Algal Research 95 (2026) 104605

Contents lists available at ScienceDirect

Algal Research

journal homepage: www.elsevier.com/locate/algal

# The roles of transposable elements and gene family dynamics in shaping diversity and evolution in diatoms

Tianze Zheng [a,b], Xu Zhang [a], Tianren Liu [a], Xinzhu Liu [a], Chris Bowler [b], Xin Lin [a,c,*]

[a] State Key Laboratory of Marine Environmental Science and College of Ocean and Earth Sciences, Xiamen University, Xiamen, China
[b] Institut de Biologie de l'ENS (IBENS), Département de Biologie, École Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France
[c] Advanced Institute for Marine Studies, Fujian Ocean Innovation Center, Xiamen, China

ARTICLE INFO

*Keywords:*
Comparative genomics
Diatoms
Evolution
Long-read sequencing
*Phaeodactylum tricornutum*
Transposable elements

ABSTRACT

Diatoms are key players in aquatic ecosystems, having evolved through secondary endosymbiosis. Using long-read sequencing, we investigated how transposable elements (TEs) and gene family dynamics have shaped diatom diversification from inter-lineage to intra-species scales. Across diatom lineages, we identified ecological adaptation-linked expansions, including polyamine synthesis genes for silicification and glutathione S-transferases for oxidative stress resistance. Centric diatoms showed lineage-specific expansion of flotation-associated microtubule genes, while pennate diatoms expanded motility-related actin and myosin genes. At the intra-species level, distinct *Phaeodactylum tricornutum* strains revealed genomic adaptations correlated with their unique features, including strain-specific expansion and contraction in the cruciform strain's morphological genes and the Baltic Sea isolate's amine metabolism genes. Our estimates of major lineage divergence times in diatoms (~202 Myr and ~ 173 Myr) were highly consistent with the two deep whole-genome duplication (WGD) events (~200 Myr and ~ 170 Myr). At these evolutionary nodes, gene families showed extensive lineage-specific expansions and contractions, likely linking ancient polyploidy to subsequent gene content evolution. Substantial TE expansions occurred more recently (0.5–5 Ma), with most diatoms showing recent bursts of Long Terminal Repeat Retrotransposons (LTR-RTs) and araphid pennate diatoms displaying more ancient TE insertion peaks. This likely reflects the progressive loss of ancient TE copies, leaving only recent TE insertions detectable. Our findings provide genomic evidence for the adaptive evolution of diatoms, highlighting the crucial roles of TEs and gene family dynamics in shaping their morphological diversity and environmental adaptations, and suggesting a potential connection between WGDs, gene family dynamics, and TE insertions in genome evolution.

## 1. Introduction

Diatoms are ecologically dominant unicellular algae that originated through secondary endosymbiosis and contribute ~20% of global primary production [1,2]. Fossil records suggest that diatoms first appeared during the Cretaceous [3,4]. Diatoms are generally classified as centric or pennate [5]. Centric diatoms typically have silica frustules that are radially or cylindrically shaped, lacking a raphe, and exhibiting limited mobility [6]. Pennate diatoms exhibit bilateral symmetry and typically possess a raphe that enables motility [7].

Among pennate diatoms, *Phaeodactylum tricornutum* has emerged as a model species due to its morphological plasticity, physiological adaptability, and availability of multiple natural strains [8]. Previous studies have revealed extensive genetic and transcriptomic diversity among its strains [9] [10]. To investigate intraspecific genome evolution in a controlled framework, we focused on three *P. tricornutum* strains (PtSCS, PtECS, and Pt4) that represent distinct ecological, morphological, and genomic backgrounds within the species. Pt4 is a Baltic Sea strain adapted to low salinity and low light environments and has been widely recognized as an ecophysiologically distinct lineage [11,12]. PtSCS and PtECS originate from the South China Sea and East China Sea, respectively, but exhibit marked morphological divergence: PtSCS is predominantly fusiform [13], while PtECS is characterized by a cruciform shape. Here, for the first time, we report the genome sequencing data of the cruciform strain of *P. tricornutum*. Together, these strains capture substantial intraspecific diversity across contrasting morphotypes and ecological niches.

Comparative genomics provides a powerful framework to investigate

* Corresponding author.
*E-mail address:* xinlinulm@xmu.edu.cn (X. Lin).

https://doi.org/10.1016/j.algal.2026.104605
Received 27 October 2025; Received in revised form 15 January 2026; Accepted 20 February 2026
Available online 24 February 2026

genome architecture, genetic diversity, and evolutionary mechanisms across species [14]. With the advent of long-read sequencing, which produces highly complete genome assemblies, it has become possible to systematically study lineage-specific adaptations and evolutionary mechanisms [15]. Recent advances in plant comparative genomics are exemplified by genomic analyses elucidating the origin of *Citrus* species [16], in-depth studies of *Allium fistulosum* [17], and the assembly of the *Malus baccata* genome, which revealed adaptive gene family evolution [18]. Short-read comparative analyses have revealed genomic signatures underlying marine-to-freshwater transitions in diatoms [19]. But the contributions of gene family expansions and contractions to diatom diversification remain poorly understood.

In eukaryotes, whole-genome duplication (WGD), is a major mechanism generating new templates for evolutionary innovation [20,21] with biased retention of dosage-sensitive genes (e.g., transcription factors) typically causing gene family expansions/contractions [22]. Transposable elements (TEs) are recognized as major drivers of genome architecture and eukaryotic evolution by introducing mutations and genetic novelty [23]. In diatoms, which experience highly variable environmental conditions and rapid ecological diversification, these processes may play an especially important role in facilitating metabolic flexibility, regulatory innovation, and niche adaptation. However, how WGDs, TE activity, and gene family dynamics interact to shape diatom evolution remains poorly understood. Characterizing and comparing TEs using short-read sequencing remains challenging due to difficulties in resolving repetitive flanking regions essential for accurate TE detection [24]. Long-read sequencing technologies such as PacBio and Oxford Nanopore (ONT) provide critical advantages in resolving repetitive sequences, thereby enabling more comprehensive TE identification and revealing hidden genomic diversity, such as structural variants and TE-rich regions [25]. Recent studies using long-read sequencing have uncovered extensive TE repertoires and their regulatory roles across eukaryotes [26–28].

Previous studies, particularly in plants, have shown that TE bursts often follow WGD, a pattern initially framed by McClintock's "genome shock" hypothesis, which proposed that WGD could trigger immediate TE activation [29]. More recent studies, however, suggest that TE accumulation was often a slower, long-term process shaped by relaxed purifying selection in polyploids [30] [31]. TE proliferation was also frequently associated with rapid genome size increases, which may occur independently or in synergy with WGD [32].

Diatoms have experienced multiple ancient whole-genome duplications [33], yet how these events interact with subsequent TE proliferation and gene family evolution remains unclear. Addressing these gaps requires integrating long-read sequencing with comparative genomic approaches across multiple diatom lineages, which offers a unique opportunity to examine how these processes collectively drive ecological diversification in diatoms. Based on the above background, the scientific question of this study is: (1) How do TE landscapes vary across diatom lineages? (2) how do gene family dynamics influence adaptive evolution in diatoms, and what gene families have expanded or contracted in association with ecological adaptations in diatoms? (3) What is the relationship between WGDs, gene family dynamics, and TE proliferation?

In this study, we compiled long-read sequencing data from 29 diatoms and representative outgroups (one coccolithophore, three green algae, and four red algae) to analyze TE diversity and gene family expansion/contraction across lineages. In parallel, we generated ONT-based genomes for three geographically distinct *P. tricornutum* strains to examine how TE dynamics and gene family variation contribute to functional and physiological differentiation within a model species. We further investigated the relationships between WGDs, gene family dynamics, and TE proliferation to understand their coordinated roles in shaping genome evolution. Together, this comparative genomics with strain-level analyses provides new insights into the evolutionary forces shaping diatom diversity.

## 2. Materials and methods

### 2.1. Genome sequencing of P. tricornutum

Three *P. tricornutum* strains—PtSCS, PtECS, and Pt4—were sequenced (Fig. 1). Pt4 was isolated from the Baltic Sea and deposited in the Culture Collection of Algae and Protozoa (CCAP 1052/6) [9]. PtSCS was collected from the South China Sea in 2004 and deposited in the Center for Collections of Marine Bacteria and Phytoplankton, Xiamen University (CCMBP106) [34]. PtECS, a cruciform strain from the Yangtze River estuary in the East China Sea, was deposited in the same collection (CCMBP267).

Cultures were treated with an antibiotic cocktail, maintained in f/2 enriched seawater at 19 °C under a 12 h light/dark cycle, and harvested during exponential growth. High-quality genomic DNA was extracted following the protocol described in [35]. Cells were harvested by centrifugation and lysed in SDS buffer. The lysates were subjected to ten freeze–thaw cycles, followed by digestion with proteinase K and treatment with RNase A. DNA was then purified by phenol–chloroform extraction and isopropanol precipitation, and finally resuspended in 10 mM Tris-HCl (pH 8.0).

Sequencing was performed at Biomarker Technologies (Beijing, China) using Oxford Nanopore Technology (ONT). Long reads (up to ~2 Mbp) were generated with the SQK-LSK109 ligation kit. Raw signals were collected using MinKNOW and subsequently base-called with Guppy (Oxford Nanopore Technologies; specific version information not available from archived data), and converted to FASTQ format for downstream analysis [25,36].

### 2.2. Genome assembly and annotation

We conducted de novo genome assembly using filtered reads, employing NextDenovo followed by NextPolish-based long-read correction (two rounds) and Pilon-mediated short-read polishing (three rounds) with Illumina data. Gene structural annotation was conducted using Braker2 [37] and EVidenceModeler (EVM) processes, to incorporate transcriptomic data. Additionally, gene function annotation was conducted using EggNOG-mapper [38].

### 2.3. Genomic data compiled and analysis of long-read sequencing data

We analyzed long-read genome assemblies from diatoms and representative outgroups, retaining only high-contiguity genomes with reliable gene annotations to ensure robust inference of TE landscapes and gene family dynamics. We analyzed published genomic data from 29 diatoms sequenced with Nanopore/PacBio platforms, including the *P. tricornutum* strain Pt1 isolated from the British Irish Sea [35]. Only diatom genomes meeting stringent quality criteria were included in this study: those based on long-read assemblies with sufficient contiguity and supported by high-quality gene annotation files to ensure accurate TE and gene family analysis, as well as reliable TE insertion-time inference.

The dataset also incorporated one coccolithophore, three green algae, and four red algae obtained from NCBI, JGI, and Ensembl databases. We performed comparative genomic analysis on selected diatoms and outgroups (one coccolithophore, three green algae, and three red algae) with high-quality genomes and GFF annotations (Table 1).

Genome completeness was assessed using BUSCO v5.4+ with both universal eukaryotic (eukaryota_odb10) and lineage-specific datasets: stramenopiles_odb10 for diatoms, chlorophyta_odb10 for green algae, and rhodophyta_odb12 for red algae, ensuring the most appropriate evolutionary context for each taxonomic group. BUSCO completeness assessment revealed generally high genome quality across the analyzed species, with scores ranging from 76.0% to 100%. Most diatom genomes showed excellent assembly quality (92–99% BUSCO), with *Nitzschia inconspicua* being complete at 100%.

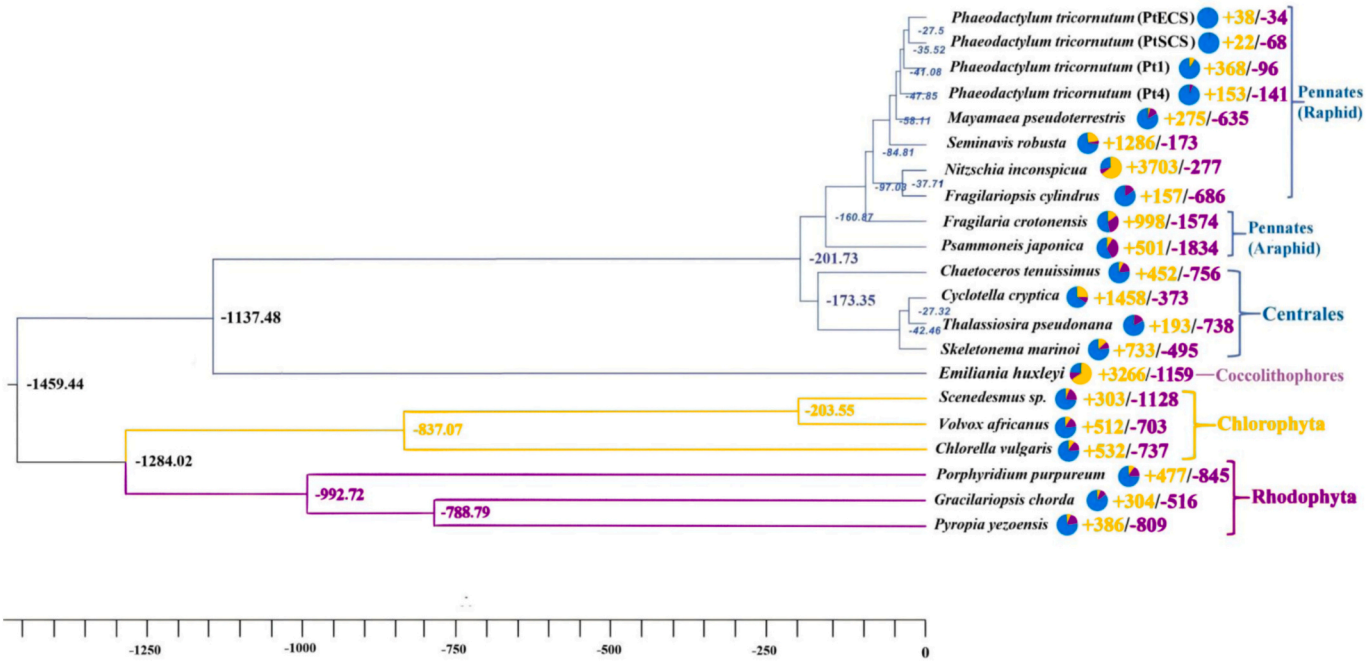


**Fig. 1.** Analysis of divergence times and gene family evolution in 14 diatom species, based on long-read sequenced, well-annotated nuclear genomes from both pennate and centric lineages, one coccolithophore, three green algae, and three red algae. Different phyla are colour-coded. The expansion (+) and contraction (−) of gene families are displayed on the phylogenetic branches. Pie charts following species names show contraction (red), expansion (green), and unchanged (blue) gene families, while numbers following each node indicate estimated divergence times for those taxa. One unit of length equals one million years. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

### *2.4. Comparative genomic analysis of diatoms and selected coccolithophore, red and green algal species*

We conducted a comparative genomic analysis of 21 species, including 14 diatoms with high-quality annotated genomes (pennate and centric), one coccolithophore, three green algae, and three red algae as outgroups. The diatoms included four *P. tricornutum* strains (Pt1, Pt4, PtSCS, PtECS). We performed gene family clustering analysis using Diamond (v2.1.8) to compare sequences and identify homologous proteins, followed by OrthoFinder [59] to cluster gene families and identify shared and unique ones.

We reconstructed the phylogeny and estimated divergence times using a Bayesian approach in MCMCtree (PAML package; [60]). The analysis was based on OrthoFinder-derived alignments filtered with TrimAl (gt = 0.6, cons = 60), with the species tree constructed using RAxML-HPC-PTHREADS and rooted in MEGA. Divergence times were estimated using the MCMCtree tool. Gene family expansions and contractions were analyzed using CAFE5 [61].

We performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses on expanded/contracted gene families across different evolutionary nodes. In addition to analyses at key evolutionary nodes, we also conducted species-specific GO and KEGG analyses for the model diatom *P. tricornutum*, *N. inconspicua* (selected due to its extensive gene family expansion), and the two araphid pennate diatoms. The araphid pennate diatom *Psammoneis japonica* and *Fragilaria crotonensis* displayed unique genomic and TE features, suggesting their potential transitional position in diatom evolution between centric and pennate diatoms. To explore their evolutionary relationship within the centric-pennate diatom divergence, we performed whole-genome duplication (WGD) analysis using the 'WGD' tool, calculating synonymous substitutions per synonymous site (Ks) [62].

### *2.5. TE annotation and divergence analysis of diatoms and representative red algal and green algal species*

TE annotation was performed using a sequential, integrated pipeline. First, comprehensive de novo TE identification and initial classification were conducted using the Extensive de novo TE Annotator (EDTA, v2.2.2) [63], which detects TEs based on structural features including terminal repeats, target site duplications, and homology. Second, DeepTE [64], a deep learning-based classifier, was applied to refine the classification of TEs that remained unclassified or ambiguously annotated by EDTA. This complementary approach maximizes both the sensitivity of TE detection and the accuracy of superfamily assignment. Based on the integrated annotation results, we compiled characteristic parameters of TEs across different species of diatoms, green algae, and red algae.

To investigate TE dynamics in diatoms, we analyzed LTR-RT insertion patterns in a subset of 12 diatom species selected based on two criteria: (i) relatively high LTR-RT abundance, ensuring sufficient signal for insertion time inference, and (ii) high-quality, near-chromosome-level genome assemblies, minimizing biases associated with assembly fragmentation in repetitive regions. These species were selected from the available diatom genomes to balance phylogenetic diversity while ensuring reliable estimation of LTR-RT insertion dynamics and capturing key evolutionary and functional variation within diatoms. We dated Long Terminal Repeat Retrotransposons (LTR-RTs) insertion time using a sequence identity-based consensus approach, where the nucleotide identity between individual TE copies and their subfamily consensus sequences serves as a proxy for time since insertion. To further place these estimates on an absolute timescale, we applied a molecular clock calibration based on the mutation rate of *P. tricornutum* [65]. Using RAxML, we constructed a phylogenetic tree based on the RT domains of full-length LTR-RTs.

**Table 1**
Overview of genomic data from published studies for selected diatoms, red algae, and green algae, including species name, sequencing technology used (e.g., Nanopore, PacBio, Illumina), genome size in megabase pairs (Mbp), number of contigs, the length of the largest contig in Mbp, the sequence length of the shortest contig at 50% of the total assembly length (N50) in Mbp, GC content percentage, Data source, and the score of Busco.

| Species | Sequencing technology | Size (Mbp) | Contigs | Largest length (Mb) | N50 (Mb) | GC content | Annotation | Data source | Busco (%) |
|---|---|---|---|---|---|---|---|---|---|
| *Phaeodactylum tricornutum* (Pt1) [35] | Nanopore | 32.93 | 45 | 2.884 | 1.161 | 47.36% | √ | NCBI:PRJNA487263 | 93.0 |
| *Mayamaea pseudoterrestris* [39] | Nanopore | 30.63 | 41 | 2.995 | 1.068 | 40.60% | √ | DDBJ: BROI01000001–BROI01000041 | 95.0 |
| *Seminavis robusta* [40] | PacBio;Illumina | 125.57 | 4752 | 0.318 | 0.051 | 39.94% | √ | ENA:PRJEB36614 | 99.0 |
| *Fragilariopsis cylindrus* [41] | PacBio | 80.54 | 271 | 5.926 | 1.296 | 36.44% | √ | DDBJ/EMBL:PRJEB15040 | 94.0 |
| *Nitzschia inconspicua* [42] | PacBio | 99.92 | 125 | 6.575 | 3.618 | 38.79% | √ | NCBI:PRJNA675887 | 100 |
| *Fragilaria crotonensis* [43] | Nanopore; Illumina | 62.11 | 881 | 0.902 | 0.089 | 43.23% | √ | NCBI:PRJNA807324 | 76.0 |
| *Thalassiosira pseudonana* [35] | Nanopore | 33.82 | 52 | 2.762 | 1.385 | 46.95% | √ | NCBI:PRJNA487263 | 96 |
| *Volvox africanus* [44] | PacBio;Illumina | 129.33 | 448 | 6.702 | 1.357 | 39.21% | √ | DDBJ:DRA010672 | 82.2 |
| *Scenedesmus* sp. [45] | PacBio | 39.97 | 80 | 4.854 | 1.281 | 25.26% | √ | NCBI:PRJNA637367 | 91.1 |
| *Chlorella vulgaris* [46] | PacBio;Illumina | 40.44 | 45 | 5.423 | 2.825 | 40.05% | √ | DDBJ:SIDB00000000 | 99.0 |
| *Porphyridium purpureum* [47] | Nanopore | 22.19 | 52 | 5.718 | 1.854 | 49.70% | √ | NCBI:PRJNA560054 | 84 |
| *Gracilariopsis chorda* [48] | PacBio | 92.18 | 1211 | 1.357 | 0.22 | 35.33% | √ | NCBI:PRJNA560054 | 93 |
| *Chaetoceros tenuissimus* [49] | Nanopore; Illumina | 41.15 | 87 | 4.456 | 2.427 | 34.58% | √ | NCBI:PRJNA361418 | 95 |
| *Cyclotella cryptica* [50] | Nanopore; Illumina | 166.00 | 662 | 2.498 | 0.494 | 32.96% | √ | NCBI:PRJNA628076 | 96 |
| *Skeletonema marinoi* [51] | PacBio;Hi-C | 63.00 | 52 | 5.945 | 3.005 | 38.64% | √ | NCBI:PRJNA960877 | 92 |
| *Psammoneis japonica* | PacBio;Illumina | 89.00 | 597 | 1.222 | 0.378 | 41.31% | √ | NCBI:PRJNA476996 | 78 |
| *Pyropia yezoensis* [52] | PacBio;Hi-C; Illumina | 104.00 | 28 | 43.6 | 34.3 | 32.82% | √ | NCBI:PRJNA589917 | 92 |
| *Attheya* sp. [53] | PacBio | 155.83 | 712,539 | 0.07 | 0.0029 | 43.34% | | NCBI:PRJNA517804 | |
| *Cyclotella meneghiniana* [54] | PacBio | 117.82 | 397,078 | 0.069 | 0.0029 | 43.40% | | NCBI:PRJNA825288 | |
| *Fistulifera pelliculosa* | Nanopore | 30.16 | 8063 | 0.18 | 0.021 | 41.55% | | NCBI:PRJDB12112 | |
| *Amphora coffeaeformis* [53] | PacBio | 42.3 | 3718 | 0.22 | 0.025 | 47.71% | | NCBI:PRJNA517804 | |
| *Minidiscus variabilis* | PacBio | 76.19 | 584 | 1.57 | 0.36 | 40.08% | | JGI:CCMP495 v1.0 | |
| *Nitzschia putrida* [53] | PacBio RSII; Illumina | 47.13 | 234 | 3.79 | 0.55 | 43.21% | | NCBI:PRJNA517804 | |
| *Skeletonema costatum* [55] | PacBio; Illumina | 51.13 | 1282 | 0.76 | 0.098 | 39.37% | | NCBI:PRJNA647329 | |
| *Thalassiosira oceanica* [53] | PacBio | 83.51 | 48,312 | 0.049 | 0.004 | 43.79% | | NCBI:PRJNA517804 | |
| *Chaetoceros muellerii* [53] | PacBio | 37.74 | 8639 | 0.21 | 0.02 | 36.85% | | NCBI:PRJNA517804 | |
| *Chaetoceros calcitrans* [53] | PacBio | 23.09 | 12,915 | 0.096 | 0.017 | 39.52% | | NCBI:PRJNA517804 | |
| *Chaetoceros neogracile* [53] | PacBio | 34.98 | 12,407 | 0.22 | 0.024 | 41.96% | | NCBI:PRJNA517804 | |
| *Skeletonema menzelii* [53] | PacBio | 38.93 | 34,909 | 0.096 | 0.012 | 44.72% | | NCBI:PRJNA517804 | |
| *Skeletonema dohrnii* [53] | PacBio | 59.40 | 43,142 | 0.099 | 0.006 | 45.17% | | NCBI:PRJNA517804 | |
| *Cylindrothec fusiformis* [53] | PacBio | 49.97 | 27,045 | 0.347 | 0.017 | 45.71% | | NCBI:PRJNA517804 | |
| *Navicula incerta* [53] | PacBio | 60.9 | 21,614 | 0.0625 | 0.003 | 44.78% | | NCBI:PRJNA517804 | |
| *Navicula pelliculosa* [53] | PacBio | 28.12 | 3258 | 0.181 | 0.022 | 49.01% | | NCBI:PRJNA517804 | |
| *Epithemia pelagica* | PacBio | 60.52 | 23 | 6.983 | 3.86 | 41.71% | | NCBI:PRJEB56202 | |
| *Craspedostauros australis* [53] | PacBio | 74.83 | 90 | 3.87 | 1.72 | 44.68% | | NCBI:PRJNA517804 | |
| *Galdieria partita* [56] | PacBio | 17.97 | 82 | 0.365 | 0.225 | 33.42% | √ | NCBI:PRJDB12742 | 95 |
| *Porphyra umbilicalis* [57] | PacBio | 87.89 | 2126 | 0.971 | 0.202 | 39.36% | √ | DDBJ:MXAK00000000 | 80 |
| *Emiliania huxleyi* [58] | PacBio | 190 | 600 | 3.53 | 0.68 | 63.31% | √ | NCBI:PRJNA789191 | 77.7 |

## 3. Results

### 3.1. Estimation of divergence time and gene family expansions and contractions among diatoms

To reconstruct diatom evolutionary history, we analyzed 21 genomes, including 14 diatoms representing major centric and pennate lineages, alongside one coccolithophore, three green algae, and three red algae. Divergence time estimates place the origin of diatoms at ~202 Myr, with centric diatoms emerging earlier (~173 Myr) and pennate diatoms diversifying later, including *Psammoneis japonica* (~161 Myr) and *Fragilaria crotonensis* (~97 Myr), consistent with previous evolutionary frameworks [66].

In diatoms, glutathione metabolism genes—particularly glutathione S-transferases—showed widespread expansion, as did polyamine metabolism and carotenoid biosynthesis pathways, both central to frustule silicification and photoprotection. In centric diatoms, gene families related to microtubules and cellular protrusions were expanded. Pennate diatoms displayed expansions of actin- and myosin-related genes underlying motility (Supplementary Table S2).

The pennate araphid diatoms *P. japonica* and *F. crotonensis* showed substantial gene family contractions (1834 and 1574, respectively). *C. cryptica* expanded 1458 gene families, while *S. robusta* and *N. inconspicua* expanded 1286 and 3703 gene families, respectively (Fig. 1).

Functional enrichment pointed to expansions of fucoxanthin chlorophyll *a*/c binding proteins, HSP70s, carbonic anhydrases, and Rubisco at key evolutionary nodes (Supplementary Table S5). We observed a striking gene family expansion in *N. inconspicua*, with functional enrichment analyses revealing that the expanded families were primarily associated with spliceosome, ribosome, RNA metabolism, and cell cycle regulation (KEGG), as well as nitrogen compound transport, protein localization, energy metabolism, and organelle organization (GO) (Fig. S3, S4). In *P. tricornutum*, genes involved in DNA integration and histone modifications showed pronounced expansions, suggesting lineage-specific regulatory innovations (Supplementary Table S2).

### 3.2. The diversity of TEs in diatoms and comparison with other algal species

To examine TE distribution patterns across algal lineages, we analyzed and compared TE content in diverse diatoms, green algae, and red algae. *Cyclotella cryptica* had the highest overall TE and LTR-RT proportions among diatom genomes (53.37% and 34.98%, respectively). *Attheya* sp. ranked second in total TE content (47.84%) and displayed the highest proportion of TIR elements (15.84%) among all species analyzed (Fig. 2; Supplementary Table S1). *Chaetoceros australis* had the second-highest LTR-RT proportion (17.95%).

Species with larger genomes, such as *C. cryptica* (166 Mb) and *Attheya* sp. (155.8 Mb), contained substantially higher proportions of TEs compared to species with smaller genomes. Notably, *F. crotonensis*, an araphid pennate diatom, exhibited distinctive TE distribution patterns with LTR-RTs accounting for up to 90% of its total TE content.

### 3.3. Phylogenetic analysis of LTR-RTs indicates divergent insertion patterns among diatoms

We conducted a phylogenetic analysis using conserved domain sequences to elucidate the distribution and evolution of LTR-RTs across different diatom species. *C. cryptica*, *F. crotonensis* and *P. japonica* exhibited high LTR-RT diversity. In contrast, *Thalassiosira pseudonana* exhibited markedly lower LTR-RT diversity (Fig. 3).

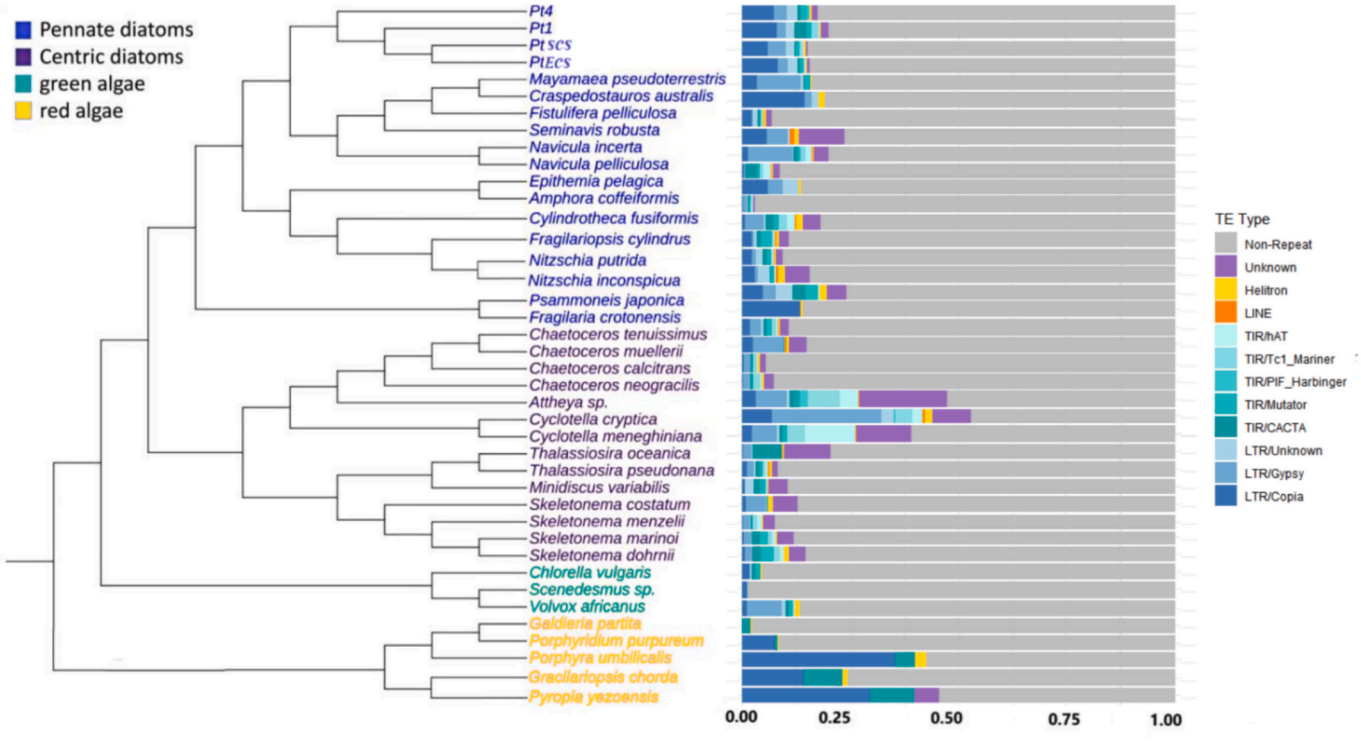


**Fig. 2.** The content of TEs in 32 species of diatoms, three green algae and five red algae species. The left side of the figure shows the phylogenetic relationship of centric diatoms, pennate diatoms, green algae, and red algae, represented in different colour areas; the right side represents the proportion of TEs in the genome of each species, with different colors distinguishing different types of TEs (subdivided into superfamilies). The phylogenetic relationship between different diatoms and other algae species is based on NCBI common tree (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

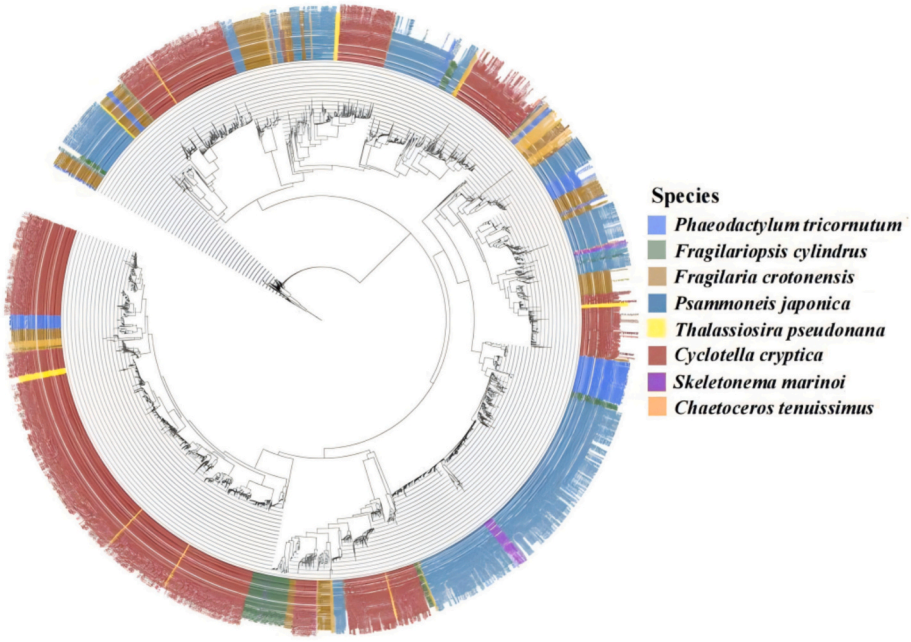


**Fig. 3.** The phylogenetic analysis of LTR-RTs across eight diatom species (including five representative pennate and centric species) based on conserved domain sequences. The tree was constructed using conserved RT domain sequences. Different colored labels represent different diatom species.

### 3.4. Diversity in LTR-RT insertion patterns among diatoms

Twelve selected diatom species from the available diatom genomes represent evolutionary and functional variation within diatoms while enabling reliable estimation of LTR-RT insertion dynamics. Insertion times were estimated using sequence identity between LTR-RT elements and their ancestral sequences based on EDTA annotations. Higher sequence identity indicates more recent insertions, whereas lower identity reflects older events that have accumulated mutations over time.

Most diatoms exhibited a recent burst of LTR-RT activity. *Attheya* sp. showed the most prominent recent peak, while *C. cryptica* and *Psammoneis japonica*—despite their high TE content—lacked detectable recent insertions. Interestingly, the araphid diatoms *F. crotonensis* and *P. japonica* displayed insertion patterns consistent with more ancient evolutionary events (Fig. 4).

Based on the base substitution mutation rate of *P. tricornutum* ($4.77 \times 10^{-10}$), mutations per site per generation [65], we inferred the

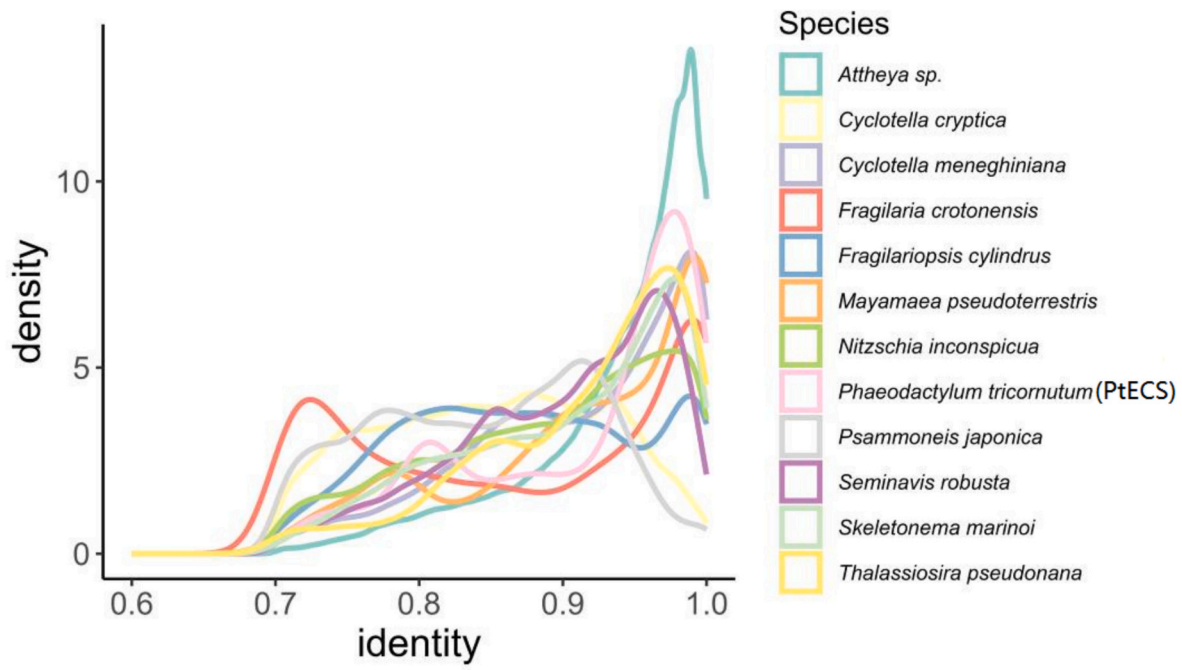


**Fig. 4.** Density plot of sequence identity between LTR-RT of diatoms. The X-axis shows identity, reflecting how genetically similar the LTR sequences are to ancestral ones. A higher identity score signifies closer resemblance to the ancestral sequences, suggesting a more recent divergence. The Y-axis depicts the density, representing the frequency of LTR-RT elements sharing an identity level (indicated by the X-axis value). Peaks on the graph signify widespread transposon replication events across the genome.

insertion time distributions of Copia and Gypsy LTR-RTs across six diatom genomes (Fig. 5). The results revealed pronounced recent expansions with distinct patterns among species: in *C. cryptica* and *T. pseudonana*, Gypsy elements dominated with expansions extending to ~3–4 Myr; *F. crotonensis* and *P. japonica* showed continuous Gypsy accumulation accompanied by localized Copia peaks; while in *N. inconspicua* and *P. tricornutum*, expansions were almost exclusively driven by Copia with more recent insertions (~0.5–1.5 Myr).

### 3.5. Genome comparison of different P. tricornutum strains

To compare genomic features across *P. tricornutum* strains, we analyzed their genome assemblies and structural variations. Relative to the Pt1 reference genome [67], Pt4 harbored the highest number of structural variations (5520), followed by PtSCS (4974) and PtECS (3647). Assembly statistics also differed: Pt4 contained the largest number of contigs (68), whereas PtSCS and PtECS had 30 and 36, respectively. The longest contigs were slightly larger in PtSCS and PtECS (2.81 Mbp) than in Pt4 (2.69 Mbp). PtSCS showed the highest N50 (1.52 Mbp), while PtECS had the lowest (1.3 Mbp). In contrast, GC content was nearly identical across strains (~48.5%), and TE content was broadly similar, though PtSCS exhibited the lowest proportion (15.6%) (Table 2).

**Table 2**
Genome assembly parameters of different *P. tricornutum* strains include average sequencing depth, genome size in megabase pairs (Mbp), number of contigs, the length of the largest contig in Mbp, the sequence length of the shortest contig at 50% of the total assembly length (N50) in Mbp, GC content percentage, TE content, and detected SVs).

| Strain | Pt4 | PtSCS | PtECS |
|---|---|---|---|
| Average depth | 139× | 168× | 99× |
| Data size (Mbp) | 4856.75 | 5362.84 | 3150.76 |
| Genome size (Mbp) | 30.5 | 29.15 | 29.07 |
| Contigs | 68 | 30 | 36 |
| Largest contig length (Mbp) | 2.69 | 2.81 | 2.81 |
| N50 (Mbp) | 1.31 | 1.52 | 1.3 |
| GC content | 48.53% | 48.59% | 48.51% |
| TE content | 17.72% | 15.60% | 16.11% |
| Structural variations (SV) | 5520 | 4974 | 3647 |

### 3.6. Gene family contraction and expansion in different P. tricornutum strains

To investigate gene family dynamics across *P. tricornutum* strains, we conducted a comparative genomics analysis and reconstructed their phylogeny using divergence time estimation. Pt4 diverged earliest (~41.1 Myr), followed by Pt1 (~35.5 Myr), while PtSCS and PtECS separated more recently (~27.5 Myr) (Fig. 6).

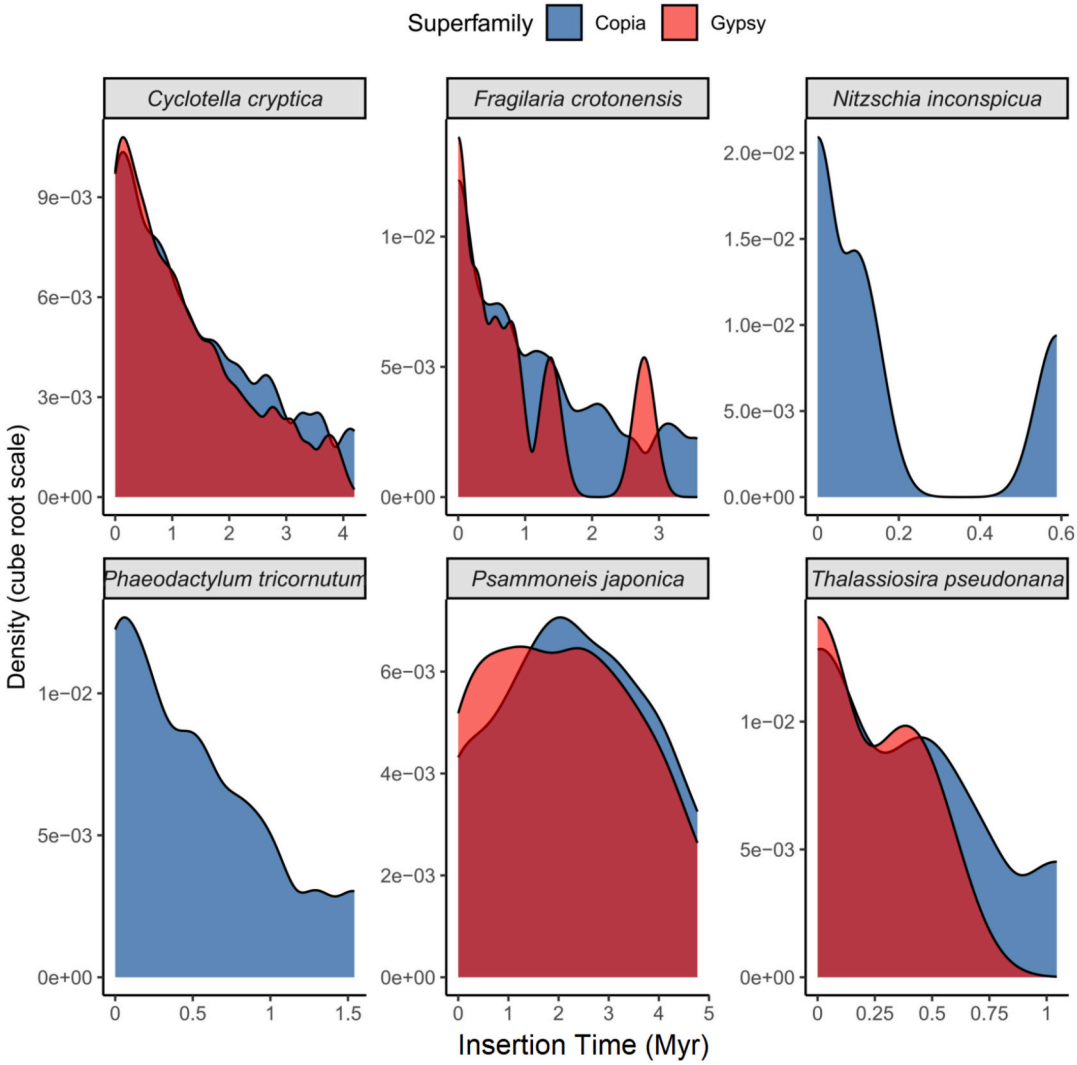


**Fig. 5.** Insertion time distributions of Copia and Gypsy retrotransposons across six diatom genomes based on the mutation rate of *P. tricornutum* [65].

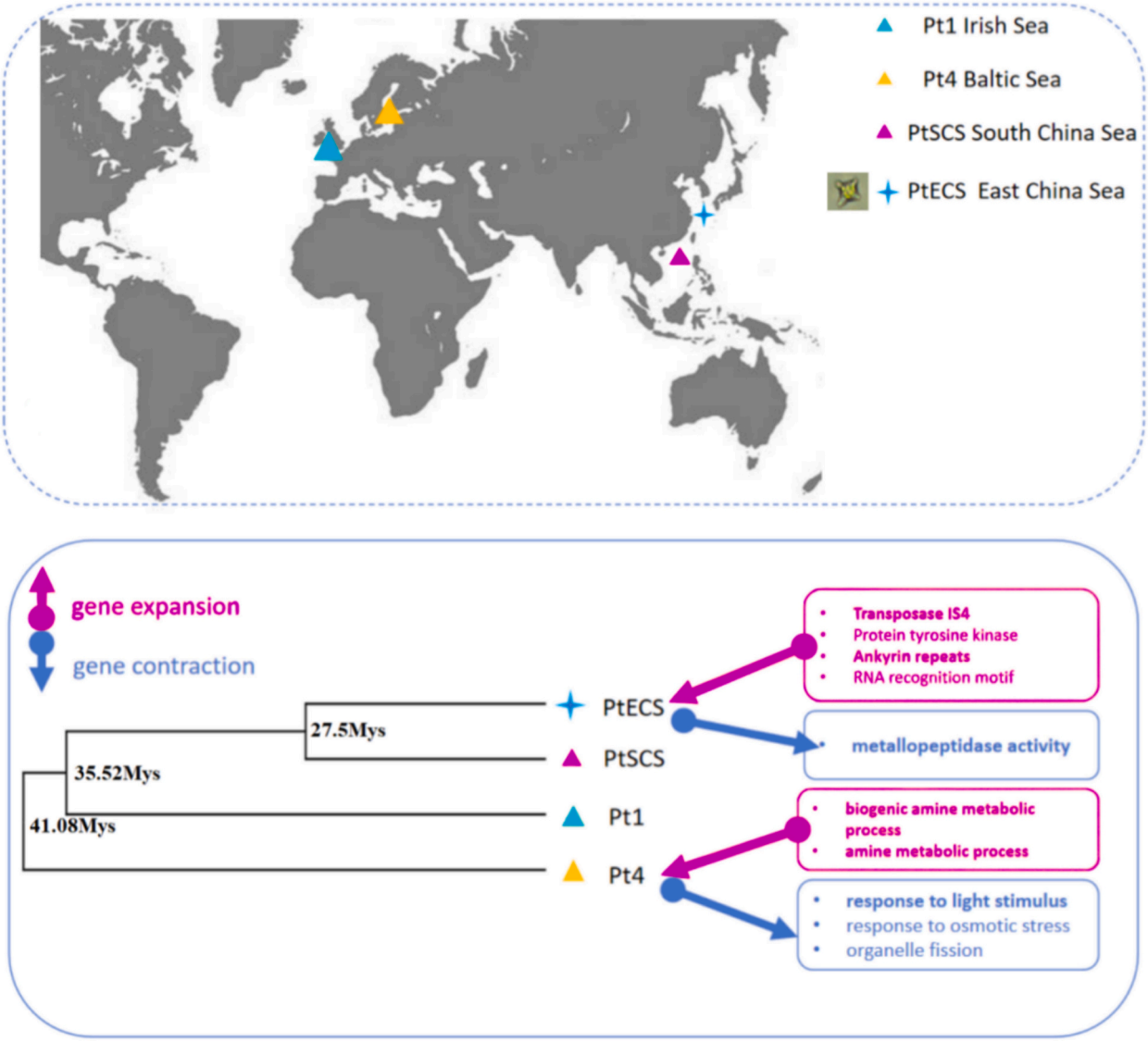


**Fig. 6.** Gene contraction and expansion in different strains *P. tricornutum* isolated from different regions. The upper panel illustrates the distribution of various *P. tricornutum* strains from different locations on the map. Gene functions that have undergone expansion or contraction are indicated in the lower panel with red arrows for expansion and blue arrows for contraction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Although gene family expansion and contraction were assessed across all strains, PtECS and Pt4 were selected for detailed presentation because they represent the most divergent phenotypic and ecological profiles within the species, particularly with respect to cell morphology and light-related responses.

Strain-specific patterns of gene family change were evident. In PtECS, transposase IS4, ankyrin repeats, and protein tyrosine kinase families expanded, whereas metallopeptidases contracted. In Pt4, gene families related to amine metabolism expanded, while light response genes contracted (Fig. 6). Notably, the pyruvate orthophosphate dikinase (PPDK) family, linked to $C_4$ metabolism, expanded in Pt4, whereas other strains showed expansions of genes associated with $C_3$ metabolism (Supplementary Table S4).

### 3.7. TE diversity among different P. tricornutum strains

We analyzed TE composition across *P. tricornutum* strains using EDTA and DeepTE. LTR-RTs were the most abundant TE class. They accounted for 12–12.9% of the genome and represented approximately 75–80% of the total TE content. Among the strains, PtECS showed the highest proportion of LTR-RTs, whereas PtSCS had the lowest.

Given their genomic dominance and the documented activity of LTR/Copia elements in *P. tricornutum* [68], we focused our subsequent polymorphism analyses primarily on LTR-TEs. TIR elements contributed 1.97–3.21% of the genome, representing the second most abundant TE class, and were therefore also included in our TE insertion analysis (Fig. 7C). In contrast, LINEs were rare (<0.5%) and other TE types showed relatively low abundance (Fig. 7A). Overall, total TE content remained relatively stable among strains (15.6–17.7%), but our estimates were markedly higher than the 6.4% previously reported for the reference genome (Fig. 7A).

A total of 364 LTR/Copia sequences were shared across all four strains, while strain-specific sequences numbered 13 in Pt1, 3 in Pt4, 6 in PtSCS, and 5 in PtECS. Pairwise comparisons revealed stronger similarity between Pt1 and Pt4 (25 shared sequences) and between PtSCS and PtECS (20 shared sequences) (Fig. 7B). All strains exhibited genomic signatures of recent TE activity, but their patterns differed. PtSCS and PtECS showed distinct LTR-RT insertion peaks at 80% sequence identity,

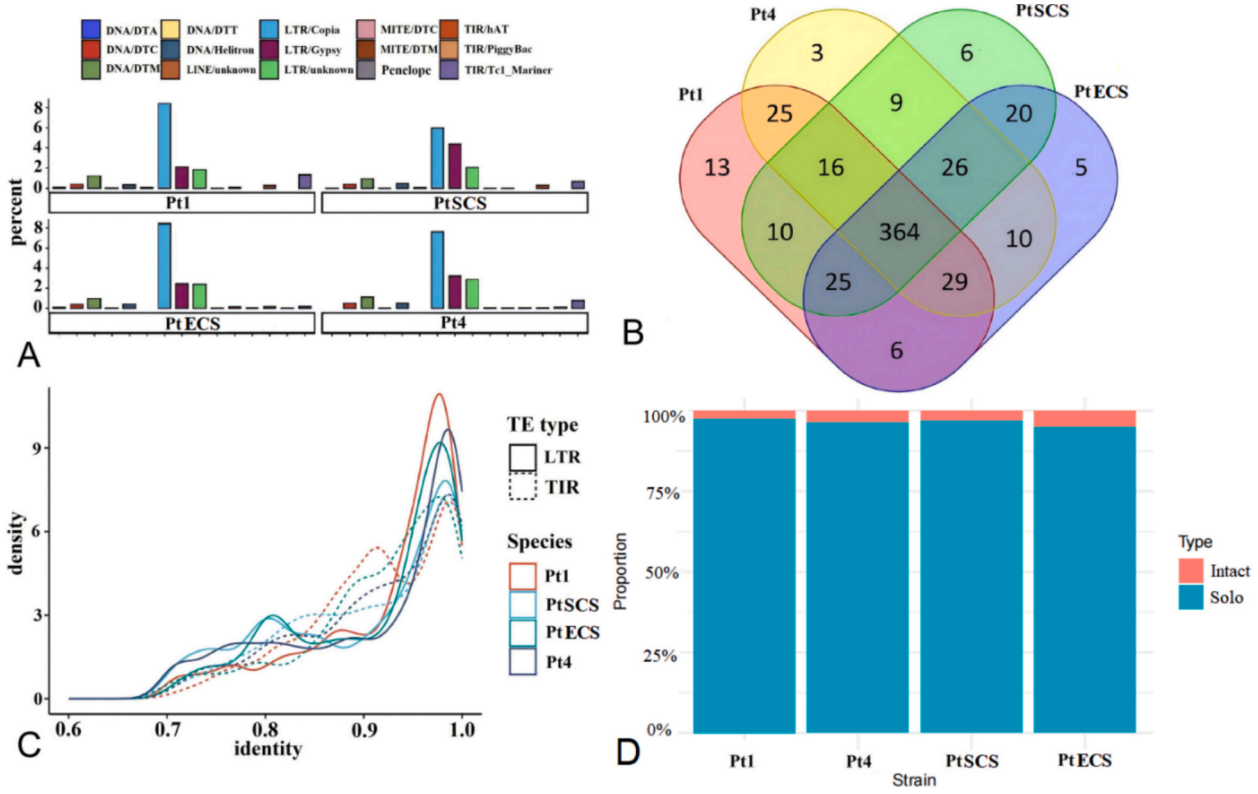


**Fig. 7.** TE diversity and sequence identity across *P. tricornutum* strains. TE proportion of *P. tricornutum* bar graph illustrates the distribution percentages of various transposon types within *P. tricornutum*, with each colour signifying a distinct type of TEs (A). LTR/Copia element variability in different strains chart enumerates both the common and unique LTR/Copia elements found in different *P. tricornutum* strains (B). Density plot of sequence identity delineates the sequence identity between transposons from four *P. tricornutum* strains compared to their original sequences (C). Proportion of intact versus solo LTRs across four *P. tricornutum* strains. The stacked bars show relative abundance of intact LTR-RTs (red) and solo-LTRs (turquoise) in Pt1, Pt4, PtSCS, and PtECS. Solo-LTRs predominate in all strains, indicating extensive LTR decay and recombination-driven fragmentation (D). Colors correspond to different *P. tricornutum* strains or TE categories as indicated in the legend. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

whereas Pt1 displayed a pronounced TIR expansion at 90% identity, suggesting a recent large-scale insertion event (Fig. 7C). Phylogenetic analysis of *P. tricornutum* LTR/Copia elements (Fig. 8) confirmed the seven CoDi groups (CoDi1–CoDi7) described by [68] and identified 30 unclassified elements that likely represent novel retrotransposon families. These novel elements showed strain-specific distributions: 5 in Pt1, 7 in Pt4, and 9 each in the Chinese strains PtSCS and PtECS.

Among all strains, the intact LTR-RTs (the incomplete form) constituted a minor fraction of the total LTR-RT content, with the remainder being predominantly solo LTRs (the complete form). The Intact LTR content of the PtECS strain was the highest, and it was also the strain with the latest stage of differentiation (Fig. 7D).

## 4. Discussion

### *4.1. Diatom evolutionary events and the potential association between contraction/expansion of gene families and morpho-physiological traits*

Diatoms and coccolithophores have evolved specialized adaptations through gene family expansions that underpin their ecological success (Fig. 9). Expansions of fucoxanthin–chlorophyll protein (FCP) genes enhance light harvesting in the blue–green spectrum (450–570 nm) characteristic of marine environments [69]. Duplications of carbonic anhydrase and RubisCO genes support $CO_2$-concentrating mechanisms (CCMs), facilitating efficient carbon fixation under low-$CO_2$ conditions [70,71]. Expansions of heat shock protein (HSP) families further contribute to resilience under thermal stress, a key trait in dynamic marine habitats [72,73].

Diatoms exhibited notable gene family expansions in polyamine metabolism, essential for silica frustule formation—a defining adaptive trait with major ecological significance [74]. They also expanded glutathione metabolism genes, particularly glutathione S-transferases (GSTs), which are central to oxidative stress defense, detoxification, and cellular signaling [75]. This expansion likely conferred resilience against diverse stressors, including UV radiation, nutrient limitation, and fluctuations in salinity and temperature.

As a sulfur-containing compound, glutathione is closely linked to sulfur metabolism through DMSP and DHPS pathways [76,77]. Both metabolites play pivotal roles in diatom–bacteria interactions, serving as major sources of carbon and sulfur for marine bacteria. Our genomic analyses revealed that diatoms have increased copy numbers of genes for DHPS synthesis, but not of those involved in DMSP synthesis (Supplementary Fig. S1, Fig. S2, Table S3). This supports the emerging view that diatoms are key producers of DHPS in the ocean [78].

Our whole-genome analysis confirms that centric diatoms diversified prior to pennate diatoms, consistent with prior SSU rRNA studies [79,80] (Fig. 1). Our study revealed expanded gene families related to microtubules and cellular protrusions in centric diatoms, likely enhancing their buoyancy (Fig. 9). In contrast, pennate diatoms exhibited bilateral symmetry and glide motility driven by actin-myosin-mediated raphe movement [79,81,82]. In pennate diatoms, genes related to actin, myosin, and cytoskeleton showed expansion, further underpinning their motility (Fig. 9). The expansion of distinct gene families in centric and pennate diatoms could reflect their adaptation to different ecological niches.

*N. inconspicua* showed significant gene expansion, without recent TE bursts or WGD. This expansion shows strong functional enrichment in core cellular processes (Figs. S3, S4), suggesting adaptive retention

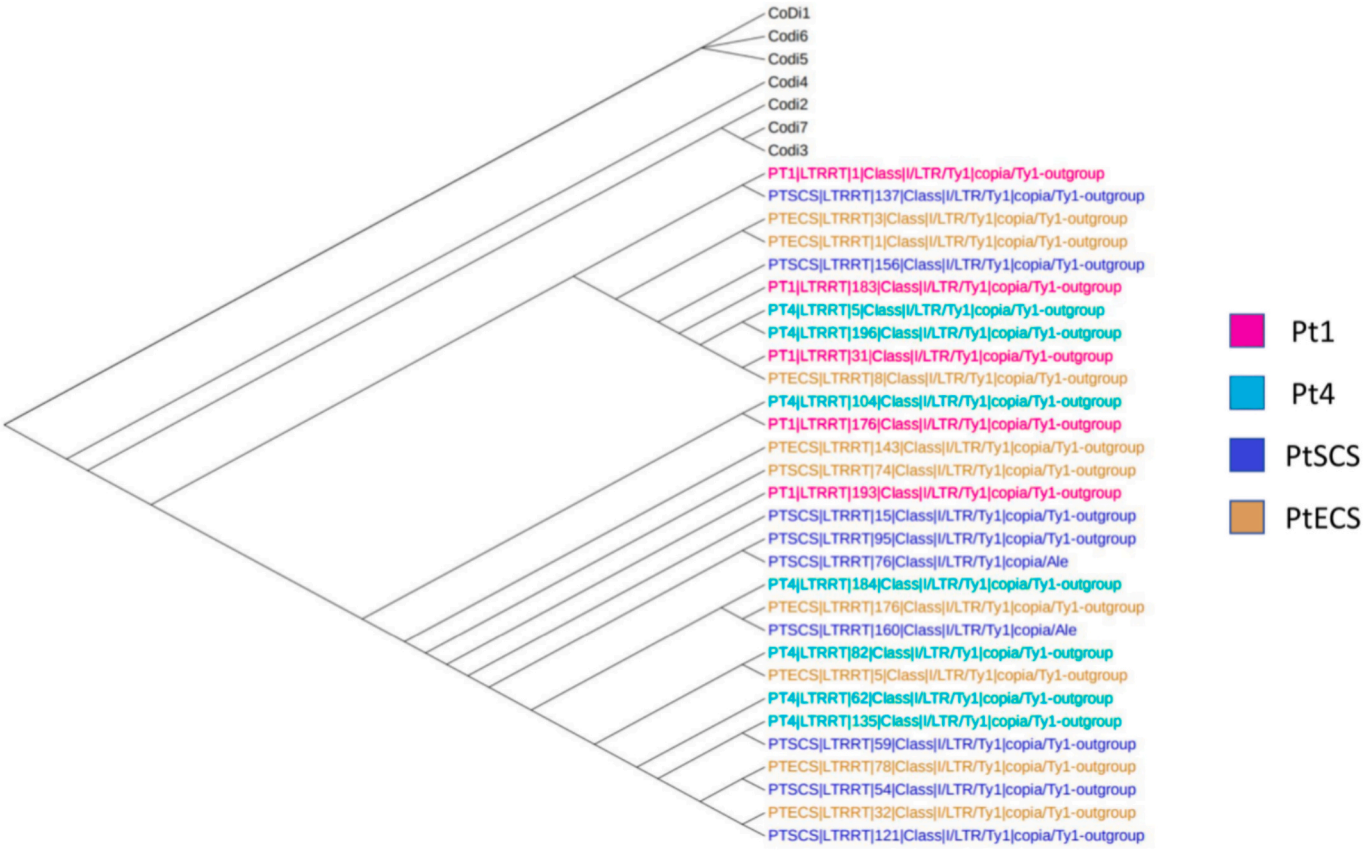


**Fig. 8.** Phylogenetic tree of RT domains from LTR/Copia transposons in *P. tricornutum*, including known typical CoDi groups (CoDi1-CoDi7) and newly discovered LTR/Copia sequences from four strains. Colors represent different strains: red (Pt1), green (Pt4), blue (PtSCS), and brown (PtECS). The phylogenetic tree illustrates the evolutionary relationships between these newly discovered LTR/Copia sequences and known CoDi groups. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

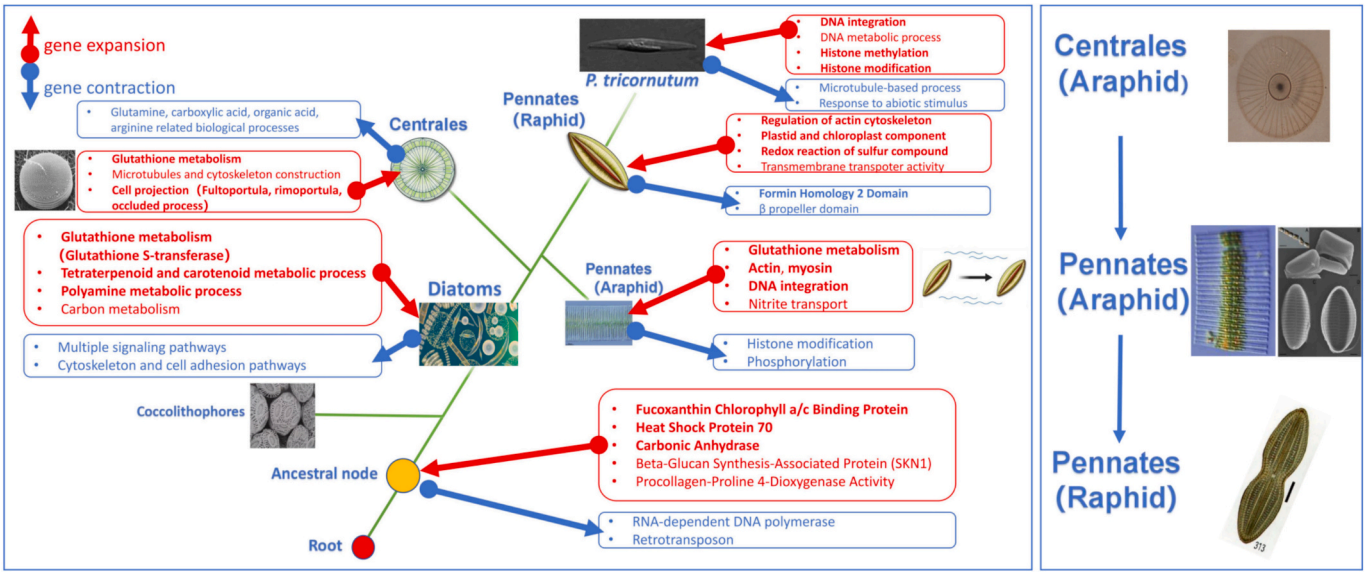


**Fig. 9.** Schematic of gene function contraction and expansion in diatoms based on the whole-genome comparative analysis. In the left panel, red arrows symbolize genes that have expanded, while blue arrows represent genes that have contracted. Green lines denote evolutionary relationships. The descriptions inside the boxes provide a general summary of the gene functions that have contracted or expanded at the current node compared to its ancestral node. Some image materials are from Flickr (https://www.flickr.com/photos/tags/flickr/). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

linked to its ecological niche. Future studies with chromosome-level assemblies and comparative analyses of closely related species would help clarify the mechanistic basis of this expansion.

### 4.2. The impact of TEs on diatom evolution: evidence of TE diversity and insertion timing in diatom genomes

The TE content in the genomes of pennate diatoms has remained stable throughout evolution (below 30%), whereas centric diatoms exhibit greater variability, ranging from 5% to 50% (Fig. 2). The variation in TE proportions in centric diatoms partially reflects the variation in TE activity within their genomes, which may have different impacts on the stability of their genome structures. Our results demonstrate a relatively high proportion of LTR retrotransposons (LTR-RTs) in most diatom genomes, consistent with previous studies [83,84]. This conserved genomic feature mirrors the prevalence of LTR-RTs across diverse eukaryotic organisms, where these mobile elements are recognized as major drivers of structural and functional genome evolution [85]. Interestingly, the araphid pennate diatom *F. crotonensis* displayed a particularly remarkable LTR-RTs dominance, accounting for over 90% of its total TE content. This suggests that LTRs contribute significantly to genome architecture evolution in *F. crotonensis* (Fig. 2).

The raphe-lacking pennate araphid diatoms represent transitional forms in diatom evolution. Comparative genomic analysis showed that the pennate araphid diatoms *P. japonica* and *F. crotonensis*, diverged approximately 160.87 and 97 million years ago, respectively (Fig. 1). Our results suggest that TEs, particularly LTR-RTs, likely facilitated the evolutionary shift from centric to araphid pennate diatoms, as demonstrated by LTR diversity and unique LTR insertion timings. Firstly, the higher LTR- RTs diversity in *P. japonica* and *F. crotonensis* compared to other diatom species suggests increased potential for TE-mediated genomic plasticity [86] (Fig. 3). Secondly, LTR insertion time analysis revealed two distinct patterns of diatom genome reorganization: while most diatom genomes showed recent LTR insertion peaks, *P. japonica* and *F. crotonensis* displayed a unique pattern with a peak in a more ancient period (Fig. 4). This suggests that during early divergence, the genomes of these two araphid pennate diatoms underwent a major LTR transposon insertion event.

Our LTR/Copia and LTR/Gypsy insertion time analysis revealed that the vast majority of detectable LTR insertions across diatom genomes occurred within the last 2 million years, with only *P. japonica* retaining a substantial signature of ancient (3–5 Myr) insertions (Fig. 5). The predominance of recent insertions in most species indicates not a lack of ancient TE copies, but the progressive evolutionary elimination of old TE copies through illegitimate recombination, genome rearrangement, and purifying selection [87]. The retention of ancient LTRs in araphid pennates (*P. japonica* and *F. crotonensis*), potentially due to reduced deletion, selective pressure, or past massive proliferation, underscores the important role of historical TE activity in their early evolution (Fig. 4).

### 4.3. The relationships of WGD, gene family dynamic, and TE insertion, and the hidden genomic mysteries

Parks *et al.*, [33] identified two deep WGD events occurring at ~200 Myr and ~ 170 Myr, which established an ancient polyploid ancestry for most extant diatoms. Our estimates of major lineage divergence times in diatoms (~202 Myr and ~ 173 Myr) are highly consistent with these timings (Fig. 10) and further reinforce the robustness of these events as major genomic landmarks. These two periods likely coincide with critical phases of early diatom diversification, when polyploidy may have facilitated genomic innovation, ecological expansion, and the emergence of lineage-specific traits.

Building on previous findings that WGDs precede large-scale gene family reorganization [88,89], we propose that large-scale gene family expansions in diatoms may reflect WGD-driven retention patterns, potentially linked to ecological adaptation and the diversification of diatoms. Our analysis added a functional dimension to this temporal framework by identifying the specific gene families that were significantly expanded or contracted, directly illustrating the lineage-specific genomic restructuring.

The temporal relationship between WGD and TE expansion has long been debated, but it is generally accepted that TE proliferation occurs after WGD. The "genome shock" hypothesis proposed that WGD triggers immediate bursts of TE activity [29]. However, increasing evidence suggests that large-scale TE accumulation is typically not an abrupt burst but rather a gradual process. Empirical studies in *Atlantic salmon* further showed that some TE families became more active after WGD and contributed to cis-regulatory evolution [31]. TE proliferation has also been linked to rapid genome size increases, either independently or jointly with WGD, as reported in maize [32] and *Corydoradinae catfishes* [90].

Our LTR insertion time analysis revealed a striking temporal pattern. The vast majority of detectable LTR/Copia and LTR/Gypsy insertions across diatom genomes occurred within the last 0.5–5 million years (Fig. 10). This period substantially postdates the ancient whole-genome duplication (WGD) events (~200–170 Myr). It also postdates the major gene family turnover associated with these polyploidy events (Fig. 10).

This apparent temporal gap between WGD (~200–170 Myr) and observable TE proliferation (<5 Myr) is striking and requires careful interpretation. Critical methodological limitations compound these interpretive challenges. Our TE insertion time estimates assume a constant substitution rate from *P. tricornutum* [65], yet TE mutation rates vary substantially among taxa [91]. Moreover, our results readily detected recent TE insertions (<5 Myr) in modern diatom genomes, any hypothetical TE bursts coinciding with the ancient WGDs (100–200 Myr) likely have been largely erased by the pervasive loss of TE copies over time, thereby leaving minimal genomic signatures [87]. The

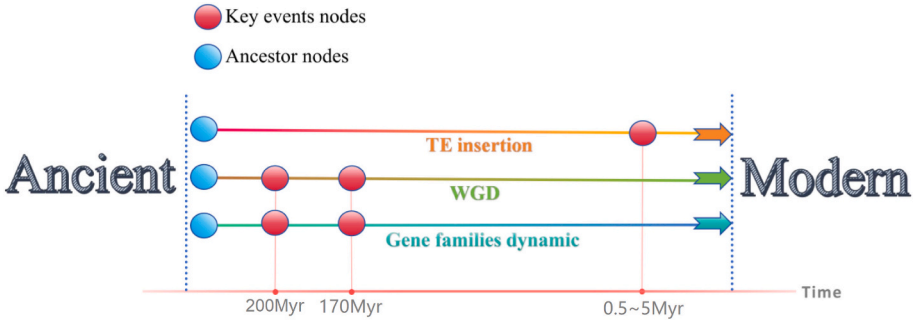


**Fig. 10.** Temporal relationship between WGD, Gene family dynamics, and TE Insertion in diatoms. Two ancient WGD events (~200 Myr and ~ 170 Myr) coincide with divergence of major ancestral nodes and likely drove large-scale gene family expansions and contractions. TE insertions (0.5–5 Myr) occurred later than WGD and gene family dynamics, consistent with TE proliferation lagging behind WGD. The timeline is based on mutation rates [65] and reflects relative, not absolute, timing. Red nodes mark key genomic events; blue nodes indicate ancestral divergence points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

mutation rate uncertainty, combined with the progressive deletion of ancient TE copies, make it difficult to determine when the ancient TE insertions occurred.

Despite these limitations, our data provide the evidence in diatoms for distinct temporal scales of genome evolution following WGD: gene family reorganization likely occurred relatively rapidly, leaving stable phylogenetic signatures, whereas TE-mediated genome dynamics represent ongoing processes whose ancient history has likely been progressively obscured (Fig. 10). Future studies incorporating improved TE age dating methods and broader phylogenetic sampling will be essential to refine these temporal relationships and test the generality of these patterns across diatom lineages.

### 4.4. Differences in functional genes and TEs among P. tricornutum strains

Gene families associated with DNA integration and histone modification have expanded in *P. tricornutum*, suggesting that DNA transposition and epigenetic regulation potentially play important roles in its environmental adaptability (Fig. 9). Our phylogenetic analysis identifies Pt4 and PtECS as the earliest-diverged and most recently derived strains of *P. tricornutum*, respectively (Fig. 6). PtECS has a unique cruciform structure that changes with temperature [92]. PtECS strain's distinct cruciform shape appears associated with the expansion and contraction of specific gene families (Fig. 6). In PtECS, the expansion of gene families associated with protein tyrosine kinase and ankyrin repeats could aid cell signaling and differentiation, and maintain cellular integrity [93,94] while the contraction of metallopeptidase-related gene families can influence the extracellular matrix and cell structure [95]. This association underscores the significance of gene family dynamics in shaping strain-specific morphologies.

The Pt4 strain, isolated from the low-salinity Baltic Sea, is adapted to low light and exhibits reduced non-photochemical quenching [12]. Consistent with its niche, we observed contraction in gene families involved in osmotic stress, light response, and organelle fission. Notably, while short-read sequencing revealed an increased copy number of the nitrate assimilation gene Phatr3_EG02286 in Pt4 [12]. Our long-read-based analysis further identified expansions in gene families related to amine metabolism (Fig. 6). These results indicate potential differences in nitrogen metabolism between Pt4 and other strains, warranting further investigation. Moreover, Pt4 exhibited the most significant reduction in dissolved inorganic carbon affinity under high $CO_2$ (1000 ppm) among *P. tricorntum* strains, suggesting a unique carbon metabolism strategy [96]. Pt4 exhibited an expansion of C4-related PPDK genes, whereas other strains had more C3-related genes (Supplementary Table S4), further supporting divergent carbon assimilation pathways across strains.

Our results demonstrate that TEs potentially drive divergence and adaptation in *P. tricornutum* strains. The widespread presence of LTR/Copia elements across all *P. tricornutum* strains suggests their key role in genome evolution (Fig. 7A). Notably, PtECS—identified as the most recently evolved strain (Fig. 6) — exhibits the highest LTR-RT content and the lowest TIR element abundance, consistent with the hypothesis that recent transposable element expansions drive genomic adaptation to environmental changes [97]. Strains from China (PtSCS and PtECS) share a higher number of common LTR/Copia elements and similar LTR insertion peaks during earlier periods, underscoring the role of LTR-RTs in *P. tricornutum* regional strain differentiation (Fig. 7B, C).

Our study, utilizing long-read sequencing, significantly advances the characterization of TEs in *P. tricornutum*. While [68,83] first identified CoDi elements constituting ~5.4% of the genome, subsequent work by [98] revealed a higher proportion (~7.6%). Our study extends these findings by discovering previously unclassified sequences that form a potentially novel CoDi branch (Fig. 8). Notably, the Chinese strains PtSCS and PtECS each contained nine identical novel sequences - a conserved pattern absents in other geographical variants. This distinct regional signature in transposon profiles suggests LTR/Copia retrotransposons contribute to geographical differentiation, potentially reflecting adaptive evolution to unique environmental pressures in Chinese coastal waters. The relatively higher retention of intact LTRs in PtECS may indicate either more recent transposition activity or reduced recombination rates in this lineage, preserving potentially autonomous elements that could retain transposition capacity (Fig. 7D).

Our molecular dating suggests an unusually deep divergence (~41 Myr) among *P. tricornutum* strains. While the absolute date requires caution, it is consistent with the known existence of four deeply diverged, globally distributed clades that exhibit significant genomic, morphological (oval, fusiform, and triradiate and cruciform), and physiological differentiation, indicating possible cryptic speciation [9,92].

### 4.5. Limitations and perspectives

Although our study integrates whole-genome duplication, gene family dynamics, and TE activity within a unified comparative framework, several limitations should be acknowledged when interpreting their relationships. While the inferred diatom WGD events date back over 170 million years, most detectable TE insertions reflect much more recent activity (<5 Myr). We speculate that ancient TE insertions potentially associated with early WGD events may have been progressively erased through sequence decay and deletion, leaving limited detectable signatures in extant genomes. In addition, TE insertion times were estimated using a constant substitution rate derived from *P. tricornutum*, although mutation rates are known to vary across taxa. Together, these factors constrain our ability to reconstruct the deep evolutionary history of TE activity in diatoms.

Future work will benefit from several key improvements. First, broader taxon sampling across additional diatom lineages will be necessary to assess the generality of the observed patterns beyond the species examined here. Second, more refined approaches for estimating TE insertion times, particularly for older events, are needed to reduce uncertainties associated with constant-rate assumptions. Finally, integrating complementary lines of evidence beyond temporal comparisons will be required to more robustly evaluate potential causal relationships among WGD, gene family dynamics, and TE activity.

## 5. Conclusion

Through long-read sequencing and comparative genomics, this study reveals the crucial roles of TEs and gene family dynamics in diatom evolutionary success. Comparative genomic analyses showed that diatom-specific expansions in polyamine metabolism and GST gene families underpin adaptive traits such as silica wall formation and oxidative stress resistance in diatoms. Divergent gene family profiles between centric and pennate diatoms correlate with their distinct morphologies and motility strategies, while TE insertion timing suggests TEs played important roles in the evolutionary transition to raphid pennate diatoms.

Strain-level analyses of *P. tricornutum* reveal geographic adaptation signatures, with morphology-related gene expansions in East China Sea isolates and metabolic gene contractions in Baltic Sea isolates. TE diversity correlates with geographic distribution, demonstrating how TE-mediated genome plasticity drives local adaptation.

Our phylogenomic analyses reveal temporal consistency between major lineage divergences (~202 and ~ 173 Myr) and ancient whole-genome duplications (~200 and ~ 170 Myr). Gene family analyses reveal lineage-specific expansions and contractions at these divergence nodes. TE expansions show more recent timescales, reflecting progressive deletion of ancient copies while ongoing accumulation generates the recent insertion peaks observed across diatom lineages.

## Animal and human rights statement

This study did not involve any human participants or animals. Therefore, no ethical approval was required.

## CRediT authorship contribution statement

**Tianze Zheng:** Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation. **Xu Zhang:** Writing – review & editing, Validation, Supervision, Methodology. **Tianren Liu:** Writing – original draft. **Xinzhu Liu:** Visualization. **Chris Bowler:** Writing – review & editing, Supervision. **Xin Lin:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no conflicts of interest related to this work.

## Acknowledgements

This study was supported by the National Natural Science Foundation of China (2576144) and the Natural Science Foundation of Fujian Province of China (2024J01020).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.algal.2026.104605.

## Data availability

Data will be made available on request.

## References

[1] P.G. Falkowski, R.T. Barber, V. Smetacek, Biogeochemical controls and feedbacks on Ocean primary production, Science (New York, N.Y.) 281 (5374) (1998) 200–207, https://doi.org/10.1126/science.281.5374.200.

[2] Angela Falciatore, Marianne Jaubert, Jean-Pierre Bouly, Benjamin Bailleul, Thomas Mock, Diatom molecular research comes of age: model species for studying phytoplankton biology and diversity, Plant Cell 32 (3) (2020) 547–572, https://doi.org/10.1105/tpc.19.00158.

[3] Karolina Bryłka, Matt P. Ashworth, Andrew J. Alverson, Daniel J. Conley, The cretaceous diatom database: a tool for investigating early diatom evolution, J. Phycol. 60 (5) (2024) 1090–1104, https://doi.org/10.1111/jpy.13499.

[4] Wade R. Roberts, Adam M. Siepielski, Andrew J. Alverson, Diatom abundance in the polar oceans is predicted by genome size, PLoS Biol. 22 (8) (2024) e3002733, https://doi.org/10.1371/journal.pbio.3002733.

[5] F.E. Round, R.M. Crawford, D.G. Mann, Diatoms: Biology and Morphology of the Genera, Cambridge University Press, 1990.

[6] Patricia A. Sims, David G. Mann, Linda K. Medlin, Evolution of the diatoms: insights from fossil, biological and molecular data, Phycologia 45 (4) (2006) 361–402, https://doi.org/10.2216/05-22.1.

[7] Stanley A. Cohn, Roy E. Weitzell, Ecological considerations of diatom cell motility. I. Characterization of motility and adhesion in four diatom species, J. Phycol. 32 (6) (1996) 928–939, https://doi.org/10.1111/j.0022-3646.1996.00928.x.

[8] Chris Bowler, Andrew E. Allen, Jonathan H. Badger, Jane Grimwood, Kamel Jabbari, Alan Kuo, Uma Maheswari, Cindy Martens, Florian Maumus, Robert P. Otillar, Edda Rayko, Asaf Salamov, Klaas Vandepoele, Bank Beszteri, Ansgar Gruber, Marc Heijde, Michael Katinka, Thomas Mock, Klaus Valentin, Fréderic Verret, John A. Berges, Colin Brownlee, Jean-Paul Cadoret, Anthony Chiovitti, Chang Jae Choi, Sacha Coesel, Alessandra De Martino, J. Chris Detter, Colleen Durkin, Angela Falciatore, Jérome Fournet, Miyoshi Haruta, Marie J.J. Huysman, Bethany D. Jenkins, Katerina Jiroutova, Richard E. Jorgensen, Yolaine Joubert, Aaron Kaplan, Nils Kröger, Peter G. Kroth, Julie La Roche, Erica Lindquist, Markus Lommer, Véronique Martin-Jézéquel, Pascal J. Lopez, Susan Lucas, Manuela Mangogna, Karen McGinnis, Linda K. Medlin, Anton Montsant, Marie-Pierre Oudot-Le Secq, Carolyn Napoli, Miroslav Obornik, Micaela Schnitzler Parker, Jean-Louis Petit, Betina M. Porcel, Nicole Poulsen, Matthew Robison, Leszek Rychlewski, Tatiana A. Rynearson, Jeremy Schmutz, Harris Shapiro, Magali Siaut, Michele Stanley, Michael R. Sussman, Alison R. Taylor, Assaf Vardi, Peter von Dassow, Wim Vyverman, Anusuya Willis, Lucjan S. Wyrwicz, Daniel S. Rokhsar, Jean Weissenbach, E. Virginia Armbrust, Beverley R. Green, Yves Van de Peer, Igor V. Grigoriev, The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes, Nature 456 (7219) (2008) 239–244, https://doi.org/10.1038/nature07410.

[9] Achal Rastogi, Fabio Rocha Jimenez Vieira, Anne-Flore Deton-Cabanillas, Alaguraj Veluchamy, Catherine Cantrel, Gaohong Wang, Pieter Vanormelingen, Chris Bowler, Gwenael Piganeau, Hu Hanhua, Leila Tirichine, A genomics approach reveals the global genetic polymorphism, structure, and functional diversity of ten accessions of the marine model diatom Phaeodactylum Tricornutum, ISME J. 14 (2) (2020) 347–363, https://doi.org/10.1038/s41396-019-0528-3.

[10] Timothée Chaumier, Feng Yang, Eric Manirakiza, Ouardia Ait-Mohamed, Yue Wu, Udita Chandola, Bruno Jesus, Gwenael Piganeau, Agnès Groisillier, Leila Tirichine, Genome-wide assessment of genetic diversity and transcript variations in 17 accessions of the model diatom *Phaeodactylum tricornutum*, ISME Commun. 4 (1) (2024) ycad008, https://doi.org/10.1093/ismeco/ycad008.

[11] Benjamin Bailleul, Alessandra Rogato, Alessandra de Martino, Sacha Coesel, Pierre Cardol, Chris Bowler, Angela Falciatore, Giovanni Finazzi, An atypical member of the light-harvesting complex stress-related protein family modulates diatom responses to light, Proc. Natl. Acad. Sci. USA 107 (42) (2010) 18214–18219, https://doi.org/10.1073/pnas.1007703107.

[12] Achal Rastogi, Fabio Rocha Jimenez Vieira, Anne-Flore Deton-Cabanillas, Alaguraj Veluchamy, Catherine Cantrel, Gaohong Wang, Pieter Vanormelingen, Chris Bowler, Gwenael Piganeau, Hu Hanhua, Leila Tirichine, A genomics approach reveals the global genetic polymorphism, structure, and functional diversity of ten accessions of the marine model diatom Phaeodactylum Tricornutum, ISME J. 14 (2) (2020) 347–363, https://doi.org/10.1038/s41396-019-0528-3.

[13] Ruiping Huang, Jiancheng Ding, Jiazhen Sun, Yang Tian, Chris Bowler, Xin Lin, Kunshan Gao, Physiological and molecular responses to ocean acidification among strains of a model diatom, Limnol. Oceanogr. 65 (12) (2020) 2926–2936, https://doi.org/10.1002/lno.11565.

[14] J.A.M. Graves, Background and overview of comparative genomics, ILAR J. 39 (2–3) (1998) 48–65, https://doi.org/10.1093/ilar.39.2-3.48.

[15] Konstantina Athanasopoulou, Michaela A. Boti, Panagiotis G. Adamopoulos, Paraskevi C. Skourou, Andreas Scorilas, Third-generation sequencing: the spearhead towards the radical transformation of modern genomics, Life (Basel, Switzerland) 12 (1) (2021) 30, https://doi.org/10.3390/life12010030.

[16] Lun Wang, Fa He, Yue Huang, Jiaxian He, Shuizhi Yang, Jiwu Zeng, Chongling Deng, Xiaolin Jiang, Yiwen Fang, Shaohua Wen, Rangwei Xu, Huiwen Yu, Xiaoming Yang, Guangyan Zhong, Chuanwu Chen, Xiang Yan, Changfu Zhou, Hongyan Zhang, Zongzhou Xie, Robert M. Larkin, Xiuxin Deng, Qiang Xu, Genome of wild mandarin and domestication history of mandarin, Mol. Plant 11 (8) (2018) 1024–1037, https://doi.org/10.1016/j.molp.2018.06.001.

[17] Nanqiao Liao, Zhongyuan Hu, Jinshan Miao, Xiaodi Hu, Xiaolong Lyu, Haitian Fang, Yi-Mei Zhou, Ahmed Mahmoud, Guancong Deng, Yi-Qing Meng, Kejia Zhang, Yu-Yuan Ma, Yuelin Xia, Meng Zhao, Haiyang Yang, Yong Zhao, Ling Kang, Yiming Wang, Jing-Hua Yang, Yan-Hong Zhou, Ming-Fang Zhang, Jing-Quan Yu, Chromosome-level genome assembly of bunching onion illuminates genome evolution and flavor formation in *Allium* crops, Nat. Commun. 13 (1) (2022) 6690, https://doi.org/10.1038/s41467-022-34491-3.

[18] Xilong Chen, Shiming Li, Dong Zhang, Mingyu Han, Xin Jin, Caipin Zhao, Songbo Wang, Libo Xing, Juanjuan Ma, Jingjing Ji, Na An, Sequencing of a wild apple (*Malus baccata*) genome unravels the differences between cultivated and wild apple species regarding disease resistance and cold tolerance, G3 Genes|Genomes|Genetics 9 (7) (2019) 2051–2060, https://doi.org/10.1534/g3.119.400245.

[19] Wade R. Roberts, Elizabeth C. Ruck, Kala M. Downey, Eveline Pinseel, Andrew J. Alverson, Resolving marine-freshwater transitions by diatoms through a fog of gene tree discordance, Syst. Biol. (2023) syad038, https://doi.org/10.1093/sysbio/syad038.

[20] Romain Koszul, Gilles Fischer, A prominent role for segmental duplications in modeling eukaryotic genomes, C. R. Biol. 332 (2–3) (2009) 254–266, https://doi.org/10.1016/j.crvi.2008.07.005.

[21] Michael Lynch, Allan G. Force, The origin of interspecific genomic incompatibility via gene duplication, Am. Nat. 156 (6) (2000) 590–605, https://doi.org/10.1086/316992.

[22] Steven Lockton, Brandon S. Gaut, Plant conserved non-coding sequences and paralogue evolution, Trends Genet. 21 (1) (2005) 60–65, https://doi.org/10.1016/j.tig.2004.11.013.

[23] Clément Gilbert, Jean Peccoud, Richard Cordaux, Transposable elements and the evolution of insects, Annu. Rev. Entomol. 66 (2021) 355–372, https://doi.org/10.1146/annurev-ento-070720-074650.

[24] Maria Anisimova (Ed.), *Evolutionary Genomics: Statistical and Computational Methods*. Vol. 1910. Methods in Molecular Biology, Springer New York, New York, NY, 2019.

[25] Peter E. Warburton, Robert P. Sebra, Long-read DNA sequencing: recent advances and remaining challenges, Annu. Rev. Genomics Hum. Genet. 24 (2023) 109–132, https://doi.org/10.1146/annurev-genom-101722-103045.

[26] Aurélie Kapusta, Alexander Suh, Evolution of bird genomes-a transposon's-eye view, Ann. N. Y. Acad. Sci. 1389 (1) (2017) 164–185, https://doi.org/10.1111/nyas.13295.

[27] Christoph Stritt, Michele Wyler, Elena L. Gimmi, Martin Pippel, Anne C. Roulin, Diversity, dynamics and effects of long terminal repeat retrotransposons in the model grass Brachypodium Distachyon, New Phytol. 227 (6) (2020) 1736–1748, https://doi.org/10.1111/nph.16308.

[28] Liyi Zhang, Jiang Hu, Xiaolei Han, Jingjing Li, Yuan Gao, Christopher M. Richards, Caixia Zhang, Yi Tian, Guiming Liu, Hera Gul, Dajiang Wang, Yu Tian, Chuanxin Yang, Minghui Meng, Gaopeng Yuan, Guodong Kang, Yonglong Wu,

Kun Wang, Hengtao Zhang, Depeng Wang, Peihua Cong, A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour, Nat. Commun. 10 (1) (2019) 1494, https://doi.org/10.1038/s41467-019-09518-x.
[29] B. McClintock, The significance of responses of the genome to challenge, Science 226 (4676) (1984) 792–801, https://doi.org/10.1126/science.15739260 (New York, N.Y.).
[30] Pierre Baduel, Leandro Quadrana, Ben Hunter, Kirsten Bomblies, Vincent Colot, Relaxed purifying selection in Autopolyploids drives transposable element over-accumulation which provides variants for local adaptation, Nat. Commun. 10 (1) (2019) 5818, https://doi.org/10.1038/s41467-019-13730-0.
[31] Øystein Monsen, Lars Grønvold, Alex Datsomor, Thomas Harvey, James Kijas, Alexander Suh, Torgeir R. Hvidsten, Simen Rød Sandve, The role of transposon activity in shaping Cis-regulatory element evolution after whole-genome duplication, Genome Res. 35 (3) (2025) 475–488, https://doi.org/10.1101/gr.278931.124.
[32] Vicki L. Chandler, Volker Brendel, The maize genome sequencing project, Plant Physiol. 130 (4) (2002) 1594–1597, https://doi.org/10.1104/pp.015594.
[33] Matthew B. Parks, Teofil Nakov, Elizabeth C. Ruck, Norman J. Wickett, Andrew J. Alverson, Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta), Am. J. Bot. 105 (3) (2018) 330–347, https://doi.org/10.1002/ajb2.1056.
[34] Ruiping Huang, Jiancheng Ding, Jiazhen Sun, Yang Tian, Chris Bowler, Xin Lin, Kunshan Gao, Physiological and molecular responses to ocean acidification among strains of a model diatom, Limnol. Oceanogr. 65 (12) (2020) 2926–2936, https://doi.org/10.1002/lno.11565.
[35] Gina V. Filloramo, Bruce A. Curtis, Emma Blanche, John M. Archibald, Re-examination of two diatom reference genomes using long-read sequencing, BMC Genomics 22 (1) (2021) 379, https://doi.org/10.1186/s12864-021-07666-3.
[36] David Deamer, Mark Akeson, Daniel Branton, Three decades of Nanopore sequencing, Nat. Biotechnol. 34 (5) (2016) 518–524, https://doi.org/10.1038/nbt.3423.
[37] Tomáš Brůna, Katharina J. Hoff, Alexandre Lomsadze, Mario Stanke, Mark Borodovsky, BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database, NAR Genomics Bioinform. 3 (1) (2021) lqaa108, https://doi.org/10.1093/nargab/lqaa108.
[38] Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, Peer Bork, Fast genome-wide functional annotation through Orthology assignment by eggNOG-mapper, Mol. Biol. Evol. 34 (8) (2017) 2115–2122, https://doi.org/10.1093/molbev/msx148.
[39] Shigekatsu Suzuki, Shuhei Ota, Takahiro Yamagishi, Akihiro Tuji, Haruyo Yamaguchi, Masanobu Kawachi, Rapid transcriptomic and physiological changes in the freshwater pennate diatom *Mayamaea pseudoterrestris* in response to copper exposure, DNA Res. 29 (6) (2022) dsac037, https://doi.org/10.1093/dnares/dsac037.
[40] Cristina Maria Osuna-Cruz, Gust Bilcke, Emmelien Vancaester, Sam De Decker, Atle M. Bones, Per Winge, Nicole Poulsen, Petra Bulankova, Bram Verhelst, Sien Audoor, Darja Belisova, Aikaterini Pargana, Monia Russo, Frederike Stock, Emilio Cirri, Tore Brembu, Georg Pohnert, Gwenael Piganeau, Maria Immacolata Ferrante, Thomas Mock, Lieven Sterck, Koen Sabbe, Lieven De Veylder, Wim Vyverman, Klaas Vandepoele, The Seminavis Robusta genome provides insights into the evolutionary adaptations of benthic diatoms, Nat. Commun. 11 (1) (2020) 3320, https://doi.org/10.1038/s41467-020-17191-8.
[41] Pirita Paajanen, Jan Strauss, Cock van Oosterhout, Mark McMullan, Matthew D. Clark, Thomas Mock, Building a locally diploid genome and transcriptome of the diatom Fragilariopsis Cylindrus, Sci. Data 4 (2017) 170149, https://doi.org/10.1038/sdata.2017.149.
[42] Aaron Oliver, Sheila Podell, Agnieszka Pinowska, Jesse C. Traller, Sarah R. Smith, Ryan McClure, Alex Beliaev, Pavlo Bohutskyi, Eric A. Hill, Ariel Rabines, Hong Zheng, Lisa Zeigler Allen, Alan Kuo, Igor V. Grigoriev, Andrew E. Allen, David Hazlebeck, Eric E. Allen, Diploid genomic architecture of Nitzschia Inconspicua, an elite biomass production diatom, Sci. Rep. 11 (1) (2021) 15592, https://doi.org/10.1038/s41598-021-95106-3.
[43] Brittany N. Zepernick, David J. Niknejad, Gwendolyn F. Stark, Alexander R. Truchon, Robbie M. Martin, Karen L. Rossignol, Hans W. Paerl, Steven W. Wilhelm, Morphological, physiological, and transcriptional responses of the freshwater diatom Fragilaria Crotonensis to elevated pH conditions, Front. Microbiol. 13 (2022) 1044464, https://doi.org/10.3389/fmicb.2022.1044464.
[44] Kayoko Yamamoto, Takashi Hamaji, Hiroko Kawai-Toyooka, Ryo Matsuzaki, Fumio Takahashi, Yoshiki Nishimura, Masanobu Kawachi, Hideki Noguchi, Yohei Minakuchi, James G. Umen, Atsushi Toyoda, Hisayoshi Nozaki, Three genomes in the algal genus Volvox reveal the fate of a haploid sex-determining region after a transition to Homothallism, Proc. Natl. Acad. Sci. USA 118 (21) (2021) e2100712118, https://doi.org/10.1073/pnas.2100712118.
[45] Natalia Calixto Mancipe, Evelyn M. McLaughlin, Brett M. Barney, Genomic analysis and characterization of *ScenedesmusGlucoliberatum* PABB004: an unconventional sugar-secreting green alga, J. Appl. Microbiol. 132 (3) (2022) 2004–2019, https://doi.org/10.1111/jam.15311.
[46] Michela Cecchin, Luca Marcolungo, Marzia Rossato, Laura Girolomoni, Emanuela Cosentino, Stephan Cuine, Yonghua Li-Beisson, Massimo Delledonne, Matteo Ballottari, Chlorella vulgaris genome assembly and annotation reveals the molecular basis for metabolic acclimation to high light conditions, Plant J. 100 (6) (2019) 1289–1305, https://doi.org/10.1111/tpj.14508.
[47] JunMo Lee, Dongseok Kim, Debashish Bhattacharya, Hwan Su Yoon, Expansion of Phycobilisome linker gene families in mesophilic red algae, Nat. Commun. 10 (1) (2019) 4823, https://doi.org/10.1038/s41467-019-12779-1.
[48] JunMo Lee, Eun Chan Yang, Louis Graf, Ji Hyun Yang, Huan Qiu, Udi Zelzion, Cheong Xin Chan, Timothy G. Stephens, Andreas P.M. Weber, Ga Hun Boo, Sung Min Boo, Kyeong Mi Kim, Younhee Shin, Myunghee Jung, Seung Jae Lee, Hyung-Soon Yim, Jung-Hyun Lee, Debashish Bhattacharya, Hwan Su Yoon, Analysis of the draft genome of the red seaweed Gracilariopsis Chorda provides insights into genome size evolution in Rhodophyta, Mol. Biol. Evol. 35 (8) (2018) 1869–1886, https://doi.org/10.1093/molbev/msy081.
[49] Yuki Hongo, Kei Kimura, Yoshihiro Takaki, Yukari Yoshida, Shuichiro Baba, Genta Kobayashi, Keizo Nagasaki, Takeshi Hano, Yuji Tomaru, The genome of the diatom Chaetoceros Tenuissimus carries an ancient integrated fragment of an extant virus, Sci. Rep. 11 (2021) 22877, https://doi.org/10.1038/s41598-021-00565-3.
[50] Wade R. Roberts, Kala M. Downey, Elizabeth C. Ruck, Jesse C. Traller, Andrew J. Alverson, Improved reference genome for *Cyclotella cryptica* CCMP332, a model for cell wall morphogenesis, salinity adaptation, and lipid production in diatoms (Bacillariophyta), G3 (Bethesda, Md.) 10 (9) (2020) 2965–2974, https://doi.org/10.1534/g3.120.401408.
[51] Shuya Liu, Qing Xu, Nansheng Chen, Expansion of photoreception-related gene families may drive ecological adaptation of the dominant diatom species *Skeletonema marinoi*, Sci. Total Environ. (2023) 165384, https://doi.org/10.1016/j.scitotenv.2023.165384.
[52] Dongmei Wang, Xinzi Yu, Kuipeng Xu, Guiqi Bi, Min Cao, Ehud Zelzion, Chunxiang Fu, Peipei Sun, Yang Liu, Fanna Kong, Guoying Du, Xianghai Tang, Ruijuan Yang, Junhao Wang, Lei Tang, Lu Wang, Yingjun Zhao, Yuan Ge, Yunyun Zhuang, Zhaolan Mo, Yu Chen, Tian Gao, Xiaowei Guan, Rui Chen, Weihua Qu, Bin Sun, Debashish Bhattacharya, Yunxiang Mao, Pyropia Yezoensis genome reveals diverse mechanisms of carbon acquisition in the intertidal environment, Nat. Commun. 11 (1) (2020) 4028, https://doi.org/10.1038/s41467-020-17689-1.
[53] David R. Nelson, Khaled M. Hazzouri, Kyle J. Lauersen, Ashish Jaiswal, Amphun Chaiboonchoe, Alexandra Mystikou, Fu Weiqi, Sarah Daakour, Bushra Dohai, Amnah Alzahmi, David Nobles, Mark Hurd, Julie Sexton, Michael J. Preston, Joan Blanchette, Michael W. Lomas, Khaled M.A. Amiri, Kourosh Salehi-Ashtiani, Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution, Cell Host Microbe 29 (2) (2021) 250–266.e8, https://doi.org/10.1016/j.chom.2020.12.005.
[54] Roberts Wr, Ec Ruck, Km Downey, E. Pinseel, Aj Alverson, Resolving marine-freshwater transitions by diatoms through a fog of gene tree discordance, Syst. Biol. 72 (5) (2023), https://doi.org/10.1093/sysbio/syad038.
[55] Maria Sorokina, Emanuel Barth, Mahnoor Zulfiqar, Michiel Kwantes, Georg Pohnert, Christoph Steinbeck, Draft genome assembly and sequencing dataset of the marine diatom Skeletonema Cf. Costatum RCC75, Data in Brief 41 (2022) 107931, https://doi.org/10.1016/j.dib.2022.107931.
[56] Shunsuke Hirooka, Takeshi Itabashi, Takako M. Ichinose, Ryo Onuma, Takayuki Fujiwara, Shota Yamashita, Lin Wei Jong, Reiko Tomita, Atsuko H. Iwane, Shin-Ya Miyagishima, Life cycle and functional genomics of the unicellular red alga Galdieria for elucidating algal and plant evolution and industrial use, Proc. Natl. Acad. Sci. USA 119 (41) (2022) e2210665119, https://doi.org/10.1073/pnas.2210665119.
[57] Susan H. Brawley, Nicolas A. Blouin, Elizabeth Ficko-Blean, Glen L. Wheeler, Martin Lohr, Holly V. Goodson, Jerry W. Jenkins, Crysten E. Blaby-Haas, Katherine E. Helliwell, Cheong Xin Chan, Tara N. Marriage, Debashish Bhattacharya, Anita S. Klein, Yacine Badis, Juliet Brodie, Yuanyu Cao, Jonas Collén, Simon M. Dittami, Claire M.M. Gachon, Beverley R. Green, Steven J. Karpowicz, Jay W. Kim, Ulrich Johan Kudahl, Senjie Lin, Gurvan Michel, Maria Mittag, Bradley J.S.C. Olson, Jasmyn L. Pangilinan, Yi Peng, Huan Qiu, Shengqiang Shu, John T. Singer, Alison G. Smith, Brittany N. Sprecher, Volker Wagner, Wenfei Wang, Zhi-Yong Wang, Juying Yan, Charles Yarish, Simone Zäuner-Riek, Yunyun Zhuang, Yong Zou, Erika A. Lindquist, Jane Grimwood, Kerrie W. Barry, Daniel S. Rokhsar, Jeremy Schmutz, John W. Stiller, Arthur R. Grossman, Simon E. Prochnik, Insights into the red algae and eukaryotic evolution from the genome of Porphyra Umbilicalis (Bangiophyceae, Rhodophyta), Proc. Natl. Acad. Sci. USA 114 (31) (2017) E6361–E6370, https://doi.org/10.1073/pnas.1703088114.
[58] Alastair Skeffington, Axel Fischer, Sanja Sviben, Magdalena Brzezinka, Michał Górka, Luca Bertinetti, Christian Woehle, Bruno Huettel, Alexander Graf, André Scheffel, A joint proteomic and genomic investigation provides insights into the mechanism of calcification in Coccolithophores, Nat. Commun. 14 (1) (2023) 3749, https://doi.org/10.1038/s41467-023-39336-1.
[59] David M. Emms, Steven Kelly, OrthoFinder: phylogenetic Orthology inference for comparative genomics, Genome Biol. 20 (1) (2019) 238, https://doi.org/10.1186/s13059-019-1832-y.
[60] Mario Dos Reis, Dating microbial evolution with MCMCtree, Methods Mol. Biol. (Clifton, N.J.) 2569 (2022) 3–22, https://doi.org/10.1007/978-1-0716-2691-7_1.
[61] Fábio K. Mendes, Dan Vanderpool, Ben Fulton, Matthew W. Hahn, CAFE 5 models variation in evolutionary rates among gene families, Bioinf. (Oxf.) 36 (22−23) (2021) 5516–5518, https://doi.org/10.1093/bioinformatics/btaa1022.
[62] Arthur Zwaenepoel, Yves Van de Peer, Wgd-simple command line tools for the analysis of ancient whole-genome duplications, Bioinf. (Oxf.) 35 (12) (2019) 2153–2155, https://doi.org/10.1093/bioinformatics/bty915.
[63] Weijia Su, Shujun Ou, Matthew B. Hufford, Thomas Peterson, A tutorial of EDTA: extensive de novo TE annotator, Methods Mol. Biol. (Clifton, N.J.) 2250 (2021) 55–67, https://doi.org/10.1007/978-1-0716-1134-0_4.
[64] Haidong Yan, Aureliano Bombarely, Song Li, DeepTE: a computational method for de novo classification of transposons with convolutional neural network, Bioinf. (Oxf.) 36 (15) (2020) 4269–4275, https://doi.org/10.1093/bioinformatics/btaa519.

[65] Marc Krasovec, Sophie Sanchez-Brosseau, Gwenael Piganeau, First estimation of the spontaneous mutation rate in diatoms, Genome Biol. Evol. 11 (7) (2019) 1829–1837, https://doi.org/10.1093/gbe/evz130.
[66] Angela Falciatore, Marianne Jaubert, Jean-Pierre Bouly, Benjamin Bailleul, Thomas Mock, Diatom molecular research comes of age: model species for studying phytoplankton biology and diversity[OPEN], Plant Cell 32 (3) (2020) 547–572, https://doi.org/10.1105/tpc.19.00158.
[67] Chris Bowler, Andrew E. Allen, Jonathan H. Badger, Jane Grimwood, Kamel Jabbari, Alan Kuo, Uma Maheswari, Cindy Martens, Florian Maumus, Robert P. Otillar, Edda Rayko, Asaf Salamov, Klaas Vandepoele, Bank Beszteri, Ansgar Gruber, Marc Heijde, Michael Katinka, Thomas Mock, Klaus Valentin, Fréderic Verret, John A. Berges, Colin Brownlee, Jean-Paul Cadoret, Anthony Chiovitti, Chang Jae Choi, Sacha Coesel, Alessandra De Martino, J. Chris Detter, Colleen Durkin, Angela Falciatore, Jérome Fournet, Miyoshi Haruta, Marie J.J. Huysman, Bethany D. Jenkins, Katerina Jiroutova, Richard E. Jorgensen, Yolaine Joubert, Aaron Kaplan, Nils Kröger, Peter G. Kroth, Julie La Roche, Erica Lindquist, Markus Lommer, Véronique Martin-Jézéquel, Pascal J. Lopez, Susan Lucas, Manuela Mangogna, Karen McGinnis, Linda K. Medlin, Anton Montsant, Marie-Pierre Oudot-Le Secq, Carolyn Napoli, Miroslav Obornik, Micaela Schnitzler Parker, Jean-Louis Petit, Betina M. Porcel, Nicole Poulsen, Matthew Robison, Leszek Rychlewski, Tatiana A. Rynearson, Jeremy Schmutz, Harris Shapiro, Magali Siaut, Michele Stanley, Michael R. Sussman, Alison R. Taylor, Assaf Vardi, Peter von Dassow, Wim Vyverman, Anusuya Willis, Lucjan S. Wyrwicz, Daniel S. Rokhsar, Jean Weissenbach, E. Virginia Armbrust, Beverley R. Green, Yves Van de Peer, Igor V. Grigoriev, The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes, Nature 456 (7219) (2008) 239–244, https://doi.org/10.1038/nature07410.
[68] Florian Maumus, Andrew E. Allen, Corinne Mhiri, Hu Hanhua, Kamel Jabbari, Assaf Vardi, Marie-Angèle Grandbastien, Chris Bowler, Potential impact of stress activated retrotransposons on genome evolution in a marine diatom, BMC Genomics 10 (1) (2009) 624, https://doi.org/10.1186/1471-2164-10-624.
[69] Tianjun Cao, Yu Bai, Paul Buschbeck, Qiaozhu Tan, Michael B. Cantrell, Yinjuan Chen, Yanyou Jiang, Run-Zhou Liu, Nana K. Ries, Xiaohuo Shi, Yan Sun, Maxwell A. Ware, Fenghua Yang, Huan Zhang, Jichang Han, Lihan Zhang, Jing Huang, Martin Lohr, Graham Peers, Xiaobo Li, An unexpected hydratase synthesizes the Green light-absorbing pigment Fucoxanthin, Plant Cell 35 (8) (2023) 3053–3072, https://doi.org/10.1093/plcell/koad116.
[70] Sebastià Capó-Bauçà, Concepción Iñiguez, Jeroni Galmés, The diversity and coevolution of Rubisco and CO2 concentrating mechanisms in marine Macrophytes, New Phytol. 241 (6) (2024) 2353–2365, https://doi.org/10.1111/nph.19528.
[71] John R. Reinfelder, Carbon concentrating mechanisms in eukaryotic marine phytoplankton, Annu. Rev. Mar. Sci. 3 (2011) 291–315, https://doi.org/10.1146/annurev-marine-120709-142720.
[72] Sivakamavalli Jeyachandran, Hethesh Chellapandian, Kiyun Park, Ihn-Sil Kwak, A review on the involvement of heat shock proteins (extrinsic chaperones) in response to stress conditions in aquatic organisms, Antioxidants (Basel, Switzerland) 12 (7) (2023) 1444, https://doi.org/10.3390/antiox12071444.
[73] Jeffrey M. Rousch, Scott E. Bingham, Milton R. Sommerfeld, Protein expression during heat stress in Thermo-intolerant and Thermo-tolerant diatoms, J. Exp. Mar. Biol. Ecol. 306 (2) (2004) 231–243, https://doi.org/10.1016/j.jembe.2004.01.009.
[74] Hung-Yun Lin, Chung-Hsiao Liu, Yong-Ting Kang, Sin-Wei Lin, Hsin-Yun Liu, Chun-Ting Lee, Yu-Chen Liu, Man-Chun Hsu, Ya-Yun Chien, Shao-Ming Hong, Yun-Hsuan Cheng, Bing-You Hsieh, Han-Jia Lin, Enhancing the spermidine synthase-based polyamine biosynthetic pathway to boost rapid growth in marine diatom Phaeodactylum Tricornutum, Biomolecules 14 (3) (2024) 372, https://doi.org/10.3390/biom14030372.
[75] Chinyere Aloke, Olalekan Olugbenga Onisuru, Ikechukwu Achilonu, Glutathione S-transferase: a versatile and dynamic enzyme, Biochem. Biophys. Res. Commun. 734 (2024) 150774, https://doi.org/10.1016/j.bbrc.2024.150774.
[76] Chun-Yang Li, Hai-Yan Cao, Rocky D. Payet, Jonathan D. Todd, Yu-Zhong Zhang, Dimethylsulfoniopropionate (DMSP): from biochemistry to global ecological significance, Ann. Rev. Microbiol. 78 (1) (2024) 513–532, https://doi.org/10.1146/annurev-micro-041222-024055.
[77] Graham Noctor, Amna Mhamdi, Sejir Chaouch, Yi Han, Jenny Neukermans, Belen Marquez-Garcia, Guillaume Queval, Christine H. Foyer, Glutathione in plants: an integrated overview, Plant Cell Environ. 35 (2) (2012) 454–484, https://doi.org/10.1111/j.1365-3040.2011.02400.x.
[78] Lijuan Yan, Yongjun Liu, Mechanistic insights into the anaerobic degradation of globally abundant Dihydroxypropanesulfonate catalyzed by the DHPS-Sulfolyase (HpsG), J. Chem. Inf. Model. 62 (11) (2022) 2880–2888, https://doi.org/10.1021/acs.jcim.2c00174.
[79] L.K. Medlin, W.H. Kooistra, R. Gersonde, U. Wellbrock, Evolution of the diatoms (Bacillariophyta). II. Nuclear-encoded small-subunit rRNA sequence comparisons confirm a paraphyletic origin for the centric diatoms, Mol. Biol. Evol. 13 (1) (1996) 67–75, https://doi.org/10.1093/oxfordjournals.molbev.a025571.
[80] Edward C. Theriot, Matt Ashworth, Elizabeth Ruck, Teofil Nakov, Robert K. Jansen, A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research, Plant Ecology and Evolution 143 (3) (2010) 278–296, https://doi.org/10.5091/plecevo.2010.418.
[81] Metin G. Davutoglu, Veikko F. Geyer, Lukas Niese, Johannes R. Soltwedel, Marcelo L. Zoccoler, Valeria Sabatino, Robert Haase, Nils Kröger, Stefan Diez, Nicole Poulsen, Gliding motility of the diatom Craspedostauros Australis coincides with the intracellular movement of Raphid-specific Myosins, Communications Biology 7 (1) (2024) 1187, https://doi.org/10.1038/s42003-024-06889-w.
[82] N.C. Poulsen, I. Spector, T.P. Spurck, T.F. Schultz, R. Wetherbee, Diatom gliding is the result of an actin-myosin motility system, Cell Motil. Cytoskeleton 44 (1) (1999) 23–33, https://doi.org/10.1002/(SICI)1097-0169(199909)44:1%3C23::AID-CM2%3E3.0.CO;2-D.
[83] Florian Maumus, Andrew E. Allen, Corinne Mhiri, Hu Hanhua, Kamel Jabbari, Assaf Vardi, Marie-Angèle Grandbastien, Chris Bowler, Potential impact of stress activated retrotransposons on genome evolution in a marine diatom, BMC Genomics 10 (1) (2009) 624, https://doi.org/10.1186/1471-2164-10-624.
[84] C. Vitte, O. Panaud, LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model, Cytogenet. Genome Res. 110 (1–4) (2005) 91–107, https://doi.org/10.1159/000084941.
[85] Maud I. Tenaillon, Jesse D. Hollister, Brandon S. Gaut, A triptych of the evolution of plant transposable elements, Trends Plant Sci. 15 (8) (2010) 471–478, https://doi.org/10.1016/j.tplants.2010.05.003.
[86] Elena Casacuberta, Josefa González, The impact of transposable elements in environmental adaptation, Mol. Ecol. 22 (6) (2013) 1503–1517, https://doi.org/10.1111/mec.12170.
[87] Katrien M. Devos, James K.M. Brown, Jeffrey L. Bennetzen, Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis, Genome Res. 12 (7) (2002) 1075–1079, https://doi.org/10.1101/gr.132102.
[88] Jean-Francois Gout, Yue Hao, Parul Johri, Olivier Arnaiz, Thomas G. Doak, Simran Bhullar, Arnaud Couloux, Fréderic Guérin, Sophie Malinsky, Alexey Potekhin, Natalia Sawka, Linda Sperling, Karine Labadie, Eric Meyer, Sandra Duharcourt, Michael Lynch, Dynamics of gene loss following ancient whole-genome duplication in the cryptic Paramecium complex, Mol. Biol. Evol. 40 (5) (2023) msad107, https://doi.org/10.1093/molbev/msad107.
[89] Yangxin Zhang, Kecheng Qian, Yu Qiaoming, Xiangxiang Chen, Jiakai Liang, Zhiguang Liu, Zhuoxuan Dong, Yunxiao Liu, Yaqiang Sun, Zhenhua Guo, Fengwang Ma, Tao Zhao, Subgenomic divergence and functional innovation following whole-genome duplication in Maleae species of Rosaceae, Plant J. 123 (6) (2025) e70499, https://doi.org/10.1111/tpj.70499.
[90] Sarah Marburger, Markos A. Alexandrou, John B. Taggart, Simon Creer, Gary Carvalho, Claudio Oliveira, Martin I. Taylor, Whole genome duplication and transposable element proliferation drive genome expansion in Corydoradinae catfishes, Proc. R. Soc. B Biol. Sci. 285 (1872) (2018) 20172732, https://doi.org/10.1098/rspb.2017.2732.
[91] E.E. Marchani, J. Xing, D.J. Witherspoon, L.B. Jorde, A.R. Rogers, Estimating the age of retrotransposon subfamilies using maximum likelihood, Genomics 94 (2009) 78–82.
[92] Liyan He, Xiaotian Han, Yu. Zhiming, A rare Phaeodactylum Tricornutum cruciform Morphotype: culture conditions, transformation and unique fatty acid characteristics, PLoS One 9 (4) (2014) e93922, https://doi.org/10.1371/journal.pone.0093922.
[93] T. Hunter, Signaling–2000 and beyond, Cell 100 (1) (2000) 113–127, https://doi.org/10.1016/s0092-8674(00)81688-8.
[94] Amit Kumar, Jochen Balbach, Folding and stability of Ankyrin repeats control biological protein function, Biomolecules 11 (6) (2021) 840, https://doi.org/10.3390/biom11060840.
[95] Gillian Murphy, Hideaki Nagase, Progress in matrix metalloproteinase research, Mol. Asp. Med. 29 (5) (2008) 290–308, https://doi.org/10.1016/j.mam.2008.05.002.
[96] Teng Huang, Hu Fan, Yufang Pan, Chenjie Li, Hu. Hanhua, Pyruvate orthophosphate Dikinase is required for the acclimation to high bicarbonate concentrations in Phaeodactylum Tricornutum, Algal Res. 72 (2023) 103131, https://doi.org/10.1016/j.algal.2023.103131.
[97] Guillaume Bourque, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, Zsuzsanna Izsvák, Henry L. Levin, Todd S. Macfarlan, Dixie L. Mager, Cédric Feschotte, Ten things you should know about transposable elements, Genome Biol. 19 (1) (2018) 199, https://doi.org/10.1186/s13059-018-1577-z.
[98] Achal Rastogi, Uma Maheswari, Richard G. Dorrell, Fabio Rocha Jimenez Vieira, Florian Maumus, Adam Kustka, James McCarthy, Andy E. Allen, Paul Kersey, Chris Bowler, Leila Tirichine, Integrative analysis of large scale transcriptome data draws a comprehensive landscape of Phaeodactylum Tricornutum genome and evolutionary origin of diatoms, Sci. Rep. 8 (1) (2018) 4834, https://doi.org/10.1038/s41598-018-23106-x.