. **2004**. Chemotaxis of Silicibacter sp. strain TM1040 toward dinoflagellate products. *Applied and environmental microbiology* **70**: 4692–4701.

Minh BQ, Schmidt H, Chernomor O, Schrempf D, Woodhams M, von Haeseler A, Lanfear

R. **2019**. *IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era*. Evolutionary Biology.

Mönnich J, Tebben J, Bergemann J, Case R, Wohlrab S, Harder T. **2020**. Niche-based assembly of bacterial consortia on the diatom *Thalassiosira rotula* is stable and reproducible. *The ISME journal*.

Munday PL. **2004**. Competitive coexistence of coral-dwelling fishes: The lottery hypothesis revisited. *Ecology* **85**: 623–628.

Nawrocki E. **2009**. Structural RNA homology search and alignment using covariance models.

Paradis E, Claude J, Strimmer K. **2004**. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290.

Paradis E, Schliep K. **2018**. ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**: 526–528.

Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. **2020**. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature biotechnology*.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. **2018**. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology* **36**: 996–1004.

Rappé MS, Giovannoni SJ. **2003**. The uncultured microbial majority. *Annual review of microbiology* **57**: 369–394.

Rines JBE, Hargraves PE. **1988**. The Chaetoceros Ehrenberg (Bacillariophyceae) flora of Narragansett Bay, Rhode Island, U.S.A. *Bibliotheca Phycologica* **79**: 1–196.

Rodríguez-Gijón A, Nuy JK, Mehrshad M, Buck M, Schulz F, Woyke T, Garcia SL. **2022**. A Genomic Perspective Across Earth's Microbiomes Reveals That Genome Size in Archaea and Bacteria Is Linked to Ecosystem Type and Trophic Strategy. *Frontiers in microbiology* **12**: 4194.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. **2014**. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* **12**: 87.

Sapp M, Schwaderer AS, Wiltshire KH, Hoppe H-G, Gerdts G, Wichels A. **2007**. Species-Specific Bacterial Communities in the Phycosphere of Microalgae? *Microbial Ecology* **53**: 683–699.

Segev E, Wyche TP, Kim KH, Petersen J, Ellebrandt C, Vlamakis H, Barteneva N, Paulson JN, Chai L, Clardy J, *et al.* 2016. Dynamic metabolic exchange governs a marine algal-bacterial interaction. *eLife* **5**.

Seymour JR, Amin SA, Raina J-B, Stocker R. **2017**. Zooming in on the phycosphere: The ecological interface for phytoplankton-bacteria relationships. *Nature microbiology* **2**: 17065.

Shibl AA, Isaac A, Ochsenkühn MA, Cárdenas A, Fei C, Behringer G, Arnoux M, Drou N, Santos MP, Gunsalus KC, *et al.* 2020. Diatom Modulation of Microbial Consortia Through Use of Two Unique Secondary Metabolites. *bioRxiv*: 2020.06.11.144840.

Simon Garnier, Noam Ross, Bob Rudis, Marco Sciaini, Cédric Scherer. **2018**. *Package 'viridis'*.

Stock W, Willems A, Mangelinckx S, Vyverman W, Sabbe K. **2022**. Selection constrains lottery assembly in the microbiomes of closely related diatom species. *ISME Communications* **2**: 1–10.

Tang YZ, Koch F, Gobler CJ. **2010**. Most harmful algal bloom species are vitamin B1 and B12 auxotrophs. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 20756–20761.

Tsirogiannis C, Sandel B. **2016**. PhyloMeasures: A package for computing phylogenetic biodiversity measures and their statistical moments. *Ecography* **39**: 709–714.

Tsirogiannis C, Sandel B, Kalvisa A. **2014**. New Algorithms for Computing Phylogenetic Biodiversity. In: Algorithms in Bioinformatics. Springer Berlin Heidelberg, 187–203.

Vellend, Mark, William K. Cornwell, Karen Magnuson-Ford, and Arne Ø. Mooers. **2011**. Measuring phylogenetic biodiversity. In: Magurran AE, McGill BJ, eds. Biological Diversity: Frontiers in Measurement and Assessment. Oxford University Press, 194–207.

Webb CO. **2000**. Exploring the Phylogenetic Structure of Ecological Communities: An Example for Rain Forest Trees. *The American naturalist* **156**: 145–155.

Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. **2003**. Phylogenies and Community Ecology. *Annual review of*.

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, *et al.* 2019. Welcome to the Tidyverse. *Journal of Open Source Software* **4**: 1686.

Wickham H, Chang W, Others. **2008**. ggplot2: An implementation of the Grammar of Graphics. *R package version 0. 7, URL: http://CRAN. R-project. org/package= ggplot2* **3**.

Wickham H, Others. **2007**. Reshaping data with the reshape package. *Journal of statistical software* **21**: 1–20.

Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ. **2015**. Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science* **347**: 1257594–1257594.

Chapter 4 Conclusion

4.1 Summary of results

4.1.1 The Psammoneis japonica genome retains evidence of horizontal gene transfer from members of its bacterial consortia.

I sequenced the nuclear genome and transcriptome of *Psammoneis japonica*, a chain-forming, benthic pennate diatom. The nuclear genome of *P. japonica* is 91.4 Mbp in length, with 15,170 predicted genes making up 27% of the total genome, repetitive elements accounting for 33% of the genome, and other non-coding elements comprising the remaining 40% of the genome. Repetitive elements were found to have a positive relationship with genome size. The partial metagenome of *P. japonica* revealed a diverse microbial community of at least 25 associated bacterial taxa, including four near-complete genomes for novel species of Planctomycetota, α-proteobacteria, and Bacteroidota. The *P. japonica* genome contains genes and potential pseudogenes which were transferred from several cohabiting bacteria. A total of 17 genic and 40 intergenic HGT candidate proteins were found. Three intergenic ORFs were found to form a potential pseudogene in the intergenic regions of the *P. japonica* genome. Several of these HGT candidate proteins are located in regions with transposon densities higher than the average for the genic and intergenic regions of the *P. japonica* genome.

4.1.2 Diatom transcriptomes reveal evidence for co-evolutionary dynamics and high bacterial diversity

I assembled bacterial 16S sequences from 274 diatom transcriptomes to investigate bacterial diversity and community phylogenetics across a broad diversity of diatoms. I found a high degree of dissimilarity in phylogenetic beta-diversity between diatom bacterial communities, even in closely related species of diatoms. Ordination analysis of phylogenetic beta-diversity demonstrated distinct groupings of diatom microbiomes by salinity. PERMANOVA analysis confirmed that beta-diversity was significantly different between these groups. Our results support that diatom phycosphere communities are more similar within salinity levels,

while still maintaining high diversity within and across genera. I also investigated the cophylogenetic concordance of diatom-bacteria associations and analyzed these associations in relation to phylogeny, salinity, and other factors. Significant cophylogenetic concordance was found between diatoms of the genus Chaetoceros and their bacterial partners, suggesting that some diatom–bacteria relationships are maintained over evolutionary time scales.

4.2 Future work

Through this research, we've explored the bacterial metagenome of one diatom, *Psammoneis japonica,* and the associated bacterial 16S sequences of 273 other diatoms. I identified the associated metagenome of *P. japonica* and found strong evidence of horizontal gene transfer between its lineage and the lineages of its bacterial consortia. As this project was initially conceived to only investigate the genome of *P. japonica,* the recovered metagenome is most likely incomplete in nature. Based on this, metagenome-focused sequencing should be pursued for a more complete picture of the bacteria that associate with this diatom. It would also be of interest to obtain new *P. japonica* cultures to compare wild type metagenomes of *P. japonica* and the metagenome of existing cultures. This would also allow us to examine *P. japonica* and determine whether it maintains a stable core microbiome over time under the conditions of culturing (Behringer *et al.*, 2018; Mönnich *et al.*, 2020). Similarly for our large scale study, broad-scale sequencing effort should be made to examine the metagenomes across the diatom tree of life. This sequencing could initially focus on the genus *Chaetoceros* to determine if the positive cophylogenetic results from this dissertation can be replicated with better sampling of both *Chaetoceros* species and their associated bacteria. Additionally, the genus *Thalassiosira* would be another diatom clade to investigate in depth. Members of *Thalassiosira* are already known to have gone through several, independent freshwater–marine transitions and would serve as an ideal model system to investigate how metagenomes change during these transitions (Alverson *et al.*, 2007).

4.3 References

Alverson AJ, Jansen RK, Theriot EC. **2007**. Bridging the Rubicon: Phylogenetic analysis reveals repeated colonizations of marine and fresh waters by thalassiosiroid diatoms. *Molecular phylogenetics and evolution* **45**: 193–210.

Behringer G, Ochsenkühn MA, Fei C, Fanning J, Koester JA, Amin SA. **2018**. Bacterial Communities of Diatoms Display Strong Conservation Across Strains and Time. *Frontiers in microbiology* **9**: 659.

Mönnich J, Tebben J, Bergemann J, Case R, Wohlrab S, Harder T. **2020**. Niche-based assembly of bacterial consortia on the diatom *Thalassiosira rotula* is stable and reproducible. *The ISME journal*.