**The Dissertation Committee for Mengjie Yu Certifies that this is the approved version of the following dissertation:**

# Tempo and Mode of Diatom Plastid Genome Evolution

**Committee:**

__________________________________
Edward C. Theriot, Supervisor

__________________________________
Robert K. Jansen, Co-Supervisor

__________________________________
John W. La Claire

__________________________________
David L. Herrin

__________________________________
Robin R. Gutell

# Tempo and Mode of Diatom Plastid Genome Evolution

**by**

**Mengjie Yu**

**Dissertation**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Doctor of Philosophy**

**The University of Texas at Austin**

**December 2017**

# Dedication

This dissertation is dedicated to my parents, Xinyou Yu and Huiping Zhang, my husband Forest Pfeiffer, and my extended family for their endless support and love during this seven year journey.

## Acknowledgements

I am indebted to my supervisor, Edward Theriot, for all his guidance and support during my doctoral training. His hardworking, dedication to science, and willingness to share made a huge impact in my life! I am thankful to my co-supervisor, Robert Jansen, for providing guidance, enlightening discussion and funding support for my project. I am also thankful to my committee members John La Claire and David Herrin for sharing their knowledge through phycology and plant molecular biology class, which helped me lay the foundation for my research. I am grateful for the discussion with my committee member Robin Gutell on bioinformatics and alternative career choices in industry.

I would like to thank all the current and former Theriot lab members, especially Matt Ashworth for training me the basic lab skills and providing diatom strain samples for my project, Teofil Narkov for hands on instruction on phylogenetic analysis, Mariska Brady for helpful discussion. I would also like to thank all the wonderful people in Jansen lab, Tracey Ruhlman for her assistance in plastid DNA extraction, Jin Zhang and Mao-Lun Weng for their help and training in genome assembly and rate analysis, Seongjun Park and Erika Schwarz for their technical help and discussions.

Finally, I would like to acknowledge the President of King Abdulaziz University, Prof. Abdulrahman O.Alyoubi, for funding support, the Genome Sequencing and Analysis Facility (GSAF) at the University of Texas at Austin for performing Illumina sequencing, the Texas Advanced Computing Center (TACC) at the University of Texas at Austin for access to supercomputers. And all the collaborators I had during graduate school, their attitude towards scientific research kept inspiring me along this journey.

# Abstract

## Tempo and Mode of Diatom Plastid Genome Evolution



Mengjie Yu, PhD

The University of Texas at Austin, 2017

Supervisors: Edward C. Theriot & Robert K. Jansen



Diatoms are mostly photosynthetic eukaryotes within the heterokont lineage. Their plastids were derived through secondary endosymbiosis of a red alga. Despite years of phylogenetic research, relationships among major groups of diatoms still remain uncertain. Additional plastid genome (plastome) sequences can not only provide more insight into diatom plastid evolution, but also assess phylogenetic relationships among the major lineages of diatoms. In my dissertation, I have more than doubled the available plastome sequences. My work in the plastome evolution in Thalassiosirales, one of the more comprehensively studied orders in terms of both genetics and morphology, showed highly conserved gene content and gene order within this order. I also documented the first instance of the loss of photosynthetic genes *psaE*, *psaI* and *psaM* in *Rhizosolenia imbricata*. By extensively sampling the diatoms with critical phylogenetic positions, I presented the largest genome scale phylogeny yet published for diatoms based on 103 shared plastid-coding genes from 40 diatoms and *Triparma laevis* as the outgroup. The most recent diatom classification posits that there are three major clades of diatoms: Coscinodiscophyceae (informally radial centrics), Mediophyceae (bi- or multipolar centrics), and Bacillariophyceae (pennates). Phylogenetic analysis of plastome data recovered the radial centric *Leptocylindrus* as the sister group to the remaining diatoms and recovered the polar diatoms *Attheya* plus

*Biddulphia* in a clade sister to pennate diatoms. Statistical analysis comparing this optimal tree to trees constraining diatoms to the existing classification strongly rejected monophyly for the Coscinodiscophyceae and Mediophyceae. Extensive plastome rearrangements and variable gene content were observed among the 40 diatom species. *Astrosyne radiata*, recovered on the longest terminal branch, experienced extensive gene loss. The nucleotide substitution rates of plastid protein coding genes were estimated, and their patterns were compared across different gene categories. Relationships between substitution rates and plastome characteristics, such as indels, genome size, genome rearrangement, were examined. The analyses also revealed a strong positive correlation between sequence divergence and gene order change in diatom plastomes.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1: Introduction

Diatoms are mostly photosynthetic eukaryotes within the heterokont lineage. They are unicellular organisms with delicate siliceous walls, forming a monophyletic group within the heterokont algae. The plastid of diatoms was derived when a eukaryotic cell engulfed a red alga through secondary endosymbiosis. Variable plastid genome sizes and extensive genome rearrangements have been observed. However, little is known about plastid genome evolution within order- or family-level clades, and extensive plastid genome studies across the diatom phylogeny are still lacking. The research in this dissertation focused on two main areas of plastid genome evolution in diatoms. First, this dissertation addressed the mode of plastid genome evolution in diatoms. Gene content, genome size and genome rearrangement were examined within and across the diatom phylogeny, and the genome scale phylogeny was discussed. Second, this dissertation focused on the tempo of plastid genome evolution in diatoms. The pattern of mutation rate of plastid genes was examined, and correlations between genome rearrangement and mutation rates were tested.

In Chapter 2, extensive sampling was conducted within Thalassiosirales, one of the more comprehensively studied diatom orders in terms of both genetics and morphology. Seven complete diatom plastid genomes are reported here including four Thalassiosirales: *Thalassiosira weissflogii*, *Roundia cardiophora*, *Cyclotella* sp. WC03_2, *Cyclotella* sp. L04_2, and three additional non-Thalassiosirales species *Chaetoceros simplex*, *Cerataulina daemon*, and *Rhizosolenia imbricata*. The sizes of the seven genomes varied from 116,459 to 129,498 bp, and their genomes are compact and lack introns. We found the larger size of the plastid genomes of Thalassiosirales compared to other diatoms was due primarily to expansion of the inverted repeat. Gene content within Thalassiosirales was more conserved compared to other diatom lineages. Gene order within Thalassiosirales was found to be highly conserved except for the extensive genome rearrangement in *Thalassiosira oceanica*. *Cyclotella nana*, *Thalassiosira weissflogii* and *Roundia cardiophora* shared an identical gene order, which was inferred to be the ancestral order for the Thalassiosirales, differing from that of the other two *Cyclotella* species by a single inversion. A few gene loss patterns were also discovered. The genes *ilvB* and *ilvH* were missing in all six diatom plastid genomes except for *Cerataulina daemon*, suggesting an independent gain of these genes in this species. The *acpP1* gene was missing in all Thalassiosirales, suggesting that its loss may be a synapomorphy for the order and this gene may have been functionally transferred to the nucleus. Three genes involved in photosynthesis, *psaE*, *psaI*, *psaM*, are missing in *Rhizosolenia imbricata,* which represents the first documented instance of the loss of photosynthetic genes from diatom plastid genomes.

In Chapter 3, we expanded our taxon sampling across the major clades of diatom phylogeny. We reported another 18 diatom plastome sequences ranging in size from 119,120 to 201,816 bp. We found that *Plagiogramma staurophorum* had the largest plastome sequenced so far due to large inverted repeats and a 2,971 bp group II intron insertion in *petD* gene. We also found that the continuation of the pattern of *psaE, psaI* and *psaM* genes loss in *Rhizosolenia fallax*., the closely related species of *Rhizosolenia imbricate*. Based on 103 shared plastid-coding gene from 40 diatoms and *Triparma laevis* as the outgroup, we reported the largest genome scale phylogeny yet published for diatoms. From our phylogeny, *Leptocylindrus* was recovered as sister to the remaining diatoms and the clade of *Attheya* plus *Biddulphia* was recovered as sister to pennate diatoms, strongly rejecting monophyly for two of the three proposed classes of diatoms.

In Chapter 4, we explored the patterns of plastid genes mutation rates in 40 diatom species across the diatom phylogeny. We found most accelerated rates in the long branch bearing species *Astrosyne radiata* and *Proboscia* sp. Consistent with previous studies, dN and dS rate in genes integral to photosynthesis were much lower than other groups, while the replicative DNA helicase gene *dnaB* showed the highest dN and dS value. A significant positive correlation was observed between dN, dS and dN/dS and the number of indels. However, no obvious correlation was found between the substitution rates and plastid genome size. Significant correlation between pairwise mutation rates and genome rearrangement measured by inversion distance were detected, with the long branch species *Astrosyne radiata* showing the highest correlation score.

# Chapter 2: Conserved gene order and expanded inverted repeats characterize plastid genomes of Thalassiosirales

This Chapter is published in *PloS ONE* 9(9): e107854. Mengjie Yu is co-first and corresponding author.

## Introduction

Diatoms are unicellular organisms with delicate siliceous walls, forming a monophyletic group within the heterokont algae (Evans et al., 2004; Julius and Theriot, 2010; Round and Crawford, 1984; Theriot et al., 2011). Most diatoms are photosynthetic and are responsible for one quarter of global net primary production, and they are the main biological mediators of the silica cycle in the oceans (Nelson et al., 1995). The completion of nuclear and plastid genome sequences for three diatoms, *Cyclotella nana* Hustedt (Armbrust et al., 2004) (formerly *Thalassiosira pseudonana* Hasle & Heimdal (Alverson et al., 2011)), *Phaeodactylum tricornutum* Bohlin(Bowler et al., 2008), and *Thalassiosira oceanica* Hasle (Lommer et al., 2010), allowed the exploration of their evolutionary history in a genomic context. For example, one environmentally-driven gene transfer event has been reported in *T. oceanica*, where the *petF* gene encoding ferredoxin was transferred from the plastid to the nucleus (Lommer et al., 2010). Replacing the iron-sulfur protein ferredoxin by iron-free flavodoxin presumably contributed to the ecological success of *T. oceanica* in iron limited environments (Lommer et al., 2010).

Understanding possible adaptive events such as the transfer of *petF* requires a dense taxon sampling of the trait of interest over a well-resolved phylogeny. The Thalassiosirales Glezer & Makarova are the only diatom order with a moderately well-resolved phylogeny that has been used to formally examine the evolution of ecological, morphological and genetic traits, particularly with regard to adaptation across marine and freshwater environments (Alverson, 2007; Nakov et al., 2014).

Fifteen diatom plastid genomes have been sequenced so far (Brembu et al., 2013; Galachyants et al., 2012; Lommer et al., 2010; Oudot-Le Secq et al., 2007; Ruck et al., 2014; Tanaka et al., 2011). The overall organization of these genomes is conserved with all of them having a large single copy region (LSC), small single copy region (SSC), and two inverted repeats (IR). However, the plastid genomes range from ~ 116 to 165 kb, and they show extensive genome rearrangements, gene loss, duplication and functional transfers of genes to the nucleus (Ruck et al., 2014). The first introns in diatom plastid genome were reported in the *rnl* and *atpB* genes of *Seminavis robusta* (Brembu et al., 2013), and extrachromosomal plasmids were found in several diatom plastid genomes (Brembu et al., 2013; Ruck et al., 2014).

In this study, plastid genome sequences are reported for four more thalassiosiralean diatoms (*Thalassiosira weissflogii* (Grunow) G. Fryxell & Hasle, *Cyclotella* (F.T. Kützing) A. de Brébisson sp. L04_2, *Cyclotella* (F.T. Kützing) A. de Brébisson sp. WC03_2 and *Roundia cardiophora* (Round) Makarova) and representatives of three other diatom orders, Chaetoceratales Round & Crawford (*Chaetoceros simplex* Ostenfeld), Hemiaulales Round & Crawford (*Cerataulina daemon* (Greville) Hasle in Hasle & Syvertsen) and

Rhizosoleniales Silva (*Rhizosolenia imbricata* Brightwell). Gene content, genome size and gene order are compared across the genomes to better understand plastid genome evolution within Thalassiosirales.

## Materials and Methods

### Diatom strains and culture conditions

Seven diatom strains from different sources were examined (Appendix Table A.1). There were no permissions required for those collection sites, and there are no endangered/protected diatoms. All DNA were extracted from cultured materials, several of which are already publicly available. *Cerataulina daemon*, *Roundia cardiophora* and *Rhizosolenia imbricata* were grown in marine f/2 medium (Guillard, 1983)in a Percival model I-36LL incubation chamber (Percival, Boone, Iowa, USA) at 21 $^{o}$C; *Cyclotella sp. L04_2* and *Cyclotella sp. WC03_2* were grown in COMBO medium (Interlandi and Kilham, 1998) on a window-lit lab bench; *Thalassiosira weissflogii* and *Chaetoceros simplex* were grown in f/2 medium (Guillard, 1983) on a window-lit lab bench. The incubator was illuminated with fluorescent lights using a 12:12 hour light:dark photoperiod.

### DNA extraction

Diatom cells were pelleted in a Sorvall RC-5B refrigerated superspeed centrifuge (DuPont Company, Newton, CT, USA) for 20 minutes at 7649 × g from a culture in the late logarithmic phase of growth. Cells were lysed using a PARR Cell Disruption Bomb

(Parr Instrument Company, Moline, IL, USA) filled with nitrogen gas at 1500 psi. Isolation of DNA was performed following Doyle and Doyle (Doyle and Doyle, 1987)with modifications. Cetyl trimethylammonium bromide (CTAB) buffer was augmented with 3% PVP and 3% beta-mercaptoethanol (Sigma, St. Louis MO, USA). Organic phase separation was repeated until the aqueous fraction was clear. DNA pellets were resuspended in ~200 μL DNase-free water. Following treatment with RNase A (ThermoScientific, Lafayette, CO, USA) samples were again subjected to phase separation with chloroform, and DNA was recovered by ethanol precipitation. Samples were resuspended in DNase-free water, evaluated for concentration by NanoDrop and stored at -20$^{o}$ C.

**DNA sequencing and genome assembly**

Paired-end (PE) libraries with insert sizes of 400 bp were prepared at the Genome Sequence and Analysis Facility (GSAF) at the University of Texas at Austin. Illumina HiSeq 2000 paired-end platform (Illumina, San Diego, CA, USA) was used to sequence total genomic DNA. The PE Illumina reads were assembled with Velvet v.1.2.08 (Zerbino and Birney, 2008; Zerbino et al., 2009) using multiple *k*-mers ranging from 71 to 83. Plastid contigs were identified by BLAST analyses of the assembled contigs against published diatom plastid genomes from NCBI. The boundaries between inverted repeats and single copy regions were confirmed bioinformatically or using PCR and Sanger sequencing. The latter two techniques were also utilized to fill gaps in the plastid genome sequences. The PCR primers used for Sanger sequencing were designed by Primer3

(Untergasser et al., 2012) in Geneious R6 v.6.1.6 (Drummond and al, 2010) (Appendix Table A.2).

**Genome annotations and analyses**

Plastid genomes were annotated using Dual Organellar GenoMe Annotator (DOGMA) (Wyman et al., 2004), followed by manual corrections for start codons using Geneious R6 v.6.1.6. tRNA genes were predicted using DOGMA (Wyman et al., 2004) and tRNAscan-SE 1.21 (Schattner et al., 2005). Boundaries of rRNA genes, tmRNA *ssra* gene and signal recognition particle RNA *ffs* gene were delimited by direct comparison to sequenced diatom orthologues with Geneious R6 v.6.1.6 (Drummond and al, 2010). Circular plastid genome maps were generated with Organellar GenomeDraw (OGDraw) (Lohse et al., 2007). Repeated sequences were identified by performing BlastN v.2.2.28+ comparisons of each plastid genome against itself with an e-value cutoff of $1e^{-10}$ and at least 90 percent sequence identity. Annotated plastid genomes are available from GenBank using accession numbers KJ958479 – KJ958485. Genome rearrangements were estimated with MAUVE after eliminating one copy of the inverted repeat (Darling et al., 2004). Numbers of genome inversions were inferred by GRIMM (Tesler, 2002).

**Identification of genes transferred to the nucleus and signal peptides**

Genes absent from plastid genomes were searched for by BLAST searches in *Cyclotella nana* nuclear genome against assembled contigs of transcriptome assemblies of *T. weissflogii* (MMETSP0878) and *Rhizosolenia setigera* (MMETSP0789) from the

Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) website (http://marinemicroeukaryotes.org/) and nuclear assembly of *T. oceanica* (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AGNL01#contigs) using BLASTN with an e-value cutoff of $1e^{-10}$. The previous reported nuclear copy of *acp* gene in *Cyclotella nana* (XM_002290970) was used as the query sequence to search for the missing *acp* genes. SignalP was used to predict signal peptides and cleavage sites (Petersen et al., 2011).

### Phylogenetic analysis

Sequences of 20 plastid genes (*psaA*, *psbC*, *petD*, *petG*, *atpA*, *atpG*, *rbcL*, *rbcS*, *rpoA*, *rpoB*, *rps14*, *rpl33*,*rnl*, *rns*, *ycf89*, *sufB*, *sufC*, *dnaK*, *dnaB*, *clpC*) from 22 diatom taxa were aligned with MAFFT (Katoh et al., 2005). This included 15 published diatom plastid genomes and the seven genomes sequenced in this study. All sequences were included, and protein-coding genes were partitioned by gene and codon position. A maximum likelihood tree was constructed with RAxML7.2.8 (Stamatakis, 2006a), using the substitution model GTR+G+I and "-f a" option, and 1000 bootstrap replicates were performed to evaluate support for clades.

## Results

### General features of plastid genomes

All seven sequenced plastid genomes mapped as single circles with two IRs dividing the genome into LSC and SSC regions (Figure 2.1). The genomes are compact and lack introns. The three rRNA subunits (5S, 16S and 23S) are in the IR. Twenty-seven tRNAs together with two other RNAs, transfer-messenger RNA (*ssra*) and plastid signal recognition particle RNA (*ffs*), are found in all genomes. Nucleotide composition is highly conserved, with G+C content ranging from 30-32% (Appendix Table A.3). Four pairs of overlapping genes are present in the seven diatom genomes; *sufC-sufB* by 1 bp; *psbD-psbC* by 53 bp; *atpD-atpF* by 4bp versus 1 bp in *Rh. imbricata*; and *rpl4-rpl23* by 17 bp in the two the *Cyclotellas* versus 8 bp in the other species (Appendix Table A.3). The number of protein-coding genes ranges from 122 to 130. All protein-coding genes use the standard plastid-bacterial genetic code except for *psbC* in *Ro. cardiophora*, which uses ACG as the start codon instead of ATG. General features of the seven plastid genomes are compared with the two published thalassiosiralean genomes in Appendix Table A.3.

**Gene loss**

The protein-coding gene complement of the six Thalassiosirales plastid genomes is almost identical with 125 shared genes. A few notable exceptions were found. *ycf66* in *Ro. cardiophora* is a pseudogene as evidenced by several internal stop codons. The *acpP1* (acyl carrier protein) gene and the *syfB* (Phenylalanyl-tRNA synthetase) gene are missing in all Thalassiosirales (Figure 2.2; Appendix Table A.4). *acpP1* is present in all three sequenced non-Thalassiosirales diatoms; however, *syfB* is missing only in the more distantly related *Rh. imbricata* (Figure 2.2; Appendix Table A.4). The *ycf42* gene is missing in both *Ce. daemon* and *Ch. simplex*. The *ilvB* and *ilvH* genes, the large and small subunits of acetolactate synthase, are only found in *Ce. daemon* (Figure 2.2; Appendix Table A.4). Several genes are missing from *Rh. imbricata*, including three photosynthetic genes (*psaE*, *psaI* and *psaM*), the protein translation elongation factor Tu (*tufA*), *syfB* and *ycf35*.

**Functional gene transfer from plastid to nucleus**

One ORF with 83.41% identity to the *Cyclotella nana* hypothetical plastid targeted acyl carrier protein gene *acp3* (XM_002290970) was found in the assembled transcriptome contig (MMETSP0878-20121228|7451_1) of *T. weissflogii*. The canonical signal peptide cleavage site ASAFVP, same as the signal peptide cleavage site of the *acp3* gene in *Cyclotella nana*, was found and indicated plastid targeting after cleaving between the endoplasmic reticulum (ER) signal peptide and transit peptide (Appendix Figure A.1). However, SignalP did not indicate the presence of a signal peptide (Appendix Figure A.1). BLAST analyses of the nuclear *acp3* gene of *Cyclotella nana* against the *T. oceanica* nuclear genome revealed one ORF with 86.64% identity. The canonical signal peptide cleavage site ASAFAP was found (Appendix Figure A2.1), and SignalP indicated peptide signaling to the ER. Searches for the missing *syfB* gene using gene sequences from the closely related species *Ce. daemon* and *Ch. simplex* against the nuclear genome of *Cy. nana* and *T. oceanica* and the transcriptome assembly of *T. weissflogii* did not identify any matches. Searching the annotated transcriptome data on the MMETSP website of a related species *Rhizosolenia setigera* Brightwell CCMP 1694 showed several contigs

(MMETSP0789-20121207|1125_1, MMETSP0789-20121207|12246-1 *etc.*) annotated as elongation factor Tu domain or elongation factor Tu binding domain.

**Genome size and repetitive DNA**

The size of the seven sequenced diatom plastid genomes ranges from ~ 116 kb in *Chaetoceros* to ~ 129 kb in *Cyclotella* (Appendix Table A.3). Plastid genomes of the Thalassiosirales are larger than the three non-Thalassiosirales species (*Ch. simplex*, *Ce. daemon* and *Rh. imbricata*, Appendix Table A.3). The sizes of the LSC of the Thalassiosirales are similar to other diatoms sequenced here, however, the sizes of the SSC (24-27 kb) are smaller (27-40 kb) (Figure 2.3, Appendix Table A.3). The IRs of Thalassiosirales tend to be larger, ranging from 18 to 23 kb, compared to 7 kb in *Ch. simplex* and *Ce. daemon* to 16 kb in *Rh. imbricata* (Figure 2.3, Appendix Table A.3). The plastid genomes are compact with small intergenic spacer regions averaging 87-155 bp (Appendix Table A.3). BLASTN analysis of each plastid genome against itself revealed only five short tandem repeats in Thalassiosirales with lengths ranging from 79 to 90 bp (Appendix Table A.5).

The *rrnS-trnI-trnA-rnL-rrn5* gene cluster comprises the core of the IR. In Thalassiosirales, genes at the boundaries of IRs and single copy regions are the same, except for *T. oceanica*, which has an IR expanded through the *clpC* gene in SSC (Figure 2.3). The Chaetocerotales (*Ch. simplex*) and Hemiaulales (*Ce. daemon*) plastid genomes are smaller than the other diatoms examined. The IR of *Ch. simplex* is 7403 bp, which is slightly larger than the IR of *Ce. daemon* at 7004 bp (Figure 2.3). The IR of *Ch. simplex* includes one more gene (*acpP*) than *Ce. daemon*. The IR of Rhizosoleniales (*Rh. imbricata*) is larger than *Ch. simplex* and *Ce. daemon*.

**Ancestral plastid genome organization of Thalassiosirales**

To reconstruct the ancestral plastid genome organization of Thalassiosirales, shared inversions and ancestral IR/SSC and IR/LSC boundaries were identified. The Mauve alignment identified thirty-two locally collinear blocks (LCBs) shared by the nine diatom plastid genomes examined (Appendix Table A.6). Gene order within Thalassiosirales is very conserved, except for *T. oceanica* (Figure 2.4). *Cyclotella nana*, *T. weissflogii* and *Ro. cardiophora* have identical gene orders. Likewise, *Cyclotella* sp. L04_2 and *Cy.* sp. W03_2 have identical gene orders. The gene order of these two groups differs by only a single inversion of five adjacent LCBs (-19)(-15)(-14)(-9)(-10) between *rpl 19* and *rpl 20* in the LSC region (Appendix Table A.6; Figure 2.4). The plastid genome of *T. oceanica* is much more rearranged than other members of Thalassiosirales. GRIMM analysis estimated that ten inversions could explain the different gene orders between *Ro. cardiophora* and *T. oceanica* (Appendix Figure A.2). Based on the most parsimonious reconstruction, the ancestral gene order of Thalassiosirales is the same as that of *Ro. cardiophora*, *T. weissflogii* and *Cy. nana*. The ancestral IR/LSC and IR/LSC boundaries in Thalassiosirales are shared by *Ro. cardiophora*, *T. weissflogii*, *Cy. nana*, *Cy.* sp. L04_2 and *Cy.* sp. WC03_2.

**Genome rearrangements between Thalassiosirales and the other three diatoms sequenced**

Twenty inversions were inferred between the ancestral Thalassiosirales condition and *Rh. imbricata* (Appendix Table A.7). Fourteen inversions were inferred between the Thalassiosirales ancestral gene order and *Ce. daemon*, and seventeen inversions were inferred between the Thalassiosirales ancestral gene order and *Ch. simplex* (Appendix Table A.7). Among those inversions two inverted gene blocks, (8) to (-8) and (23) to (-23), are shared by all three non-Thalassiosirales (Appendix Table A.7). In addition, two inversions, (10)(9) to (-9)(-10) and (30)(31)(32)(27)(26)(25) to (-25)(-26)(-27)(-32)(-31)(-30), are shared by *Ce. daemon* and *Ch. simplex* (Appendix Figure A.3). *Chaetoceros simplex* and *Ce. daemon* gene orders are more similar to each other than either is to *Rh. imbricata* (Figure 2.4, Appendix Table A.7). The most extensive genome rearrangement occurs between *T. oceanica* and *Rh. imbricata*, which differ by twenty-five inversions (Appendix Table A.7).

**Discussion**

The Thalassiosirales is a well-supported monophyletic diatom order common in marine, brackish, and freshwater habitats. Due to the monophyletic origin, we expect that the plastid genomes within this order will share many features in terms of gene content, genome size and gene order. All Thalassiosirales plastid genomes are very compact, lacking introns and having only a few short repeats. In contrast, genome organization of outgroup species varies considerably. The Thalassiosirales show a much higher level of conservation of genome organization compared to a recent comparison of a more phylogenetically diverse assemblage of diatoms (Ruck et al., 2014). Denser sampling of this order provides valuable insights into the dynamics of plastid genome evolution within a single order.

**Conserved gene content within Thalassiosirales**

The plastid genomes of Thalassiosirales have 126-127 protein-coding genes, together with 3 rRNAs and 27 tRNAs (Appendix Table A.3). Gene content variation is limited in the order with only few notable gene losses/transfers compared to other diatoms (Figure 2.2). The *acpP1* and *syfB* genes are absent from all Thalassiosirales. It is well known that plastid genes tend to undergo a sequential process of transfer from the plastid to the nucleus (Jansen and Ruhlman, 2012). Centralized regulation of plastid metabolism in the nucleus has been suggested as a potential driving force for these transfers (Lommer et al., 2010). A nuclear encoded plastid targeted acyl carrier protein gene was reported in *Cyclotella nana* (Oudot-Le Secq et al., 2007) and *Synedra acus* (Galachyants et al., 2012). Previous research showed that a conserved amino acid motif AXAFXP at the cleavage site of the signal peptide was crucial for plastid targeting (Gruber et al., 2007). A nuclear encoded, plastid targeted acyl carrier gene was located in the nuclear genomes of *T. weissflogii* and *T. oceanica* with a canonical AXAFXP motif (Appendix Figure A.1). Searching the transcriptome data of *Cyclotella meneghiniana* from the MMETSP website also revealed an ORF (CAMNT_0012963711) with 84.91% identity with the *acp3* gene in *Cyclotella nana,* and with an ASAFVP signal peptide cleavage motif indicating plastid targeting (data not shown). These results suggest that *acpP1* in Thalassiosirales likely represents a functional transfer from the plastid to the nucleus.

Transfer of *petF* from the plastid to the nucleus is unique to a single species of Thalassiosirales, *T. oceanica* (Brembu et al., 2013; Galachyants et al., 2012; Kowallik et al., 1995; Oudot-Le Secq et al., 2007; Tanaka et al., 2011). It was suggested that this transfer may have been driven by an adaptation to a low iron environment (Lommer et al., 2010). To test whether this transfer is environmentally driven or limited to a single species, denser taxon sampling of species throughout the diatom phylogeny in different environments with varying amounts of iron is needed. The sequencing of the plastid genome of *Skeletonema*, the closest relative of *T. oceanica* (Alverson et al., 2007), and other diatoms living in the open water with low iron concentration will enhance the understanding of the forces causing the transfer of the *petF* gene. Another possible gene loss/transfer within Thalassiosirales is *ycf66*, which is a pseudogene in *Ro. cardiophora* as suggested by the presence of several internal stop codons. However, more nuclear data are needed to test whether this gene is lost completely or it has been transferred to the nucleus.

**Variation of gene content in non-Thalassiosirales species**

There are large differences in gene content in non-Thalassiosirales plastid genomes (Figure 2.2). The large and small subunits of acetolactate synthase, *ilvB* and *ilvH*, are reported present in all sequenced red algal plastid genomes (Janouškovec et al., 2013). There has been a history of repeated loss of these genes among the 16 diatom genomes

(Ruck et al., 2014). Among the seven new plastid genomes reported here, *ilvB* and *ilvH* are absent in all species except *Cerataulina daemon*. The most parsimonious reconstruction of gene gain/losses suggests that these genes were reacquired independently by this species. More plastid genomes need to be sampled to better understand the loss/gain history of these genes across the diatom tree.

The *ycf42* gene is missing from the plastid genomes of both *Ce. daemon* and *Chaetoceros simplex*. This gene was reported lost from the plastid genome of *Fistulifera* sp. JPCC DA0580 (Tanaka et al., 2011), *Leptocylindrus danicus* and *Cylindrotheca closterium* (Ruck et al., 2014), and has been pseudogenized in the plastid genomes of *Asterionellopsis glacialis*, *Asterionella formosa*, *Eunotia naegelii* and *Didymosphenia geminata* (Figure 2) (Ruck et al., 2014). More nuclear genome sequences are needed to determine whether *ycf42* has been transferred to the nucleus or has simply been lost.

The *ycf35* gene is missing from the *Rh. imbricata* plastid genome, representing the first case of the loss of this gene from a diatom. The *tufA* gene, encoding chloroplast protein synthesis elongation factor Tu, is also missing in *Rh. imbricata*. In the green algal ancestor of land plants, *tufA* was transferred from the plastid to the nucleus (Baldauf and Palmer, 1990). It is possible that *tufA* in *Rh. imbricata* has been functionally transferred to the nucleus but more nuclear data for this species is needed to confirm the transfer.

The most noteworthy gene losses are from the *Rh. imbricata* plastid genome where the three photosynthetic genes *psaE*, *psaI* and *psaM* are missing. It is well-known that parasitic prokaryotes and eukaryotes have experienced extensive genome size reduction due to loss of genes that are no longer functional (Moran, 2001; Vivares et al., 2002). The plastid genome of non-photosynthetic euglenoid flagellate *Astasia longa* lost all photosynthetic genes from its plastid genome except for *rbcL* (Gockel and Hachtel, 2000). The non-photosynthetic parasitic flowering plant *Epifagus virginiana* only contains 42 genes, all genes for photosynthesis and chlororespiration, together with many tRNA and RNA polymerase genes have been lost (Wolfe et al., 1992). But the loss of photosynthetic genes from plastid genomes of non-parasitic plants or algae is rare (Green, 2011). There are two possible explanations for the loss of *psaE*, *psaI* and *psaM* from the *Rh. imbricata* plastid genome. First, these genes may have been functionally transferred to the nucleus. Second, several studies have documented the presence of the endosymbiont, diazotrophic cyanobacterium *Richelia intracellularis* living within the siliceous frustules of several *Rhizosolenia* species, including *Rh. clevei* and *Rh. hebetat*a (Ashworth et al., 2013; Madhu et al., 2013; Villareal, 1990). So, it is possible that the missing photosynthetic genes of *Rh. imbricata* have been horizontally transferred to the endosymbiont, similar to the situation that occurred in the sea slug (Rumpho et al., 2008). However, without nuclear genome/transcriptome data for *Rh. imbricata* or evidence that a cyanobacterial endosymbiont genome has acquired these genes, it is not possible to determine which of these explanations is more likely.

**Genome size**

Plastid genome size varies among diatoms, ranging from 116,251 bp in *Synedra acus* (Galachyants et al., 2012) to 165,809 bp in *Cylindrotheca closterium* (Ruck et al., 2014). Expansion/contraction/loss of the IR, gene loss and duplication, and reduced size of the introns and intergenic spacer regions are the major factors contributing to variation in genome size (Jansen and Ruhlman, 2012). The large genome of *Cylindrotheca closterium* is mainly due to expanded intergenic spacer regions, which accounts for up to one quarter of the *Cylindrotheca* plastid genome (Ruck et al., 2014). It has been previously reported that the larger plastid genome size of *T. oceanica* compared to the *Cyclotella nana* is due to the expansion of the inverted repeat (Lommer et al., 2010). Thalassiosirales have larger plastid genomes than the three sequenced non-Thalassiosirales diatom in this study (Figure 1, Appendix Table A.3), and most of the diatom species sequenced by Ruck *et al.* (Ruck et al., 2014). The low number of repeats and the larger IRs in Thalassiosirales compared other species (Appendix Table A.3, Figure 2.3) indicates that their larger genome size is due to expansion of the IR.

**Genome rearrangements**

Evolutionary events can alter the gene order through inversion, expansion/contraction of the IR, gene duplication/loss, and transposition. Inversions caused by recombination between repeated sequences are considered the major mechanism for gene order changes in plastid genomes (Jansen and Ruhlman, 2012). There have been numerous rearrangements among published diatom genomes (Ruck et al., 2014), however, only two species of Thalassiosirales were previously sampled. Completion of plastid genomes of four additional members of the Thalassiosirales and additional diatom species from other lineages shows that gene order within Thalassiosirales is highly conserved with the exception of *T. oceanica*. The sequenced Thalassiosirales plastid genomes have three different gene order patterns. The first and most common pattern is shared by *Ro. cardiophora*, *T. weissflogii* and *Cyclotella nana* and it represents the ancestral gene order for the order. The second pattern occurs in the two freshwater *Cyclotella* species, which have one inversion in the LSC region that may be a synapomorphy for this clade (Figure 2.2, Appendix Tables A.6 - A.7). The third pattern is represented by *T. oceanica*, which is distinct from the rest of the Thalassiosirales. The genome has ten inversions relative to the ancestral genome arrangement for the order (Figure 2.2, Appendix Table A.7). The IR boundary of *T. oceanica* is also distinct from the rest of the Thalassiosirales (Figure 2.3). IR boundary shifts are a common phenomenon (Goulding et al., 1996) and is likely one of the factors contributing to the extensive rearrangements in *T. oceanica*. Alverson *et al.* (Alverson et al., 2007) examined the molecular phylogeny of Thalassiosirales and found that *T. weissflogii* and Cyclotella species group together, while *T. oceanica* is more phylogenetically distant from the Thalassiosirales that share similar gene order. To examine whether the gene order change is gradual or punctuated, a wider sampling of plastid genomes across the rest of the Thalassiosirales will be needed to elucidate gene order evolution in this order.

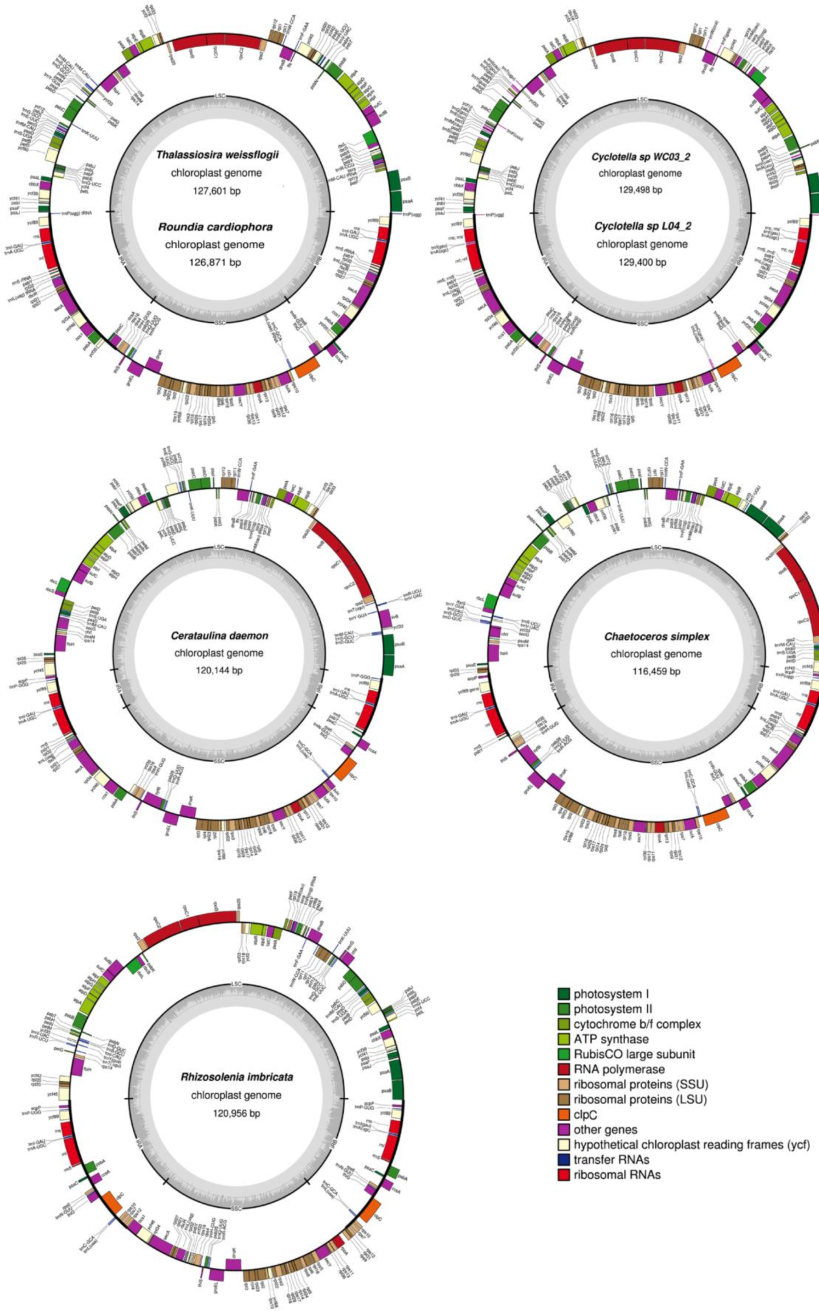



Figure 2.1. Plastid genome maps of seven newly sequenced diatom species. Species that share the same circular map have the same gene order. Genes on the outside are transcribed clockwise; those on the inside counterclockwise. The ring of bar graphs on the inner circle display GC content in dark grey.

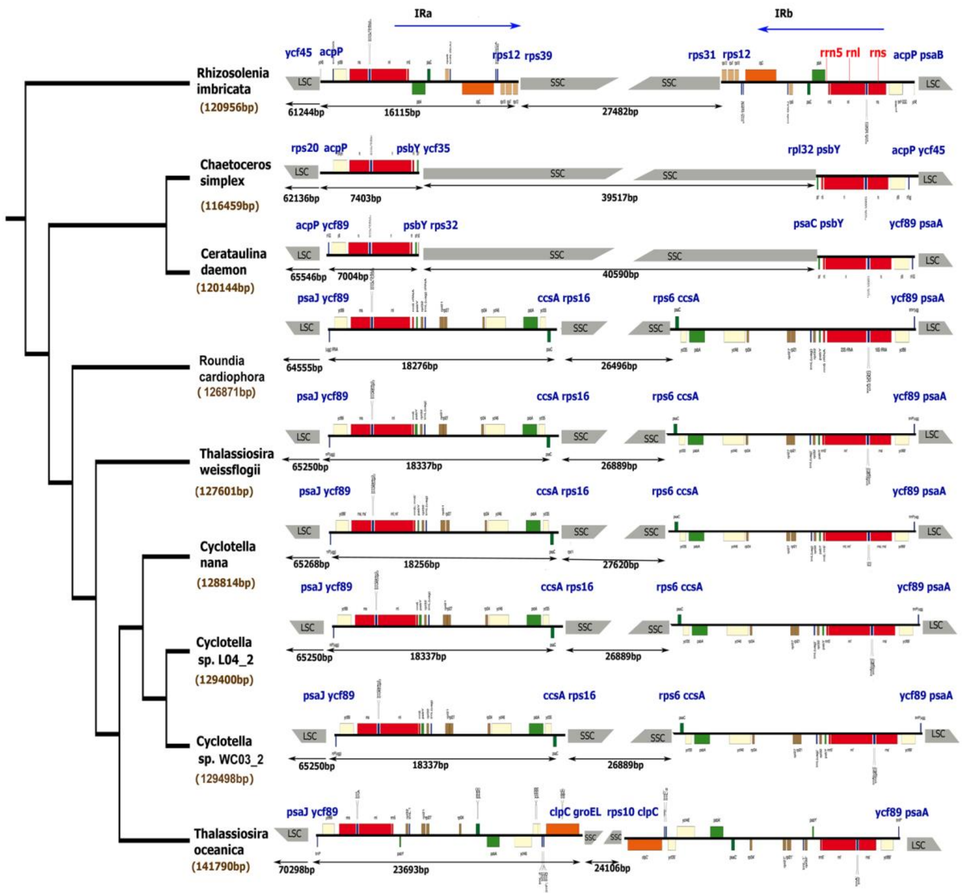



Figure 2.2. Phylogeny of Thalassiosirales and other diatom species based on twenty plastid protein-coding genes with gene/intron loss and plastid genome rearrangement events mapped on the branches. Number of genome inversions within Thalassiosirales were estimated based on Thalassiosirales ancestral genome using GRIMM. Taxa in bold are new genomes sequenced in this study.

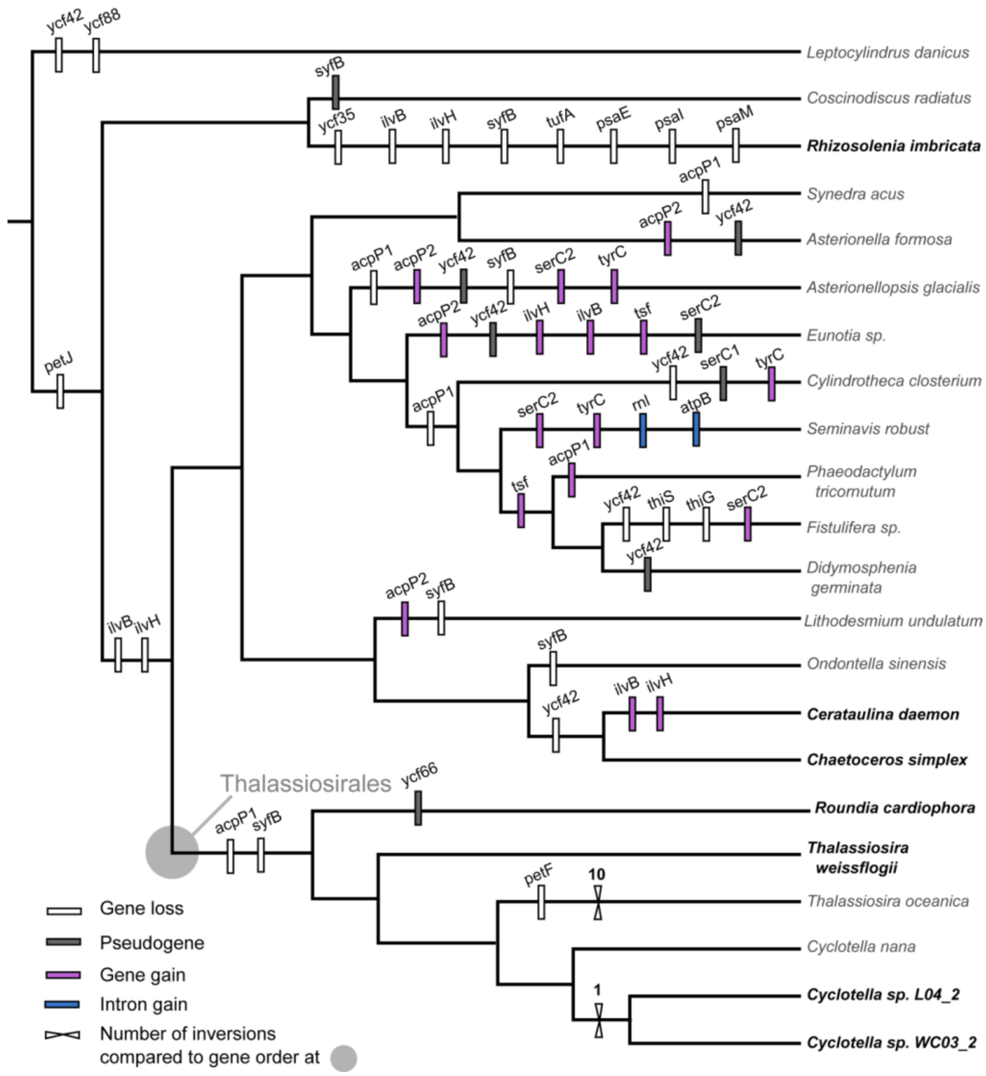



Figure 2.3. Comparison of inverted repeat boundaries in the seven diatom species newly sequenced for this study plus the two previously sequenced Thalassiosirales. Tree is that of Figure 2 with previously sequenced outgroup taxa pruned for visual simplicity. The numbers in brown indicate plastid genome size; the numbers in black below each genome fragment indicate the sizes of the LSC, IR and SSC, respectively. Protein coding genes at the IR boundaries are listed in blue. Three red gene blocks are *rrn5*, *rns* and *rnl*, respectively. Names in bold are Thalassiosirales. Underscored names are for taxa newly sequenced for this study.


Figure 2.4. Gene order comparison of the plastid genomes of seven diatoms sequenced for this study plus previously sequenced Thalassiosirales. Alignments were performed in Geneious R6 with mauveAligner. Taxon names in bold are members of the Thalassiosirales. Names underscored are those sequenced for this study.

# Chapter 3: Analysis of Forty Plastid Genomes Resolves Relationships in Diatoms and Identifies Genome-scale Evolutionary Patterns

This Chapter is published for publication in Advances in Botanical Research Volumn 85, 2018, Pages 129-155, where Mengjie Yu is first and corresponding author.

**Introduction**

Diatoms are photoautotrophic eukaryotic, single celled heterokont algae and play an important role in the global geological cycle, being responsible for one quarter of primary production, as well as being the primary biological mediators of the silica cycle in the oceans (Nelson, Treguer et al. 1995). They have delicate siliceous cell walls, which have been utilized to morphologically define taxa. Diatoms were traditionally classified into two major groups, centrics and pennates, with the former typically exhibiting radial or bi-(multi) polar symmetry, and the latter normally with bilateral symmetry. The pennates may or may not have a pair of slits in the cell wall (*i.e.* raphe). The centrics are paraphyletic, and have been divided into groups based on their wall outline (circular vs. triangular or quadrate). The pennates can be further divided into two groups, the raphe-bearing ("raphid") pennates and those without raphe slits ("araphid"). Traditional morphological studies showed considerable disagreement in diatom classification. Among those classification schemes, three strikingly different hypotheses were proposed. Steinecke (Steinecke 1931) proposed that centrics and pennates were each monophyletic sister taxa, and raphid pennates were monophyletic and nested within araphid pennates. In stark contrast, Simonsen (Simonsen 1979) concluded that centrics were paraphyletic, and araphids were monophyletic and nested within paraphyletic raphids. In disagreement with the previous two classifications, Round and Crawford (Round and Crawford 1981; Round and Crawford 1984) later argued that the three major lineages (centrics, araphid pennates and raphid pennates) were derived independently, and were thus each monophyletic.

Molecular phylogenies were similar to traditional phylogenies in that relationships varied from study to study, without a clear consensus as to arrangement of radial and (bi- or multi-) polar centrics (Theriot, Cannone et al. 2009; Theriot, Ashworth et al. 2010). Again, a few studies have produced radically different topologies, and relationships among diatoms are still a matter of debate (CHESNICK, KOOISTRA et al. 1997). Here, we cite only a range of results to illustrate our point. Araphid monophyly, as proposed by Round and Crawford (Round and Crawford 1981; Round and Crawford 1984), was supported by analysis of the *coxI* gene dataset with limited taxon sampling (Ehara, Inagaki et al. 2000). Centric monophyly was recovered using the nuclear-encoded small subunit ribosomal RNA (SSU) dataset (Van de Peer, Van der Auwera et al. 1996). These studies led to a reclassification of diatoms with Medlin *et al.* (Medlin and Kaczmarska 2004) naming the bulk of radial centrics as the Coscinodiscophyceae, the bi- and multi-polar centrics plus the order Thalassiosirales as the Mediophyceae and the pennates as the Bacillariophyceae. Each was argued to be monophyletic based on analysis of nuclear-encoded SSU. This classification, referred as the CMB hypothesis, has been under debate because different taxon sampling, alignments and optimality criteria can yield different results with radials being either monophyletic or not and polars (plus Thalassiosirales) being monophyletic or not (CHESNICK, KOOISTRA et al. 1997; Alverson, Jansen et al. 2009; Theriot, Ashworth et al. 2015). Incongruence in phylogeny was also reported using diatom plastid protein-encoded genes versus nuclear encoded SSU (Theriot, Ashworth et al. 2010).

The variations in results have led to inclusion of more sources of molecular data for resolving diatom relationships. The focus has been primarily on plastid genes due to the challenges of using nuclear data. The nuclear genome of eukaryotes is composed largely of multiple copy genes, making it difficult to reliably determine orthology. A more complex issue is that the diatom nuclear genome may be a chimeric assemblage due to multiple horizontal gene transfer events through diatom evolutionary history (Bowler, Allen et al. 2008). In contrast, the plastome is largely composed of single copy genes, with limited horizontal gene transfer events (Ruck, Nakov et al. 2014). Plastid protein coding genes are also easily aligned across a wide range of diatoms (Theriot, Ashworth et al. 2015). A recent study testing the phylogenetic informativeness using a broader suite of diatom plastid genes showed that the addition of plastid data adds signal instead of noise, and these same authors suggested that a phylogenomic study of plastid genes would provide valuable information for resolving the diatom phylogeny (Theriot, Ashworth et al. 2015).

Advances in sequencing technology have opened the door for generating genomic sequences more cheaply and quickly to better understand diatom evolution. The plastome organization potentially provides insights into diatom evolution. The first two diatom plastomes were sequenced in 2007 (Oudot-Le Secq, Grimwood et al. 2007), since then the number of sequenced diatom plastid genomes has increased greatly. Although the overall organization of these plastomes is conserved, all with a large single copy region (LSC), a small single copy region (SSC), and two inverted repeats (IR). Sequencing of phylogenetically diverse diatoms showed remarkable variation in genome size, gene content and gene order (Ruck, Nakov et al. 2014), with expansion of the IR and intergenic regions being the primary cause of plastome size variation (Ruck, Nakov et al. 2014; Sabir, Yu et al. 2014). Extensive plastome sequencing in Thalassiosirales, an order with a moderately well-resolved multi-gene phylogeny, showed a high level of conservation of genome organization among closely related species (Sabir, Yu et al. 2014). One environmentally-driven gene transfer event was reported in *T. oceanica*, where the *petF* gene encoding ferredoxin was transferred from the plastid to the nucleus, contributing to the ecological success of *T. oceanica* in iron limited environment by replacing the iron-sulfur protein with iron-free flavodoxin (Lommer, Roy et al. 2010). A plastid to nuclear gene transfer event of the acyl carrier protein gene *acpP* was also reported in all Thalassiosirales (Sabir, Yu et al. 2014).

Due to the limited number of plastome sequences available, phylogenomics has previously not been an option for resolving questions about diatom systematics. In addition to the paucity of diatom plastome data, the lack of genomes from potential outgroups meant early attempts at phylogenomics were unrooted. Thus monophyly of the Coscinodiscophyceae, which previous single and multi-gene phylogenies recover as either monophyletic or a basal grade, could not be tested. The sister group to pennate diatoms, which recovered as the bipolar diatom *Attheya* in a phylogeny with nine nuclear and plastid genes (Sorhannus and Fox 2011), could also not be tested as the genome was not available.

A phylogenetic framework with more complete taxonomic sampling is necessary to identify and understand patterns and processes of diatom plastome evolution.

In this study, we nearly doubled the number of sequenced plastomes and added critical taxa such as *Attheya*. We also included the recently sequenced genome of *Triparma*, a close relative of diatoms (Tajima, Saitoh et al. 2016), to provide a more in-depth examination of diatom plastome evolution and to resolve phylogenetic relationships among major clades.

## Materials and Methods

### Diatom strains and DNA extraction.

Eighteen diatom strains were collected from different sources (Appendix Table B.1). Taxon sampling was based on Theriot *et al*. (Theriot, Ashworth et al. 2015). The medium and cultivation methods are described in Appendix Table B.1. All DNAs were extracted from cultured materials. Diatom cells were pelleted in a Sorvall RC-5B refrigerated superspeed centrifuge (DuPont Company, Newton, CT, USA) for 20 minutes at 7649 × g from a culture in the late logarithmic phase of growth. Cells were lysed using a PARR Cell Disruption Bomb (Parr Instrument Company, Moline, IL, USA) filled with nitrogen gas at 1500 psi. Isolation of DNA was performed following Doyle and Doyle (Doyle 1987) with modifications. Cetyl trimethylammonium bromide (CTAB) buffer was augmented with 3% PVP and 3% beta-mercaptoethanol (Sigma, St. Louis MO, USA). Organic phase separation was repeated until the aqueous fraction was clear. DNA pellets were resuspended in ~200 μL DNase-free water. Following treatment with RNase A

(ThermoScientific, Lafayette, CO, USA) samples were again subjected to phase separation with chloroform and DNA was recovered by ethanol precipitation. Samples were resuspended in DNase-free water, evaluated for concentration by NanoDrop and stored at -20$^{o}$ C.

**DNA sequencing and genome assembly**.

Paired-end (PE) libraries with insert sizes of 400 bp were prepared at the Genome Sequence and Analysis Facility (GSAF) at the University of Texas at Austin. Illumina HiSeq 2000 platform (Illumina, San Diego, CA, USA) was used to sequence total genomic DNA. The PE Illumina reads were assembled with Velvet v.1.2.08 (Zerbino and Birney 2008; Zerbino, McEwen et al. 2009) using multiple odd number *k*-mers ranging from 71 to 83 on stampede supercomputer at the Texas Advanced Computing Center (TACC). Plastid contigs were identified by BLAST analyses of the assembled contigs against publicly available diatom plastid genomes from NCBI. The boundaries between inverted repeats and single copy regions were confirmed using Motif search in Geneious R6 v6.1.6 (Drummond and al 2010). Bowtie2 mapping (Langmead and Salzberg 2012) was utilized to fill gaps in the plastid genome sequences.

**Genome annotations and analyses**.

Plastid genomes were annotated using Dual Organellar GenoMe Annotator (DOGMA) (Wyman, Jansen et al. 2004), followed by manual corrections for start codons using Geneious R6 v.6.1.6. tRNA genes were predicted using DOGMA (Wyman, Jansen et al. 2004) and tRNAscan-SE 1.21 (Schattner, Brooks et al. 2005). Boundaries of rRNA genes, tmRNA *ssra* gene and signal recognition particle RNA *ffs* gene were delimited by direct comparison to sequenced diatom orthologs with Geneious R6 v.6.1.6 (Drummond and al 2010). The length of total genome, IR, SSC and LSC were shown in Appendix Table B.3. Genome length variation was analyzed using APE library in R (Paradis, Claude et al. 2004).

**Phylogenetic analysis**.

Sequences of 103 shared plastid protein-encoding genes from 40 diatom taxa and the outgroup *Triparma laevis* were aligned with MAFFT (Katoh, Kuma et al. 2005) based on translated protein sequences. This included twenty two published diatom plastid genomes, one outgroup species *Triparma laevis* and the eighteen plastid genomes newly sequenced in this study. Three different partitioning schemes were analyzed including no partitioning (one partition), partition by codon position (3 partitions), and partition by codon position and gene functional group (21 partitions). Genes in each functional group were listed in Supplementary Table S2. A maximum likelihood tree for each partition was computed on TACC Stampede supercomputer using RAxML 8.2.9 (Stamatakis 2014) with the substitution model GTR+G and "-f a" option. 1000 bootstrap replicates were performed. The probabilities conferred upon the molecular data by trees in which Araphids, Mediophyceae, Coscinodiscophyceae, and Coscinodiscophyceae plus Mediophyceae were each constrained as monophyletic were tested using the AU (Approximately Unbiased) and SH (Shimodara-Hasegawa) tests (Shimodaira 2002).

To test the possibility of recombination in diatom plastid genomes, eleven conserved gene order blocks occurring in most diatoms were identified (Appendix Table B.3). Gene blocks 1 to 4 and 6 to 10 were concatenated due to short sequence length. Four resulting concatenated sequence alignments (gene blocks 1-4, gene block 5, gene blocks 6-10 and gene block 11) were used to construct phylogenetic trees using RAxML with codon partition. SH tests (Shimodaira 2002) were run among the four resulting trees to test the congruency with the concatenated tree using 103 protein coding genes.

### Gene Order Analysis

Genome rearrangements were estimated with MAUVE after eliminating one copy of the inverted repeat (IRB copy) (Darling, Mau et al. 2004). The rearrangement distances between gene orders were measured by Genome Rearrangements in Man and Mouse (GRIMM) and visualized using d3heatmap library in R (Tesler 2002). Correlation between substitution rates (estimated from branch lengths on the ML tree) and genome rearrangement distances were analyzed using Pearson correlation coefficient and Pearson test with Bonferroni multiple testing correction. The gene order tree with varying branch lengths to best fit the constrained ML sequence tree was constructed using PAUP v 4.0b10 (Swofford 2003) not allowing negative branch lengths.

### Gene Content Analysis

Gene loss and gain events were mapped to the ML cladogram using Dollo parsimony in MacClade v4.08 (Maddison and Maddison 2000) based on the gene content comparison table (Supplementary Table S5). The presence and absence of genes were encoded as 1 and 0, respectively. Gene pseudogenization events were encoded as 2, and the states (absent, present, and pseudogenized) were treated as ordered. Dollo parsimony was used as an approximation of the assumption that genes were more likely to be lost from the plastome than gained, and that functioning genes are more likely to become pseudogenes than the reverse.

## Results

### Phylogenomic Analysis.

All partition schemes yielded trees with identical topologies and very similar branch lengths and bootstrap (BS) support values (Figure 3.1; Supplementary Figures. B.1-B.3). We present the results of the dataset partitioned by functional category and codon position. The maximum likelihood tree has 100% BS support values on most nodes (Figure 3.1). Raphid pennate diatoms (labeled "Raphid") were recovered as a monophyletic group sister to a clade of araphid pennate diatoms ("Araphid 2") with 100% BS support. Within raphid diatoms, *Eunotia naegelii* was sister to the rest of the raphid diatoms with 100% BS support. The model diatom *Phaeodactylum tricornutum* was recovered as sister to *Didymosphenia germinata*, but with only 52% BS support. Within Araphid 2, *Astrosyne radiata* was recovered on an extremely long branch. Araphid 1 was sister to Araphid 2 plus the Raphid group with 100% BS.

The Mediophyceae (bi- and multi-polar diatoms plus the Thalassiosirales) were contained in three clades (“Polar 1”, “Polar 2” and “Polar 3”) and was paraphyletic. *Attheya longicornis* formed Polar clade 3 with the two *Biddulphia* species, and together were sister to the pennate diatoms (Araphid 1 and 2, plus Raphid) with 100% BS support. The clade Polar 2 was sister to the Polar 1 clade with 94% BS support. The Thalassiosirales (including the euryhaline model diatom *Cyclotella nana* Hustedt, which was sister to two undescribed freshwater species of *Cyclotella*), were in Polar 1 clade, and were monophyletic with 100% BS support. *Eunotogramma* sp. and *Lithodesmium undulatum* were sequentially related to the Thalassiosirales with 100% BS support. *Biddulphia* plus *Attheya* formed a clade with 100% BS support, and that clade was sister to pennates with 100% BS support.

The radial centrics of the Coscinodiscophyceae (Radial 1, 2 and 3) formed a basal grade. Within Radial 3 *Guinardia striata* was nested within *Rhizosolenia spp*. with low BS support. The two remaining radial centrics groups, *Proboscia* sp. (Radial 2) and *Leptocylindrus danicus* (Radial 1) formed a grade at the base of the tree with each node having 100% BS support.

Monophyly of Araphids, Mediophyceae, Coscinodiscophyceae, and Mediophyceae plus Coscinodiscophyceae were each strongly rejected in favor of the best unconstrained tree by AU and SH tests (P values < 0.005).

Comparison of the maximum likelihood tree constructed by 4 different gene order blocks revealed the conservation of five internal branches separating major clades as indicated by arrows in Figure 3.1 and red lines in Supplementary Figure 3.4. All trees in

Supplementary Figure 3.4 showed the following relationships: *Leptocylindrus danicus* sister to the rest of diatoms; Polar diatoms paraphyletic with *Biddulphia* plus *Attheya* sister to pennates; Raphids monophyletic within the monophyletic pennates. These relationships were consistent with the tree constructed using 103 concatenated genes in Figure 3.1. The SH tests also showed none of those trees was significantly worse than the concatenated tree.

**Genome Size**.

Plastome length varied across clades (Figure 3.2) with *Plagiogramma staurophorum* exhibiting the largest size of 201,816 bp among all sequenced diatoms (Supplementary Table B.4). The Araphid 1 group (indicated in red), where *P. staurophorum* was recovered, showed relatively larger genome size compared to other groups (Figure 3.2). Large variation in IR length was found in Araphid 2 (violet) and Raphid (purple) groups, where the longest IR was almost 2-3 times longer than the shortest (Figure 3.2). Sister to Araphid and Raphid groups, the Polar 3 clade (brown) displayed a relatively conserved genome length, with little variation within the LSC, SSC and IR.

Polar 1 (light green) and Polar 2 (dark green) groups also showed relatively conserved genome lengths, with *Eunotogramma sp.* and *Plagiogrammopsis van heurckii* showing the largest genome size in the Polar 1 and Polar 2 clades, respectively. The Radial 3 group (dark blue) had relatively conserved genome length ranging from 118,120 bp to 125,283 bp (Appendix Table B.4). *Triparma laevis*, the outgroup species, showed the longest LSC and the shortest IR in the dataset (Figure 3.2; Appendix table B.4).

IR length showed more variation across the groups than the length of LSC and SSC (Figure 3.2). Phylogenetic independent contrast analysis showed that IR length contributed to the majority of the plastome size variation with $R^2$ = 0.6875. In comparison, the LSC and SSC contributed a relatively smaller portion, with $R^2$ = 0.2959 and 0.1036, respectively (Appendix Figure B.5).

**Gene Content.**

Dollo parsimony was used to optimize gene losses and gains on the diatom phylogeny as an approximation of the higher likelihood that genes are lost from the plastome rather than gained (Figure 3.3). Three genes involved in light-independent chlorophyll a biosynthesis, *chlB*, *chlL* and *chlN*, together with RNA polymerase omega subunit *rpoZ*, were entirely absent in the forty sequenced diatom plastid genomes. In contrast, two hypothetical plastid ORFs with unknown functions (*ycf89* and *ycf90*) were absent in the outgroup species *Triparma laevis* but present in all 40 diatom plastomes (Figure 3.3).

Other genes appear to have undergone multiple losses, such as elongation factor Ts *tsf*, which was lost 11 times, and the acetolactate synthase large and small subunits *IlvB* and *IlvH*, which were lost 10 times.

Loss through pseudogenization was relatively rare. The phenylalanyl-tRNA synthetase beta chain gene *syfB* showed seven losses and one pseudogenization event. The gene *ycf66* underwent one pseudogenization event but no losses. The gene *ycf42* was an exception with four pseudogenization events.

The branches with the largest number of gene losses (*Proboscia* sp. and *Astrosyne radiata*, 11 each) were also those with the greatest amount of inferred nucleotide substitution based on branch lengths (cf. Figures 3.1, 3.3).

Finally, introns were detected in *atpB* in Radial 2 species *Proboscia sp.* and in *petD* in Araphid 1 species in *Plagiogramma staurophorum*. A Conserved Domain Database (Marchler-Bauer and Bryant 2004) search of these introns revealed a reverse transcriptase with group II intron origin with E-values of 5.24e-44 and 7.89e-40 for *atpB* and *petD*, respectively (Appendix Figures B.6 and B.7). Blast comparisons of the intron-encoded proteins against NCBI revealed that the top hits were green algae reverse transcriptase with 50% and 54% nucleotide sequence identity, respectively.

**Gene Order**.

The 40 diatom plastomes exhibit various degrees of gene order rearrangement (Figure 3.4; Appendix Table B.6). The MAUVE alignment identified 42 locally collinear blocks (LCBs) shared by the plastid genomes examined (Appendix Table B.7). Closely related species share more similar gene orders. Identical gene orders were found in Radial 3, Polar 1, Polar 3 and Raphid groups. The mostly extensive sampled Polar 1 clade showed six very similar gene orders, with four Thalassiosirales (*Roundia cardiophora*, *Thalassiosira weissflogii*, *Discostella pseudostelligera*, *Cyclotella nana*) having exactly the same gene order, and the two closely related *Cyclotella* taxa differ by one gene block inversion (Appendix Table B.7).

Gene order and sequence divergence was strongly positively correlated in some regions of the tree. Approximately 40% of the Bonferroni corrected P values of the Pearson correlation between pairwise branch length and gene order rearrangement distances were significant (Appendix Table B.8). For example, *Astrosyne radiata*, which had the longest branch in the sequence tree (Figure 1), also exhibited a high level of gene order rearrangement and had a high correlation value of 0.71 (Appendix Table B.8). Similarly, *Proboscia sp.* had the next longest branch and also exhibited high levels of gene order rearrangement (Figures 3.1, 3.4).

## Discussion

The advent of sequencing technology and powerful computers made it possible to sequence the whole plastomes in a short amount of time at a reasonable cost. Given the phylogenetic diversity of diatoms, it is critical that a wider diversity be studied for their genomic properties to better understand their evolutionary history. In this study, we sampled extensively across the diatom phylogeny, especially taxa whose phylogenetic placement remains controversial. Our results provide deeper insights into diatom phylogeny and the dynamics of the plastome evolution.

### **Phylogeny of diatoms**.

Medlin *et al.* (Medlin and Kaczmarska 2004; Medlin 2017) proposed a classification with three monophyletic classes based primarily on SSU rDNA sequence analysis, Coscinodiscophyceae (radial centric), Mediophyceae (polar centric), and Bacillariophyceae (araphid and raphid pennate) or the CMB hypothesis. This hypothesis has not been widely accepted. In their higher level classification of eukaryotes, Adl *et al.*(Adl, Simpson et al. 2005) explicitly considered the Coscinodiscophyceae and Mediophyceae to be paraphyletic. Since then other studies have recovered the "grade" hypothesis, in which the Coscinodiscophyceae and Mediophyceae are each paraphyletic, some have recovered one or the other as monophyletic (Theriot, Ruck et al. 2011). The foundational problem is that the taxon sampling and molecular sampling to date have simply not generated a robust result. The CMB hypothesis is only 7 steps longer than the grade hypothesis (the most parsimonious hypothesis, L=14094 steps) using SSU data alone, for example (Theriot, Ashworth et al. 2010). Theriot *et al.* (Theriot, Ashworth et al. 2010) analyzed SSU, *rbcL* and *psbC* for 136 diatoms under ML; the optimal solution was again the grade hypothesis, but it was not statistically significantly different than the CMB hypothesis. In short, for most data and taxon sets in the diatom literature, it takes little to turn the CMB hypothesis into the grade hypothesis, and vice versa.

In the resulting search for more genes that might provide information about the diatom phylogeny, Theriot *et al.* (Theriot, Ashworth et al. 2015) found that individual plastid genes return results that disagree with traditional views, the CMB hypothesis, the grade hypothesis and indeed even with one another. In instances where plastids are biparentally inherited, there is the possibility that species hybridization could lead to recombination in the plastome, and to conflict between gene trees (D'Alelio and Ruggiero ; Sullivan, Schiffthaler et al. 2017). Such instances might result in different plastid genes yielding different but strongly supported trees. The individual gene trees recovered by Theriot *et al*. (Theriot, Ashworth et al. 2015), however, were not robustly supported. After studying the potential for saturation, and analyzing signal/noise ratios, they argued that individual plastid genes could be concatenated. Doing so, they recovered the grade hypothesis with strong support. Their conclusion was that the signal in the individual genes was low, but that it was additive. While the noise levels were high, they were not correlated and did not sum to a positively misleading signal. Thus, incongruence among plastid genes seemed to be best explained simply by noise.

We examined the potential for plastome hybridization as a source of misleading signal by analyzing four subsets of the plastome genome: two large blocks of genes that each seem to be inherited as a single locus and two concatenated subsets of smaller blocks of genes with each smaller block acting as a single locus returned trees rejecting the CMB topology in the same manner (*Leptocylindrus* sister to all other diatoms; *Attheya* plus *Biddulphia* sister to pennates). We cannot reject the hypothesis that (relatively minor) examples of plastome hybridization are occurring and may affect some parts of the tree. But it seems certain there are not two or more different strong signals for different relationships, and it seems certain that signal for the tree in Fig. 1 comes from across plastome.

We also tested the 103 combined plastid genes with three different partitions. All phylogenetic analyses showed the same tree topology with slightly different bootstrap support (Appendix Figures B.1 - B.3). The resulting ML tree partitioned by codon and gene functional group showed the Coscinodicophyceae ("radial centrics") and

Mediophyceae ("bi- and multi-polar centrics") were not monophyletic, while the Bacillariophyceae (raphid diatoms) were monophyletic with high bootstrap support (Figure 3.1). The AU tests of araphid pennate monophyly suggested by Simonsen (Simonsen 1979) and the CMB monophyly suggested by Medlin (Medlin 2017), were both strongly rejected with P values less than 0.05.

Our results also show that *Cyclotella nana* (*Thalassiosira pseudonana*), the model marine diatom abundant in the world's oceans and freshwaters, is more closely related to the euryhaline genus *Cyclotella* (Figure 3.1), which is congruent with Alverson *et al.* (Alverson, Beszteri et al. 2011). Another model marine diatom *Phaeodactylum tricornutum* is sister to *Didymosphenia geminata* in the Raphid clade with low bootstrap support (Figure 3.1). Raphid diatoms are a diverse clade and are currently under-sampled. More extensive taxon sampling in this group may further elucidate the phylogenetic position of this model organism.

**Plastome Evolution**.

Plastome size varies within diatoms, with sizes ranging from 116,251 to 201,816 bp (Figure 2, Appendix Table B.3). Several factors such as expansion or contraction of the IR, loss and duplication of genes, gain of introns and expansion of intergenic spacer regions are responsible for variation in genome sizes (Jansen and Ruhlman 2012). It has been previously reported that the larger plastid genome size in *Thalassiosirales* was mainly due to expansion of the IR (Sabir, Yu et al. 2014). Our study reports the largest diatom plastome at 201,816 bp in *Plagiogramma staurophorum* (Appendix Table B.3). This species also has the largest IR among diatoms at 34,888 bp (Figure 2 and Appendix Table B.3). The large size of the genome is mainly due to the IR expansion. An introduction of a 2,971 bp group II intron in *petD* also contributed to the larger size of *P. staurophorum*. This is consistent with our phylogenetic independent contrast analysis that IR length contributed to the majority of the plastome size variation (Appendix Figure B.5).

Our extensive sampling across diatom phylogeny also showed the similarity of genome sizes within clades (Figure 3.2), which is consistent with previous finding that species within Thalassiosirales having similar plastid genome size (Sabir, Yu et al. 2014). Ruck *et al.* (Ruck, Nakov et al. 2014) reported that larger intergenic space regions and the introduction of foreign genes played an important role in the expansion of plastome size. Within the Araphid 1 clade, the introduction of *SerC*1 gene probably contributed to the relative larger size of *Psammoneis obaidii*.

Massive numbers of gene losses occur across diatom plastomes (Figure 3.3). The four gene losses [*chlB, chlL, chlN* and *rpo*Z] together with two hypothetical protein gains [*ycf 89* and *ycf90*] appear to be synapomorphies for diatoms. Gene loss in plastomes is often associated with a functional gene transfer to the nucleus. Acyl carrier protein *acpP1*, the gene involved in the lipid metabolism pathway, was reported missing in all *Thalassiosirales* and a hypothetical transfer from plastid to nucleus transfer was proposed (Sabir, Yu et al. 2014). In this study, expanded taxon sampling in the Polar 1 group again confirmed the order-wide loss of *acpP1* in all *Thalassiosirales* and *Eunotogramma* (Figure 3.3), and we found the gene loss event occurred at the split between *Lithodesmium* and *Thalassiosirales*. Ferredoxin gene *petF*, an ecologically driven plastid to nucleus transfer in *T. oceanica* (Lommer, Specht et al. 2012), is also absent from the *Astrosyne radiata* plastome. *Astrosyne radiata* has not only undergone extensive gene order rearrangement and sequence divergence (Figure 3.1 and Appendix Figure B.7), it has also experienced extreme morphological divergence, having entirely lost the symmetry of pennate morphological structure (Ashworth, Ruck et al. 2012). Gene loss was suggested as a pervasive source of genetic change that potentially causes adaptive phenotype diversity (Albalat and Canestro 2016). Our gene content comparison showed massive gene loss (11 losses) in the *A. radiata* plastome. The connection between plastid evolution and morphological evolution suggests that perhaps the nuclear genome of *A. radiata* also experienced radical change.

Another long branch bearing species, *Proboscia. sp*., has also experienced massive gene loss (Figure 3.4, 10 losses) and a rare instance of an intron gain in *atpB*. However, in this case gene losses seem only weakly correlated with gene order rearrangement. *Actinocyclus* and *Coscinodiscus* are morphologically similar, identical in gene order and exhibit two losses each of functional genes (one due to pseudogenenization in *Coscinodiscus*). In contrast, the extensively sampled diatom order Thalassiosirales showed a pattern of stasis in gene content and gene order except for *T. oceanica*, which has a high degree of reorganization but only one gene loss and one gene gain. The branch leading to *Rhizosolenia fallax* and *R. imbricata* exhibits the next highest level of gene loss (5 losses), but very few gene order changes (Figure 3.4)**.**

Photosysnthetic gene loss is rare in diatom plastomes. Three noteworthy gene losses reported in diatom plastomes were the photosynthetic genes *psaE, psaI* and *psaM* missing from *Rhizosolenia imbricata* (Sabir, Yu et al. 2014). Our study also documented the loss of *psaE*, *psaI* and *psaM* in *Rh. fallax*, a species sister to *Rh. imbricata* but these genes are present in *Rh. setigera*, an earlier diverging *Rhizosolenia* in the Radial 3 clade (Figure 3.3). This indicates that the loss of these three photosynthetic genes occurred at the split between *Guinardia* and the more recently derived *Rhizosolenia* species.

There has been a history of repeated loss of the acetolactate synthase large and small subunits, *ilvB* and *ilvH* among diatom plastomes (Ruck, Nakov et al. 2014; Sabir, Yu et al. 2014). The tRNA synthetase gene, *syfB*, has a similar history of repeated loss in several diatom plastid genomes (Figure 3.3). A pseudogene copy is retained in *Coscinodiscus radiatus* indicating that losses are ongoing. The translation factor gene *tsf* shows a similar pattern (Figure 3.3). Ruck *et al.* (Ruck, Nakov et al. 2014) proposed a single deep plastid-to-nuclear transfer of *tsf*. In our study, we also found repeated losses of *tsf*, but data are not available at this time to determine if there have been multiple transfers to the nucleus.

Group II introns are mostly found in plants, fungi, eubacteria and archaea. The first group II intron encoding intronic maturase was found in tRNA-Met in the red alga *Gracilaria* (Janouškovec, Liu et al. 2013). There were reports of a group II intron in the *atpB* gene of the diatoms *Seminavis robusta* and *psaA* gene of *Toxarium undulatum* (Brembu, Winge et al. 2013; Ruck, Linard et al. 2016). We found two additional group II introns, one in *petD* gene in *Plagiogramma staurophorum*, and another in *atpB* gene in *Proboscia sp.* Both reverse transcriptases within the introns are most similar to reverse transcriptase in green algae. There have been studies reporting genes of green algal origin in diatom nuclear genomes (Bowler, Allen et al. 2008), and an endosymbiotic gene transfer from green algae was proposed (Moustafa, Beszteri et al. 2009). More intensive molecular investigation would likely reveal more evidence for the origin and evolution of those introns.

Highly conserved gene order within clades and extensively rearranged gene orders across groups have been reported in previous diatom plastome studies (Ruck, Nakov et al. 2014; Sabir, Yu et al. 2014). Our extended sampling further confirmed the conservation of gene order in closely related species and extensive rearrangement in distantly related species (Figure 3.4). Correlations between rates of nucleotide substitution and genomic rearrangements were detected in angiosperms (Jansen, Cai et al. 2007; Weng, Blazier et al. 2013). A significant positive correlation between nucleotide substitution and gene order rearrangement is present on the long branch leading to *A. radiata* (Appendix Table B.7). The longest branch in Polar 1 group, *T. oceanica*, also showed a significant correlation between sequence divergence and genome rearrangement (Appendix Table B.7).

Doubling the size of available plastome data of diatoms has greatly expanded our understanding of plastome evolution across this large and diverse photosynthetic clade. With the inclusion of *Triparma laevis* as the outgroup, we strongly rejected the CMB hypothesis of diatom classification. Our data suggests that Radial diatoms evolved as a grade, Polar diatoms and Araphid diatoms are paraphyletic, and Raphid diatoms are monophyletic and nested within the pennates. The 103 combined plastid gene data set also strongly suggests that *Attheya* together with the *Biddulphia* group is the sister to the pennate diatoms. Our expanded sampling again confirmed that expansion of IR played the major role of plastome size variation. Gene content and order of closely related species is much more conserved than distantly related species. Extensive gene loss events were also observed. Our study also shows a strong positive correlation between sequence divergence and genome rearrangement in diatoms, a phenomenon that has been documented in flowering plants (Jansen, Cai et al. 2007; Weng, Blazier et al. 2013; Schwarz, Ruhlman et al. 2017). Expanded studies of the sequence divergence in terms of substitution rates will provide more insights into the driving force for diatom plastome evolution.

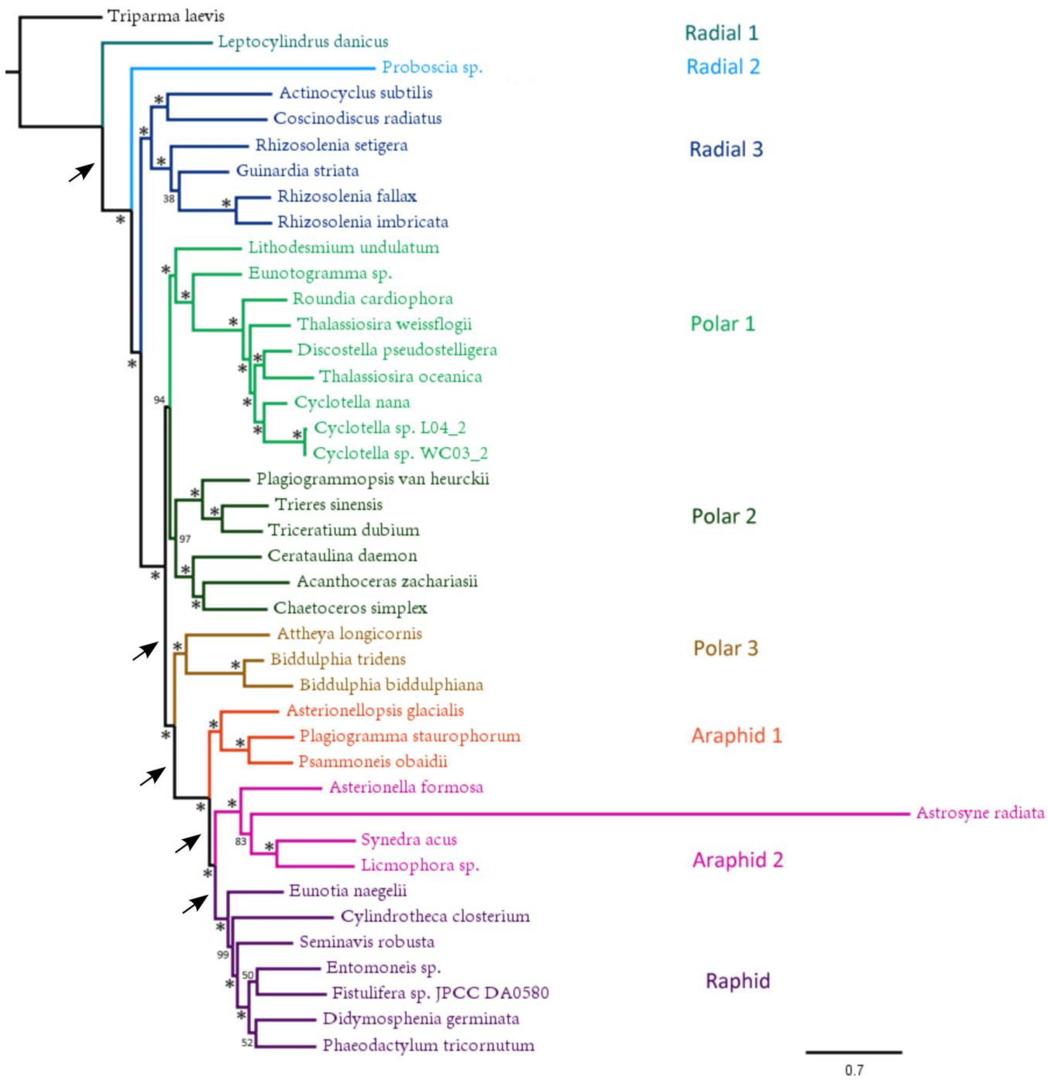



Figure 3.1. Maximum likelihood tree inferred from 103 shared plastid genes of 40 diatom species and the outgroup *Triparma laevis*. Branch lengths are proportional to the number of nucleotide changes as indicated by the scale bar (0.7 substitutions per site). Asterisks at nodes indicate 100% bootstrap support; numbers indicate bootstrap support values. Different colors indicate different diatom groups. The arrows indicate consistent branches separating different clades in gene order combination analysis.

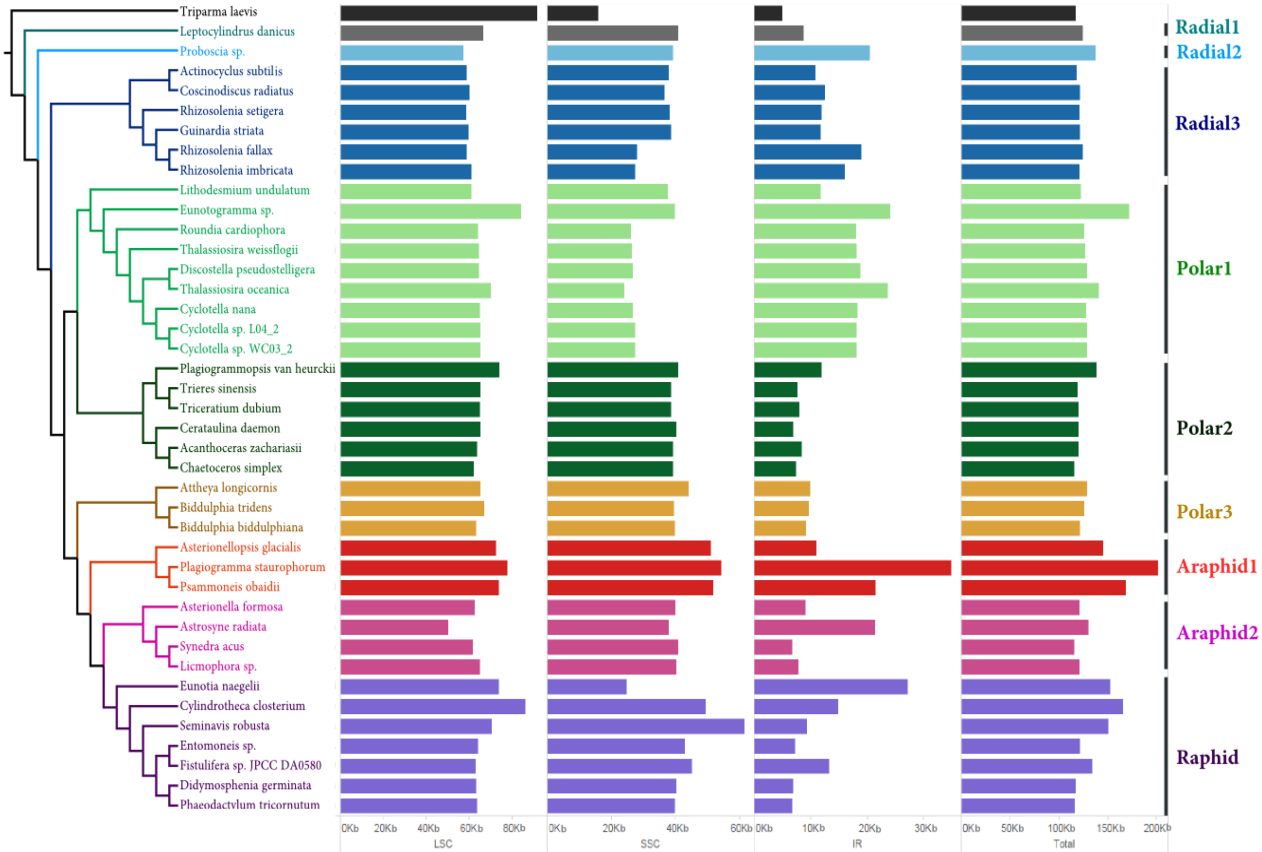



Figure 3.2. Genome length variation across 40 diatom species and the outgroup *Triparma laevis*. Colors indicate different diatom groups same as Figure 1. LSC = large single copy, SSC = small single copy, IR = inverted repeat. The length of LSC, SSC and IR were scaled differently. Scale on x axis in kilobases (Kb).

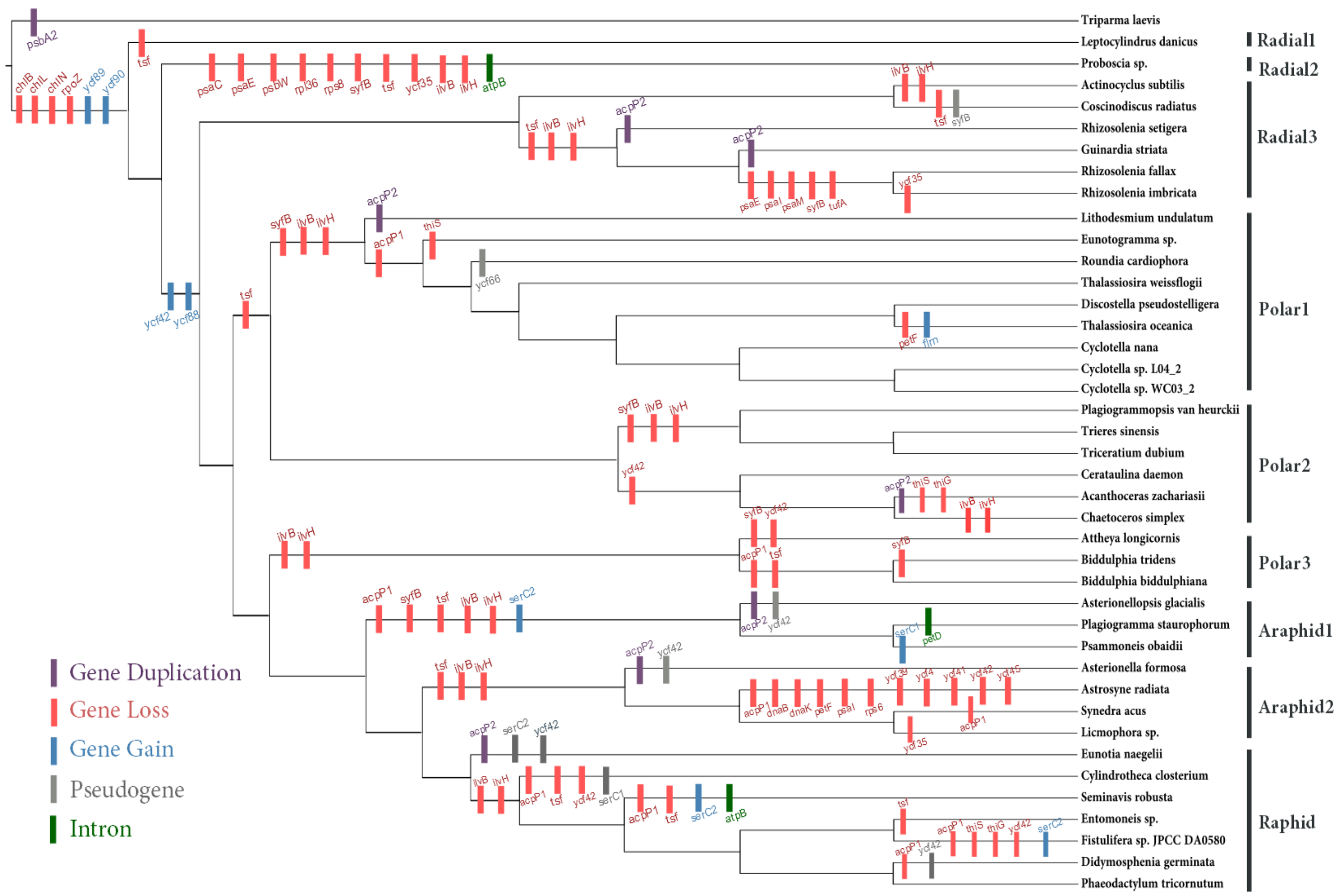



Figure 3.3 Gene and intron loss and gain events mapped on the cladogram of the ML plastid gene tree using Dollo parsimony.

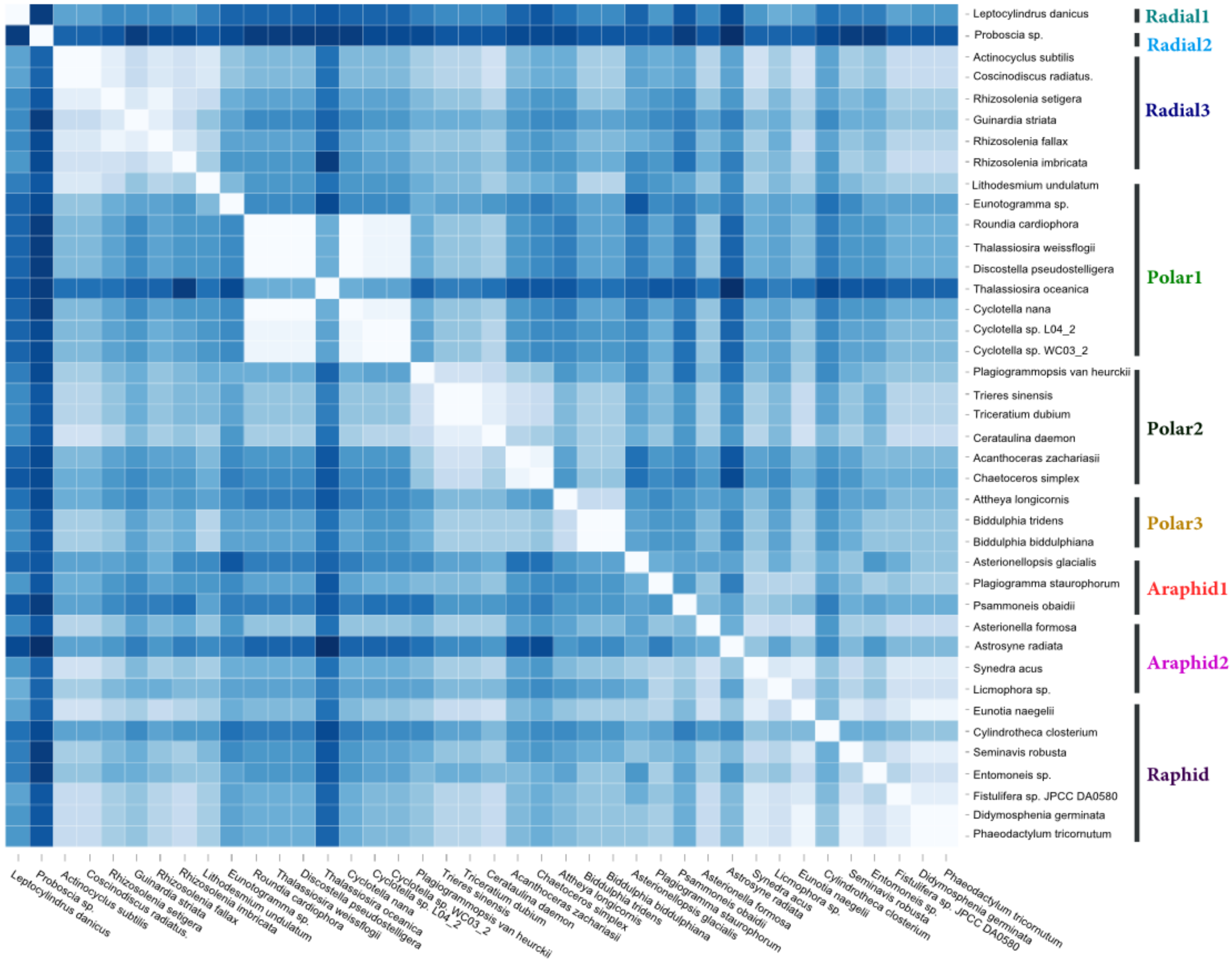



Figure 3.4. Heatmap of pairwise genomic rearrangement distance estimated by GRIMM. The intensity of the color is proportional to the degree of genome rearrangement. Dark blue indicates higher degree of genome rearrangement, and light color indicates lower degree of genome rearrangement.

# Chapter 4: Correlation Between Plastome Nucleotide Substitution Rates and Genome Organization across Diatoms

## Introduction

Knowledge of genome architecture evolution and nucleotide substitution rates is essential for our understanding of molecular sequence evolution and estimation of phylogenetic relationships. Synonymous substitutions (*dS*) are largely invisible to natural selection, while nonsynonymous substitutions (*dN*) may be under selective pressure. The ratio of non-synonymous (*dN*) and synonymous (*dS*) substitution rates is an indicator of selection. Variable *dN*/*dS* ratios among lineages may indicate adaptive evolution or relaxed selective constraints. Thus comparing rates of nucleotide substitutions provides a powerful tool for understanding the mechanisms of DNA sequence evolution.

Comparison of nucleotide substitution rates across functional groups of genes provides insight into plastome evolution. Genes encoding subunits that are integral to photosynthesis, such as ATP synthase (*atp* genes), cytochrome b6f complex (*pet* genes) and photosystems I and II (*psa* and *psb* genes) have been shown to have lower rates of nucleotide substitution than other functional groups in angiosperms (dicot and monocot) and conifers (Buschiazzo et al., 2012; Chang et al., 2006; Guisinger et al., 2008a; Guisinger et al., 2010). Ribosomal protein (*rpl* and *rps*) genes and RNA polymerase (*rpo*) genes were shown to have accelerated mutation rates (Blazier et al., 2016; Guisinger et al., 2011).

In addition to rate differences between gene functional groups, rate variation relative to genomic features such as genome rearrangements can also provide insights into the forces shaping plastome evolution. Evolutionary events can alter the gene order through gene duplication usually via expansion of the inverted repeat (IR), inversions, insertions and deletions (indels). Previous studies have identified a significant positive correlation between rates of nucleotide substitution and gene order changes in angiosperms plastid genomes, bacterial genomes, and arthropod mitochondrial genomes (Belda et al., 2005; Jansen et al., 2007; Shao et al., 2003; Weng et al., 2013; Xu et al., 2006). Disruption of DNA repair, recombination and replication (DNA-RRR) systems has been suggested to cause highly elevated nucleotide substitution rates and genome rearrangements (Jansen and Ruhlman, 2012). A recent study revealed a significant correlation between *dN* of DNA-RRR genes and plastome complexity in an angiosperm family (Zhang et al., 2016). Previous studies shwoed a negative correlation between genome size and substitution rates in previous plastome studies (Schwarz et al., 2017; Wu and Chaw, 2014).

Large-scale sequencing now allows us to compare thousands of genes in all domains of life. Factors affecting rates of sequence evolution in plastid genomes have extensively examined including speciation rates (Barraclough and Savolainen, 2001), generation time (Chang et al., 2006), gene function , and gene copy number (Wolfe et al., 1987). Synonymous and non-synonymous substation rates vary widely within and between taxa. Survey of 25 gene families in four grass species showed significantly heterogeneous dN/dS ratio across the branches, with majority of the ratio less than1.0 suggesting of selective constraint on amino acid substitution (Zhang et al., 2001). Comparing conifer to angiosperms, significantly slower evolution rates were found in conifer, however with higher dN/dS ratio indicating higher adaptation (Buschiazzo et al., 2012). Study of variation in substitution rates among genes and lineages in ferns revealed faster gene substitution rates in ferns than seed plants (Wolf et al., 2011). Order specific nucleotide acceleration was found in Poaceae within the monocot (Guisinger et al., 2010). Dramatically lower substitution rates were also in conifers than in angiosperms, and those differences vary across functional categories of genes (Buschiazzo et al., 2012).

Other factors related to life history have also been proposed to influence substitution rates. Previous studies on angiosperms (Barraclough and Savolainen, 2001; Duchene and Bromham, 2013), birds(Lanfear et al., 2010) and reptiles found a correlation between synonymous substitution and net diversification, suggesting a possible causal link between mutation rate and net diversification. Study show that tree and shrubs with long generation time has lower rate of mutation, while herbaceous plants with short generation time has higher rates of mutation (Smith and Donoghue, 2008).

Pattern of genome architecture change also seem to be associated with mutation rates. Previous studies showed a tendency of plastid genes in close proximity revealed similar changes of selection (Wicke et al., 2014). Strong correlation between high sequence divergence and low GC contents were detected(Wicke et al., 2014). Studies in ciliates found more fragmented genomes having significantly elevated mutation rate(Zufall et al., 2006).

Diatoms are the most species-rich group of phytoplankton in the ocean (Kooistra et al., 2007), originating about 250 Ma (Sorhannus, 2007). They are diploid and mainly dominated by asexual reproduction. Diatoms have a high capacity to accumulate mutations. Whole genome sequencing showed that the difference in genetic sequence diversity between model diatom species *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* is comparable to that between mammals and fish (Bowler et al., 2008). Diatoms reflect a fundamentally different evolutionary path from higher plants, green and red algae because they are derived from a secondary endosymbiosis between a non-photosynthetic eukaryote and a red alga. Diatoms offer an ideal opportunity to examine the patterns of nucleotide substitution rates for a secondary endosymbiosic lineage.

Studies on diatom substitution rates are advancing our knowledge on its evolution. Previous work has shown that *dS* and *dN* were lower in diatom plastid genes than in nuclear genes, and there was a negative correlation between the *dS* in plastid genes and the degree of codon usage bias (Sorhannus and Fox, 1999). The ecologically important transporters (*SITs*), which import silicic acid from the environment into the diatom cell, experienced strong purifying selection among 45 marine and freshwater thalassiosiroid diatoms (Alverson, 2007). Analyzing gene expression profiles in three genera of diatoms revealed positive selection in orphan genes and genes encoding protein-binding domains and transcriptional regulators (Koester et al., 2013). Whole genome sequencing of the cold-adapted pennate diatom *Fragiolariopsis cylindrus* revealed an association between *dN*/*dS* and condition-dependent expressions and a correlation between diversifying selection and allelic differentiation (Mock et al., 2017).

So far, no studies have yet compared substitution rates of diatom plastid genes on genome scale. Here, we carried out the first comparative study of substitution rates of diatom plastid genes in a genome scale. We explored the pattern of diatom plastid gene substitution rates. Our study also examined the correlation pattern between plastome mutation rates and potential genome features, such as genome size, indels, and genome rearrangement. This work advances our current understanding of diatom plastid genome evolution and the forces shaping the tempo and mode of diatom plastid genome evolution.

## Methods

### Gene Sequence Alignment and Phylogenetic Analysis

Plastid protein-coding genes were extracted from the 40 complete diatom plastomes across diatom phylogeny together with the outgroup species *Triparma laevis*. The 103 shared plastid gene sequences were aligned with MAFFT (Katoh et al., 2005). Protein-coding genes were partitioned by codon and gene functional category. A maximum likelihood tree was constructed with RAxML7.2.9 (Stamatakis, 2006b), with the substitution model GTR+G and "-f a" option. 1000 bootstrap replicates were performed. The maximum likelihood tree was then used as a constraint tree for estimating substitution rates.

### Nucleotide Substitution Rates

Nucleotide substitution rates (*dN* and *dS*) were estimated using the codeml function implemented in PAML (Yang, 2007). Codon frequencies were determined by the F3×4 model. Gapped regions were excluded with the parameter cleandata = 1 to avoid spurious rate inference. Pairwise rates with the outgroup species *Triparma laevis* were estimated with the parameter runmode = -2. Mutation rates were estimated for both the concatenated sequence and the sequences in different functional groups as listed in Supplementary Table 4.1.

**Plastid Genome Complexity Analysis**

The number of indels for the concatenated 103 protein coding genes was calculated using a custom python script in which *Triparma laevis* was used as a reference. Whole genome alignment among the forty diatom species was performed using the ProgressiveMauve algorithm in Mauve v2.3.1 (Darling et al., 2004). The same copy of IR (IRb) was removed from the plastid genome where two copies were present. The locally collinear blocks (LCBs) identified by the Mauve alignment were numbered with positive or negative sign based on strand orientation to identify synapomorphic genome rearrangements and estimate genome rearrangement distance. Inversion (IV) distances were estimated using GRIMM (Tesler, 2002). Genome size included only one copy of the IR for consistency.

**Correlation between Substitution Rates and Genome Characteristics**

Pairwise *dN* and *dS* values were calculated for each taxon compared to the outgroup species *Triparma laevis,* and corresponding *dN*/ *dS* ratios were calculated. Correlations of *dN* and *dS* with plastome size and number of indels for each genome were tested. Phylogenetic Generalized Least Squares was performed using the 'ape' and 'nlme' packages in R. The constraint tree was utilized with outgroup taxa pruned.

Pairwise nucleotide substitution rate and inversion distance for each diatom species were collected as vectors. The correlations between two vectors were calculated. The correlation between the rate of nucleotide substitution and the rate of genome rearrangement were tested using Pearson test. The resulting p-values were Bonferroni corrected using the built-in p.adjust function to remove the effect of multi-hypothesis testing.

## Results

### Substitution Rate in a Phylogenetic Context

Most clades were recovered with strong bootstrap support (see bootstrap support values in Figure 3.1). The radial centrics of the Coscinodiscophyceae (Radial 1, 2 and 3) formed a basal grade. The Mediophyceae (bi- and multi-polar diatoms plus the Thalassiosirales) were contained in three clades (“Polar 1”, “Polar 2” and “Polar 3”) and was paraphyletic. Araphid 1 was sister to Araphid 2 plus the Raphid group. Within Araphid 2, *Astrosyne radiata* was recovered on an extremely long branch. Raphid pennate diatoms (labeled “Raphid”) were recovered as a monophyletic group sister to a clade of araphid pennate diatoms (“Araphid 2”).

The *dN* and *dS* trees showed very similar pattern in branch length variation. The most accelerated lineage was branch 63 leading to *Astrosyne radiata* (Figure 4.1). Branch 4 leading to *Proboscia sp.* also showed accelerated rates in both *dN* and *dS*. Comparing substitution rates in different functional groups also showed the most accelerated rates on branch 63 (Appendix Figures C.1 and C.2).

### Rate Variation in Functional Groups of Genes

Gene sequences in each functional category were concatenated to estimate *dS* and *dN*. The patterns of nonsynonymous and synonymous substitution rates in different functional groups were similar (Figure 4.2). RNA polymerase genes had the highest median values of *dN* and *dS* among the major gene categories. Ribosomal protein genes (*rpl* and *rps)* also had high median values of *dN* and *dS*. Both the *dN* and *dS* median values of the genes integral to photosynthesis, such as *psa*, *psb*, *pet* and *ATP* genes, were much lower than the other groups. Comparisons of the individual genes in the other gene category (Appendix Table C.1) showed that the highest *dN* and *dS* was for *dnaB*, the replicative DNA helicase gene (Appendix Figures C.3 and C.4).

*dN* and *dS* were highly positively correlated for genes involved in similar functions. Photosystem I *psa* genes and photosystem II *psb* genes had correlation coefficients of 0.95 and 0.96 for *dN* and *dS*, respectively (Appendix Figures C.5 and C.6). Ribosomal protein small subunit (*rps*) genes and large subunit (*rpl*) genes had correlation coefficients of 0.96 and 0.97 for *dN* and *dS*, respectively. RNA polymerase genes were also highly correlated with *rpl* genes in both *dN* and *dS*.

**Correlation between Substitution Rates and Plastome Characteristics**

A significant positive correlation was observed between *dN*, *dS*, *dN*/*dS* and the number of indels (Figure 4.3). Both *dN* and *dS* showed positive correlation with the number of indels, while the ratio of *dN* and *dS* showed negative correlation with indel number. All correlations were significant with p-values less than 0.05 and small standard errors (Figure 4.3). No obvious correlation was found between the substitution rate and plastid genome size (Appendix Figure C.7). However, *Astrosyne radiata*, which had the highest *dN* and *dS*, among diatoms, showed a relatively small genome size compared with the rest of diatoms.

Correlations of pairwise mutation rate and genome inversion distance were tested among 40 diatom plastid genomes. Our results showed that *dN* had 25 significant correlations among the 40 pairwise comparisons at the significance level of 0.05 (Figure 4.4, Appendix Table C.2). *dS* and *dN*/*dS* had 18 and 13 significant correlations among the 40 pairwise comparisons, respectively. Polar 1 group, the mostly extensively sampled group of diatoms (indicated by the light green color in Figure had the largest percentage of significant correlation, with 7 out of 9 in both *dN* and *dS* and 6 out of 9 in *dN*/*dS*. *Astrosyne radiata*, the long-branch diatom, showed significant correlations in all of *dN* ($p$=2.41e-06), *dS* ($p$=2.23e-03) and *dN*/*dS* ($p$=3.55e-04) (Appendix Table C.2). *Astrosyne* gene order inversion distance also showed the highest correlation coefficient of 0.7431 with *dN*.

## Discussion

Identifying the pattern of nucleotide substitution underlying plastid gene evolution is key to understanding the mutational and selective cores responsible for diatom plastid genome evolution. In our study, over one hundred plastid genes were examined across 40 diatom species. The ribosome subunit and RNA polymerase genes showed accelerated nucleotide substitution rates compared to photosystem genes *psa*, *psb*, *pet* and *ATP* genes. Positive correlations were uncovered between substitution rates and number of indels and genome rearrangements. By using genomic scale sequences of an understudied yet important group in the tree of life, our study sheds light on the pattern and the forces shaping molecular evolution in diatom plastid genomes.

**Lineage specific mutation rates**

Lineage specific mutation rates were reported in previous studies. A general elevation of nucleotide substitution rates were observed in carnivorous versus non-carnivorous *Lentibulariaceae*, the plants exhibiting the most sophisticated implementation of carnivorous syndrome (Wicke et al., 2014). Studies in the marine cyanobacterium *Prochlorococcus* found significantly lower genome-wide average dN/dS ratio in high-light-adapted groups versus those in the closely related sister group *Synechoccus* (Hu and Blanchard, 2009). The authors argued that the lower dN/dS ratios suggest ingrelatively larger effective population size, which is consistent with their ocean abundance observation of *Prochlorococcu* (Hu and Blanchard, 2009). Among major groups of gymnosperms, significantly slower synonymous and nonsynonymous substitution rates were found in *cycad* comparing to *Pinaceae* (Wu and Chaw, 2015). Conifers had lower level of substitution rates compared to angiosperms, and it is proposed that reduced levels of nucleotide mutation coupled with large effective population size were the main contribution factor (Buschiazzo et al., 2012). Among seed plants, acceleration of non-synonymous rate in the subtree Euphorbia was also detected (Lee et al., 2011). The tufA genes, which encodes the elongation factor Tu, was found evolving at a fast pace in green algae *Coleochaetophyceae*, compared with other sister clades (Lemieux et al., 2016).

In this study, we found significantly higher mutation rate in the long branch bearing species *Astrosyne radiata*, comparing to the rest of the diatoms (Figure 1). Extensive gene loss was also found in *Astrosyne* (Chapter 2). Our results suggest unprecedented evolutionary events might be going on the branch leading to *Astrosyne*. Additional taxon sampling around the *Astrosyne* might help us better elucidate the evolutionary changes in araphid 2 clade.

**Differential Mutation Rates in Gene Functional Groups**

Gene essentiality is the most studied factor for mutation rate variation, with the idea that essential genes are subject to stronger selective constraint than non-essential genes (Wilson et al., 1977). Several studies in various organisms have demonstrated that variation in nucleotide substitutions is correlated with expression levels in which highly expressed genes evolve at a slower rate (Drummond et al., 2006; Sharp, 1991; Shields et al., 1988). Studies in plants *Picea* also showed negative correlation between substitution rate and gene expression, underlying that highly expressed genes might undergone greater selective constraints than lowly expressed genes (De La Torre et al., 2015). However, study over 3,000 mouse essential genes showed the relative importance of factors in determining mammalian protein evolution in descending order are gene compactness, gene essentiality, gene expression level (Liao et al., 2006). Studies of evolutionary rate in mammals and flies showed little correlation with expression level, but the rates of adjacent protein domains tend to fluctuate together (Du et al., 2013).

Diatom plastid genes mainly fall into two categories, those involved in the photosynthetic apparatus and in the transcription-translation apparatus. The first category mainly includes photosynthesis genes *psa* (photosystem I), *psb* (photosystem II), *pet* (cytochrome $b_6$/f complex) and *atp* (chloroplast ATP synthase). The second category involves RNA polymerase and ribosome proteins. It was found in conifers that genes involved in signal transduction and regulation of transcription and nucleic acid seem more likely to evolve under reduced constraint; whereas genes involved in translation, protein assembly, chlorophyll biosynthesis and cellular organization are under strong selective constraint (Buschiazzo et al., 2012). It was suggested that genes involved in signal transduction and regulation of transcription experienced adaptive selection which allow for responsiveness and plasticity to defend themselves against herbivores and pathogens; whearas genes in translation, cellular organization and chlorophyll biosynthesis were under strong selective constraint due to the fact that those processes are highly conserved (Buschiazzo et al., 2012) . Similar patterns were also found in angiosperms (Chang et al., 2006; Guisinger et al., 2008a; Guisinger et al., 2010). Studies in unicellular green alga *Ostreococcus* showed that faster evolving genes encode significantly more membrane or secretion associated genes, as cell surface modification is driven by selection on resistance to viruses (Jancek et al., 2008). In our study, the results also showed similar pattern that genes involved in photosynthesis had relatively lower substitution rates than genes in transcription-translation apparatus (Figure 4.3).

**Correlation between Substitution Rates and Plastome Characteristics**

**Indels**

Indels are thought to be a major driving force in sequence evolution (Britten, 1986). Previous studies in a broad range of eukaryotes and bacteria revealed that mutation rate is substantially elevated in regions surrounding sites that have undergone a short insertion or deletion mutations (Hodgkinson and Eyre-Walker, 2011; Hollister et al., 2010; Tian et al., 2008; Zhu et al., 2009). On 50bp either side of an indel, the mutation rate increased 30-fold in yeasts (Tian et al., 2008) and 6-fold in humans (Hodgkinson and Eyre-Walker, 2011). An “indel-induced mutation” hypothesis was introduced stating that presence of an indel is induces a high mutation rate (Tian et al., 2008). Indels were also reported associated within regions of repetitive DNA (Dettman et al., 2016). Studies in the carnivorous plant *Lentibulariaceae* showed a strong correlation of indels and substitution rates across plastid non-coding regions (Wicke et al., 2014). Similar to previously published results, we found both *dN* and *dS* had significant positive correlations with the number of indels in coding regions (Figure 4.3).

Studies in cotton showed the ratio of substitution rate and indel increased as divergence time increased (Xu et al., 2012).

**Size**

Previous studies of diatoms showed that the plastome size variation is mainly due to IR expansion and the introduction of foreign genes (Ruck et al., 2014; Sabir et al., 2014). An inverse relationship between mutation rate per base pair and genome size was proposed by Drake *et al.* (Drake, 1991; Drake et al., 1998). Bradwell *et al.* (Bradwell et al., 2013) showed a negative correlation between mutation rate and genome size in Riboviruses. Lynch et al. (Lynch et al., 2006) hypothesized that low organelle DNA substitution rates contribute to a more permissive environment leading to organelle genome expansion and high mutation rates resulting in genome contraction. Tests of this hypothesis in flowering plant organelle genomes showed some mixed results (Schwarz et al., 2017; Sloan et al., 2012). Significant correlation were found in gymnosperm *Picea* between gene family size and rates of sequence divergence (De La Torre et al., 2015). Negative association were found between dS values and cpDNA size in Cupressophytes, a conifer clade, but no association as detected for dn or dN/dS values(Wu and Chaw, 2014).

Our analyses did not show any significant relationship between mutation rates and genome size (Appendix Figure C.7). In fact, *Astrosyne radiata*, the diatom species that experienced multiple gene loss (Figure 3.3) showed the highest mutation rate in both *dN* and *dS* even though it has a relatively small genome size. More extensive sampling in the Araphid 2 clade (where the long branch bearing species *Astrosyne* belongs) would likely provide more information on the fast evolving mutation rate and plastid genome size.

**Genome Rearrangement**

Previous studies have shown a positive correlation between genome rearrangement and nucleotide substitution rates (Guisinger et al., 2008b; Schwarz et al., 2017; Weng et al., 2013). In our result, significant correlations between genome rearrangement and substitution rates were also observed in diatoms (Figure 4.4, Supplementary Table S4.2). The results also showed that *dN* had the largest number of significant correlations among all the pairwise comparisons. One possible mechanism could be improper DNA repair leading to genome rearrangement and increased nucleotide substitution. It has been suggested that genes involved in DNA replication, recombination, and repair (DNA-RRR) systems may be responsible for elevated nucleotide substitution rates and increased genome rearrangement in plastid (Guisinger et al., 2008b; Zhang et al., 2016). DNA repair mechanism is also proposed to explain the rearrangement and mutation rate in plant mitochondria (Christensen, 2013). Completed sequences for additional highly rearranged diatom plastid genomes, and characterization of genes involved in DNA repair in diatoms are need to better understand the highly accelerated substitution patterns.

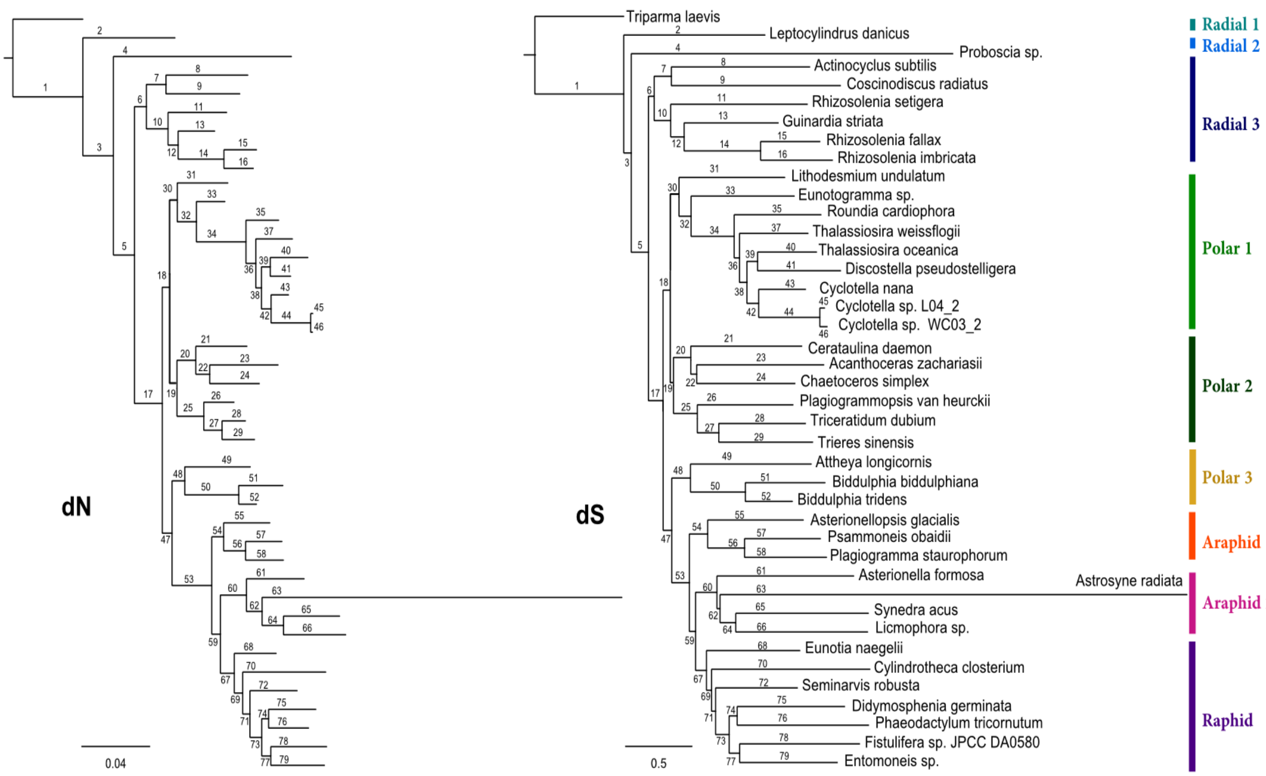



Figure 4.1. *dN* and *dS* trees estimated using maximum likelihood and 103 concatenated protein coding gene sequences. The bars at the base of each tree indicates the number of nucleotide substitutions per codon. *dN* and dS trees are on different scale. Numbers on the branches in the *dN* tree are branch numbers.

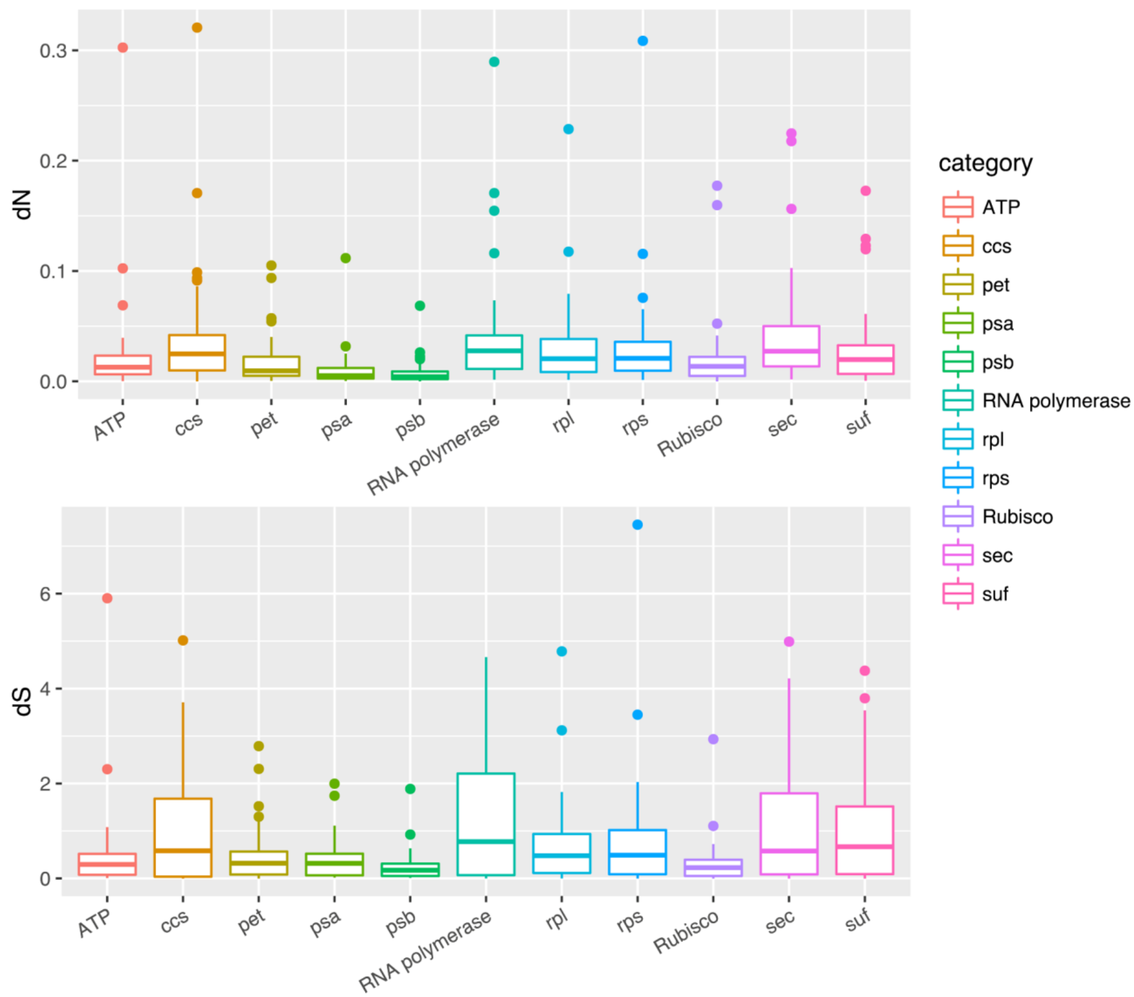



Figure 4.2. Boxplot of the number of nonsynonymous (*dN*) and synonymous (*dS*) substitutions for functional groups of genes.

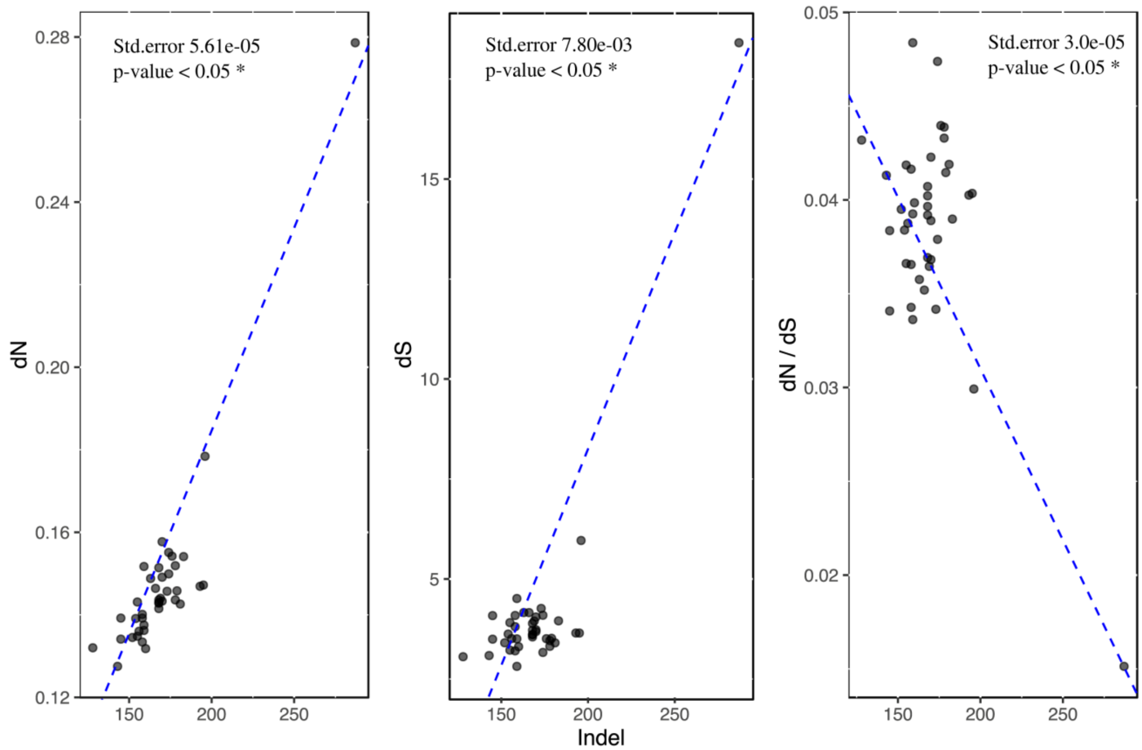



Figure 4.3. Relationship between the number of indels and substitution rates. Scatterplots were constructed and the regression line (dashed blue) and statistical values are shown. X-axis gives the number of indels in each species.

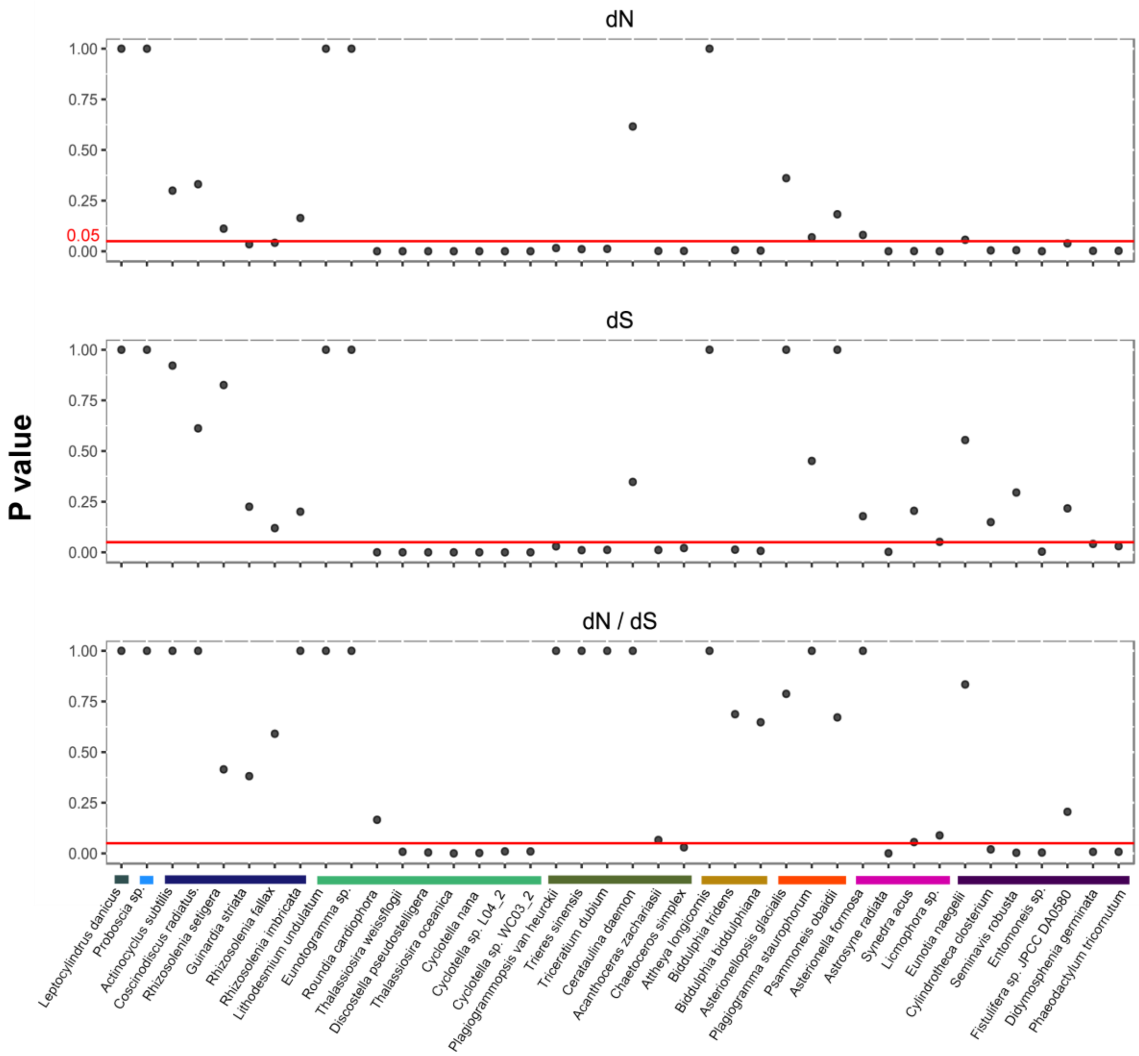
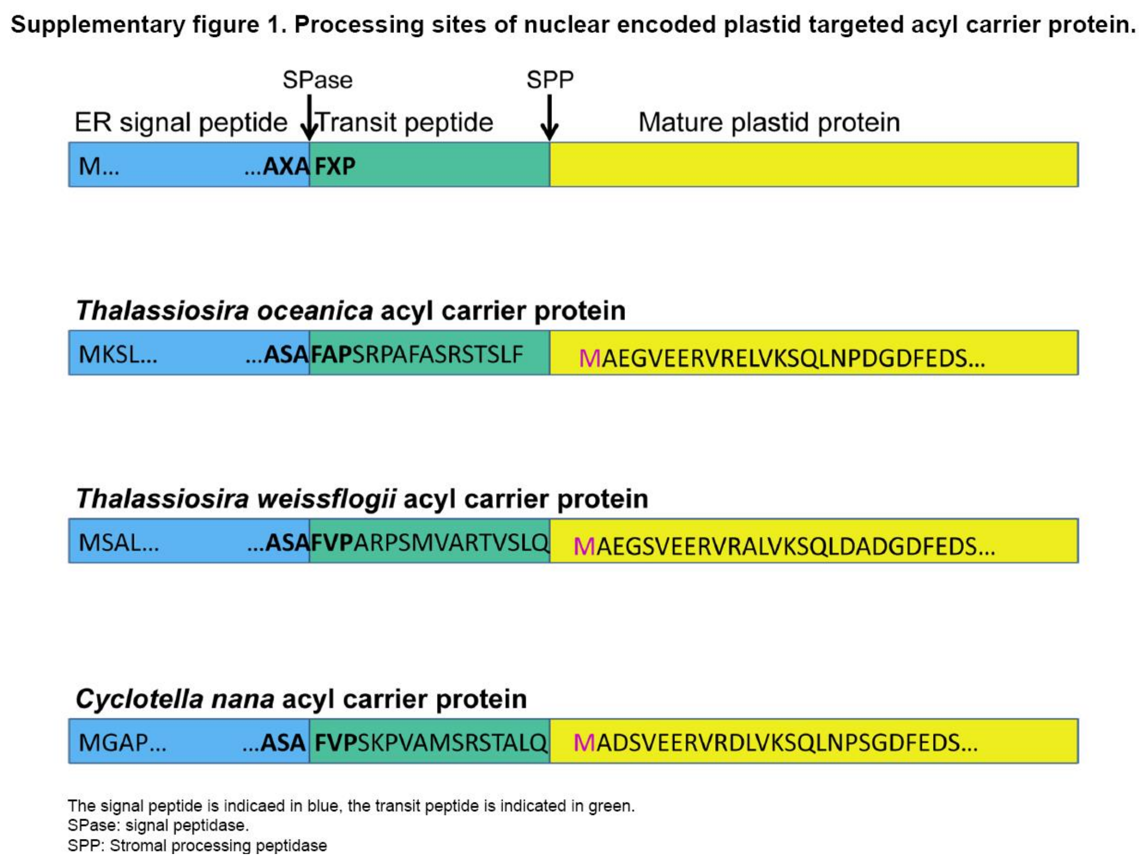



Figure 4.4. P values for pairwise correlation of substitution rate and genome inversion distance in each diatom. Alpha = 0.05 (red horizontal line) was used to access the significance level. The colored bar indicates different clades of diatoms. From left to right: radial 1, radial 2, radial 3, polar 1, polar 2, polar 3, araphid 1, araphid 2, raphid.

# Appendix

## Chapter 2



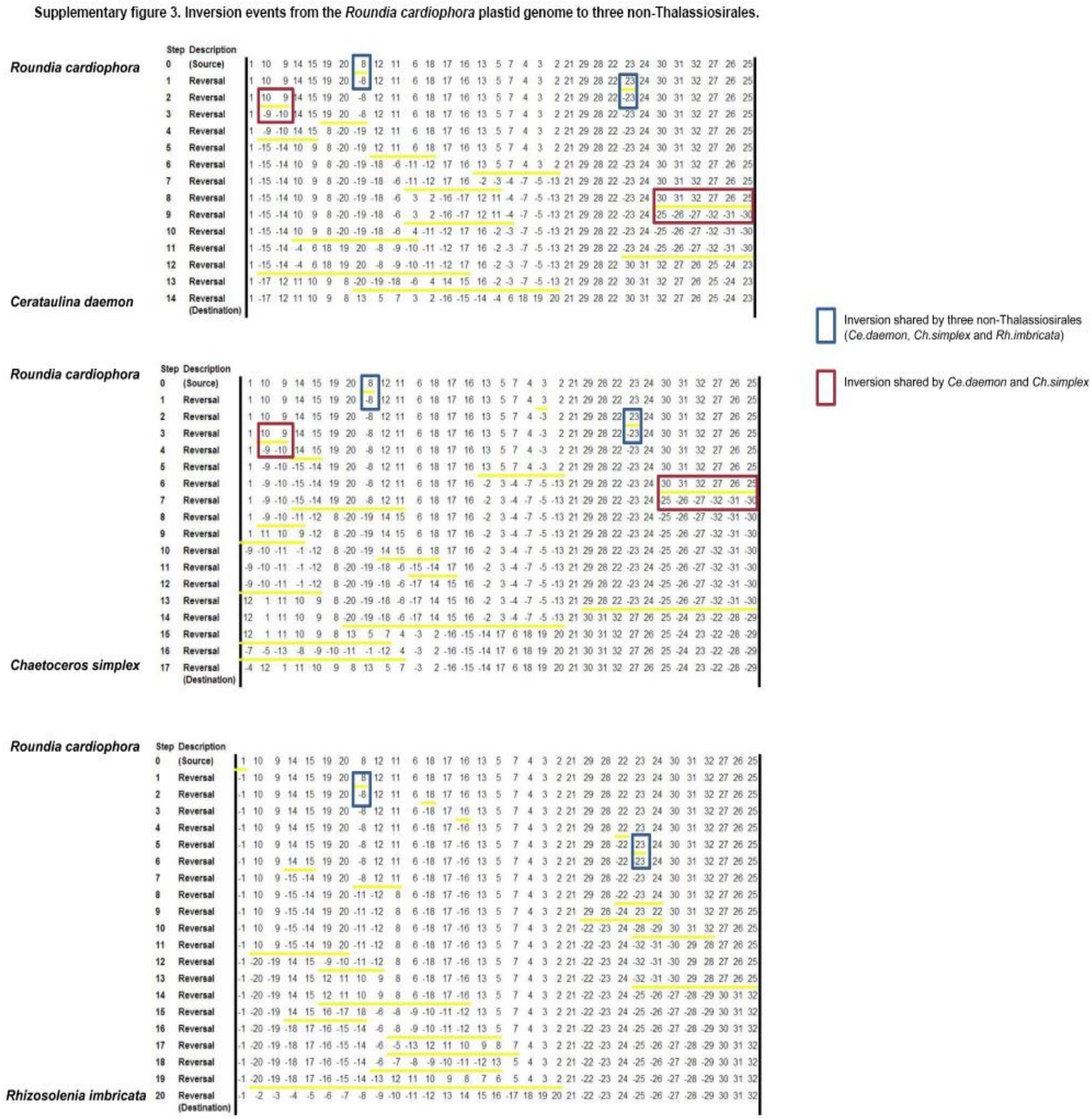

Appendix Figure A.1. Processing sites of nuclear encoded plastid targeted acyl carrier protein. The signal peptide (blue) is removed by signal peptidase (SPase) and the transit peptide (green) is removed by stromal processing peptidase (SPP). The signal peptide and transit peptide junction site show a canonical AXAFXP motif.

**Supplementary figure 2. Inversion events from the *Roundia cardiophora* plastid genome to *Thalassiosira oceanica* plastid genome.**

| Step | Description | |
|---|---|---|
| 0 | *Roundia cardiophora* | 1 10 9 14 15 19 20 8 12 11 6 18 17 16 13 5 7 4 3 2 21 29 28 22 23 24 30 31 32 27 26 25 |
| 1 | Reversal | 1 10 -14 -9 15 19 20 8 12 11 6 18 17 16 13 5 7 4 3 2 21 29 28 22 23 24 30 31 32 27 26 25 |
| 2 | Reversal | 1 10 -14 -9 15 19 20 8 12 11 -18 -6 17 16 13 5 7 4 3 2 21 29 28 22 23 24 30 31 32 27 26 25 |
| 3 | Reversal | 1 10 -14 -9 15 19 20 8 12 11 -18 -6 17 16 13 -7 -5 4 3 2 21 29 28 22 23 24 30 31 32 27 26 25 |
| 4 | Reversal | 1 10 -14 -20 -19 -15 9 8 12 11 -18 -6 17 16 13 -7 -5 4 3 2 21 29 28 22 23 24 30 31 32 27 26 25 |
| 5 | Reversal | 1 10 -14 -20 -19 -15 9 8 12 11 -18 -6 17 16 13 -7 -5 4 3 2 21 29 -24 -23 -22 -28 30 31 32 27 26 25 |
| 6 | Reversal | 1 10 -14 -20 -19 -15 9 8 12 11 -18 -6 17 16 13 -7 -5 4 3 2 21 29 -24 -23 -22 -28 30 -26 -27 -32 -31 25 |
| 7 | Reversal | 1 10 -14 -20 -19 -15 9 8 12 11 -18 -6 17 16 13 -7 -5 4 3 2 21 29 -24 -23 -22 -28 26 -30 -27 -32 -31 25 |
| 8 | Reversal | 1 10 -14 -20 -19 -15 9 8 12 11 -18 -6 17 16 13 -7 -5 4 3 2 21 29 -24 -23 -22 -28 26 -30 -25 31 32 27 |
| 9 | Reversal | 1 10 -14 -20 -19 -15 9 8 12 11 -3 -4 5 7 -13 -16 -17 6 18 2 21 29 -24 -23 -22 -28 26 -30 -25 31 32 27 |
| 10 | Reversal | 1 10 -14 -20 -19 -15 9 17 16 13 -7 -5 4 3 -11 -12 -8 6 18 2 21 29 -24 -23 -22 -28 26 -30 -25 31 32 27 |
| | *Thalassiosira oceanica* | |

Note: Only one IR is included in this analysis

Appendix Figure A.2. Inversion events from the *Roundia cardiophora* plastid genome to *Thalassiosira oceanica* plastid genome.




Appendix Figure A.3. Inversion events from the *Roundia cardiophora* plastid genome to three non-Thalassiosirales.

| Taxon | Source/locality | GenBank Accession |
|---|---|---|
| *Cerataulina daemon* | Atlantic coast, FL, USA<br>Approx. 26.9º N, -80.0º W | KJ958484 |
| *Chaetoceros simplex* | CCMP 200 | KJ958479 |
| *Cyclotella sp. L04_2* | Lake Ohrid, Macedonia | KJ958480 |
| *Cyclotella sp. WC03_2* | Waller Creek, TX, USA<br>30.12 º N, 97.43 º W | KJ958481 |
| *Thalassiosira weissflogii* | CCMP 1336 | KJ958485 |
| *Rhizosolenia imbricata* | Harbor Branch Oceanographic Institute boat dock, FL, USA<br>Approx. 27.5º N, -80.3º W | KJ958482 |
| *Roundia cardiophora* | Achang Reef, Guam, USA<br>13.249º N, 144.697º W | KJ958483 |

Abbreviation: CCMP (National Center for Culture of Marine Phytoplankton)

Appendix Table A.1. Taxa used for plastid genome sequencing with source and GenBank accession numbers.

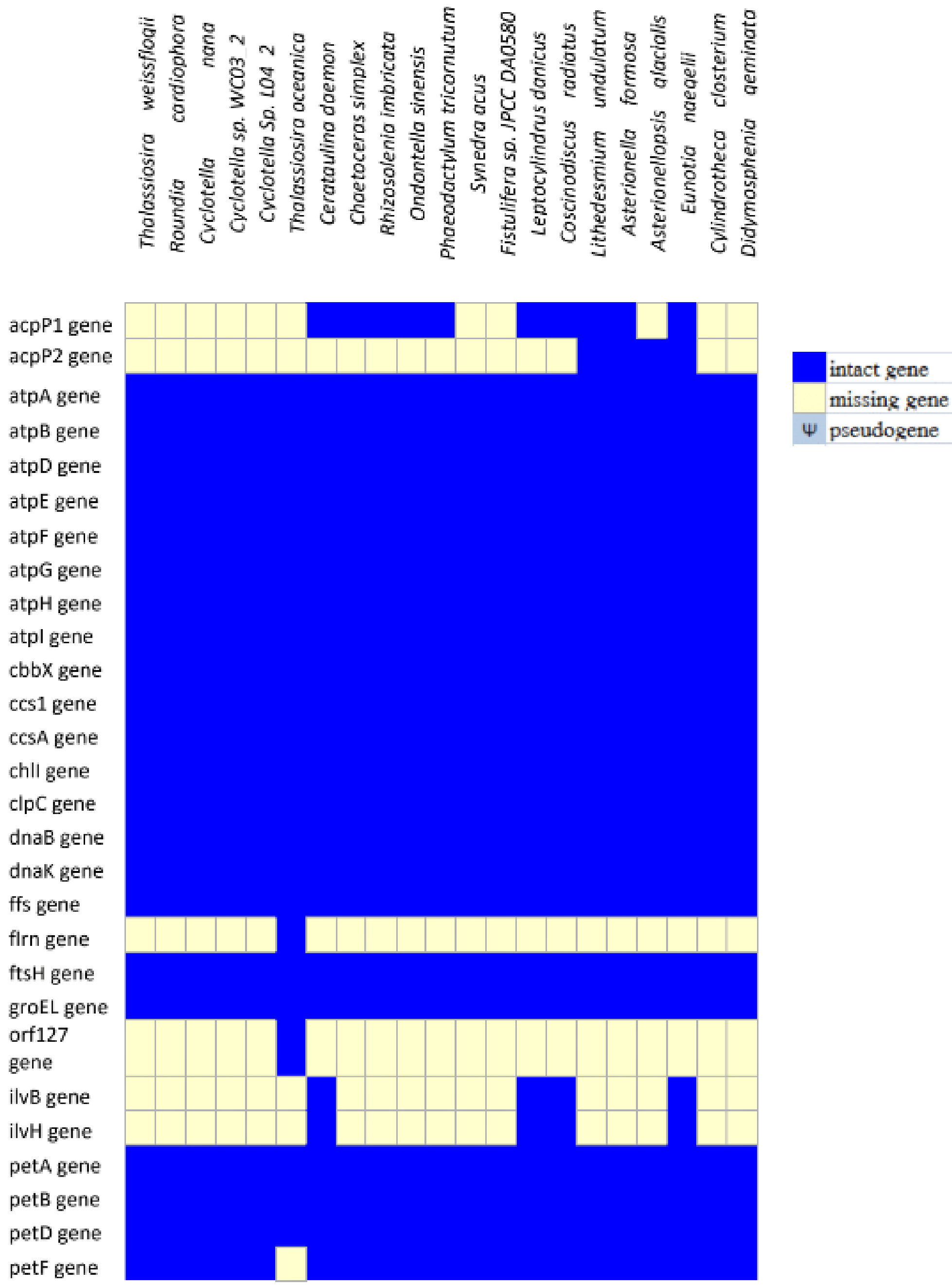
| Primer name | Sequence (5' → 3') |
|---|---|
| *Cerataulina_psaA_trnK_F* | TGA CCT GGT TGT GCC CAT TT |
| *Cerataulina_psaA_trnK_R* | ACC AAA CTG AGC TAT ATC CCG T |
| *Cerataulina_trnP_ycf45_f* | GAA CCT ACG ACA CCC TGG TC |
| *Cerataulina_trnP_ycf45_R* | ACA AGA GAT ATT AAA AAG GCA ACG A |
| *Cerataulina_psaC-psbX_F* | ACG AGT TGT TTC TGC GCC TA |
| *Cerataulina_psaC-psbX_R* | TGC ACC TGT TTT AAT CGC AGC |
| *Cerataulina_psbY_rbcR_F* | TGC ACC TGT TTT AAT CGC AGC |
| *Cerataulina_psbY_rbcR_R* | TCA GCA GCA CGT GTA AAG CT |
| *Cyclotella_L04_2_petG_F* | TCA AAT TGA TTT CCA CGA CGA T |
| *Cyclotella_L04_2_psaI_R* | ACC AAC AAG TGG TAC AAG AA |

Appendix Table A.2. PCR Primers used for finishing diatom plastid genome sequencing and confirming boundaries between inverted repeats and single copy regions.

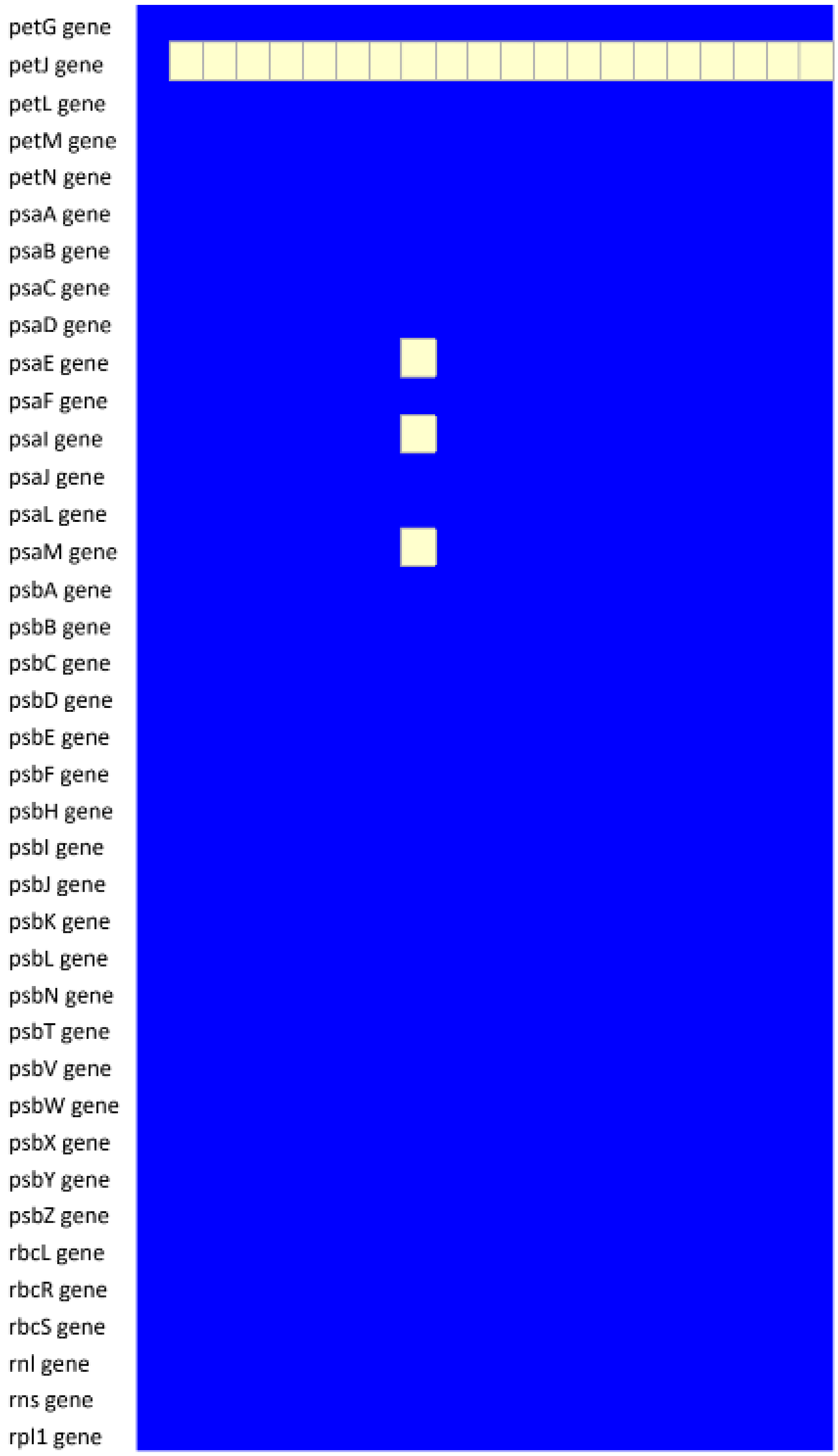
| | *T. weissflogii* | *Cy. sp. L04_2* | *Cy. sp. WC03_2* | *Cy. nana* | *T. oceanica* | *Ro. cardiophora* | *Ch. simplex* | *Ce. daemon* | *Rh. imbricata* |
|---|---|---|---|---|---|---|---|---|---|
| **Size (bp)** | **127,601** | **129,400** | **129,498** | **128,814** | **141,790** | **126,871** | **116,459** | **120,144** | **120,956** |
| SSC | 26,496 | 27,620 | 27,602 | 26,889 | 24,106 | 26,274 | 39,517 | 40,590 | 27,482 |
| LSC | 64,555 | 65,268 | 65,210 | 65,250 | 70,298 | 64,387 | 62,136 | 65,546 | 61,244 |
| IR | 18,276 | 18,256 | 18,261 | 18337 | 23,693 | 18,105 | 7,403 | 7,004 | 16,115 |
| G+C content | 30.8% | 30.3% | 30.0% | 30.7% | 30.4% | 31.0% | 32.1% | 31.2% | 31.8% |
| Protein coding genes | 127 | 127 | 127 | 127 | 126[a] | 126[b] | 128[c] | 130[d] | 122[e] |
| rRNA genes | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| tRNA genes | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 |
| Other RNAs | 2 | 2 | 2 | 2 | 2+flrn | 2 | 2 | 2 | 2 |
| genome coding for genes % | 85.18% | 85.25% | 84.88% | 85.56% | 79.67% | 85.16% | 87.34% | 84.56% | 79.46% |
| Gene density (genes/kb) | 1.41 | 1.39 | 1.39 | 1.38 | 1.30 | 1.42 | 1.45 | 1.41 | 1.41 |
| Average IGS (bp) | 106.08 | 106.06 | 108.76 | 103.31 | 155.82 | 104.57 | 87.27 | 109.79 | 145.30 |
| Overlapping genes | *sufC-sufB*: 1nt<br>*atpF-atpD:* 4nt<br>*psbC-psbD:* 53nt<br>*rpl4-rpl23:* 8nt | *sufC-sufB:* 1nt<br>*atpF-atpD:* 4nt<br>*psbC-psbD:* 53nt<br>*rpl4-rpl23:* 17nt | *sufC-sufB:* 1nt<br>*atpF-atpD:* 4nt<br>*psbC-psbD:* 53nt<br>*rpl4-rpl23:* 17nt | *sufC-sufB*: 1nt<br>*atpF-atpD:* 4nt<br>*psbC-psbD:* 53nt<br>*rpl4-rpl23:* 8nt | *sufC-sufB:* 1nt<br>*atpF-atpD:* 4nt<br>*psbC-psbD*: 53nt<br>*rpl4-rpl23:* 8nt | *sufC-sufB:* 1nt<br>*atpF-atpD*: 4nt<br>*psbC-psbD:* 53nt<br>*rpl4-rpl23:* 8nt | *sufC-sufB:* 1nt<br>*atpF-atpD:* 4nt<br>*psbC-psbD:* 53nt<br>*rpl4-rpl23:* 8nt | *sufC-sufB:* 1nt<br>*atpF-atpD:* 4nt<br>*psbC-psbD:* 53nt<br>*rpl4-rpl23:* 8nt | *sufC-sufB:* 1nt<br>*atpF-atpD:* 1nt<br>*psbC-psbD:* 53nt<br>*rpl4-rpl23:* 8nt |

Abbreviation: *Thalassiosira (T.), Cyclotella (Cy.), Roundia (Ro.), Chaetoceros (Ch.), Cerataulina(Ce.), Rhizosolenia(Rh.)*
a: missing *petF*, has orf127
b: *ycf66* is a pseudogene
c: missing *ycf42*, has *acpP1* and *syfB*
d: missing *ycf42*, has *acpP1* and *syfB, ilvB, ilvH*
e: missing *psaE, psaI*, *psaM*, *ycf35, tufA*, *syfB*, has *acpP1*.

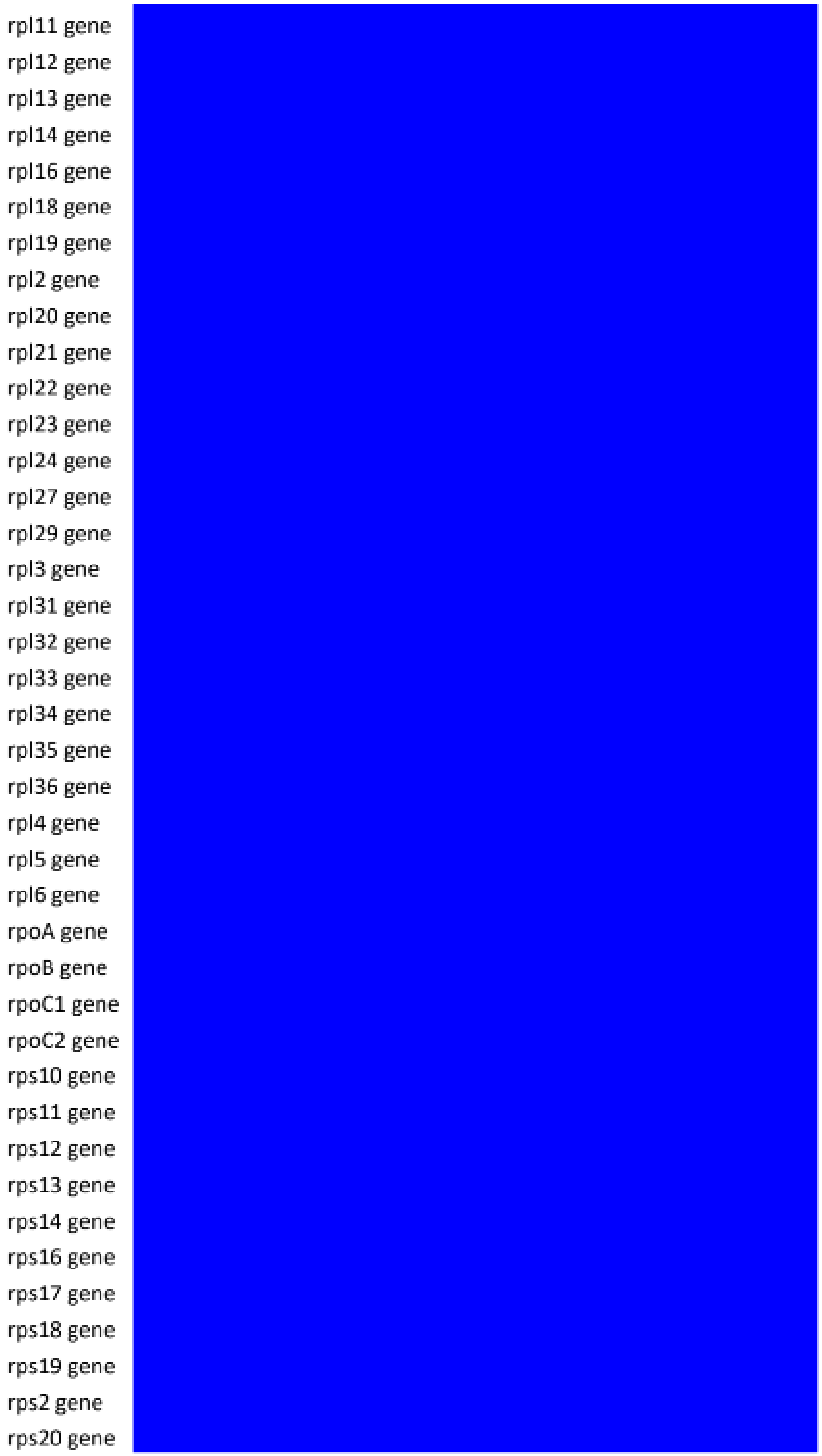
Appendix Table A.3. Plastid genome features of seven sequenced diatoms in comparison with *Cyclotella nana* and *Thalassiosira oceanica*.



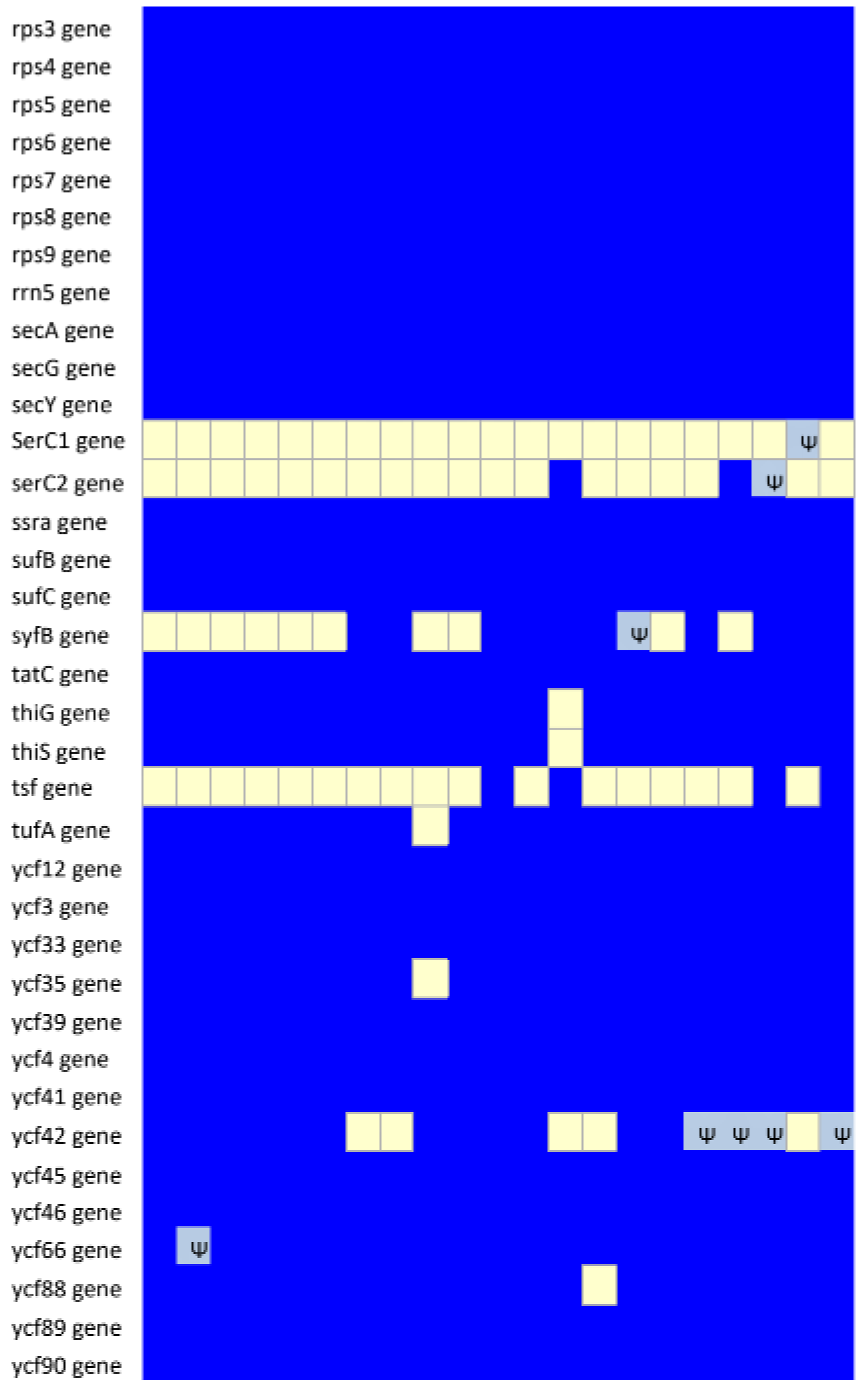

Appendix Table A.4. Gene content comparison of seven sequence diatom plastid genomes with other published diatom plastid genomes. Intact genes are indicated by dark blue, pseudogenes as light blue, and missing genes in light yellow.




Appendix Table A.4. Continued



Appendix Table A.4. Continued




Appendix Table A.4. Continued

| Species | Identity | Alignment length | Number of mismatches | Number of gap opens | Start1 | End1 | Start2 | End2 | E-value | Bit score |
|---|---|---|---|---|---|---|---|---|---|---|
| *Cy. sp. W03_2* | 100 | 84 | 0 | 0 | 65293 | 65376 | 65211 | 65294 | $1e^{-36}$ | 152 |
| *Cy. sp. W03_2* | 100 | 82 | 0 | 0 | 83554 | 83635 | 83472 | 83553 | $2e^{-35}$ | 149 |
| *Cy. sp. L04_2* | 100 | 79 | 0 | 0 | 65268 | 65346 | 65190 | 65268 | $7e^{-34}$ | 143 |
| *T. oceanica* | 96.67 | 90 | 3 | 0 | 29941 | 30030 | 18564 | 18475 | $2e^{-35}$ | 149 |
| *T. oceanica* | 91.25 | 80 | 7 | 0 | 6626 | 6705 | 5376 | 5297 | $1e^{-24}$ | 113 |

Generic abbreviations are: *Cyclotella* (*Cy.*), *Thalassiosira* (*T.*).

Appendix Table A.5. Predicted repeat pairs in seven sequenced diatom plastid genomes.

| | |
|---|---|
| *Rhizosolenia imbricata* | -1 -2 -3 -4 -5 -6 -7 -8 -9 -10 -11 -12 13 14 15 16 -17 18 19 20 21 -22 -23 24 -25 -26 -27 -28 -29 30 31 32 |
| *Chaetoceros simplex* | -4 12 1 11 10 9 8 13 5 7 -3 2 -16 -15 -14 17 6 18 19 20 21 30 31 32 27 26 25 -24 23 -22 -28 -29 |
| *Cerataulina daemon* | 1 -17 12 11 10 9 8 13 5 7 3 2 -16 -15 -14 -4 6 18 19 20 21 29 28 22 30 31 32 27 26 25 -24 23 |
| *Cyclotella nana* | 1 10 9 14 15 19 20 8 12 11 6 18 17 16 13 5 7 4 3 2 21 29 28 22 23 24 30 31 32 27 26 25 |
| *Thalassiosira weissflogii* | 1 10 9 14 15 19 20 8 12 11 6 18 17 16 13 5 7 4 3 2 21 29 28 22 23 24 30 31 32 27 26 25 |
| *Roundia cardiophora* | 1 10 9 14 15 19 20 8 12 11 6 18 17 16 13 5 7 4 3 2 21 29 28 22 23 24 30 31 32 27 26 25 |
| *Cyclotella sp.W03_2* | 1 -19 -15 -14 -9 -10 20 8 12 11 6 18 17 16 13 5 7 4 3 2 21 29 28 22 23 24 30 31 32 27 26 25 |
| *Cyclotella sp. L04_2* | 1 -19 -15 -14 -9 -10 20 8 12 11 6 18 17 16 13 5 7 4 3 2 21 29 28 22 23 24 30 31 32 27 26 25 |
| *Thalassiosira oceanica* | 1 10 -15 -14 -21 -20 -16 9 18 17 -7 -5 4 3 -11 -12 -13 -8 6 19 2 22 31 30 -25 -24 -23 -29 27 -32 -26 33 28 |

Note: Only one IR is included in this analysis.
Highlighted area indicates the one single inversion between *Roundia cardiophora* plastid genome and *Cyclotella* sp. W03_2 and *Cyclotella* sp. L04_2 plastid genomes.

Appendix Table A.6. The permutation of number coded Locally Colinear Block (LCB) for each plastid genome. Negative number indicates an inversion of the given LCB.

| | *T. weissflogii* | *Cy. sp. L04_2* | *Cy. sp. WC03_2* | *Cy. nana* | *T. oceanica* | *Ro. cardiophora* | *Ch. simplex* | *Ce. daemon* | *Rh. imbricata* |
|---|---|---|---|---|---|---|---|---|---|
| ***T. weissflogii*** | | | | | | | | | |
| ***Cy. sp. L04_2*** | 1 | | | | | | | | |
| ***Cy. sp.WC03_2*** | 1 | 0 | | | | | | | |
| ***Cy. nana*** | 0 | 1 | 1 | | | | | | |
| ***T. oceanica*** | 10 | 11 | 11 | 10 | | | | | |
| ***Ro. cardiophora*** | 0 | 1 | 1 | 0 | 10 | | | | |
| ***Ch. simplex*** | 17 | 18 | 18 | 17 | 22 | 17 | | | |
| ***Ce. daemon*** | 14 | 15 | 15 | 14 | 19 | 14 | 8 | | |
| ***Rh. imbricata*** | 20 | 21 | 21 | 20 | 25 | 20 | 14 | 12 | |

Abbreviation: *Thalassiosira (T.), Cyclotella (Cy.), Roundia (Ro.), Chaetoceros (Ch.), Cerataulina(Ce.), Rhizosolenia(Rh.).* The zero inversion in yellow indicates those three plastid genome *Cy. nana*, *T. weissflogii* and *Ro. cardiophora* have the same gene order.

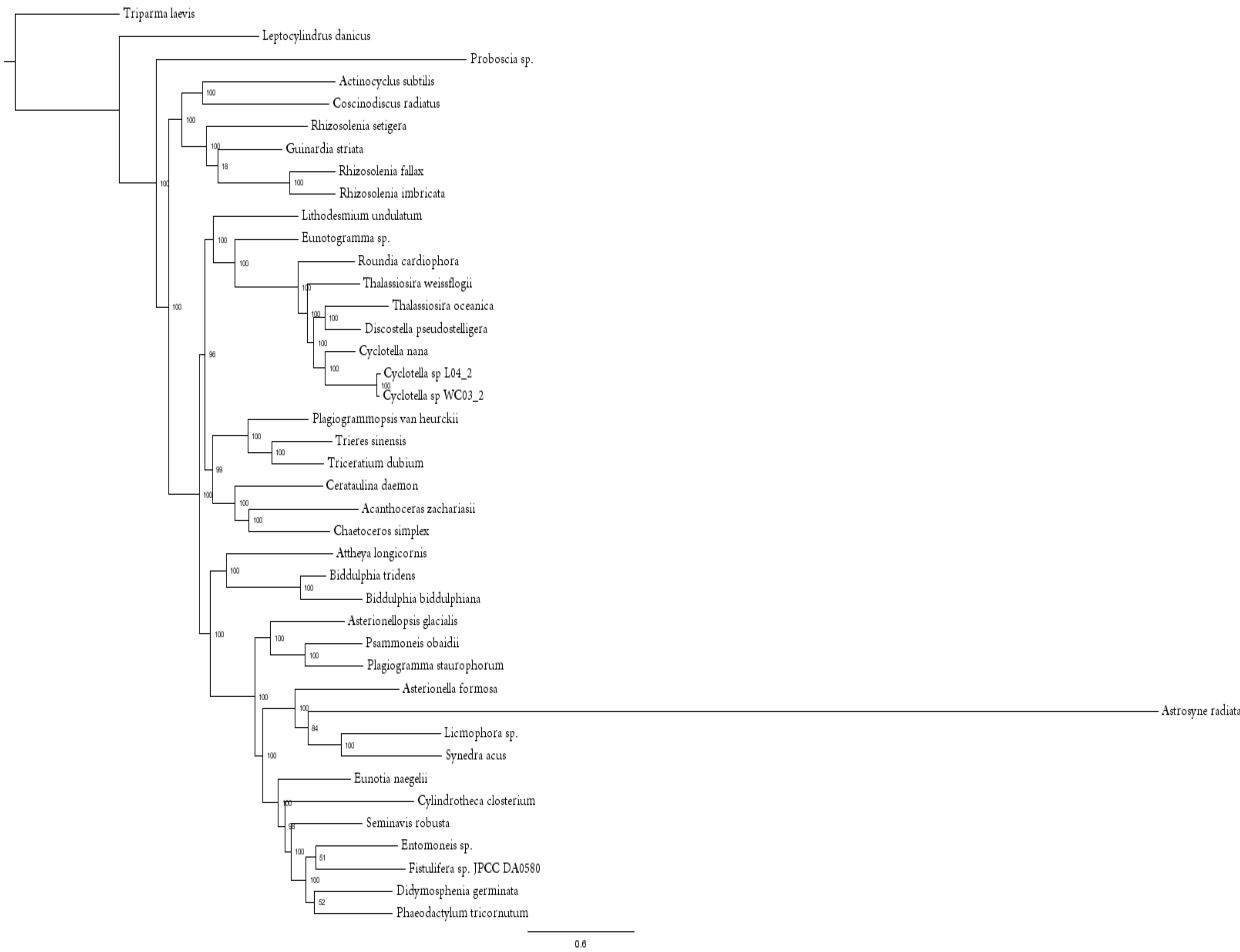
Appendix Table A.7. Pairwise number of inversions inferred by GRIMM.

| LCB number | Genes names |
|---|---|
| 1 | *psaA, psaB* |
| 2 | *psaF, psaJ* |
| 3 | *ycf90, psbI* |
| 4 | *petB, psaD* |
| 5 | *psbD, psbC* |
| 6 | *secG, psaM* |
| 7 | *ycf12, psbZ* |
| 8 | *dnaB, rpl12* |
| 9 | *psbX, psbV* |
| 10 | *rpl19, ssra* |
| 11 | *petA, ycf3* |
| 12 | *rps18, rps2* |
| 13 | *psbK, psaI* |
| 14 | *rbcS, atpA* |
| 15 | *psbB, psbH* |
| 16 | *petN, ycf33* |
| 17 | *petG* |
| 18 | *rps14, ftsH* |
| 19 | *psaE, rpl20* |
| 20 | *ycf45, acpP* |
| 21 | *ycf89, rrn5* |
| 22 | *psbA* |
| 23 | *psaC* |
| 24 | *ccsA* |
| 25 | *rps6, thiG* |
| 26 | *clpC* |
| 27 | *rps10, rps12* |
| 28 | *ccs1, ycf46* |
| 29 | *rpl34, rpl32* |
| 30 | *rps16, groEL* |
| 31 | *dnaK, rpl16* |
| 32 | *rpl18, rpl31* |

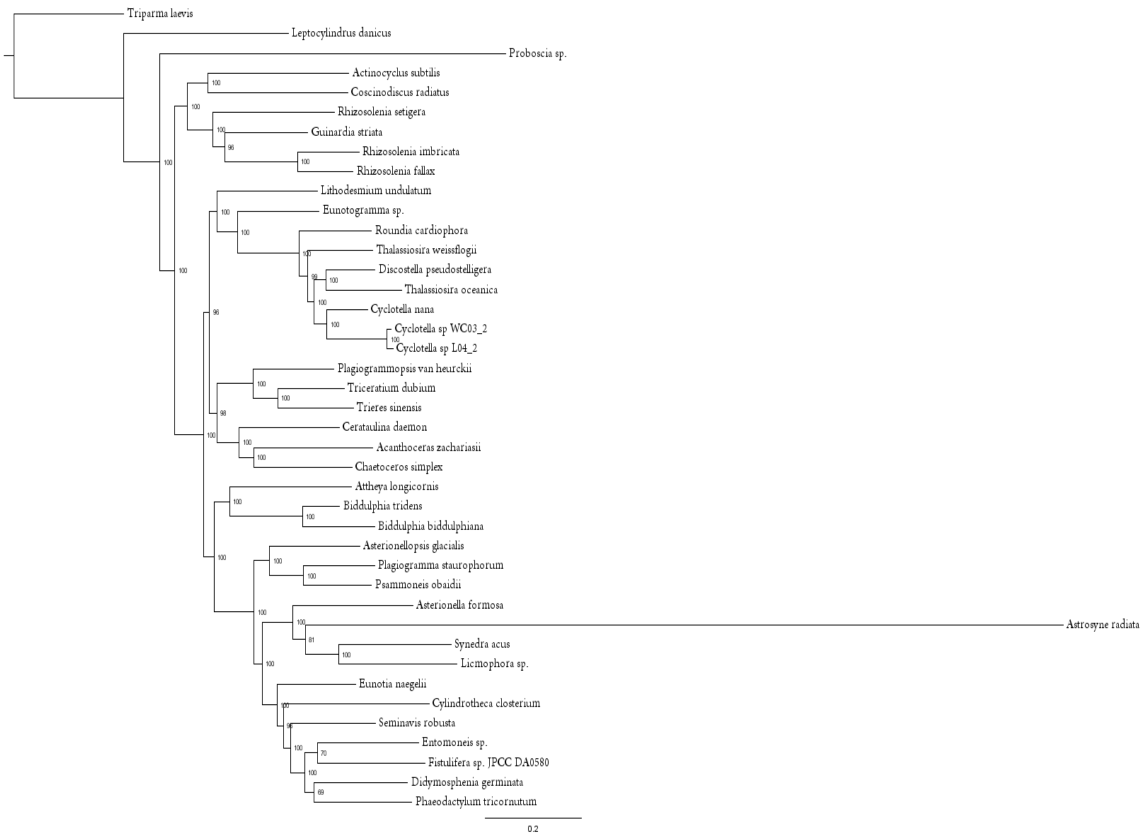
Appendix Table A.8. Genes at the boundary of each Locally Colinear Block (LCB).

## Chapter 3

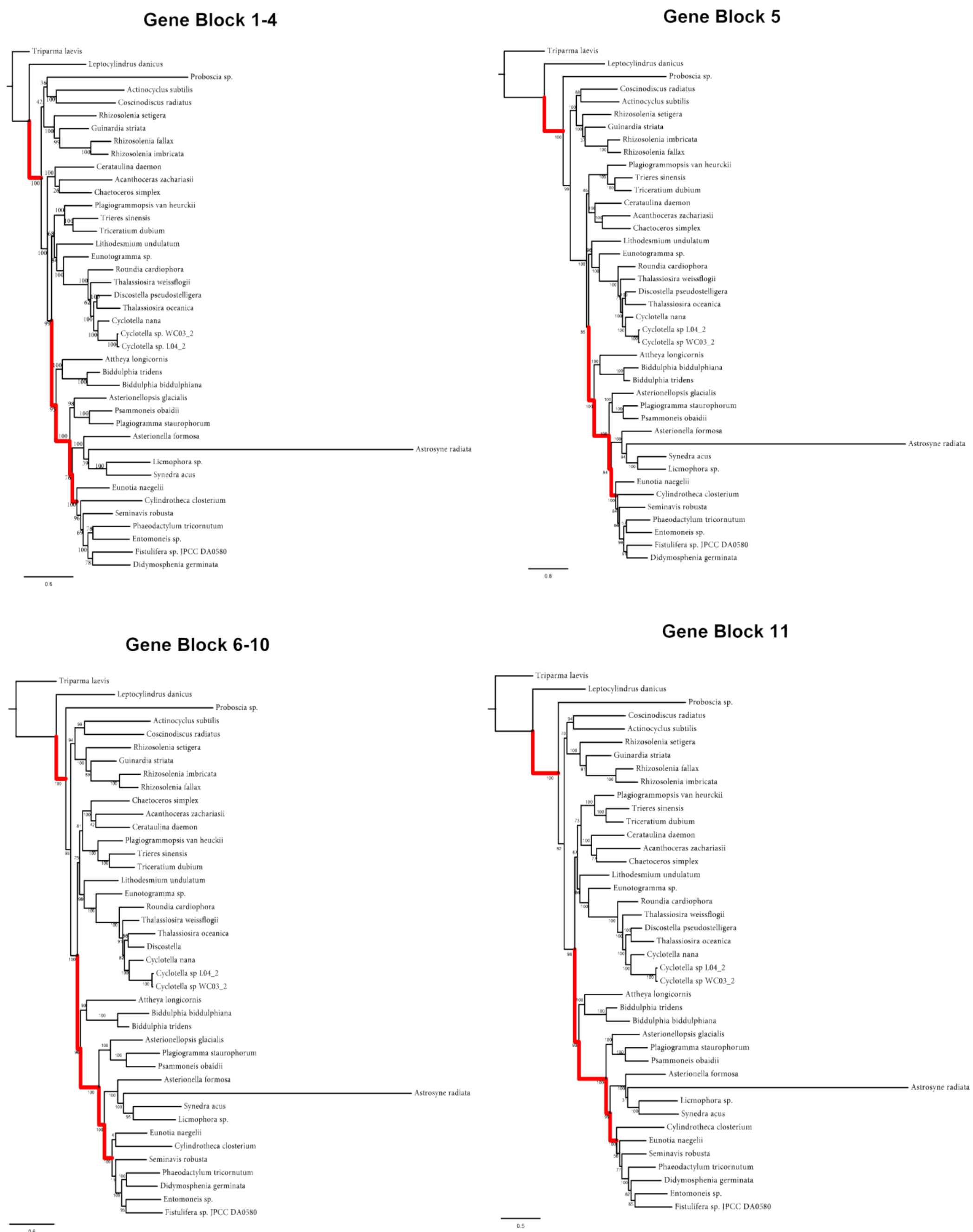

Appendix Figure B.1. Maximum likelihood tree from analysis of 103 shared plastid genes with no partition.



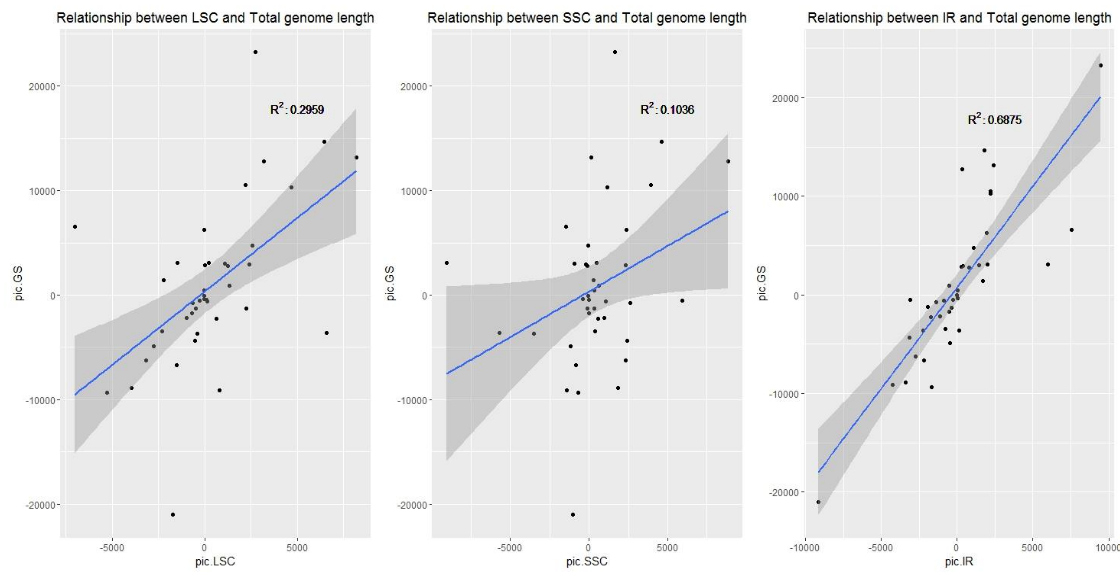

Appendix Figure B.2. Maximum likelihood tree from analysis of 103 shared plastid genes with codon partition.



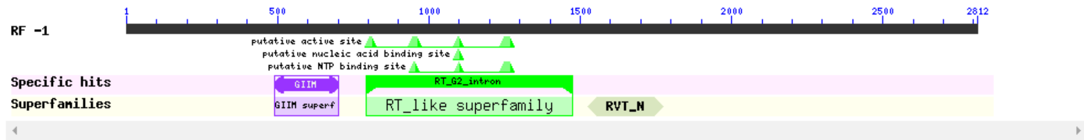

Appendix Figure B.3. Maximum likelihood tree from analysis of 103 shared plastid genes with gene category partition.




Appendix Figure B.4. Comparison of maximum likelihood tree constructed from 4 different gene blocks with codon partition. The 5 branches in red represent the consistent branches separating Radial 1 from the rest of clades, separating Polar 2 from Polar 3 and the Pennate, separating Polar 3 from the Pennate, separating Araphid1 from Araphid 2 and Raphid, separating Araphid 2 from Raphid, respectively. The branches in red are consistent with the corresponding branches with arrow in Figure 3.1.



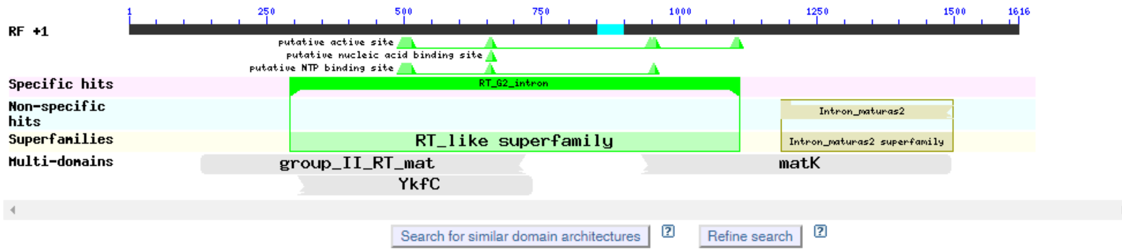

Appendix Figure B.5. Relationship between total genome size and LSC, SSC and IR respectively after applying phylogenetic independent contrast analysis. The blue line indicates the regression line. The shaded area indicates 95% of confidence interval. The coefficient of determination is indicated by R squared.




**List of domain hits**

| Name | Accession | Description | Interval | E-value |
|---|---|---|---|---|
| [+] RT_G2_intron | cd01651 | RT_G2_intron: Reverse transcriptases (RTs) with group II intron origin. RT transcribes DNA ... | 794-1474 | 5.24e-44 |
| [+] RVT_N super family | cl16337 | N-terminal domain of reverse transcriptase; This domain is found at the N-terminus of ... | 1526-1774 | 4.52e-22 |
| [+] GIIM | pfam08388 | Group II intron, maturase-specific domain; This region is found mainly in various bacterial ... | 488-700 | 3.25e-09 |

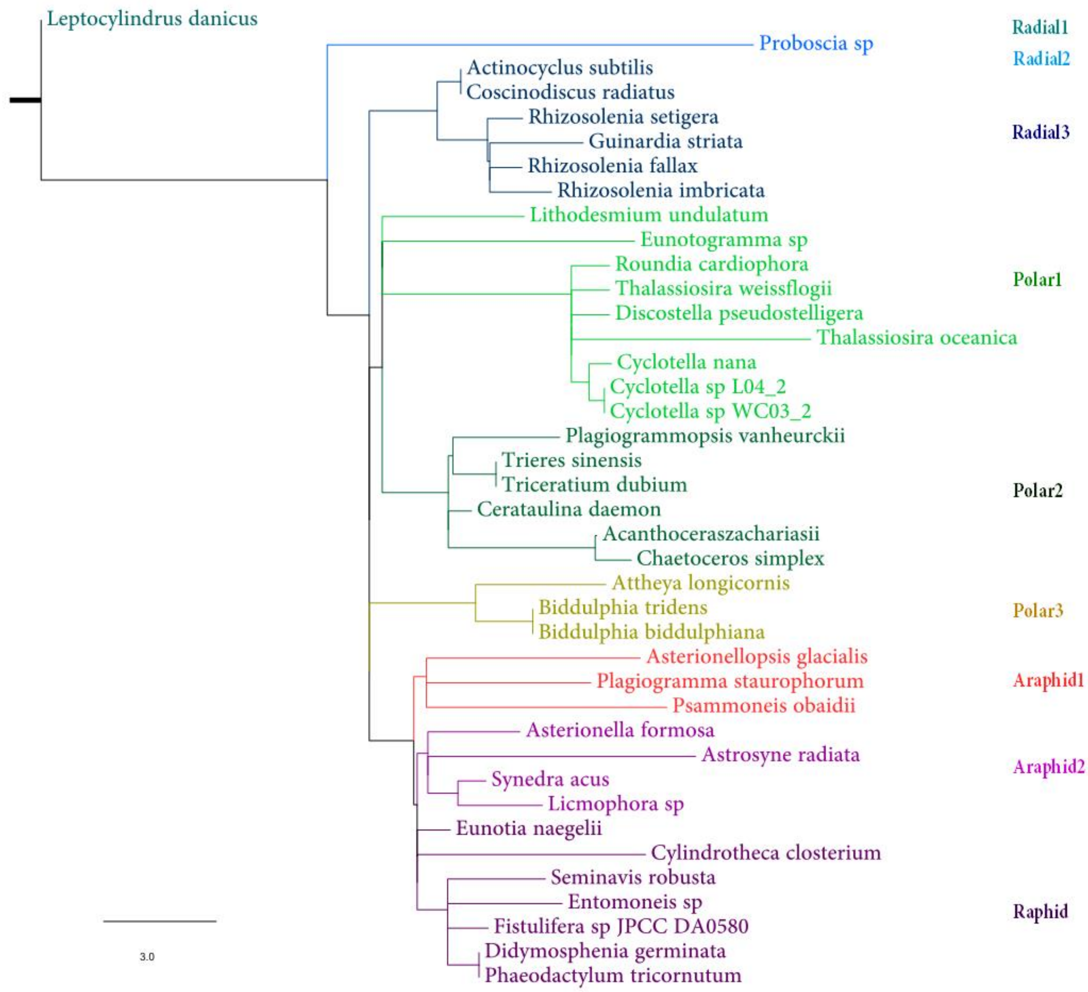
Appendix Figure B.6. Conserved domain search result of *atpB* group II intron in *Proboscia sp.*




**List of domain hits**

| Name | Accession | Description | Interval | E-value |
|---|---|---|---|---|
| [+] RT_G2_intron | cd01651 | RT_G2_intron: Reverse transcriptases (RTs) with group II intron origin. RT transcribes DNA ... | 292-1110 | 7.98e-40 |
| [+] Intron_maturas2 | pfam01348 | Type II intron maturase; Group II introns use intron-encoded reverse transcriptase, maturase ... | 1186-1497 | 1.59e-08 |
| [+] group_II_RT_mat | TIGR04416 | group II intron reverse transcriptase/maturase; Members of this protein family are ... | 130-720 | 1.30e-30 |
| [+] matK | CHL00002 | maturase K | 931-1494 | 2.68e-12 |
| [+] YkfC | COG3344 | Retron-type reverse transcriptase [Mobilome: prophages, transposons]; | 307-732 | 3.05e-08 |

Appendix Figure B.7. Conserved domain search result of *petD* group II intron in *Plagiogramma staurophorum.*




Appendix Figure B.8. The gene order tree constructed using gene order inversion distance and 103 protein coding genes as constraint. Different colors indicate different diatom groups.

| Taxon | Source/locality | Culture condition |
|---|---|---|
| *Acanthoceras zachariasii* | Lake Okoboji, Iowa, USA | 20-24ºC, 0 ppt, WC |
| *Actinocyclus subtilis* | University of Guam Marine Lab outflows, Guam, USA | 27ºC, 32 ppt, f/2 |
| *Astrosyne radiata* | Gab Gab Beach, Guam, USA | 27ºC, 32 ppt, f/2 |
| *Attheya longicornis* | CCMP 214 | 4ºC, 32 ppt, f/2 |
| *Biddulphia biddulphiana* | Gab Gab Beach, Guam, USA | 27ºC, 32 ppt, f/2 |
| *Biddulphia tridens* | Long Beach, California, USA | 20-24ºC, 32 ppt, f/2 |
| *Discostella pseudostelligera* | Upper Bull Shoals Lake, Missouri, USA | 20-24ºC, 0 ppt, WC |
| *Entomoneis sp.* | Jeddah, Saudi Arabia | 27ºC, 40 ppt, f/2 |
| *Eunotogramma sp.* | Atlantic Coast, South Florida, USA | 20-24ºC, 32 ppt, f/2 |
| *Guinardia striata* | Port O'Connor, Texas, USA | 20-24ºC, 32 ppt, f/2 |
| *Licmophora sp.* | Duba, Saudi Arabia | 27ºC, 40 ppt, f/2 |
| *Plagiogramma staurophorum* | Taelayag Beach, Guam, USA | 27ºC, 32 ppt, f/2 |
| *Plagiogrammopsis van heurckii* | Moss Landing, California, USA | 14ºC, 32 ppt, f/2 |
| *Proboscia sp.* | Duba, Saudi Arabia | 27ºC, 40 ppt, f/2 |
| *Psammoneis obaidii* | Markaz Al Shoaibah, Saudi Arabia | 27ºC, 40 ppt, f/2 |
| *Rhizosolenia fallax* | Duba, Saudi Arabia | 27ºC, 40 ppt, f/2 |
| *Rhizosolenia setigera* | Lady's Island, South Carolina, USA | 20-24ºC, 32 ppt, f/2 |
| *Triceratium dubium* | Al-Wajh, Saudi Arabia | 27ºC, 40 ppt, f/2 |

Appendix Table B.1. Taxa used for plastid genome sequencing with source.

| **Category** | **Genes** |
|---|---|
| Photosystem | *psaA, psaB, psaD, psaF, psaJ, psaL, psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI , psbJ, psbK, psbL, psbN, psbT, psbV, psbX, psbY, psbZ* |
| Cytochrome b/f complex | *petA, petB, petD, petG, petL, petM, petN* |
| ATP synthase | *atpA, atpB, atpD, atpE, atpF, atpG, atpH, atpI* |
| RubisCo subunit | *rbcL, rbcS, rbcR* |
| RNA polymerase | *rpoA, rpoB, rpoC1, rpoC2* |
| Ribosomal proteins | *rpl1, rpl2, rpl3, rpl4, rpl5, rpl6, rpl11, rpl12, rpl13, rpl14, rpl16, rpl18, rpl19, rpl20, rpl21, rpl22, rpl23, rpl24, rpl27, rpl29, rpl31, rpl32, rpl33, rpl34, rpl35, rps2, rps3, rps4, rps5, rps7, rps9, rps10, rps11, rps12, rps13, rps14, rps16, rps17, rps18, rps19, rps20* |
| Other genes | *cbbX, ccs1, ccsA, chlI, clpC, dnaB, ftsH, groEL, secA, secG, secY, sufB, sufC, tatC, ycf3, ycf12, ycf46* |

Appendix Table B.2. 103 shared protein coding genes partitioned by functional groups.

| Gene Block | Genes |
|---|---|
| 1 | *sufC, sufB, rbcL, rbcS* |
| 2 | *atpI, atpH, atpG, atpF, atpD, atpA* |
| 3 | *secG, psaD, petB, petD* |
| 4 | *rpl12, rpl1, rpl11, dnaB* |
| 5 | *petA, tatC, atpE, atpB, ycf3,rps18,rpl33,rps20,rpoB, rpoC1, rpoC2* |
| 6 | *psbD, psbC* |
| 7 | *psbB, psbT, psbN, psbH* |
| 8 | *psbJ, psbL, psbF, psbE* |
| 9 | *Rpl34, secA, rpl27, rpl21, rbcR* |
| 10 | *dnaK, rpl3, rpl4, rpl23, rpl2, rps19* |
| 11 | *rpl22 rps3,rpl16,rpl29,rps17,rpl14,rpl24,rpl5,rps8,rpl6,rpl18,rps5,secY, rpl36,rps13,rps11,rpoA,rpl13,rps9,rpl31,rps12,rps7,tufA,rps10* |

Appendix Table B.3. Genes in conserved gene order blocks among most of diatom plastid genomes.

| Name | LSC | SSC | IR | Total | Clade |
|---|---|---|---|---|---|
| Acanthoceras zachariasii | 63924 | 39368 | 8550 | 120392 | Polar2 |
| Actinocyclus subtilis | 59040 | 38042 | 11019 | 119120 | Radial3 |
| Asterionella formosa | 62681 | 40193 | 9182 | 121238 | Araphid2 |
| Asterionellopsis glacialis | 72585 | 51181 | 11129 | 146024 | Araphid1 |
| Astrosyne radiata | 50213 | 37953 | 21433 | 131032 | Araphid2 |
| Attheya longicornis | 65290 | 44231 | 10022 | 129565 | Polar1 |
| Biddulphia biddulphiana | 63612 | 40024 | 9246 | 122128 | Polar1 |
| Biddulphia tridens | 66995 | 39752 | 9774 | 126295 | Polar1 |
| Cerataulina daemon | 65546 | 40590 | 7004 | 120144 | Polar2 |
| Chaetoceros simplex | 62136 | 39517 | 7403 | 116459 | Polar2 |
| Coscinodiscus radiatus | 60402 | 36643 | 12584 | 122213 | Radial3 |
| Cyclotella sp. WC03_2 | 65292 | 27684 | 18261 | 129498 | Polar1 |
| Cyclotella nana | 65250 | 26889 | 18338 | 128814 | Polar1 |
| Cyclotella sp. L04_2 | 65268 | 27620 | 18256 | 129400 | Polar1 |
| Cylindrotheca closterium | 86398 | 49671 | 14870 | 165809 | Raphid |
| Didymosphenia germinata | 63610 | 40370 | 6996 | 117972 | Raphid |
| Discostella pseudostelligera | 64734 | 26735 | 18896 | 129261 | Polar1 |
| Entomoneis sp. | 64114 | 43246 | 7348 | 122056 | Raphid |
| Eunotia naegelii | 73679 | 24857 | 27185 | 152906 | Raphid |
| Eunotogramma sp. | 84201 | 39912 | 24102 | 172317 | Polar1 |
| Fistulifera sp. JPCC DA0580 | 62994 | 45264 | 13330 | 134918 | Raphid |
| Guinardia striata | 59711 | 38870 | 11782 | 122145 | Radial3 |
| Leptocylindrus danicus | 66724 | 40981 | 8754 | 125213 | Radia1 |
| Licmophora sp. | 64999 | 40389 | 7898 | 121184 | Araphid2 |
| Lithodesmium undulatum | 61086 | 37854 | 11860 | 122660 | Polar1 |
| Phaeodactylum tricornutum | 63674 | 39871 | 6912 | 117369 | Raphid |
| Plagiogramma staurophorum | 77767 | 54273 | 34888 | 201816 | Araphid1 |
| Plagiogrammopsis van heurckii | 74042 | 41125 | 12069 | 139305 | Polar1 |
| Proboscia sp. | 57631 | 39450 | 20584 | 138249 | Radia2 |
| Psammoneis obaidii | 73911 | 51965 | 21523 | 168922 | Araphid1 |
| Rhizosolenia fallax | 59165 | 28184 | 18967 | 125283 | Radial3 |
| Rhizosolenia imbricata | 61244 | 27482 | 16115 | 120956 | Radial3 |
| Rhizosolenia setigera | 58541 | 38332 | 12069 | 121011 | Radial3 |
| Roundia cardiophora | 64387 | 26274 | 18105 | 126871 | Polar1 |
| Seminavis robusta | 70540 | 61497 | 9434 | 150905 | Raphid |
| Synedra acus | 61724 | 40937 | 6795 | 116251 | Araphid2 |
| Thalassiosira oceanica | 70298 | 24106 | 23693 | 141790 | Polar1 |
| Thalassiosira weissflogii | 64555 | 26494 | 18276 | 127601 | Polar1 |
| Triceratium dubium | 65233 | 38936 | 8106 | 120381 | Polar1 |
| Trieres sinensis | 65346 | 38908 | 7725 | 119704 | Polar1 |
| Triparma laevis | 91600 | 15946 | 4984 | 117514 | outgroup |

Appendix Table B.4. Genome size comparison of forty diatom plastid genomes together with the outgroup species *Tripama*

| | Triparma laevis | Leptocylindrus danicus | Proboscia sp. | Actinocyclus subtilis | Coscinodiscus radiatus | Rhizosolenia setigera | Guinardia striata | Rhizosolenia fallax | Rhizosolenia imbricata | Lithedesmium undulatum | Eunotogramma sp. | Roundia cardiopohra | Thalassiosira weissflogii | Thalassiosira oceanica | Discostella pseudostelligera | Cyclotella nana | Cyclotella sp. WC03_2 | Cyclotella Sp. L04_2 | Plagiogrammopsis van heurckii | Trieres sinensis | Triceratium dubium | Cerataulina daemon | Acanthoceras zachariasii | Chaetoceros simplex | Attheya longicornis | Biddulphia tridens | Biddulphia biddulphiana | Asterionellopsis glacialis | Plagiogramma staurophorum | Psammoneis obaidii | Asterionella formosa | Astrosyne radiata | Licmophora sp. | Synedra acus | Eunotia naegelii | Cylindrotheca closterium | Seminavis robusta | Phaeodactylum tricornutum | Didymosphenia geminata | Entomoneis sp. | Fistulifera sp. JPCC DA0580 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| acpP1 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| acpP2 gene | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| atpA gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| atpB gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| atpD gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| atpE gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| atpF gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| atpG gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| atpH gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| atpI gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| cbbX gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ccs1 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ccsA gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| chlB gene | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| chlI gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| chlL gene | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| chlN gene | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| clpC gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| dnaB gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| dnaK gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| flrn gene | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ftsH gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| groEL gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ilvB gene | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| ilvH gene | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| petA gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| petB gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

1 intact gene
0 missing gene
Ψ pseudogene

Intact genes are indicated by dark blue, pseudogenes as light blue, and missing genes in light yellow

Appendix Table B.5. Gene content comparison of forty diatom plastid genomes together with the outgroup species *Triparma laevis*..

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| petD gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| petF gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| petG gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| petL gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| petM gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| petN gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psaA gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psaB gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psaC gene | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psaD gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psaE gene | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psaF gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psaI gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psaJ gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psaL gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psaM gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbA gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbA2 gene | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| psbB gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbC gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbD gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbE gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbF gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbH gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbI gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbJ gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbK gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbL gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbN gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbT gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbV gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbW gene | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbX gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbY gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| psbZ gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rbcL gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rbcR gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rbcS gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl1 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Appendix Table B.5. Continued

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rpl11 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl12 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl13 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl14 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl16 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl18 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl19 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl2 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl20 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl21 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl22 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl23 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl24 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl27 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl29 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl3 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl31 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl32 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl33 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl34 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl35 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl36 gene | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl4 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl5 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpl6 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpoA gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpoB gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpoC1 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpoC2 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rpoZ gene | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| rps10 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps11 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps12 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps13 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps14 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps16 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps17 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps18 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps19 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Appendix Table B.5. Continued

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rps2 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps20 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps3 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps4 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps5 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps6 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps7 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps8 gene | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rps9 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| secA gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| secG gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| secY gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| SerC1 gene | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | Ψ | 0 | 0 | 0 | 0 | 0 |
| SerC2 gene | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | Ψ | 0 | 1 | 0 | 0 | 0 | 1 |
| sufB gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| sufC gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| syfB gene | 1 | 1 | 0 | 1 | Ψ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| tatC gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| thiG gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| thiS gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| tsf gene | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| tufA gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf12 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf3 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf33 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf35 gene | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf39 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf4 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf41 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf42 gene | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | Ψ | 1 | 1 | Ψ | 0 | 1 | 1 | Ψ | 0 | 1 | 1 | Ψ | 1 | 0 |
| ycf45 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf46 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf66 gene | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Ψ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf88 gene | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf89 gene | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ycf90 gene | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Appendix Table B.5. Continued

| Species | Gene Collinear Block Order |
|---|---|
| Leptocylindrus danicus | 27 26 -28 -11 -14 -12 -22 -18 -17 -5 -6 -7 -8 -9 -10 1 2 3 4 -16 -15 -23 21 20 19 13 25 24 42 39 33 34 35 36 37 32 -31 -30 -29 -38 -40 -41 |
| Probosica sp | -1 2 3 4 -5 -6 -7 -8 -9 -10 11 -12 -13 -14 15 16 17 18 -19 -20 -21 22 -23 -24 25 -26 -27 -28 29 30 31 -32 33 34 35 36 37 -38 -39 -40 -41 -42 |
| Actinocyclus subtilis | 2 3 4 -13 -19 -20 -21 -1 -11 -14 -12 -22 -18 -17 -5 -6 -7 -8 -9 -10 23 15 16 28 27 26 25 24 42 41 40 39 38 29 30 31 33 34 35 36 37 32 |
| Coscinodiscus radiatus | 2 3 4 -13 -19 -20 -21 -1 -11 -14 -12 -22 -18 -17 -5 -6 -7 -8 -9 -10 23 15 16 28 27 26 25 24 42 41 40 39 38 29 30 31 33 34 35 36 37 32 |
| Rhizosolenia setigera | 2 3 4 -13 -19 -20 -21 -1 -11 -14 -12 -22 -18 -17 -5 -6 -7 -8 -9 -10 23 15 16 28 27 26 25 24 42 41 -32 -37 -36 -35 -34 -33 29 30 31 -38 -39 -40 |
| Guinardia striata | 2 3 4 -13 23 15 16 28 27 26 10 9 8 7 6 5 17 18 22 12 14 11 1 21 20 19 -41 -42 -24 -25 -32 -37 -36 -35 -34 -33 29 30 31 -38 -39 -40 |
| Rhizosolenia fallax | 2 3 4 -13 -19 -20 -21 -1 -11 -14 -12 -22 -18 -17 -5 -6 -7 -8 -9 -10 -26 -27 -28 -16 -15 -23 25 24 42 41 -32 -37 -36 -35 -34 -33 29 30 31 -38 -39 -40 |
| Rhizosolenia imbricata | -4 -3 -2 -13 -19 -20 -21 -1 -11 -14 -12 -22 -18 -17 -5 -6 -7 -8 -9 -10 -26 -27 -28 -16 -15 -23 25 24 42 41 -32 -37 -36 40 39 38 29 30 31 33 34 35 |
| Lithodesmium undulatum | 2 3 4 -13 -19 -20 -21 -1 -11 18 22 12 14 -17 -5 -6 -7 -8 -9 -10 23 15 16 28 27 26 25 29 30 31 33 34 35 36 37 32 -38 -39 -40 -41 -42 -24 |
| Eunotogramma sp | 22 12 14 11 16 28 27 26 2 3 4 -13 -19 -20 -21 -18 -17 -5 -6 -7 -8 -9 -10 15 -23 1 25 24 42 41 40 39 38 30 31 33 34 35 36 37 32 -29 |
| Roundia cardiophora | 2 3 4 17 18 22 -26 -27 -28 -16 12 14 10 9 8 7 6 5 11 23 -15 1 21 20 19 13 25 -38 -39 -40 -41 -42 -24 29 30 31 33 34 35 36 37 32 |
| Thalassiosira weissflogii | 2 3 4 17 18 22 -26 -27 -28 -16 12 14 10 9 8 7 6 5 11 23 -15 1 21 20 19 13 25 -38 -39 -40 -41 -42 -24 29 30 31 33 34 35 36 37 32 |
| Discostella pseudostelligera | 2 3 4 17 18 22 -26 -27 -28 -16 12 14 10 9 8 7 6 5 11 23 -15 1 21 20 19 13 25 -38 -39 -40 -41 -42 -24 29 30 31 33 34 35 36 37 32 |
| Thalassiosira oceania | 2 3 4 17 18 28 27 26 16 22 23 -15 -1 21 20 19 -5 -6 -7 -8 -9 -10 -14 -12 11 13 25 -38 -39 24 42 41 40 32 -31 -30 -29 33 34 35 36 37 |
| Cyclotella_nana | 2 3 4 17 18 22 -26 -27 -28 -16 12 14 10 9 8 7 6 5 11 23 -15 1 21 20 19 13 25 -38 -39 -40 -41 -42 -24 29 30 31 33 34 35 36 37 32 |
| Cyclotella sp. L04_2 | 2 3 4 16 28 27 26 -22 -18 -17 12 14 10 9 8 7 6 5 11 23 -15 1 21 20 19 13 25 -38 -39 -40 -41 -42 -24 29 30 31 33 34 35 36 37 32 |
| Cyclotella sp WC03_2 | 2 3 4 16 28 27 26 -22 -18 -17 12 14 10 9 8 7 6 5 11 23 -15 1 21 20 19 13 25 -38 -39 -40 -41 -42 -24 29 30 31 33 34 35 36 37 32 |
| Plagiogrammopsis van heurckii | 2 3 4 10 9 8 7 6 5 17 18 22 12 14 1 21 20 19 13 15 16 23 -26 -27 -28 11 25 24 42 41 40 39 38 -29 30 31 33 34 35 36 37 32 |
| Trieres sinensis | 2 3 4 10 9 8 7 6 5 17 18 22 12 14 1 21 20 19 13 15 16 28 27 26 23 11 25 -38 -39 -40 -41 -42 -24 29 30 31 33 34 35 36 37 32 |
| Triceratium dubium | 2 3 4 10 9 8 7 6 5 17 18 22 12 14 121 20 19 13 15 16 28 27 26 23 11 25 -38 -39 -40 -41 -42 -24 29 30 31 33 34 35 36 37 32 |
| Cerataulina daemon | 2 3 4 -23 10 9 8 7 6 5 17 18 22 12 14 1 21 20 19 13 15 16 28 27 26 11 25 -38 -39 -40 -41 -42 -24 29 30 31 33 34 35 36 37 32 |
| Acanthoceras zachariasii | 10 9 8 2 3 4 7 6 5 17 18 22 12 14 1 21 20 19 13 15 16 28 27 26 23 11 25 29 30 31 33 34 35 36 37 32 24 42 41 40 39 38 |
| Chaetoceros simplex | 10 9 8 2 3 4 7 6 5 17 18 22 12 14 1 -19 -20 -21 13 15 16 28 27 26 23 11 25 29 30 31 33 34 35 36 37 32 24 42 41 40 3938 |
| Attheya logicornis | 2 3 4 13 -19 -20 -21 23 15 16 1 28 27 26 22 12 14 -18 -17 -5 -6 -7 -8 -9 -10 11 25 -39 -40 -41 -42 -24 29 30 31 33 34 35 36 37 32 38 |
| Biddulphia tridens | 2 3 4 13 -19 -20 -21 23 15 16 28 27 26 1 -14 -12 -22 -18 -17 -5 -6 -7 -8 -9 -1011 25 29 30 31 33 34 35 36 37 32 24 42 41 40 39 38 |
| Biddulphia biddulphiana | 2 3 4 13 -19 -20 -21 23 15 16 28 27 26 1 -14 -12 -22 -18 -17 -5 -6 -7 -8 -9 -10 11 25 29 30 31 33 34 35 36 37 32 24 42 41 40 39 38 |
| Asterionellopsis glacialis | 2 3 4 13 -23 -1 21 20 19 15 16 -11 -14 -12 -22 -18 -17 10 9 8 7 6 5 -26 -27 -28 25 -38 33 34 35 36 37 -39 -40 -41 -42 -24 -31 -30 -29 32 |
| Plagiogramma staurophorum | -21 1 23 -4 -3 -2 10 9 8 7 6 5 17 18 22 12 14 11 -16 -15 28 27 26 20 19 13 25 -38 -39 -40 -41 -42 -24 -29 30 31 33 34 35 36 37 32 |
| Psammoneis obaidii | 7 6 5 2 3 4 -13 -19 -20 -21 1 23 15 16 -11 17 18 22 12 14 -26 -27 -28 10 9 8 25 -38 32 29 30 31 33 34 35 36 37 24 42 41 40 39 |

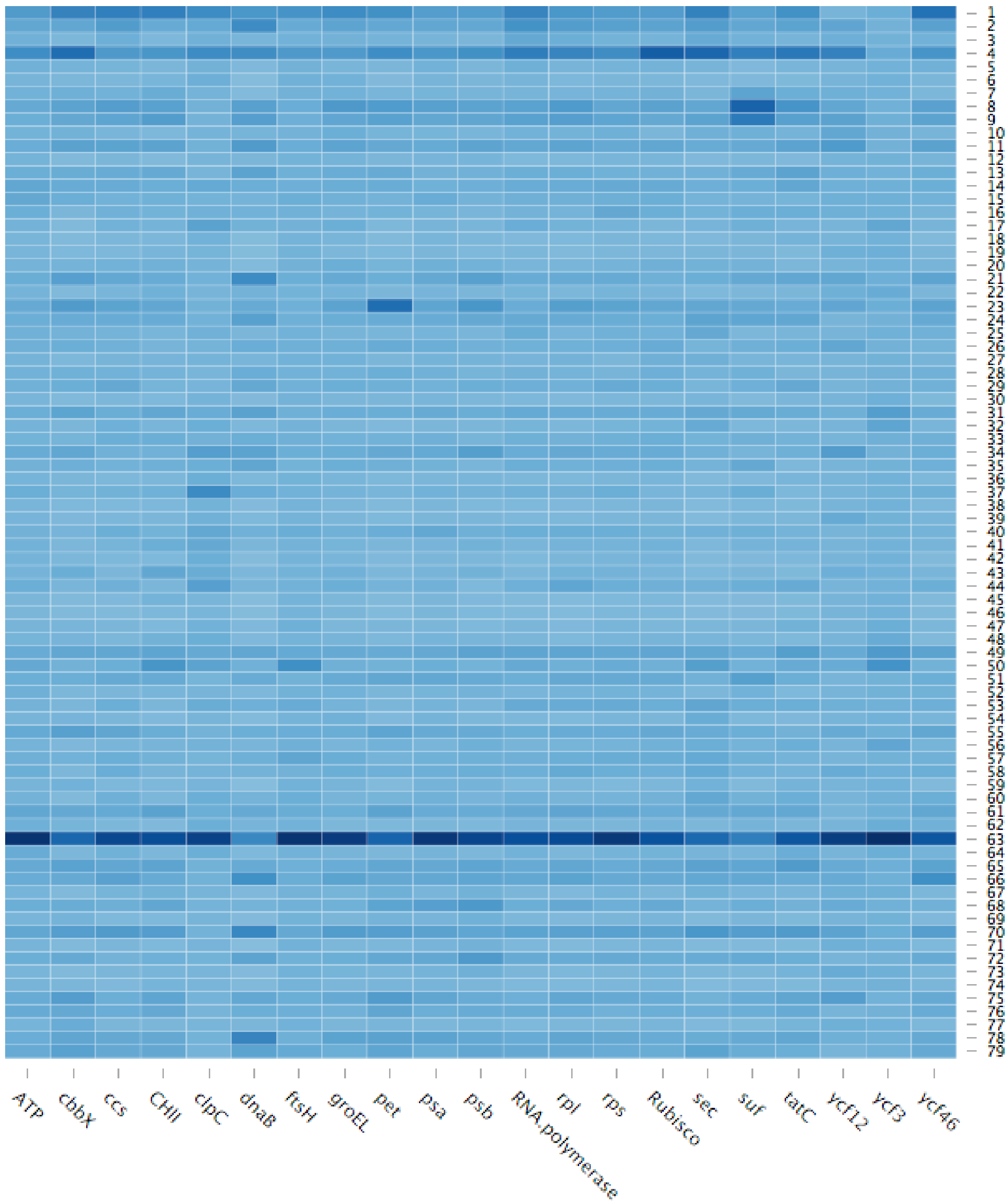
Appendix Table B.6. The permutation of number coded Locally Colinear Block (LCB) for each plastid genome. Negative number indicates an inversion of the given LCB. The species with same gene order are highlighted in same color.

| Species | Pearson correlation | Cosine similarity | Bonferroni corrected P Values |
|---|---|---|---|
| Leptocylindrus.danicus | 0.300527725 | 0.966511547 | 1 |
| Proboscia.sp. | 0.298994885 | 0.982463104 | 1 |
| Actinocyclus.subtilis | 0.36307644 | 0.916780204 | 0.923980051 |
| Coscinodiscus.radiatus | 0.362684359 | 0.916656178 | 0.930368861 |
| Rhizosolenia.setigera | 0.380395451 | 0.927302829 | 0.676033793 |
| Guinardia.striata | 0.424447589 | 0.940498767 | 0.28317551 |
| Rhizosolenia.fallax | 0.43351619 | 0.930083763 | 0.233349191 |
| Rhizosolenia.imbricata | 0.414041479 | 0.932793743 | 0.35138452 |
| Lithodesmium.undulatum | 0.217643241 | 0.910789935 | 1 |
| Eunotogramma.sp. | 0.174424289 | 0.923694794 | 1 |
| Roundia.cardiophora | 0.624747486 | 0.941415891 | 0.000850738 |
| Thalassiosira.weissflogii | 0.639665289 | 0.942700294 | 0.000465779 |
| Discostella.pseudostelligera | 0.643394807 | 0.942455083 | 0.000398647 |
| Thalassiosira.oceanica | 0.721714345 | 0.957219391 | 8.65E-06 |
| Cyclotella.nana | 0.655192983 | 0.942926895 | 0.000240263 |
| Cyclotella.sp.L04_2 | 0.665370634 | 0.946527429 | 0.000152478 |
| Cyclotella.sp.WC03_2 | 0.665124243 | 0.946283918 | 0.000154197 |
| Plagiogrammopsis.van.heurckii | 0.502569716 | 0.94662253 | 0.044494692 |
| Trieres.sinensis | 0.51426211 | 0.938811013 | 0.032433879 |
| Triceratium.dubium | 0.511005484 | 0.937279255 | 0.035459705 |
| Cerataulina.daemon | 0.391329462 | 0.918228801 | 0.550395549 |
| Acanthoceras.zachariasii | 0.5359583 | 0.953510683 | 0.017490891 |
| Chaetoceros.simplex | 0.53281735 | 0.953794818 | 0.019175989 |
| Attheya.longicornis | 0.254110385 | 0.941266996 | 1 |
| Biddulphia.tridens | 0.483796883 | 0.945525515 | 0.072253036 |
| Biddulphia.biddulphiana | 0.500576854 | 0.952005709 | 0.04690607 |
| Asterionellopsis.glacialis | 0.256530321 | 0.940181273 | 1 |
| Plagiogramma.staurophorum | 0.42909215 | 0.95077029 | 0.256619127 |
| Psammoneis.obaidii | 0.265190737 | 0.942800462 | 1 |
| Asterionella.formosa | 0.404419665 | 0.947345538 | 0.426515999 |
| Astrosyne.radiata | 0.712869078 | 0.982525667 | 1.42E-05 |
| Synedra.acus | 0.449584292 | 0.937522651 | 0.163478869 |
| Licmophora.sp. | 0.525025317 | 0.962497795 | 0.023999057 |
| Eunotia.naegelii | 0.373246764 | 0.905145087 | 0.770579275 |
| Cylindrotheca.closterium | 0.435591957 | 0.961946827 | 0.223074753 |
| Seminavis.robusta | 0.402372733 | 0.926795655 | 0.444153936 |
| Entomoneis.sp. | 0.548764845 | 0.956853379 | 0.011907802 |
| Fistulifera.sp.JPCC.DA0580 | 0.404570266 | 0.924337063 | 0.425242218 |
| Didymosphenia.germinata | 0.470392001 | 0.919442225 | 0.100486588 |
| Phaeodactylum.tricornutum | 0.471696255 | 0.919940409 | 0.097368916 |

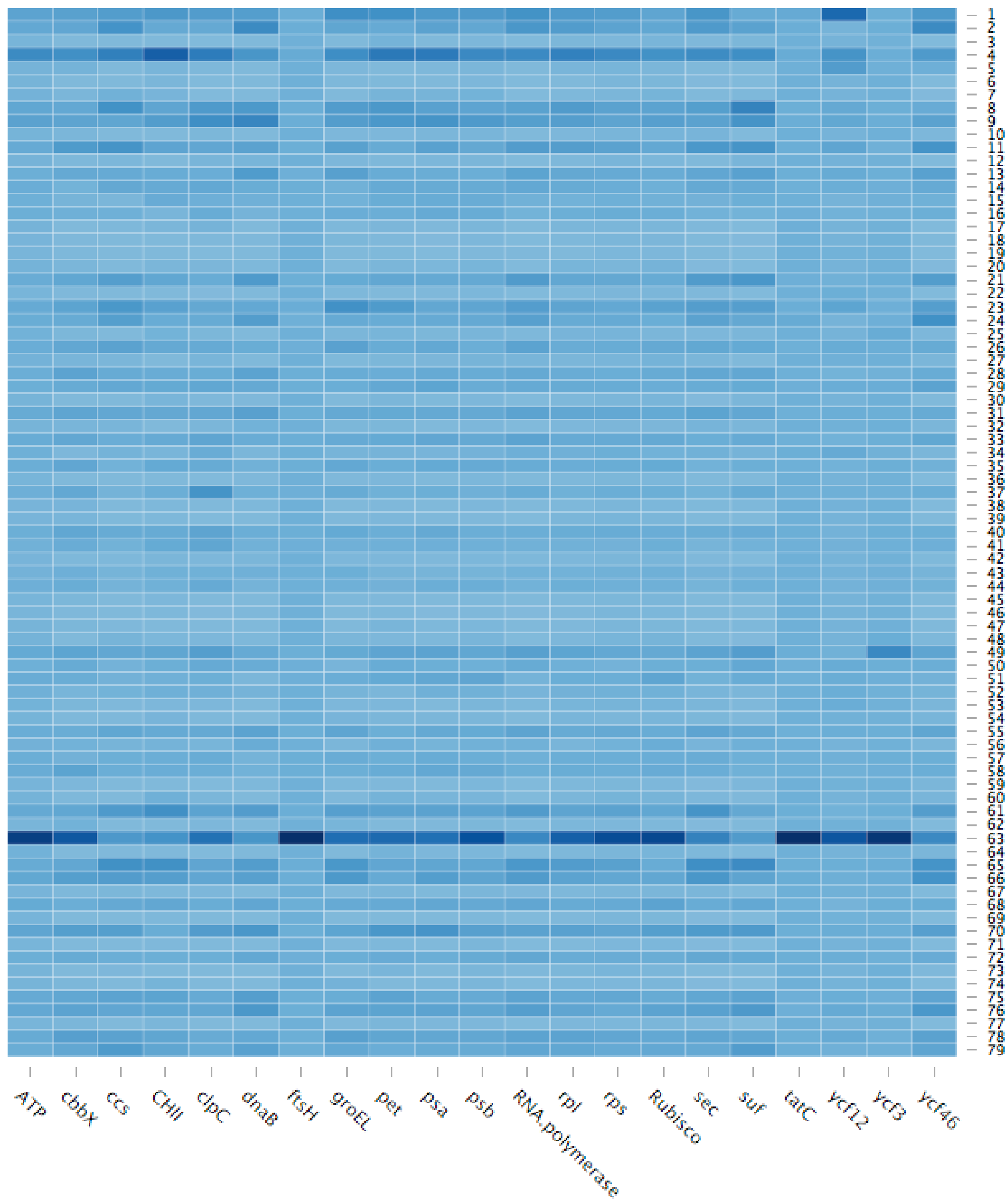
Pink highlights indicate significant P values.

Appendix Table B.7. Correlation test score between pairwise branch length estimated from maximum likelihood tree and gene order inversion distance inferred by GRIMM.

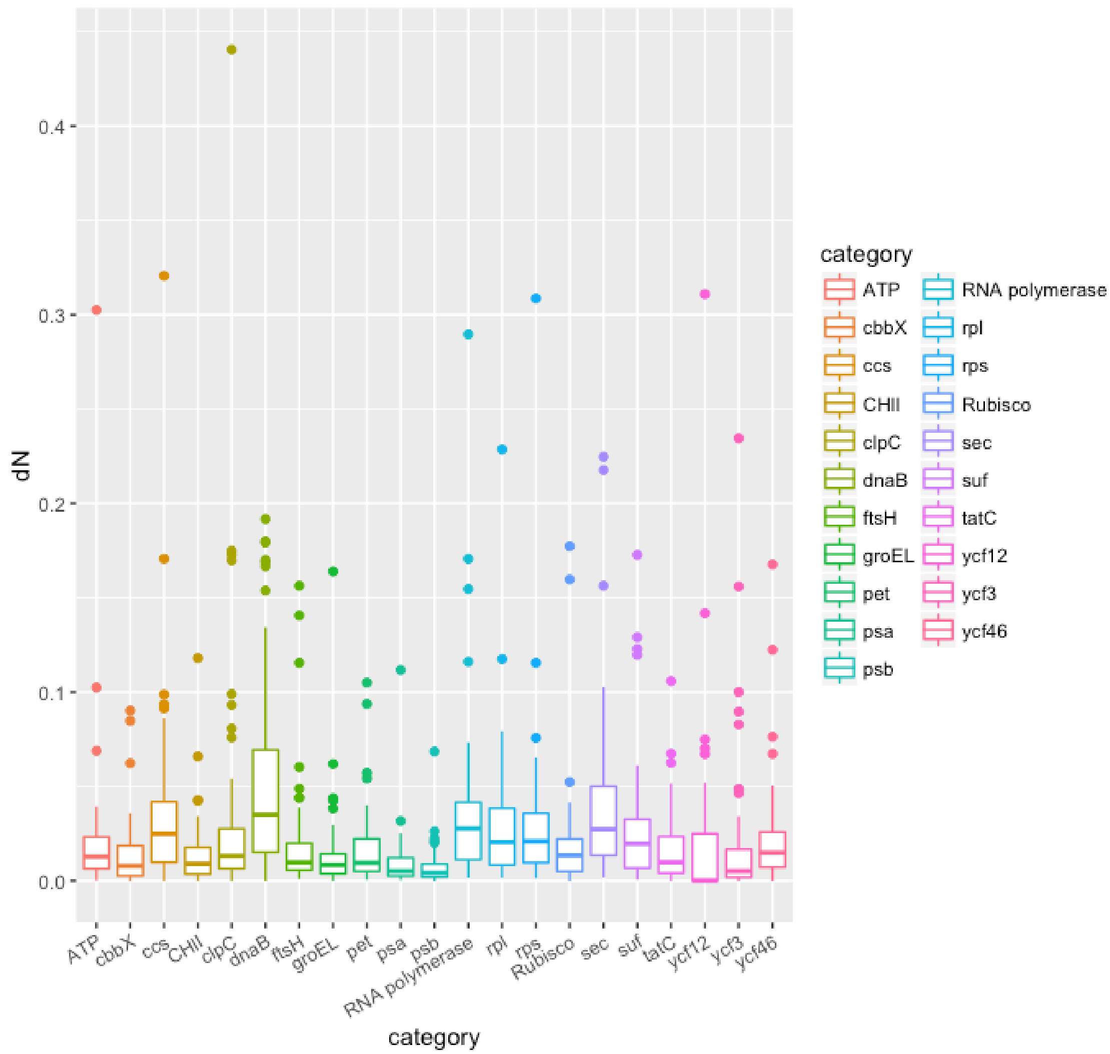
**Chapter 4**



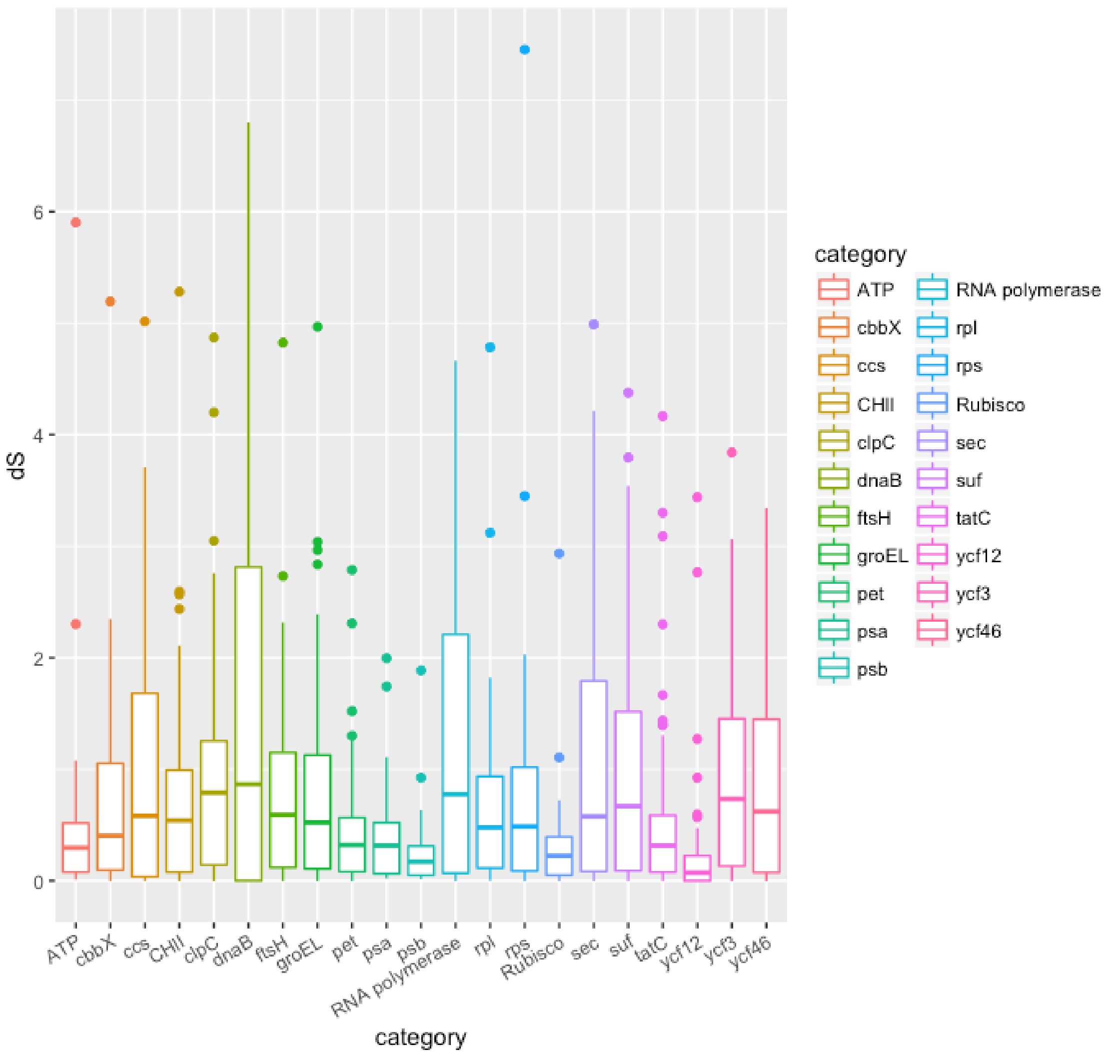

Appendix Figure C.1. Heatmap of non-synonymous substitution rates on different branches across gene functional groups. Branch numbers on the y-axis correspond to the branch labels on the phylogeny in from Figure 4.1 with the outgroup taxa removed. The intensity of the color is proportional to the value of *dN* with darker values having higher *dN*.



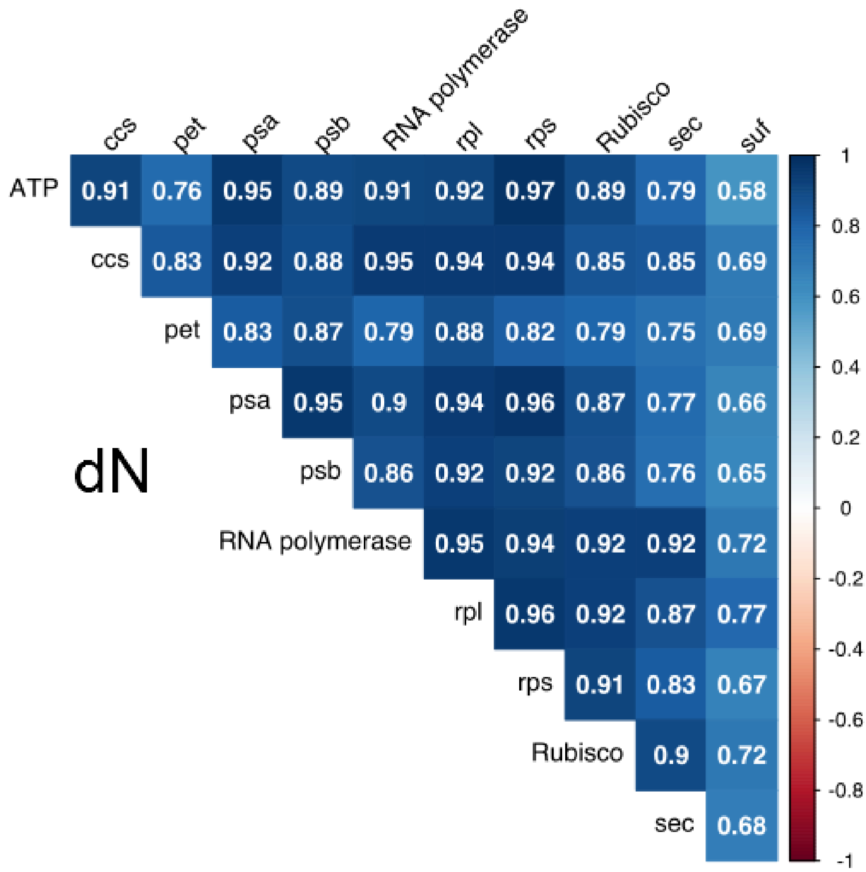

Appendix Figure C.2. Heatmap of synonymous substitution rates on different branches across different gene functional groups. Branch numbers on the y-axis correspond to the branch labels on the phylogeny in Figure 4.1 with the outgroup taxa removed. The intensity of the color is proportional to the value of *dS* with darker values having higher *dS*.



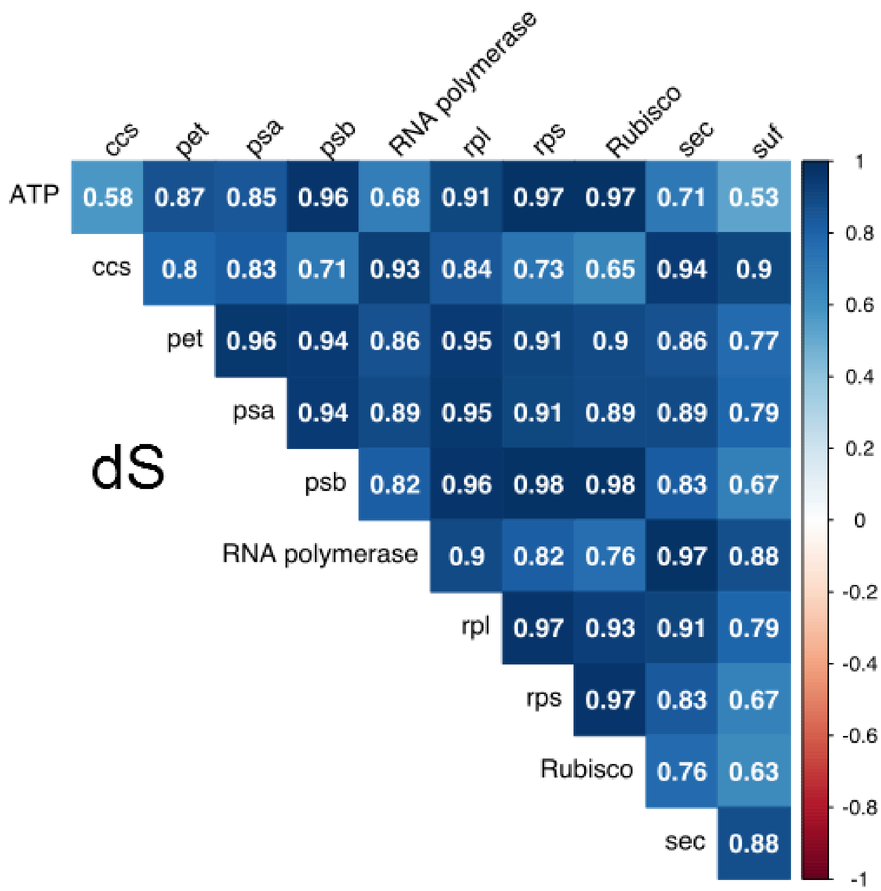

Appendix Figure C.3. Boxplot of the number of nonsynonymous (*dN*) substitutions for groups of genes and individual genes.



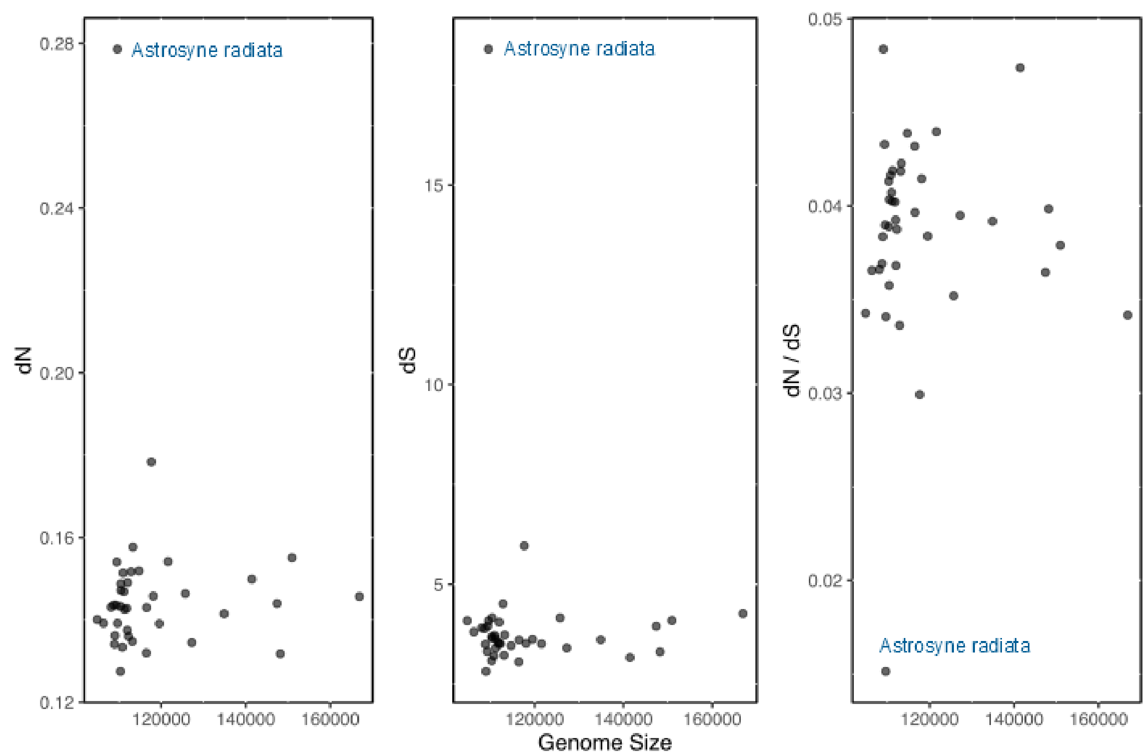

Appendix Figure C.4. Boxplot of the number of synonymous (*dS*) substitutions for groups of genes and individual genes.




Appendix Figure C.5. Heatmap of correlation of nonsynonymous (*dN*) substitution rates among major gene functional groups. The numbers in white represent correlation coefficient. The colors are proportional to the color bar on the right.




Appendix Figure C.6. Heatmap of correlation of synonymous (*dS*) substitution rates among major gene functional groups. The numbers in white represent correlation coefficient. The colors are proportional to the color bar on the right.




Appendix Figure C.7. Relationship between the plastid genome size (only one IR was included) and substitution rates.

| Category | Genes |
|---|---|
| Photosystem I | *psaA, psaB, psaD, psaF, psaJ, psaL* |
| Photosystem II | *psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI , psbJ, psbK, psbL, psbN, psbT, psbV, psbX, psbY, psbZ* |
| Cytochrome b/f complex | *petA, petB, petD, petG, petL, petM, petN* |
| ATP synthase | *atpA, atpB, atpD, atpE, atpF, atpG, atpH, atpI* |
| RubisCo subunit | *rbcL, rbcS, rbcR* |
| RNA polymerase | *rpoA, rpoB, rpoC1, rpoC2* |
| Ribosomal proteins large subunit | *rpl1, rpl2, rpl3, rpl4, rpl5, rpl6, rpl11, rpl12, rpl13, rpl14, rpl16, rpl18, rpl19, rpl20, rpl21, rpl22, rpl23, rpl24, rpl27, rpl29, rpl31, rpl32, rpl33, rpl34, rpl35* |
| Ribosomal proteins small subunit | *rps2, rps3, rps4, rps5, rps7, rps9, rps10, rps11, rps12, rps13, rps14, rps16, rps17, rps18, rps19, rps20* |
| Cytochrome c biogenesis | *ccs1, ccsA* |
| Protein translocase | *secA, secG, secY* |
| Fe-S cluster assembly protein | *sufB, sufC* |
| Other genes | *cbbX, chlI, clpC, dnaB, ftsH, groEL, tatC, ycf3, ycf12, ycf46* |

Appendix Table C.1. List of functional groups of genes with indication of which gene belongs in each category.

| Species | dN Cor | dN Adjusted P value | dS Cor | dS Adjusted P value | W Cor | W Adjusted P value | Clade |
|---|---|---|---|---|---|---|---|
| Leptocylindrus.danicus | 0.249362463 | 1 | 0.274406108 | 1 | -0.0947601 | 1 | Radial1 |
| Proboscia.sp. | 0.281901059 | 1 | 0.209321483 | 1 | 0.1019521 | 1 | Radial2 |
| Actinocyclus.subtilis | 0.421786386 | 0.299431933 | 0.363209145 | 0.921825966 | 0.2509063 | 1 | Radial3 |
| Coscinodiscus.radiatus | 0.416940345 | 0.331100498 | 0.38572864 | 0.612029705 | 0.2493974 | 1 | Radial3 |
| Rhizosolenia.setigera | 0.466018696 | 0.111588386 | 0.369400225 | 0.825877165 | 0.4058141 | 0.41484544 | Radial3 |
| Guinardia.striata | 0.511819208 | 0.034681049 | 0.43514724 | 0.225242417 | 0.4100105 | 0.3813526 | Radial3 |
| Rhizosolenia.fallax | 0.504714175 | 0.042023163 | 0.463100951 | 0.119579107 | 0.3876155 | 0.59064546 | Radial3 |
| Rhizosolenia.imbricata | 0.449402086 | 0.164155514 | 0.440441621 | 0.200582541 | 0.2543502 | 1 | Radial3 |
| Lithodesmium.undulatum | 0.173043419 | 1 | 0.163370113 | 1 | -0.1890741 | 1 | Polar1 |
| Eunotogramma.sp. | 0.142648235 | 1 | 0.161402882 | 1 | -0.1539923 | 1 | Polar1 |
| Roundia.cardiophora | 0.674043907 | 0.000102068 | 0.689209635 | 4.90E-05 | 0.4489244 | 0.16594097 | Polar1 |
| Thalassiosira.weissflogii | 0.697095747 | 3.28E-05 | 0.695149186 | 3.63E-05 | 0.5610295 | 0.00811869 | Polar1 |
| Discostella.pseudostelligera | 0.696753636 | 3.34E-05 | 0.699876785 | 2.84E-05 | 0.5778036 | 0.00468999 | Polar1 |
| Thalassiosira.oceanica | 0.775774078 | 2.64E-07 | 0.778485249 | 2.16E-07 | 0.6735405 | 0.00010451 | Polar1 |
| Cyclotella.nana | 0.710748719 | 1.59E-05 | 0.713458405 | 1.37E-05 | 0.6004613 | 0.00212756 | Polar1 |
| Cyclotella.sp.L04_2 | 0.702493306 | 2.48E-05 | 0.714705524 | 1.28E-05 | 0.5544628 | 0.00998511 | Polar1 |
| Cyclotella.sp.WC03_2 | 0.702450644 | 2.48E-05 | 0.714486026 | 1.30E-05 | 0.5549535 | 0.00983337 | Polar1 |
| Plagiogrammopsis.van.heurckii | 0.539178885 | 0.015901802 | 0.517654516 | 0.029528124 | 0.2912593 | 1 | Polar2 |
| Trieres.sinensis | 0.55324712 | 0.010370159 | 0.55223179 | 0.010701918 | 0.3504453 | 1 | Polar2 |
| Triceratium.dubium | 0.54893446 | 0.011846083 | 0.548404246 | 0.012039974 | 0.3496521 | 1 | Polar2 |
| Cerataulina.daemon | 0.385386993 | 0.61597059 | 0.414586181 | 0.347493932 | 0.1296736 | 1 | Polar2 |
| Acanthoceras.zachariasii | 0.607803552 | 0.001625466 | 0.549880351 | 0.011507131 | 0.4872444 | 0.06623323 | Polar2 |
| Chaetoceros.simplex | 0.607150507 | 0.001665304 | 0.529659953 | 0.021014678 | 0.5170309 | 0.03004423 | Polar2 |
| Attheya.longicornis | 0.324697507 | 1 | 0.330594406 | 1 | 0.2832113 | 1 | Polar3 |
| Biddulphia.tridens | 0.573014743 | 0.00550208 | 0.545305193 | 0.013231303 | 0.3795144 | 0.68713137 | Polar3 |
| Biddulphia.biddulphiana | 0.590314317 | 0.003053483 | 0.565746504 | 0.006978508 | 0.3827313 | 0.64734071 | Polar3 |
| Asterionellopsis.glacialis | 0.41273536 | 0.360865861 | 0.250482164 | 1 | 0.3720415 | 0.78756207 | Araphid1 |
| Plagiogramma.staurophorum | 0.48560559 | 0.069037318 | 0.401567136 | 0.451263897 | 0.0826206 | 1 | Araphid1 |
| Psammoneis.obaidii | 0.444644599 | 0.182706417 | 0.278596095 | 1 | 0.3807752 | 0.67129707 | Araphid1 |
| Asterionella.formosa | 0.479432935 | 0.080560614 | 0.445734299 | 0.178303945 | 0.2112848 | 1 | Araphid2 |
| Astrosyne.radiata | 0.743113847 | 2.41E-06 | 0.598559319 | 0.002278767 | 0.6461316 | 0.00035515 | Araphid2 |
| Synedra.acus | 0.62298395 | 0.000911655 | 0.439424521 | 0.205129545 | 0.4939714 | 0.05574699 | Araphid2 |
| Licmophora.sp. | 0.682795027 | 6.72E-05 | 0.496816073 | 0.051773684 | 0.4754451 | 0.08887504 | Araphid2 |
| Eunotia.naegelii | 0.493281213 | 0.05675078 | 0.390963517 | 0.554255499 | 0.3688571 | 0.8339466 | Raphid |
| Cylindrotheca.closterium | 0.58102035 | 0.004207043 | 0.453658639 | 0.148968941 | 0.5317408 | 0.01978609 | Raphid |
| Seminavis.robusta | 0.574853202 | 0.005176409 | 0.422466096 | 0.295205369 | 0.5885402 | 0.00324857 | Raphid |
| Entomoneis.sp. | 0.660163357 | 0.00019283 | 0.586512136 | 0.003485367 | 0.5802484 | 0.00431866 | Raphid |
| Fistulifera.sp.JPCC.DA0580 | 0.507540489 | 0.038952023 | 0.436830462 | 0.21713278 | 0.4393604 | 0.20541924 | Raphid |
| Didymosphenia.germinata | 0.601524138 | 0.002047095 | 0.505160094 | 0.041524793 | 0.562476 | 0.00775239 | Raphid |
| Phaeodactylum.tricornutum | 0.599266452 | 0.002221449 | 0.516731197 | 0.030295096 | 0.5620021 | 0.00787073 | Raphid |

Appendix Table C.2. Correlation coefficient and adjusted P values for correlation between substitution rates and genome rearrangement measured by inversion distance. Red entry indicates significant p values.

# References

## Chapter 2


Alverson, A.J. (2007). Strong purifying selection in the silicon transporters of marine and freshwater diatoms. Limnol Oceanogr *52*, 1420-1429.

Alverson, A.J., Beszteri, B., Julius, M.L., and Theriot, E.C. (2011). The model marine diatom *Thalassiosira pseudonana* likely descended from a freshwater ancestor in the genus *Cyclotella*. BMC Evol Biol *11*, 125.

Alverson, A.J., Jansen, R.K., and Theriot, E.C. (2007). Bridging the Rubicon: Phylogenetic analysis reveals repeated colonizations of marine and fresh waters by thalassiosiroid diatoms. Mol Phylogenet Evol *45*, 193-210.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M.*, et al.* (2004). The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. Science *306*, 79-86.

Ashworth, J., Coesel, S., Lee, A., Armbrust, E.V., Orellana, M.V., and Baliga, N.S. (2013). Genome-wide diel growth state transitions in the diatom Thalassiosira pseudonana. Proc Natl Acad Sci U S A *110*, 7518-7523.

Baldauf, S.L., and Palmer, J.D. (1990). Evolutionary transfer of the chloroplast tufA gene to the nucleus. Nature *344*, 262-265.

Barraclough, T., and Savolainen, V. (2001). Evolutionary rates and species diversity in flowering plants. Evolution *55*, 677 - 683.

Belda, E., Moya, A., and Silva, F.J. (2005). Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria. Mol Biol Evol *22*, 1456-1467.

Blazier, J.C., Ruhlman, T.A., Weng, M.-L., Rehman, S.K., Sabir, J.S.M., and Jansen, R.K. (2016). Divergence of RNA polymerase α subunits in angiosperm plastid genomes is mediated by genomic rearrangement. *6*, 24595.

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P.*, et al.* (2008). The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature *456*, 239-244.

Bradwell, K., Combe, M., Domingo-Calap, P., and Sanjuán, R. (2013). Correlation Between Mutation Rate and Genome Size in Riboviruses: Mutation Rate of Bacteriophage Qβ. Genetics *195*, 243-251.

Brembu, T., Winge, P., Tooming-Klunderud, A., Nederbragt, A.J., Jakobsen, K.S., and Bones, A.M. (2013). The chloroplast genome of the diatom Seminavis robusta: New features introduced through multiple mechanisms of horizontal gene transfer. Mar Genomics *21*, 00080-00089.

Britten, R.J. (1986). Rates of DNA sequence evolution differ between taxonomic groups. Science *231*.

Buschiazzo, E., Ritland, C., Bohlmann, J., and Ritland, K. (2012). Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. BMC Evol Biol *12*, 8.

Chang, C., Lin, H., Lin, I., Chow, T., Chen, H., Chen, W., Cheng, C., Lin, C., Liu, S., and Chaw, S. (2006). The chloroplast genome of Phalaenopsis aphrodite (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol Biol Evol *23*, 279 - 291.

Christensen, A.C. (2013). Plant Mitochondrial Genome Evolution Can Be Explained by DNA Repair Mechanisms. Genome Biol Evol *5*, 1079-1086.

Darling, A.C., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res *14*, 1394-1403.

De La Torre, A.R., Lin, Y.-C., Van de Peer, Y., and Ingvarsson, P.K. (2015). Genome-Wide Analysis Reveals Diverged Patterns of Codon Bias, Gene Expression, and Rates of Sequence Evolution in Picea Gene Families. Genome Biol Evol *7*, 1002-1015.

Dettman, J.R., Sztepanacz, J.L., and Kassen, R. (2016). The properties of spontaneous mutations in the opportunistic pathogen Pseudomonas aeruginosa. BMC Genomics *17*, 27.

Doyle, J., and Doyle, J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissues. Phytochem bull *19*, 11 - 15.

Drake, J.W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. P Natl Acad Sci USA *88*, 7160-7164.

Drake, J.W., Charlesworth, B., Charlesworth, D., and Crow, J.F. (1998). Rates of Spontaneous Mutation. Genetics *148*, 1667.

Drummond, A.J., and al, e. (2010). Geneious v5.5. Available at http://wwwgeneiouscom.

Drummond, D.A., Raval, A., and Wilke, C.O. (2006). A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol *23*, 327-337.

Du, X., Lipman, D.J., and Cherry, J.L. (2013). Why Does a Protein's Evolutionary Rate Vary over Time? Genome Biol Evol *5*, 494-503.

Duchene, D., and Bromham, L. (2013). Rates of molecular evolution and diversification in plants: chloroplast substitution rates correlate with species-richness in the Proteaceae. BMC Evol Biol *13*, 65.

Evans, K.M., Bates, S.S., Medlin, L.K., and Hayes, P.K. (2004). Microsatellite marker development and genetic variation in the toxic marine diatom *Pseudo-nitzschia multiseries* (Bacillariophyceae). J Phycol *40*, 911-920.

Galachyants, Y.P., Morozov, A.A., Mardanov, A.V., Beletsky, A.V., and Ravin, N.V. (2012). Complete Chloroplast Genome Sequence of Freshwater Araphid Pennate Diatom Alga Synedra acus from Lake Baikal. International Journal of Biology *4*, 27-35.

Gockel, G., and Hachtel, W. (2000). Complete Gene Map of the Plastid Genome of the Nonphotosynthetic Euglenoid Flagellate Astasia longa. Protist *151*, 347-351.

Goulding, S., Olmstead, R., Morden, C., and Wolfe, K. (1996). Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet *252*, 195 - 206.

Green, B.R. (2011). Chloroplast genomes of photosynthetic eukaryotes. Plant J *66*, 34-44.

Gruber, A., Vugrinec, S., Hempel, F., Gould, S.B., Maier, U.G., and Kroth, P.G. (2007). Protein targeting into complex diatom plastids: functional characterisation of a specific targeting motif. Plant Mol Biol *64*, 519-530.

Guillard, R.R.L. (1983). Culture of phytoplankton for feeding marine intvertebrates. In Culture of Marine Invertebrates, C.J. Berg, ed. (New York, Hutchinson Ropss Publication Company), pp. 108-132.

Guisinger, M., Kuehl, J., Boore, J., and Jansen, R. (2008a). Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc Natl Acad Sci U S A *105*, 18424 - 18429.

Guisinger, M., Kuehl, J., Boore, J., and Jansen, R. (2011). Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol *28*, 583 - 600.

Guisinger, M.M., Chumley, T.W., Kuehl, J.V., Boore, J.L., and Jansen, R.K. (2010). Implications of the Plastid Genome Sequence of Typha (Typhaceae, Poales) for Understanding Genome Evolution in Poaceae. J Mol Evol *70*, 149-166.

Guisinger, M.M., Kuehl, J.V., Boore, J.L., and Jansen, R.K. (2008b). Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc Natl Acad Sci U S A *105*, 18424-18429.

Hodgkinson, A., and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. Nat Rev Genet *12*, 756-766.

Hollister, J.D., Ross-Ibarra, J., and Gaut, B.S. (2010). Indel-Associated Mutation Rate Varies with Mating System in Flowering Plants. Mol Biol Evol *27*, 409-416.

Hu, J., and Blanchard, J.L. (2009). Environmental Sequence Data from the Sargasso Sea Reveal That the Characteristics of Genome Reduction in Prochlorococcus Are Not a Harbinger for an Escalation in Genetic Drift. Mol Biol Evol *26*, 5-13.

Interlandi, S.J., and Kilham, S.S. (1998). Assessing the effects of nitrogen deposition on mountain waters: A study of phytoplankton community dynamics. Water Science & Technology *38*, 139-146.

Jancek, S., Gourbière, S., Moreau, H., and Piganeau, G. (2008). Clues about the Genetic Basis of Adaptation Emerge from Comparing the Proteomes of Two Ostreococcus Ecotypes (Chlorophyta, Prasinophyceae). Mol Biol Evol *25*, 2293-2300.

Janouškovec, J., Liu, S.-L., Martone, P.T., Carré, W., Leblanc, C., Collén, J., and Keeling, P.J. (2013). Evolution of Red Algal Plastid Genomes: Ancient Architectures, Introns, Horizontal Gene Transfer, and Taxonomic Utility of Plastid Markers. PLoS ONE *8*, e59001.

Jansen, R., and Ruhlman, T. (2012). Plastid Genomes of Seed Plants. In Genomics of Chloroplasts and Mitochondria, R. Bock, and V. Knoop, eds. (Springer Netherlands), pp. 103-126.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., dePamphilis, C.W., Leebens-Mack, J., Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K*., et al.* (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. P Natl Acad Sci USA *104*, 19369-19374.

Julius, M.L., and Theriot, E.C. (2010). The diatoms: A primer. In The Diatoms: Applications for the Environmental and Earth Sciences, J. Smol, and E.F. Stoermer, eds. (Cambridge University Press), pp. 8-22.

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res *33*, 511 - 518.

Koester, J.A., Swanson, W.J., and Armbrust, E.V. (2013). Positive selection within a diatom species acts on putative protein interactions and transcriptional regulation. Mol Biol Evol *30*, 422-434.

Kooistra, W.H.C.F., Gersonde, R., Medlin, L.K., and Mann, D.G. (2007). CHAPTER 11 - The Origin and Evolution of the Diatoms: Their Adaptation to a Planktonic Existence A2 - Falkowski, Paul G. In Evolution of Primary Producers in the Sea, A.H. Knoll, ed. (Burlington, Academic Press), pp. 207-249.

Kowallik, K., Stoebe, B., Schaffran, I., Kroth-Pancic, P., and Freier, U. (1995). The chloroplast genome of a chlorophylla+c-containing alga,Odontella sinensis. Plant Molecular Biology Reporter *13*, 336-342.

Lanfear, R., Ho, S.Y.W., Love, D., and Bromham, L. (2010). Mutation rate is linked to diversification in birds. Proc Natl Acad Sci USA *107*.

Lee, E.K., Cibrian-Jaramillo, A., Kolokotronis, S.-O., Katari, M.S., Stamatakis, A., Ott, M., Chiu, J.C., Little, D.P., Stevenson, D.W., McCombie, W.R*., et al.* (2011). A Functional Phylogenomic View of the Seed Plants. PLoS Genet *7*, e1002411.

Lemieux, C., Otis, C., and Turmel, M. (2016). Comparative Chloroplast Genome Analyses of Streptophyte Green Algae Uncover Major Structural Alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. Frontiers in Plant Science *7*, 697.

Liao, B.-Y., Scott, N.M., and Zhang, J. (2006). Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate of Mammalian Proteins. Mol Biol Evol *23*, 2072-2080.

Lohse, M., Drechsel, O., and Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet *52*, 267 - 274.

Lommer, M., Roy, A.S., Schilhabel, M., Schreiber, S., Rosenstiel, P., and LaRoche, J. (2010). Recent transfer of an iron-regulated gene from the plastid to the nuclear

genome in an oceanic diatom adapted to chronic iron limitation. BMC Genomics *11*.
Lynch, M., Koskella, B., and Schaack, S. (2006). Mutation Pressure and the Evolution of Organelle Genomic Architecture. Science *311*, 1727-1730.
Madhu, N.V., Meenu, P., Ullas, N., Ashwini, R., and Rehitha, T.V. (2013). Occurrence of cyanobacteria (Richelia intracellularis)-diatom (Rhizosolenia hebetaa) consortium in the Palk Bay, southeast coast of India. Indian Journal of Geo-Marine Sciences *42*, 453-457.
Mock, T., Otillar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., Ward, B.J.*, et al.* (2017). Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature *541*, 536-540.
Moran, N. (2001). Microbial minimalism: genome reduction in bacterial pathogens. Cell *108*, 583 - 586.
Nakov, T., Alverson, A.J., and Theriot, E. (2014). Using phylogeny to model cell size evolution in marine and freshwater diatoms.
Nelson, D.M., Treguer, P., Brzezinski, M.A., Leynaert, A., and Queguiner, B. (1995). Production and Dissolution of Biogenic Silica in the Ocean - Revised Global Estimates, Comparison with Regional Data and Relationship to Biogenic Sedimentation. Global Biogeochem Cy *9*, 359-372.
Oudot-Le Secq, M.-P., Grimwood, J., Shapiro, H., Armbrust, E., Bowler, C., and Green, B. (2007). Chloroplast genomes of the diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana and comparison with other plastid genomes of the red lineage. Molecular Genetics and Genomics *277*, 427-439.
Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Meth *8*, 785-786.
Round, F.E., and Crawford, R.M. (1984). The lines of evolution of the Bacillariophyta. 2. The Centric Series. Proc R Soc Lond Ser B-Biol Sci *221*, 169-&.
Ruck, E.C., Nakov, T., Jansen, R.K., Theriot, E.C., and Alverson, A.J. (2014). Serial gene losses and foreign DNA underlie size and sequence variation in the plastid genomes of diatoms. Genome Biol Evol *6*, 644-654.
Rumpho, M.E., Worful, J.M., Lee, J., Kannan, K., Tyler, M.S., Bhattacharya, D., Moustafa, A., and Manhart, J.R. (2008). Horizontal gene transfer of the algal nuclear gene psbO to the photosynthetic sea slug Elysia chlorotica. P Natl Acad Sci USA *105*, 17867-17871.
Sabir, J.S.M., Yu, M., Ashworth, M.P., Baeshen, N.A., Baeshen, M.N., Bahieldin, A., Theriot, E.C., and Jansen, R.K. (2014). Conserved Gene Order and Expanded Inverted Repeats Characterize Plastid Genomes of Thalassiosirales. PLoS ONE *9*, e107854.

Schattner, P., Brooks, A.N., and Lowe, T.M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res *33*, W686-689.

Schwarz, E.N., Ruhlman, T.A., Weng, M.L., Khiyami, M.A., Sabir, J.S.M., Hajarah, N.H., Alharbi, N.S., Rabah, S.O., and Jansen, R.K. (2017). Plastome-Wide Nucleotide Substitution Rates Reveal Accelerated Rates in Papilionoideae and Correlations with Genome Features Across Legume Subfamilies. J Mol Evol *84*, 187-203.

Shao, R., Dowton, M., Murrell, A., and Barker, S.C. (2003). Rates of gene rearrangement and nucleotide substitution are correlated in the mitochondrial genomes of insects. Mol Biol Evol *20*, 1612-1619.

Sharp, P.M. (1991). Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: codon usage, map position, and concerted evolution. J Mol Evol *33*, 23-33.

Shields, D.C., Sharp, P.M., Higgins, D.G., and Wright, F. (1988). "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol *5*, 704-716.

Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P., Wu, M., McCauley, D.E., Palmer, J.D., and Taylor, D.R. (2012). Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. Plos Biology *10*, e1001241.

Smith, S.A., and Donoghue, M.J. (2008). Rates of molecular evolution are linked to life history in flowering plants. Science *322*.

Sorhannus, U. (2007). A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. Mar Micropaleontol *65*, 1-12.

Sorhannus, U., and Fox, M. (1999). Synonymous and nonsynonymous substitution rates in diatoms: a comparison between chloroplast and nuclear genes. J Mol Evol *48*, 209-212.

Stamatakis, A. (2006a). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688 - 2690.

Stamatakis, A. (2006b). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688-2690.

Tanaka, T., Fukuda, Y., Yoshino, T., Maeda, Y., Muto, M., Matsumoto, M., Mayama, S., and Matsunaga, T. (2011). High-throughput pyrosequencing of the chloroplast genome of a highly neutral-lipid-producing marine pennate diatom, Fistulifera sp. strain JPCC DA0580. Photosynth Res *109*, 223-229.

Tesler, G. (2002). GRIMM: genome rearrangements web server. Bioinformatics *18*, 492 - 493.

Theriot, E.C., Ruck, E.C., Ashworth, M., Nakov, T., and Jansen, R.K. (2011). Status of the pursuit of the diatom phylogeny: Are traditional views and new molecular

paradigms really that different? In The Diatom World, J. Seckbach, and J.P. Kociolek, eds. (Springer), p. 600.
Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R., Bergelson, J., and Chen, J.-Q. (2008). Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature *455*, 105-108.
Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. (2012). Primer3--new capabilities and interfaces. Nucleic Acids Res *40*, 22.
Villareal, T.A. (1990). Laboratory Culture and Preliminary Characterization of the Nitrogen-Fixing Rhizosolenia-Richelia Symbiosis. Mar Ecol *11*, 117-132.
Vivares, C.P., Gouy, M., Thomarat, F., and Metenier, G. (2002). Functional and evolutionary analysis of a eukaryotic parasitic genome. Curr Opin Microbiol *5*, 499-505.
Weng, M.-L., Blazier, J.C., Govindu, M., and Jansen, R.K. (2013). Reconstruction of the Ancestral Plastid Genome in Geraniaceae Reveals a Correlation Between Genome Rearrangements, Repeats and Nucleotide Substitution Rates. Mol Biol Evol.
Wicke, S., Schäferhoff, B., dePamphilis, C.W., and Müller, K.F. (2014). Disproportional Plastome-Wide Increase of Substitution Rates and Relaxed Purifying Selection in Genes of Carnivorous Lentibulariaceae. Mol Biol Evol *31*, 529-545.
Wilson, A.C., Carlson, S.S., and White, T.J. (1977). Biochemical evolution. Annu Rev Biochem *46*, 573-639.
Wolf, P.G., Der, J.P., Duffy, A.M., Davidson, J.B., Grusz, A.L., and Pryer, K.M. (2011). The evolution of chloroplast genes and genomes in ferns. Plant Mol Biol *76*, 251-261.
Wolfe, K., Li, W., and Sharp, P. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA *84*, 9054 - 9058.
Wolfe, K.H., Morden, C.W., and Palmer, J.D. (1992). Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. P Natl Acad Sci USA *89*, 10648-10652.
Wu, C.-S., and Chaw, S.-M. (2015). Evolutionary Stasis in Cycad Plastomes and the First Case of Plastome GC-Biased Gene Conversion. Genome Biol Evol *7*, 2000-2009.
Wu, C.S., and Chaw, S.M. (2014). Highly rearranged and size-variable chloroplast genomes in conifers II clade (cupressophytes): evolution towards shorter intergenic spacers. Plant Biotechnol J *12*, 344-353.
Wyman, S., Jansen, R., and Boore, J. (2004). Automatic annotation of organellar genomes with DOGMA. Bioinformatics *20*, 3252 - 3255.
Xu, Q., Xiong, G., Li, P., He, F., Huang, Y., Wang, K., Li, Z., and Hua, J. (2012). Analysis of Complete Nucleotide Sequences of 12 Gossypium Chloroplast Genomes: Origin and Evolution of Allotetraploids. PLoS ONE *7*, e37128.

Xu, W., Jameson, D., Tang, B., and Higgs, P.G. (2006). The relationship between the rate of molecular evolution and the rate of genome rearrangement in animal mitochondrial genomes. J Mol Evol *63*, 375-392.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol *24*, 1586 - 1591.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res *18*, 821-829.

Zerbino, D.R., McEwen, G.K., Margulies, E.H., and Birney, E. (2009). Pebble and Rock Band: Heuristic Resolution of Repeats and Scaffolding in the Velvet Short-Read <italic>de Novo</italic> Assembler. PLoS ONE *4*, e8407.

Zhang, J., Ruhlman, T.A., Sabir, J.S.M., Blazier, J.C., Weng, M.-L., Park, S., and Jansen, R.K. (2016). Coevolution between Nuclear-Encoded DNA Replication, Recombination, and Repair Genes and Plastid Genome Complexity. Genome Biol Evol *8*, 622-634.

Zhang, L., Pond, S.K., and Gaut, B.S. (2001). A Survey of the Molecular Evolutionary Dynamics of Twenty-Five Multigene Families from Four Grass Taxa. J Mol Evol *52*, 144-156.

Zhu, L., Wang, Q., Tang, P., Araki, H., and Tian, D. (2009). Genomewide Association between Insertions/Deletions and the Nucleotide Diversity in Bacteria. Mol Biol Evol *26*, 2353-2361.

Zufall, R.A., McGrath, C.L., Muse, S.V., and Katz, L.A. (2006). Genome Architecture Drives Protein Evolution in Ciliates. Mol Biol Evol *23*, 1681-1687.

## Chapter 3


Adl, S. M., A. G. B. Simpson, et al. (2005). "The new higher level classification of eukaryotes with emphasis on the taxonomy of protists." The Journal of Eukaryotic Microbiology **52**(5): 399-451.

Albalat, R. and C. Canestro (2016). "Evolution by gene loss." Nature Reviews Genetics 17(7): 379-391.

Alverson, A. J., B. Beszteri, et al. (2011). "The model marine diatom *Thalassiosira pseudonana* likely descended from a freshwater ancestor in the genus *Cyclotella*." BMC Evolutionary Biology 11: 125.

Alverson, A. J., R. K. Jansen, et al. (2009). "Response to Medlin and Kaczmarska (2008)." Molecular Phylogenetics and Evolution 50(2): 409-410.

Ashworth, M. P., E. C. Ruck, et al. (2012). "A revision of the genus Cyclophora and description of Astrosyne gen. nov. (Bacillariophyta), two genera with the pyrenoids contained within pseudosepta." Phycologia 51(6): 684-699.

Bowler, C., A. E. Allen, et al. (2008). "The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes." Nature 456(7219): 239-244.

Brembu, T., P. Winge, et al. (2013). "The chloroplast genome of the diatom Seminavis robusta: New features introduced through multiple mechanisms of horizontal gene transfer." Mar Genomics 21(13): 00080-00089.

CHESNICK, J. M., W. H. C. F. KOOISTRA, et al. (1997). "Ribosomal RNA Analysis Indicates a Benthic Pennate Diatom Ancestry for the Endosymbionts of the Dinoflagellates <i>Peridinium foliaceum</i> and <i>Peridinium balticum</i> (Pyrrhophyta)." Journal of Eurkaryotic Microbiology 44(4): 314-320.

D'Alelio, D. and M. V. Ruggiero Interspecific plastidial recombination in the diatom genus Pseudo-nitzschia, J Phycol. 2015 Dec;51(6):1024-8. doi: 10.1111/jpy.12350. Epub 2015 Oct 23.

Darling, A. C., B. Mau, et al. (2004). "Mauve: multiple alignment of conserved genomic sequence with rearrangements." Genome Research 14(7): 1394-1403.

Doyle, J. J. (1987). "A rapid DNA isolation procedure for small quantities of fresh leaf tissue." Phytochem bull 19: 11-15.

Drummond, A. J. and e. al (2010). "Geneious v5.5." Available at http://www.geneious.com.

Ehara, M., Y. Inagaki, et al. (2000). "Phylogenetic analysis of diatom coxI genes and implications of a fluctuating GC content on mitochondrial genetic code evolution." Current Genetics 37(1): 29-33.

Janouškovec, J., S.-L. Liu, et al. (2013). "Evolution of Red Algal Plastid Genomes: Ancient Architectures, Introns, Horizontal Gene Transfer, and Taxonomic Utility of Plastid Markers." PLoS ONE 8(3): e59001.

Jansen, R. and T. Ruhlman (2012). Plastid Genomes of Seed Plants. Genomics of Chloroplasts and Mitochondria. R. Bock and V. Knoop, Springer Netherlands. 35: 103-126.
Jansen, R. K., Z. Cai, et al. (2007). "Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns." Proceedings of the National Academy of Sciences 104(49): 19369-19374.
Katoh, K., K. Kuma, et al. (2005). "MAFFT version 5: improvement in accuracy of multiple sequence alignment." Nucleic Acids Research 33(2): 511 - 518.
Langmead, B. and S. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods 9(4): 357 - 359.
Lommer, M., A. S. Roy, et al. (2010). "Recent transfer of an iron-regulated gene from the plastid to the nuclear genome in an oceanic diatom adapted to chronic iron limitation." BMC Genomics 11.
Lommer, M., M. Specht, et al. (2012). "Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation." Genome Biology 13(7): R66.
Maddison, W. and D. Maddison (2000). "MacClade 4: Analysis of Phylogeny and Character Evolution."
Marchler-Bauer, A. and S. H. Bryant (2004). "CD-Search: protein domain annotations on the fly." Nucleic Acids Research 32(suppl_2): W327-W331.
Medlin, L. K. (2017). "Evolution of the diatoms: IX. Two datasets resolving monophyletic Classes of diatoms are used to explore the validity of adding short clone library sequences to the analysis." European Journal of Phycology: 1-14.
Medlin, L. K. and I. Kaczmarska (2004). "Evolution of the diatoms V: Morphological and cytological support for the major clades and a taxonomic revision." Phycologia 43(3): 245-270.
Moustafa, A., B. Beszteri, et al. (2009). "Genomic footprints of a cryptic plastid endosymbiosis in diatoms." Science 324(5935): 1724 - 1726.
Nelson, D. M., P. Treguer, et al. (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): 359-372.
Oudot-Le Secq, M.-P., J. Grimwood, et al. (2007). "Chloroplast genomes of the diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana and comparison with other plastid genomes of the red lineage." Molecular Genetics and Genomics 277(4): 427-439.
Paradis, E., J. Claude, et al. (2004). "APE: Analyses of Phylogenetics and Evolution in R language." Bioinformatics 20(2): 289-290.
Round, F. E. and R. M. Crawford (1981). "The lines of evolution of the Bacillariophyta .1. Origin." Proceedings of the Royal Society of London Series B-Biological Sciences 211(1183): 237-&.

Round, F. E. and R. M. Crawford (1984). "The lines of evolution of the Bacillariophyta. 2. The Centric Series." Proceedings of the Royal Society of London Series B-Biological Sciences 221(1223): 169-&.

Ruck, E. C., S. R. Linard, et al. (2016). "Hoarding and horizontal transfer led to an expanded gene and intron repertoire in the plastid genome of the diatom, Toxarium undulatum (Bacillariophyta)." Current Genetics: 1-9.

Ruck, E. C., T. Nakov, et al. (2014). "Serial gene losses and foreign DNA underlie size and sequence variation in the plastid genomes of diatoms." Genome Biol Evol 6(3): 644-654.

Sabir, J. S. M., M. Yu, et al. (2014). "Conserved Gene Order and Expanded Inverted Repeats Characterize Plastid Genomes of Thalassiosirales." PLoS ONE 9(9): e107854.

Schattner, P., A. N. Brooks, et al. (2005). "The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs." Nucleic Acids Research 33(Web Server issue): W686-689.

Schwarz, E. N., T. A. Ruhlman, et al. (2017). "Plastome-Wide Nucleotide Substitution Rates Reveal Accelerated Rates in Papilionoideae and Correlations with Genome Features Across Legume Subfamilies." Journal of Molecular Evolution 84(4): 187-203.

Shimodaira, H. (2002). "An approximately unbiased test of phylogenetic tree selection." Systematic Biology 51: 492 - 508.

Simonsen, R. (1979). "The diatom system: Ideas on phylogeny." Bacillaria 2: 9-71.

Sorhannus, U. and M. G. Fox (2011). "Phylogenetic Analyses of a Combined Data Set Suggest that the Attheya Lineage is the Closest Living Relative of the Pennate Diatoms (Bacillariophyceae)." Protist(0).

Stamatakis, A. (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." Bioinformatics 30(9): 1312-1313.

Steinecke, F. (1931). Die Phylogenie der Algophyten: Versuch einer morphologischen Begründung des natürtichen Systems der Algen, Niemeyer.

Sullivan, A. R., B. Schiffthaler, et al. (2017). "Interspecific Plastome Recombination Reflects Ancient Reticulate Evolution in Picea (Pinaceae)." Molecular Biology and Evolution 34(7): 1689-1701.

Swofford, D. (2003). "PAUP*: Phylogenetic Analysis Using Parsimony. (* and other methods)." ver. 4.0b10 edn.

Tajima, N., K. Saitoh, et al. (2016). "Sequencing and analysis of the complete organellar genomes of Parmales, a closely related group to Bacillariophyta (diatoms)." Current Genetics 62(4): 887-896.

Tesler, G. (2002). "GRIMM: genome rearrangements web server." Bioinformatics 18: 492 - 493.

Theriot, E., E. Ruck, et al. (2011). Status of the Pursuit of the Diatom Phylogeny: Are Traditional Views and New Molecular Paradigms Really That Different? The Diatom World. J. Seckbach and P. Kociolek, Springer Netherlands. 19: 119-142.

Theriot, E. C., M. Ashworth, et al. (2010). "A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research " Plant Ecology and Evolution 143(3): 278-296.

Theriot, E. C., M. P. Ashworth, et al. (2015). "Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling." Molecular Phylogenetics and Evolution 89: 28-36.

Theriot, E. C., J. J. Cannone, et al. (2009). "The limits of nuclear-encoded SSU rDNA for resolving the diatom phylogeny." European Journal of Phycology 44(3): 277-290.

Van de Peer, Y., G. Van der Auwera, et al. (1996). "The evolution of stramenopiles and alveolates as derived by "substitution rate calibration" of small ribosomal subunit RNA." Journal of Molecular Evolution 42(2): 201-210.

Weng, M.-L., J. C. Blazier, et al. (2013). "Reconstruction of the Ancestral Plastid Genome in Geraniaceae Reveals a Correlation Between Genome Rearrangements, Repeats and Nucleotide Substitution Rates." Molecular Biology and Evolution.

Wyman, S., R. Jansen, et al. (2004). "Automatic annotation of organellar genomes with DOGMA." Bioinformatics 20(17): 3252 - 3255.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Research 18(5): 821-829.

Zerbino, D. R., G. K. McEwen, et al. (2009). "Pebble and Rock Band: Heuristic Resolution of Repeats and Scaffolding in the Velvet Short-Read <italic>de Novo</italic> Assembler." PLoS ONE 4(12): e8407.

## Chapter 4

Alverson, A. J. (2007). "Strong purifying selection in the silicon transporters of marine and freshwater diatoms." Limnology and Oceanography 52(4): 1420-1429.

Barraclough, T. and V. Savolainen (2001). "Evolutionary rates and species diversity in flowering plants." Evolution 55: 677 - 683.

Belda, E., A. Moya, et al. (2005). "Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria." Molecular Biology and Evolution 22(6): 1456-1467.

Blazier, J. C., T. A. Ruhlman, et al. (2016). "Divergence of RNA polymerase α subunits in angiosperm plastid genomes is mediated by genomic rearrangement." 6: 24595.

Bowler, C., A. E. Allen, et al. (2008). "The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes." Nature 456(7219): 239-244.

Bradwell, K., M. Combe, et al. (2013). "Correlation Between Mutation Rate and Genome Size in Riboviruses: Mutation Rate of Bacteriophage Qβ." Genetics 195(1): 243-251.

Britten, R. J. (1986). "Rates of DNA sequence evolution differ between taxonomic groups." Science 231.

Buschiazzo, E., C. Ritland, et al. (2012). "Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms." BMC Evolutionary Biology 12(1): 8.

Chang, C., H. Lin, et al. (2006). "The chloroplast genome of Phalaenopsis aphrodite (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications." Molecular Biology and Evolution 23(2): 279 - 291.

Christensen, A. C. (2013). "Plant Mitochondrial Genome Evolution Can Be Explained by DNA Repair Mechanisms." Genome Biology and Evolution 5(6): 1079-1086.

Darling, A. C., B. Mau, et al. (2004). "Mauve: multiple alignment of conserved genomic sequence with rearrangements." Genome Research 14(7): 1394-1403.

De La Torre, A. R., Y.-C. Lin, et al. (2015). "Genome-Wide Analysis Reveals Diverged Patterns of Codon Bias, Gene Expression, and Rates of Sequence Evolution in Picea Gene Families." Genome Biology and Evolution 7(4): 1002-1015.

Dettman, J. R., J. L. Sztepanacz, et al. (2016). "The properties of spontaneous mutations in the opportunistic pathogen Pseudomonas aeruginosa." BMC Genomics 17(1): 27.

Drake, J. W. (1991). "A constant rate of spontaneous mutation in DNA-based microbes." Proceedings of the National Academy of Sciences 88(16): 7160-7164.

Drake, J. W., B. Charlesworth, et al. (1998). "Rates of Spontaneous Mutation." Genetics 148(4): 1667.

Drummond, D. A., A. Raval, et al. (2006). "A single determinant dominates the rate of yeast protein evolution." Molecular Biology and Evolution 23(2): 327-337.

Du, X., D. J. Lipman, et al. (2013). "Why Does a Protein's Evolutionary Rate Vary over Time?" Genome Biology and Evolution 5(3): 494-503.
Duchene, D. and L. Bromham (2013). "Rates of molecular evolution and diversification in plants: chloroplast substitution rates correlate with species-richness in the Proteaceae." BMC Evolutionary Biology 13(1): 65.
Guisinger, M., J. Kuehl, et al. (2008). "Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions." Proceedings of the National Academy of Sciences of the United States of America 105: 18424 - 18429.
Guisinger, M., J. Kuehl, et al. (2011). "Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage." Molecular Biology and Evolution 28: 583 - 600.
Guisinger, M. M., T. W. Chumley, et al. (2010). "Implications of the Plastid Genome Sequence of Typha (Typhaceae, Poales) for Understanding Genome Evolution in Poaceae." Journal of Molecular Evolution 70(2): 149-166.
Guisinger, M. M., J. V. Kuehl, et al. (2008). "Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions." Proceedings of the National Academy of Sciences of the United States of America 105(47): 18424-18429.
Hodgkinson, A. and A. Eyre-Walker (2011). "Variation in the mutation rate across mammalian genomes." Nature Reviews Genetics 12(11): 756-766.
Hollister, J. D., J. Ross-Ibarra, et al. (2010). "Indel-Associated Mutation Rate Varies with Mating System in Flowering Plants." Molecular Biology and Evolution 27(2): 409-416.
Hu, J. and J. L. Blanchard (2009). "Environmental Sequence Data from the Sargasso Sea Reveal That the Characteristics of Genome Reduction in Prochlorococcus Are Not a Harbinger for an Escalation in Genetic Drift." Molecular Biology and Evolution 26(1): 5-13.
Jancek, S., S. Gourbière, et al. (2008). "Clues about the Genetic Basis of Adaptation Emerge from Comparing the Proteomes of Two Ostreococcus Ecotypes (Chlorophyta, Prasinophyceae)." Molecular Biology and Evolution 25(11): 2293-2300.
Jansen, R. and T. Ruhlman (2012). Plastid Genomes of Seed Plants. Genomics of Chloroplasts and Mitochondria. R. Bock and V. Knoop, Springer Netherlands. 35: 103-126.
Jansen, R. K., Z. Cai, et al. (2007). "Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns." Proceedings of the National Academy of Sciences 104(49): 19369-19374.
Katoh, K., K. Kuma, et al. (2005). "MAFFT version 5: improvement in accuracy of multiple sequence alignment." Nucleic Acids Research 33(2): 511 - 518.

Koester, J. A., W. J. Swanson, et al. (2013). "Positive selection within a diatom species acts on putative protein interactions and transcriptional regulation." Molecular Biology and Evolution 30(2): 422-434.

Kooistra, W. H. C. F., R. Gersonde, et al. (2007). CHAPTER 11 - The Origin and Evolution of the Diatoms: Their Adaptation to a Planktonic Existence A2 - Falkowski, Paul G. Evolution of Primary Producers in the Sea. A. H. Knoll. Burlington, Academic Press: 207-249.

Lanfear, R., S. Y. W. Ho, et al. (2010). "Mutation rate is linked to diversification in birds." Proc Natl Acad Sci USA 107.

Lee, E. K., A. Cibrian-Jaramillo, et al. (2011). "A Functional Phylogenomic View of the Seed Plants." Plos Genetics 7(12): e1002411.

Lemieux, C., C. Otis, et al. (2016). "Comparative Chloroplast Genome Analyses of Streptophyte Green Algae Uncover Major Structural Alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae." Frontiers in Plant Science 7: 697.

Liao, B.-Y., N. M. Scott, et al. (2006). "Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate of Mammalian Proteins." Molecular Biology and Evolution 23(11): 2072-2080.

Lynch, M., B. Koskella, et al. (2006). "Mutation Pressure and the Evolution of Organelle Genomic Architecture." Science 311(5768): 1727-1730.

Mock, T., R. P. Otillar, et al. (2017). "Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus." Nature 541(7638): 536-540.

Ruck, E. C., T. Nakov, et al. (2014). "Serial gene losses and foreign DNA underlie size and sequence variation in the plastid genomes of diatoms." Genome Biol Evol 6(3): 644-654.

Sabir, J. S. M., M. Yu, et al. (2014). "Conserved Gene Order and Expanded Inverted Repeats Characterize Plastid Genomes of Thalassiosirales." PLoS ONE 9(9): e107854.

Schwarz, E. N., T. A. Ruhlman, et al. (2017). "Plastome-Wide Nucleotide Substitution Rates Reveal Accelerated Rates in Papilionoideae and Correlations with Genome Features Across Legume Subfamilies." Journal of Molecular Evolution 84(4): 187-203.

Shao, R., M. Dowton, et al. (2003). "Rates of gene rearrangement and nucleotide substitution are correlated in the mitochondrial genomes of insects." Molecular Biology and Evolution 20(10): 1612-1619.

Sharp, P. M. (1991). "Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: codon usage, map position, and concerted evolution." Journal of Molecular Evolution 33(1): 23-33.

Shields, D. C., P. M. Sharp, et al. (1988). ""Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons." Molecular Biology and Evolution 5(6): 704-716.

Sloan, D. B., A. J. Alverson, et al. (2012). "Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates." Plos Biology 10(1): e1001241.

Smith, S. A. and M. J. Donoghue (2008). "Rates of molecular evolution are linked to life history in flowering plants." Science 322.

Sorhannus, U. (2007). "A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution." Marine Micropaleontology 65(1-2): 1-12.

Sorhannus, U. and M. Fox (1999). "Synonymous and nonsynonymous substitution rates in diatoms: a comparison between chloroplast and nuclear genes." Journal of Molecular Evolution 48(2): 209-212.

Stamatakis, A. (2006). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." Bioinformatics 22(21): 2688-2690.

Tesler, G. (2002). "GRIMM: genome rearrangements web server." Bioinformatics 18: 492 - 493.

Tian, D., Q. Wang, et al. (2008). "Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes." Nature 455(7209): 105-108.

Weng, M.-L., J. C. Blazier, et al. (2013). "Reconstruction of the Ancestral Plastid Genome in Geraniaceae Reveals a Correlation Between Genome Rearrangements, Repeats and Nucleotide Substitution Rates." Molecular Biology and Evolution.

Wicke, S., B. Schäferhoff, et al. (2014). "Disproportional Plastome-Wide Increase of Substitution Rates and Relaxed Purifying Selection in Genes of Carnivorous Lentibulariaceae." Molecular Biology and Evolution 31(3): 529-545.

Wilson, A. C., S. S. Carlson, et al. (1977). "Biochemical evolution." Annual Review of Biochemistry 46: 573-639.

Wolf, P. G., J. P. Der, et al. (2011). "The evolution of chloroplast genes and genomes in ferns." Plant Molecular Biology 76(3): 251-261.

Wolfe, K., W. Li, et al. (1987). "Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs." Proc Natl Acad Sci USA 84: 9054 - 9058.

Wu, C.-S. and S.-M. Chaw (2015). "Evolutionary Stasis in Cycad Plastomes and the First Case of Plastome GC-Biased Gene Conversion." Genome Biology and Evolution 7(7): 2000-2009.

Wu, C. S. and S. M. Chaw (2014). "Highly rearranged and size-variable chloroplast genomes in conifers II clade (cupressophytes): evolution towards shorter intergenic spacers." Plant Biotechnol J 12(3): 344-353.

Xu, Q., G. Xiong, et al. (2012). "Analysis of Complete Nucleotide Sequences of 12 Gossypium Chloroplast Genomes: Origin and Evolution of Allotetraploids." PLoS ONE 7(8): e37128.

Xu, W., D. Jameson, et al. (2006). "The relationship between the rate of molecular evolution and the rate of genome rearrangement in animal mitochondrial genomes." Journal of Molecular Evolution 63(3): 375-392.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Molecular Biology and Evolution 24(8): 1586 - 1591.

Zhang, J., T. A. Ruhlman, et al. (2016). "Coevolution between Nuclear-Encoded DNA Replication, Recombination, and Repair Genes and Plastid Genome Complexity." Genome Biology and Evolution 8(3): 622-634.

Zhang, L., S. K. Pond, et al. (2001). "A Survey of the Molecular Evolutionary Dynamics of Twenty-Five Multigene Families from Four Grass Taxa." Journal of Molecular Evolution 52(2): 144-156.

Zhu, L., Q. Wang, et al. (2009). "Genomewide Association between Insertions/Deletions and the Nucleotide Diversity in Bacteria." Molecular Biology and Evolution 26(10): 2353-2361.

Zufall, R. A., C. L. McGrath, et al. (2006). "Genome Architecture Drives Protein Evolution in Ciliates." Molecular Biology and Evolution 23(9): 1681-1687.

# Vita

Mengjie Yu graduated with her bachelor degree in Biotechnology in Jinan University, Guangzhou, China in 2010. She attended The University of Texas at Austin in 2010 fall to pursue PhD degree in Plant Biology focusing on phylogenomics. She received Integrative Biology recruitment fellowship and graduate school continuing fellowship in her first and last year at UT. She received multiple awards from the phycological society: International Phycological Congress financial support award in 2013, International Phycological Society Paul C. Silva Travel Award in 2014, International Diatom Symposium Student Travel Grant in 2014, Phycological Society Student Research Grant in 2014, Phycological Society of America Hoshaw Travel award in 2015. She was awarded the best student poster award in the 23rd International Diatom Symposium in Nanjing, China. During the summers of 2015 and 2016, she interned at Takeda Pharmaceutical Computation Biology group and Dow Agrosciences Mathematics and Statistics group, respectively. In 2017 spring, she completed M.S. in Statistics from The University of Texas at Austin Statistics and Data Science Department. Mengjie Yu currently works as a Data Scientist at Intel Corporation in California.

Permanent email: annayu2010@gmail.com

This dissertation was typed by Mengjie Yu.