Nucleotide substitution rates of diatom plastid encoded protein genes are correlated with genome architecture

Mengjie Yu[1&], Tracey A. Ruhlman[1,&], Nahid H. Hajrah[2], Mohammad A. Khiyami[3], Mumdooh J. Sabir[4], Mohammed H. Alblowi[5], Alawiah M. Alhebshi[2], Abdulrahman L. Al-Malki[2], Jamal S. M. Sabir[2], Edward C. Theriot[1], and Robert K. Jansen[1,2*]

[1]*Department of Integrative Biology, University of Texas at Austin, Austin TX 78712, USA*

[2]*Center of Excellence for Bionanoscience Research, King Abdulaziz University (KAU), Jeddah 21589, Saudi Arabia*

[3]*King Abdulaziz City for Science and Technology, Riyadh, 11442, Saudi Arabia*

[4]*Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia*

[&]Co-first authors

[5]*Laboratory Departments, Ministry of Environment, Agriculture and Water, Riyadh, Saudi Arabia*

[*]corresponding author, E-mail: jansen@austin.utexas.edu, Tel.: 512-471-8827.

E-mail addresses: annaymj_2010@utexas.edu (M. Yu), truhlman@austin.utexas.edu (T.A. Ruhlman), nhajrah260@gmail.com (N.H. Hajrah), mkhiyami@kacst.edu.sa (M.A. Khiyami), msabir599@hotmail.com (M.J. Sabir), blewevet@gmail.com (M.H. Alblowi), aalhebshi@kau.edu.sa (A.M. Alhebshi), alalmalki@kau.edu.sa (A.L. Al-Malki), jsabir2622@gmail.com (J.S.M. Sabir), etheriot@austin.utexas.edu (E.C. Theriot), jansen@austin.utexas.edu (R.K. Jansen).

Abstract

Diatoms are the largest group of heterokont algae with more than 100,000 species. They are photosynthetic, unicellular eukaryotes that contribute ~ 45% of global primary production and inhabit marine, aquatic and terrestrial ecosystems. Despite their ubiquity and environmental significance very few diatom plastid genomes (plastomes) have been sequenced and studied. This study explored the pattern of diatom plastid nucleotide substitution rates across the entire suite of plastome protein-coding genes for 40 taxa representing the major clades. Substitution rate acceleration was lineage specific with the highest rates in the araphid 2 taxon *Astrosyne radiata* and radial 2 taxon *Proboscia* sp. Rate heterogeneity was also evident in different functional classes of genes. Similar to land plants, proteins genes involved in photosynthetic metabolism have substantially lower rates than those involved in transcription and translation. Significant positive correlations were identified between rates and measures of genomic rearrangement, but not plastome size. This work advances the current understanding of diatom plastomes and provides a foundation for future studies of their evolution.

Keywords: Bacillariophyta; plastid genome; evolutionary rates; genomic rearrangements; plastome size

Introduction

Diatoms are photosynthetic, unicellular eukaryotes of the heterokont algal lineage. As such their plastids were derived from a secondary endosymbiotic event, in which a nonphotosynthetic eukaryote phagocytized a red alga about 250 million years ago [1]. Diatoms have since colonized freshwater, marine and terrestrial habitats contributing ~ 45% of global primary production [2–4] and as much as 20% of global carbon fixation via photosynthesis [5,6].

Despite their ubiquity and the environmental significance of diatom photosynthesis very few diatom plastid genomes (plastomes) have been sequenced and studied. Unlike plant plastomes, with more than 2900 species represented in the public databases (NCBI accessed February 4, 2019), just 40 diatom taxa have been sequenced thus far [7]. Plastome sequences have provided insight into relationships within the monophyletic diatom lineage [7–11], supported the common ancestry of diatoms and rhodophytes [12] and been used to explore variation in plastome structure and gene content across orders, genera and species [7,9,10,13–16].

Within the diatom cytoplasm are numerous or singular plastids of variable shape [17,18]. Of four diatom species examined, each plastid contained a single nucleoid [19] comprising copies of the plastome monomer or unit-genome, RNA and proteins [20]. Diatom plastid genes are densely arrayed on both strands of the unit-genome, representing one full complement of the gene space and intergenic regions, which can be repeated many times such that the plastome of an individual may comprise many copies of the unit-genome. Although this unit is often diagrammed as a circular molecule, the plastome more likely contains a collection of circular, linear and linear-branched molecules that each comprises two to many copies of the monomer [21]. All diatom plastomes sequenced to date include a large inverted repeat (IR) separated by

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

large and small single copy regions (LSC and SSC, respectively). Apart from the common quadripartite structure, diatom plastomes exhibit a range of gene order arrangements, and gains and losses of genes and introns [7,9,10]. Gene order changes can arise through gene duplication, often via expansion of the IR, but also via inversions and insertions and deletions (indels) in both IR and SC regions.

Although phylogenetic reconstructions have relied on diatom plastome coding sequences, evolutionary rate analyses are lacking. Synonymous and nonsynonymous nucleotide substitution rates (*dS* and *dN*, respectively) can vary widely within and between taxa. Comparison of nucleotide substitution rates across functional groups of genes and between lineages provides insight into the evolution of plastomes [22]. Genes encoding subunits that are integral to photosynthesis, such as cytochrome b_6f complex (PET) and photosystems I and II (PSA and PSB) have lower rates of nucleotide substitution than other functional groups in angiosperms and conifers [23–26]. Accelerated substitution rates have been detected in ribosomal protein (RPL and RPS) genes and RNA polymerase (RPO) genes [24, 26–30].

In addition to rate differences between gene functional groups, rate variation relative to genomic features such as rearrangements in gene order can also provide insight into the forces shaping plastome evolution. Previous studies have identified a significant positive correlation between rates of nucleotide substitution and gene order changes in angiosperm plastid genomes [30,31], bacterial genomes [32] and arthropod mitochondrial genomes [33,34]. Disruption of DNA repair, recombination and replication (DNA-RRR) systems has been suggested to cause highly elevated nucleotide substitution rates and genome rearrangements [24,35]. A recent study revealed a significant correlation between *dN* of nuclear encoded DNA-RRR genes that are targeted to plastomes and measures of plastome complexity in one angiosperm family [36].

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Previous studies also showed a negative correlation between genome size and substitution rates in plastomes [26,37].

Diatoms are the most species-rich constituent of oceanic phytoplankton [38] and reflect a fundamentally different evolutionary path than land plants or the green and red algae, offering an ideal opportunity to examine the patterns of plastome nucleotide substitution rates in a secondary endosymbiotic lineage. Previous work in diatoms comparing four plastid and six nuclear genes has shown that *dS* and *dN* were lower in plastid genes than in nuclear genes, and *dS* was negatively correlated to the degree of codon usage bias in plastid genes [39]. However, no diatom study to date has investigated nucleotide substitution rates of all shared plastome protein-coding genes. The present study explored the pattern of diatom plastid nucleotide substitution rates across the entire suite of plastome genes. Correlation between plastome substitution rates and genome features, such as plastome size, number of indels and genome rearrangement were examined. This work advances the current understanding of diatom plastomes and provides a foundation for future studies of their evolution.

Materials and methods

Gene sequence alignment and phylogenetic analysis

Plastid protein-coding genes were extracted from all available complete diatom plastomes (40 taxa) together with the outgroup species *Triparma laevis* (Bolidophyceae) (S1 Table). The 103 shared plastid gene sequences were aligned with Multiple Alignment using Fast Fourier Transform [MAFFT, 40]. Protein-coding genes were partitioned by functional category following Yu et al. [7]. A maximum likelihood tree was constructed with RAxML7.2.9 [41], with the substitution model GTR+G and -f a option. One thousand bootstrap replicates were

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

performed to assess strength of support for clades. 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 [42]. Gapped regions were excluded with the parameter cleandata = 1 to avoid spurious rate inference. Pairwise rates were calculated relative to the outgroup species *Triparma laevis* and estimated with the parameter runmode = -2. All shared plastome genes (103) were concatenated for nucleotide substitution rate estimation and separate estimations were calculated on individual genes or concatenated sequences of genes in different functional groups as listed in S2 Table.

Plastome features for correlation analyses

The number of indels for the concatenated 103 protein-coding genes was calculated using a custom Python script. *Triparma laevis* (Bolidophyceae) was used as a reference. Indels within aligned protein-coding regions were summed using a custom Python script resulting in a single value for each taxon; only intact genes were included (in-frame indels). Whole genome alignment among the 40 diatom species was performed using the ProgressiveMauve algorithm in Mauve v2.3.1 [43]. The same IR copy (IRb) was removed from all plastomes. The locally collinear blocks (LCBs) identified by Mauve were numbered with positive or negative sign based on strand orientation to estimate genome rearrangement distance (S3 Table). Inversion (IV) distances were estimated using GRIMM (S4 Table) [44]. The feature 'plastome size' excludes one copy of the IR for each taxon.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Correlation between substitution rates and genome characteristics

Pairwise *dN* and *dS* values were calculated for the 103 shared genes from each taxon relative to the outgroup *Triparma laevis*. Correlation of *dN* and *dS* with plastome size and indel number for each plastome was tested. Phylogenetic Generalized Least Squares was performed using the 'ape' and 'nlme' packages in R. The ML constraint tree was utilized with outgroup taxa pruned. The correlation between *dN* and *dS* with IV distance was tested using the Pearson test. The resulting p-values were Bonferroni corrected using the built-in p.adjust function to account for the effect of multi-hypothesis testing.

Results

Phylogenetic relationships and branch lengths

Phylogenetic analysis of 40 diatoms (S1 Table) for the concatenated 103 gene data set (S2 Table) recovered a fully resolved ML tree with strong bootstrap (> 95) support for all but three clades (Figs 1, S1). 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) was paraphyletic and contained in three clades (polar 1, 2 and 3). Araphid 1 was sister to araphid 2 plus the raphid group. Within araphid 2, *Astrosyne radiata* was recovered on an extremely long branch (Fig 1). Raphid pennate diatoms were recovered as a monophyletic group sister to a clade of araphid pennate diatoms (araphid 2). The maximum likelihood tree (Fig. S1) was then used as a constraint tree for estimating substitution rates.

The *dN* and *dS* trees had very similar patterns with regard to branch length variation. The most accelerated lineage was branch 63 leading to *Astrosyne radiata* (Fig 1). Branch 4, leading to *Proboscia sp.*, also showed acceleration in both *dN* and *dS*.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Rate variation in functional groups of genes

Gene sequences in each functional class (S2 Table) were concatenated to estimate *dN* and *dS*. The pattern of *dN* and *dS* was similar within functional classes (Figs 2, S2-S3). RNA polymerase genes (RPO) had the highest median values of *dN* and *dS* among the major gene categories. Large and small subunit ribosomal protein genes (RPL and RPS, respectively) also had high median values of *dN* and *dS*. Median values of both *dN* and *dS* were lower in genes integral to photosynthesis (i.e. PSA, PSB, PET and ATP genes) than the other groups. Among all classes and individual genes (S2 Table) *dnaB*, encoding the replicative DNA helicase, had the highest *dN* and *dS* (Figs S2-S3). Among all gene classes, branch 63, leading to *A. radiata* showed the greatest acceleration in both *dN* and *dS* (Fig 1).

Correlation of substitution rates and plastome characteristics

Significant correlation was observed between overall *dN* and *dS* and the number of indels ($p < 0.05$; Fig 3). No significant correlation was found between the substitution rate and plastome size (Fig S4). However, *Astrosyne radiata*, which has a relatively small plastome among diatoms, had the highest overall *dN* and *dS* (Fig 1, S5 Table).

Correlation of pairwise substitution rates and inversion distance (S6 Table) was tested in the 40 diatom plastomes. Significant correlation ($p<0.05$) was found between *dN* and inversion distance in 25 of 40 pairwise comparisons (Fig 4; S6 Table). Among the 40 plastomes, *dS* was significantly correlated with inversion distance in 18 pairwise comparisons. The polar 1 group had the largest proportion of significant correlations between substitution rates and inversion distance, with seven of nine sampled taxa significantly correlated in both *dN* and *dS*. *Astrosyne radiata*, which produced the longest branch in the diatom phylogeny (Figs 1, S1), also showed

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

significant correlation of *dN* (p=2.41e-06) and *dS* (p=2.23e-03) with inversion distance (Fig 4; S6 Table).

Discussion

In this study, 103 plastid genes were examined across 40 diatom species, most of which were recently published by our group [7,9,10]. The ribosomal subunit and RNA polymerase genes showed accelerated nucleotide substitution rates compared to photosystem genes. Positive correlations were uncovered between substitution rates and number of indels and inversion distance, proxies of genome rearrangement. Using all shared plastome protein-coding sequences from taxa in an understudied yet important group may help illuminate patterns and forces shaping molecular evolution in diatom plastomes.

Lineage specific nucleotide substitution rates

Lineage specific substitution rates were reported in previous studies utilizing plastome sequences. Several studies have demonstrated lineage specific acceleration of substitution rates in angiosperms [24–26,45–48] and gymnosperms [49,50]. The plastid encoded *tufA* gene, encoding the transcript elongation factor Tu, was found to be evolving at a fast pace in the charophyte algal class *Coleochaetophyceae*, compared with sister clades [51].

Here, substantially higher substitution rates were detected in *Astrosyne radiata* and *Proboscia* sp. relative to all other included diatoms (Fig 1). Extensive gene loss was also found in both *Astrosyne* and *Proboscia* [7]. Additional taxon sampling of relatives of both of these genera may help better elucidate the changes in the araphid 2 and radial 2 clades.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Variable substitution rates in gene functional classes

Gene essentiality is a widely studied factor in substitution rate variation, with the idea that essential genes are subject to stronger selective constraint than non-essential genes [52–54]. Several studies utilizing nuclear sequences have demonstrated that rates of nucleotide substitution are associated with expression levels where highly or more widely expressed genes evolve at a slower rate in plants [55–57] and animals [58–60] supporting the notion that these genes may evolve under greater selective constraints than genes with more limited expression profiles.

Substitution rates in plastome sequences demonstrated significant differences in both *dS* and, more drastically, *dN* between housekeeping and photosynthetic genes among 283 angiosperms [61]. Significant differences in *dN,* but not *dS,* were noted among sets of genes involved in the photosynthesis apparatus, photosynthetic metabolism, gene expression and a group of miscellaneous genes in *Phalaenopsis* [23]. An increase in *dN* was noted for RPS and RPL genes relative to photosynthetic genes in Poaceae and four Saxifragales species while *dS* values were similar across gene types [25,28]. Plastome nucleotide substitution rates in two *Silene* species, *S. conica* and *S. noctiflora*, showed little elevation across photosynthetic genes but marked acceleration in genes involved in translation relative to more conserved plastomes in the genus. Nonsynonymous rate accelerations in RPO, RPL and RPS genes were disproportionately large [27]. Although substitution rates in Geraniaceae are higher overall than in other angiosperms, RPL, RPS and RPO genes were the most significantly different in *dN* relative to genes involved in photosynthesis, which exhibited similar but weaker patterns [24]. These findings concur with those reported previously in maize and rice, where acceleration in *dN* was the predominant phenomenon varying RPO, RPS/L and ATP sequences [62].

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Previous studies examining substitution rate heterogeneity among different functional classes of genes have not been reported for heterokont algae. Like land plants, diatom plastid genes mainly fall into two general classes, those encoding proteins involved in photosynthetic metabolism (PSA, PSB, PET, ATP) and those with roles in transcription and translation (RPS, RPL, RPO). The finding that genes involved in photosynthesis had relatively lower overall substitution rates than genes in transcription-translation apparatus confirms that rate heterogeneity by functional class is a shared feature of diatoms and land plant plastomes (Figs 2, S2-S3).

Correlation between substitution rates and plastome characteristics

Diatom plastomes are gene dense, with very little space dedicated to non-coding sequences and most are devoid of large repeat sequences [63]. However, on average diatom plastomes are sized similarly to those of seed plants as they encode more genes [7,9,10,64]. Previous studies in diatoms showed that variation in the unit-genome size is mainly due to expansion and contraction of the IR, gene loss and the introduction of foreign DNA of unknown origin [7,9,10]. These studies did not consider the relationship of nucleotide substitution rates and expansion or contraction of the plastome monomer. An inverse relationship between rates and genome size was proposed for bacteriophages, *Escherichia coli* and single-celled eukaryotes [65] as well as organelle genomes in animals and plants [66].

Tests of this hypothesis in seed plant plastomes have thus far supported an inverse correlation between substitution rates and plastome size, however whether nonsynonymous or synonymous changes are responsible for this relationship varied depending on the plant group studied and perhaps in part due to differences in analytical approach. A significant negative correlation between plastome size and *dN* was reported among four Saxifragales species [28].

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Although a negative correlation between both *dS* and *dN* and plastome size was noted among legumes**,** only the *dN* correlation was significant [26] while *dS* but not *dN* was significantly correlated to plastome size in cupressophytes [37]. Further examination suggested that accelerated substitution rates together with decreased noncoding content reduced plastome size of cupressophytes [50].

No significant correlation between substitution rates and plastome size was detected among the diatom plastomes examined here (Fig S4). However, *Astrosyne radiata*, the diatom species that exhibited numerous gene losses [7] had the highest *dN* and *dS* and a relatively small plastome (Fig S4). Expanded sampling in the araphid 2 clade, which includes *Astrosyne*, could provide more information on the relationship between accelerated rates of nucleotide substitution and plastome size variation in diatoms.

Study of evolutionary rates in other genetic compartments in *Astrosyne* may be particularly interesting. *Astrosyne* is a highly unusual diatom from a morphological perspective. Although it is placed among araphid pennates, diatoms with an elongate sternum and bilateral symmetry, it has fully reverted to the ancestral radial symmetry (where all structures are rotationally arranged and symmetric about a single point in the center) of diatoms in radial 3, such as *Coscinodiscus* and *Actinocyclus*. This represents the largest degree of incongruence between molecular and morphological data in the diatom tree. With nuclear small subunit ribosomal RNA sequences also showing accelerated evolutionary rates in *Astrosyne* [67], perhaps a systemic change in evolutionary rate has occurred in this lineage across the plastome, and nuclear and perhaps mitochondrial genomes.

Apart from overall size expansion, correlations between both indels and genome rearrangements, i.e. inversions and gene order changes, and nucleotide substitution rates have

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

been demonstrated for seed plants. Branch lengths, representing total substitution rates for 81 plastid genes shared among 61 seed plant taxa, were positively correlated to numbers of gene and intron losses, gene order changes and indels [31]. The angiosperm family Geraniaceae includes some of the most variable plastomes [68] and has been used to explore the relationship between substitution rates and changes in plastome structure and content. Earlier studies suggested a positive correlation between rates and rearrangements in the unit-genome [24] that was later confirmed. Increases in the degree of plastid genome rearrangements, estimated by inversions and IR boundary shifts, were correlated with the acceleration in *dN* but not *dS*. Furthermore, the degree of genome rearrangement was significantly correlated with the number of repeats larger than 60 bp [30]. Positive correlations of *dN* and *dS* with both indels and rearrangements have been reported for legumes [26] and similar observations have been made in unrelated angiosperm lineages [27,69]. Both *dN* and *dS* have been positively correlated to inversions in cupressophytes [50] and in *Cephalotaxus* total substitutions were correlated to both indels and repeats. The strongest correlation was between repeats and substitutions, and repeats were also correlated with indels [70]. In a similar comparison carried out in the monocot family Araceae the strongest correlation was between repeats and indels, followed by total substitutions and indels with repeats and substitutions also showing significant correlation [71].

Significant correlations between branch length and gene order changes were recently reported for several of the taxa in this study, including *Astrosyne radiata* and *Proboscia* sp. [7]. Significant correlations between indels and inversion distance, proxies for genome rearrangement, and both *dN* and *dS*, were observed among the 40 diatom species studied here (Figs 3-4; S6 Table). Repeats were not evaluated for diatoms here, but overall the diatom plastome monomer is depauperate in repeated sequences [63]. An investigation of diatoms in the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

order Thalassiosirales included repeat analyses that identified a single tandem repeat in each of five Thalassiosirales species with lengths ranging from 79 to 90 base pairs (bp; [10]).

Several studies have identified positive correlations between accumulation of small dispersed repeats and the extent of genome rearrangement in different lineages of chlorophyte algae. Repeat units as small as 7 bp and up to 21 bp were identified by Reputer [72] and manually refined to exclude overlapping or embedded repeats [73–75]. Microhomologies, i.e. repeats as little as 5 bp, have been shown to produce plastome rearrangements in plants [76,77]. The most recent investigation of diatom plastome repeats restricted its analysis to motifs ≥ 16 bp [10]. Future analyses that consider SDR < 16 bp may uncover a correlation between repeats and other measures plastome diversity in diatoms.

In agreement with earlier findings both *dN* and *dS* had significant positive correlations with the number of indels in coding regions in diatom plastomes (Fig 3). Previous studies revealed that nucleotide diversity is substantially elevated in regions surrounding sites that have undergone short insertion or deletion mutations [78,79]. Within 50 bp either side of an indel, the mutation rate increased 30-fold in yeasts [78] and 6-fold in humans [80]. The indel-induced mutation hypothesis was proposed stating that presence of an indel induces increased mutation in proximal sequences [78], which was supported by findings from 262 bacterial genomes [79]. An alternative hypothesis posits error-prone DNA polymerase recruitment to restart replication following polymerase stalling at repeated sequences is responsible for nucleotide diversity [81]. In addition, stalled replication forks tend to incur both single and double strand breaks, which also results in error-prone repair [82], increasing the likelihood of nucleotide substitutions. Although the association between repeats and substitution rates was not investigated here, such a

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

relationship could underlie the observed correlation of both indels and inversions and substitutions.

These previous studies employed eukaryotic nuclear genomes [78] and bacterial genomes [79,81]. Plastomes of algae and land plants comprise several to many copies of the unit-genome, arrayed as linear, branched molecules and tightly associated in nucleoids [83]. In addition to the IR of each monomer, all of the sequence represented in a single unit, genic and intergenic, are repeated and available as substrate for recombination-dependent replication (RDR), an error-prone pathway to restart stalled replication forks [21,82,84]. Just as RDR leads to the documented inversion of the single copy regions depending on which copies, IRb or IRa of different monomer copies, participate in the reaction, inversions may arise in any region of the plastome, facilitated by large or even very small repeats. Although several studies cited herein have shown an association between the number of repeats in a plastome monomer and monomer rearrangement in plant plastomes, the possibility remains that recombination events may be occurring both between and within copies; a possibility that is difficult, if not impossible to resolve using short read sequencing technologies or shotgun cloning approaches [84]. The relationship between repeats, with respect to the entirety of the plastome present within a single plastid or plastid nucleoid, and changes in both unit-genome structure and substitution rates remains unresolved for diatoms. Future studies that employ single molecule, long read sequencing technology may be useful to study alternative arrangements in diatom plastome structure [84].

This investigation found that correlation of nucleotide substitution rates with plastome rearrangements and the number of indels in protein-coding genes is a phenomenon that is observed in diatoms and seed plants alike. Addressing this observation further to explore causal

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

relationships between rates and plastome features will require expanded plastome sampling within and between diatom lineages. Nuclear transcriptome sequences have been used to examine coevolution of nuclear and plastome genes in plants [46,48,85,86]. Investigations that include nuclear and plastome data would further our understanding of evolution in diatom plastomes.

Supporting information

S1 Table. **Taxa included in diatom plastome analyses and NCBI accession numbers.**

(PDF)

S2 Table. Plastid genes and functional groups included in rates analyses.

(PDF)

S3 Table. **Local collinear blocks (LCBs) for each of the 40 diatom plastomes identified by Mauve.**

(PDF)

S4 Table. **Pairwise inversion distance inferred by GRIMM of 40 diatom plastomes.**

(PDF)

S5 Table. **Nucleotide substitution rates and plastome features included in correlation analyses**.

(PDF)

S6 Table. Correlation coefficient and adjusted P-values for correlation between substitution rates and plastome rearrangement measured by inversion distance.

(PDF)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

S1 Fig. Maximum likelihood phylogeny of 40 diatom species and outgroup *Triparma laevis*. Phylogeny was used as constraint tree in substitution rate analyses. Bootstrap values less than 100% are indicated at the nodes. * indicates bootstrap of 100%. Scale is substitutions per codon.

(JPG)

S2 Fig. Distribution of nonsynonymous (*dN*) substitution rates for groups of genes and individual genes. The top and bottom lines of each box represent the 75th and 25th percentiles, respectively and the middle line in each box represents the 50th percentile. The whisker lines represent the minimum to the maximum points and the points outside of the whisker lines are outliers.

(PNG)

S3 Fig. Distribution of synonymous (*dS*) substitution rates for groups of genes and individual genes. The top and bottom lines of each box represent the 75th and 25th percentiles, respectively and the middle line in each box represents the 50th percentile. The whisker lines represent the minimum to the maximum points and the points outside of the whisker lines are outliers.

(PNG)

S.4 Fig. Relationship between the plastome size (one IR copy was excluded) and substitution rate.

(PNG)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Acknowledgements

We are grateful to 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 (UT Austin) for Illumina sequencing, the Texas Advanced Computing Center (TACC) at UT Austin for access to supercomputers and Erika Schwarz, Mao-Lun Weng and Jin Zhang for advice on rate analyses.

Authors contributions

Conceived and designed the experiments: RKJ, ECT, JMSM, ALA-M, MHA

Performed analyses and interpreted results: MY, TAR, MJS, AMA, ALA, ECT, RKJ

Secured funding for project: JMSM, NHH, MAK, RKJ, MHA

Wrote the paper: TAR, MY, RKJ, JSMS, ECT

References

1. Sorhannus U. A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. Marine Micropaleontology. 2007; 65: 1–12.
2. Nelson DM, Treguer P, Brzezinski MA, Leynaert A, Queguiner B. Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation, Global Biogeochemistry Cycles. 1995; 9: 359–372.
3. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: integrating terrestrial and oceanic components, Science. 1998; 281: 237–240.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

4. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, et al. The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism, Science. 2004; 306: 79–86.

5. Mann D G. The species concept in diatoms. Phycologia. 1999; 38:437–495.

6. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, et al. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes, Nature. 2008; 456: 239–244.

7. Yu M, Ashworth M, Hajrah NH, Khiyami MA, Sabir MJ, Alhebshi AM, et al. Evolution of the plastid genomes in diatoms Advamces in Botanical Research. 2018; 85:129–155.

8. Theriot EC, Ashworth M, Nakov T, Ruck EC, Jansen RK. A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. Plant Ecology and Evolution. 2010; 143: 278–296.

9. 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 Biology and Evolution. 2014; 6: 644–654.

10. Sabir JS, 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.

11. Theriot EC, Ashworth M, Nakov T, Ruck EC, Jansen RK. Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling, Molecular Phylogenetics and Evolution. 2015; 89: 28–36.

12. Martin W, Stoebe B, Goremykin V, Hapsmann S, Hasegawa M, Kowallik KV. Gene transfer to the nucleus and the evolution of chloroplasts. Nature. 1998; 393: 162–165.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

13. Kowallik KV, Stoebe B, SchaVran I, Kroth-Pancic P, Freier U. The chloroplast genome of a chlorophyll a + c- containing alga, *Odontella sinensis*. Plant Molecular Biology Reporter. 1995; 13; 336–342.

14. Oudot-Le Secq MP, Grimwood J, Shapiro H, Armbrust EV, Bowler C, Green BR. Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* and comparison with other plastid genomes of the red lineage. Molecular Genetics and Genomics. 2007; 277: 427–439.

15. Lommer M, Roy AS, 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.

16. 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. Photosynthesis Research. 2011; 109: 223–229.

17. Bedoshvili YD, Popkova TP, Likhoshway YV. Chloroplast structure of diatoms of different classes. Cell Tissue Biology. 2009; 3: 297–310.

18. Cooper JT, Malsy JP. Speciation in diatoms: Patterns, mechaninsms and environmental change. In: Pawel M, editor. Speciation: Natural Processes, Genetics and Biodiversity. Nova Science Publishers, New York, USA. 2013; pp 1–6.

19. T. Kuroiwa, T. Suzuki, K. Ogawa, S. Kawano. The chloroplast nucleus: Distribution, number, size, and shape, and a model for the multiplication of the chloroplast genome during chloroplast development, Plant Cell Physiology. 1981; 22: 381–396.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

20. Sato N. Origin and evolution of plastids: Genomic view on the unification and diversity of plastids. In: Wise RR, Hoober JK, editors. The Structure and Function of Plastids, Advances in Photosynthesis and Respiration, vol 23. Springer, Dordrecht. 2007; pp 75–102.

21. Oldenburg DJ, A.J Bendich. DNA maintenance in plastids and mitochondria of plants,. Frontiers in Plant Science. 2015; 6: 883.

22. Bromham L, Hua X, Lanfear R, Cowman PF. Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants, American Naturalist. 2015; 185: 507–524.

23. Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH et al. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): Comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Molecular Biology aand Evolution. 2006; 23: 279–291.

24. Guisinger MM, Kuehl JV, Boore JL, Jansen RK. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proceedings of the National Academy of Sciences USA. 2008; 105: 18424–18429.

25. Guisinger MM, Chumley TW, Kuehl JV, Boore JL, Jansen RK. Implications of the plastid genome sequence of *Typha* (Typhaceae, Poales) for understanding genome evolution in Poaceae. Journal of Molecular Evolution. 2010; 70: 149–166.

26. Schwarz EN, Ruhlman TA, Weng ML, Khiyami MA, Sabir JSM, Hajarah NH, et al. Plastome-wide nucleotide substitution rates reveal accelerated rates in Papilionoideae and correlations with genome features across legume subfamilies. Journal of Molecular Evolution. 2017; 84: 187–203.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

27. Sloan DB, Alverson AJ, Wu M, Palmer JD, Taylor DR. Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence in the angiosperm genus *Silene*. Genome Biology Evolution. 2012; 4: 294–306.

28. Dong W, Xu C, Cheng T, Zhou S. Complete chloroplast genome of *Sedum sarmentosum* and chloroplast genome evolution in Saxifragales. PLoS One 2013; 8: e77965.

29. Park S, Ruhlman TA, Weng ML, Hajrah NH, Sabir JSM, Jansen RK. Contrasting patterns of nucleotide substitution rates provide insight into dynamic evolution of plastid and mitochondrial genomes of *Geranium*. Genome Biology and Evolution. 2017; 9: 1766–1780.

30. Weng ML, Blazier JC, Govindu M, Jansen RK. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats and nucleotide substitution rates, Mol. Biol. Evol. 31 (2013) 645–659.

31. Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, et al. 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 USA. 2007; 104: 19369–19374.

32. Belda E, Moya A, Siva F.J. Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria. Molecular Biology and Evolution. 2005; 22: 1456–1467.

33. Shao R, Dowton M, Murrell A, Barker S.C. Rates of gene rearrangement and nucleotide substitution are correlated in the mitochondrial genomes of insects. Molecular Biology and Evolution. 2003; 20: 1612–1619.

34. Xu W, Jameson D, Tang B, Higgs P.G. The relationship between the rate of molecular evolution and the rate of genome rearrangement in animal mitochondrial genomes. Journal of Molecular Evolution. 2006; 63: 375–392.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

35. Jansen RK, Ruhlman TA. Plastid genomes of seed plants. In: Bock R, Knoop V, editors. Advances in Photosynthesis and Respiration, Volume 35: Genomics of Chloroplasts and Mitochondria. Springer, Netherlands. 2012; pp 103–126.

36. Zhang J, Ruhlman TA, Sabir JSM, Blazier JC, Weng ML, Park S, et al. Coevolution between nuclear-encoded DNA replication, recombination, and repair genes and plastid genome complexity. Genome Biology and Evolution. 2016; 8: 622–634.

37. Wu CS, Chaw SM. Highly rearranged and size-variable chloroplast genomes in conifers II clade (cupressophytes): Evolution towards shorter intergenic spacers. Plant Biotechnology Journal 12 (2014) 344–353.

38. Kooistra WHCF, Gersonde R, Medlin LK, Mann D.G. The origin and evolution of the diatoms: Their adaptation to a planktonic existence. In: Falkowski PG, Knoll AH, editors. Evolution of Primary Producers in the Sea. Burlington, Academic Press. 2007; pp 207–249.

39. Sorhannus U, Fox M. Synonymous and nonsynonymous substitution rates in diatoms: A comparison between chloroplast and nuclear genes. Journal of Molecular Evolution. 1999; 48: 209–212.

40. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. Nucleic Acids Research. 2005; 33: 511–518.

41. Stamatakis A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22: 2688–2690.

42. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 2007; 24: 1586–1591.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

43. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Research. 2004; 14: 1394–1403.

44. Tesler G. GRIMM: genome rearrangements web server. Bioinformatics. 2002; 18: 492–493.

45. Weng ML, Ruhlman TA, Gibby M, Jansen RK. Phylogeny, rate variation, and genome size evolution of *Pelargonium* (Geraniaceae). Molecular Phylogenetics and Evolution. 2012; 64: 654–670.

46. Sloan DB, Triant DA, Wu M, Taylor DR. Cytonuclear interactions and relaxed selection accelerate sequence evolution in organelle ribosomes. Molecular Biology and Evolution. 2014; 3: 673–682.

47. Blazier JC, Ruhlman TA, Weng ML, Rehman SK, Sabir JS, Jansen R.K. Divergence of RNA polymerase α subunits in angiosperm plastid genomes is mediated by genomic rearrangement. Scientific Reports. 2016; 6: 24595.

48. Weng ML, Ruhlman TA, Jansen RK. Plastid–nuclear interaction and accelerated coevolution in plastid ribosomal genes in Geraniaceae. Genome Biology and Evolution. 2016; 8: 1824–1838.

49. Wu CS, Chaw SM. Evolutionary stasis in cycad plastomes and the first case of plastome GC-biased gene conversion. Genome Biology and Evolution. 2015; 7: 2000–2009.

50. Wu CS, Chaw SM. Large-scale comparative analysis reveals the mechanisms driving plastomic compaction, reduction, and inversions in Conifers II (Cupressophytes). Genome Biology and Evolution. 2016; 8: 3740–3750.

51. Lemieux C, Otis C, Turmel M. Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. Frontiers in Plant Science 2016; 7: 697.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

52. Wilson AC, Carlson SS, White TJ. Biochemical evolution. Annual Review of Biochemstry. 1977; 46: 573–639.

53. Zhang J, Yang JR. Determinants of the rate of protein sequence evolution. Nature Reviews Genetics. 2015; 16: 409–420.

54. Havird JC, Sloan DB. The roles of mutation, selection, and expression in determining relative rates of evolution in mitochondrial versus nuclear genomes. Molecular Biology and Evolution. 2016; 11: 3042–3053.

55. Wright SI, Yau CB, Looseley M, Meyers BC. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*,]. Molecular Biology Evolution. 2004; 21: 1717–1726.

56. Ingvarsson PK. Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*. Molecular Biology and Evolution. 2007; 24: 836–844.

57. De La Torre AR, Lin YC, Van de Peer Y, Ingvarsson PK. Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in *Picea* gene families. Genome Biology and Evolution. 2015; 7: 1002–1015.

58. Shields DC, Sharp PM, Higgins DG, Wright F. Silent sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. Molecular Biology and Evolution. 1988; 5: 704–716.

59. Drummond DA, Raval A, Wilke CO. A single determinant dominates the rate of yeast protein evolution. Molecular Biology and Evolution. 2006; 23: 327–337.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

60. Shen Y, Lv Y, Huang L, Liu W, Wen M, Tang T, Zhang R, et al. Testing hypotheses on the rate of molecular evolution in relation to gene expression using microRNAs. Proceedings of the National Academy of Sciences USA. 2011; 108: 15942–15947.

61. Wicke S, Schneeweiss GM. Next-generation organellar genomics: Potentials and pitfalls of high-throughput technologies for molecular evolutionary studies and plant systematics. In: Hörandl E, Appelhans MS, editors. Next-Generation Sequencing in Plant Systematics International Association for Plant Taxonomy (IAPT), 2015: pp 1–42.

62. Gaut BS, Muse SV, Clegg MT. Relative rates of nucleotide substitution in the chloroplast genome. Molecular Phylogenetics and Evolution. 1994; 2: 89–96.

63. Green BR. Chloroplast genomes of photosynthetic eukaryotes. Plant Journal. 2011; 66: 34–44.

64. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, et al. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proceedings of the National Academy of Sciences USA. 2002; 99: 12246–12251.

65. Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of spontaneous mutation. Genetics. 1998; 148: 1667–1686.

66. Lynch M, Koskella B, Schaack S. Mutation pressure and the evolution of organelle genomic architecture. Science. 2006; 311: 1727-1730.

67. Ashworth MP, Ruck EC, Lobban CS, Romanovicz DK, Theriot EC. A revision of the genus *Cyclophora* and description of *Astrosyne gen. nov*. (Bacillariophyta), two genera with the pyrenoids contained within pseudosepta. Phycologia. 2012; 51: 684-699.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

68. Ruhlman TA, Jansen RK. Aberration or analogy? The atypical plastomes of Geraniaceae. Advances in Botanical Research. 2018; 85: 223–262.

69. Straub SC, Moore MJ, Soltis PS, Soltis DE, Liston A, Livshultz T. Phylogenetic signal detection from an ancient rapid radiation: Effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. Molecular Phylogenetics and Evolution. 2014; 80: 169–185.

70. Yi X, Gao L, Wang B, Su YJ, Wang T. The complete chloroplast genome sequence of *Cephalotaxus oliveri* (Cephalotaxaceae): evolutionary comparison of *Cephalotaxus* chloroplast DNAs and insights into the loss of inverted repeat copies in gymnosperms, Genome Biology and Evolution. 2013; 5: 688–698.

71. Ahmed I, Biggs PJ, Matthews PJ, Collins LJ, Hendy MD, Lockhart P.J. Mutational dynamics of aroid chloroplast genomes. Genome Biology and Evolution. 2012; 4: 1316-1323.

72. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research. 2001; 29: 4633–4642.

73. Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, et al. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. Plant Cell. 2002; 14: 2659–2679.

74. Pombert JF, Otis C, Lemieux C, Turmel M. The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages. Molecular Biology and Evolution. 2005; 22: 1903–1918.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

75. Pombert JF, Lemieux C, Turmel M. The complete chloroplast DNA sequence of the green alga *Oltmannsiellopsis viridis* reveals a distinctive quadripartite architecture in the chloroplast genome of early diverging ulvophytes. BMC Biology. 2006; 4: 3.

76. Maréchal A, Parent JS, Véronneau-Lafortune F, Joyeux A, Lang BF, Brisson N. Whirly proteins maintain plastid genome stability in *Arabidopsis*. Proceedings of the National Academy of Sciences USA. 2009: 106: 14693–14698.

77. Zampini É, Lepage É, Tremblay-Belzile S, Truche S, Brisson N. Organelle DNA rearrangement mapping reveals U-turn-like inversions as a major source of genomic instability in *Arabidopsis* and humans. Genome Research. 2015; 25: 645–654.

78. Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, et al. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature. 2008; 455: 105–108.

79. Zhu L, Wang Q, Tang P, Araki H, Tian D. Genome wide association between insertions/deletions and the nucleotide diversity in bacteria. Molecular Biology and Evolution. 2009; 26: 2353–2361.

80. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes, Nature Reviews Genetics. 2011; 12: 756–766.

81. McDonald MJ, Wang WC, Huang HD, Leu J.Y. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. PLoS Biology. 2011; 9: e1000622.

82. Maréchal A, Brisson N. Recombination and the maintenance of plant organelle genome stability New Phytologist. 2010; 186: 299–317.

83. Bendich AJ. Circular chloroplast chromosomes: the grand illusion, Plant Cell. 2004; 16: 1661–1666.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

84. Ruhlman TA, Zhang J, Blazier JC, Sabir JSM, Jansen RK. Recombination-dependent replication and gene conversion homogenize repeat sequences and diversify plastid genome structure. American Journal of Botany. 2017; 104: 559–572.

85. Zhang J, Ruhlman TA, Sabir JSM, Blazier JC, Jansen RK. Coordinated rates of evolution between interacting plastid and nuclear genes in Geraniaceae. Plant Cell. 2015; 27: 563–573.

86. Rockenbach K, Havird JC, Monroe JG, Triant DA, Taylor DR, Sloan DB. Positive selection in rapidly evolving plastid-nuclear enzyme complexes. Genetics. 2016; 204: 1507–1522.

87. 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. Marine Genomics. 2013; 21: 17–27.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Fig. 1. **Maximum likelihood *dN* and *dS* trees estimated from 103 concatenated protein-coding gene sequences**. The bars at the base of each tree indicate the number of nucleotide substitutions per codon and the *dN* and *dS* trees are on different scale. Numbers on the branches on the trees are branch numbers. The constraint tree is presented in Fig. B.1.

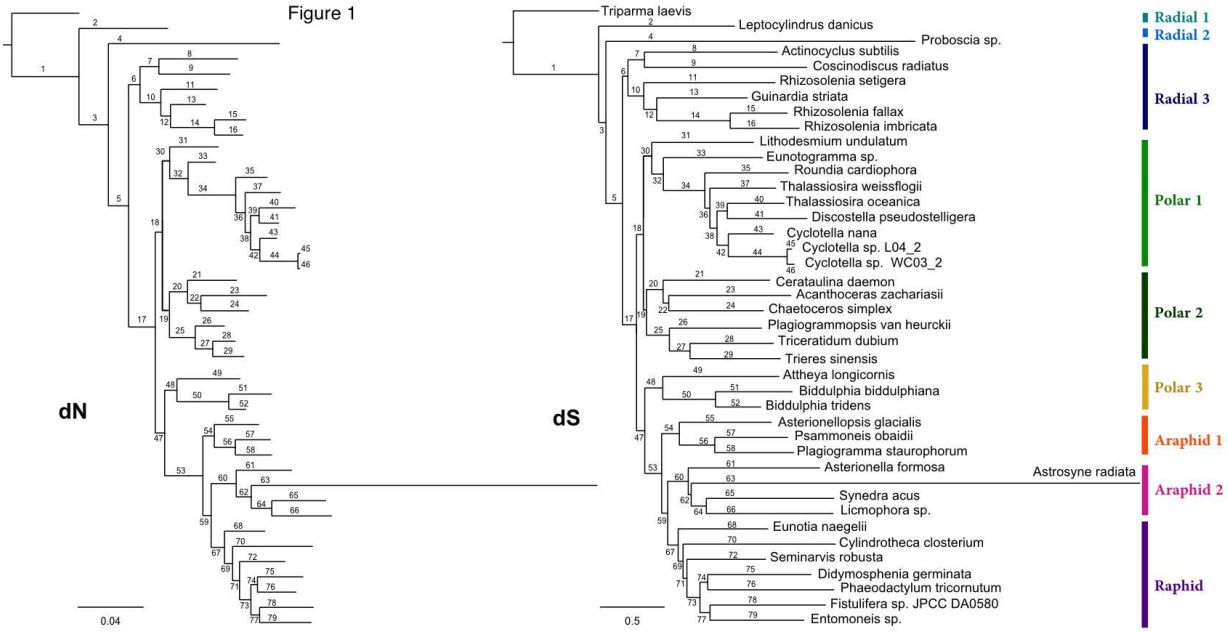
Fig. 2. **Distribution of the nonsynonymous (*dN*) and synonymous (*dS*) substitution rate for functional groups of genes across all included diatoms**. The top and bottom lines of each box represent the 75th and 25th percentiles, respectively and the middle line in each box represents the 50th percentile. The whisker lines represent the minimum to the maximum points and the points outside of the whisker lines are outliers.

Fig. 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.

Fig. 4. **Pairwise correlation of substitution rate and plastome inversion distance in diatoms**. 0.05 (red horizontal line) was used to assess the level of significance; P-values were plotted on the X axis. Colored bars indicate different clades of diatoms and correspond to Fig. 1. From left to right: radial 1, radial 2, radial 3, polar 1, polar 2, polar 3, araphid 1, araphid 2, raphid.

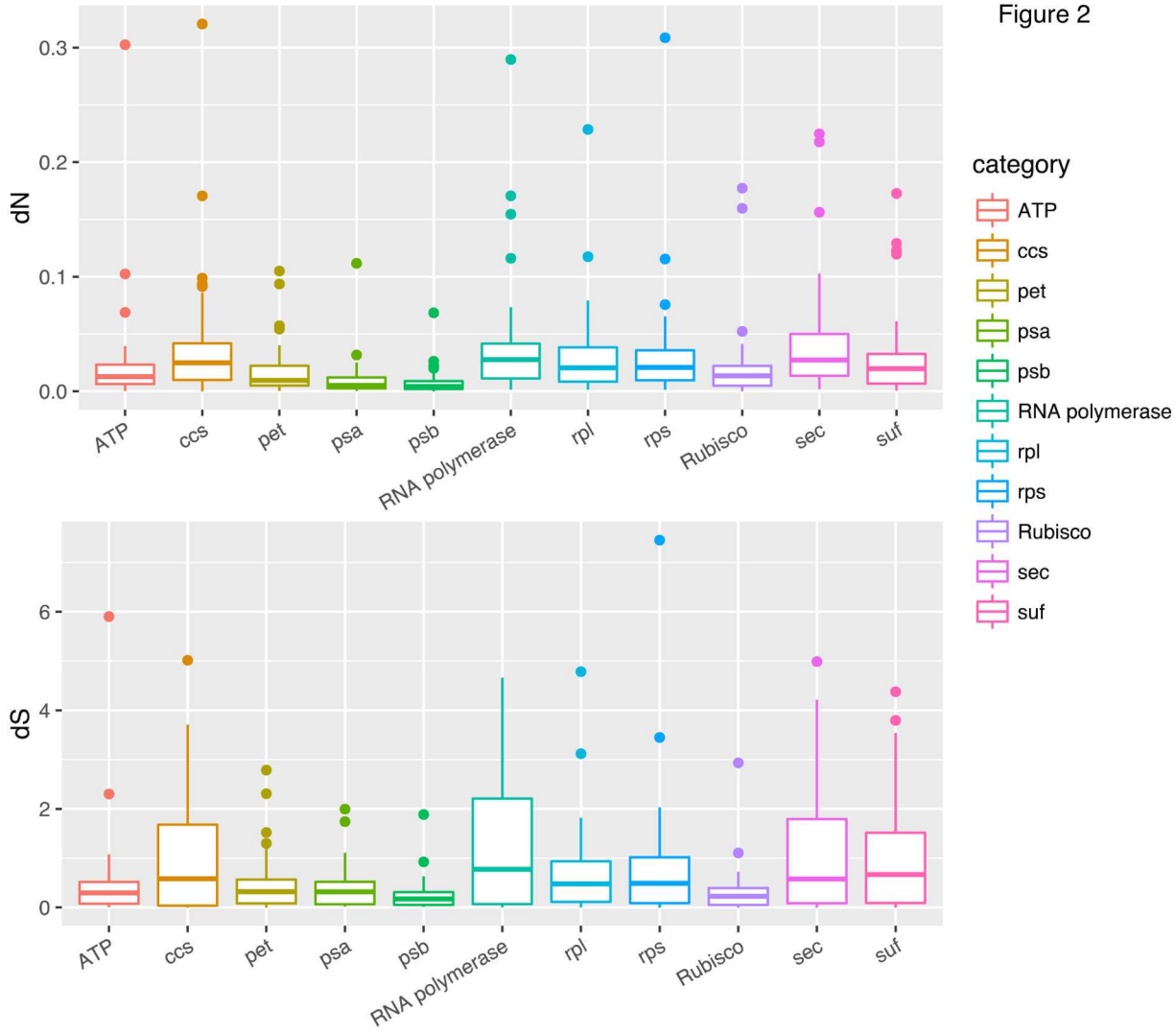
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Figure 1



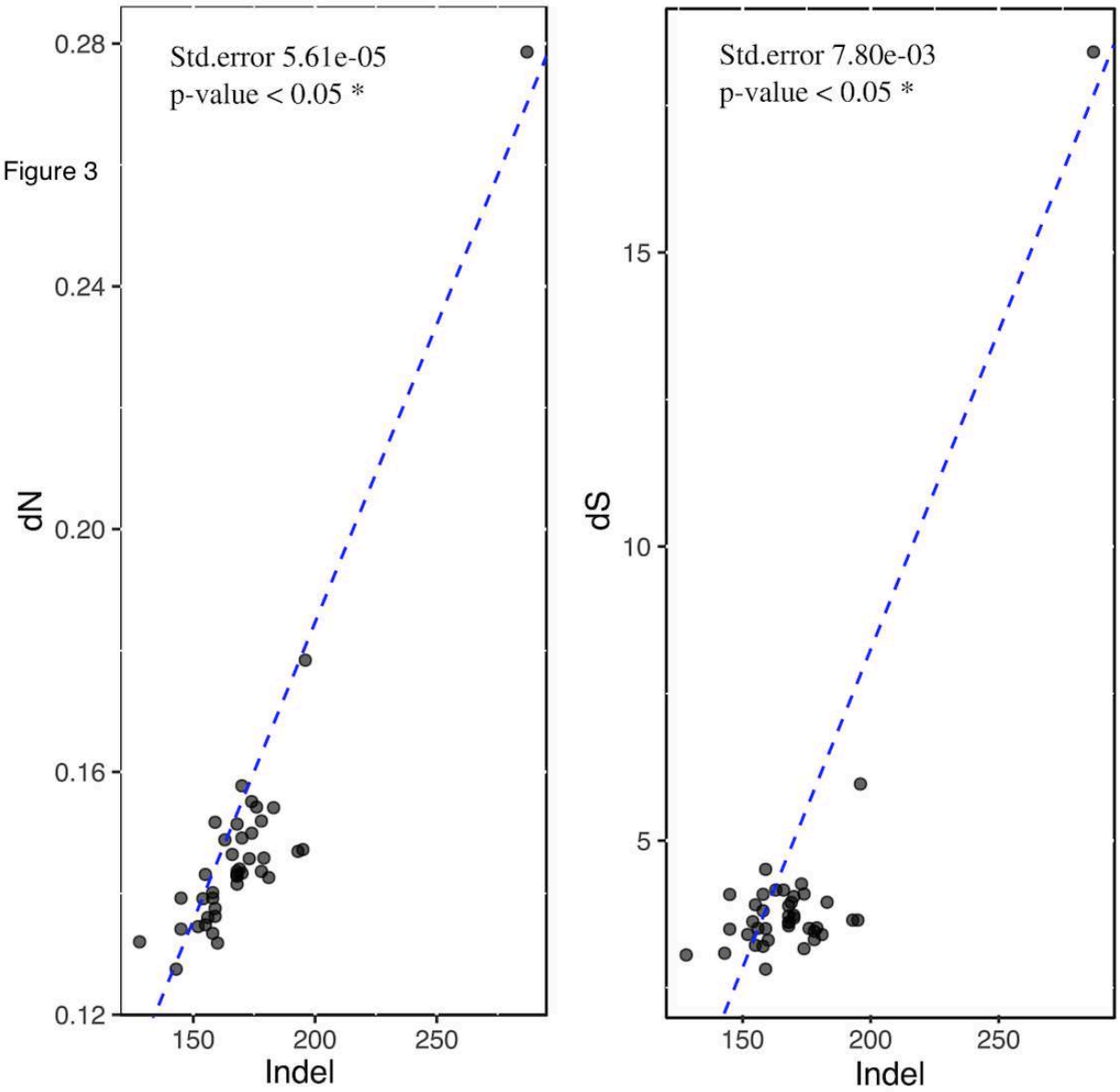
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Figure 2



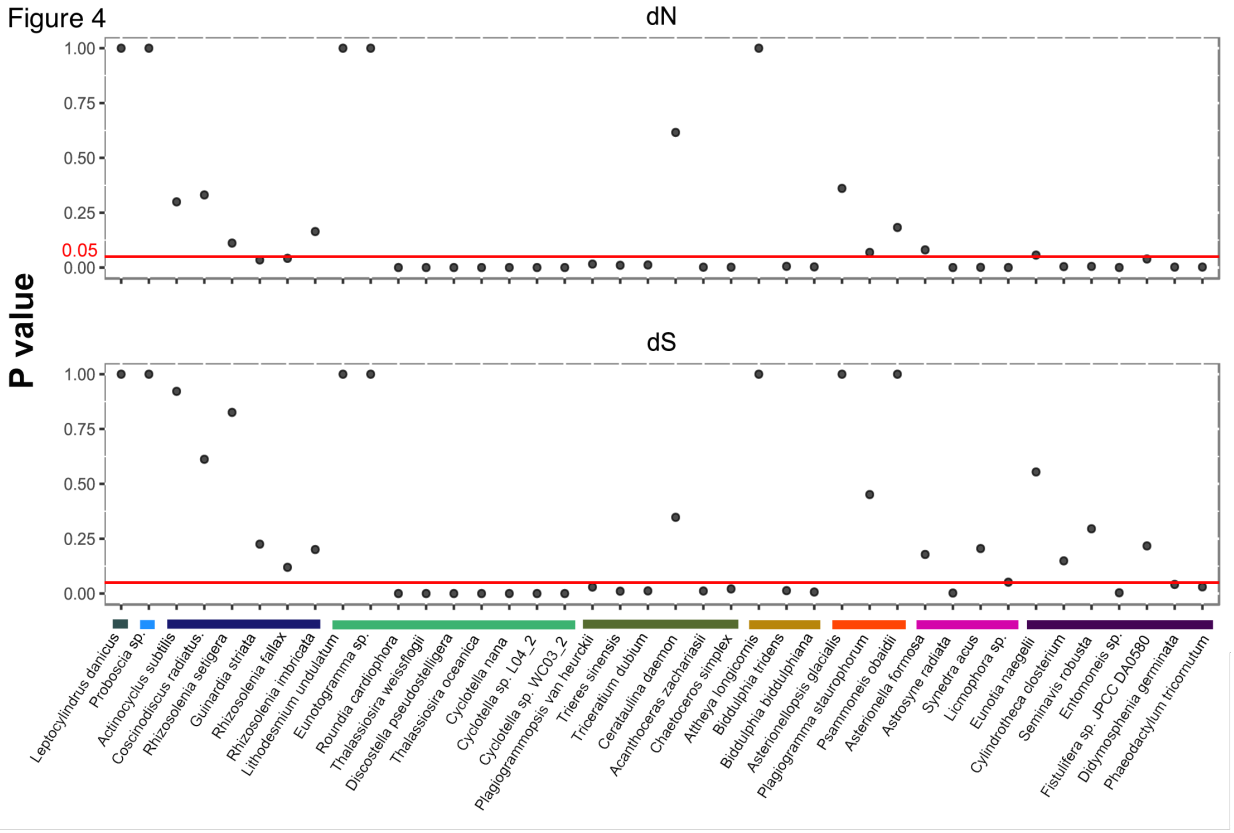
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Figure 3



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019

Figure 4



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27636v1 | CC BY 4.0 Open Access | rec: 5 Apr 2019, publ: 5 Apr 2019