Contents lists available at ScienceDirect

Science of the Total Environment

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

Long-read genome sequencing provides novel insights into the harmful algal bloom species *Prymnesium parvum*

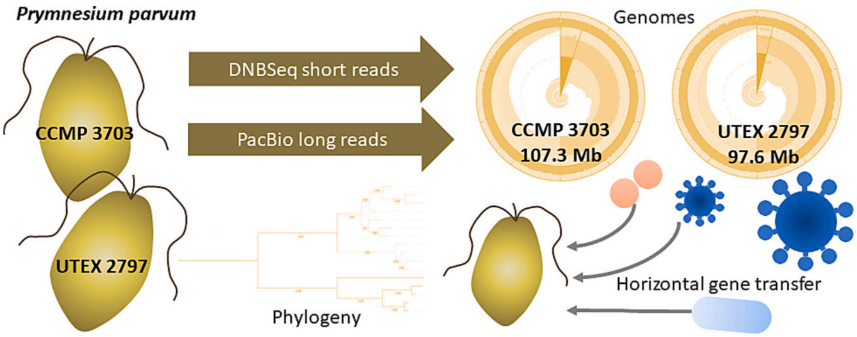
Jianbo Jian [a,b], Zhangyan Wu [b], Arisbe Silva-Núñez [a,c], Xiaohui Li [b], Xiaomin Zheng [b], Bei Luo [b], Yun Liu [b], Xiaodong Fang [b], Christopher T. Workman [a], Thomas Ostenfeld Larsen [a], Per Juel Hansen [d], Eva C. Sonnenschein [a,e,*]

[a] *Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark*
[b] *BGI-Genomics, BGI-Shenzhen, Shenzhen, China*
[c] *Tecnologico de Monterrey, School of Engineering and Science, Monterrey, Nuevo León, Mexico*
[d] *Marine Biological Section, University of Copenhagen, Helsingør, Denmark*
[e] *Department of Biosciences, Swansea University, Swansea, United Kingdom*

HIGHLIGHTS

- High-quality genomes of *Prymnesium parvum* strains UTEX 2797 and CCMP 3037 were assembled using advanced sequencing techniques.
- The genomes present some of highest contiguous assembled contig sequences to date.
- The genomes carry crucial pathways for the evolution of *P. parvum* and may support monitoring and managing algal blooms.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Abasiofiok Mark Ibekwe

Keywords:
Harmful algal blooms
Comparative genomics
Horizontal gene transfer
Polyketide synthase

ABSTRACT

Prymnesium parvum is a toxin-producing haptophyte that causes harmful algal blooms worldwide, which are often associated with massive fish-kills and subsequent economic losses. In here, we present nuclear and plastid genome assemblies using PacBio HiFi long reads and DNBseq short reads for the two *P. parvum* strains UTEX 2797 and CCMP 3037, representing producers of type A prymnesins. Our results show that the *P. parvum* strains have a moderate haptophyte genome size of 97.56 and 107.32 Mb. The genome assemblies present one of highest contiguous assembled contig sequences to date consisting of 463 and 362 contigs with a contig N50 of 596.99 kb and 968.39 kb for strain UTEX 2797 and CCMP 3037, respectively. The assembled contigs of UTEX 2797 and CCMP 3037 were anchored to 34 scaffolds, with a scaffold N50 of 5.35 Mb and 3.61 Mb, respectively, accounting for 93.2 % and 97.9 % of the total length. Each plastid genome comprises a circular contig. A total of 20,578 and 19,426 protein-coding genes were annotated for UTEX 2797 and CCMP 3037. The expanded gene family analysis showed that starch and sucrose metabolism, sulfur metabolism, energy metabolism and ABC transporters are involved in the evolution of *P. parvum*. Polyketide synthase (PKS) genes responsible for the production of

* Corresponding author at: Department of Biosciences, Swansea University, Swansea, United Kingdom.
E-mail address: e.c.sonnenschein@swansea.ac.uk (E.C. Sonnenschein).

https://doi.org/10.1016/j.scitotenv.2023.168042
Received 21 August 2023; Received in revised form 20 October 2023; Accepted 20 October 2023
Available online 26 October 2023

secondary metabolites such as prymnesins displayed different expression patterns under nutrient limitation. Overlap with repeats and horizontal gene transfer may be two contributing factors to the high number of PKS genes found in this species. The two high quality *P. parvum* genomes will serve as valuable resources for ecological, genetic, and toxicological studies of haptophytes that can be used to monitor and potentially manage harmful blooms of ichthyotoxic *P. parvum* in the future.

1. Introduction

Prymnesium parvum is a haptophyte microalga with a global distribution. It can be found in freshwater environments such as rivers, lakes, and ponds, as well as in brackish water environments such as embayments and estuaries (Granéli et al., 2012; Salcher et al., 2011). Its ability to cause harmful algal blooms (HABs) on a global level has become a growing concern (Wagstaff et al., 2018; Svendsen et al., 2018; Blossom et al., 2014). For over a century, *P. parvum* blooms have led to large-scale fish kills worldwide with the first being reported in 1920 (Svendsen et al., 2018; Wagstaff et al., 2021). It produces toxins referred to as prymnesins, which belong to the polyketides and are biosynthesized by complex, multi-domain enzymes, the polyketide synthases (PKSs). The prymnesins are classified into three types (type A, B and C) based on the length of the carbon backbone of the compound (Binzer et al., 2019), and many prymnesin analogs have been described (Rasmussen et al., 2016; Taylor et al., 2020; Taylor et al., 2021).

Many studies have investigated *P. parvum* and the prymnesins (Binzer et al., 2019; Driscoll et al., 2023; La Claire, 2006; Lundgren et al., 2016; Manning and La Claire, 2010). All strains seem to produce one of the 3 types of prymnesins (Binzer et al., 2019), but we still have limited knowledge of the factors affecting their production, biosynthesis and their toxicity to different types of organisms. It is however, accepted that environmental factors like salinity, pH, light and nutrient availability may affect prymnesin production (Franco et al., 2019). It has recently been shown that low salinity combined with low pH decreases the toxicity (Caron et al., 2023). Light is required for the production of prymnesins, and since growth rates of *P. parvum* increases with light intensity so does the production of prymnesins (Medić et al., 2022). However, prymnesins that are released into the water have been shown to be photo-labile and exposure of a filtrate from a dense *P. parvum* culture to high light levels leads to a reduction of acute toxicity to fish (Taylor et al., 2021). Nitrogen and phosphorus limitation affect expression of photosynthetic genes and polyketide synthase genes in *P. parvum* (Anestis et al., 2021; Anestis et al., 2022; Liu et al., 2015).

Genomic and transcriptomic approaches have been used to study the biology of *P. parvum*. Comparative analysis of transcriptomes of four algal species, *P. parvum*, *Chrysochromulina brevifilum*, *Chrysochromulina ericina* and *Phaeocystis antarctica*, found that species clustered with phylogeny and nutritional modes meaning that genomic factors relate to both evolutionary relationships and trophic ecology (Koid et al., 2014). Using expressed sequence tags of *P. parvum*, putative genes involved in the biosynthesis and secretion of prymnesin toxins were identified (La Claire, 2006). Proteomic analysis studying the effect of iron on *P. parvum* showed that production of ABC transporters and related genes including ribulose biphosphate carboxylase (RuBisCo), malate dehydrogenase, manganese superoxide dismutase and serine threonine kinase and two Fe-independent oxidative stress response proteins increased in response to iron limitation (Mamunur Rahman et al., 2014). Due to the lack of a *P. parvum* reference genome, previous studies explored the PKSs-related genes through transcriptomic data. Numerous type I PKSs (37 to 109) were identified from transcripts of nine different strains (Anestis et al., 2021) and 15 PKSs were found in *P. parvum* in six transcriptomic samples (Kohli et al., 2016). PKS genes exhibited different expression patterns under phosphorus limitation and higher PKS gene expression was consistent with higher toxicity (Liu et al., 2015). However, the exact relationship between PKS genes and toxin production remains unclear (Manning and La Claire, 2010).

Recently, genomes of 15 strains of *P. parvum* were reported, with two assembled from a combination of Illumina, Nanopore and Hi-C sequence data and the remaining 13 based on Illumina data. The genomic data revealed an exceptional level of genome diversity and identified candidate gene families involved in the biosynthesis of toxic metabolites (Wisecaver et al., 2023). Genomic diversity can be driven by horizontal gene transfer (HGT) between organisms of the same and different kingdoms as well as by viruses. With the recent revolution in algal genome sequencing, many viruses infecting eukaryotic algae had been identified as DNA viruses have the ability to enter into the host genome (Feschotte and Gilbert, 2012). For example, nucleocytoplasmic large DNA viruses exhibit a tight assimilation into the host genome, thereby substantially impacting genome composition and providing new genetic material to algal lineages (Moniruzzaman et al., 2020). As of 2020, a total of 61 viruses have been documented to infect algae (Short et al., 2020) and some viruses even have been shown to transfer between different hosts (Dolja and Koonin, 2011; Dolja and Koonin, 2018). Viruses play a crucial role as members of the biosphere, contributing to its genetic diversity and dynamics, however further investigations are necessary to fully understand their impact on algal diversity. Genomic information of haptophytes and particularly *P. parvum* is still limited, however necessary as a resource to understand haptophyte evolution and toxin production and HAB formation of *P. parvum*. *P. parvum* UTEX 2797 and CCMP3037 originate both from US freshwater systems and are well studied laboratory strains. Prior to genome sequencing, they were specifically investigated for identification of natural products, including prymnesins. Strains CCMP3037 and UTEX 2797 represent producers of type A prymnesins (Binzer et al., 2019), however belong to different phylogenetic sub-groups within type A based on internal transcribed spacers. In here, we report high-quality genomes of two strains of *P. parvum* based on long reads sequencing.

2. Methods

2.1. Cultivation and genomic DNA extraction

The two *Prymnesium parvum* strains CCMP 3037 and UTEX 2797 were obtained from the NCMA - National Center for Marine Algae and Microbiota (https://ncma.bigelow.org/) and UTEX - Culture Collection of Algae at the University of Texas in Austin (https://utex.org/pages/search-results/collections-living-algal-strains), respectively. The algae were maintained in sterile-filtered f/2 media (Guillard and Ryther, 1962) at constant light of 70 μmol photons m^{-2} s^{-1}. For upscaling, the strains were cultivated in f/2 media in a 10-L Nalgene bottles (Thermo Scientific, NY, USA). After 2 weeks' culture, the biomasses were harvested by centrifugation and freeze-dried. Then, the total genomic DNA was extracted using the DNeasy Plant Kit (Qiagen).

2.2. Library preparation and sequencing

The short insert size (300-500 bp) libraries were prepared following the protocol of MGI Library Prep Reagents and sequenced with 150 bp paired-end by MGISEQ-2000 platform. The adapter, PCR duplicates, N content (N bases >1 %) and low-quality (quality value ≤10 with low-quality base >20 %) reads were removed by the software of SOAPnuke version 1.5.3 (Chen et al., 2018). The PacBio circular consensus sequencing (CCS) HiFi library was prepared using the SMRTbell Prep Kit 2.0 and sequenced on PacBio Sequal II platform.

2.3. Genome survey and genome assembly

To assist sequencing strategy, the genome survey firstly was performed to confirm the genome size, GC content, heterozygosity and repeat content. The k-mer distribution profiling was computed by employing jellyfish software on quality-filtered short DNB-Seq reads (Marcais and Kingsford, 2011). The genomes of two strains of *Prymnesium parvum* were characterized using K-mer profiling through the application of Genomescope 1.0 (Vurture et al., 2017). The long high quality sequence reads were assembled using hifiasm v0.7 with default parameters (Cheng et al., 2021). The plastid genome was assembled into a complete circular plastome using GetOrganelle pipline with default parameters (Jin et al., 2020).

2.4. Genome curation

The initial assembly of *P. parvum* strain CCMP 3037 and *P. parvum* strain UTEX 2797 consisted of 602 and 709 contigs, respectively, with a combined length of 137.57 MB and 120.16 MB. The initial assembled size exceeds the estimated genome size. To eliminate duplicated sequences originating from low abundance, heterozygous regions or contaminant reads, we employed the PurgeHaplotigs pipeline to curate the heterozygous diploid genome (Roach et al., 2018). A total of 217 and 127 relatively short sequences, with a combined length of 17.6 Mb and 8.4 Mb respectively, were eliminated from *P. parvum* strain CCMP 3037 and *P. parvum* strain UTEX 2797. The assembled sequences of the two curated strains underwent a filtration process based on GC content, taxonomic annotations, and organelle data to eliminate any potential contaminants. The contigs were subjected to a search using mega-blast against the GenBank nucleotide (nt) database, as well as Minimap2 with haptophyte organelle data available on NCBI. Hits with an identity of 90 % and an aligned length of 500 were utilized for filtering, followed by manual verification of the taxonomic hits. The contigs with GC content below 45 % were derived from either organelle data or bacterial sources. With the implementation of these process approaches, the assembled genome of *P. parvum* CCMP 3037 was revised from 119.97 MB (385 contigs) to 107.32 MB (362 contigs), while that of *P. parvum* UTEX 2797 was reduced from 111.78 MB (578 contigs) to 97.56 MB (463 contigs).

2.5. Genome annotation

The newly assembled genomes of two strains of *P. parvum* were utilized for repeat identification and gene structure prediction. Firstly, through the homology-based methods, the similarly repetitive sequences were identified using RepeatModeler v2.0 (Flynn et al., 2020). (http://www.repeatmasker.org/RepeatModeler/) and LTR_FINDER v1.07 (Xu and Wang, 2007) (http://tlife.fudan.edu.cn/ltr_finder/) based on the repeat sequence database of RepBase v21.12 (http://www.girinst.org/repbase) (Bao et al., 2015). Combing ab initio method, the repetitive sequences were predicted by RepeatMasker 3.3.0 (Saha et al., 2008). After repeat annotation, the gene annotation was performed with three different algorithms (homology-based, RNA-Seq data-based and ab initio prediction). In ab initio prediction, the repeated-masked genome sequences were used to predict coding regions of genes using AUGUSTUS v3.2.3 (Stanke et al., 2006) and SNAP (Korf, 2004).

A total of 10 published available homology protein sequence data sets from *A. protothecoides*, *C. reinhardtii*, *T. pseudonana*, *G. theta*, *C. vulgaris*, *C. tobini*, *D. lutheri*, *E. huxleyi*, *Pavlovales* sp. CCMP2436, and *T. lutea* were mapped to the two assembled genomes using TBLASTn (Kent, 2002). Three libraries of transcriptomic data were downloaded from NCBI database, including PolyA (SRP06612, SRP050355), Random (SRP042159) and cDNA (SRP026696). A total of 30 runs of sequencing data, totaling approximately 68.55 Gb, were aligned to the two genome sequences using hista2.2.1(Kim et al., 2019). The average mapping rate of thirty RNA-Seq datasets ranged from 46.71 % to 90.2 %, with a mean value of 73.56 %.With all of the aligned result, the transcripts were constructed using stringtie2.1.6 (Kovaka et al., 2019) and the gtf format was converted to gff with cufflinks. Finally, all gene prediction data combing three kinds of evidence were integrated into a consensus gene set using the MAKER pipeline (v3.31.8) (Holt and Yandell, 2011). The plastid genome was annotated using the software of GeSeq (Tillich et al., 2017). The gene set were evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO version 5.1.2) with the "eukaryote_odb10" database with 255 conserved core eukaryotic genes.

2.6. Gene family and phylogenomic analysis

The amino acid sequences from the two newly sequenced genomes were compared with four available haptophyte species (*C. tobini*, *D. lutheri*, *E. huxleyi* and *Pavlovales* sp. CCMP2436) and eight other representative species (*T. pseudonana*, *P. tricornutum*, *A. anophagefferens*, *C. reinhardtii* V5, *C. desiccate*, *F. kawagutii*, *Symbiodinium* sp., *P. purpureum* and *C. paradoxa*) using OrthoFinder-2.3.11 (Emms and Kelly, 2019). The gene sets of the 14 genomes were initially processed prior to gene clustering. Genes encoding proteins <50 amino acids in length were excluded, and only the longest transcript was retained when multiple spliced transcripts existed within a single gene. The protein sequences were utilized to cluster orthogroups, which were subsequently employed in inferring single copy homologous groups using the MSA program. A total of 979 orthogroups underwent multiple sequence alignment via MAFFT (v7.310) (Katoh and Standley, 2013) followed by gap position removal with the software of trimal (v1.4.1) (Capella-Gutierrez et al., 2009). The phylogenetic tree was inferred by (Randomized Axelerated Maximum Likelihood) RAxML-8.2.11 (version8.2.12) with PROTGAMMALGX program model (Stamatakis, 2014). Then, the species of *C. paradoxa* was rooted as outgroup using TreeBest (https://github.com/Ensembl/treebest). The MCMCtree of PAML (version 4.9j) was used to infer the species divergence time (Yang, 2007). Three calibrated divergence times were derived from the TimeTree database (http://www.timetree.org/): *T. pseudonana-P. tricornutum* (132.0–227.9 Mya), *C. reinhardtii-P. tricornutum* (781.1–803.1 Mya), and *A. anophagefferens-E. huxleyi* (824.3–1900.0 Mya). The gene family expansion and contraction of 15 genomes were identified with CAFE v4.2 pipeline (De Bie et al., 2006) and the expanded gene families were applied to KEGG and GO enrichment for obtain their functions. Principal components analysis was conducted to confirm the species clade classification based on InterPro.

2.7. RNA-Seq data analysis

Three publicly available data samples (SRP042159: SRR1296917, SRR1296973, and SRR1296769) were downloaded from the NCBI Sequence Read Archive under accession number SRA166613. These samples represent axenic (replete treatment), N-limited, and P-limited treatments, respectively. The gene expression analysis was conducted using *P. parvum* UTEX 2797 as the reference genome. The downloaded data was filtered by SOAPnuke version 1.5.3(Chen et al., 2018) with low quality reads N content (N bases >5 %) and low-quality (quality value ≤15 with low-quality base >20 %) removed. The clean reads were mapped to the reference genome with HISAT v2.1.0 (Kim et al., 2015) and reference gene with Bowtie2 (Langmead and Salzberg, 2012), respectively. Then, the gene expression was quantified using RSEM (Li and Dewey, 2011). In this study, the differentially expressed genes (DEGs) were identified using the PossionDis method (Audic and Claverie, 1997). The genes significantly differentially expressed between the three samples were determined with the threshold (false discovery rate (FDR) ≤0.001 and fold change ≥2).

2.8. The identification of polyketide synthase genes and putative HGT

The gene clusters responsible for the biosynthesis of secondary metabolites, particularly the putative PKS genes of interest in the two *P. parvum* genomes, were predicted using the bacterial, fungal and plant versions of antiSMASH 7 beta available on https://antismash.secondarymetabolites.org/ (Blin et al., 2021). With the input of FASTA and GFF annotation files, all three versions of antiSMASH were utilized to predict PKS in the algae. The results indicated that the bacterial and fungal versions yielded the same outcome, while only a limited number of secondary metabolites were predicted by plantiSMASH. The conserved domains of PKS from the Pfam database were utilized to determine the domain architecture for protein sequences of either PKS or hybrid PKS-NRPS.

Putative HGT genes from prokaryotes were identified using the HGTphyloDetect pipeline (Yuan et al., 2023), with modifications made to blastp by Diamond. The parameters of AI (Alien Index) $\geq$ 45 and out_pct $\geq$ 90 % were considered for potential HGT candidates. The target species genes were analysed using blastp with diamond and an $1E^{-10}$ against the NCBI non-redundant(nr) protein database (20230420). The identified putative genes were utilized to cluster ortholog groups using OrthoFinder-2.3.11 (Emms and Kelly, 2019). The homologous protein sequences of target species and Nr databases were subjected to multiple sequence alignment using MAFFT v7.310 (Katoh and Standley, 2013) and trimmed with Trimal-1.4.1 (Capella-Gutierrez et al., 2009) using the gappyout parameter. The topologies of putative and mapped genes in the database were inferred using IQ-TREE multicore version 2.2.0 (Nguyen et al., 2015). Subsequently, the phylogenies were rooted in middle branches and visualized through iTOL v6 (Letunic and Bork, 2021). Finally, gene expression profiling was employed to confirm the viral genes that have been transferred.

3. Results and discussion

3.1. Genome analysis of P. parvum

The diversity of phenotypes and adaptations to ecological niches are remarkable in algae, which is encoded in their distinctly different genomes (Blaby-Haas and Merchant, 2019). Specifically, the genome sequences of *P. parvum* will provide a foundation for evolutionary understanding and biosynthesis of prymnesins. Two different strains of *P. parvum*, CCMP 3037 and UTEX 2797, were collected for whole-genome sequencing. Before long read sequencing, a genome survey of short read data was performed to estimate the genome size and heterozygosity. A total of 21.22 Gb and 13.69 Gb of clean DNBSeq short read data were sequenced for *P. parvum* CCMP 3037 and UTEX 2797, respectively. The fact that the Q20 quality of reads 1 and reads 2 in the sequenced data of two strains exceeded 96.5 % indicates the production of high-quality data. The Q30 value of total reads were approximately 92 % and 91 % for CCMP 3037 and UTEX 2797, respectively (Table S1). The genomes of *P. parvum* CCMP 3037 and UTEX 2797 were assessed using K-mer analysis tools (jellyfish and GenomeScope, Kmer = 21). The results indicated that the genome sizes of these two strains were approximately 113.28 Mb and 108.96 Mb with a heterozygosity of 2.21 % and 0.95 %, respectively (Figs. S1, S2). The PacBio long reads were generated using the circular consensus sequencing (CCS) model with barcoding for the two strains, followed by splitting into HiFi reads.

A total of 3.47 Gb and 3.18 Gb clean long read data were generated for strains CCMP 3037 and UTEX 2797, respectively (Table S2). The HiFi reads were utilized for Hifiasm-based assembly of primary genome sequences, resulting in 137.57 MB and 120.16 MB for strains CCMP 3037 and UTEX 2797, respectively (Table S3). After genome curation including removal of the lower GC content, contaminant and organelle-related sequences, the final nuclear genome assembly of strain CCMP 3037 comprised approximately 107.32 Mb across 362 contigs (Table 1, Table S4 and Fig. 1). The final nuclear genome of strain UTEX 2797 was 97.56 Mb, consisting of 463 contigs (Table 1 and Fig. 1). For strain CCMP 3037, the contig N50 was 968,388 bp with a maximum assembled contig length of 5,352,942 bp. In addition, for strain UTEX 2797, the contig N50 was slightly lower at 596,989 bp but still impressive with a maximum assembled contig length of 3,608,513 bp (Table 1, Table S4). Nuclear genome sequences of six haptophyte genera have been published until now (Carrier et al., 2018; Hovde et al., 2019; Hovde et al., 2015; Hulatt et al., 2021; Read et al., 2013; Wisecaver et al., 2023). The genome sizes of *P. parvum* are significantly larger than four of the published haptophyte genomes (from 45.3 Mb for *Diacronema lutheri* to 65.77 Mb for *Chrysochromulina parva*), but smaller than that of *Emiliania huxleyi* (155.93 Mb).

Table 1
The assembly and annotation statistics for *P. parvum* CCMP 3037, UTEX 2797 and two published strains.

	CCMP 3037 (this study)	UTEX 2797 (this study)	12B1 (Wisecaver et al., 2023)	UTEX 2797 (Wisecaver et al., 2023)
Data type	PacBio HiFi, DNBseq	PacBio HiFi, DNBseq	Illumina, ONT, Hi-C	Illumina, ONT, Hi-C
Assembled genome size (bp)	107,321,770	97,558,583	93,538,114	197,592,770 (two *haploids*)
Contig N50 (length)	968,388	596,989	852,115	548,273
Max contig length	5,352,942	3,608,513	3,281,684	3,504,997
Number of contigs	362	463	225	585
Scaffold N50 *	3,786,890	3,230,053	3,203,049	3,431,116
Scaffold L50	10	11	11	21
Plastid genome size (bp)	107,830	107,831	NA	NA
Repetitive elements (%)	29.67	24.76	29.4	35.5
GC content (%)	57.4	57.6	57.7	57.3
Coding gene	20,578	19,426	24,964	47,239
Gene BUSCO (complete %)	77.2	81.2	84.7	84.7
Gene BUSCO (complete dup%)	5.1	4.5	7.8	65.1

The quality of the *P. parvum* assemblies was higher than most of the current haptophyte genomes, which had a contig N50 length of <30 kb except *D. lutheri* (852.26 kb). The N50 and number of contigs for two newly sequenced strains were comparable to those of recently published strains (12B1 with a contig N50 of 852,115 bp and 225 contigs; UTEX 2797 with a contig N50 of 548,273 bp and 585 contigs) assembled from Illumina, Nanopore, and Hi-C data (Wisecaver et al., 2023). However, the previously assembled UTEX 2797 genome of 197.6 Mbp was found to consist of two haplotypes with a two-fold increase in genome size and the majority of BUSCOs (75.5 %) being duplicated, which may be attributed to the presence of heterozygous sequences (Wisecaver et al., 2023). Using RagTag v2.1.0 (Alonge et al., 2019), as a reference guide to anchor contigs to nearly chromosome level, the two published UTEX 2797 (ONT) and 12B1 were employed to assist in anchoring the UTEX 2797 (HiFi) contigs. As a result, a total of 318 contigs with 88,647,005 bp (90.9 %) and 335 contigs with 90,945,659 bp (93.2 %) were successfully anchored to nearly chromosome level. The better improved anchoring based on 12B1 suggested the presence of some duplicated or incorrectly positioned haplotypes in UTEX 2797 (ONT). The contigs of CCMP 3037 were assembled into 34 scaffolds (97.9 %) with a total length of 105,105,890 bp using the 12B1 reference genome. Finally, the assembled contigs of UTEX 2797 and CCMP 3037 were anchored to 34 scaffolds with a scaffold N50 of 5.35 Mb and 3.61 Mb, respectively (Table 1). The published genome of UTEX 2797 (ONT) is nearly double

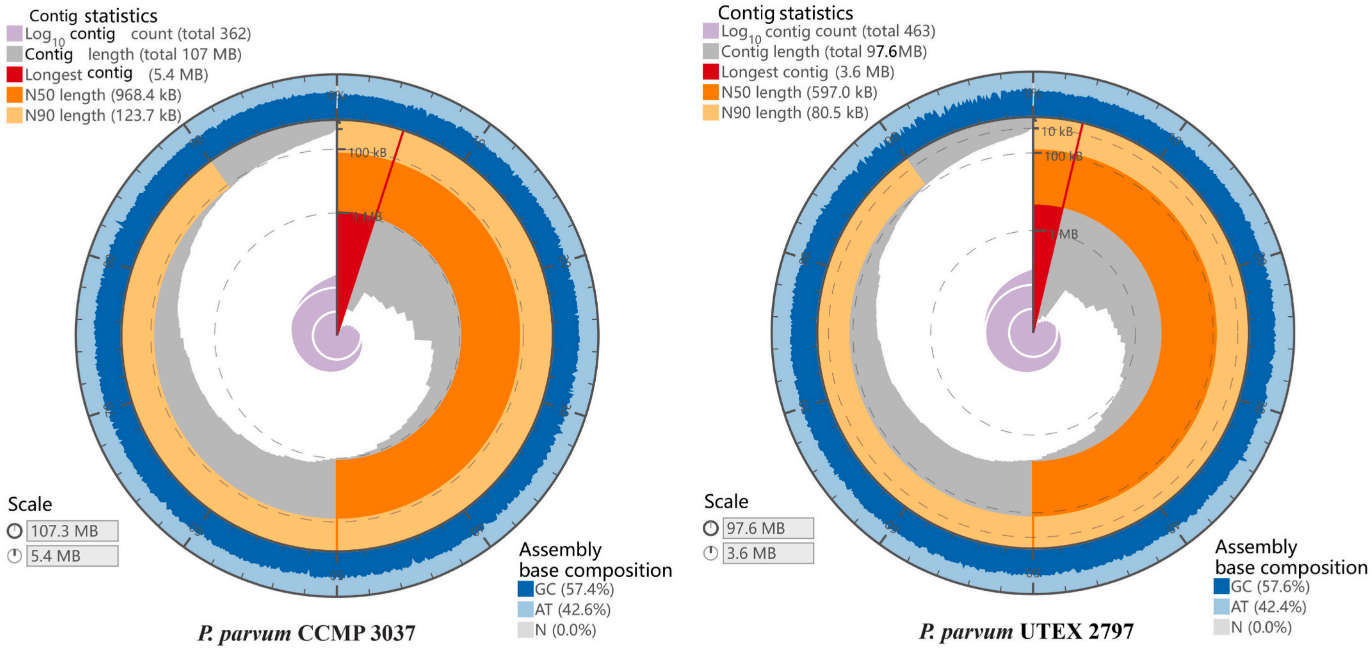


Fig. 1. The genome characteristic of *P. parvum* CCMP 3037 and UTEX 2797.

the size of that of the herein presented UTEX 2797 (HiFi) (Wisecaver et al., 2023), potentially attributed to the presence of two haploid genome sequences in UTEX 2797 (ONT). Genomic alignment revealed an approximately 1:2 relationship, with some regions being absent in UTEX 2797 (ONT) (Fig. S3). The synteny and collinearity analyses revealed a strong 2:1 syntenic gene relationship between UTEX 2797 (ONT) and UTEX 2797 (HiFi), as depicted in the genomic plot (Fig. 2b). Additionally, a robust 1:1:1 synteny pattern was observed among the genes of 12B1, CCMP 3037, and UTEX 2797 (HiFi) (Fig. 2c).

The GC content of the two *P. parvum* strains was 57.4 % and 57.6 %, which is lower than that of the five published haptophyte genomes (58.7 %–73.3 %) (Hulatt et al., 2021), but slightly higher than *Pavlovales* sp. (55.2 %) (Fig. 3a). The GC distribution revealed significant differences among haptophytes (Fig. 3a). The large variation of GC content within the haptophytes of nearly 15 % suggests a broad genetic diversity within this algal group, possibly due to evolutionary adaptation (Smarda et al., 2014).

The nuclear haptophyte genomes that have been published prior to 2018 were sequenced using short reads or Sanger sequencing, with the exception of *D. lutheri* (sequenced using PacBio CLR model) (Hulatt et al., 2021). Only a few haptophyte genomes were available in the NCBI database, and no *Prymnesium* plastid genomes have been included to date. In contrast, the herein presented genomes were sequenced with PacBio HiFi long reads and represent one of the highest contiguous sequences with lower number of contigs to date. The N50 scaffold of the two newly HiFi sequenced strains was slightly higher than that of the published genomes primarily based on nanopore sequencing (Wisecaver et al., 2023). The genome contiguity of 13 Illumina-based genomes was low (Wisecaver et al., 2023). Long read genome assemblies increase the contiguity metrics (by average contig N50 length) by a dozen-fold in comparison to the previous short read-based genomes (Roberts et al., 2020). The circular sequences of plastid genomes from both strains of *P. parvum* were successfully assembled with a total length of 107,830 bp and 107,831 bp for CCMP 3037 and UTEX 2797, respectively. While the plastid genomes of the two *P. parvum* strains were similar in length, there was a significant difference in size between their assembled nuclear genomes, with a discrepancy of approximately 10 Mb. This suggests that the plastid genomes are more conserved than their nuclear counterparts are.

3.2. *Comparison of repeat and gene features*

With homology-based and *de novo* methods, 37.18 Mb and 29.06 Mb of repetitive sequences were identified in CCMP 3037 and UTEX 2797, respectively (Table S5). After merging and filtering overlapping sequences, a total of 29.7 % of repetitive sequences with approximately half being LTRs (long-terminal repeats) were identified in the CCMP 3037 nuclear genome (Table S6). In UTEX 2797, a total of 24.8 % (24.16 Mb) of repetitive sequences were detected with 14.4 % being LTRs (14.05 Mb LTR) (Table S7). The repeat content of *P. parvum* is a little higher than that of the *D. lutheri* genome (22.9 %), but significantly lower than that of the *E. huxleyi* genome (64 %) (Read et al., 2013). All current haptophyte nuclear genomes except the large *E. huxleyi* genome have a low percentage of repetitive sequences in comparison to land plants, which generally have >40 % of repetitive sequences. As in seed plants, the genome size of *P. parvum* is positively correlated with the repetitive content, but repeat turnover is around 10 Gb (Novak et al., 2020). Previously, repetitive sequences were labeled as “junk”, however, a lot of studies revealed that these elements effect biological function and play important roles in evolution (Gemmell, 2021).

A total of 20,578 and 19,426 protein-coding genes were annotated for CCMP 3037 and UTEX 2797 combing homology-based, RNA-Seq data-based and ab initio prediction methods (Table 1). The total amount of protein-coding nucleotides was 33.22 Mb and 31.02 Mb representing 30.9 % and 31.8 % of the CCMP 3037 and UTEX 2797 genomes, respectively. In contrast, roughly 61.2 % and 40 % of the genomes of the smaller haptophyte species *D. lutheri* (43.5 Mb) and *Chrysochromulina tobini* (59 Mb) encode proteins (Hovde et al., 2015; Hulatt et al., 2021). A small genome usually means that the repeat content is low such as in the green algae *Prototheca wickerhamii*, which has a genome size below 20 Mb with 3 % repetitive sequences (Guo et al., 2022). The comparison of gene features in mRNA (gene), exon, and intron revealed similar distribution patterns among the two strains of *P. parvum*, but significantly different ones from the patterns found in *C. tobini*, *D. lutheri*, *E. huxleyi*, and *Pavlovales* sp. (Fig. 3b-e). The larger peaks in gene and intron length observed in *P. parvum* suggested higher quality genomes compared to the other four haptophyte genomes. The analysis of haptophyte genome structures and gene features patterns revealed significant differences, which were also observed between pennate and centric

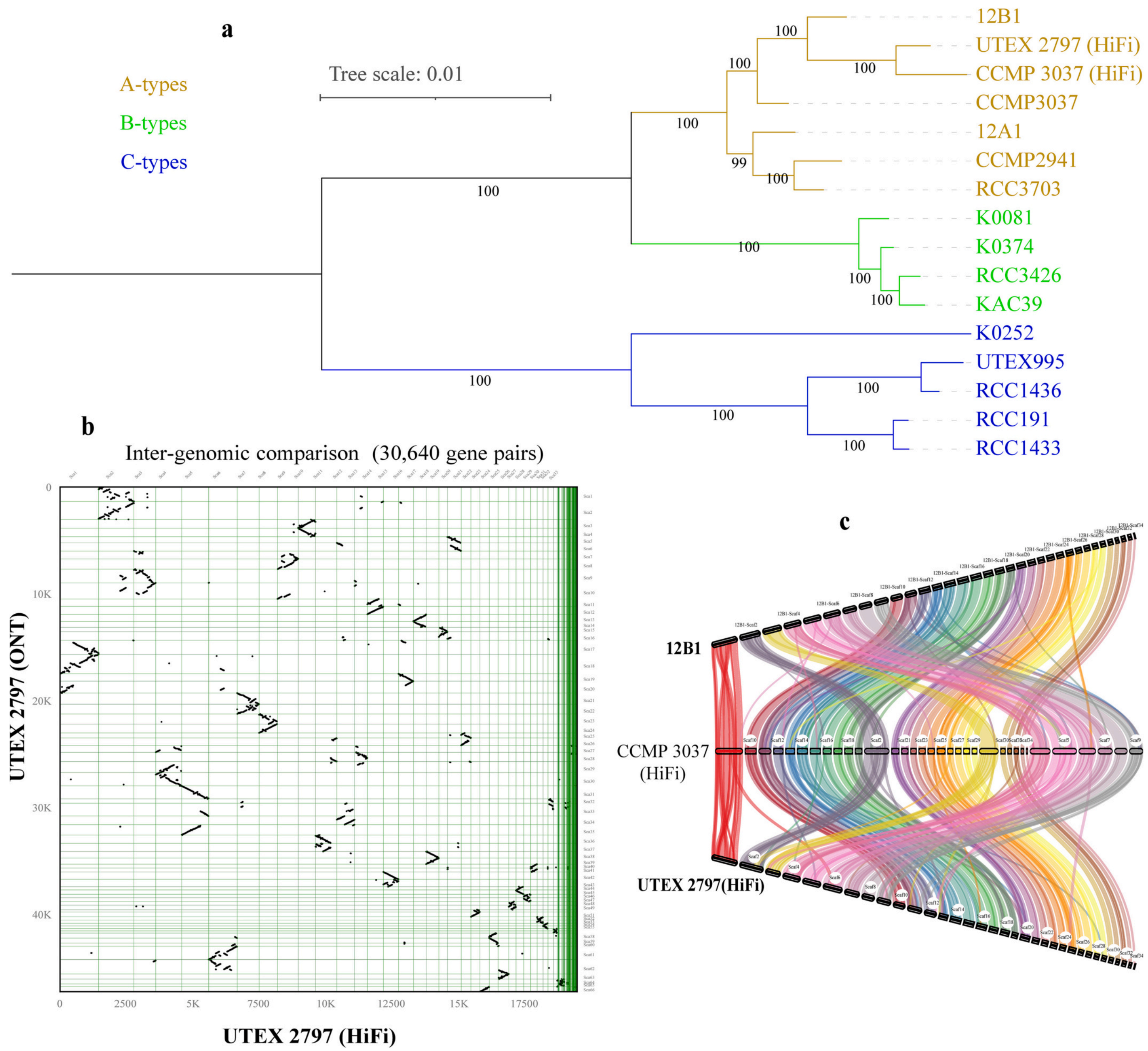


Fig. 2. The comparison among different *P. parvum* strains. a: the species phylogenetic tree based on single copy genes. b: Macrosynteny of two UTEX 2797 genomes. c: synteny plot of the genomes of 12B1, CCMP 3037 (HiFi) and UTEX 2797 (HiFi).

diatoms (Bowler et al., 2008). The quality of the gene annotation was evaluated with the BUSCO V5 analysis based on eukaryota_odb10 (a total of 255 gene set). 77.2 % and 81.2 % of genes were evaluated to be complete for the predicted genes in CCMP 3037 and UTEX 2797, respectively (Table 1 and Table S8), thus representing a high quality close to that of the *D. lutheri* genome annotation (80.80 %). The complete BUSCOs using the "eukaryote_odb10" database of all the haptophyte genomes were nearly below 80 % as in four diatom species (*P. tricornutum*, *F. cylindrus*, *C. nana*, and *C. cryptica*) (Roberts et al., 2020). However, the complete BUSCOs of the long-read assembled *P. parvum* genomes had a higher value compared to the other four haptophyte genomes (from 51.8 % to 72.9 %) based on short-read or Sanger sequencing data (Hulatt et al., 2021). The genomic characteristics and gene features exhibit remarkable similarity across different strains of *P. parvum*. To confirm the strain types of CCMP 3037 and UTEX 2797 according to Binzer et al. (2019), the available 15 strains of *P. parvum* were used for comparative analysis. CCMP 3037 and UTEX 2797 were classified as A-type using both all the ortholog groups and single-copy gene orthologs (Fig. S4 and Fig. 2a).

The protein-coding genes were compared to protein sequences in seven common databases (NR, Swissprot, KEGG, KOG, TrEMBL, Interpro and GO). The KEGG enrichment showed that most of the genes were related with metabolic pathways (Figs. S5 and S6). A total of 655 and 671 genes encoded for secondary carbohydrate metabolism in the two strains agreeing with their important role in global carbon cycle. 149 and 130 genes were enriched in metabolism of terpenoids and polyketides in the two strains, which are significantly more than found in *P. wickerhamii* (~30) (Guo et al., 2022). Hundreds of genes of amino acid and lipid metabolism were found in the *P. parvum* genomes. As primary metabolic products, fatty acids are essential membrane constituents and energy sources. Polyketides are proposed to have a common evolutionary origin with fatty acids in eukaryotes (Kohli et al., 2016). The polyketides can function as toxins and are synthesized not only for chemical defense, but also to immobilize motile prey through the use of

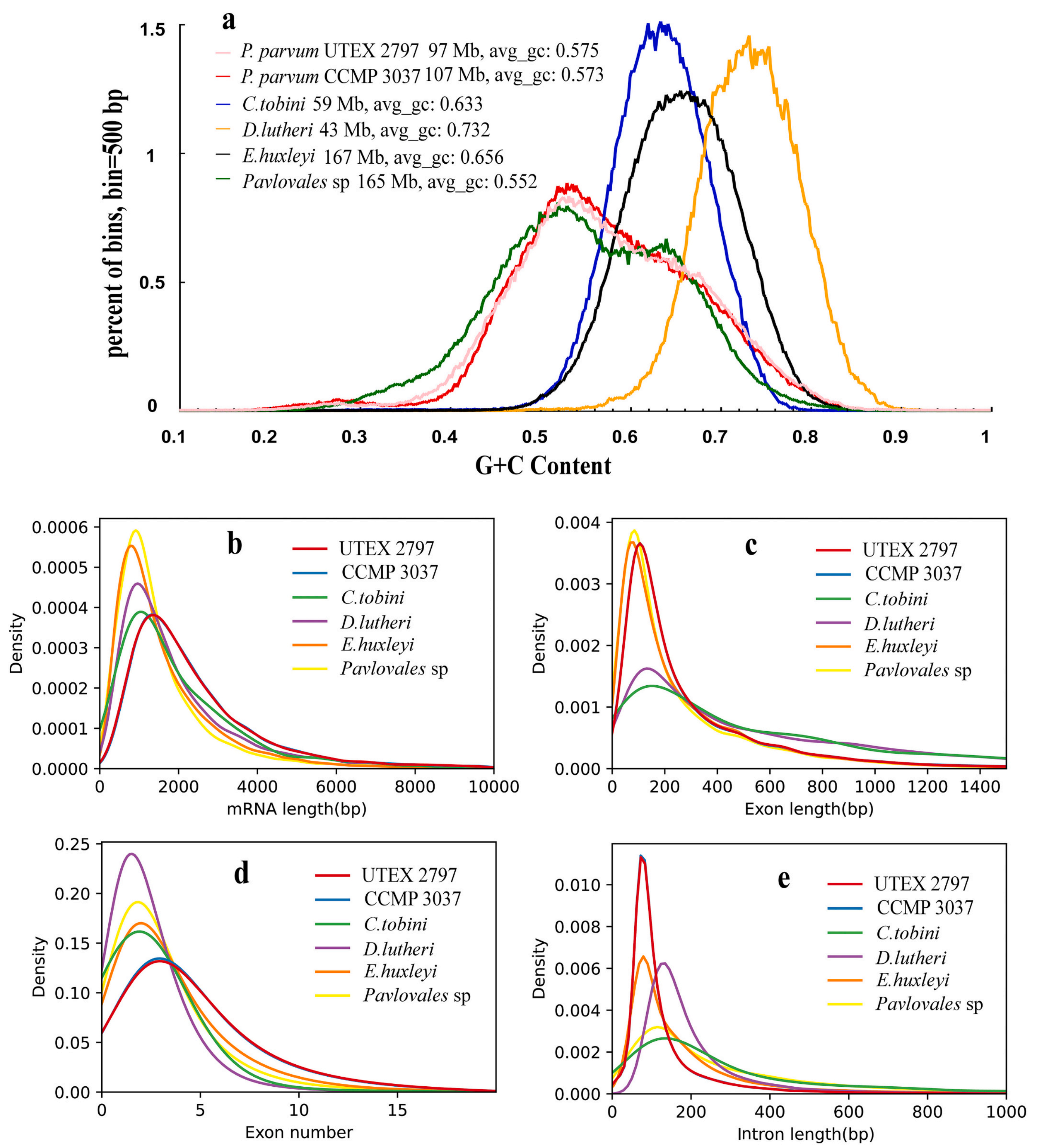


Fig. 3. The GC content and gene features of *P. parvum* CCMP 3037 and UTEX 2797 and four published haptophyte species. a: GC plot of sliding windows for haptophyte species. b: mRNA length. c: exon length. d: exon number. e: intron length.

prymnesins, which contribute to the adverse effects of harmful algal blooms (HABs) (Skovgaard and Hansen, 2003). A total of 8094 and 7950 genes were annotated in the five databases as shown in the Venn diagrams (Figs. S7 and S8), indicating significant differences in functional annotation across different databases. This highlights the need for further algal research. The functional annotation of protein coding genes showed that 85 % of genes were annotated in the two newly sequenced *P. parvum* genomes (Table S9). However, <50 % of genes could be annotated in Swissprot, KEGG, KOG and GO. The functional annotation of green algae is higher than 95 % with representative in *P. wickerhamii* (Guo et al., 2022) and 94.7 % in *Prasinoderma colonial* (Li et al., 2020). The relatively low percentage of functional annotation calls for further investigation, such as the inclusion of more genome data or conducting functional verification experiments. In the two plastid genomes, a total number of 208 genes were annotated including 121 CDS encoding photosystems, ATP synthases, RNA polymerases and other proteins, as well as 28 ribosomal RNA (rRNA) and 59 transfer RNA (tRNA) in both strains (Figs. S9 and S10). In addition, two inverted repeats were also identified for the two plastid genomes.

3.3. Analysis of gene families and phylogeny

To reveal insights into the evolution of *P. parvum*, the protein-coding genes of the two *P. parvum* strains and 13 representative algal species, including four haptophytes (*C. tobini*, *D. lutheri*, *E. huxleyi*, *Pavlovales* sp.), three diatoms (*Thalassiosira pseudonana*, *Phaeodactylum tricornutum*, *Aureococcus anophagefferens*), two green algae (*Chlamydomonas reinhardtii* V5, *Chlorella desiccate*), two dinoflagellates (*Fugacium kawagutii*, *Symbiodinium* sp.), the red alga *Porphyridium purpureum* and the glaucophyte *Cyanophora paradoxa*), were used to identify and compare groups of orthologous genes. The 311,893 genes originating from the 15 genomes were grouped into a total of 40,738 gene families (Table S10 and Fig. S11) with 7523 (*P. purpureum*) up to 34,062 (*E. huxleyi*) clustered gene families per genome (Table S10).

We identified 2876 homologous gene families shared by these six genomes, and 4391 gene families that were specific to *P. parvum* CCMP 3037 and UTEX 2797 when comparing the two *P. parvum* strains with the four other haptophytes (Fig. S12). 117 and 85 gene families were unique to CCMP 3037 and UTEX 2797, respectively (Fig. S11). The abundance of gene families specific to *P. parvum* suggests significant differences compared to the four published haptophyte species. Among the 4391 gene families specific to *P. parvum*, the most significantly enriched pathway was associated with plant-pathogen interaction, which was consistent with the ecological and physiological characteristics of this organism (Driscoll et al., 2023). A total of 844 out of the 40,738 identified orthogroups were common to all 15 species, and only nine were single copy orthologs found in all genomes. A previous study on algal evolution found only five single-copy orthologs in 12 algal species (Hulatt et al., 2021). It could be suggested that these algae are polyphyletic, resulting in only a few single copy orthologs. Gene replication and loss events occur frequently leading to single copy ortholog genes being rare, especially when comparing very diverse, primitive algal species. All orthogroups were used to infer single copy homologous groups and 979 genes were used to generate a species tree and phylogenomic reconstruction. The phylogenetic tree showed that the two newly sequenced *P. parvum* strains were sister with *C. tobini* and in a single clade with *E. huxleyi* (Fig. S13). The topology of the evolutionary tree for the haptophyte species was consistent with previous work based on 1152 orthogroups (Hulatt et al., 2021).

The diatoms, but not the dinoflagellates, grouped with the haptophytes, which is in conflict with that dinoflagellates and haptophytes are in a monophyletic clade based on plastid *psbA* gene phylogeny (Yoon et al., 2002). This could potentially provide phylogenomic evidence of separate origins of the nuclear and plastid genomes. With calibrated times, the *Prymnesium* genus diverged from *C. tobini* ~ 498.9 million years ago (Mya) and the two strains of *P. parvum* separated from each other approximately 21.1 Mya (Fig. 4). The haptophytes as a sister of diatoms and dinoflagellates diverged approximately 868.2 Mya. To clarify the adaptation of different species, gene family expansions and contractions were performed showing 359 gene families expanding and 724 gene families contracting in the species of *P. parvum*. Of the 359 expanded gene families, only 7 gene families are significantly ($p \leq 0.01$) expanded including genes enriched in starch and sucrose metabolism, followed by sulfur metabolism, energy metabolism and ABC transporters (Table S11). An analysis of gene expression also found that the *P. parvum* up-regulated the ABC transporters as a response to iron depletion (Mamunur Rahman et al., 2014). These expanded genes were implicated in responding to alterations in carbon and energy availability, as well as sulfur emissions. A small subset of significantly contracted genes (10 in total) was identified (p ≤0.01), with 8 of these genes involved in NHL

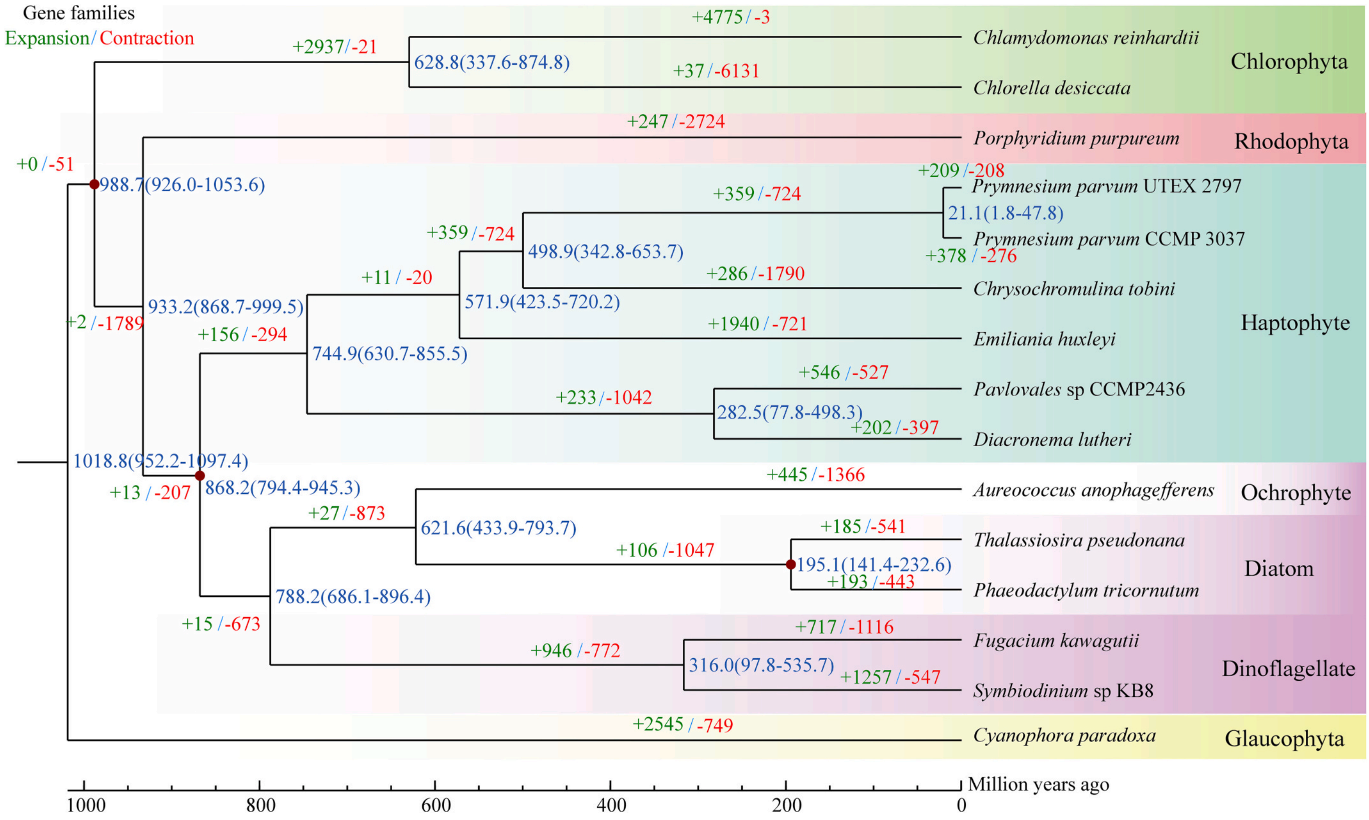


Fig. 4. Phylogenetic tree and gene family expansion and contraction for *Prymnesium parvum.*

repeat (IPR001258).

To confirm the patterns of functional diversity, a principal component analysis (PCA) was applied to analyze the number of InterPro domains (top 10 % as conserved protein domains) in the algal species used in the phylogenetic tree. The PCA results showed that the diatoms and haptophytes formed a discrete cluster that was separate from other algae (Fig. 5a), with *P. purpureum* being an outlier. Heatmaps generated from the top 10 % of InterPro entries showed that *P. purpureum* and *C. paradoxa* formed a clade with the dinoflagellates (Fig. 5b). Given the distinct distributions of rhodophytes, dinoflagellates and chlorophytes in the PCA analysis, the data suggests that significant functional diversification had occurred since the major lineages split. The PCA and heatmap analysis revealed that haptophytes exhibit greater similarity with diatoms and ochrophyte, as opposed to dinoflagellates. These findings contradict the phylogenetic tree, which suggests that diatoms, ochrophyte, and dinoflagellates belonged to the same clade. According to traditional classification, these three types are classified into stramenopiles. In the phylogenetic tree, haptophytes are closely related to and serve as an outgroup of the stramenopiles, suggesting a shared ancient origin. However, diversification occurred within the clade of stramenopiles. The genome size of dinoflagellates (>800 Mb) may have undergone expansion through repetitive elements while the other genomes have a size below 200 Mb. The gene number in dinoflagellate genomes is also significantly higher in comparison to other species within the same clade. Therefore, the diatoms and ochrophytes exhibited a higher similarity with haptophytes based on the InterPro domains. Given the importance of polyketides related to HABs, we focused on significant expansions of polyketide-related InterPro domains in *Symbiodinium* sp., *F. kawagutii*, *E. huxleyi*, *P. parvum* CCMP 3037 and UTEX 2797. Dozens of genes related to polyketide biosynthesis were identified in *P. parvum*, while only a limited number were detected in other haptophytes except for *E. huxleyi*, which harbored 45 of such genes. This result suggests that PKS may be involved in the formation of blooms caused by *P. parvum* and *E. huxleyi*, while *C. tobini, D. lutheri* and *Pavlovales* sp. do not cause bloom (Camara Dos Reis et al., 2023). The involvement of more PKSs in *P. parvum* may contribute to the production of toxic substances, while their synthesis in *E. huxleyi* may exhibit a plethora of pharmacologically significant activities (Vicente et al., 2021). The two dinoflagellates also showed expanded polyketide-related InterPro, which has been reported previously (Kohli et al., 2016).

3.4. Analysis of polyketide synthase genes and their differential expression

Prymnesins are large polyketides, meaning that they are biosynthesized by polyketide synthases (PKSs) (Anestis et al., 2021). Analysis with InterPro and Pfam scans identified 27 and 22 PKS genes in the genomes of CCMP 3037 and UTEX 2797, respectively (Fig. 6a). In *E. huxleyi*, 23 PKS genes were identified, while in the other haptophyte species only one or two PKS genes were detected. A slightly higher number of PKS genes can be annotated for the dinoflagellate species, *F. kawagutii* (29) and *Symbiodinium* sp. (32). The other algal genomes analyzed herein, including chlorophytes, diatoms, rhodophytes and glaucophytes, exhibited a limited number or absence of PKS genes (Fig. 6a). These findings demonstrates the broad yet inconsistent distribution of PKSs in algae, suggesting diverse origins. Only 3 and 2 biosynthetic gene clusters (BGCs) were detected for CCMP 3037 and UTEX 2797, respectively, using plantiSMASH (Table S12). A total of 28 and 24 BGCs for secondary metabolites were identified in CCMP 3037 and UTEX 2797, respectively, using antiSMASH 7 and antiSMASH fungal version. The sequences of 22 BGCs showed high similarity between the two strains, while five unique BGCs were found in CCMP 3037 and two unique BGCs were found in UTEX 2797 (Table S13). The limited inclusion of algal data in the plantiSMASH database suggests that its secondary metabolite prediction may not be applicable to algae. In CCMP 3037, the nine identified PKS BGCs comprised 13 biosynthetic or biosynthesis-related PKS genes including four type I and type III PKS and one PKS-NRPS-like hybrid (Table S13). In UTEX 2797, there were seven PKS BGCs containing a total of ten biosynthetic or biosynthetic-related PKS genes, including two type I and three type III PKSs, one hybrid

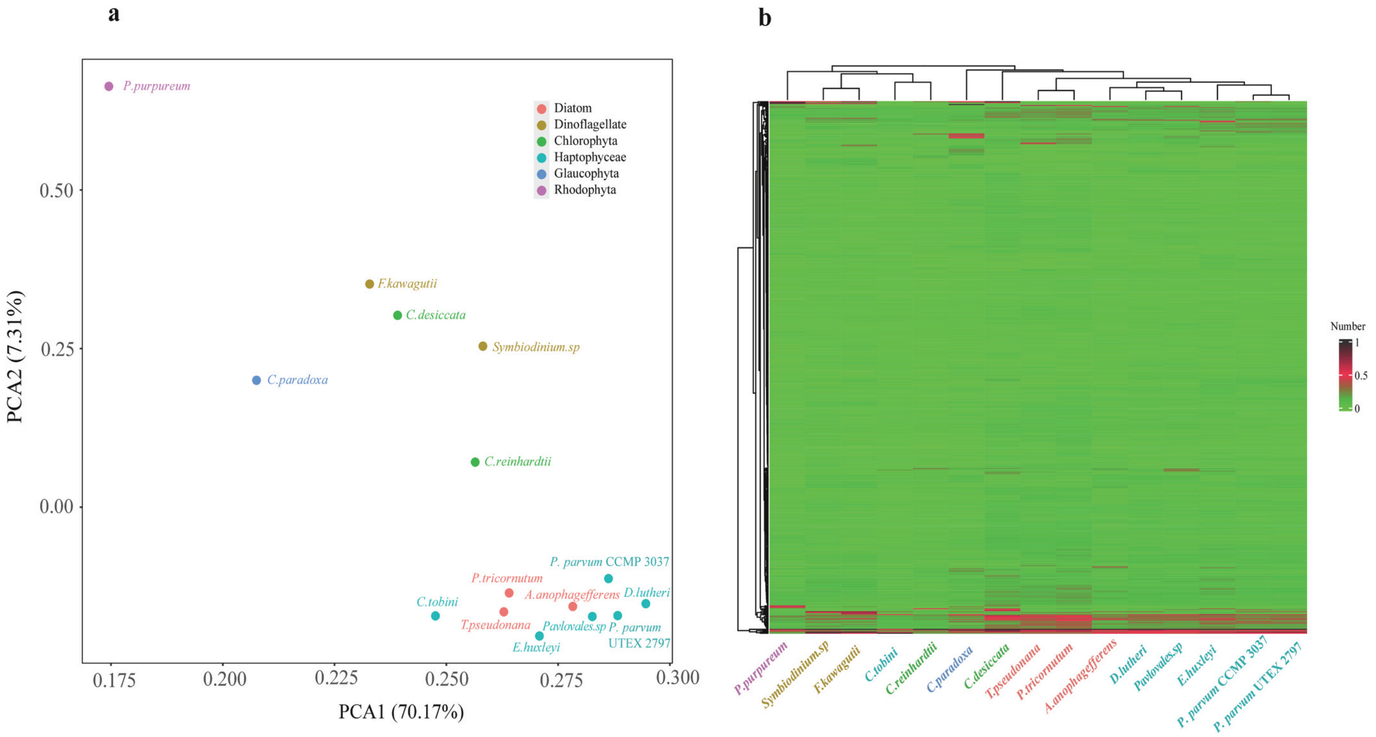


Fig. 5. PCA and heatmap analysis of top 10 % InterPro domains in 15 algal genomes. a: PCA analysis of top 10 % of InterPro domains. b: Heatmap of top 10 % the number of InterPro domains.

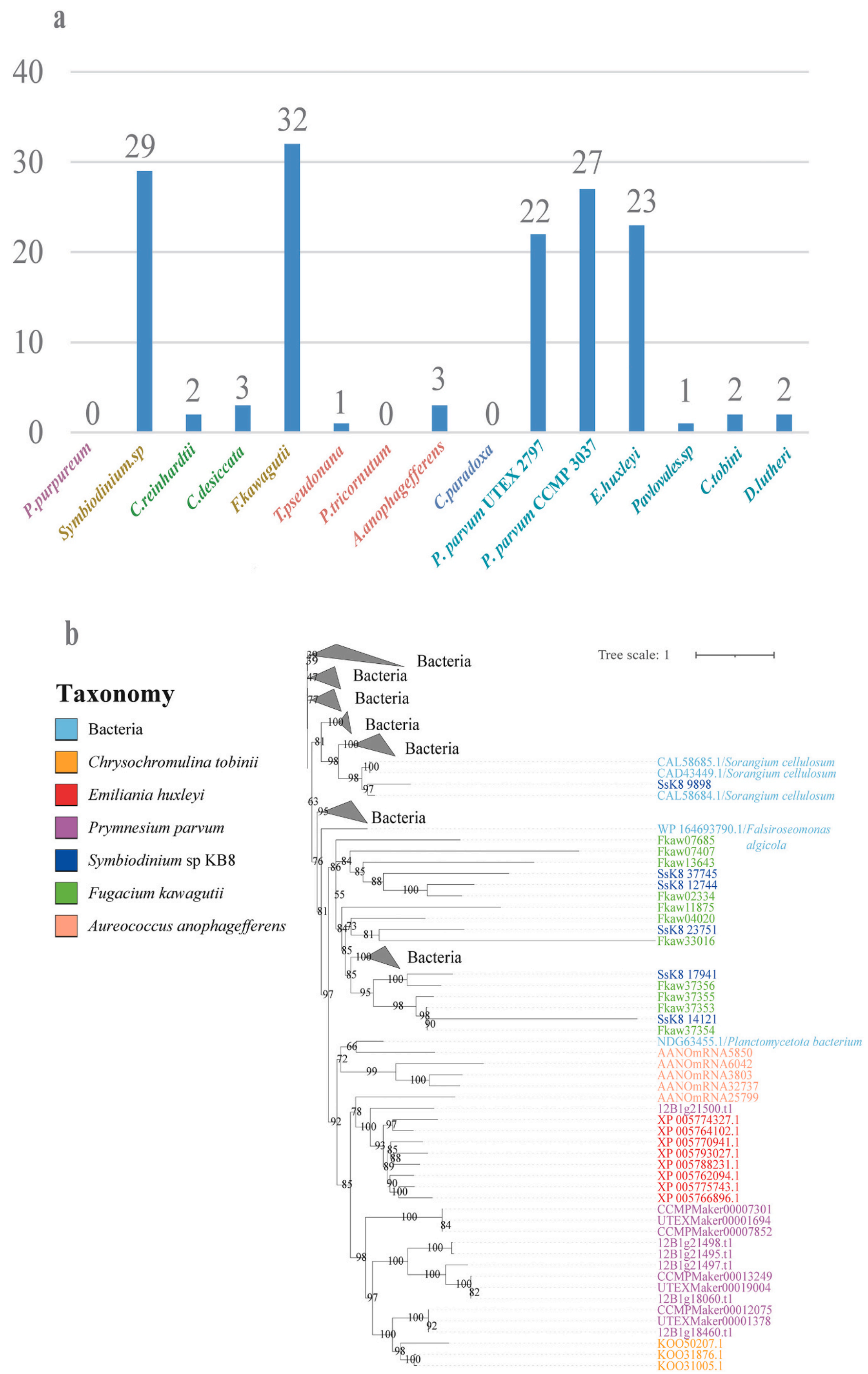


Fig. 6. Putative PKS genes in *P. parvum*. a: Number of PKS genes identified in Pfam. b: Phylogenetic tree of putative HGT PKS genes.

hglE-ketosynthase and type I PKS, as well as one PKS-AT (Table S13). In *P. parvum* CCMP 3037, three type III PKS genes are co-located in region 4.1 BGC, while two type I PKS genes are found to be co-located in two separate BGCs (region 103.1 and region 124.1). Similar gene locations were also observed in *P. parvum* UTEX 2797. Without a reference genome for *P. parvum*, a previous study utilized transcriptomic data to identify 37–109 type I PKSs across nine distinct strains (Anestis et al., 2021), which is significantly more than identified in the genome sequences. However, only 15 PKSs were found in another study based on six transcriptomes (Kohli et al., 2016). The discrepancy in numbers may be attributed to differences in databases or identification methods utilized. The abundance of PKSs identified in transcriptomes suggests that a single PKS gene may generate multiple distinct transcripts through alternative splicing or other post-transcriptional processes. The higher number of PKSs in four haptophytes (*E. huxleyi, P. parvum* and *Gephyrocapsa oceanica and Isochrysis galbana*) has been demonstrated previously (Kohli et al., 2016). The PKS gene families exhibited significant expansion in dinoflagellates (*Symbiodinium* sp. *and F. kawagutii*) and haptophytes (*P. parvum and E. huxleyi*), which was consistent with previous studies (Kohli et al., 2016). Furthermore, the PKS genes of *P. parvum* CCMP 3037 and UTEX 2797 showed a high degree of overlap with LTRs or Tandem Repeats, with 24 out of 27 and 21 out of 22 genes exhibiting this characteristic respectively.

To understand the expression of PKS genes in *P. parvum*, three published transcriptomes of the strain UOBS-LP0109 studied under different growth conditions were utilized: 1) an axenic nutrient replete treatment (axenic), 2) a nitrogen limited (N-limited) treatment, and 3) a phosphorus limited (P-limited) treatment (Keeling et al., 2014) and compared to the newly assembled reference genomes. After filtering out low-quality reads, 42.48 M (axenic), 23.52 M (N-limited) and 37.52 M (P-limited) reads were used for subsequent analysis. For the axenic, N-limited, and P-limited treatments, the percentages of uniquely mapped reads to the UTEX 2797 reference genome were 84.42 %, 87.43 % and 78.71 %, respectively. The sequencing saturation curve indicated that the downloaded reads were sufficient for conducting gene expression analysis (Fig. S14). Using KEGG enrichment of the 1316 differential genes shared between all three samples (Fig. S15), the highest enrichment pathway was biosynthesis of amino acid and followed by carbon metabolism (Fig. S16), which was consistent with the involvement of many genes in the tricarboxylic acid cycle as reported in previous studies (Liu et al., 2015). The FPKM (Fragments per kilo base per million mapped reads) showed that most of the PKS-related genes (type I and III PKS and PKS hybrids) were highly expressed except Maker00001694, Maker00016521 and Maker00020608 (Tables S14) in the nutrient replete treatment. Most of the PKS-related genes exhibit significant differential expression except for Maker00001694 (FPKM below 1 in all samples) and Maker00016497 (high FPKM but <2-fold change in differential expression) (Tables S14 and Fig. S17). The expression levels of two PKS genes, Maker00012819 and Maker00017916, were found to be higher in the N-limited treatment compared to nutrient replete (axenic) and P-limitation. The expression of PKSs (Maker00012477 and Maker00016497) was found to be significantly downregulated under both nitrogen and phosphorus limitation. The expression patterns of PKS genes exhibited inconsistency in the P-limited treatment, suggesting a distinct regulatory model in response to varying nutrient availabilities. The majority of PKS genes exhibited higher expression levels in the nutrient replete treatment compared to those in the nutrient limited treatments, except for three specific PKS genes (Fig. S17). The expression of the PKS gene, Maker00020748, was found to be highly expressed in the control, suggesting a potentially complex pattern of gene expression that deviates from the expected regulation during bloom formation.

3.5. Horizontal gene transfer

Horizontal gene transfer (HGT) is well studied and widely recognized in bacteria and archaea (Soucy et al., 2015). Despite its low probability of occurrence, HGT has been demonstrated to play a significant role in the evolution of algae by facilitating the acquisition of adaptive functions (Van Etten and Bhattacharya, 2020). A total of 10,860 genes were identified as putative HGT events from prokaryotes in the target 16 algae genomes (Table S15). A high number (1009, 2068 and 1476) and percentage (2.26 %, 5.39 % and 5.96 %) of HGT genes were identified in dinoflagellates and glaucophytes. Lower, but still high number (512, 599) and percentage (4.94 %, 5.4 %) of HGT events have been found in diatoms. The chlorophytes exhibited the lowest percentage and number of HGT events (115 and 180 with 1.2 % and 0.92 %). The significance of putative gene numbers in different clades of algae further demonstrates the high diversity among them. Haptophytes (*E. huxleyi* (1125 with 3 %) and *P. parvum* (530–760 with 2.17 %–3.91 %) had higher number of HGT and percentages of events compared to the green algae, and other haptophytes including *Pavlovales* sp., *C. tobini* and *D. lutheri*. This phenomenon is also consistent with the higher number of PKS in *E. huxleyi* and *P. parvum*. Enrichment GO analysis revealed that both CCMP 3037 and UTEX 2797 exhibited significant enrichment of genes involved in catalytic activity (GO:0003824), particularly hydrolase activity (GO:0016787), as well as catalytic activity acting on nucleic acids (GO:0140640) and RNA (GO:0140098). Additionally, sulfuric ester hydrolase activity (GO:0008484) and amidase activity (GO:0004040) were also enriched (Table S16). With KEGG enrichment analysis, numerous genes were found to be enriched in pathways related to biosynthesis of secondary metabolites, fatty acid metabolism, carbon metabolism and amino acid biosynthesis (Figs. S18 and S19). All these pathways are involved in polyketide biosynthesis. The putative genes were clustered using OrthoFinder, revealing that the first orthogroups are associated with PKS-related genes, including 12B1 (12), UTEX2797 (14), and CCMP 3037 (11) (Fig. S20). These genes included IPR003439 (ABC transporter-like), IPR000873 (AMP-dependent synthetase/ligase), and IPR009081 (Phosphopantetheine binding ACP domain).

Several putative HGT PKS genes from UTEX 2797, CCMP 3037 and 12B1 were identified as having been transferred from bacteria and grouped together in the same clade (Fig. 6b). Eight PKS genes of *E. huxleyi* were also identified as transferred from bacteria. The haplotype PKS genes were grouped together. Furthermore, 11 and 5 dinoflagellate HGT PKS genes were identified in *F. kawagutii* and *Symbiodinium* sp., respectively, located in two distinct branches of the phylogenetic tree. Therefore, the high number of PKSs in algae could have evolved or been acquired through various evolutionary processes including HGT e.g., from bacteria. The presence of multiple HGT PKSs in dinoflagellates*, E. huxleyi* or *P. parvum* suggests that HGT may play a significant role in facilitating adaptation to specific environments. The biosynthesis of polyketides and toxins in *P. parvum* and dinoflagellate species contributes to their ecological success (Verma et al., 2019). The PKS genes have undergone independent expansion events, such as those involving LTR or HGT.

3.6. Analysis of putative viruses

With the most diverse and abundant characteristics, viruses possess infectious properties that allow for horizontal transmission between individuals and across species (Breitbart and Rohwer, 2005). The presence of algal viruses has been demonstrated in host algal species, providing a new source of genetic material through endogenization (Moniruzzaman et al., 2020; Rozenberg et al., 2020; Short et al., 2020; Van Etten and Meints, 1999). A total of three and one putative virus genes were identified in CCMP 3037 and UTEX 2797, respectively. The Alien Index (AI) scores of these putative genes were higher than 55. The two strains share the putative virus genes, namely glycosyltransferase from *Tunisvirus fontaine* 2. All the putative viruses were classified as Varidnaviria, with a genome consisting of double-stranded DNA (dsDNA). Based on the phylogenetic tree, it can be inferred that putative gene Maker00016477 was likely to have been acquired from bacteria rather than a virus, as it clustered with bacterial clades (Fig. 7). All known viruses infecting haptophytes have been described as lytic viruses. A total of 12 hits were identified in association with the *E. huxleyi* virus, which was consistent with laboratory experiments where stable coexistence between surviving *E. huxleyi* hosts and their viruses was observed for over a year (Thyrhaug et al., 2003). The genome of the *E. huxleyi* virus, which is approximately 400 kb in size, belongs to the family Coccolithovirus (Nissimov et al., 2011). Similar virus HGT results have also been found in a recent study (Wisecaver et al., 2023). The two virus HGT genes showed high expression in replete, N-limited, and P-limited treatments. The gene (Maker00010126) exhibited the highest similarity with the hypothetical protein of *E. huxleyi* virus 203, and the Maker00010126 expression value (FPKM ~3998.45) was approximately 40-fold and 115-fold higher than that under N-limited and replete treatments, respectively. suggesting a high gene regulation under P-limitation. In here, we observed virus-host transfer or virus–host relationships, which was consistent with the findings of metatranscriptomics (Moniruzzaman et al., 2017) and genomic data (Wisecaver et al., 2023). Further research should be conducted on the function and mechanism of the virus.

4. Conclusions

Here, we utilized PacBio HiFi long reads to sequence and assemble

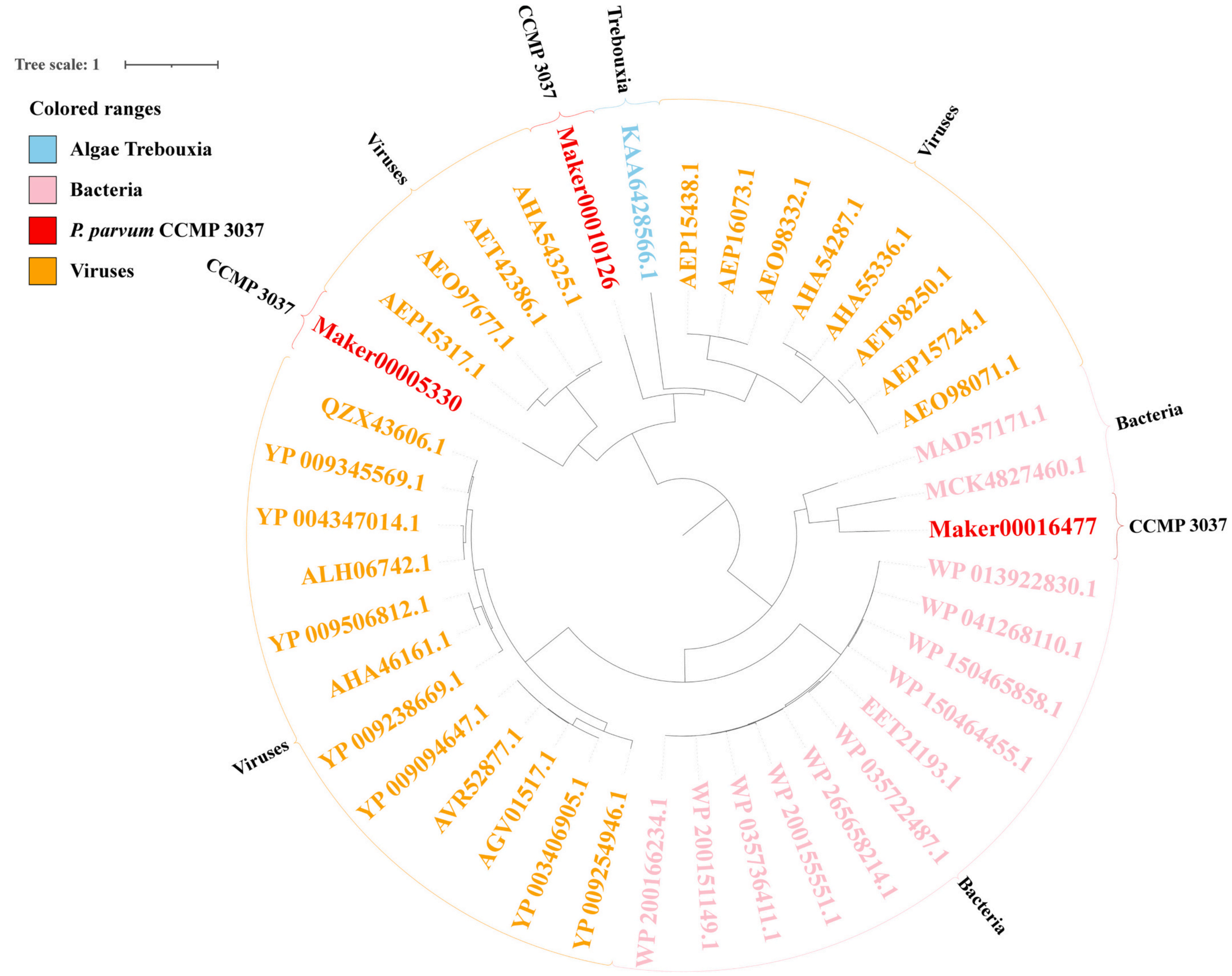


Fig. 7. Phylogenetic tree of putative HGT virus genes of *P. parvum* CCMP 3037.

two genomes of *P. parvum*, namely strains CCMP 3037 and UTEX 2797, resulting in high-quality sequences. The distinctive genomic features of *P. parvum* encompassed its genome size and GC content. The successful completion of the nuclear genome sequencing using HiFi long reads and the acquisition of a circular plastid sequence for *P. parvum* represents significant milestones in scientific research of haptophyte genomics, marking the first time such achievements have been accomplished. The numerous specific gene families in *P. parvum* underlined the genotypic difference in comparison to the four currently published haptophyte genomes. The phylogenetic analysis of 16 algal genomes grouped the haptophytes into the same clade, while dinoflagellate species were classified into a clade with the diatoms. The expanded gene families of *P. parvum* were enriched in starch and sucrose metabolism, followed by sulfur metabolism, energy metabolism and ABC transporters. The PCA analysis and heatmap of the number of InterPro domains suggested that a significant functional diversification had occurred in the Stramenopiles. The analysis of polyketide synthase genes revealed a substantial number of PKS genes in *P. parvum* and the extensive diversity may have resulted from horizontal gene transfer from bacteria. The identification of the HGT viral gene represented the initial observation of a virus-host transfer and highlights its potential functional role in *P. parvum*. In summary, our work provides a valuable genomic resource and information for elucidating the genetic basis of algal diversity, evolution, and metabolic modeling to gain insights into harmful algal bloom events.

CRediT authorship contribution statement

Jianbo Jian: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Supervision; Validation; Visualization; Roles/Writing - original draft; Writing - review & editing. Zhangyan Wu: Investigation, Writing - review & editing. Andrea Arisbé Silva Nuñez: Investigation; Methodology; Writing - review & editing. Xiaomin Zheng: Investigation, Writing - review & editing. Xiaohui Li: Investigation, Writing - review & editing.

Bei Luo: Investigation, Writing - review & editing. Yun Liu: Investigation, Writing - review & editing. Xiaodong Fang: Supervision, Formal analysis, Writing - review & editing. Christopher Workman: Supervision, Formal analysis, Writing - review & editing. Thomas Ostenfeld Larsen: Supervision, Writing - review & editing. Per Juel Hansen: Writing - review & editing. Eva C. Sonnenschein: Conceptualization, Project administration, Resources, Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that support the findings of this study have been deposited into CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) with accession number CNP0004284.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2023.168042.

References

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F.J., et al., 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 20, 224.

Anestis, K., Kohli, G.S., Wohlrab, S., Varga, E., Larsen, T.O., Hansen, P.J., et al., 2021. Polyketide synthase genes and molecular trade-offs in the ichthyotoxic species *Prymnesium parvum*. Sci. Total Environ. 795, 148878.

Anestis, K., Wohlrab, S., Varga, E., Hansen, P.J., John, U., 2022. The relationship between toxicity and mixotrophy in bloom dynamics of the ichthyotoxic *Prymnesium parvum*. Authorea Preprints.

Audic, S., Claverie, J.M., 1997. The significance of digital gene expression profiles. Genome Res. 7, 986–995.

Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6, 11.

Binzer, S.B., Svenssen, D.K., Daugbjerg, N., Alves-de-Souza, C., Pinto, E., Hansen, P.J., et al., 2019. A-, B- and C-type prymnesins are clade specific compounds and chemotaxonomic markers in *Prymnesium parvum*. Harmful Algae 81, 10–17.

Blaby-Haas, C.E., Merchant, S.S., 2019. Comparative and functional algal genomics. Annu. Rev. Plant Biol. 70, 605–638.

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M. H., et al., 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res. 49, W29–W35.

Blossom, H.E., Rasmussen, S.A., Andersen, N.G., Larsen, T.O., Nielsen, K.F., Hansen, P.J., 2014. *Prymnesium parvum* revisited: relationship between allelopathy, ichthyotoxicity, and chemical profiles in 5 strains. Aquat. Toxicol. 157, 159–166. https://doi.org/10.1016/j.aquatox.2014.10.006.

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

Breitbart, M., Rohwer, F., 2005. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 13, 278–284.

Camara Dos Reis, M., Romac, S., Le Gall, F., Marie, D., Frada, M.J., Koplovitz, G., et al., 2023. Exploring the phycosphere of *Emiliania huxleyi*: from bloom dynamics to microbiome assembly experiments. Mol. Ecol. 1–16.

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.

Caron, D.A., Lie, A.A.Y., Buckowski, T., Turner, J., Frabotta, K., 2023. The effect of pH and salinity on the toxicity and growth of the golden alga, *Prymnesium parvum*. Protist 174, 125927.

Carrier, G., Baroukh, C., Rouxel, C., Duboscq-Bidot, L., Schreiber, N., Bougaran, G., 2018. Draft genomes and phenotypic characterization of *Tisochrysis lutea* strains. Toward the production of domesticated strains with high added value. Algal Res. 29, 1–11.

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., et al., 2018. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience 7, 1–6.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H., 2021. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat. Methods 18, 170–175.

De Bie, T., Cristianini, N., Demuth, J.P., Hahn, M.W., 2006. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22, 1269–1271.

Dolja, V.V., Koonin, E.V., 2011. Common origins and host-dependent diversity of plant and animal viromes. Curr. Opin. Virol. 1, 322–331.

Dolja, V.V., Koonin, E.V., 2018. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res. 244, 36–52.

Driscoll, W.W., Wisecaver, J.H., Hackett, J.D., Espinosa, N.J., Padway, J., Engers, J.E., et al., 2023. Behavioural differences underlie toxicity and predation variation in blooms of *Prymnesium parvum*. Ecol. Lett. 26 (5), 677–691.

Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20, 238.

Feschotte, C., Gilbert, C., 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. 13, 283–296.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., et al., 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. U. S. A. 117, 9451–9457.

Franco, M.E., Hill, B.N., Brooks, B.W., Lavado, R., 2019. Prymnesium parvum differentially triggers sublethal fish antioxidant responses *in vitro* among salinity and nutrient conditions. Aquat. Toxicol. 213, 105214.

Gemmell, N.J., 2021. Repetitive DNA: genomic dark matter matters. Nat. Rev. Genet. 22, 342.

Granéli, E., Edvardsen, B., Roelke, D.L., Hagström, J.A., 2012. The ecophysiology and bloom dynamics of *Prymnesium* spp. Harmful Algae 14, 260–270.

Guillard, R.R., Ryther, J.H., 1962. Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (cleve) Gran. Can. J. Microbiol. 8, 229–239.

Guo, J., Jian, J., Wang, L., Xiong, L., Lin, H., Zhou, Z., et al., 2022. Genome sequences of two strains of *Prototheca wickerhamii* provide insight into the protothecosis evolution. Front. Cell. Infect. Microbiol. 12, 797017.

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

Hovde, B.T., Deodato, C.R., Hunsperger, H.M., Ryken, S.A., Yost, W., Jha, R.K., et al., 2015. Genome sequence and transcriptome analyses of *Chrysochromulina tobin*: metabolic tools for enhanced algal fitness in the prominent order prymnesiales (Haptophyceae). PLoS Genet. 11, e1005469.

Hovde, B.T., Deodato, C.R., Andersen, R.A., Starkenburg, S.R., Barlow, S.B., Cattolico, R. A., 2019. *Chrysochromulina*: genomic assessment and taxonomic diagnosis of the type species for an oleaginous algal clade. Algal Res. 37, 307–319.

Hulatt, C.J., Wijffels, R.H., Posewitz, M.C., 2021. The Genome of the Haptophyte *Diacronema lutheri* (*Pavlova lutheri*, Pavlovales): A Model for Lipid Biosynthesis in Eukaryotic Algae. Genome Biol. Evol. 13.

Jin, J.J., Yu, W.B., Yang, J.B., Song, Y., dePamphilis, C.W., Yi, T.S., et al., 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 21, 241.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., et al., 2014. The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 12, e1001889.

Kent, W.J., 2002. BLAT–the BLAST-like alignment tool. Genome Res. 12 (4), 656–664. https://doi.org/10.1101/gr.229202.

Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360.

Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37, 907–915.

Kohli, G.S., John, U., Van Dolah, F.M., Murray, S.A., 2016. Evolutionary distinctiveness of fatty acid and polyketide synthesis in eukaryotes. ISME J. 10, 1877–1890.

Koid, A.E., Liu, Z., Terrado, R., Jones, A.C., Caron, D.A., Heidelberg, K.B., 2014. Comparative transcriptome analysis of four prymnesiophyte algae. PloS One 9, e97801.

Korf, I., 2004. Gene finding in novel genomes. BMC Bioinformatics 5, 59.

Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L., Pertea, M., 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20, 278.

La Claire . 2nd, J.W, 2006. Analysis of expressed sequence tags from the harmful alga, *Prymnesium parvum* (Prymnesiophyceae, Haptophyta). Mar. Biotechnol. (N.Y.) 8, 534–546.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Letunic, I., Bork, P., 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 49, W293–W296.

Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323.

Li, L., Wang, S., Wang, H., Sahu, S.K., Marin, B., Li, H., et al., 2020. The genome of *Prasinoderma coloniale* unveils the existence of a third phylum within green plants. Nat. Ecol. Evol. 4, 1220–1231.

Liu, Z., Koid, A.E., Terrado, R., Campbell, V., Caron, D.A., Heidelberg, K.B., 2015. Changes in gene expression of *Prymnesium parvum* induced by nitrogen and phosphorus limitation. Front. Microbiol. 6, 631.

Lundgren, V.M., Glibert, P.M., Graneli, E., Vidyarathna, N.K., Fiori, E., Ou, L., et al., 2016. Metabolic and physiological changes in *Prymnesium parvum* when grown under, and grazing on prey of, variable nitrogen:phosphorus stoichiometry. Harmful Algae 55, 1–12.

Mamunur Rahman, M., Azizur Rahman, M., Maki, T., Nishiuchi, T., Asano, T., Hasegawa, H., 2014. A marine phytoplankton (*Prymnesium parvum*) up-regulates ABC transporters and several other proteins to acclimatize with Fe-limitation. Chemosphere 95, 213–219.

Manning, S.R., La Claire, J.W., 2010. Prymnesins: toxic metabolites of the golden alga, *Prymnesium parvum* Carter (Haptophyta). Mar. Drugs 8, 678–704.

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

Medić, N., Varga, E., Waal, DbVd, Larsen, T.O., Hansen, P.J., 2022. The coupling between irradiance, growth, photosynthesis and prymnesin cell quota and production in two strains of the bloom-forming haptophyte, *Prymnesium parvum*. Harmful Algae 112, 102173.

Moniruzzaman, M., Wurch, L.L., Alexander, H., Dyhrman, S.T., Gobler, C.J., Wilhelm, S. W., 2017. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. Nat. Commun. 8, 16054.

Moniruzzaman, M., Weinheimer, A.R., Martinez-Gutierrez, C.A., Aylward, F.O., 2020. Widespread endogenization of giant viruses shapes genomes of green algae. Nature 588, 141–145.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.

Nissimov, J.I., Worthy, C.A., Rooks, P., Napier, J.A., Kimmance, S.A., Henn, M.R., et al., 2011. Draft genome sequence of the coccolithovirus EhV-84. Stand Genomic Sci. 5, 1–11.

Novak, P., Guignard, M.S., Neumann, P., Kelly, L.J., Mlinarec, J., Koblizkova, A., et al., 2020. Repeat-sequence turnover shifts fundamentally in species with large genomes. Nat. Plants 6, 1325–1329.

Rasmussen, S.A., Meier, S., Andersen, N.G., Blossom, H.E., Duus, J.O., Nielsen, K.F., et al., 2016. Chemodiversity of ladder-frame Prymnesin Polyethers in *Prymnesium parvum*. J. Nat. Prod. 79, 2250–2256.

Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., et al., 2013. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. Nature 499, 209–213.

Roach, M.J., Schmidt, S.A., Borneman, A.R., 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19, 460.

Roberts, W.R., Downey, K.M., Ruck, E.C., Traller, J.C., Alverson, A.J., 2020. Improved reference genome for *Cyclotella cryptica* CCMP332, a model for cell wall morphogenesis, salinity adaptation, and lipid production in diatoms (Bacillariophyta). G3 (Bethesda) 10, 2965–2974.

Rozenberg, A., Oppermann, J., Wietek, J., Fernandez Lahore, R.G., Sandaa, R.A., Bratbak, G., et al., 2020. Lateral gene transfer of anion-conducting Channelrhodopsins between green algae and giant viruses. Curr. Biol. 30 (4910-4920 e5).

Saha, S., Bridges, S., Magbanua, Z.V., Peterson, D.G., 2008. Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res. 36, 2284–2294.

Salcher, M.M., Pernthaler, J., Posch, T., 2011. Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria 'that rule the waves' (LD12). ISME J. 5, 1242–1252.

Short, S.M., Staniewski, M.A., Chaban, Y.V., Long, A.M., Wang, D., 2020. Diversity of viruses infecting eukaryotic algae. Curr. Issues Mol. Biol. 39, 29–62.

Skovgaard, A., Hansen, P.J., 2003. Food uptake in the harmful alga *Prymnesium parvum* mediated by excreted toxins. Limnol. Oceanogr. 48, 1161–1166.

Smarda, P., Bures, P., Horova, L., Leitch, I.J., Mucina, L., Pacini, E., et al., 2014. Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc. Natl. Acad. Sci. U. S. A. 111, E4096–E4102.

Soucy, S.M., Huang, J., Gogarten, J.P., 2015. Horizontal gene transfer: building the web of life. Nat. Rev. Genet. 16, 472–482.

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

Stanke, M., Schoffmann, O., Morgenstern, B., Waack, S., 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7, 62.

Svendsen, M.B.S., Andersen, N.R., Hansen, P.J., Steffensen, J.F., 2018. Effects of harmful algal blooms on fish: insights from *Prymnesium parvum*. Fishes 3, 11.

Taylor, R.B., Hill, B.N., Bobbitt, J.M., Hering, A.S., Brooks, B.W., Chambliss, C.K., 2020. Suspect and non-target screening of acutely toxic *Prymnesium parvum*. Sci. Total Environ. 715, 136835.

Taylor, R.B., Hill, B.N., Langan, L.M., Chambliss, C.K., Brooks, B.W., 2021. Sunlight concurrently reduces *Prymnesium parvum* elicited acute toxicity to fish and prymnesins. Chemosphere 263, 127927.

Thyrhaug, R., Larsen, A., Thingstad, T.F., Bratbak, G., 2003. Stable coexistence in marine algal host-virus systems. Mar. Ecol. Prog. Ser. 254, 27–35.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., et al., 2017. GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 45, W6–W11.

Van Etten, J., Bhattacharya, D., 2020. Horizontal gene transfer in eukaryotes: not if, but how much? Trends Genet. 36, 915–925.

Van Etten, J.L., Meints, R.H., 1999. Giant viruses infecting algae. Annu. Rev. Microbiol. 53, 447–494.

Verma, A., Barua, A., Ruvindy, R., Savela, H., Ajani, P.A., Murray, S.A., 2019. The genetic basis of toxin biosynthesis in dinoflagellates. Microorganisms 7.

Vicente, B., Matos, J., Gomes, R., Sapatinha, M., Afonso, C., Rodrigues, T., et al., 2021. Production and bioaccessibility of *Emiliania huxleyi* biomass and bioactivity of its aqueous and ethanolic extracts. J. Appl. Phycol. 33, 3719–3729.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., et al., 2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33, 2202–2204.

Wagstaff, B.A., Hems, E.S., Rejzek, M., Pratscher, J., Brooks, E., Kuhaudomlarp, S., et al., 2018. Insights into toxic *Prymnesium parvum* blooms: the role of sugars and algal viruses. Biochem. Soc. Trans. 46, 413–421.

Wagstaff, B.A., Pratscher, J., Rivera, P.P.L., Hems, E.S., Brooks, E., Rejzek, M., et al., 2021. Assessing the toxicity and mitigating the impact of harmful prymnesium blooms in eutrophic waters of the Norfolk broads. Environ. Sci. Technol. 55, 16538–16551.

Wisecaver, J.H., Auber, R.P., Pendleton, A.L., Watervoort, N.F., Fallon, T.R., Riedling, O. L., et al., 2023. Extreme genome diversity and cryptic speciation in a harmful algal-bloom-forming eukaryote. Curr. Biol. 33 (2246-2259 e8).

Xu, Z., Wang, H., 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35, W265–W268.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

Yoon, H.S., Hackett, J.D., Bhattacharya, D., 2002. A single origin of the peridinin- and fucoxanthin-containing plastids in dinoflagellates through tertiary endosymbiosis. Proc. Natl. Acad. Sci. U. S. A. 99, 11724–11729.

Yuan, L., Lu, H., Li, F., Nielsen, J., Kerkhoven, E.J., 2023. HGTphyloDetect: facilitating the identification and phylogenetic analysis of horizontal gene transfer. Brief. Bioinform. 24.