



**HAL**  
open science

# Chromosome-scale assembly of the streamlined picoeukaryote *Picochlorum* sp. SENEW3 genome reveals Rabl-like chromatin structure and potential for C4 photosynthesis

Patrick da Roza, Héloïse Muller, Geraldine Sullivan, Roy Walker, Hugh  
Goold, Robert Willows, Brian Palenik, Ian Paulsen

## ► To cite this version:

Patrick da Roza, Héloïse Muller, Geraldine Sullivan, Roy Walker, Hugh Goold, et al.. Chromosome-scale assembly of the streamlined picoeukaryote *Picochlorum* sp. SENEW3 genome reveals Rabl-like chromatin structure and potential for C4 photosynthesis. *Microbial Genomics*, 2024, 10 (4), 10.1099/mgen.0.001223 . hal-04658930

**HAL Id: hal-04658930**

**<https://hal.science/hal-04658930>**

Submitted on 22 Jul 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Chromosome-scale assembly of the streamlined picoeukaryote *Picochlorum* sp. SENEW3 genome reveals Rab1-like chromatin structure and potential for C<sub>4</sub> photosynthesis.

**Authors:** Patrick A. da Roza<sup>1,2</sup>, Héloïse Muller<sup>3</sup>, Geraldine J. Sullivan<sup>1,2</sup>, Roy S. K. Walker<sup>1,2</sup>, Hugh D. Goold<sup>1,4</sup>, Robert D. Willows<sup>2</sup>, Brian Palenik<sup>5</sup> and Ian T. Paulsen<sup>1,2,\*</sup>

<sup>1</sup> ARC Centre of Excellence in Synthetic Biology, Macquarie University, Sydney, NSW 2109, Australia

<sup>2</sup> School of Natural Sciences, Macquarie University, Sydney, Australia

<sup>3</sup> Institut Curie, PSL University, Sorbonne Université, CNRS, Nuclear Dynamics, 75005 Paris, France

<sup>4</sup> New South Wales Department of Primary Industries, Orange, NSW 2800, Australia

<sup>5</sup> Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093-0202, USA

\* Corresponding author: [ian.paulsen@mq.edu.au](mailto:ian.paulsen@mq.edu.au)

## Keywords

Photosynthetic picoeukaryote, genomic assembly, chromatin structure, C<sub>4</sub> photosynthesis, environmental adaption, *Picochlorum*.

## Repositories and data availability.

Primary and haplotype genome assemblies have been made available at the National Centre for Biotechnology Information (NCBI) under the BioProjects PRJNA839791 and PRJNA839798, and Genbank Accession numbers CP120453 - CP120466 and CP120450 - CP120452 respectively. DNA sequencing and Hi-C read data have been made available at the NCBI Sequence Read Archive (SRA) under the BioProject: PRJNA839791 and BioSample: SAMN28555558. All other supplementary data may be found in the supplementary data spreadsheet.

## Abbreviations

CAM, crassulacean acid metabolism; CCM, CO<sub>2</sub>-concentrating mechanism; CDS, protein coding sequence; CTAB, Cetyltrimethyl ammonium bromide; EDTA, ethylenediaminetetraacetic acid; ONT, Oxford Nanopore Technologies; PFGE, pulsed field gel electrophoresis; *P. SENEW3*, *Picochlorum* sp. SENEW3; PPE, photosynthetic picoeukaryote; RuBisCO, Ribulose-1,5-bisphosphate carboxylase-oxygenase; TBE, Tris/Borate/EDTA; TE, transposable elements.

## Abstract

Genome sequencing and assembly of the photosynthetic picoeukaryotic *Picochlorum* sp. SENEW3 revealed a compact genome with a reduced gene set, few repetitive sequences, and an organised Rab1-like chromatin structure. Hi-C chromosome conformation capture revealed evidence of

possible chromosomal translocations, as well as putative centromere locations. Maintenance of a relatively few selenoproteins, as compared to similarly sized marine picoprasinophytes Mamiellales, and broad halotolerance compared to others in Trebouxiophyceae, suggests evolutionary adaptation to variable salinity environments. Such adaptation may have driven size and genome minimization and have been enabled by the retention of high number of membrane transporters. Identification of required pathway genes for both CAM and C<sub>4</sub> photosynthetic carbon fixation, known to exist in the marine mamiellale pico-prasinophytes and seaweed *Ulva*, but few other chlorophyte species, further highlights the unique adaptations of this robust alga. This high-quality assembly provides a significant advance in the resources available for genomic investigations of this and other photosynthetic picoeukaryotes.

## **Impact Statement**

This work provides a fully annotated chromosomal genome assembly of euryhaline pico-sized green algae. It identifies key genomic features including putative centromeres and chromosomal translocations, reduced selenoprotein complement, CAM and C<sub>4</sub> photosynthetic carbon fixation gene sets. This study adds to the field by identifying novel genomic aspects of this organisms, affords additional utility in genomic studies via the genome's complete functional annotation, and provides a basis for further investigation of algal genomic architecture.

## **Introduction**

The green algae are a highly successful group of photosynthetic eukaryotes within the green lineage (Viridiplantae). Comprising of two divergent clades, Chlorophyta and Streptophyta (which includes the charophyte algae and land plants); the green algae display a high level of phylogenetic, ecological, and morphological diversity [1, 2]. Rapid advances in DNA sequencing over the past decade have facilitated the genomic sequencing and analysis of numerous green algal species, enabling investigation of many biological facets of these organisms, including ecological adaptations [3, 4], photosynthetic evolution [5, 6], cellular mechanisms and functions [7, 8], evolutionary phylogeny [2, 9], and the biosynthesis of valuable metabolites and compounds [4, 10].

Photosynthetic picoeukaryotes (PPEs) are a widely distributed and metabolically diverse group of photosynthetic unicellular algae approximately 1 - 3  $\mu$ M in diameter, that have independently evolved within heterokont, haptophyte and green algal lineages [11, 12]. They are important

primary producing organisms, forming key components of aquatic ecosystems and microbial food webs [13-16]. Their highly reduced size is posited as being a primary factor in their success and broad environmental distribution. Such miniaturization can reduce predation by some grazers, and results in a high cell surface area to volume ratio that enhances the efficiency of nutrient uptake and transport [11]. These attributes enable persistence in oligotrophic conditions and high-specific growth rates under nutrient-rich regimes [17, 18].

*Picochlorum* is a genus of picoeukaryotic green algae in the class Trebouxiophyceae, that have recently been of interest in microalgal research [19-21]. Isolates of the genus, predominantly inhabit marine and brackish water habits, and display broad levels of halo- and thermotolerance. Similar to the smallest known PPEs, of the Mamiellophyceae genus *Ostreococcus* [22-25], *Picochlorum* are non-flagellated pico-sized unicells with a mean diameter of 2  $\mu$ M; and possess correspondingly small genomes in the range of 13 – 15 Mbp [20, 23, 24, 26-28]. However, unlike *Ostreococcus*, isolates of *Picochlorum* exhibit great flexibility in growth conditions, and rapid growth rates under high-light, and nutrient-rich conditions [29, 30]. Due to their streamlined genomes, robust characteristics, and vigorous growth rates, *Picochlorum* isolates have recently been investigated for comparative genomics analysis [31], environmental fluctuations and toxicity [26, 28, 32], photophysiology [33], biomass and lipid production [29, 34-41], aquaculture feedstock, waste management and bioremediation [42-46], as well as biofuels, transgene expression [47-50] and most recently adaptive laboratory evolution (ALE) [51].

The isolate *Picochlorum* sp. SENEW3 (*P. SENEW3*) exhibits particularly robust characteristics under variable environmental conditions. It tolerates temperatures between 16 to 32 °C and thriving in both freshwater, three-fold seawater salinity under laboratory conditions, and seasonal salinity fluctuations of 0.2 to 13 Brix units in its natural estuary pond environment [27]. Previous examination of the *P. SENEW3* genome revealed its small and compact nature, enriched with putative salt stress response transporters, functional gene clusters and multiple instances of horizontal gene transfer related to stress adaptation [26, 52]. As such, it provides an intriguing case of picoeukaryote genome reduction under challenging environmental conditions, with robust growth and stress responses that appear to be absent in other PPEs such as *Ostreococcus*.

Previous genomic investigations have been partially limited due to fragmented short-read assembly technologies or incomplete functional annotation. However, recent advances in the

availability of long-read sequencing [53, 54], chromosome conformation capture techniques [55-59], and eukaryotic genome annotation pipelines [60-63] now enable full chromosomal assembly and functional annotation of even large, repeat rich plant genomes, improving omics analysis of genomic features [57, 64-66]. Utilising both short-read and long-read Oxford Nanopore Technologies (ONT) DNA sequencing, pulse field gel electrophoresis (PFGE), Hi-C chromatin conformation capture and functional annotation pipelines, this work presents in-depth investigation and functional analysis of the *P. SENEW3* genome with a fully annotated chromosomal assembly. This provides a key resource for further omics and molecular analysis of this robust poikilohaline picoeukaryote alga.

## Methods

### ***Picochlorum* sp. SENEW3 cell culture**

*Picochlorum* sp. SENEW3 (*P. SENEW3*) was obtained from Prof. Brian Palenik at Scripps Institution of Oceanography, UC San Diego, CA, USA and was originally isolated from a poikilohaline pond in the San Elijo Lagoon estuary in southern California [27]. Cultures were grown in f/2 medium without the addition of sodium silicate [67, 68] and with artificial seawater base used for Keller (K) medium [69]. Cultures were grown at 22 °C, 100 rpm, 100 m<sup>-2</sup> s<sup>-1</sup> white light on a 14-hour day: 10-hour night light cycle and transferred weekly to fresh medium. Cultures were periodically aseptically cell sorted with a BD Influx flow cytometer and/or treated with antibiotics (50 µg mL<sup>-1</sup> ampicillin, 10 µg mL<sup>-1</sup> gentamicin, µg mL<sup>-1</sup> kanamycin, and 100 µg mL<sup>-1</sup> neomycin) [70] to prevent bacterial or fungal growth. Cell density estimations via haemocytometer cell count were made using an Olympus BH-2 microscope with a 'Neubauer improved bright-line' hemocytometer slide as per the manufacturer instructions.

### **Pulsed-field gel electrophoresis**

Genomic DNA agarose plugs were generated based on the protocol of Hage and Houseley (2013) [71]. Briefly, *P. SENEW3* cells were cultured in f/2 medium to mid exponential growth phase cell density of 2 x 10<sup>7</sup> cells mL<sup>-1</sup>. Ten mL cell culture aliquots were centrifuged at 3,200 g for 15 mins, cell pellets were washed in 1 mL of wash buffer (10 mM Tris pH 7.6, 50 mM ethylenediaminetetraacetic acid [EDTA]) and resuspended in 50 µL of wash buffer with 1 mg mL<sup>-1</sup> lyticase. Cells were heated to 55 °C for 5 mins and mixed with 55 µL of melted 1.6% SeaKem LE agarose in dH<sub>2</sub>O at 55 °C and set in PFGE combe wells at 4 °C for 1 hr. Solid plugs were transferred to 1.5 mL microcentrifuge tubes, digested in 500 µL wash buffer with 1 mg mL<sup>-1</sup> lyticase for 3 hr at

37 °C and then in 500 µL PK buffer (100 mM EDTA, 0.2% sodium deoxycholate, 1% sodium lauryl sarcosine) with 1 mg mL<sup>-1</sup> proteinase K (Sigma-Aldrich P6556) overnight at 55 °C. Digested cell plugs were washed three times in 1 mL wash buffer for 30 mins at room temperature, resuspended in 500 µL wash buffer and stored at 4 °C.

*P. SENEW3* chromosomal DNA was separated using a Bio-Rad CHEF Mapper XA pulsed field gel electrophoresis (PFGE) system. Four agarose plugs and a NEB Yeast Chromosome PFGE marker (N0345S) plug were inserted into a 1% SeaPlaque Low Melting Point Agarose (Lonza: 50101 – 25 g) gel, prepared with 0.5 X Tris/Borate/EDTA (TBE) buffer. Each plug was run in parallel for 24, 28, 32 and 36 hours, respectively (Supp. Table 1). Following each time point, each lane was excised and visualised with a blue light gel box.

### **Genomic DNA extraction**

Genomic DNA for ONT long-read sequencing was extracted via agarose gel electrophoresis concentration and drop dialysis of PFGE agarose plugs as described in Sambrook and Russell (2001) protocol 20 [72]. Ten *P. SENEW3* PFGE plugs were placed together within a large cut-out well in 5% NuSieve GTG Agarose gel (Lonza 50081), stained with SYBR Safe DNA gel stain (Invitrogen S33102) and separated using standard agarose gel electrophoresis apparatus at 60 V for 3 hours at 4 °C. The resulting gel band was visualised using a blue-light gel box, excised, and stored with 1 mL of PFGE wash buffer at 4 °C. Gel slices were equilibrated in 12 mL of equilibration buffer (1X TBE, 100 mM NaCl, 30 µM spermine, 70 µM spermidine) for 20 mins at room temperature with gentle agitation, and then melted at 68 °C for 15 mins followed by incubation for 2 hours at 42 °C with 10 units of β-agarase I (New England BioLabs M0392S). Digested supernatant (100 µL at a time) was spotted onto the centre of a 0.05 µM MF-Millipore membrane filter (Merck Millipore VMWP02500) floating on 100 mL of transfection buffer (10 mM Tris pH 7.5, 0.1 mM EDTA, 100 mM NaCl, 30 µM spermine, 70 µM spermidine), dialysed for 1 hour at room temperature, followed by replacement with the same volume of fresh buffer and an additional 1 hour of dialysis. DNA was transferred into a new 2 mL Eppendorf tube and precipitated with 3 volumes of 100% cold ethanol and 1/10 volume 3 M Sodium Acetate for 1 hour at -30 °C. DNA precipitate was pelleted at 12,000 g for 15 mins, washed three times with 70% ethanol and resuspended in 100 µL Tris-EDTA buffer overnight at 4 °C. DNA quality and concentration were assessed via Nanodrop and Qbit dsDNA HS Assay (Invitrogen, Thermo Fisher Scientific: Q32854). Genomic DNA integrity was assessed via agarose gel electrophoresis.

## Genomic sequencing

Genomic DNA for Illumina DNA sequencing (used for genome polishing and organelle assembly) was extracted via a CTAB/Phenol/Chloroform method based off Jagielski et al (2017) [73]. In short, 100 mL of approximately  $1 \times 10^7$  cell  $\text{mL}^{-1}$  of *P. SENEW3* were centrifuged at 3,200 g at 4 °C for 15 mins, resuspended in 600  $\mu\text{L}$  of freshly prepared DNA extraction buffer (10 mM Tris-HCL pH 8.0, 1 mM EDTA, 100 mM NaCl, 2% Triton-x100, 1% SDS) and transferred to 2 mL Eppendorf tubes with 0.1 mm glass beads (cell pellet/beads, 1:1 ratio). Sample tubes were vortexed at 20 Hz for 5 mins, lysate was transferred to 5 mL Eppendorf tubes, and remaining glass beads were washed twice with 250  $\mu\text{L}$  of extraction buffer that was pooled with the initial lysate. Samples were incubated at 56 °C for 1 hour with 160  $\mu\text{g mL}^{-1}$  proteinase K, followed by the addition of 2  $\mu\text{L}$  of 2-mercaptoethanol, 500  $\mu\text{L}$  of 5 M NaCl, 400  $\mu\text{L}$  of prewarmed 10% CTAB (10% Cetyltrimethyl ammonium bromide and 0.7% NaCl (w/v) in nuclease free water) and incubated at 65 °C for 10 mins. Lysate was extracted once with chloroform: isoamyl alcohol (24:1 v/v), mixed via inversion and centrifugation at 12,000 g for 5 mins, supernatant was aspirated and extracted three times with equal volume of phenol: chloroform: isoamyl alcohol (25:24:1 v/v/v). Genomic DNA was precipitated with 3 volumes of 100% ethanol and 1/10 volume 3 M Sodium Acetate for 1 hour at -30 °C. DNA precipitate was pelleted at 12,000 g for 15 mins, washed three times with 70% ethanol and resuspended in 100  $\mu\text{L}$  Tris-EDTA buffer overnight at 4 °C. Residual RNA was degraded with 10 Units of RNase I (Thermo Scientific EN0601) at 37 °C for 1 hours followed by ethanol precipitation and wash was repeated, and DNA was resuspended in 100  $\mu\text{L}$  Tris-EDTA buffer. DNA quality and concentration was assessed via Nanodrop and Qubit dsDNA HS Assay (Invitrogen, Thermo Fisher Scientific: Q32854). Genomic DNA integrity was assessed via agarose gel electrophoresis. DNA library was prepared using TruSeq Nano DNA Kit (preparation guide part # 15041110 Rev. D) and Illumina pair-end DNA sequencing was undertaken by Macrogen Inc. (Republic of Korea).

Long-read sequencing of genomic DNA was undertaken using ONT MinION for nuclear genome assembly. One  $\mu\text{g}$  of gDNA was sequenced as per ONT protocol '1D Genomic DNA by ligation' (SQK-LSK109) version: GDE\_9063\_v109\_revD. In short, gDNA was repaired and end-prepared using NEBNext FFPDNA Repair Mix and NEBNext End repair / dA-tailing Module in accordance with manufacturer's instructions and size selected with 1X volume (1:1 / bead : sample ratio, 60  $\mu\text{L}$  total) of AMPure XP magnetic beads. Adapter ligation and clean-up was carried out with NEBNext Quick Ligation Module (E6056) T4 Ligase and purified with 0.67X volume (beads:sample, 40  $\mu\text{L}$ ) of

AMPure XP magnetic beads. DNA was then prepared for sequencing using the provided ligation sequencing kit (SQK-LSK109), primed, and loaded on to the SpotON flow cell (FLO-MIN106, v. R9.4.1). MinION sequencing was run with the MinKnow software v 3.3.2 for 48 h (base voltage - 180 mV, 1.5 hr mux scan time). Raw data was output as fast5 files and converted to fastq via MinKnow software base-calling (guppy base-caller v. 3.0.3).

### **Genome assembly**

Organellar genomes were constructed by mapping assembly of the Illumina pair-end reads. Here, Mirabait (MIRA v 4.9) [74] was used to pull out Illumina reads that mapped to the chloroplast and mitochondria of the first assembly of the *P. SENEW3* genome [26]. MIRA v 4.9 was then used to re-assemble the baited reads into chloroplast and mitochondria genomes.

The *P. SENEW3* nuclear genome was assembled *de novo* with the ONT MinION sequencing reads and was polished with both the ONT and the Illumina sequencing reads. ONT reads were assembled with Canu v 1.8 using default settings and corrected long-reads were remapped to the assembled contigs with Minimap2 v 2.17 [75], which were then polished with Racon v 1.4.3 [76]. Racon polished contigs were then polished once with Medaka v 1.0.3 [77]. The resulting contigs were iteratively corrected four times with indexed Illumina short-reads (Samtools v 1.9 [78, 79]) using Pilon v 1.23 [80]. The resulting alignments were inspected and manually corrected with Gap5 (Staden Package v 1.9) [81]. Contigs with unmapped or poor read coverage were discarded as likely mis-assemblies or contaminants.

### **Hi-C: Chromosome conformation capture**

One hundred and fifty mL of approx.  $1 \times 10^7$  cell mL<sup>-1</sup> *P. SENEW3* culture in f/2 medium were fixed directly in culture medium by adding methanol stabilised formaldehyde (Sigma-Aldrich F8775-4X25ML) to each cell culture to a final concentration of 3% (v/v). Cultures were incubated for 30 mins at room temperature (RT), and subsequently quenched with 2 volumes of 2.5 M glycine for 5 mins at RT then for 15 mins at 4 °C. Cells were centrifuged at 3,200 g for 15 mins at 4 °C, washed with 20 mL of f/2 medium, re-centrifuged at 3,200 g for 5 mins at 4 °C. Samples were then resuspended in 1 mL of f/2 medium, transferred to 1.5 mL Eppendorf tubes, centrifuged for 2 mins at 13,000 g, supernatant was aspirated and pellets were frozen and stored at - 80 °C. Samples were shipped on dry-ice to the Curie Institute, Centre of Research - UMR3664 Nuclear Dynamics – Paris France for Hi-C chromosome conformation capture processing.

The cell pellet was thawed on ice and suspended in 500  $\mu$ L TE + cComplete™, EDTA-free Protease Inhibitor Cocktail (Roche). Sample was then treated for 3 cycles of 20 sec at 6.5 M/s with 0.5 mm beads in a FastPrep apparatus (MP Biomedicals) to break the cell walls before starting the Hi-C protocol. Hi-C was carried out using the Arima-HiC+ kit (Arima Genomics) following the manufacturer's instructions. The Hi-C library were processed according to Arima Genomics recommendation using the KAPA HyperPrep Kit (Roche) and Illumina TruSeq sequencing adapters (Illumina 20020590) to make an Illumina sequencing library that was further sequenced as PE100 on a NovaSeq 6000 system at the Curie CoreTech platform.

Hi-C pair-end reads were aligned and filtered using HiC-Pro [82]. Valid pairs were then processed to build raw and balanced Hi-C matrices in .cool format using the cooler package [83]. Hi-C matrices were further manipulated and drawn at indicated resolutions using cooler, cooltools (version 0.5.0, DOI:10.5281/zenodo.5708875) and matplotlib (version 3.3.2, DOI:10.5281/zenodo.5773480) [84] python packages and library.

### **Genome annotation**

*P. SENEW3* genome annotation was undertaken separately for organellar and nuclear genomes. The chloroplast and mitochondrial assemblies were annotated with Prokka: rapid prokaryotic genome annotation v 1.14.6 [85]. Annotation of the nuclear genome was made using the 'assemblage' iterative two-pass workflow (<https://github.com/sujaikumar/assemblage>) of the MAKER2 eukaryotic genome annotation pipeline [63]. Prior to running the first pass of the MAKER2 v 3.01.03 genome annotation pipeline, separate eukaryotic *ab initio* gene predictions were run, and a custom repetitive sequence library was generated for the *P. SENEW3* nuclear genome. Augustus v 3.3.3 gene prediction [86] was trained on the first version of the *P. SENEW3* genome (psev1) [26] with the psev1 protein and transcript/ESTs as evidence for prediction training. Scipio v 1.4 [87] was used to map protein and transcript evidence to the psev1 genome to generate a gene training set. Augustus v 3.3.3 e-training and optimisation scripts were used to generate an Augustus species gene prediction parameter set. BUSCO v 3.0.2 was used to generate the in initial *ab initio* predicted gene set for the *P. SENEW3* genome, using the eukaryote core BUSCO dataset (v eukaryote\_odb9 2016-11-02) [88] and the Augustus v 3.3.3 generated psev1 species parameters set. BUSCO gene predictions were converted from gff3 file type to SNAP v 2006-07-28 [89] gene predictor compatible hmm file type with an in-house Perl script. GeneMark-

ES v 3.60 [90] with default setting was run against the *P. SENEW3* nuclear genome as the third *ab initio* gene predictor. A custom repetitive sequence library was generated for the *P. SENEW3* nuclear genome using the MAKER2 'repeat library construction-advanced' protocol [91].

BUSCO-SNAP predictions, GeneMark-ES predictions, the custom repeat masking library, pseV1 EST's and protein data sets from *Picochlorum* sp. SENEW3 v 1.0 [26], *Picochlorum* sp. SENEW3 v 2.0, *Picochlorum* nbrC102739, *Picochlorum oculatum*, *Picochlorum oklahomensis* [52], *Picochlorum soloecismus* (spDOE101) [92] and *Picochlorum renovo* v 2 [47] were imported into the first pass of the MAKER2 v 3.01.03 pipeline. MAKER2 first-pass output was used to re-train SNAP gene predictor. MAKER2 pipeline was run for an increased stringency second pass with updated SNAP, Genemark and Augustus. Putative functions and domain information were appended to MAKER2 gene predictions using BlastP (blast suite v 2.10.1+) [93] against the UniProt/SwissProt database [94] and InterProScan v 5.42.78.0 [95]. Genomic tRNAs were annotated with tRNAscan-SE v 2.0 [96].

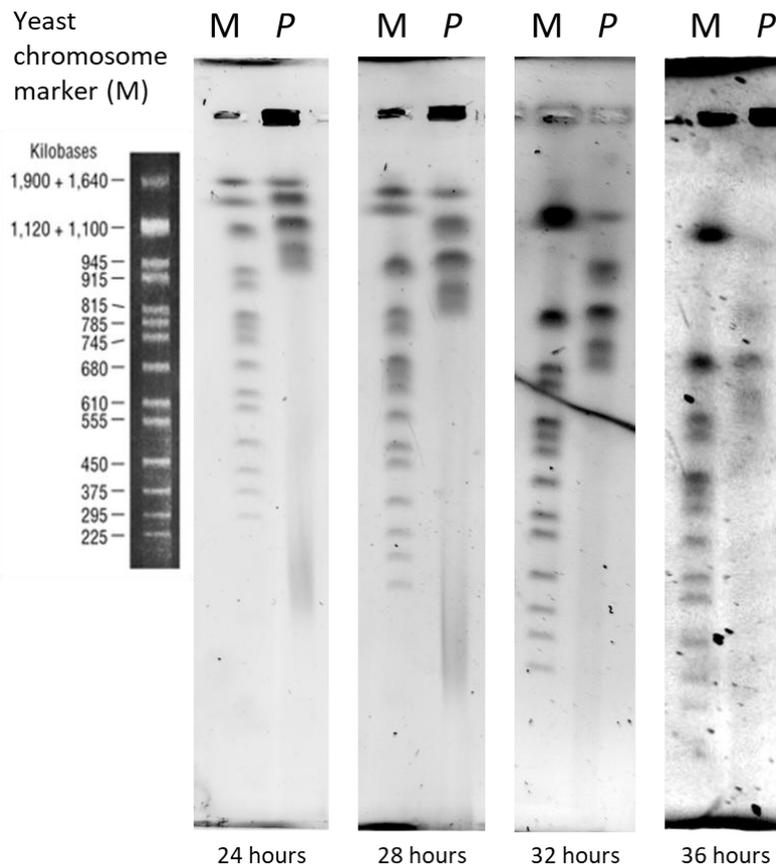
### **Genomic analysis**

The combined nuclear, chloroplast and mitochondrial genome assemblies were assessed for assembly completeness using BUSCO v 3.0.2 [97] against the eukaryota\_odb9 (2016-11-02) BUSCO gene database. RepeatModeler2 v 2.0.1 [98] was run against the combined genome to produce a custom *de novo* transposable element (TE) family library. The TE library was input into RepeatMasker v 4.1.0 (database: Dfam\_3.1, rmbblastn version 2.10.0+) [99], which was used to assess the type and coverage of repetitive elements present in the assembled genome (Supp. Table 4). Comparative gene orthology across 15 select species/strains of the phylum Chlorophyta (Supp. Table 19) with available proteomes was analysed using Orthofinder v 2.5.4 [100, 101]. Carbohydrate-active enzymes were identified via dbCAN2 meta server via protein sequence with HMMER (E-Value < 1e-15, coverage > 0.35), DIAMOND (E-Value < 1e-102), hotpep (Frequency > 2.6, Hits > 6) and CGCFinder (Distance <= 2, signature genes = CAZyme+TC) [102, 103] (Supp. Table 12). Transporters were identified using TransAAP v. 2.0 [104]. Variants calls, including identification of single nucleotide polymorphisms (SNPs), insertion and deletions (INDELs) were made using bcftools v. 1.17 (Supp. Tables 21 and 22) [79] and ploidy analysis was undertaken using nQuire [105]. Additional functional annotations of predicted proteins were identified using KEGG GhostKOALA [106], and eggNOG-mapper v. 2.1.6 [107, 108] (Supp. Table 9 – 11). A list of combined annotation available in Supp. Table 3.

## Results and Discussion

### Gene content

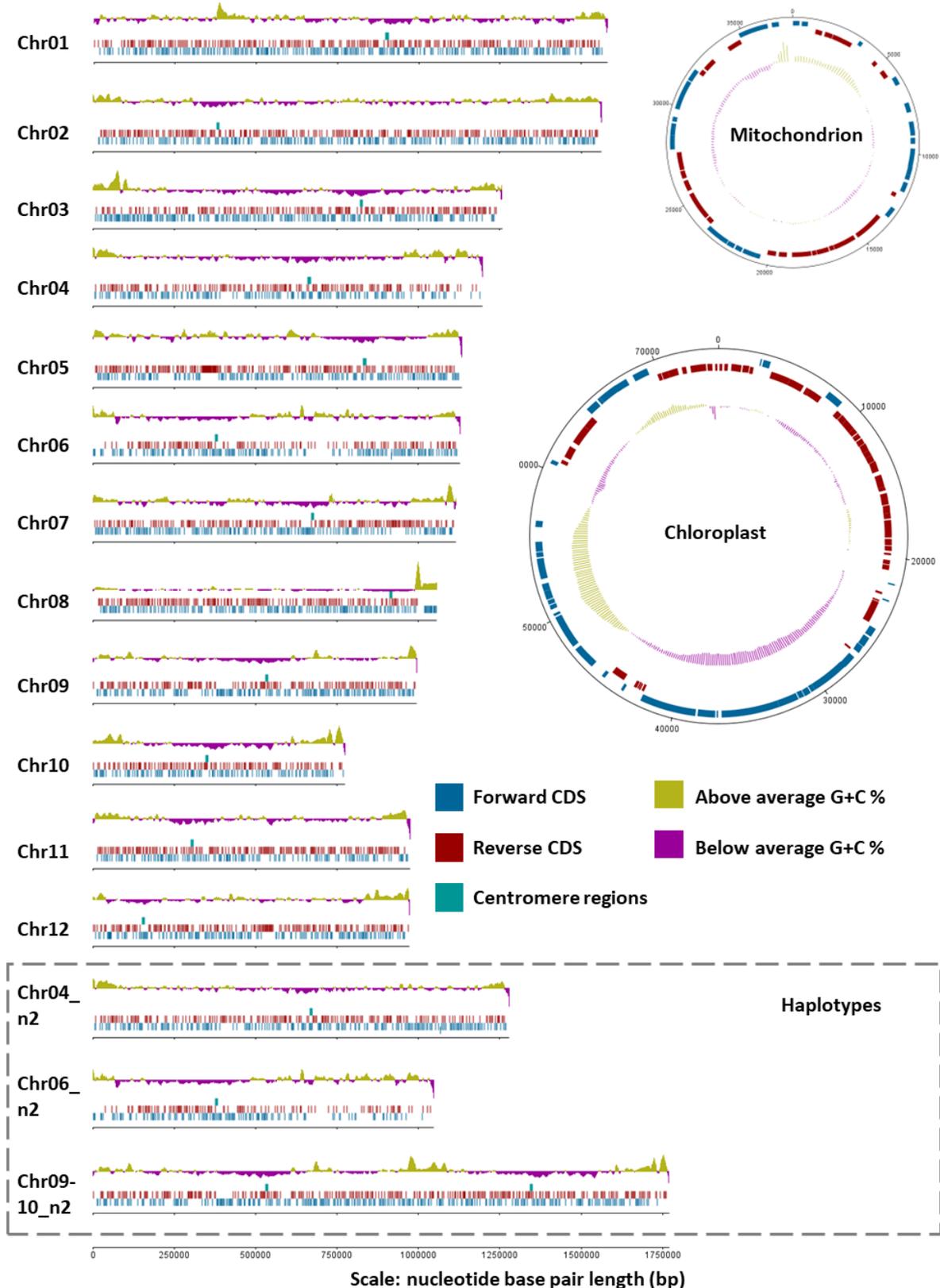
Using a combination of short reads (Illumina), long reads (ONT MinION) and proximity information provided by a Hi-C experiment (see paragraph below), we assembled the 13.75 Mbp primary haploid assembly of the *P. SENEW3* nuclear genome into 12 chromosomes (chr) ranging from 0.77 to 1.58 Mbp. These were consistent with the range of physical chromosomes separated by PFGE (Figure 1). Bioinformatic ploidy analysis suggested that the genome is at least diploid (Supp. Table 20), this is consistent with previous analysis [52]. Variant calling to determine heterozygosity, identified 964 SNPs (7.02 per 100 Kbp) and 366 INDELS (2.66 per 100 Kbp) across the primary assembly (summarised in Supp. Table St2). This rate for example is higher than that of similarly sized *Saccharomyces cerevisiae* genome of 4 SNPs per 100 Kbp [109]. It is notable that previous analysis of heterozygosity among *Picochlorum* strains has identified significant variation in genome size, ploidy, and heterozygosity across isolates. Among the putative diploids it appears that the genome of *P. SENEW* possesses relatively low heterozygosity [52]. The nuclear genome had an average G+C content of 46.3% and BUSCO (Benchmarking Universal Single-Copy Orthologs) completeness score of 90.7% (C:90.7% [S:90.4%,D:0.3%], F:1.0%, M:8.3%, n:303) [88]. Organellar 74.135 Kbp chloroplast and 37.566 Kbp mitochondria genomes had G+C contents of 32.2% and 42.4% respectively (Supp. Table 2).



**Figure 1.** Pulsed field gel electrophoresis gel slices of *P. SENEW3* chromosomal DNA (*P*) against NEB Yeast Chromosome PFGE marker N0345S (*M*) at 24-, 28-, 32- and 36-hour time points, using a Bio-Rad CHEF Mapper XA pulsed field gel electrophoresis system.

Repetitive sequence analysis of the assembled genome using Repeatmasker - based off the generated Repeatmodeler2 TE library, revealed a low density of repeat elements. TEs made up only 4.99% of the assembled nuclear genome (5.23% including organelle genomes) and were comprised primarily of unclassified repeats (1.87%) and retroelements (2.06%), including long interspersed repetitive elements (1.25%, LINES) and long terminal repeats (0.81%, LTRs). DNA transposons only made up 0.35% of the nuclear genome (Supp. Table 4). Despite this low TE count, several localised genomic regions appear to be enriched in repetitive sequences, likely the cause of difficulty in assembly and Hi-C contact mapping. Three such regions included the right arms of chr04 and ch06 resulting in fragmented Hi-C contact mapping (Figure 4B), and a ~10 Kbp region of Guanine (G) repeats resulting in a high G+C% spike at the right arm of chr08 (Figure 2). Thus, the *P. SENEW3* assembled genome is in general repeat poor, however it is likely that the repeat count is underrepresented. Nevertheless, the *P. SENEW3* genome appears to be highly compact like that of other PPE's (e.g., *O. tauri* and *Micromonas*) [9, 23]. In contrast to many other green lineage organisms, which typically contain many repeat sequences and transposable

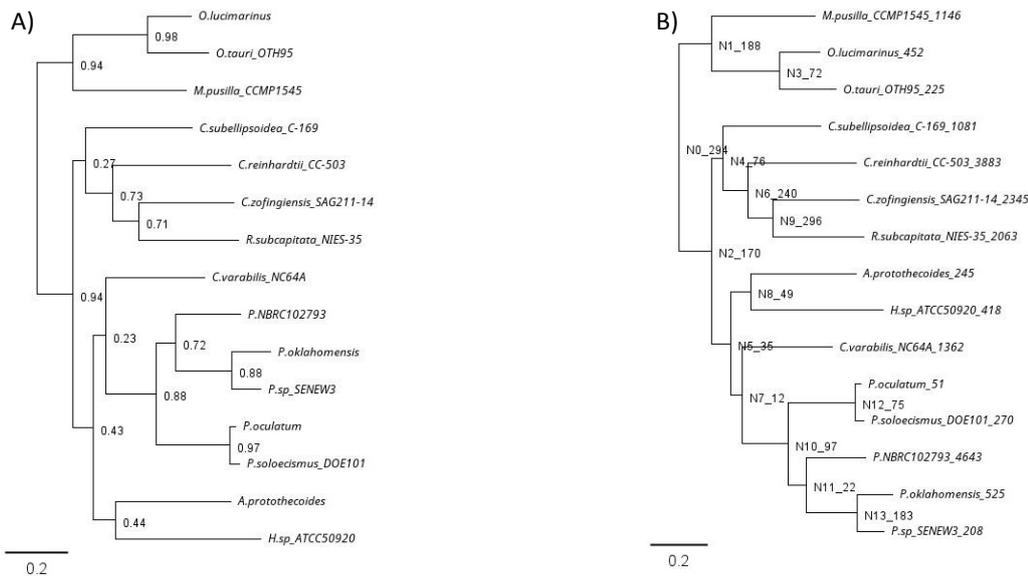
elements, the number of which is known to be positively correlated with genome size above ~10 Mbp [110, 111].



**Figure 2.** Genome chromosome and organelle map, displaying chromosomes (Chr) 1-12, chromosome haplotypes

(n2), chloroplast and mitochondria sequence lengths, created with Artemis - DNA plotter [112]. Forward protein coding sequences (CDS) shown in blue, reverse CDS in red, putative centromere regions in green, above average G+C% in yellow, and below average G+C% in purple. The Chr09-10\_n2 haplotype represents a putative fusion. Gene loci for primary and haplotype assemblies available in Supp. tables 7 and 8).

A total of 6,564 genes were predicted across the nuclear and organellar genomes. This represents fewer predicted genes than earlier assemblies [26, 52], however increased prediction stringency applied during the second pass of the MAKER2 workflow successfully removed false positive gene predictions (923) from the first run of the annotation pipeline. Final density represents one gene per 2,161 bp, with 74.35% of the genome represented by protein-coding sequence. This high gene-density is comparable to that of the Mamiellophyceae genomes of *Ostreococcus* and *Micromonas* with protein-coding sequence comprising close to 70% for most strains [1, 9, 22-24]. Whole proteome orthogroup analysis against 14 other Chlorophyta species with complete genome assemblies was used to verify the phylogeny of the assembly. This clearly placed the *P. SENEW3* within the *Picochlorum* genus and most similar to *Picochlorum oklahomensis* (Figure 3A). In addition, the relatively low number of gene duplication events in the *P. SENEW3* genome (Figure 3B) was consistent with its streamlined and compact nature. Meiosis is a key process in evolutionary diversification in eukaryotic organisms, however in several groups of green algae, sexual reproduction is either absent or yet undocumented, as appears to be the case for most planktonic species [1, 113]. The presence of core meiotic determinant genes (though many more are involved in the process) [114, 115] was examined to assess whether *P. SENEW3* is likely to undertake sexual reproduction. Several of these were identified: *DMC1*, *HOP2*, *MSH5* (and homologs *MSH2* and *MSH6*) *SPO11-2*, *HAP2* and *MER3*; yet six others were missing (Supp. Table 13), suggesting but leaving uncertain whether *P. SENEW3* can undergo meiosis.



**Figure 3.** Orthogroup analysis of 15 representative chlorophyte proteomes displaying A) an inferred species tree based on orthogroup relationships and B) number of gene duplications with greater than 0.5 support per orthogroup and species. Scale bar displays number of amino acid substitutions per site.

The occurrence of such highly reduced genome and cell size has led to the proposition that picoeukaryotic algae may be near the limit for minimal genomes of free-living photosynthetic eukaryotes [1]. One hypothesis explaining this genome reduction in PPEs is the Black Queen hypothesis (BQH), which predicts that an availability of shared metabolites in marine environments would drive the loss of redundant metabolic pathways in individual genomes/species that are otherwise available within the surrounding microbial community. Providing a selective advantage by conserving an organism's finite resources through adaptive gene loss [116-118]. However, unlike other PPEs such as *Ostreococcus*, *P. SENEW3* displays robust growth in variable environmental conditions, tolerating both euryhaline conditions and temperatures between 16 to 32 °C [26]. In such extreme or challenging environments, this pressure for adaptive gene loss should be reduced since there are typically fewer community members to produce expensive metabolites [116-118]. Hence, the large range of environmental tolerances of *P. SENEW3* and payoffs between growth, nutrient availability, and stress, presents an alternate environmental adaptation paradigm for PPEs.

### Genome assembly and centromere identification with Hi-C

Chromosome conformation capture (3C) techniques such as Hi-C allow for the examination of spatial chromatin organization by quantifying the interactions between genomic loci that may otherwise be distantly separated in sequence location [55, 56, 119, 120]. This is based on the principle that interaction frequencies decrease with genomic distance and that *cis*-interactions

(from loci located on the same chromosome) are always significantly higher than *trans*-interactions (from loci located on different chromosomes). Recently developed Hi-C based genome assembly programs [121-123] have proved to be a powerful instrument in the genomics toolkit, and are being rapidly integrated into individual organism sequencing projects [57, 65, 66, 124, 125] and large scale biodiversity sequencing projects such as the earth biogenome project [126-128]. In particular, Hi-C assembly allows for chromosome scaffolding without the need of contiguous sequences and gap filling. This alleviates difficulties related to large repetitive sequences that often stay unassembled in genomes.

For the small, repeat-poor *P. SENEW3* genome, we corrected minimal scaffolding errors via visual inspection of the Hi-C matrix mapped to our initial reference assembly (Figure 4A). This resulted in a final reference genome with 12 chromosomes, corresponding to twelve individual squares of higher interaction frequencies along the diagonal of the matrix (Figure 4A, Supp. Figure 1). Still, an abnormal signal is present in the inter-chromosomal space between chr04 and chr06 highlighted on a higher resolution map (5 kbp bins) in Figure 4B. The end of the right arm of chr06 (~327 Kbp) makes high interactions with the rest of chr06, but also with a large segment of chr04. Similarly, the end of the right arm of chr04 (~245 Kbp) makes high contacts with chr06. Nevertheless, these two regions do not interact together more than in the inter-chromosomal range. This represents a reciprocal translocation between a subset of chr04 and chr06 in the population. Either this translocation occurred early during the growth of the clone used to perform the Hi-C experiment and represent a large part of the final analysed population, or the strain is diploid, and the reciprocal translocation occurred between two of the homologs (Figure 4C). Another signal appears to be unusual in the inter-chromosomal area on the final assembly, between chr09 and chr10. The contact frequency between the two chromosomes is in between the inter-chromosomal range of frequencies with other chromosomes and the intra-chromosomal range. We explain this signal by a fusion occurring between both chromosomes but involving only part of the population, or again, the fusion of two homologues only in a diploid strain (Figure 4C). Additional Hi-C experiments with a fresh clone, could be used to decipher the validity of these hypotheses and whether the genome is inherently unstable under laboratory culture, or such rearrangements are fixed within the population. We have not undertaken the work to address these hypotheses further, favouring the rearrangement of homologs in a diploid strain (Figure 4C). This is supported by the existence of a PFGE chromosomal band above 1,640 Kbp (Figure 1), which would correspond to the ~1,770 Kbp sequence length of chr09-10\_n2 (Figure 2). We have



Chr10. **D)** Protein sequence alignment (Clustal Omega v 1.2.4) of *P. SENEW3* histone H3 homologues (PSENEW3\_00005899, \_00004192, \_00001782, \_00001261) and putative CenH3/CENP-A histone H3 variant (PSENEW3\_00002576), highlighting key CenH3 amino acid substitutions (sites 77, 85 and 108) and insertion [129].

One of the most striking architectural chromosomal features visible on Hi-C matrices are inter-chromosomal centromere interactions. These interactions are the consequence of centromere clustering, and a broadly shared organization across the eukaryotic domain of life [130]. Several groups have taken advantage of this property to identify centromeric regions in species where they were not defined [131-135]. In *P. SENEW*, the inter-chromosomal area clearly shows such a dotted pattern between chromosomes (blue arrowheads on Figure 4A). This enabled confirmation of twelve centromeric regions (primary assembly), one per chromosome, marked by high-frequency trans-interactions. Sequence bins of high-interaction chromosomal trans-signals were identified using matrices at 10 kb resolution and defined as putative centromeric regions (Supp. Table 5).

To further characterise the centromeres of *P. SENEW3*, and decipher whether they exist as point centromeres, as in budding yeasts, or short regional centromeres (below 10 kb) [130, 136] additional experiments are required. Chromatin immunoprecipitation using a tagged version of the centromeric variant of the histone H3 (CENH3/CENP-A) as a bait has been successfully used in this purpose [137, 138]. To initiate this future work, the genome was examined for a putative CENH3. A tBLASTn homology search using human H3.1 (UniProt P68431 H31\_HUMAN) as query against the *P. SENEW* genome identified five H3 homologs (Figure 4D, and Supp. Table 6). Four of the identified proteins were identical to each other and contained only four mismatches compared to human H3.1: PSENEW3\_00005899, PSENEW3\_00004192, PSENEW3\_00001782, PSENEW3\_00001261. Corresponding genes were already well annotated in our database as being H3 homologs, belonging to the four histone clusters of the genome (each comprising H3, H4, H2A and H2B). The last homolog, PSENEW3\_00002576, had diverged amino acid sequence, was found isolated on chr01, and had two introns while the others had none. Additionally, multiple alignment of this proteins revealed substitutions in PSENEW3\_00002576 (blue boxes on Figure 4D) at key positions reported to be mutated in CENH3s compared canonical H3s in other species [129], as well as a higher divergence in the N-terminal tail. Taken together, these characteristics gave sufficient evidence to annotate PSENEW3\_00002576 as a putative CENH3 in *P. SENEW3*.

Hi-C analysis of additional architectural features at 5 kb resolution indicated that the *P. SENEW3* genome possessed few other elements of chromosome organisation commonly found in other species such as chromatin compartmentalisation or topologically associated domains (TADs) [120, 139, 140]. Nevertheless, two other feature types were identified. The first of these were several clear loop anchors, the most prominent involving three regions on chr12 (Supp. Figure 1 – chr12). These interactions are reminiscent of the mating type loci of *Saccharomyces cerevisiae* [141-144], however, examination of the underlying sequence at these loci did not reveal candidate genetic elements that would explain such strong interactions. The second identifiable element was the clustering of telomeres within and between the termini of chromosomes (see green circles in Figure 4A). The separate clustering of centromeres and telomeres, typically at opposed ends of the nucleus, corresponds to the description of Rab1-like genome organisation (clustering of the centromeres on a single side of the nuclear envelope), commonly seen in the genomes of some plants, as well as yeast with similarly small genomes [145].

Chromatin conformation capture methods like Hi-C are an invaluable tool for the investigation of 3D genomic architecture as applied here, enabling final scaffolding of genome assembly, identification of chromosomal rearrangements and features such as centromeres, telomeres, and loop anchors. Chromosomal rearrangements, fusions and translocations as seen here between chr04 and chr06 and between chr09 and chr010 can occur naturally or otherwise in animals, fungi and plants [146-150]. However, the formation of dicentric chromosomes for example are known to be unstable and typically result in the deletion or inactivation of one of centromeres when fixed in a population (indicated by a “?” in Figure 4C) [151, 152]. Large structural genome rearrangements have however been previously observed in the PPEs *P. celeri* [153] and *O. tauri* [23], and can be important factors in population adaptation and divergence [154, 155] and genome evolution and diversity [147, 150]. Additionally, centromere characterisation has become invaluable in molecular and synthetic biology and has enabled the development of yeast plasmid vectors [156, 157], artificial chromosomes, neochromosomes [158-160] and episomes [161-163].

### **Selenocysteine proteins and metabolism**

Genomic selenoprotein composition has been associated with algal environmental, habitat and lifestyle adaptation [164]. Thus, we examined the selenoprotein composition of *Picochlorum* SENEW3. BlastP matching of the *P. SENEW3* proteome against the algal selenoprotein database [164] identified 11 selenoproteins (10 families) out of the 25 selenoprotein families found in algae.

This includes an additional six selenoproteins, aside from the five that have recently been identified in *P. SENEW3* [164]. In addition, the presence of the selenocysteine tRNA gene, tRNA-Sec(TCA), required for selenoprotein translation [165] was also identified. The best blastP matches for *P. SENEW3* selenoproteins against the algal selenoprotein database [164] included two glutathione peroxidases (GPXs) as well as possible selenoproteins K (SELENOK), T (SELENOT), and U (SELENOU). With additional matches to Thioredoxin reductase (TXNRD), Iron-sulfur oxidoreductase (FeSoxR), DNA-[protein]-cysteine methyltransferase selenoprotein (MDP), Methionine sulfoxide reductase A (MSRA) and Selenoprotein O (SELENOO) (Supp. Tables 15 and 16).

Selenium (Se) is an essential trace element in many organisms, it is co-translationally incorporated into the 21st amino acid selenocysteine (codon UGA) [165-167] and undertakes vital roles in both redox homeostasis and numerous cellular processes [168-171]. Despite identification of an additional six putative selenoproteins here, the Trebouxiophyceae, including *P. SENEW3*, appear to contain far fewer [164]. Among the aquatic green algae, the order Mamiellales for example have been found to possess the greatest number of selenoproteins (generally > 20). In fact, most algal groups high in selenoproteins, like the Mamiellales, appear to live in marine environments. Whereas algae and plants inhabiting non-marine environments (i.e., freshwater and terrestrial) have small selenoproteomes, often replaced with cysteine or expunged altogether [164, 172, 173]. We speculate that the ancestor of *P. SENEW3*, along with others in the genus, initially adapted to freshwater, undergoing a reduction in selenoprotein content prior to re-adapting to marine and saline lagoon environments. This seems likely considering the closely related genus *Chlorella* predominantly inhabits freshwater habitats [174].

### **Transporters**

Transporter analysis using TransAAP [104], KEGG [106] and eggNOG annotations [107, 108] revealed a high number of likely transporters (416) largely similar to those predicted by Foflonker (2015) [26] (Supp. Tables 17 and 18). Some of the most notable were the 39 annotated metal-ion transporters, with the most numerous being of the zinc (Zn<sup>2+</sup>)-iron (Fe<sup>2+</sup>) permease (ZIP) family, the CorA metal ion transporter (MIT) family, the P-type ATPase (P-ATPase) superfamily and the ATP-binding Cassette (ABC) superfamily for transport of zinc (Zn), magnesium (Mg), cobalt (Co) and Copper (Cu). Three iron transporters were also identified, including a high affinity iron permease 1, as well as transporters for the toxic heavy metals cadmium/zinc (Cd/Zn), chromate

(CrO<sup>2-</sup>), nickel (Ni), tellurium (Te), metal tolerance proteins A2 and C4 (MTPA2 and MTPC4) and arsenite (As III), though this may in fact be involved in chlorophyll biogenesis in algae [175].

Metal-ion transporters form vital components in the interactions of algae with their environment, and function in the scavenging and transport of essential cofactors and in initial response to fluctuations in extracellular metal concentrations. The abundance of these metal transporters likely represents their requirement in the “light” and “dark” reactions of photosynthesis [176, 177]. In addition, the presence of heavy metal transporters could represent potential avenues for metal tolerance in *P. SENEW* which should be investigated further. Indeed, many microalgae are being investigated for potential heavy metal phyto remediation applications [178-182]. A recent review by Goswami et al (2021) [183], investigating phyto remediation of wastewater by *Picochlorum* strains suggest a high level of strain variability in tolerance to heavy metal contaminated water.

We identified numerous sodium ion (Na<sup>+</sup>): proton (H<sup>+</sup>)/phosphate (PO<sub>4</sub><sup>3-</sup>) symporters and antiporters, potassium/sodium ion antiporters and two inward rectifier potassium channel transporters. An additional four (total of seven) mechanosensitive ion channel proteins, and nine calcium ion (Ca<sup>2+</sup>) transporters were also identified. The sodium, potassium, and calcium transporters in *P. SENEW3* likely contribute to its broad salinity tolerance [26, 28]. This euryhaline characteristic of *P. SENEW3* is most likely expanded by the additional mechanosensitive ion channels which are responsible for transducing physical force into electrochemical signals. In particular, mechanosensitive channels of small conductance (MscS) are responsible for management of cytosolic osmotic pressure and can act as emergency release valves by rapidly jettisoning cellular osmolytes under hypoosmotic shock, preventing cell lysis [184-188].

The final set of prominent transporters identified were five putative bestrophin proteins, a family of calcium (Ca<sup>2+</sup>) activated chloride (Cl<sup>-</sup>) channels that mediate the transport of monovalent anions (largely Cl<sup>-</sup>) in response to intracellular Ca<sup>2+</sup> [189, 190]. Aside from transporting Cl<sup>-</sup>, bestrophin channels have also been found to be highly permeable to bicarbonate (HCO<sub>3</sub><sup>-</sup>) [191]. HCO<sub>3</sub><sup>-</sup> is utilised as a form of inorganic carbon in algal CO<sub>2</sub>-concentrating mechanisms (CCMs) to increase the CO<sub>2</sub>:O<sub>2</sub> ratio at the active site of ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO). This enhances the efficiency of photosynthetic carbon fixation in water where CO<sub>2</sub> diffusion is limited [192, 193]. For example, in the green alga *Chlamydomonas reinhardtii* bestrophin channel



**Figure 5.** Kyoto encyclopedia of Genes and Genomes (KEGG) map of carbon fixation in photosynthetic organisms made with KEGG Mapper [200, 201]. Enzymes present in the *P. SENEW3* genome are highlighted in green, identifying genes necessary for both crassulacean acid metabolism (CAM) and C<sub>4</sub> (Hatch-Slack / C<sub>4</sub>-Dicarboxylic acid cycle) photosynthesis.

C<sub>4</sub> photosynthesis concentrates CO<sub>2</sub> for fixation by RuBisCO and is initiated by the carboxylation of phosphoenolpyruvate (PEP) by PEPCase to form a 4-carbon acid compound oxaloacetate (OAA) [198, 199]. CO<sub>2</sub> incorporated by this mechanism, lowers photorespiration and water use, improving photosynthetic efficiency under low CO<sub>2</sub>, dry and/or high temperature conditions [198, 199]. While CAM is highly prevalent in algae, C<sub>4</sub> photosynthesis is relatively infrequent [194]. For example, in the green chlorophyte algae C<sub>4</sub>-like photosynthesis has only been identified in a handful of lineages including the microalgae *Ostreococcus tauri*, *Micromonas* sp. CCMP1545 and *Micromonas* sp. RCC299 [9, 23], as well as the macroalgae *Udotea flabellum* [202], *Ulva prolifera* [203, 204] and *Ulva linza* [205, 206]. It has been postulated that CAM and C<sub>4</sub> photosynthesis likely arose in aquatic eukaryotes in response to the restricted availability of dissolved CO<sub>2</sub> in water, thereby maximising carbon acquisition and enhancing growth rates [207-209]. In *Ulva prolifera* for example, CO<sub>2</sub> is preferentially assimilated by the C<sub>3</sub> pathway under normal conditions but assimilated via a HCO<sub>3</sub><sup>-</sup> CAM mechanism under low CO<sub>2</sub> conditions and via a C<sub>4</sub> pathway under high-light irradiance [203, 204]. Further experimental work will be required to demonstrate the activity and localisation of the identified genes in CAM and C<sub>4</sub> photosynthesis. Nevertheless, we think it likely that the genetic capacity for CAM and C<sub>4</sub> photosynthesis in *P. SENEW3* could have been acquired as a response to its highly variable aquatic environment.

## Conclusion

New techniques and methodologies now enable improved characterisation of genomes, their features, and their structures. In this study we re-investigated the genome of the PPE alga *P. SENEW3* utilizing short-read, long-read, and Hi-C chromosome conformation capture sequencing, resulting in a thoroughly annotated full chromosomal assembly. The *P. SENEW3* genome contained a low repetitive sequence content and a minimal complement of genes, with few duplications. Hi-C contact mapping revealed the *P. SENEW3* genome to have a Rab1-like structure and enabled identification of putative centromeres and a candidate CenH3 variant. These results form the basis for further centromere characterisation and possible development of native plasmids and artificial chromosomes. Despite detection of additional selenoproteins, overall selenoprotein content remained low, especially in comparison to the picoeukaryotic

Mamiellophyceae algae. We speculate this could be explained by adaptation from freshwater to saline environments during their evolutionary history. Finally, alongside the C<sub>3</sub> photosynthetic pathway, we identified the necessary genes for both CAM and NAD-ME/PCK type C<sub>4</sub> photosynthesis. C<sub>4</sub> photosynthesis is still uncommon in chlorophyte algae, the genetic capability for it may partially explain the unique robustness of this organism. This new fully annotated chromosomal assembly provides an improved resource for the molecular study of this unique organism and offers further insight into its unique environmental adaptations.

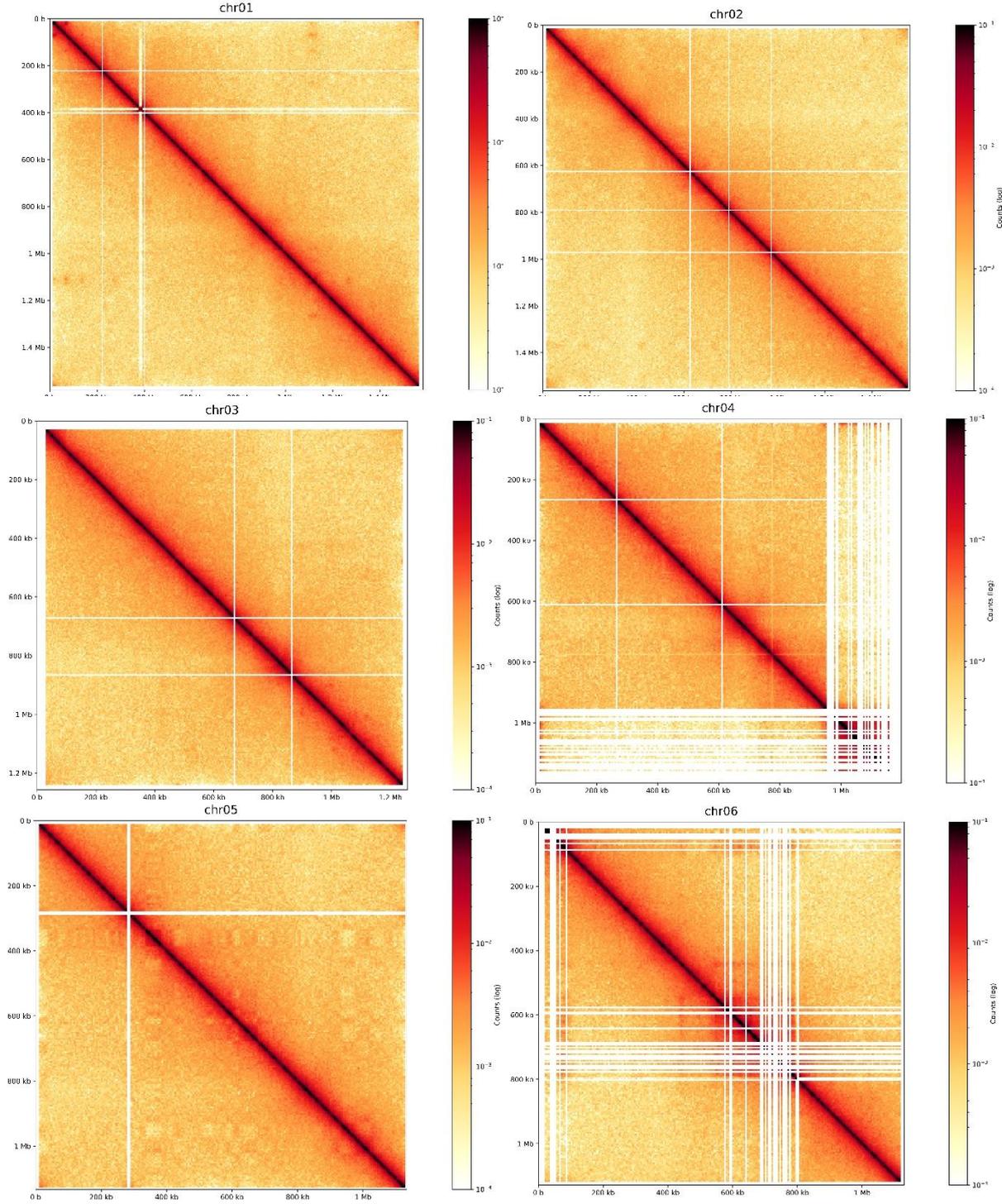
## **Author contributions**

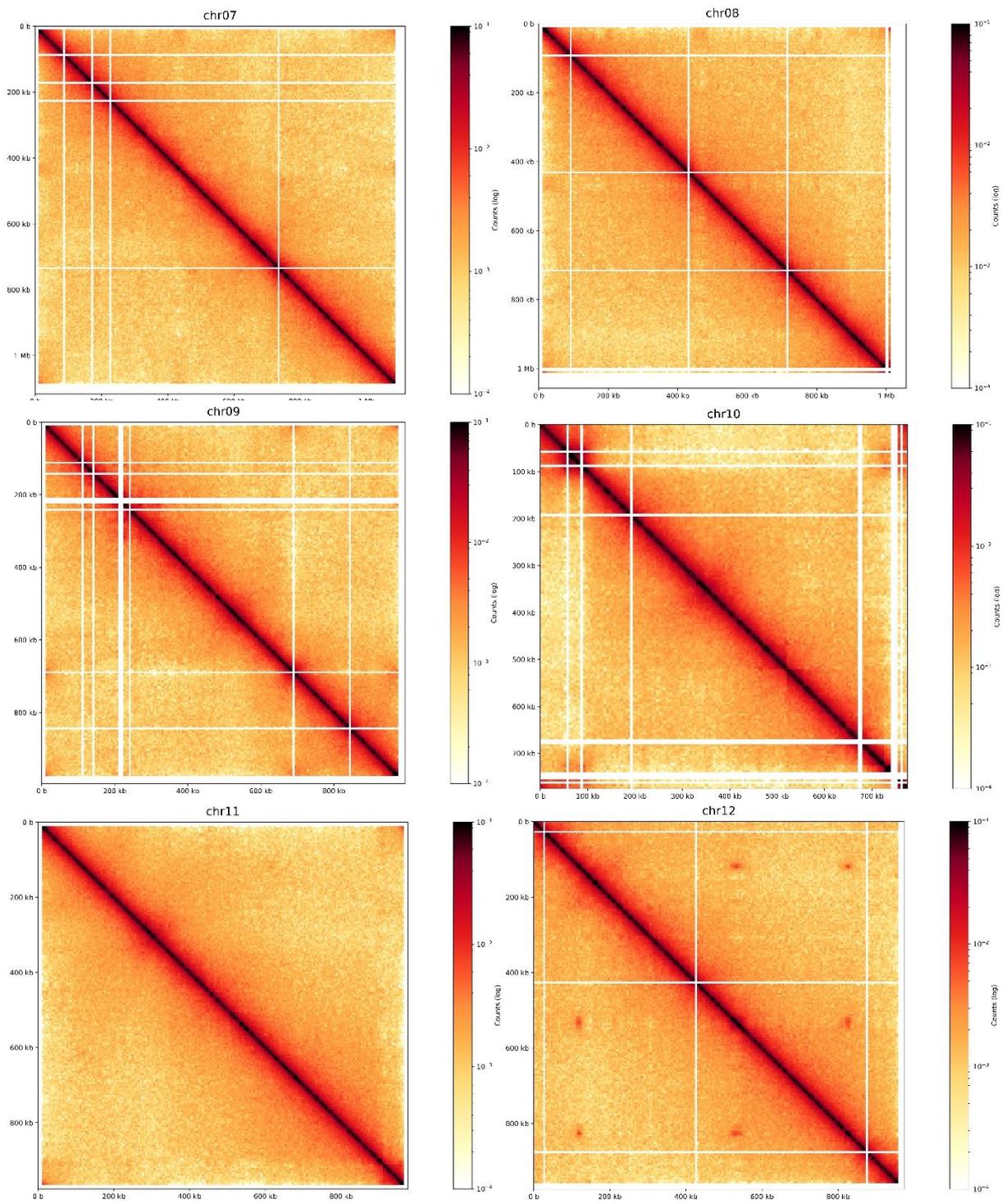
PdR, HG, BP and IP developed the concept and design of the project. PdR undertook sample preparation, sequencing, initial and final assembly, annotation, pulsed field gel electrophoresis, bioinformatic and Hi-C data analysis and produced manuscript and figures. HM undertook Hi-C sequencing, re-assembly, data analysis and contributed to initial manuscript drafts and created Hi-C figures. RSKW assisted with pulsed field gel electrophoresis and RDW with draft genome assembly and polishing. GJS assisted with bioinformatic analysis and data submission. BP provided algal strains, cultivation techniques and project consultation. PdR, HM, RSKW, BP and IP provided critical review of the manuscript.

## **Acknowledgments**

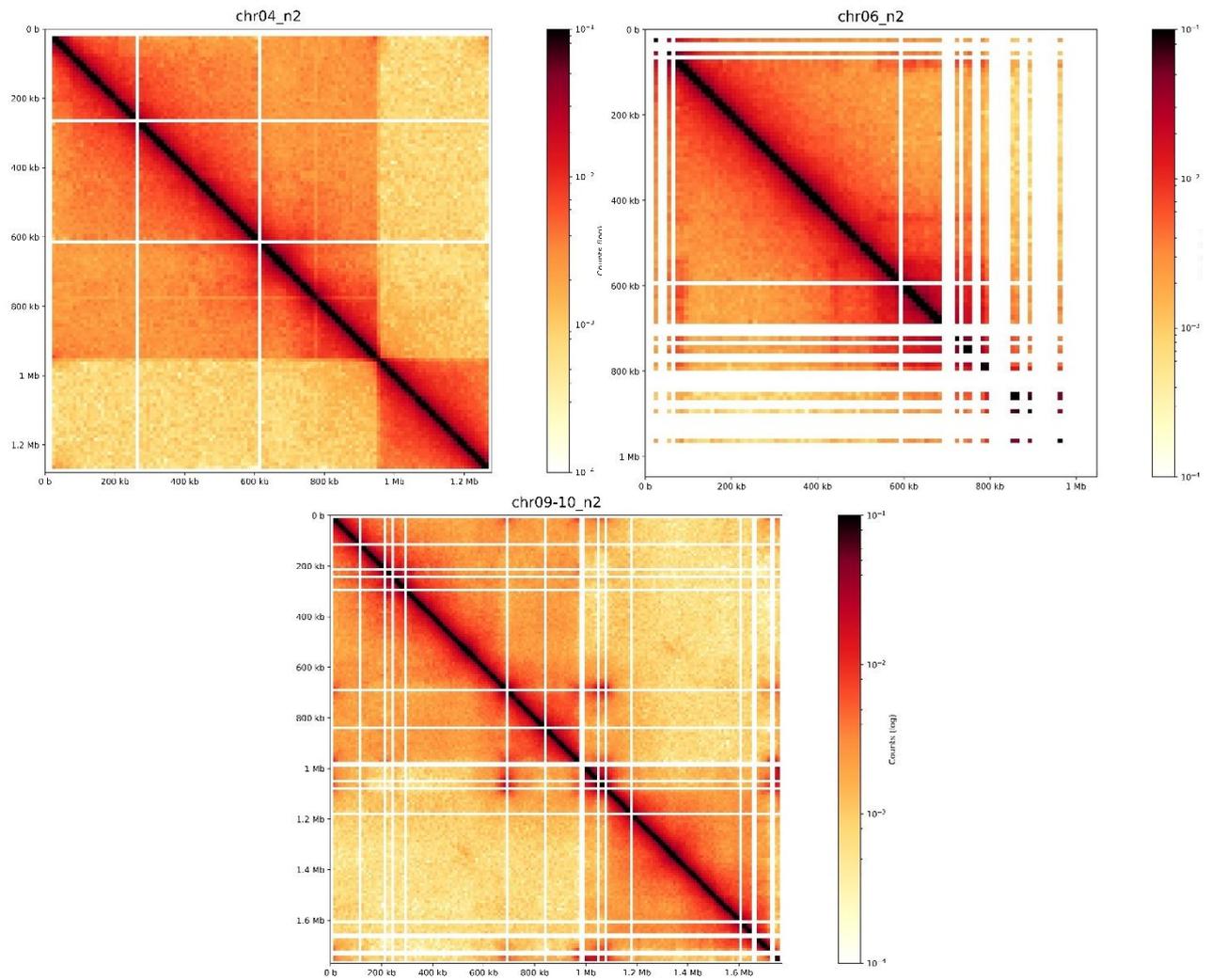
This research was supported by the Australian Research Council Centre of Excellence in Synthetic Biology (project number CE200100029). External support for Macquarie University's Synthetic Biology initiative is acknowledged from Bioplatforms Australia, the New South Wales (NSW) Chief Scientist and Engineer, and the NSW Government's Department of Primary Industries. PdR acknowledges receipt of an Australian Government Research Training Program (RTP) scholarship.

# Supplementary

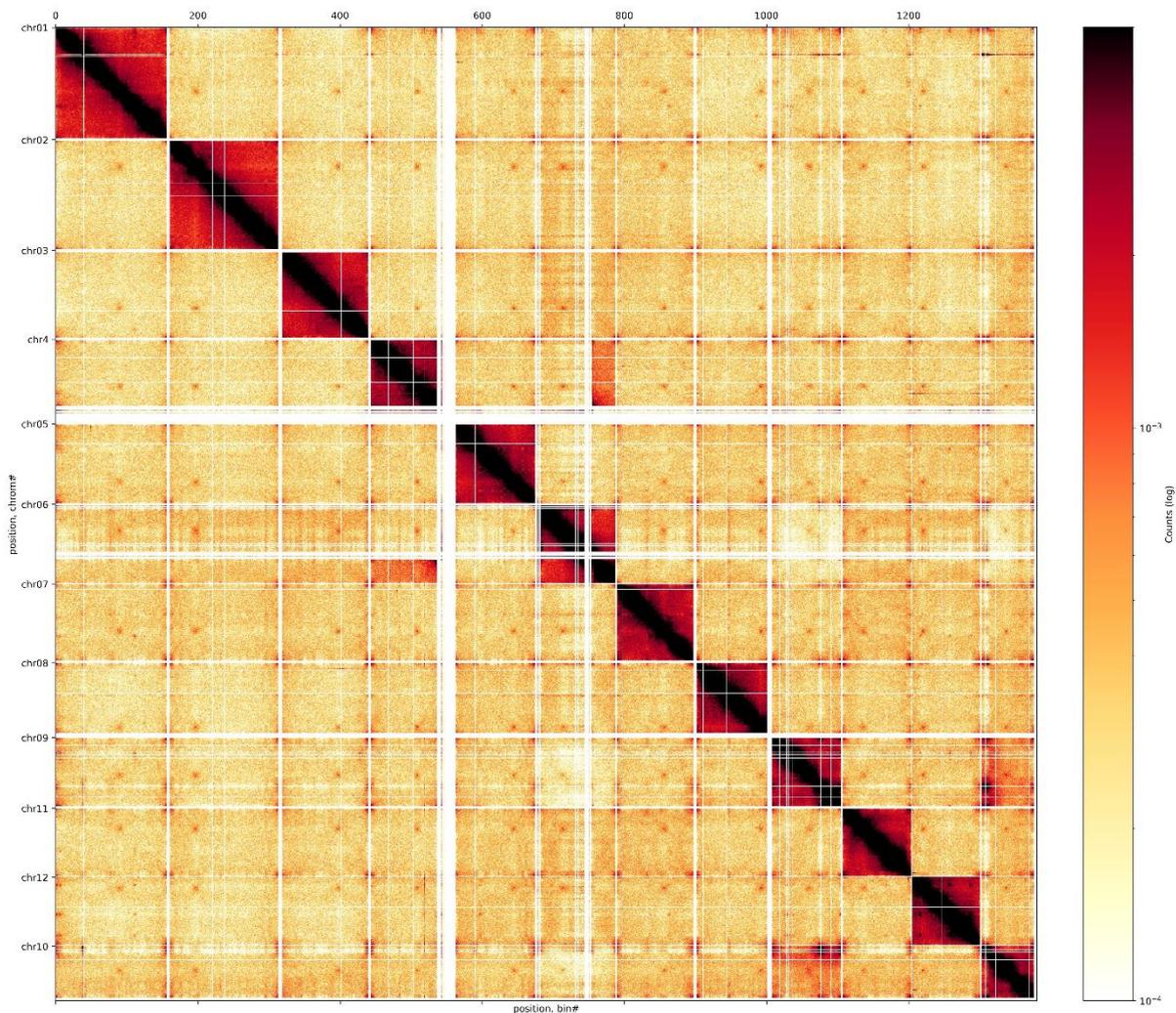




**Supplementary Figure 1.** Hi-C contact heat maps (balanced) at 5 Kbp resolution for nuclear genome primary assembly chromosomes (chr) 1 – 12. Scale (log) shows interaction frequency count per 5 Kbp bin.



**Supplementary Figure 2.** Hi-C contact heat maps (balanced) at 5 Kbp resolution for nuclear genome alternative "n2" haplotype variant chromosomes (chr) chr04\_n2, chr06\_n2 and chr09-10\_n2). Scale (log) shows interaction frequency count per 5 Kbp bin.



**Supplementary Figure 3.** Hi-C contact heat maps (balanced) of final full primary nuclear genome (chromosomes 1 – 12) at 10 Kbp resolution. Scale (log) shows interaction frequency count per 10 Kbp bin.

**Supplementary Table 1.** CHEF Mapper XA PFGE system run settings for separation of *Picochlorum* sp. SENEW3 chromosomal DNA.

Setting	Stage / Slice 1	Stage / Slice 2	Stage / Slice 3	Stage / Slice 4
Buffer	0.5 X TBE	0.5 X TBE	0.5 X TBE	0.5 X TBE
Temperature	12 °C	12 °C	12 °C	12 °C
Voltage	6	6	6	6
Calibration factor	1	1	1	1
Included angle	120°	120°	120°	120°
Switch time (linear ramped)	60 – 120 s	122.5 – 130 s	132.5 – 140 s	142.5 – 150 s
Run time	24 h	28 h	32 h	36 h

**Supplementary Tables 2 - 22**

Additional supplementary data tables are available in the supplementary data spreadsheet.

## References

1. **Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H et al.** Phylogeny and molecular evolution of the green algae. *Critical Reviews in Plant Sciences* 2012;31(1):1-46.
2. **Leliaert F, Tronholm A, Lemieux C, Turmel M, DePriest MS et al.** Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Scientific Reports* 2016;6(1):25367.
3. **Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A et al.** The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol* 2012;13(5):R39.
4. **Hirooka S, Hirose Y, Kanesaki Y, Higuchi S, Fujiwara T et al.** Acidophilic green algal genome provides insights into adaptation to an acidic environment. *Proceedings of the National Academy of Sciences* 2017;114(39):E8304-E8313.
5. **Price DC, Chan CX, Yoon HS, Yang EC, Qiu H et al.** *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science* 2012;335(6070):843-847.
6. **Suzuki S, Endoh R, Manabe R-i, Ohkuma M, Hirakawa Y.** Multiple losses of photosynthesis and convergent reductive genome evolution in the colourless green algae *Prototheca*. *Scientific Reports* 2018;8(1):940.
7. **Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ et al.** The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 2007;318(5848):245-250.
8. **Featherston J, Arakaki Y, Hanschen ER, Ferris PJ, Michod RE et al.** The 4-celled *Tetrabaena socialis* nuclear genome reveals the essential components for genetic control of cell number at the origin of multicellularity in the volvocine lineage. *Molecular Biology and Evolution* 2017;35(4):855-870.
9. **Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP et al.** Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* 2009;324(5924):268-272.
10. **Roth MS, Cokus SJ, Gallaher SD, Walter A, Lopez D et al.** Chromosome-level genome assembly and transcriptome of the green alga *Chromochloris zofingiensis* illuminates astaxanthin production. *Proceedings of the National Academy of Sciences* 2017;114(21):E4296-E4305.
11. **Potter D, Lajeunesse TC, Saunders GW, Anderson RA.** Convergent evolution masks extensive biodiversity among marine coccoid picoplankton. *Biodiversity & Conservation* 1997;6(1):99-107.
12. **Vargas Cd, Audic S, Henry N, Decelle J, Mahé F et al.** Eukaryotic plankton diversity in the sunlit ocean. *Science* 2015;348(6237):1261605.
13. **Vaulot D, Eikrem W, Viprey M, Moreau H.** The diversity of small eukaryotic phytoplankton ( $\leq 3 \mu\text{m}$ ) in marine ecosystems. *FEMS Microbiology Reviews* 2008;32(5):795-820.
14. **Caron DA, Countway PD, Jones AC, Kim DY, Schnetzer A.** Marine protistan diversity. *Annual Review of Marine Science* 2012;4(1):467-493.
15. **Metz S, Lopes dos Santos A, Berman MC, Bigeard E, Licursi M et al.** Diversity of photosynthetic picoeukaryotes in eutrophic shallow lakes as assessed by combining flow cytometry cell-sorting and high throughput sequencing. *FEMS Microbiology Ecology* 2019;95(5).
16. **Wang F, Xie Y, Wu W, Sun P, Wang L et al.** Picoeukaryotic diversity and activity in the northwestern Pacific Ocean based on rDNA and rRNA high-throughput sequencing. *Frontiers in Microbiology*, Original Research 2019;9(3259).
17. **Raven JA.** The twelfth Tansley Lecture. Small is beautiful: the picophytoplankton. *Functional Ecology* 1998;12(4):503-513.

18. **Fouilland E, Descolas-Gros C, Courties C, Collos Y, Vaquer A et al.** Productivity and growth of a natural population of the smallest free-living eukaryote under nitrogen deficiency and sufficiency. *Microbial Ecology* 2004;48(1):103-110.
19. **Butcher RW.** Contributions to our knowledge of the smaller marine algae. *Journal of the Marine Biological Association of the United Kingdom* 1952;31(1):175-191.
20. **Henley WJ, Hironaka JL, Guillou L, Buchheim MA, Buchheim JA et al.** Phylogenetic analysis of the 'Nannochloris-like' algae and diagnoses of *Picochlorum oklahomensis* gen. et sp. nov. (Trebouxiophyceae, Chlorophyta). *Phycologia* 2004;43(6):641-652.
21. **da Roza PA, Goold HD, Paulsen IT.** *Picochlorum* sp. SENEW3. *Trends in Genetics* 2021.
22. **Blanc-Mathieu R, Verhelst B, Derelle E, Rombauts S, Bouget F-Y et al.** An improved genome of the model marine alga *Ostreococcus tauri* unfolds by assessing Illumina *de novo* assemblies. *BMC Genomics* 2014;15(1):1103.
23. **Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ et al.** Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proceedings of the National Academy of Sciences* 2006;103(31):11647-11652.
24. **Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A et al.** The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences* 2007;104(18):7705-7710.
25. **Blanc-Mathieu R, Krasovec M, Hebrard M, Yau S, Desgranges E et al.** Population genomics of picophytoplankton unveils novel chromosome hypervariability. *Sci Adv* 2017;3(7):e1700239-e1700239.
26. **Foflonker F, Price DC, Qiu H, Palenik B, Wang S et al.** Genome of the halotolerant green alga *Picochlorum* sp. reveals strategies for thriving under fluctuating environmental conditions. *Environmental Microbiology* 2015;17(2):412-426.
27. **Wang S, Lambert W, Giang S, Goericke R, Palenik B.** Microalgal assemblages in a poikilohaline pond. *Journal of Phycology* 2014;50(2):303-309.
28. **Foflonker F, Ananyev G, Qiu H, Morrison A, Palenik B et al.** The unexpected extremophile: tolerance to fluctuating salinity in the green alga *Picochlorum*. *Algal Research* 2016;16:465-472.
29. **Weissman JC, Likhogrud M, Thomas DC, Fang W, Karns DAJ et al.** High-light selection produces a fast-growing *Picochlorum celeri*. *Algal Res*, Article 2018;36:17-28.
30. **Watanabe K, Fujii K.** Isolation of high-level-CO<sub>2</sub>-preferring *Picochlorum* sp. strains and their biotechnological potential. *Algal Res*, Article 2016;18:135-143.
31. **Foflonker F.** Understanding the genomic basis of stress adaptation in *Picochlorum* green algae. 2018.
32. **Hazeem LJ, Bououdina M, Rashdan S, Brunet L, Slomianny C et al.** Cumulative effect of zinc oxide and titanium oxide nanoparticles on growth and chlorophyll a content of *Picochlorum* sp. *Environmental Science and Pollution Research*, journal article 2016;23(3):2821-2830.
33. **Dimier C, Corato F, Saviello G, Brunet C.** Photophysiological properties of the marine picoeukaryote *Picochlorum* RCC 237 (Trebouxiophyceae, Chlorophyta)<sup>1</sup>. *Journal of Phycology* 2007;43(2):275-283.
34. **Dogaris I, Brown TR, Loya B, Philippidis G.** Cultivation study of the marine microalga *Picochlorum oculatum* and outdoor deployment in a novel bioreactor for high-density production of algal cell mass. *Biomass and Bioenergy* 2016;89:11-23.
35. **Dahmen I, Chtourou H, Jebali A, Daassi D, Karray F et al.** Optimisation of the critical medium components for better growth of *Picochlorum* sp. and the role of stressful environments for higher lipid production. *Journal of the Science of Food and Agriculture* 2014;94(8):1628-1638.
36. **El-Kassas HY.** Growth and fatty acid profile of the marine microalga *Picochlorum* Sp. grown under nutrient stress conditions. *The Egyptian Journal of Aquatic Research* 2013;39(4):233-239.

37. **Tanhaemami M, Alizadeh E, Sanders CK, Marrone BL, Munsky B.** Using flow cytometry and multistage machine learning to discover label-free signatures of algal lipid accumulation. *Phys Biol*, Article 2019;16(5):12.
38. **Tran D, Giordano M, Louime C, Tran N, Vo T et al.** An isolated *Picochlorum* species for aquaculture, food, and biofuel. *N Am J Aquacult*, Article 2014;76(4):305-311.
39. **de la Vega M, Díaz E, Vila M, León R.** Isolation of a new strain of *Picochlorum* sp and characterization of its potential biotechnological applications. *Biotechnology Progress* 2011;27(6):1535-1543.
40. **Yang F, Xiang W, Sun X, Wu H, Li T et al.** A novel lipid extraction method from wet microalga *Picochlorum* sp. at room temperature. *Marine Drugs* 2014;12(3):1258-1270.
41. **Edmundson SJ, Huesemann MH.** The dark side of algae cultivation: characterizing night biomass loss in three photosynthetic algae, *Chlorella sorokiniana*, *Nannochloropsis salina* and *Picochlorum* sp. *Algal Research* 2015;12:470-476.
42. **Dinesh Kumar S, Santhanam P, Nandakumar R, Ananth S, Nithya P et al.** Bioremediation of shrimp (*Litopenaeus vannamei*) cultured effluent using copepod (*Oithona rigida*) and microalgae (*Picochlorum maculatum* & *Amphora* sp.)—An integrated approach. *Desalination and Water Treatment* 2016;57(54):26257-26266.
43. **Chen T-Y, Lin H-Y, Lin C-C, Lu C-K, Chen Y-M.** *Picochlorum* as an alternative to *Nannochloropsis* for grouper larval rearing. *Aquaculture* 2012;338-341:82-88.
44. **Dinesh Kumar S, Santhanam P, Ananth S, Kaviyarasan M, Nithya P et al.** Evaluation of suitability of wastewater-grown microalgae (*Picochlorum maculatum*) and copepod (*Oithona rigida*) as live feed for white leg shrimp *Litopenaeus vannamei* post-larvae. *Aquaculture International*, journal article 2017;25(1):393-411.
45. **Dinesh Kumar S, Santhanam P, Prabhavathi P, Kanimozhi B, Abirami M et al.** Optimal conditions for the treatment of shrimp culture effluent using immobilized marine microalga *Picochlorum maculatum* (PSDK01). *Proceedings of the National Academy of Sciences, India Section B: Biological Sciences*, journal article 2018;88(3):1177-1185.
46. **von Alvensleben N, Stookey K, Magnusson M, Heimann K.** Salinity tolerance of *Picochlorum atomus* and the use of salinity for contamination control by the freshwater cyanobacterium *Pseudanabaena limnetica*. *PLOS ONE* 2013;8(5):e63569.
47. **Dahlin LR, Gerritsen AT, Henard CA, Van Wychen S, Linger JG et al.** Development of a high-productivity, halophilic, thermotolerant microalga *Picochlorum renovo*. *Communications Biology* 2019;2(1):388.
48. **Dahlin LR, Guarnieri MT.** Development of the high-productivity marine microalga, *Picochlorum renovo*, as a photosynthetic protein secretion platform. *Algal Research* 2021;54:102197.
49. **Gonzalez-Esquer CR, Wright KT, Sudasinghe N, Carr CK, Sanders CK et al.** Demonstration of the potential of *Picochlorum soloecismus* as a microalgal platform for the production of renewable fuels. *Algal Research* 2019;43:101658.
50. **Krishnan A, Likhogrud M, Cano M, Edmundson S, Melanson JB et al.** *Picochlorum celeri* as a model system for robust outdoor algal growth in seawater. *Scientific Reports* 2021;11(1):11649.
51. **Barten R, van Workum D-JM, de Bakker E, Risse J, Kleisman M et al.** Genetic mechanisms underlying increased microalgal thermotolerance, maximal growth rate, and yield on light following adaptive laboratory evolution. *BMC Biology* 2022;20(1):242.
52. **Foflonker F, Mollegard D, Ong M, Yoon HS, Bhattacharya D.** Genomic analysis of *Picochlorum* species reveals how microalgae may adapt to variable environments. *Molecular Biology and Evolution* 2018;35(11):2702-2711.
53. **Venkatesan BM, Bashir R.** Nanopore sensors for nucleic acid analysis. *Nature Nanotechnology* 2011;6(10):615-624.

54. **Rhoads A, Au KF.** PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics* 2015;13(5):278-289.
55. **Dekker J, Rippe K, Dekker M, Kleckner N.** Capturing chromosome conformation. *Science* 2002;295(5558):1306-1311.
56. **Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y et al.** Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 2012;58(3):268-276.
57. **Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO et al.** A chromosome conformation capture ordered sequence of the barley genome. *Nature* 2017;544(7651):427-433.
58. **van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A et al.** Hi-C: a method to study the three-dimensional architecture of genomes. *JoVE* 2010(39):e1869.
59. **Oluwadare O, Highsmith M, Cheng J.** An overview of methods for reconstructing 3-d chromosome and genome structures from Hi-C data. *Biological Procedures Online* 2019;21(1):7.
60. **Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M.** BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* 2021;3(1):lqaa108.
61. **Garrison E, Marth G.** Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:12073907* 2012.
62. **Cantarel BL, Korf I, Robb SMC, Parra G, Ross E et al.** MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* 2008;18(1):188-196.
63. **Holt C, Yandell M.** MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 2011;12(1):491.
64. **Pham GM, Hamilton JP, Wood JC, Burke JT, Zhao H et al.** Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience* 2020;9(9).
65. **Wang J, Liu W, Zhu D, Hong P, Zhang S et al.** Chromosome-scale genome assembly of sweet cherry (*Prunus avium* L.) cv. Tieton obtained using long-read and Hi-C sequencing. *Horticulture Research* 2020;7(1):122.
66. **Ma D, Guo Z, Ding Q, Zhao Z, Shen Z et al.** Chromosome-level assembly of the mangrove plant *Aegiceras corniculatum* genome generated through Illumina, PacBio and Hi-C sequencing technologies. *Molecular Ecology Resources* 2021;21(5):1593-1607.
67. **Guillard RR, Ryther JH.** Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (Cleve) Gran. *Canadian Journal of Microbiology* 1962;8(2):229-239.
68. **Guillard RRL.** Culture of phytoplankton for feeding marine invertebrates. In: Smith WL, Chanley MH (editors). *Culture of marine invertebrate animals: Proceedings — 1st conference on culture of marine invertebrate animals greenport*. Boston, MA: Springer US; 1975. pp. 29-60.
69. **Keller MD, Selvin RC, Claus W, Guillard RRL.** Media for the culture of oceanic ultraphytoplankton. *Journal of Phycology* 1987;23(4):633-638.
70. **Sanchez F, Geffroy S, Norest M, Yau S, Moreau H et al.** Simplified transformation of *Ostreococcus tauri* using polyethylene glycol. *Genes* 2019;10(5):399.
71. **Hage AE, Houseley J.** Resolution of budding yeast chromosomes using pulsed-field gel electrophoresis. In: Makovets S (editor). *DNA Electrophoresis (Methods in Molecular Biology (Methods and Protocols))*. Totowa, NJ.: Humana Press; 2013. pp. 195-207.
72. **Sambrook J, Russell DW, Irwin CA, Laboratory CSH, Fund RECM et al.** *Molecular cloning: a laboratory manual*, 3rd edition ed: Cold Spring Harbor Laboratory Press; 2001.
73. **Jagielski T, Gawor J, Bakula Z, Zuchniewicz K, Żak I et al.** An optimized method for high quality DNA extraction from microalga *Prototheca wickerhamii* for genome sequencing. *Plant Methods* 2017;13:77.
74. **Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG et al.** Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research* 2004;14(6):1147-1159.

75. **Li H.** Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34(18):3094-3100.
76. **Vaser R, Sovic I, Nagarajan N, Sikic M.** Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Research* 2017;27(5):737-746.
77. 2018. Medaka: Sequence correction provided by ONT Research. <https://github.com/nanoporetech/medaka> [accessed December 2019].
78. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al.** The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25(16):2078-2079.
79. **Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V et al.** Twelve years of SAMtools and BCFtools. *Gigascience* 2021;10(2).
80. **Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A et al.** Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* 2014;9(11):e112963.
81. **Bonfield JK, Whitwham A.** Gap5—editing the billion fragment sequence assembly. *Bioinformatics* 2010;26(14):1699-1703.
82. **Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J et al.** HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* 2015;16(1):259.
83. **Abdennur N, Mirny LA.** Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* 2019;36(1):311-316.
84. **Hunter JD.** Matplotlib: a 2D graphics environment. *Computing in science & engineering* 2007;9(03):90-95.
85. **Seemann T.** Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30(14):2068-2069.
86. **Stanke M, Waack S.** Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 2003;19(suppl\_2):ii215-ii225.
87. **Keller O, Odronitz F, Stanke M, Kollmar M, Waack S.** Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics* 2008;9(1):278.
88. **Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM.** BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31(19):3210-3212.
89. **Korf I.** Gene finding in novel genomes. *BMC Bioinformatics* 2004;5(1):59.
90. **Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M.** Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research* 2005;33(20):6494-6506.
91. **Campbell MS, Law M, Holt C, Stein JC, Moghe GD et al.** MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations *Plant Physiology* 2013;164(2):513-524.
92. **Gonzalez-Esquer CR, Twary SN, Hovde BT, Starkenburg SR.** Nuclear, chloroplast, and mitochondrial genome sequences of the prospective microalgal biofuel strain *Picochlorum soloecismus*. *Genome Announcements* 2018;6(4).
93. **Coordinators NR.** Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 2018;46(D1):D8-d13.
94. **Consortium TU.** UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* 2018;47(D1):D506-D515.
95. **Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S et al.** The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* 2020;49(D1):D344-D354.
96. **Lowé TM, Chan PP.** tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research* 2016;44(W1):W54-57.

97. **Seppy M, Manni M, Zdobnov EM.** BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M (editor). *Gene Prediction (Methods in Molecular Biology)*. New York, NY: Humana; 2019. pp. 227-245.
98. **Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG et al.** RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* 2020;117(17):9451-9457.
99. **Tarailo-Graovac M, Chen N.** Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* 2009;25(1):4.10.11-14.10.14.
100. **Emms DM, Kelly S.** OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;20(1):238.
101. **Emms DM, Kelly S.** OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;16(1):157.
102. **Yin Y, Mao X, Yang J, Chen X, Mao F et al.** dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* 2012;40(Web Server issue):W445-W451.
103. **Zhang H, Yohe T, Huang L, Entwistle S, Wu P et al.** dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* 2018;46(W1):W95-W101.
104. **Elbourne LDH, Tetu SG, Hassan KA, Paulsen IT.** TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Research* 2017;45(D1):D320-D324.
105. **Weiß CL, Pais M, Cano LM, Kamoun S, Burbano HA.** nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* 2018;19(1):122.
106. **Kanehisa M, Sato Y, Morishima K.** BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of Molecular Biology* 2016;428(4):726-731.
107. **Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK et al.** eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* 2018;47(D1):D309-D314.
108. **Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J.** eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution* 2021;38(12):5825–5829.
109. **Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R et al.** The Reference Genome Sequence of *Saccharomyces cerevisiae*: Then and Now. *G3 Genes/Genomes/Genetics* 2014;4(3):389-398.
110. **Lynch M, Walsh B.** *The origins of genome architecture*: Sinauer associates Sunderland, MA; 2007.
111. **Gaut BS, Ross-Ibarra J.** Selection on major components of angiosperm genomes. *Science* 2008;320(5875):484-486.
112. **Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J.** DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* 2008;25(1):119-120.
113. **Leliaert F.** Green algae: Chlorophyta and Streptophyta. In: Schmidt TM (editor). *Encyclopedia of Microbiology (Fourth Edition)*. Oxford: Academic Press; 2019. pp. 457-468.
114. **Schurko AM, Logsdon JM, Jr.** Using a meiosis detection toolkit to investigate ancient asexual "scandals" and the evolution of sex. *Bioessays* 2008;30(6):579-589.
115. **Hofstatter PG, Ribeiro GM, Porfírio-Sousa AL, Lahr DJG.** The sexual ancestor of all eukaryotes: a defense of the "meiosis toolkit". *BioEssays* 2020;42(9):2000037.
116. **Morris JJ, Lenski RE, Zinser ER.** The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *mBio* 2012;3(2):e00036-00012.

117. **Wolf YI, Koonin EV.** Genome reduction as the dominant mode of evolution. *BioEssays* 2013;35(9):829-837.
118. **Derilus D, Rahman MZ, Pinero F, Massey SE.** Synergism between the Black Queen effect and the proteomic constraint on genome size reduction in the photosynthetic picoeukaryotes. *Scientific Reports* 2020;10(1):8918.
119. **Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T et al.** Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;326(5950):289-293.
120. **Rao Suhas SP, Huntley Miriam H, Durand Neva C, Stamenova Elena K, Bochkov Ivan D et al.** A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;159(7):1665-1680.
121. **Marie-Nelly H, Marbouty M, Cournac A, Flot J-F, Liti G et al.** High-quality genome (re)assembly using chromosomal contact data. *Nature Communications* 2014;5:5695-5695.
122. **Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO et al.** Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology* 2013;31(12):1119-1125.
123. **Kaplan N, Dekker J.** High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nature Biotechnology* 2013;31(12):1143-1147.
124. **Shan T, Yuan J, Su L, Li J, Leng X et al.** First genome of the brown alga *Undaria pinnatifida*: chromosome-level assembly using PacBio and Hi-C technologies. *Frontiers in Genetics, Data Report* 2020;11.
125. **Chen H, Chu JS-C, Chen J, Luo Q, Wang H et al.** Insights into the ancient adaptation to intertidal environments by red algae based on a genomic and multiomics investigation of *Neoporphyra haitanensis*. *Molecular Biology and Evolution* 2021;39(1).
126. **Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J et al.** Earth biogenome project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences* 2018;115(17):4325-4333.
127. **Zhou Y, Shearwin-Whyatt L, Li J, Song Z, Hayakawa T et al.** Platypus and echidna genomes reveal mammalian biology and evolution. *Nature* 2021;592(7856):756-762.
128. **Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G et al.** Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 2021;592(7856):737-746.
129. **Drinnenberg IA, deYoung D, Henikoff S, Malik HS.** Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. *eLife* 2014;3:e03676.
130. **Muller H, Gil J, Jr., Drinnenberg IA.** The impact of centromeres on spatial genome architecture. *Trends in Genetics* 2019;35(8):565-578.
131. **Marbouty M, Cournac A, Flot J-F, Marie-Nelly H, Mozziconacci J et al.** Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife* 2014;3:e03318-e03318.
132. **Marie-Nelly H, Marbouty M, Cournac A, Liti G, Fischer G et al.** Filling annotation gaps in yeast genomes using genome-wide contact maps. *Bioinformatics* 2014;30(15):2105-2113.
133. **Descorps-Declère S, Saguez C, Cournac A, Marbouty M, Rolland T et al.** Genome-wide replication landscape of *Candida glabrata*. *BMC Biology* 2015;13(1):69.
134. **Varoquaux N, Liachko I, Ay F, Burton JN, Shendure J et al.** Accurate identification of centromere locations in yeast genomes using Hi-C. *Nucleic Acids Research* 2015;43(11):5331-5339.
135. **Burrack LS, Hutton HF, Matter KJ, Clancey SA, Liachko I et al.** Neocentromeres provide chromosome segregation accuracy and centromere clustering to multiple loci along a *Candida albicans* chromosome. *PLOS Genetics* 2016;12(9):e1006317-e1006317.
136. **Talbert PB, Henikoff S.** What makes a centromere? *Experimental Cell Research* 2020;389(2):111895.

137. **Henikoff S, Ahmad K, Malik HS.** The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 2001;293(5532):1098-1102.
138. **Black BE, Bassett EA.** The histone variant CENP-A and centromere specification. *Current Opinion in Cell Biology* 2008;20(1):91-100.
139. **Cremer T, Cremer M.** Chromosome territories. *Cold Spring Harbor Perspectives in Biology* 2010;2(3).
140. **Szabo Q, Bantignies F, Cavalli G.** Principles of genome folding into topologically associating domains. *Sci Adv* 2019;5(4):eaaw1668.
141. **Belton J-M, Lajoie Bryan R, Audibert S, Cantaloube S, Lassadi I et al.** The conformation of yeast chromosome III is mating type dependent and controlled by the recombination enhancer. *Cell Reports* 2015;13(9):1855-1867.
142. **Mercy G, Mozziconacci J, Scolari VF, Yang K, Zhao G et al.** 3D organization of synthetic and scrambled chromosomes. *Science* 2017;355(6329):eaaf4597.
143. **Shao Y, Lu N, Wu Z, Cai C, Wang S et al.** Creating a functional single-chromosome yeast. *Nature* 2018.
144. **Li M, Fine RD, Dinda M, Bekiranov S, Smith JS.** A Sir2-regulated locus control region in the recombination enhancer of *Saccharomyces cerevisiae* specifies chromosome III structure. *PLOS Genetics* 2019;15(8):e1008339.
145. **Hoencamp C, Dudchenko O, Elbatsh AMO, Brahmachari S, Raaijmakers JA et al.** 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science* 2021;372(6545):984-989.
146. **Ijdo JW, Baldini A, Ward DC, Reeders ST, Wells RA.** Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proceedings of the National Academy of Sciences* 1991;88(20):9051-9055.
147. **Cicconardi F, Lewis JJ, Martin SH, Reed RD, Danko CG et al.** Chromosome fusion affects genetic diversity and evolutionary turnover of functional loci but consistently depends on chromosome size. *Molecular Biology and Evolution* 2021;38(10):4449-4462.
148. **Luo J, Sun X, Cormack BP, Boeke JD.** Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. *Nature* 2018;560(7718):392-396.
149. **Sears ER, Câmara A.** A transmissible dicentric chromosome. *Genetics* 1952;37(2):125-135.
150. **Luo MC, Deal KR, Akhunov ED, Akhunova AR, Anderson OD et al.** Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proceedings of the National Academy of Sciences* 2009;106(37):15780-15785.
151. **Pobiega S, Marcand S.** Dicentric breakage at telomere fusions. *Genes Dev* 2010;24(7):720-733.
152. **Barra V, Fachinetti D.** The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. *Nature Communications* 2018;9(1):4340-4340.
153. **Becker SA, Spreafico R, Kit JL, Brown R, Likhogrud M et al.** Phased Diploid Genome Sequence for the Fast-Growing Microalga *Picochlorum celeri*. *Microbiol Resour Announc* 2020;9(20):e00087-00020.
154. **De Storme N, Mason A.** Plant speciation through chromosome instability and ploidy change: cellular mechanisms, molecular factors and evolutionary relevance. *Current Plant Biology* 2014;1:10-33.
155. **Feulner P, De-Kayne R.** Genome evolution, structural rearrangements and speciation. *Journal of Evolutionary Biology* 2017;30(8):1488-1490.
156. **Rose MD, Novick P, Thomas JH, Botstein D, Fink GR.** A *Saccharomyces cerevisiae* genomic plasmid bank based on a centromere-containing shuttle vector. *Gene* 1987;60(2):237-243.

157. **Christianson TW, Sikorski RS, Dante M, Shero JH, Hieter P.** Multifunctional yeast high-copy-number shuttle vectors. *Gene* 1992;110(1):119-122.
158. **Clarke L, Carbon J.** Isolation of a yeast centromere and construction of functional small circular chromosomes. *Nature* 1980;287(5782):504-509.
159. **Murray AW, Szostak JW.** Construction of artificial chromosomes in yeast. *Nature* 1983;305(5931):189-193.
160. **Ikeno M, Grimes B, Okazaki T, Nakano M, Saitoh K et al.** Construction of YAC-based mammalian artificial chromosomes. *Nature Biotechnology* 1998;16(5):431-439.
161. **Karas BJ, Diner RE, Lefebvre SC, McQuaid J, Phillips APR et al.** Designer diatom episomes delivered by bacterial conjugation. *Nature Communications* 2015;6(1):6925.
162. **Diner RE, Noddings CM, Lian NC, Kang AK, McQuaid JB et al.** Diatom centromeres suggest a mechanism for nuclear DNA acquisition. *Proceedings of the National Academy of Sciences* 2017;114(29):E6015-E6024.
163. **Diner RE, Bielinski VA, Dupont CL, Allen AE, Weyman PD.** Refinement of the diatom episome maintenance sequence and improvement of conjugation-based DNA delivery methods. *Frontiers in Bioengineering and Biotechnology, Methods* 2016;4.
164. **Jiang L, Lu Y, Zheng L, Li G, Chen L et al.** The algal selenoproteomes. *BMC Genomics* 2020;21(1):699-699.
165. **Serrão VHB, Silva IR, da Silva MTA, Scortecci JF, de Freitas Fernandes A et al.** The unique tRNA<sup>Sec</sup> and its role in selenocysteine biosynthesis. *Amino Acids* 2018;50(9):1145-1167.
166. **Vanda Papp L, Lu J, Holmgren A, Khanna KK.** From selenium to selenoproteins: synthesis, identity, and their role in human health. *Antioxidants & Redox Signaling* 2007;9(7):775-806.
167. **Turanov AA, Xu X-M, Carlson BA, Yoo M-H, Gladyshev VN et al.** Biosynthesis of selenocysteine, the 21st amino acid in the genetic code, and a novel pathway for cysteine biosynthesis. *Advances in Nutrition* 2011;2(2):122-128.
168. **Novoselov SV, Rao M, Onoshko NV, Zhi H, Kryukov GV et al.** Selenoproteins and selenocysteine insertion system in the model plant cell system, *Chlamydomonas reinhardtii*. *The EMBO Journal* 2002;21(14):3681-3693.
169. **Rayman MP.** Selenium and human health. *The Lancet* 2012;379(9822):1256-1268.
170. **Zoidis E, Seremelis I, Kontopoulos N, Danezis GP.** Selenium-dependent antioxidant enzymes: actions and properties of selenoproteins. *Antioxidants* 2018;7(5):66.
171. **Mariotti M, Salinas G, Gabaldón T, Gladyshev VN.** Utilization of selenocysteine in early-branching fungal phyla. *Nature Microbiology* 2019;4(5):759-765.
172. **Liang H, Wei T, Xu Y, Li L, Kumar Sahu S et al.** Phylogenomics provides new insights into gains and losses of selenoproteins among Archaeplastida. *Int J Mol Sci* 2019;20(12):3020.
173. **Lobanov AV, Fomenko DE, Zhang Y, Sengupta A, Hatfield DL et al.** Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. *Genome Biol* 2007;8(9):R198.
174. **Bock C, Krienitz L, Pröschold T.** Taxonomic reassessment of the genus *Chlorella* (Trebouxiophyceae) using molecular signatures (barcodes), including description of seven new species. *Fottea*, journal article 2011;11(2):293-312.
175. **Formighieri C, Cazzaniga S, Kuras R, Bassi R.** Biogenesis of photosynthetic complexes in the chloroplast of *Chlamydomonas reinhardtii* requires ARSA1, a homolog of prokaryotic arsenite transporter and eukaryotic TRC40 for guided entry of tail-anchored proteins. *The Plant Journal* 2013;73(5):850-861.
176. **Blaby-Haas CE, Merchant SS.** The ins and outs of algal metal transport. *Biochimica et Biophysica Acta* 2012;1823(9):1531-1552.

177. **Marchand J, Heydarizadeh P, Schoefs B, Spetea C.** Ion and metabolite transport in the chloroplast of algae: lessons from land plants. *Cellular and Molecular Life Sciences* 2018;75(12):2153-2176.
178. **Goswami RK, Agrawal K, Shah MP, Verma P.** Bioremediation of heavy metals from wastewater: a current perspective on microalgae-based future. *Letters in Applied Microbiology* 2021.
179. **Ali H, Khan E, Sajad MA.** Phytoremediation of heavy metals—Concepts and applications. *Chemosphere* 2013;91(7):869-881.
180. **Almomani F, Bhosale RR.** Bio-sorption of toxic metals from industrial wastewater by algae strains *Spirulina platensis* and *Chlorella vulgaris*: application of isotherm, kinetic models and process optimization. *Science of The Total Environment* 2021;755:142654.
181. **Shanab S, Essa A, Shalaby E.** Bioremoval capacity of three heavy metals by some microalgae species (Egyptian Isolates). *Plant Signal Behav* 2012;7(3):392-399.
182. **Yan A, Wang Y, Tan SN, Mohd Yusof ML, Ghosh S et al.** Phytoremediation: a promising approach for revegetation of heavy metal-polluted land. *Frontiers in Plant Science, Review* 2020;11(359).
183. **Goswami RK, Agrawal K, Verma P.** Phycoremediation of nitrogen and phosphate from wastewater using *Picochlorum* sp.: a tenable approach. *Journal of Basic Microbiology* 2021:1 - 17.
184. **Levina N, Töttemeyer S, Stokes NR, Louis P, Jones MA et al.** Protection of *Escherichia coli* cells against extreme turgor by activation of MscS and MscL mechanosensitive channels: identification of genes required for MscS activity. *The EMBO Journal* 1999;18(7):1730-1737.
185. **Kung C, Martinac B, Sukharev S.** Mechanosensitive channels in microbes. *Annual Review of Microbiology* 2010;64(1):313-329.
186. **Booth IR, Blount P.** The MscS and MscL families of mechanosensitive channels act as microbial emergency release valves. *Journal of Bacteriology* 2012;194(18):4802-4809.
187. **Basu D, Haswell ES.** Plant mechanosensitive ion channels: an ocean of possibilities. *Current Opinion in Plant Biology* 2017;40:43-48.
188. **Deng Z, Maksaev G, Schlegel AM, Zhang J, Rau M et al.** Structural mechanism for gating of a eukaryotic mechanosensitive channel of small conductance. *Nature Communications* 2020;11(1):3690.
189. **Owji AP, Zhao Q, Ji C, Kittredge A, Hopiavuori A et al.** Structural and functional characterization of the bestrophin-2 anion channel. *Nature Structural & Molecular Biology* 2020;27(4):382-391.
190. **Tsunenari T, Sun H, Williams J, Cahill H, Smallwood P et al.** Structure-function analysis of the bestrophin family of anion channels\*. *Journal of Biological Chemistry* 2003;278(42):41114-41125.
191. **Qu Z, Hartzell HC.** Bestrophin Cl<sup>-</sup> channels are highly permeable to HCO<sub>3</sub><sup>-</sup>. *American Journal of Physiology-Cell Physiology* 2008;294(6):C1371-C1377.
192. **Reinfelder JR.** Carbon concentrating mechanisms in eukaryotic marine phytoplankton. *Annual Review of Marine Science* 2011;3(1):291-315.
193. **Mackinder LCM, Chen C, Leib RD, Patena W, Blum SR et al.** A spatial interactome reveals the protein organization of the algal CO<sub>2</sub>-concentrating mechanism. *Cell* 2017;171(1):133-147.e114.
194. **Chi S, Wu S, Yu J, Wang X, Tang X et al.** Phylogeny of C<sub>4</sub>-photosynthesis enzymes based on algal transcriptomic and genomic data supports an archaeal/proteobacterial origin and multiple duplication for most C<sub>4</sub>-related genes. *PLOS ONE* 2014;9(10):e110154-e110154.
195. **Mukherjee A, Lau CS, Walker CE, Rai AK, Prejean CI et al.** Thylakoid localized bestrophin-like proteins are essential for the CO<sub>2</sub> concentrating mechanism of *Chlamydomonas reinhardtii*. *Proceedings of the National Academy of Sciences* 2019;116(34):16915-16920.

196. **Falkowski PG.** The role of phytoplankton photosynthesis in global biogeochemical cycles. *Photosynthesis Research* 1994;39(3):235-258.
197. **Worden AZ, Nolan JK, Palenik B.** Assessing the dynamics and ecology of marine picophytoplankton: the importance of the eukaryotic component. *Limnology and Oceanography* 2004;49(1):168-179.
198. **Sheen J.** C<sub>4</sub> gene expression. *Annual Review of Plant Physiology and Plant Molecular Biology* 1999;50(1):187-217.
199. **Matsuoka M, Furbank RT, and HF, Miyao M.** Molecular engineering of C<sub>4</sub> photosynthesis. *Annual Review of Plant Physiology and Plant Molecular Biology* 2001;52(1):297-314.
200. **Kanehisa M, Sato Y.** KEGG Mapper for inferring cellular functions from protein sequences. *Protein Science* 2020;29(1):28-35.
201. **Kanehisa M, Sato Y, Kawashima M.** KEGG mapping tools for uncovering hidden features in biological data. *Protein Science* 2022;31(1):47-53.
202. **Reiskind JB, Bowes G.** The role of phosphoenolpyruvate carboxykinase in a marine macroalga with C<sub>4</sub>-like photosynthetic characteristics. *Proceedings of the National Academy of Sciences* 1991;88(7):2883-2887.
203. **Xu J, Fan X, Zhang X, Xu D, Mou S et al.** Evidence of coexistence of C<sub>3</sub> and C<sub>4</sub> photosynthetic pathways in a green-tide-forming alga, *Ulva prolifera*. *PLOS ONE* 2012;7(5):e37438-e37438.
204. **Liu D, Ma Q, Valiela I, Anderson DM, Keesing JK et al.** Role of C<sub>4</sub> carbon fixation in *Ulva prolifera*, the macroalga responsible for the world's largest green tides. *Communications Biology* 2020;3(1):494.
205. **Zhang X, Ye N, Liang C, Mou S, Fan X et al.** De novo sequencing and analysis of the *Ulva linza* transcriptome to discover putative mechanisms associated with its successful colonization of coastal ecosystems. *BMC Genomics* 2012;13(1):565.
206. **Xu J, Zhang X, Ye N, Zheng Z, Mou S et al.** Activities of principal photosynthetic enzymes in green macroalga *Ulva linza*: functional implication of C<sub>4</sub> pathway in CO<sub>2</sub> assimilation. *Science China Life Sciences* 2013;56(6):571-580.
207. **Bowes G, Rao SK, Estavillo GM, Reiskind JB.** C<sub>4</sub> mechanisms in aquatic angiosperms: comparisons with terrestrial C<sub>4</sub> systems. *Functional Plant Biology* 2002;29(3):379-392.
208. **Maberly SC, Madsen TV.** Freshwater angiosperm carbon concentrating mechanisms: processes and patterns. *Functional Plant Biology* 2002;29(3):393-405.
209. **Giordano M, Beardall J, Raven JA.** CO<sub>2</sub> concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. *Annual Review of Plant Biology* 2005;56(1):99-131.