RESEARCH ARTICLE

Extensive chloroplast genome rearrangement amongst three closely related *Halamphora* spp. (Bacillariophyceae), and evidence for rapid evolution as compared to land plants

Sarah E. Hamsher[1,2]*, Kyle G. Keepers[3], Cloe S. Pogoda[3], Joshua G. Stepanek[4], Nolan C. Kane[3], J. Patrick Kociolek[3,5]

1 Department of Biology, Grand Valley State University, Allendale, Michigan, United States of America, **2** Annis Water Resources Institute, Grand Valley State University, Muskegon, Michigan, United States of America, **3** Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado, United States of America, **4** Department of Biology, Colorado Mountain College, Edwards, Colorado, United States of America, **5** Museum of Natural History, University of Colorado, Boulder, Colorado, United States of America

* hamshers@gvsu.edu

OPEN ACCESS

Citation: Hamsher SE, Keepers KG, Pogoda CS, Stepanek JG, Kane NC, Kociolek JP (2019) Extensive chloroplast genome rearrangement amongst three closely related *Halamphora* spp. (Bacillariophyceae), and evidence for rapid evolution as compared to land plants. PLoS ONE 14(7): e0217824. https://doi.org/10.1371/journal.pone.0217824

Editor: Berthold Heinze, Austrian Federal Research Centre for Forests BFW, AUSTRIA

Received: October 5, 2018

Accepted: May 21, 2019

Published: July 3, 2019

Data Availability Statement: Data are available in the paper and its Supporting Information files. All accession numbers for the chloroplast genome data will be available from the GenBank database for publication (accession numbers MK045450-MK045452).

Funding: The author(s) received no specific funding for this work.

Abstract

Diatoms are the most diverse lineage of algae, but the diversity of their chloroplast genomes, particularly within a genus, has not been well documented. Herein, we present three chloroplast genomes from the genus *Halamphora* (*H*. *americana*, *H*. *calidilacuna*, and *H*. *coffeaeformis*), the first pennate diatom genus to be represented by more than one species. *Halamphora* chloroplast genomes ranged in size from ~120 to 150 kb, representing a 24% size difference within the genus. Differences in genome size were due to changes in the length of the inverted repeat region, length of intergenic regions, and the variable presence of ORFs that appear to encode as-yet-undescribed proteins. All three species shared a set of 161 core features but differed in the presence of two genes, *serC* and *tyrC* of foreign and unknown origin, respectively. A comparison of these data to three previously published chloroplast genomes in the non-pennate genus *Cyclotella* (Thalassiosirales) revealed that *Halamphora* has undergone extensive chloroplast genome rearrangement compared to other genera, as well as containing variation within the genus. Finally, a comparison of *Halamphora* chloroplast genomes to those of land plants indicates diatom chloroplast genomes within this genus may be evolving at least ~4–7 times faster than those of land plants. Studies such as these provide deeper insights into diatom chloroplast evolution and important genetic resources for future analyses.

Introduction

Diatoms are single-celled eukaryotic algae with silica cell walls. They play an important role in global O_2, CO_2, and silica cycling [1,2] and have the most efficient Rubisco known [3].

Competing interests: The authors have declared that no competing interests exist.

Diatoms are the most diverse group of eukaryotic algae [4] with an estimated 100,000 species [5] and occupy many niches in marine and freshwater environments [6].

Halamphora (Cleve) Levkov [7] is a recently described genus composed of species formerly assigned to the genus *Amphora* Ehrenberg ex Kützing. Members of the genus occur across a wide ecological spectrum including fresh to hypersaline habitats [8] from the tropics to the polar regions [9,10], and are known to be prodigious oil producers [11,12]. Although diverse, recent taxonomic treatments [7,8] and broadly sampled molecular phylogenetic analyses [13,14] have combined to make *Halamphora* one of the most well studied groups of diatoms in terms of phylogenetic systematics.

Despite diatom diversity and importance as primary producers in most aquatic ecosystems, the chloroplast genomes of only 40 species have been analyzed to date and taxon sampling has primarily focused on the 'polar' centric diatoms (45% of published chloroplast genomes; [15,16,17]). Outside of this group, no two species from within the same genus have been examined prior to this study. As the cost of genomic sequencing has decreased [18], it has become more feasible to examine in greater detail the genetic makeup of biologically important, non-model organisms. This allows for salient comparisons and insight into evolutionary lifestyle mechanisms at the genetic level. Finer scale taxonomic sampling in non-model organisms, such as within genus-level comparisons, elucidates the time-scales of evolutionary processes that may occur at rates too rapid to yield meaningful comparisons at more sparse taxonomic sampling regimens.

Therefore, the purpose of this study is to explore the phylogenetic and genomic relationships between three closely related *Halamphora* species (*H*. *americana* Kociolek in Kociolek et al., *H*. *calidilacuna* Stepanek & Kociolek, and *H*. *coffeaeformis* (C.Agardh; Levkov)). We then compared overall genomic content between these newly sequenced and annotated genomes to currently published diatom plastid genomes. Strikingly, *Halamphora* demonstrates high levels of gene rearrangement in comparison to the genus *Cyclotella*. The levels of gene rearrangements in the genus *Halamphora* are comparable to the level observed in much older-diverging lineages of the major groups of land plants (dicots and monocots).

Materials and methods

Ethics statement

No permits were required for collection of benthic samples in Salt Alkaline Lake, ND or Blue Lake Warm Spring, UT. No specific permissions were required for these locations/activities. Both sites are publicly accessible and there are no regulations regarding the collection of algae from these sites. Field studies did not involve endangered or protected species.

Isolate collection and culturing

Environmental samples containing *H*. *americana* and *H*. *coffeaeformis* were collected from Salt Alkaline Lake, ND in 2011 and samples containing *H*. *calidilacuna* were collected from Blue Lake Warm Spring, UT in 2012 (Table 1). Conductivity and pH measurements were recorded from the sites at the time of collection using a YSI 556 multi-probe (YSI Incorporated, Yellow Springs, Ohio, USA). Individual cells were isolated into monoculture by micropipette serial dilution and grown in artificial brackish water medium created using the sea salts Instant Ocean (Spectrum Brands, Inc., Blacksburg, Virginia, USA). Conductivity of the medium was adjusted to 10 mS cm $^{-1}$ to approximate the conductivity at the collection sites (Table 1) and added macro- and micronutrients were based on those of WC media [19] with the Na_2SiO_3 concentration increased to 56.85 mg l^{-1}. Cultures were maintained at ~25°C, under fluorescent illumination with a 12:12 light cycle at an irradiance of ~50 μmol cm^{-2} S^{-1}.

Table 1. Isolates utilized in this study.

Taxon	Voucher	Collection Locality	Latitude (˚N)	Longitude (˚W)	pH	Conductivity (mS cm^{-1})
H. americana	JPK7977-AMPH100	Salt Alkaline Lake, Kidder Co., ND, USA	46.95092	99.53915	8.89	9.811
H. calidilacuna	JPK8506-AMPH118	Blue Lake Warm Spring, Tooele Co., UT, USA	40.50257	114.0336	7.60	9.319
H. coffeaeformis	JPK7977-AMPH101	Salt Alkaline Lake, Kidder Co., ND, USA	46.95092	99.53915	8.89	9.811

https://doi.org/10.1371/journal.pone.0217824.t001

DNA extraction, library preparation, and sequencing

Cultures were harvested by centrifugation and DNA was extracted using the Qiagen DNeasy Plant Mini Kit (Qiagen, Crawley, UK) following manufacturer's protocols.

Library preparation and sequencing followed Pogoda et al. [20]. Briefly, genomic libraries were prepared using Nextera XT DNA library prep kits (Illumina). The protocol calls for 1 ng total of input DNA and each gDNA sample was diluted to the appropriate concentration using a Qubit 3.0 fluorometer (ThermoFisher Scientific). Each sample was barcoded by the unique dual index adapters Nextera i5 and i7. Resulting libraries were cleaned using solid phase reversible immobilization (SPRI) to remove fragment sizes less than 300 base pairs via an epMotion 5075TMX automated liquid handling system. Sample quality control (QC) was conducted prior to normalizing the loading concentration of pooled samples to 1.8–2.1 pM with 1% PhiX control v3 added (Illumina). Samples were processed for paired end 151 base pair reads on the Illumina NextSeq sequencer at the University of Colorado's BioFrontiers Institute Next-Generation Sequencing Facility in Boulder, Colorado. All wet lab work was performed in the Department of Ecology and Evolutionary Biology at the University of Colorado, Boulder.

Assembly and annotation

Raw de-multiplexed data were sub-sampled to approximately 2 GB per sample and trimmed using Trimmomatic-0.36 with the following parameters: ILLUMINACLIP:NexteraPE-PE. fa:2:20:10MINLEN:140 LEADING:20 TRAILING:20 [21]. Resulting fastq files were *de novo* assembled using SPAdes v3.9 with the following parameters: SPAdes-3.9.0-Linux/bin/spades. py—careful -k 35,55,85 [22]. Diatom chloroplast contigs were identified using command-line BLAST against *Phaeodactylum tricornutum* (accession EF067920) and confirmed using a web BLAST. Annotations were initiated in DOGMA [23] and completed in NCBI's Sequin 15.10 (Bethesda, MD) using the protein-coding sequence of 23 diatom chloroplast genomes as references (S1 Table). Putative features unique to these genomes (or not found in the reference genomes) were identified using NCBI's ORF Finder (Bethesda, MD). Annotated plastid genomes are available in GenBank using accession numbers MK045450 –MK045452.

Genome content was visualized using OGDraw v1.2 [24]. Total sequence length of protein coding genes (CDS), transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and non-coding DNA was calculated as a sum of the total content of each type of feature. Web BLAST was used to determine if unique ORFs identified using NCBI's ORF Finder contained any sequence similarity to ORFs found in other species' chloroplast genomes. A gene was considered present if there was an appropriate, full length BLAST hit to a reference diatom chloroplast. Otherwise, it was assumed absent. In addition, gene density (number of protein coding genes / genome size) was calculated [25].

Chloroplast phylogeny

Twenty molecular markers (*psaA*, *psbC*, *petD*, *petG*, *atpA*, *atpG*, *rbcL*, *rbcS*, *rpoA*, *rpoB*, *rps14*, *rpl33*, *rnl*, *rns*, *ycf89*, *sufB*, *sufC*, *dnaK*, *dnaB*, *clpC*) from 26 species were aligned (S1 File) using the MAFFT algorithm [26], manually edited as necessary, and partitioned by gene and codon position (except for ribosomal DNA regions). Maximum likelihood analysis with 50 independent tree searches and 1000 rapid bootstrap replicates was performed in RAxML [27] with the graphical user interface raxmlGUI ver. 1.2 [28] using the GTR+Γ+I model of evolution.

Pairwise genetic distances utilizing a Jukes-Cantor model of evolution were calculated for the twenty-marker concatenated alignment using R-Studio v1.1.456 using the 'dist.dna' function in the R package ape [29].

Synteny

Two separate alignments of biraphid pennate and thalassiosiroid diatoms and resulting locally collinear blocks (LCBs) were estimated with MAUVE 2.4.0 [30] after eliminating one copy of the inverted repeat (IR). Rearrangement distances between LCBs were measured using GRIMM 2.02 [31].

Results and discussion

General features

Halamphora chloroplast genomes varied in size from ~120–150 kb (Table 2), representing a size difference of ~20% within the genus. Despite this difference in size, they contained similar GC content (~30%; Table 2). The *Halamphora* genomes contain two canonical inverted repeats (IRs) separated by a small (SSC) and a large single-copy (LSC) region, a structure these genomes share with other diatoms (e.g., [16,32]), some red algae (e.g., [33]), glaucophytes (e.g. [34]), some green algae (e.g., [35]), and most land plants (e.g., [36]).

The three *Halamphora* genomes shared a set of 130 protein-coding genes, three rDNAs, 27 tRNAs, one tmRNA (transfer-messenger RNA) and ffs (signal recognition particle RNA) (Fig 1 and S1–S3 Figs) with a similar gene content to other pennate diatoms (S2 Table). With the exception of ORFs (see below), the chloroplast genomes of *H. americana*, *H. calidilacuna*, and *H. coffeaeformis* are nearly identical (99%) in gene content and differed only in the presence/ absence of two genes– phosphoserine aminotransferase (*serC*) and cyclohexadienyl dehydrogenase (*tyrC*), both of which have been suggested to be of foreign or unknown origin, respectively [15].

Among the *Halamphora* genomes, *serC* was only present, and then only as a pseudogene, in *H. calidilacuna* and is presumed to be of plasmid origin. Plasmids have been found in only some pennate diatoms thus far [37,38]. *serC* has been found in five other diatom plastid genomes as a gene/pseudogene (S2 Table) and in plasmids of *Cylindrotheca* species [15,37,38]. An additional indication of the plasmid origin of this gene in *H. calidilacuna* is the presence of

Table 2. Quantification of features in three *Halamphora* species, including % non-coding sequence, % GC, the number of ORFs (the number of ORFs shared by > 1 *Halamphora* genome), and gene density.

Taxon	LSC	SSC	IR length	Genome size	tRNA	rRNA	CDS[a]	Feature encoding	Non-feature encoding	% non-coding	GC (%)	ORFs (shared)	Gene Density[b]
Halamphora americana	77,289	44,724	10,269	142,551	2202	8720	95,428	106,350	36,201	25.4	32	7 (1)	0.92
H. calidilacuna	82,227	49,698	9407	150,739	2199	8710	102,024	112,933	37,806	25.1	32	15 (1)	0.88
H. coffeaeformis	64,938	41,485	7752	121,927	2199	8708	91,032	101,939	19,988	16.4	31	2	1.07

LSC, large single copy region; SSC, small single copy region; IR, inverted repeat; CDS, coding sequence; ORF, open reading frame.

[a] Includes coding sequence and intronic ORFs

[b] Number of protein-coding genes / genome size

https://doi.org/10.1371/journal.pone.0217824.t002

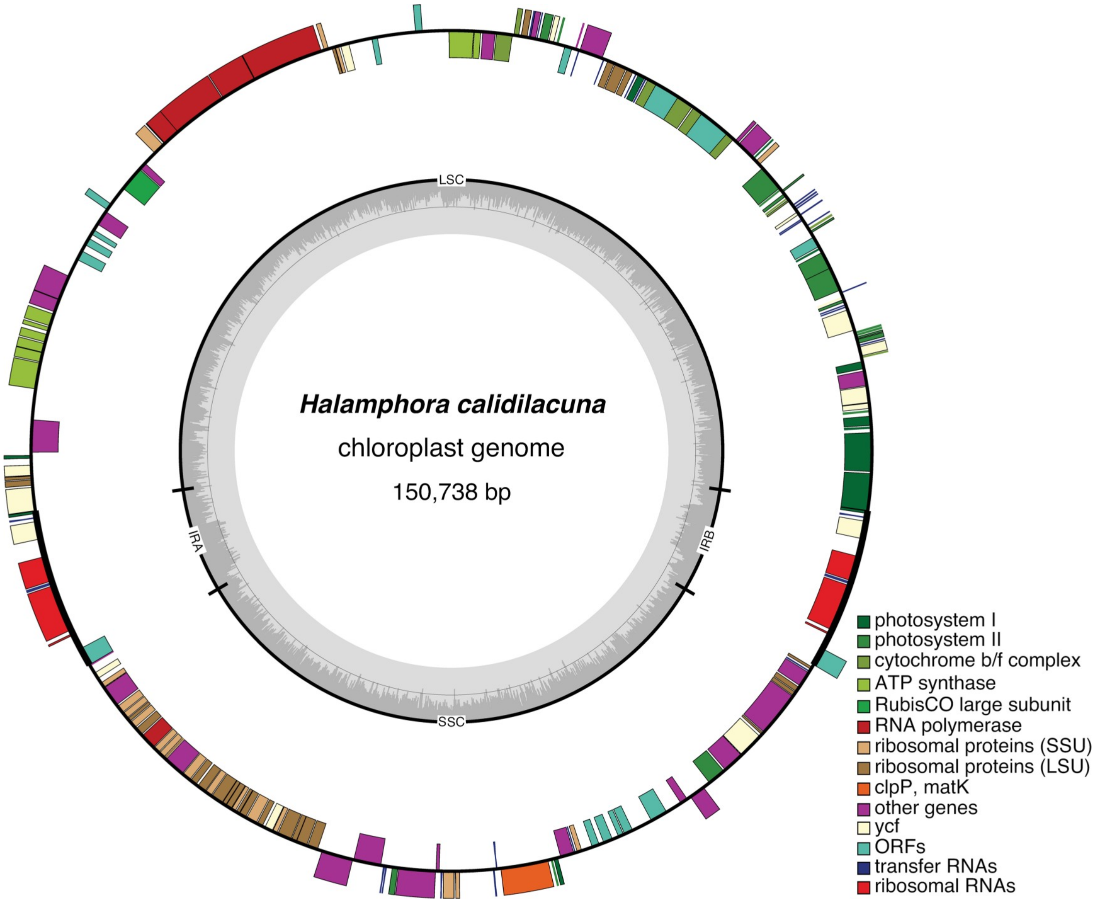

Fig 1. Plastid genome map of *Halamphora calidilacuna*. Genes on the outside are transcribed clockwise and those on the inside counterclockwise. The inner ring displays GC content in grey.

https://doi.org/10.1371/journal.pone.0217824.g001

a partial (presumably non-functional) copy of ORF484 in the chloroplast genome, an ORF also found in the pCf2 plasmid of *C*. *fusiformis* [37,38].

The presence/absence of *tyrC* follows an interesting pattern, being present (and presumably functional) in *H*. *calidilacuna*, present and presumably non-functional pseudogene in *H*. *americana*, and absent in *H*. *coffeaeformis*. Although *tyrC* has also been found in other diatoms, green algae, and bacteria, its origin is less clear [15]. The presence as a gene/pseudogene in the two most closely related *Halamphora* taxa (Fig 2) and absence in the other taxon could indicate an acquisition via plasmid or from bacterial horizontal gene transfer (HGT) prior to the split of *H*. *calidilacuna* and *H*. *americana*, followed by a subsequent loss of function in *H*. *americana*. Alternatively, it is possible that all *Halamphora* species contain *tyrC*, but it is in transition to relocating to the nucleus.

Differences in genome size are due to difference in length of the IRs and intergenic regions (Table 2). Gene density (calculated as the number of protein coding genes / genome size) was inversely proportional to genome size, with the greatest gene density in the smallest genome and vice versa (Table 2).

The largest *Halamphora* genome (*H*. *calidilacuna*; Table 2) also contains more open reading frames (ORFs) than the other congeners. Three ORFs are shared between at least two *Halamphora* genomes, but no ORF is common to all three genomes (S2 Table). Only two ORFs (ORF385, ORF484) were shared between *Halamphora* species and other diatom genomes.

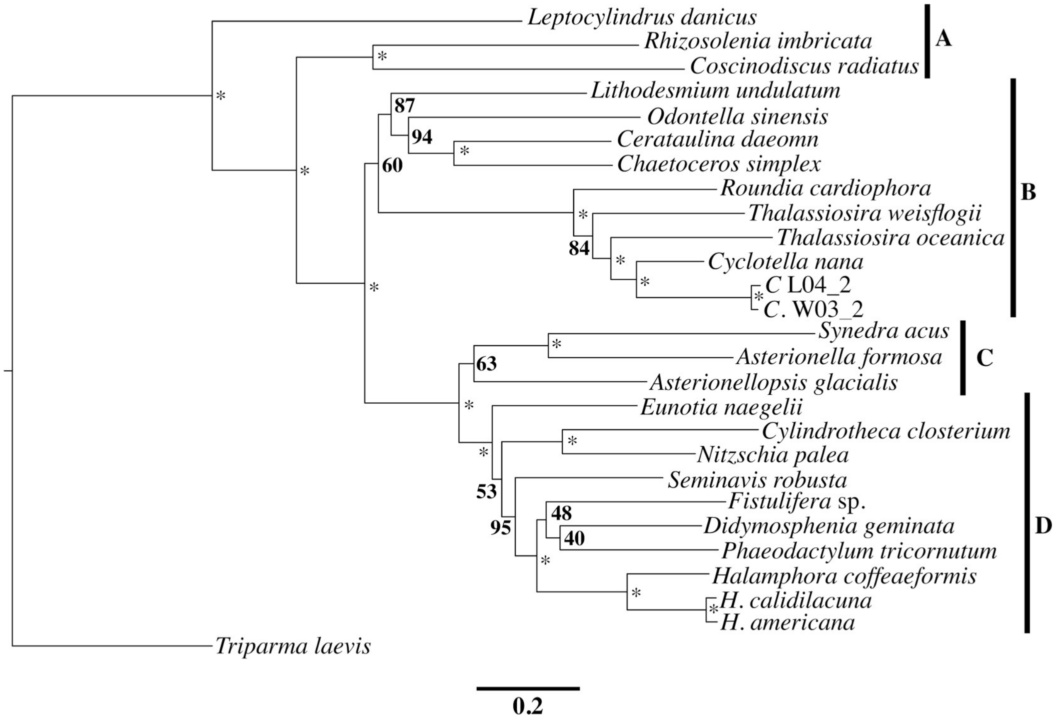

Fig 2. Maximum likelihood phylogram inferred from twenty chloroplast encoded markers (see Materials and methods). Node support is given as maximum likelihood bootstrap values (1000 bootstrap replicates). Asterisks indicate 100% support. Letters indicate morphological groups of diatoms as follows: (A) 'radial' centric; (B) 'polar' centric; (C) araphid; and (D) biraphid pennate diatoms.

https://doi.org/10.1371/journal.pone.0217824.g002

ORF385, which was free standing in the genome, was found in *H. calidilacuna*, *Seminavis robusta*, and *Asterionellopsis glacialis*. A partial (presumably non-functional) copy of ORF484, also free standing, was found in *H. calidilacuna* and within a plasmid of *Cylindrotheca fusiformis* [37,38]. Similar to the retrotransposons found in diatom mitochondrial genomes [20], two group II introns containing regions with homology to reverse transcriptases/maturases (ORF26 & ORF27) were found in *H. calidilacuna* and *S. robusta*. In *H. calidilacuna*, ORF26 was present in an intron of *petD* and ORF27 was in an intron of *petB*. These two ORFs show some similarity (~70% similarity at > 86% coverage) to group II introns found in Ulvophytes [39] and red algae [33].

The nucleotide identity between sister taxa *H. calidilacuna* and *H. americana* is 98%, and the identity between either of these taxa and *H. coffeaeformis* is 88%. In addition to overall sequence similarity, the utility of the *rbcL* (~1400 bp) and *rbcL*-3P (748 bp) as barcode markers [40] to distinguish between these closely related taxa was also evaluated. Both *rbcL* barcode markers (both phylogenetic and shorter *rbc*L-3P lengths; [40]) show the same pattern as the overall similarity; 1% divergence was observed between *H. calidilacuna* and *H. americana* and 5–6% (*rbc*L and *rbc*L-3P, respectively) divergence was observed between *H. coffeaeformis* and either of these taxa. Therefore, either of these barcoding markers could be used to distinguish between these closely related *Halamphora* species.

Seven pairs of overlapping genes were found in one or more *Halamphora* species, with some overlapping pairs also found in the Thalassiosirales. Overlapping pairs found in both *Halamphora* spp. and Thalassiosirales include: *psbD-psbC* by 53 bp; *atpF-atpD* by 4 bp; and *rpl23-rpl4* by 11 bp in *Halamphora* spp. and 8–17 bp in the Thalassiosirales. Additional overlapping pairs include: *sufB-sufC* by 1 bp (*H. americana*, *H. calidilacuna*, and Thalassiosirales), *ycf45-psaB* by 4 bp (*H. coffeaeformis* and *H. calidilacuna*), within ORF25 by 15 bp (*H. americana*); and ORF385-*serC* by 2 bp (*H. calidilacuna*). Despite the widespread occurrence of overlapping genes, the origin, evolution and ramification of these overlaps remains unknown [41]. Some overlapping genes (e.g., *psbD-psbC*) are known to cause translational coupling, i.e, the translation of the *psbC* cistron depends on the translation of the *psbD* cistron [42]. Overlapping genes may also produce novel *de novo* proteins, a process common in viral genomes that can lead to changes in pathogenicity and possibly genome evolution [43]. Another type of alternative transcription (i.e., intron retention) is important to diatoms' ability to adapt to changing nutrient conditions and not trivial in maintaining their physiology [44].

Phylogeny results

The resulting phylogeny (Fig 2) revealed 'polar' centric (B), araphid (C), and biraphid pennate (D) diatoms to be monophyletic groups and 'radial' centric (A) diatoms to be a paraphyletic group. The monophyly of araphid diatoms is most likely due to the limited taxon sampling in our tree, as this group is often paraphyletic in analyses that include a larger number of taxa in this group [16,45,46]. The monophyly of 'polar' centric diatoms may be due to taxon sampling as well [16,45,46], but some studies have shown this group to be monophyletic (e.g., [47,48]) and this is an area of ongoing research. Within the biraphid pennate diatoms, the relationships between the genus *Halamphora* and the remaining taxa largely agrees with the multi-gene phylogenies of Ruck & Theriot [49] and Stepanek & Kociolek [13]. However, a single gene (18S rDNA) phylogeny presented by Zgrundo et al. [50], which includes taxa from the genus *Fistulifera* H. Lange-Bertalot, recovered this genus within a clade consistently more closely related to the *Halamphora* than to the rest of the biraphid pennate diatoms [13,49,50]. The 20-gene phylogeny presented here continues to strongly support (BS 100) the monophyly of the genus *Halamphora* as well as the close sister relationship between *H. americana* and *H. calidilacuna* that have been recovered in several broadly-sampled multigene phylogenies of the group [13,14].

Synteny results

The MAUVE alignment of biraphid pennate diatoms resulted in 26 locally collinear blocks (LCBs) of sequence among the ten diatoms examined (Fig 3). This gene order comparison revealed inversions, translocations, and inversion/translocation combinations resulting in extensive plastid genome rearrangement across this group (Fig 3). Distances between gene orders of LCBs calculated using GRIMM revealed the gene order of *Cylindrotheca closterium* to be the most unique among these diatoms (S3 Table). Although gene order was more conserved within the clade containing *Didymosphenia geminata*, *Phaeodactylum tricornutum*, *H. americana*, *H. calidilacuna*, and *H. coffeaeformis*, no chloroplast genomes in these analyses had identical gene order (Fig 3 and S3 Table). Within *Halamphora*, gene order was not conserved and distances between gene orders of LCBs were ≥ 3 (S3 Table).

Gene order of the thalassiosiroid diatoms [15,16] was examined as a comparison to the rearrangements observed within the biraphid pennate diatoms including *Halamphora*. The MAUVE alignment of thalassiosiroid diatoms resulted in 18 locally collinear blocks (LCBs) of sequence among the six diatoms examined (Fig 4). With the exception of *T. oceanica*, gene order is more conserved in the thalassiosiroid diatoms (Fig 4) and the resulting rearrangement distances are smaller (S4 Table and Fig 5), a pattern consistent with the findings of Ruck et al. [15] and Sabir et al. [16]. In particular, there was only one inversion between *Cyclotella* species (Fig 4), in contrast to the inversions, translocations, and inversion/translocation combinations within *Halamphora* spp. (Fig 3). A comparison of Jukes-Cantor genetic distance to

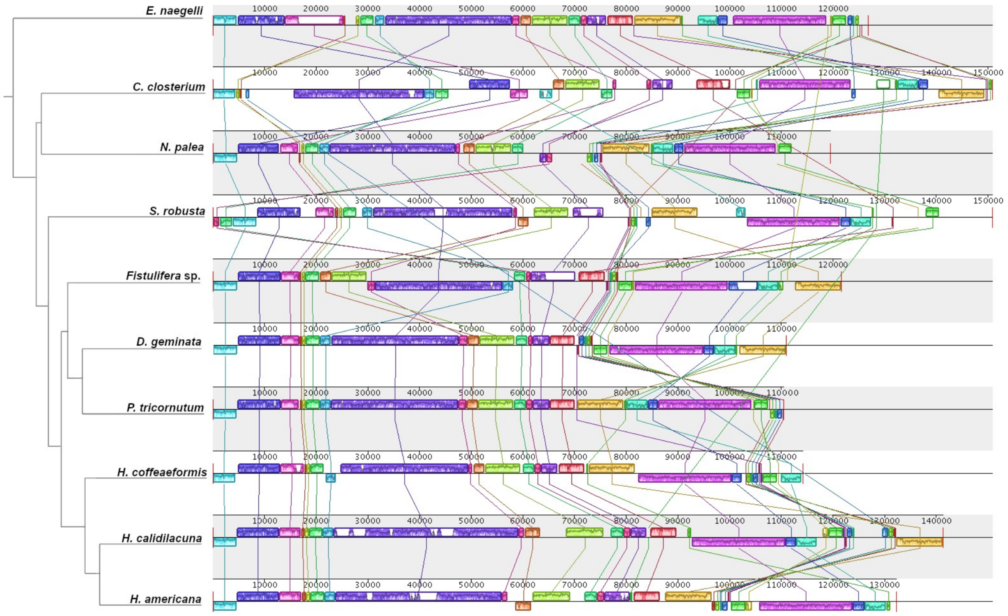

Fig 3. Gene order comparison of plastid genomes of eight biraphid pennate diatoms (three *Halamphora* spp. are from this study) with one copy of the inverted repeat removed prior to analysis. Alignment and resulting locally collinear blocks (LCBs) were generated using MAUVE. Relationships between taxa displayed as a cladogram to the left of the diagram are based on Fig 2.

https://doi.org/10.1371/journal.pone.0217824.g003

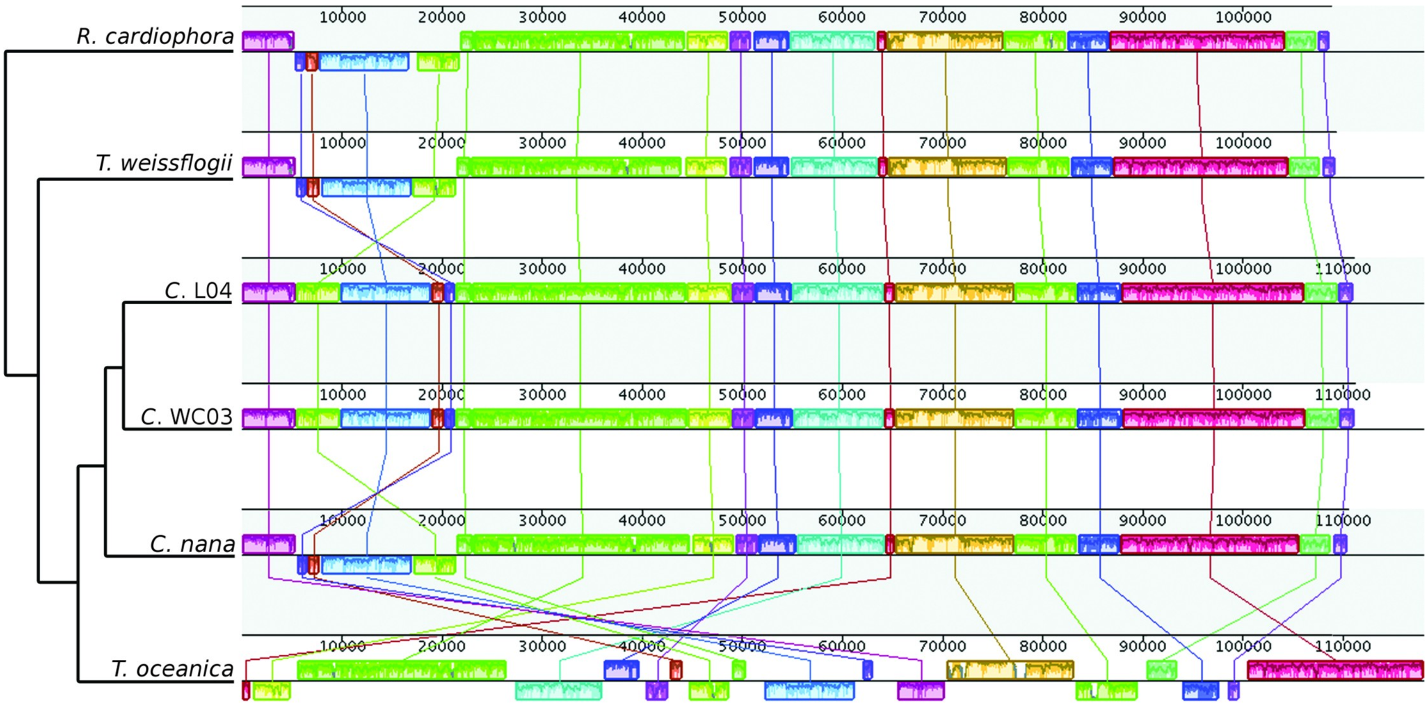

Fig 4. Gene order comparison of plastid genomes of six thalassiosiroid diatoms with one copy of the inverted repeat removed prior to analysis. Alignment and resulting locally collinear blocks (LCBs) were generated using MAUVE. Relationships between taxa displayed as a cladogram to the left of the diagram are based on Fig 2.

https://doi.org/10.1371/journal.pone.0217824.g004

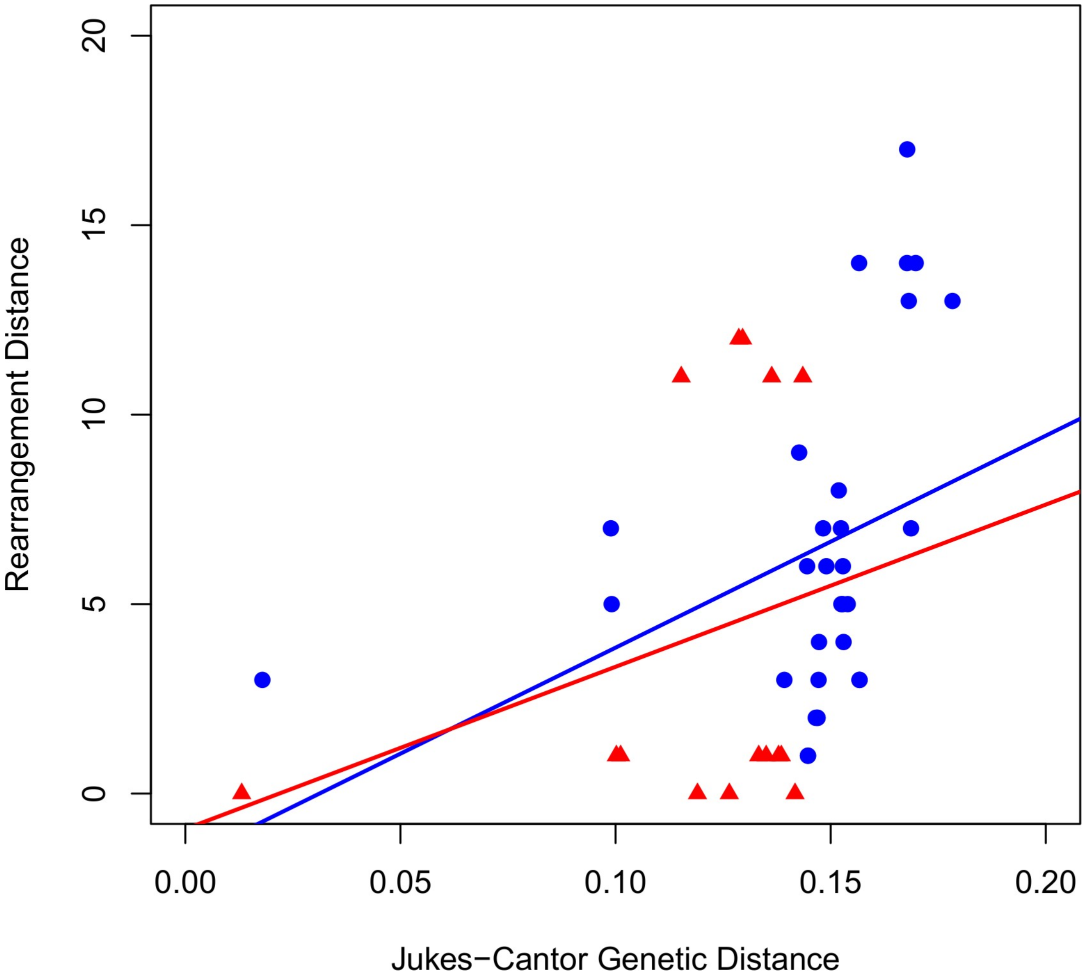

Fig 5. Plot describing the relationship between rearrangement distance and Jukes-Cantor genetic distance. Blue circles represent biraphid diatoms. Red triangles represent thalassiosiroid diatoms. Regression coefficient for the biraphid diatoms is significant ($p = 0.04$, $R^2 = 0.136$), whereas for the thalassiosiroid diatoms it is not ($p = 0.34$, $R^2 = 0.069$).

https://doi.org/10.1371/journal.pone.0217824.g005

rearrangement distance generated in GRIMM for both biraphid and thalassiosiroid diatoms (Fig 5) shows a weak but positive relationship between these two measures of evolution.

The number of inversions and translocations inferred within a genus of diatoms is surprisingly high, particularly among species of *Halamphora*, where congeners may differ by up to 7 detectable rearrangements (S3 Table). This represents a striking contrast to chloroplast genomes in most land plants, where gene order is typically highly conserved within genera, and even across diverse families, orders, and in some cases classes [51]. For example, the monocot *Oryza sativa* differs from the dicot *Arabidopsis thaliana* by only one inversion and a translocation, despite those lineages diverging an estimated 160 million years ago [52]. The level of sequence divergence (12%), differences in gene content (2 genes differ in presence/absence), and degree of genome feature rearrangement (~7 rearrangements) found in *Halamphora* is high, and in order to recapitulate this level of divergence in land plants, one must compare as disparate groups as angiosperms (i.e., *A. thaliana*: accession NC_000932) and Equisetales (i.e., *Equisetum arvense*: accession NC_014699) [52]. Based on estimates in a recent publication [45], *Halamphora* evolved ~75–100 MYA, compared to over 400 MYA between angiosperms and Equisetales [53]. The first occurrence of amphoroid diatoms in the fossil record comes from the Oamaru Formation [54], estimated to be of Eocene age (33–55 MYA). These data suggest that chloroplast genomes of this genus are evolving approximately four (based on the estimates of the first occurrence of the group) to seven (based on occurrence in the fossil record) times faster than those of land plants in multiple ways, including sequence divergence, gene content, and gene order.

Conclusions

Numerous studies have compared gene content and genome rearrangement among diatom chloroplast genomes (e.g., [15,16,17,32,55,56,57,58,59,60,61,62]), but this is the first to compare multiple genomes within a genus of biraphid pennate diatoms and several surprising patterns were identified. In regard to the phylogenetic position of *Halamphora* and the relationship between *Halamphora* species, our 20-gene phylogeny supported similar relationships to those revealed in other studies [13,14]. As with other diatoms, these chloroplast genomes are evolving relatively rapidly at the sequence level (12% divergence across the genus *Halamphora*). *Halamphora* plastid genomes also showed variation in gene content, with species incorporating two genes. Even more striking variation was observed in gene order, with multiple inversions, translocations, and inversion/translocation combinations found within this genus. *Cyclotella*, a thalassiosiroid genus, showed more conservation in gene order, with only one inversion. Overall, biraphid pennate diatoms appear to display more variation in gene order than the thalassiosiroid diatoms and significantly more variation than typical land plant chloroplasts (notable exceptions include the Geraniaceae family [63] and *Amborella* [64]). Although this pattern is conspicuous, only 0.04% of the estimated 100,000 diatom species' chloroplast genomes have been examined and therefore, additional data and comparisons are necessary before generalizations should be made regarding overall diatom chloroplast genome evolution. Although these data are preliminary (comprising only a fraction of diatom diversity), they point to a comparable degree of variation within this one genus of diatoms to the divergence among distant divisions of vascular plants. This diversification within *Halmaphora* is accompanied by a substantially higher (4–7×) rate of evolution. These remarkable intrageneric and inter-kingdom comparisons require additional data to verify the results. However, if these data are supported by additional studies, they open the door to many questions about the rate and modes of molecular evolution of the chloroplast genome in this remarkable clade.

Supporting information

S1 Table. Diatom chloroplast genomes used as reference for annotating the *Halamphora* genomes.
(DOCX)

S2 Table. Gene content comparison of three *Halamphora* spp. genomes (bold) with other published diatom plastid genomes. X, present; P, pseudogenized copy of gene is present; -, absent.
(XLSX)

S3 Table. Distances between gene orders of LCBs (generated in MAUVE) of biraphid pennate diatoms (taxa from this study are in bold) calculated using GRIMM. Smaller values indicate more similar gene order.
(DOCX)

S4 Table. Distances between gene orders of LCBs (generated in MAUVE) of thalassiosiroid diatoms calculated using GRIMM. Smaller values indicate more similar gene order.
(DOCX)

S1 Fig. Plastid genome map of *Halamphora calidilacuna*. Genes on the outside are transcribed clockwise and those on the inside counterclockwise. The inner ring displays GC content in grey.
(TIF)

S2 Fig. Plastid genome map of *Halamphora americana*. Genes on the outside are transcribed clockwise and those on the inside counterclockwise. The inner ring displays GC content in grey.
(TIF)

S3 Fig. Plastid genome map of *Halamphora coffeaeformis*. Genes on the outside are transcribed clockwise and those on the inside counterclockwise. The inner ring displays GC content in grey.
(TIF)

S1 File. Fasta alignment. The fasta alignment of twenty molecular markers (*psaA*, *psbC*, *petD*, *petG*, *atpA*, *atpG*, *rbcL*, *rbcS*, *rpoA*, *rpoB*, *rps14*, *rpl33*, *rnl*, *rns*, *ycf89*, *sufB*, *sufC*, *dnaK*, *dnaB*, *clpC*) from 26 species used to generate the phylogeny presented in this study.
(FASTA)

Author Contributions

Conceptualization: Sarah E. Hamsher, Kyle G. Keepers, Cloe S. Pogoda, Joshua G. Stepanek, Nolan C. Kane, J. Patrick Kociolek.

Data curation: Sarah E. Hamsher, Kyle G. Keepers, Cloe S. Pogoda, Joshua G. Stepanek.

Formal analysis: Sarah E. Hamsher, Kyle G. Keepers, Cloe S. Pogoda, Joshua G. Stepanek.

Funding acquisition: Nolan C. Kane, J. Patrick Kociolek.

Investigation: Sarah E. Hamsher, Kyle G. Keepers, Cloe S. Pogoda, Joshua G. Stepanek.

Methodology: Sarah E. Hamsher, Kyle G. Keepers, Cloe S. Pogoda, Joshua G. Stepanek.

Project administration: Sarah E. Hamsher, Kyle G. Keepers, Cloe S. Pogoda.

Resources: Nolan C. Kane, J. Patrick Kociolek.

Software: Kyle G. Keepers, Cloe S. Pogoda.

Supervision: Nolan C. Kane, J. Patrick Kociolek.

Validation: Sarah E. Hamsher, Kyle G. Keepers, Cloe S. Pogoda, Joshua G. Stepanek, Nolan C. Kane, J. Patrick Kociolek.

Visualization: Sarah E. Hamsher, Kyle G. Keepers, Cloe S. Pogoda, Joshua G. Stepanek.

Writing – original draft: Sarah E. Hamsher, Kyle G. Keepers, Cloe S. Pogoda, Joshua G. Stepanek.

Writing – review & editing: Sarah E. Hamsher, Kyle G. Keepers, Cloe S. Pogoda, Joshua G. Stepanek, Nolan C. Kane, J. Patrick Kociolek.

References

1. Nelson DM, Tréguer P, Brzezinski MA, Leynaert A, Quéguiner B. Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global Biogeochem Cy. 1995; 9: 359–372.

2. Struyf E, Smis A, Van Damme S, Meire P, Conley DJ. The global biogeochemical silicon cycle. Silicon. 2009; 1: 207–213. https://doi.org/10.1007/s12633-010-9035-x

3. Fabris M, Matthijs M, Rombauts S, Vyverman W, Goossens A, Baart GJE. The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff glycolytic pathway. Plant J. 2012; 70: 1004–1014. https://doi.org/10.1111/j.1365-313X.2012.04941.x PMID: 22332784

4. Mann DG. The species concept in diatoms. Phycologia. 1999; 38: 437–495. https://doi.org/10.2216/i0031-8884-38-6-437.1

5. Mann DG, Vanormelingen P. An inordinate fondness? The number, distributions, and origins of diatom species. J Eukaryot Microbiol. 2013; 60: 414–420. https://doi.org/10.1111/jeu.12047 PMID: 23710621

6. Seckbach J, Kociolek P, editors. The Diatom World. Springer Netherlands; 2011. https://doi.org/10.1007/978-94-007-1327-7

7. Levkov Z. *Amphora sensu lato*. In: Lange-Bertalot H, editor. Diatoms of Europe: diatoms of the European inland waters and comparable habitats. Ruggell: ARG Gantner Verlag KG; 2009. p. 5–916.

8. Stepanek JG, Kociolek JP. *Amphora* and *Halamphora* from coastal waters and inland waters of the United States and Japan. Bibl Diatomologica. 2018; 66: 1–260.

9. Cavalcante KP, Tremarin PI, Ludwig TAV. New records of amphoroid diatoms (Bacillariophyceae) from Cachoeira River, Northeast Brazil. Braz J Biol. 2014; 74: 257–263. https://doi.org/10.1590/1519-6984.24512 PMID: 25055112

10. Van de Vijver B, Kopalova K, Zidarova R, Levkov Z. Revision of the genus *Halamphora* (Bacillariophyta) in the Antarctic region. Plant Ecol Evol. 2014; 147: 347–391. https://doi.org/10.5091/plecevo.2014.979

11. Sheehan J, Dunahay T, Benemann J, Roessler P. A look back at the US Department of Energy's aquatic species program: biodiesel from algae. National Renewable Energy Laboratory, Golden, CO, 1998;TP-580-24190: 328 p.

12. Stepanek JG, Fields FJ, Kociolek JP. A comparison of lipid content metrics using six species from the genus *Halamphora* (Bacillariophyta). Biofuels. 2016; 7: 521–528. https://doi.org/10.1080/17597269.2016.1163216

13. Stepanek JG, Kociolek JP. Molecular phylogeny of *Amphora sensu lato* (Bacillariophyta): an investigation into monophyly and classification of the amphoroid diatoms. Protist. 2014; 165: 177–195. https://doi.org/10.1016/j.protis.2014.02.002 PMID: 24646793

14. Stepanek JG, Kociolek JP. Molecular phylogeny of the diatom genera *Amphora* and *Halamphora* (Bacillariophyta) with a focus on morphological and ecological evolution. J Phycol. 2019; 55: 442–456. https://doi.org/10.1111/jpy.12836 PMID: 30659609

15. Ruck EC, Nakov T, Jansen RK, Theriot EC, Alverson AJ. Serial gene losses and foreign DNA underlie size and sequence variation in the plastid genomes of diatoms. Genome Biol Evol. 2014; 6: 644–654. https://doi.org/10.1093/gbe/evu039 PMID: 24567305

16. Sabir JSM, Yu M, Ashworth MP, Baeshen NA, Baeshen MN, Bahieldin A, et al. Conserved gene order and expanded inverted repeats characterize plastid genomes of Thalassiosirales. PLOS One. 2014; 9: e107854. https://doi.org/10.1371/journal.pone.0107854 PMID: 25233465

17. Yu M, Ashworth MP, Hajrah NH, Khiyami MA, Sabir MJ, Alhebshi AM, et al. Evolution of the plastid genomes in diatoms. Adv Bot Res. 2018; *in press*. https://doi.org/10.1016/bs.abr.2017.11.009

18. Schuster SC. Next-generation sequencing transforms today's biology. Nature Methods. 2007; 5: 16. https://doi.org/10.1038/nmeth1156 PMID: 18165802

19. Guillard RRL, Lorenzen CJ. Yellow-green algae with Chlorophyllide. J Phycol. 1972; 8: 1014.

20. Pogoda CS, Keepers KG, Hamsher SE, Stepanek JG, Kane NC, Kociolek JP. Comparative analysis of the mitochondrial genomes of six newly sequenced diatoms reveals group II introns in the barcoding region of cox1. Mitochondr DNA Part A. 2019; 30: 43–51. https://doi.org/10.1080/24701394.2018.1450397 PMID: 29527965

21. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

22. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19: 455–477. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

23. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004; 20: 3252–3255. https://doi.org/10.1093/bioinformatics/bth352 PMID: 15180927

24. Lohse M, Drechsel O, Kahlau S, Bock R. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 2013; 41: W575–W581. https://doi.org/10.1093/nar/gkt289 PMID: 23609545

25. Lee J, Yang EC, Graf L, Yang JH, Qiu H, Zelzion U, et al. Analysis of the draft genome of the red seaweed *Gracilariopsis chorda* provides insights into genome size evolution in Rhodophyta. Mol Biol Evol. 2018; 35: 1869–1886. https://doi.org/10.1093/molbev/msy081 PMID: 29688518

26. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30: 772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

27. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

28. Silvestro D, Michalak I. raxmlGUI: a graphical front-end for RAxML. Org Div Evol. 2012; 12: 335–337. https://doi.org/10.1007/s13127-011-0056-0

29. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analysis in R. Bioinformatics. 2019; 35: 526–528. https://doi.org/10.1093/bioinformatics/bty633 PMID: 30016406

30. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004; 14: 13941403. https://doi.org/10.1101/gr.2289704 PMID: 15231754

31. Tesler G. GRIMM: genome rearrangements web server. Bioinformatics. 2002; 18: 492–493. https://doi.org/10.1093/bioinformatics/18.3.492 PMID: 11934753

32. Imanian B, Pombert J-F, Keeling PJ. The complete plastid genomes of the two 'dinotoms' *Durinskia baltica* and *Kryptoperidinium foliaceum*. PLOS One. 2010; 5: e10711. https://doi.org/10.1371/journal.pone.0010711 PMID: 20502706

33. Muñoz-Gómez SA, Mejía-Franco FG, Durnin K, Colp M, Grisdale CJ, Archibald JM, et al. The new red algal sybphylum Proteorhodophytina comprises the largest and most divergent plastid genomes known. Curr Biol. 2017; 27: 1677–1684. https://doi.org/10.1016/j.cub.2017.04.054 PMID: 28528908

34. Reyes-Prieto A, Russell S, Figueroa-Martinez F, Jackson C. Comparative plastid genomics of Glaucophytes. Adv Bot Res. 2018; 85: 95–127. https://doi.org/10.1016/bs.abr.2017.11.012

35. Letsch MR, Lewis LA. Chloroplast gene arrangement variation within a closely related group of green algae (Trebouxiophyceae, Chlorophyta). Mol Phylo Evol. 2012; 64: 524–532. https://doi.org/10.1016/j.ympev.2012.05.027 PMID: 22659018

36. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 1986; 5: 2043–2049. PMID: 16453699

37. Hildebrand M, Corey DK, Ludwig JR, Kukel A, Feng T-Y, Volcani BE. Plasmids in diatom species. J Bacteriol. 1991; 173: 5924–5927. 0021-9193/91/185924-04$02.00/0 https://doi.org/10.1128/jb.173.18.5924-5927.1991 PMID: 1885558

38. Hildebrand M, Hasegawa P, Ord RW, Thorpe VS, Glass CA, Volcani BE. Nucleotide sequence of diatom plasmids: identification of open reading frames with similarity to site-specific recombinases. Plant Mol Biol. 1992; 19: 759–770. PMID: 1322740

39. Turmel M, Otis C, Lemieux C. Mitochondrion-to-chloroplast DNA transfers and intragenomic proliferation of chloroplast group II introns in *Gloeotilopsis* Green Algae (Ulotrichales, Ulvophyceae). Genome Biol Evol. 2017; 8: 2789–2805. https://doi.org/10.1093/gbe/evw190 PMID: 27503298

40. Hamsher SE, Evans KM, Mann DG, Poulíčková A, Saunders GW. Barcoding diatoms: exploring alternatives to COI-5P. Protist. 2011; 162: 405–422. https://doi.org/10.1016/j.protis.2010.09.005 PMID: 21239228

41. Huvet M, Stumpf MPH. Overlapping genes: a window on gene evolvability. BMC Genomics. 2014; 15: 721. https://doi.org/10.1186/1471-2164-15-721

42. Adachi Y, Kuroda H, Yukawa Y, Sugiura M. Translation of partially overlapping *psbD-psbC* mRNAs in chloroplasts: the role of 5´-processing and translational coupling. Nucleic Acids Res. 2012; 40: 3152–3158. https://doi.org/10.1093/nar/gkr1185 PMID: 22156163

43. Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. J Virology. 2009; 82: 10719–10736. https://doi.org/10.1128/JVI.00595-09 PMID: 19640978

44. Rastogi A, Maheswari U, Dorrell RG, Rocha Jimenez Vieira F, Maumus F, Kustka A, et al. Integrative analysis of large scale transcriptome data draws a comprehensive landscape of *Phaeodactylum tricornutum* genome evolutionary origin of diatoms. Sci Rep. 2018; 8: 4834. https://doi.org/10.1038/s41598-018-23106-x PMID: 29556065

45. Nakov T, Beaulieu JM, Alverson AJ. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). New Phytol. 2018; 219: 462–473. https://doi.org/10.1111/nph.15137 PMID: 29624698

46. Parks MB, Nakov T, Ruck EC, Wickett NJ, Alverson AJ. Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). Am J Bot. 2018; 105: 330–347. https://doi.org/10.1002/ajb2.1056 PMID: 29665021

47. Ashworth MP, Nakov T, Theriot EC. Revisiting Ross and Sims (1971): toward a molecular phylogeny of the Biddulphiaceae and Eupodiscaceae (Bacillariophyceae). J Phycol. 2013; 49: 1207–1222. https://doi.org/10.1111/jpy.12131 PMID: 27007638

48. Gargas CB, Theriot EC, Ashworth MP, Johansen JR. Phylogenetic analysis reveals that the 'Radial Centric' diatom *Orthoseira* Thwaites (Orthoseiraceae, Bacillariophyta) is a member of a 'Multipolar' diatom lineage. Protist 2018; 169: 803–825. https://doi.org/10.1016/j.protis.2018.08.005 PMID: 30448592

49. Ruck EC, Theriot EC. Origin and evolution of the canal raphe system in diatoms. Protist. 2011; 162: 723–737. https://doi.org/10.1016/j.protis.2011.02.003 PMID: 21440497

50. Zgrundo A, Lemke P, Pniewski F, Cox EJ, Latala A. Morphological and molecular phylogenetic studies on *Fistulifera saprophila*. Diatom Res. 2013; 28: 431–443. https://doi.org/10.1080/0269249X.2013.833136

51. Jansen R. K., & Ruhlman T. A. (2012). Plastid genomes of seed plants. In *Genomics of chloroplasts and mitochondria* (pp. 103–126). Springer, Dordrecht

52. Xu J-H, Liu Q, Hu W, Wang T, Xue Q, Messing J. Dynamics of chloroplast genomes in green plants. Genomics. 2015; 106: 221–231. https://doi.org/10.1016/j.ygeno.2015.07.004 PMID: 26206079

53. Soltis PS, Soltis DE, Savolainen V, Crane PR, Barraclough TG. Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. P Natl Acad Sci USA. 2002; 99: 4430–4435.

54. Schrader H-J. Die Pennaten Diatomeen aus dem Obereozän von Oamaru, Neuseeland. Beihefte zur Nova Hedwigia. 1969; 28: 1–124.

55. Kowallik KV, Stoebe B, Schaffran I, Kroth-Pancic P, Freier U. The chloroplast genome of a chlorophyll *a*+*c*-containing alga, *Odontella sinensis*. Plant Mol Biol Rep. 1995; 13: 336–342.

56. Oudot-Le Secq MP, Grimwood J, Shapiro H, Armbrust EV, Bowler C, Green BR. Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage. Mol Genet Genomics. 2007; 277: 427–439. https://doi.org/10.1007/s00438-006-0199-4 PMID: 17252281

57. Lommer M, Roy A-S, Schilhabel M, Schreiber S, Rosenstiel P, LaRoche J. 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. 2010; 11: 718. https://doi.org/10.1186/1471-2164-11-718 PMID: 21171997

58. Ravin NV, Galachyants YP, Mardanov AV, Beletsky AV, Petrova DP, Sherbakova TA, et al. Complete sequence of the mitochondrial genome of a diatom alga *Synedra acus* and comparative analysis of diatom mitochondrial genomes. Curr Genet. 2010; 56: 215–223. https://doi.org/10.1007/s00294-010-0293-3 PMID: 20309551

59. Tanaka T, Fukuda Y, Yoshino T, Maeda Y, Muto M, Matsumoto M, et al. High-throughput pyrosequencing of the chloroplast genome of a highly neutral-lipid-producing marine pennate diatom, *Fistulifera* sp. strain JPCC DA0580. Photosynth Res. 2011; 109: 223–229. https://doi.org/10.1007/s11120-011-9622-8 PMID: 21290260

60. Brembu T, Winge P, Tooming-Klunderud A, Nederbragt AJ, Jakobsen KS, Bones AM. The chloroplast genome of the diatom *Seminavis robusta*: new features introduced through multiple mechanisms of horizontal gene transfer. Mar Genomics. 2014; 16: 17–27. https://doi.org/10.1016/j.margen.2013.12.002 PMID: 24365712

61. Galachyants YP, Zakharova YR, Petrova DP, Morozov AA, Sidorov IA, Marchenkov AM, et al. Sequencing of the complete genome of an araphid pennate diatom *Synedra acus* subsp. *radians* from Lake Baikal. Dokl Biochem Biophys. 2015; 461: 84–88. https://doi.org/10.1134/S1607672915020064 PMID: 25937221

62. Crowell RM, Nienow JA, Cahoon AB. The complete chloroplast and mitochondrial genomes of the diatom *Nitzschia palea* (Bacillariophyceae) demonstrate high sequence similarity to the endosymbiont organelles of the dinotom *Durinskia baltica*. J Phycol. 2018; 55: 352–364. https://doi.org/10.1111/jpy.12824 PMID: 30536677

63. Guisinger MM, Kuehl JV, Boore JL, Jansen, RK. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol. 2010; 28: 583–600. https://doi.org/10.1093/molbev/msq229 PMID: 20805190

64. Goremykin VV, Hirsch-Ernst KI, Wölfl S, Hellwig FH. (2003). Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol Biol Evol. 2003; 20: 1499–1505. https://doi.org/10.1093/molbev/msg159 PMID: 12832641