Genome sequence, phylogenetic analysis, and structure-based annotation reveal metabolic potential of Chlorella sp. SLA-04

Calvin L.C. Goemann, Royce Wilkinson, William Henriques, Huyen Bui, Hannah M. Goemann, Ross P. Carlson, Sridhar Viamajala, Robin Gerlach, Blake Wiedenheft

Genome sequence, phylogenetic analysis, and structure-based annotation reveals metabolic potential of *Chlorella sp.* SLA-04

Calvin L. C. Goemann[a,c], Royce Wilkinson[a], William Henriques[a], Huyen Bui[b,c], Hannah M. Goemann[a,c], Ross P. Carlson[b,c], Sridhar Viamajala[d], Robin Gerlach[b,c], Blake Wiedenheft[a*]

[a]Department of Microbiology and Cell Biology, Montana State University, Bozeman, Montana, USA

[b]Department of Chemical and Biological Engineering, Montana State University, Bozeman, Montana, USA

[c]Center for Biofilm Engineering, Montana State University, Bozeman, Montana, USA

[d]Department of Chemical Engineering, University of Toledo, Toledo Ohio, USA

[*]Correspondence: bwiedenheft@gmail.com

Running Head: *Chlorella sp.* SLA-04 genomic analysis

Abstract

Algae are a broad class of photosynthetic eukaryotes that are phylogenetically and physiologically diverse. Most of the phylogenetic diversity has been inferred from 18S rDNA sequencing since there are only a few complete genomes available in public databases. Here we use ultra-long-read Nanopore sequencing to determine a gapless, telomere-to-telomere complete genome sequence of *Chlorella sp.* SLA-04, previously described as *Chlorella sorokiniana* SLA-04. *Chlorella sp.* SLA-04 is a green alga that grows to high cell density in a wide variety of environments – high and neutral pH, high and low alkalinity, and high and low salinity. SLA-04's ability to grow in high pH and high alkalinity media without external CO_2 supply is favorable for large-scale algal biomass production. Phylogenetic analysis performed using ribosomal DNA and conserved protein sequences consistently reveal that *Chlorella sp.* SLA-04 forms a distinct lineage from other strains of *Chlorella sorokiniana*. We complement traditional genome annotation methods with high throughput structural predictions and demonstrate that this approach expands functional prediction of the SLA-04 proteome. Genomic analysis of the SLA-04 genome identifies the genes capable of utilizing TCA cycle intermediates to replenish cytosolic acetyl-CoA pools for lipid production. We also identify a complete metabolic pathway for sphingolipid anabolism that may allow SLA-04 to readily adapt to changing environmental conditions and facilitate robust cultivation in mass production systems. Collectively, this work clarifies the phylogeny of *Chlorella sp*. SLA-04 within Trebouxiophyceae and demonstrates how structural predictions can be used to improve annotation beyond sequence-based methods.

1. Introduction

Algae are diverse photosynthetic eukaryotes that assimilate six times more nitrogen than terrestrial plants, perform 45% of global oxygenic photosynthesis, and 50% of total CO_2 fixation [1-4]. The phylogenetic and physiological diversity of algae is mirrored by the diversity of aquatic ecosystems where algae are found (e.g., temperatures 0-56ºC, pH 0-11.5, and salt concentrations 0.02-3M) [5, 6]. However, it is currently unclear how selective pressures from disparate aquatic environments shape algal genotypes and corresponding phenotypes.

Genomic analysis of algae across different environments is essential to understand how algae survive and adapt to changing conditions. As of September 2022, there are 170 green algae genomes in the National Center for Biotechnology Information (NCBI) database. Most of these genomes are highly fragmented and only 16 contain defined chromosome sequences. This resource limitation restricts efforts to confidently identify genotypes that drive functional diversity. Understanding the phylogenetic and functional diversity of algae will benefit from a broader collection of complete algal genomes that can be used for comparative analyses.

To date, most algal genomes have been sequenced using short-read sequencing platforms [7, 8]. While these platforms are accurate, PCR bias and short read lengths frequently result in incomplete assemblies and omission of repetitive regions, such as retroelements that can represent up to 90% of a eukaryotic genome [9-12]. In contrast, ultra-long-read sequencing platforms generate megabase-length reads that span repetitive regions, and these methods do not rely on PCR [13, 14]. While the error rates for long-read technologies are higher, consensus correction software significantly increases accuracy as a function of read depth [15, 16]. Long-read sequence data with >100x coverage often results in highly accurate chromosome level assemblies [17].

Here we determine a complete genome sequence of *Chlorella. sp.* SLA-04, previously described as *Chlorella sorokiniana* SLA-04, using ultra-long-read Nanopore sequencing. *Chlorella sp.* SLA-04 is an oleaginous green alga that was isolated from a brackish, alkaline lake (100mM HCO_3^-, pH 10) in eastern Washington (USA) [18]. SLA-04 has growth and composition characteristics that are attractive for biofuel production. Phylogenetic analysis performed using either ribosomal DNA or conserved protein sequences reveal that SLA-04 forms a distinct lineage from other strains of the same species. While sequenced-based annotation methods confidently assign putative function to a majority of the SLA-04 genes, 44.2% of the genes were classified as “hypothetical”. Using a high-throughput structure prediction software (i.e., Alphafold) we predict the 3D protein structures encoded by the “hypothetical” proteome. Structural predictions are queried against the protein data bank (PDB) to identify functional homologs and expand functional predictions for 677 SLA-04 genes that are widely conserved in algae and a subset of genes that are unique to *Chlorella sp.* SLA-04. Collectively, this work expands our phylogenetic and functional understanding of algae.

2. Methods

2.1 Strain information

2.1.1 Strain cultivation conditions

Chlorella sp. SLA-04 cultures were grown in Modified Bold’s Basal Medium pH 8.7 (Bolds) containing $NaNO_3$ (2.94 mM), K_2HPO_4 (1.43 mM), KH_2PO_4 (1.43 mM), $MgSO_4{\cdot}7H_2O$ (0.30 mM), $CaCl_2{\cdot}2H_2O$ (0.17 mM), NaCl (0.42 mM), 1 mL/L of EDTA solution, 1 mL/L iron solution, and 2 mL/L trace metal solution. The EDTA solution contained 50 g/L Na_2EDTA and 31 g/L KOH. The iron solution contained 4.98 g/L $FeSO_4{\cdot}7H_2O$ and 1 mL/L concentrated H_2SO_4. The trace metal solution contained H_3BO_3 (9.7 mM), $MnCl_2{\cdot}4H_2O$ (1.26 mM), $ZnCl_2$ (0.15 mM), $CuCl_2{\cdot}2H_2O$ (0.11 mM), $Na_2MoO_4{\cdot}2H_2O$ (0.07 mM), $CoCl_2{\cdot}6H_2O$ (0.06 mM), $NiCl_2{\cdot}6H_2O$ (0.04 mM), V_2O_5 (0.01 mM) and KBr (0.08 mM) [19]. Media was titrated to pH 8.7 with NaOH. Cultures were grown at 23°C with 400 $\mu mol{\cdot}m^{-2}{\cdot}s^{-1}$ light intensity cycling with a 14:10 hour light/dark cycle using T5 fluorescent tubes.

2.1.2 Axenic isolation

An axenic culture of *Chlorella sp.* SLA-04 was generated through antibiotic treatment modified from Mustapa et al. [20]. Briefly, clonal isolates of SLA-04 were streaked onto plates containing Bolds agar with 700 µg/mL ampicillin and 200 µg/mL cefotaxime. Individual colonies were picked and grown in liquid Bolds. Plate antibiotic treatment and isolation was repeated a second time and algal colonies were picked and grown in liquid Bolds with 10 µg/mL tetracycline. Axenic cultures were confirmed using amplicon sequencing of the V4 region of the 16S rRNA on an Illumina MiSeq platform (**Sup. Fig. 1**) [21]. A small percentage of 16S rRNA reads within the axenic culture mapped to *Ralstonia* species, but *Ralstonia* sequences are common contaminants of DNA purification columns [22]. Axenic stock cultures of SLA-04 were maintained on agar slants and in liquid culture of Bolds with 10 µg/mL tetracycline.

2.2 Genome sequencing and assembly

2.2.1 High molecular weight DNA extraction

Chlorella sp. SLA-04 was grown in Bolds to log phase. Cells from 100 mL of culture were spun down at 3,000 x g for 10 minutes and transferred to a liquid N_2 chilled mortar. The pellet was ground for 30 minutes with liquid N_2 added as needed. Ground algal biomass was transferred to 15 mL of lysis buffer (20 mM EDTA pH 8.25, 10 mM Tris, 500 mM guanidine HCl, 200 mM NaCl, 1% w/w Triton X-100, and 9 mg hemicellulase) and incubated at 37ºC for 1 hour. To remove RNA, RNase A was added to a final concentration of 20 µg/mL and the mixture was incubated for an additional 30 minutes at 37°C. To remove proteins, Proteinase K (12 µg) was added and the mixture was incubated at 50°C for 2 hours. Samples were pelleted at 15,000 x g at 4°C for 20 minutes and the supernatant was added to an equilibrated Qiagen Genomic-tip 100/G gravity column (Cat. No.10243). DNA was washed and eluted per the manufacturer's instructions. DNA was precipitated with 3.5 mL molecular grade isopropyl alcohol (IPA). The DNA pellet was collected by centrifugation (5850 x g) at 4°C for 30 minutes and resuspended in molecular grade water.

2.2.2 RNA extraction

Chlorella sp. SLA-04 RNA was extracted using the Qiagen RNeasy powerplant RNA isolation kit (Cat. No.13500-50) with modifications. SLA-04 culture was grown to log phase ($2x10^7$ cells/mL) in Bolds. Cells from 200 mL of culture were pelleted at 3,000 x g for 5 minutes at 4°C and the pellet was ground for 10 minutes in a liquid N_2 chilled mortar using a pestle before being transferred to 1 mL Trizol in a PowerBead tube provided in the Qiagen RNeasy powerplant kit. Potassium acetate (1 M) was added to a final concentration of 0.2 M to improve polysaccharide precipitation. Phenolic separation solution (50 µL) from the Qiagen powerplant kit was added. The sample was shaken at 6.5 meters per second for 60 seconds using a MP Biomedicals FastPrep-24 Classic bead beater, incubated at room temp for 20 minutes, and centrifuged at 13,000 x g for 10 minutes at 4°C. Protein and cell debris was removed by adding 200 µL chloroform, shaking for 30 seconds, centrifuging at 13,000 x g for 10 minutes at 4°C and transferring the aqueous phase to a new tube. RNA was extracted and purified from the aqueous phase using the RNeasy powerplant RNA isolation kit per manufacturer's instructions starting at the inhibitor removal step (step 6). All glassware was washed with 2% diethylpyrocarbonate (DEPC) in water (v/v) and autoclaved before use.

2.2.3 DNA and RNA sequencing

The *Chlorella sp.* SLA-04 genome was sequenced using Nanopore sequencing. High molecular weight DNA was sequenced using a MinION sequencer and the ligation sequencing kit (SQK-LSK109 by Oxford Nanopore) per manufacturer's instructions. Briefly, 1 µg of genomic DNA was used for library preparation that included end preparation and ligation to Nanopore adapters. 350 ng of the DNA library was used and sequenced on the MinION sequencer. Sequencing generated 1.97 M reads totaling 5.64 gigabases (Gb) of data.

Chlorella sp. SLA-04 mRNA sequencing was performed using a MinION sequencer and the direct RNA sequencing kit (SQK-RNA002 by Oxford Nanopore) per manufacturer's instructions. mRNA was isolated and purified from total RNA using magnetic Dynabeads coated with Oligo$(dT)_{25}$ adapters. Purification using magnetic beads yielded 338 ng of mRNA from 22 µg of total RNA. 338 ng of mRNA was used for Nanopore library preparation that included reverse transcription of cDNA scaffold and ligation of Nanopore adapters. 168 ng of mRNA library was used and sequenced on the MinION Nanopore sequencer. Sequencing generated 1.12 M reads totaling 1.35 Gb of data.

2.2.4 Genome assembly and annotation

Nanopore raw DNA reads (5.64 GB) were basecalled and analyzed for quality using MinKNOW 4.1.22 software in fast basecalling mode. Quality reads were submitted to CANU v1.9, FLYE v2.5-g0c3de5b, and Miniasm v0.3-r179 programs for long-read genome assembly [23-25]. CANU, FLYE, and Miniasm assembly pipelines each generated contig level assemblies with minor gaps. Manual curation of assemblies was performed by aligning contigs from the three assembly pipelines using minimap2 v2.17-

r954-dirty [26]. Each gap within the individual assemblies was completely bridged by contigs from at least one other genome assembly. Assemblies were merged to form a chromosome level genome with no gaps. Chromosomes were considered complete if they had flanking telomeric repeats $(CCCTAAA)_n$ on both termini with no gaps. Coverage depth of DNA reads was calculated using samtools v1.13 [27].

RepeatMasker v4.0.9 was used to identify repetitive elements in the *Chlorella sp.* SLA-04 genome using Repbase and Dfam as reference databases [28-30]. LTRharvest v1.5.10 was used to identify and fetch the sequences of long-terminal repeats in the SLA-04 genome by performing a six open reading frame (ORF) translation of the genome and scanning each ORF longer than 75 amino acids for homology to LTR retrotransposon protein domains from Pfam [31-33]. To identify active retrotransposons, we extracted 13 sequences that contained *gag* and *pol* domains flanked by long-terminal repeats and aligned them in MAFFT v7.429 [34]. To verify LTR retrotransposon families, we aligned the capsid domains of diverse families of LTR retrotransposons from Repbase and those identified in SLA-04. SLA-04 retrotransposons clustered with retrotransposons from Copia retrotransposons found in plants.

Raw mRNA Nanopore reads were mapped to the SLA-04 genome using minimap2 and a consensus sequence and transcript counts were generated using Stringtie v2.0 and Gffread v0.11.6 [35, 36]. The mRNA consensus data was used to aid the *de novo* annotation prediction software MAKER v3.01.03 to generate a transcriptome of SLA-04 [37, 38]. Gene functions were assigned by submitting protein sequence to the Kyoto Encyclopedia for Genes and Genomes (KEGG) Automatic Annotation Server (KAAS) and Eggnog mapper webservers with default settings [39, 40]. Genome completeness was assessed by BUSCO v5.4.2 using 1519 single copy orthologs found in the Chlorophyta clade [41, 42]. A schematic of the SLA-04 genome (**Fig. 1**) was generated using the ggplot2 v3.3.3 package within RStudio v1.3.959 [43]. GC content was determined using the RStudio seqinr package v4.2-5 [44].

Structural-based protein annotation was performed following the ColabFold-FoldSeek-EggNOG (CoFFE) pipeline [45]. Briefly, amino acid sequences were submitted to the ColabFold Alphafold2-batch notebook through google colab with default settings [46-48]. The highest-ranking predicted protein data base (pdb) structure with predicted local distance difference test on the Cα atoms (pLDDT) > 70 was selected for each analyzed protein. PDB structures were submitted to Foldseek v3.915ef7d for structural comparison against the AlphaFold database of protein structures. For each SLA-04 protein we kept the AlphaFold database hit with the greatest bit score. We enforced a e-value cutoff of 1xe-05. Sequences of AlphaFold database best hits of each SLA-04 protein were annotated using EggNOG mapper webserver with default settings.

2.3 Comparative analyses

2.3.1 Phylogenetic analysis

Phylogenetic analysis was performed on all chromosome level green algal genomes (n=10) (**Sup Fig. 3**) and all Trebouxiophyceae genomes (n=25) (**Fig. 2**) within the NCBI database. 18S rDNA sequences were identified and extracted from chromosome level algal genomes (n=10) using SSU-align v0.1.1 [49, 50]. Ribosomal sequences were trimmed with trimal v1.4.rev22 with the –nogap flag, aligned with MAFFT with the –auto flag, and a maximum likelihood phylogenetic tree of the 18S multiple sequence alignment was generated using IQtree v1.6.1 using default settings [51, 52]. Phylogeny based on orthologous proteins was performed on all Trebouxiophyceae genomes (n=25) using OrthologR v0.4.0 in RStudio [53]. Briefly, OrthologR performed a reciprocal best hit blast for orthology inference between algal species. Orthologous sequences were aligned based on amino acid sequence using the Needleman-Wunsch method and codons were aligned using pal2nal [54]. The Comeron's method was used to predict dN/dS estimations. Orthologous protein sequences were aligned with MAFFT and a maximum likelihood phylogenetic tree was generated using IQtree [51, 52]. Trees were rendered using the RStudio ggtree v2.0.4 package [55].

3. Results

3.1 General genome characteristics

3.1.1 *Chlorella sp.* SLA-04 genome assembly statistics

To determine a telomere-to-telomere genome of *Chlorella sp.* SLA-04, high molecular weight DNA was extracted from SLA-04 cultures for long-read sequencing using a MinION sequencer from Oxford Nanopore (**Sup. Fig 1**). Nanopore sequencing resulted in 5.64 Gb of sequence data in a total of 1.97 M reads (**Sup. Table 1**). Sequencing reads were assembled using a combination of FLYE, Miniasm, and

CANU [23-25], resulting in a 38.8 Mb complete genome with an average of 117x coverage. The SLA-04 genome consists of 13 linear chromosomes, ranging from 1.20 to 5.25 Mbp in length with an average GC content of 65.87%, and circular mitochondrion and chloroplast sequences (**Fig. 1, Sup. Table 1**). This chromosomal architecture is consistent with previously published *Chlorella* genomes, which contain 12-16 nuclear chromosomes [56-58].

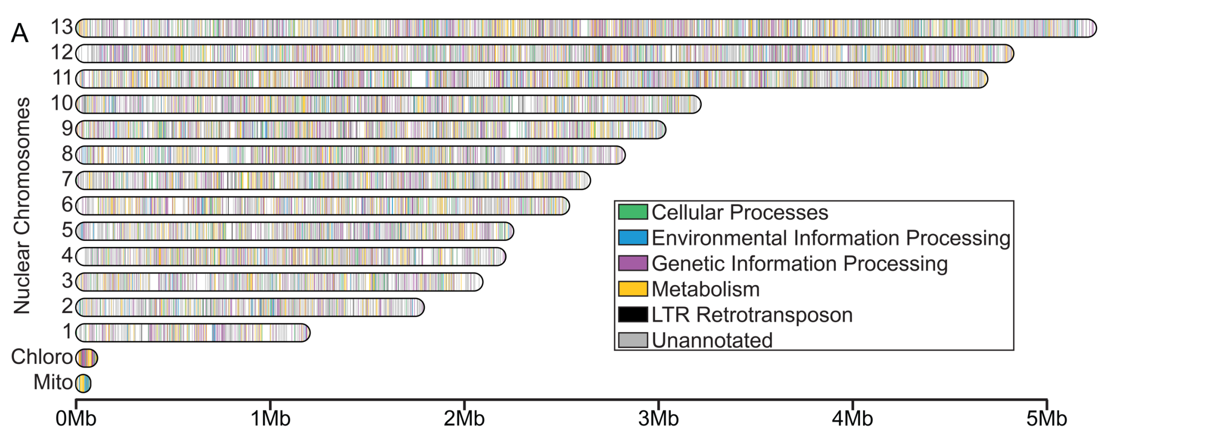

Figure 1. Complete genome of *Chlorella sp.* SLA-04. (A) Chromosome level assembly of *Chlorella sp.* SLA-04 reveals thirteen nuclear chromosomes, and circular chloroplast and mitochondrial genomes. Genes were annotated with KEGG automatic annotation server (KAAS) and color coded according to six different functional categories.

3.1.2 Genome completeness analysis

Genome completeness was evaluated using several distinct but complementary metrics. First, BUSCO was used to calculate the percentage of orthologous Chlorophyta genes within the *Chlorella sp.* SLA-04 genome [42]. The SLA-04 genome contains 93.4% of Chlorophyta orthologs (n=1418) which surpasses the 90% BUSCO cutoff considered to represent "good" genome assemblies (**Sup. Table 1**) [59]. Second, repetitive elements were identified and mapped to unique chromosomal locations, indicating that this assembly provides sufficient coverage to unambiguously assign chromosomal location to repetitive elements that have historically frustrated complete genome assemblies (**Fig. 1, Sup. Fig 2, Sup. Table 2)** [28, 60]. Finally, each nuclear chromosome is composed of a single contiguous well-supported sequence flanked by telomeric repeats $(CCCTAAA)_n$ [61] (**Fig. 1**). Collectively, these metrics suggest that the SLA-04 genome constitutes a complete telomere-to-telomere algal genome.

3.1.3 Annotation of *Chlorella sp.* SLA-04

Chlorella sp. SLA-04 mRNA was extracted, and directly sequenced using a MinION sequencer. Sequences were submitted to MAKER v3.01.02 and aligned to the SLA-04 genome to predict protein coding sequences (CDSs) and splice sites [38, 62]. The SLA-04 genome contains 10,310 CDSs that span 45.73% of the genome with an average gene length of 3,475 bp (**Sup. Table 1**). EggNOG mapper was used to assign gene function resulting in 5,757 (55.8%) CDSs with predicted function (**Fig. 2A**) [40]. Annotated gene functions appear evenly distributed across the nuclear and plastid chromosomes with little functional clustering (**Fig. 1**). Genes coding for proteins that participate in protein synthesis, amino acid metabolism and chaperones are the most abundant gene categories in the SLA-04 genome (**Fig. 2**).

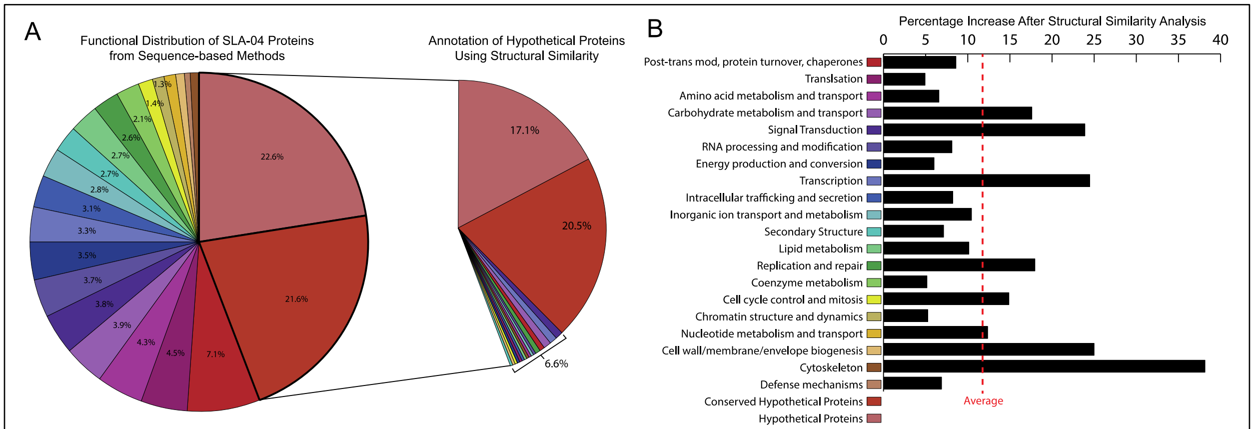

Figure 2. Annotation using protein structural similarity improves functional understanding of SLA-04 proteome. A) 10,310 protein sequences from SLA-04 were submitted to Eggnog mapper for annotation using sequence-based methods resulting in 5,757 sequences (55.8%) with predicted function. Remaining sequences (2,226 conserved hypothetical and 2,327 hypothetical proteins) were submitted to Colabfold structural prediction and resulting structures were compared to the Alphafold and Uniprot structural databases using Foldseek to identify structural homologs of hypothetical SLA-04 proteins. Sequences of structural homologs were submitted to Eggnog Mapper resulting in the annotation of an additional 6.6% of the SLA-04 proteome. B) The percent increase in annotation after structural similarity analysis was calculated by dividing the number of proteins annotated using structural prediction methods by the number of original annotated proteins from sequence-based methods. Proteins were grouped based on their Clusters of Orthologous Genes (COG) functionality.

3.1.4 Annotation using protein 3D structural homology

The 4,553 (44.2%) CDSs that could not be assigned putative functions using sequence-based annotation, consisted of “hypothetical” and “conserved hypothetical” proteins with unknown functions.

Hypothetical proteins are CDSs with no predicted function and no homology to sequences in the EggNOG database, whereas conserved hypothetical proteins are annotated CDSs that have homology to sequences within the EggNOG database, but the function of these homologous sequences remains unknown. To further improve annotation of the SLA-04 genome, we used the ColabFold-FoldSeek-EggNOG (CoFFE) pipeline to predict the 3D structures of hypothetical and conserved hypothetical proteins to query the protein databank for structural homologs [45]. Consistent with functional importance, 3D structures of proteins are more conserved during evolution than primary amino-acid sequences [63]. Thus, we hypothesized that the integration of structure prediction into genome annotation would improve our functional understanding of the SLA-04 genome. In fact, structure-based annotation resulted in high-confidence functional annotations for 677 of the CDSs that were previously assigned to the hypothetical and conserved hypothetical categories using Eggnog mapper [40] (**Fig. 2A**). Structure-based annotation identifies proteins primarily associated with formation of the cytoskeleton and cell wall, transcription, and signal transduction (**Fig. 2B**).

3.1.5 Phylogenetic analysis of *Chlorella sp.* SLA-04

Initial taxonomic classification of *Chlorella sp.* SLA-04 was based on 5.8S and internal transcribed spacer (ITS) regions of the ribosomal small subunit [18]. ITS regions contain highly variable sequences and can be used to distinguish genera; however, phylogenetic analysis performed using a single gene can limit species level classification and lack statistical support for certain phylogenetic nodes [64-66]. Phylogenomic analysis performed using orthologous proteins or 18S rDNA sequences, reveal that SLA-04 shares a closest common ancestor with *Chlorella vulgaris UTEX 395* (**Fig. 3, Sup. Fig. 3**). Further, the *Chlorell*a genus (n=8) forms a paraphyletic clade interspersed with *Micractinium conductrix SAG 241.80* (**Fig. 3**). These data suggest that the current taxonomy of Trebouxiophyceae may not capture the true phylogenetic history of these algae and we anticipate that the addition of more complete algal genome sequences will provide new taxonomic perspectives.

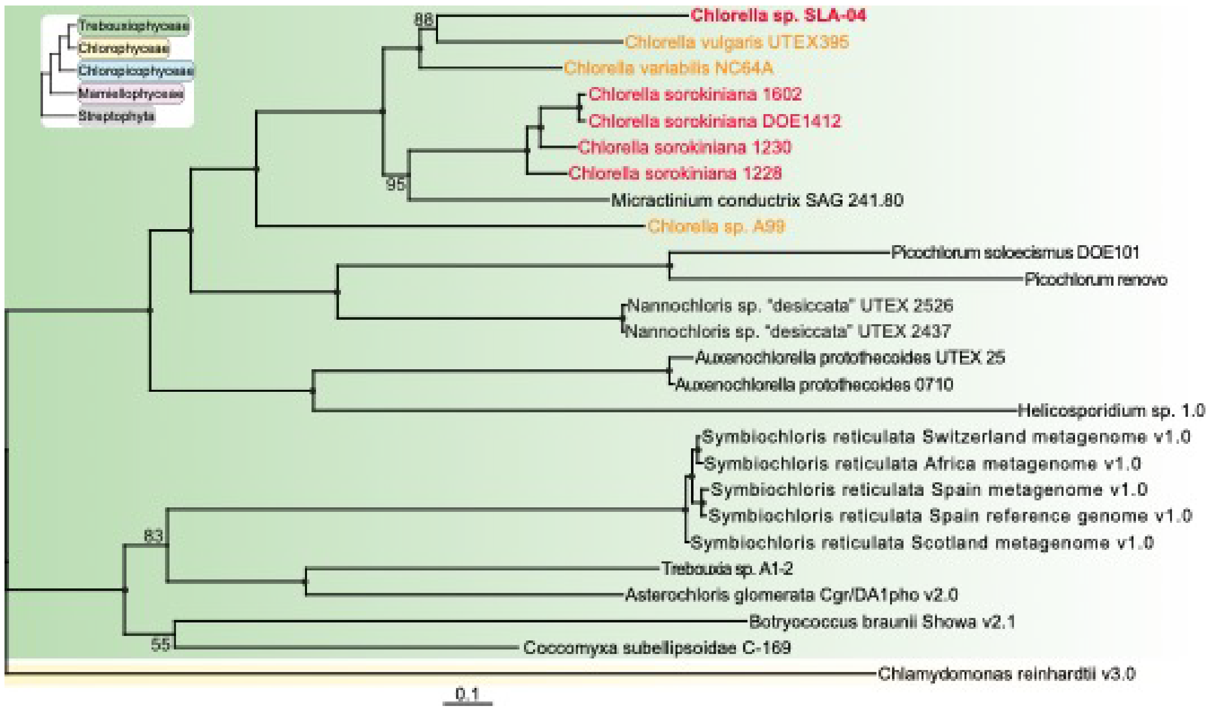

Figure 3. Phylogenomic analysis of green algae performed using orthologous genes indicates that SLA-04 is distinct from other *Chlorella sorokiniana* species. Maximum likelihood phylogenetic analysis based on a concatenation of 215 orthologous proteins sequences of available Trebouxiophyceae genomes from NCBI and PhycoCosm (JGI) databases. *Chlamydomonas reinhardtii* was used as an outgroup. Tree scale indicates the number of amino acid substitutions per site between samples. Bootstrap values of 100 are represented by black dots on nodes. Bootstrap values less than 100 are provided. Classes of algae are denoted via color on tree branches. Red text denotes *Chlorella sorokiniana* species. Orange text represents remaining algae classified within the *Chlorella* genus.

3.2 Genomic analysis of metabolic pathways in *Chlorella sp*. SLA-04

To evaluate the metabolic capability encoded within the *Chlorella sp*. SLA-04 genome, we submitted the complete chromosome level genome to the Kegg Automatic Annotation Server (KAAS) and KofamKOALA for metabolic mapping [39, 67]. These programs annotate and map proteins to established metabolic pathways. According to this analysis, the SLA-04 genome contains all genes necessary for core metabolic pathways including photosynthesis, fatty acid synthesis, carbohydrate synthesis, and the citric acid cycle.

Genes involved in lipid synthesis are critical for biofuel production and stress response to environmental changes [68-70]. The SLA-04 genome encodes ATP citrate lyase EC 2.3.3.8 (ACLY), involved in replenishing cytosolic acetyl-CoA pools from citrate (**Fig. 4**). This enzyme is present in some, but not all green algal genomes (**Sup. Fig. 4**). Cytosolic acetyl-CoA pools are necessary for fatty acid synthesis within the endoplasmic reticulum. The presence and transcription of ATP citrate lyase ACLY within the SLA-04 genome suggests that this organism can utilize TCA cycle intermediates for fatty acid synthesis in addition to using photosynthate directly (**Sup. Table 3**).

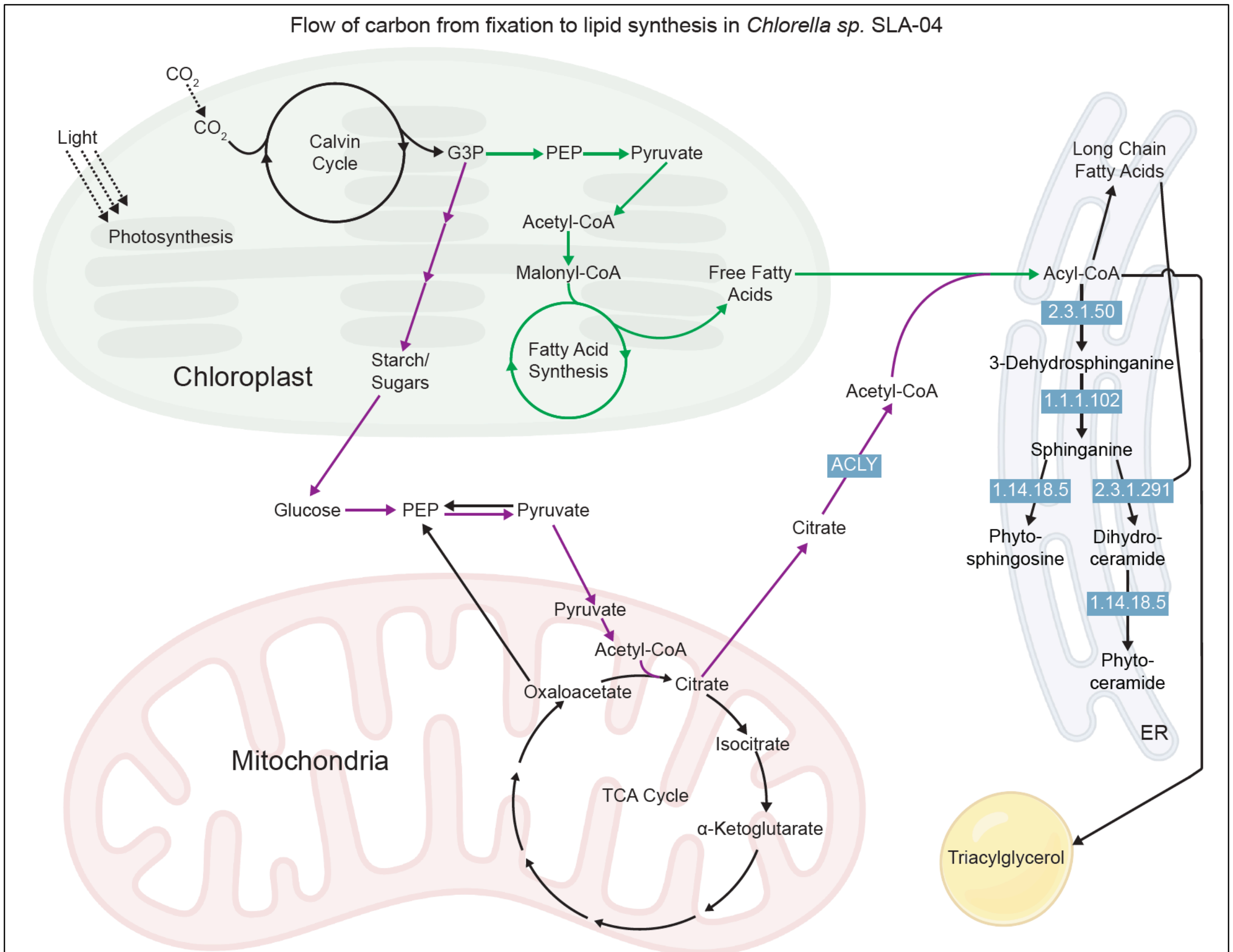

Figure 4. Flow of carbon from fixation to lipid synthesis in *Chlorella sp.* SLA-04. Schematic of carbon flow from fatty acid biosynthesis and TCA cycle to triacylglycerol and sphingolipid biosynthesis in algae. Green arrows represent the flow of carbon from photosynthesis directly to fatty acids within the chloroplast. Purple arrows represent the flow of carbon from photosynthesis to fatty acids through starch synthesis and the TCA cycle. Black arrows represent additional pathways involved in central carbon metabolism. Blue boxes show enzymes of interest. Enzymes are labeled using KEGG and International Union of Biochemistry and Molecular Biology (IUBMB) nomenclature.

Metabolic network analysis of SLA-04 genes involved in lipid synthesis also reveals the presence and transcription of genes coding for the proteins necessary for sphingolipid biosynthesis (**Fig. 4, Sup. Table 3**). Similar to ACLY, enzymes involved in sphingolipid synthesis are only present in a subset of green algal genomes (**Sup. Fig. 4**). Phytoceramide and phytosphingosine are two sphingolipids synthesized within the endoplasmic reticulum (ER) of land plants and algae that remodel the cell membranes to maintain fluidity in response to temperature and osmotic stress [68, 69, 71]. The ability to synthesize phytoceramide and phytosphingosine suggest that SLA-04 maintains the ability to synthesize and remodel cell membrane lipids in response to changes in temperature and osmotic stress.

4. Discussion

Algae thrive in diverse aquatic ecosystems and identifying the genotypes responsible for growth across diverse environments requires a comprehensive understanding of genomic and metabolic variation across habitats. Using ultra-long-read Nanopore sequencing, we determined a complete chromosome level assembly of the *Chlorella sp.* SLA-04 genome (**Fig. 1**). Annotation of the SLA-04 genome using sequence-based methods resulted in an incomplete representation of SLA-04 metabolism with just over half of the proteins having assigned functions. While traditional sequence-based annotation methods identified core metabolic pathways within SLA-04, a large portion of the proteome remained hypothetical with unknown function (**Fig. 2A**).

The structure of a protein dictates its function [72]. By predicting the structure of hypothetical proteins and comparing them to a structural database we were able to predict the function of proteins that were unrecognizable at the sequence level. While structural annotation identified additional proteins within all functional categories, some categories such as cytoskeleton and cell wall proteins were overrepresented in this analysis (**Fig. 2**). The overrepresentation of cytoskeleton and cell wall proteins could be a consequence of functional diversification or an underrepresentation of cytoskeleton and cell wall protein sequences within databases used for sequence-based annotation (NCBI or EggNOG). Regardless, we found that using structure-based annotation methods aid in generating a more complete representation of algal metabolism. Interestingly, 37.6% of SLA-04 proteins continue to escape functional predictions following structural annotation, and 20.5% of these sequences are conservation within other algae and plants. Evolutionary conservation and retention of proteins across algae and plants suggests these proteins are functionally important, and this analysis helps focus future efforts on conserved aspect of the proteome that remain functionally uncharacterized.

While annotating the SLA-04 proteome, we identified lipid synthesis pathways that may allow SLA-04 to respond to changing environmental conditions. ATP citrate lyase (ACLY) generates cytosolic acetyl-CoA from citrate and is a critical enzyme for fatty acid (FA) biosynthesis and modification [73-75]. Algal fatty acids are initially synthesized in the chloroplast and exported into the cytosol where they are modified into triacylglycerols (TAG) or sphingolipids within the ER [69, 76-80] (**Fig. 4**). Modification and synthesis of additional fatty acids within the cytosol and ER requires cytosolic acetyl-CoA. The presence and transcription of genes encoding ACLY suggests that SLA-04 can utilize starch-derived TCA intermediates as a precursor for lipid synthesis. The ability to interconvert between storage molecules enables the metabolic flexibility to synthesize desired biomolecules regardless of the molecular resources available within a cell. Since starches and lipids are important precursors for bioproducts like biodiesel and ethanol, this work demonstrates how genome sequences may improve our ability to link bioproduct production to specific organisms and environmental conditions.

Acetyl-CoA has many cytosolic functions, but we hypothesize that SLA-04 may generate additional cytosolic acetyl-CoA in part to synthesize long term storage lipids and sphingolipids. Sphingolipids are membrane lipids synthesized within the ER that contain long chain FAs. Sphingolipids modify the algal cell membrane to maintain membrane fluidity in response to temperature and osmotic stress [68-70]. The synthesis of sphingolipids such as phytoceramide and phytosphingosine requires cytosolic acetyl-CoA to modify and elongate free FAs that originated from the chloroplast [81] (**Fig. 4**). Similar to *ACLY*, genes encoding enzymes responsible for phytoceramide and phytosphingosine biosynthesis are present and being transcribed in the SLA-04 genome (**Fig. 4, Sup. Table 3**). Together, the capacity for sphingolipid biosynthesis and cytosolic acetyl-CoA generation by ACLY suggests that SLA-04 may have the metabolic capacity to modulate cell membrane fluidity to rapidly adapt to diurnal and seasonal temperature and osmotic changes (*e.*g., alkalinity and/or salinity) within the environment [82, 83].

5. Conclusion

Chlorella sp. SLA-04 grows well in a wide range of media conditions such as (i) high (>100 mM) [84] and low (<5 mM) [18] alkalinities, (ii) circumneutral pH [85] and high (>10.2) pH [84], and (iii) in media containing 0 to 35,000 ppm salt [86]. The genome analysis presented here, together with prior experimental results indicate that SLA-04 has robust and adaptable growth characteristics, with potential for use in diverse engineered growth systems for biofuel and bioproduct applications. More broadly, this comprehensive view of metabolism within SLA-04 has demonstrated how complete algal genomes can be leveraged to discover new genotypes and identify algal strains most suitable for production of desired bioproducts.

Keywords:

Chlorella sp., Ultra-long-read sequencing, Complete genome, Protein 3D structural annotation, ATP citrate lyase, Sphingolipids

Abbreviations:

NCBI, National Center for Biotechnology Information; KEGG, Kyoto Encyclopedia for Genes and Genomes; KAAS, KEGG Automatic Annotation Server; ITS, Internal transcribed spacer; TCA, Citric Acid Cycle; ACLY, ATP citrate lyase; ER, Endoplasmic reticulum; TAG, Triacylglycerol; FA, Fatty acids

Data Access:

The raw and processed genomic and transcriptomic data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under BioProject accession number PRJNA766673 and BioSample accession number SAMN21857598.

Competing interest statement:

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

B.W. is the founder of SurGene and VIRIS Detection Systems and is an inventor on patent applications related to CRISPR-Cas systems and applications thereof.

Acknowledgements:

This work was supported by the Bioenergy Technologies Office (BETO) of the U.S. Department of Energy (DOE) for the Productivity Enhanced Algae and ToolKits (PEAK) program [grant number #DE-EE0008247].

Author contributions:

Conceptualization, B.W., C.G., R.C., R.G., S.V.; Methodology, B.W., C.G., R.W.; Investigation & Data Collection, C.G, H.G., R.W.; Genomics & Bioinformatic Analysis, C.G., H.B, H.G., W.H.; Writing – Original Draft, B.W., C.G; Writing – Review & Editing, B.W., C.G., H.B., H.G., R.C., R.G, R.W., S.V., W.H.

References:

[1] M.M.M. Kuypers, H.K. Marchant, B. Kartal, The microbial nitrogen-cycling network, Nature Reviews Microbiology, 16 (2018) 263-276.

[2] J.C.G. Walker, The Oxygen Cycle, Springer Berlin Heidelberg1980, pp. 87-104.

[3] M.J. Behrenfeld, Biospheric Primary Production During an ENSO Transition, Science, 291 (2001) 2594-2597.

[4] C.B. Field, Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components, Science, 281 (1998) 237-240.

[5] J.H. Evans, The Survival of Freshwater Algae During Dry Periods: Part I. An Investigation of the Algae of Five Small Ponds, The Journal of Ecology, 46 (1958) 149.

[6] P. Varshney, P. Mikulic, A. Vonshak, J. Beardall, P.P. Wangikar, Extremophilic micro-algae and their potential contribution in biotechnology, Bioresour Technol, 184 (2015) 363-372.

[7] R. Blanc-Mathieu, B. Verhelst, E. Derelle, S. Rombauts, F.-Y. Bouget, I. Carré, A. Château, A. Eyre-Walker, N. Grimsley, H. Moreau, B. Piégu, E. Rivals, W. Schackwitz, Y. Van De Peer, G. Piganeau, An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies, BMC Genomics, 15 (2014) 1103.

[8] S.S. Merchant, S.E. Prochnik, O. Vallon, E.H. Harris, S.J. Karpowicz, G.B. Witman, A. Terry, A. Salamov, L.K. Fritz-Laylin, L. Marechal-Drouard, W.F. Marshall, L.H. Qu, D.R. Nelson, A.A. Sanderfoot, M.H. Spalding, V.V. Kapitonov, Q. Ren, P. Ferris, E. Lindquist, H. Shapiro, S.M. Lucas, J. Grimwood, J. Schmutz, P. Cardol, H. Cerutti, G. Chanfreau, C.L. Chen, V. Cognat, M.T. Croft, R. Dent, S. Dutcher, E. Fernandez, H. Fukuzawa, D. Gonzalez-Ballester, D. Gonzalez-Halphen, A. Hallmann, M. Hanikenne, M. Hippler, W. Inwood, K. Jabbari, M. Kalanon, R. Kuras, P.A. Lefebvre, S.D. Lemaire, A.V. Lobanov, M. Lohr, A. Manuell, I. Meier, L. Mets, M. Mittag, T. Mittelmeier, J.V. Moroney, J. Moseley, C. Napoli, A.M. Nedelcu, K. Niyogi, S.V. Novoselov, I.T. Paulsen, G. Pazour, S. Purton, J.P. Ral, D.M. Riano-Pachon, W. Riekhof, L. Rymarquis, M. Schroda, D. Stern, J. Umen, R. Willows, N. Wilson, S.L. Zimmer, J. Allmer, J. Balk, K. Bisova, C.J. Chen, M. Elias, K. Gendler, C. Hauser, M.R. Lamb, H. Ledford, J.C. Long, J. Minagawa, M.D. Page, J. Pan, W. Pootakham, S. Roje, A. Rose, E. Stahlberg, A.M. Terauchi, P. Yang, S. Ball, C. Bowler, C.L. Dieckmann, V.N. Gladyshev, P. Green, R. Jorgensen, S. Mayfield, B. Mueller-Roeber, S. Rajamani, R.T. Sayre, P. Brokstein, I. Dubchak, D. Goodstein, L. Hornick, Y.W. Huang, J. Jhaveri, Y. Luo, D. Martinez, W.C.A. Ngau, B. Otillar, A. Poliakov, A. Porter, L. Szajkowski, G. Werner, K. Zhou, I.V. Grigoriev, D.S. Rokhsar, A.R. Grossman, The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions, Science, 318 (2007) 245-250.

[9] S. Koren, M.C. Schatz, B.P. Walenz, J. Martin, J.T. Howard, G. Ganapathy, Z. Wang, D.A. Rasko, W.R. McCombie, E.D. Jarvis, A.M. Phillippy, Hybrid error correction and de novo assembly of single-molecule sequencing reads, Nature Biotechnology, 30 (2012) 693-700.

[10] C. Kingsford, M.C. Schatz, M. Pop, Assembly complexity of prokaryotic genomes using short reads, BMC Bioinformatics, 11 (2010) 21.

[11] S.L. Amarasinghe, S. Su, X. Dong, L. Zappia, M.E. Ritchie, Q. Gouil, Opportunities and challenges in long-read sequencing data analysis, Genome Biology, 21 (2020).

[12] S. Mehrotra, V. Goyal, Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function, Genomics Proteomics Bioinformatics, 12 (2014) 164-171.

[13] A. Payne, N. Holmes, V. Rakyan, M. Loose, BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files, Bioinformatics, 35 (2018) 2193-2198.
[14] D. Deamer, M. Akeson, D. Branton, Three decades of nanopore sequencing, Nature Biotechnology, 34 (2016) 518-524.
[15] Y. Wang, Y. Zhao, A. Bollas, Y. Wang, K.F. Au, Nanopore sequencing technology, bioinformatics and applications, Nature Biotechnology, 39 (2021) 1348-1365.
[16] J.C. Dohm, P. Peters, N. Stralis-Pavese, H. Himmelbauer, Benchmarking of long-read correction methods, NAR Genomics and Bioinformatics, 2 (2020).
[17] J.M. Sutton, J.D. Millwood, A. Case Mccormack, J.L. Fierst, Optimizing experimental design for genome sequencing and assembly with Oxford Nanopore Technologies, Gigabyte, 2021 (2021) 1-26.
[18] A. Vadlamani, S. Viamajala, B. Pendyala, S. Varanasi, Cultivation of Microalgae at Extreme Alkaline pH Conditions: A Novel Approach for Biofuel Production, ACS Sustainable Chemistry & Engineering, 5 (2017) 7284-7294.
[19] H.C. Bold, The Morphology of Chlamydomonas chlamydogama, Sp. Nov, Bulletin of the Torrey Botanical Club, 76 (1949) 101.
[20] M. Mustapa, N. Sallehudin, M. Mohamed, N. Mohammad Noor, R. Raus, Decontamination of Chlorella sp. Culture using antibiotics and antifungal cocktail treatment, ARPN Journal of Engineering and Applied Sciences, 11 (2016).
[21] H.M. Goemann, J.D. Gay, R.C. Mueller, E.N.J. Brookshire, P. Miller, B. Poulter, B.M. Peyton, Aboveground and belowground responses to cyanobacterial biofertilizer supplement in a semi-arid, perennial bioenergy cropping system, Global Change Biology. Bioenergy, 13 (2021) 1908-1923.
[22] L.S. Weyrich, A.G. Farrer, R. Eisenhofer, L.A. Arriola, J. Young, C.A. Selway, M. Handsley-Davis, C.J. Adler, J. Breen, A. Cooper, Laboratory contamination over time during low-biomass sample analysis, Molecular Ecology Resources, 19 (2019) 982-996.
[23] H. Li, Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences, Bioinformatics, 32 (2016) 2103-2110.
[24] S. Koren, B.P. Walenz, K. Berlin, J.R. Miller, N.H. Bergman, A.M. Phillippy, Canu: scalable and accurate long-read assembly via adaptivek-mer weighting and repeat separation, Genome Research, 27 (2017) 722-736.
[25] M. Kolmogorov, J. Yuan, Y. Lin, P.A. Pevzner, Assembly of long, error-prone reads using repeat graphs, Nature Biotechnology, 37 (2019) 540-546.
[26] H. Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics, 34 (2018) 3094-3100.
[27] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, A. Whitwham, T. Keane, S.A. McCarthy, R.M. Davies, H. Li, Twelve years of SAMtools and BCFtools, GigaScience, 10 (2021).
[28] A.F. Smit, R. Hubley, P. Green, RepeatMasker, 1996.
[29] W. Bao, K.K. Kojima, O. Kohany, Repbase Update, a database of repetitive elements in eukaryotic genomes, Mobile DNA, 6 (2015).
[30] J. Storer, R. Hubley, J. Rosen, T.J. Wheeler, A.F. Smit, The Dfam community resource of transposable element families, sequence models, and genome annotations, Mobile DNA, 12 (2021).

[31] S. El-Gebali, J. Mistry, A. Bateman, S.R. Eddy, A. Luciani, S.C. Potter, M. Qureshi, L.J. Richardson, G.A. Salazar, A. Smart, L. Erik, L. Hirsh, L. Paladin, D. Piovesan, C. Silvio, R.D. Finn, The Pfam protein families database in 2019, Nucleic Acids Research, 47 (2019) D427-D432.
[32] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite, Trends Genet, 16 (2000) 276-277.
[33] D. Ellinghaus, S. Kurtz, U. Willhoeft, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons, BMC Bioinformatics, 9 (2008) 18.
[34] J. Rozewicki, S. Li, K.M. Amada, D.M. Standley, K. Katoh, MAFFT-DASH: integrated protein sequence and structural alignment, Nucleic Acids Res, 47 (2019) W5-w10.
[35] G. Pertea, M. Pertea, GFF Utilities: GffRead and GffCompare, F1000Research, 9 (2020) 304.
[36] M. Pertea, G.M. Pertea, C.M. Antonescu, T.-C. Chang, J.T. Mendell, S.L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, Nature Biotechnology, 33 (2015) 290-295.
[37] B.L. Cantarel, I. Korf, S.M.C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sanchez Alvarado, M. Yandell, MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes, Genome Research, 18 (2007) 188-196.
[38] M.S. Campbell, C. Holt, B. Moore, M. Yandell, Genome Annotation and Curation Using MAKER and MAKER-P, Current Protocols in Bioinformatics, 48 (2014).
[39] Y. Moriya, M. Itoh, S. Okuda, A.C. Yoshizawa, M. Kanehisa, KAAS: an automatic genome annotation and pathway reconstruction server, Nucleic Acids Res, 35 (2007) W182-185.
[40] C.P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale, Molecular Biology and Evolution, 38 (2021) 5825-5829.
[41] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics, 31 (2015) 3210-3212.
[42] M.e. Manni, Matthew, M. Seppey, Felipe, Evgeny, BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes, arXiv pre-print server, (2021).
[43] W. Hadley, Ggplot2: Elegrant graphics for data analysis, Springer2016.
[44] D. Charif, J.R. Lobry, SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis, Springer Berlin Heidelberg2007, pp. 207-232.
[45] F. Ruperti, N. Papadopoulos, J. Musser, D. Arendt, Beyond sequence similarity: cross-phyla protein annotation by structural prediction and alignment, Cold Spring Harbor Laboratory, 2022.
[46] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature, 596 (2021) 583-589.
[47] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A.

Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer, Cold Spring Harbor Laboratory, 2021.
[48] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all, Nature Methods, 19 (2022) 679-682.
[49] S.R. Eddy, E.P. Nawrocki, Structural rna homology search and alignment using covariance models, 2009.
[50] J.J. Cannone, S. Subramanian, M.N. Schnare, J.R. Collett, L.M. D'Souza, Y. Du, B. Feng, N. Lin, L.V. Madabusi, K.M. Müller, N. Pande, Z. Shang, N. Yu, R.R. Gutell, The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs, BMC Bioinformatics, 3 (2002) 2.
[51] K. Katoh, D.M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability, Molecular Biology and Evolution, 30 (2013) 772-780.
[52] L.-T. Nguyen, H.A. Schmidt, A. Von Haeseler, B.Q. Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies, Molecular Biology and Evolution, 32 (2015) 268-274.
[53] H.-G. Drost, A. Gabel, I. Grosse, M. Quint, Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis, Molecular Biology and Evolution, 32 (2015) 1221-1231.
[54] M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments, Nucleic Acids Research, 34 (2006) W609-W612.
[55] G. Yu, D.K. Smith, H. Zhu, Y. Guan, T.T.Y. Lam, ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data, Methods in Ecology and Evolution, 8 (2017) 28-36.
[56] B.T. Hovde, E.R. Hanschen, C.R. Steadman Tyler, C.-C. Lo, Y. Kunde, K. Davenport, H. Daligault, J. Msanne, S. Canny, S.-I. Eyun, J.-J.M. Riethoven, J. Polle, S.R. Starkenburg, Genomic characterization reveals significant divergence within Chlorella sorokiniana (Chlorellales, Trebouxiophyceae), Algal Research, 35 (2018) 449-461.
[57] T. Higashiyama, S. Maki, T. Yamada, Molecular organization of Chlorella vulgaris chromosome I: presence of telomeric repeats that are conserved in higher plants, Molecular and General Genetics MGG, 246 (1995) 29-36.
[58] M.B. Arriola, N. Velmurugan, Y. Zhang, M.H. Plunkett, H. Hondzo, B.M. Barney, Genome sequences of Chlorella sorokiniana UTEX 1602 and Micractinium conductrix SAG 241.80: implications to maltose excretion by a green alga, The Plant Journal, 93 (2018) 566-586.
[59] E.R. Hanschen, B.T. Hovde, S.R. Starkenburg, An evaluation of methodology to determine algal genome completeness, Algal Research, 51 (2020) 102019.
[60] M. Tarailo-Graovac, N. Chen, Using RepeatMasker to identify repetitive elements in genomic sequences, Curr Protoc Bioinformatics, Chapter 4 (2009) Unit 4.10.
[61] J. Fulnečková, T. Hasíková, J. Fajkus, A. Lukešová, M. Eliáš, E. Sýkorová, Dynamic Evolution of Telomeric Sequences in the Green Algal Order Chlamydomonadales, Genome Biology and Evolution, 4 (2012) 248-264.
[62] C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects, BMC Bioinformatics, 12 (2011) 491.

[63] K. Illergård, D.H. Ardell, A. Elofsson, Structure is three to ten times more conserved than sequence--a study of structural response in protein cores, Proteins, 77 (2009) 499-508.
[64] I. Álvarez, J.F. Wendel, Ribosomal ITS sequences and plant phylogenetic inference, Molecular Phylogenetics and Evolution, 29 (2003) 417-434.
[65] K.M. Evans, A.H. Wortley, D.G. Mann, An Assessment of Potential Diatom "Barcode" Genes (cox1, rbcL, 18S and ITS rDNA) and their Effectiveness in Determining Relationships in Sellaphora (Bacillariophyta), Protist, 158 (2007) 349-364.
[66] F. Delsuc, H. Brinkmann, H. Philippe, Phylogenomics and the reconstruction of the tree of life, Nature Reviews Genetics, 6 (2005) 361-375.
[67] T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, H. Ogata, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold, Bioinformatics, 36 (2020) 2251-2252.
[68] Y. Li, Y. Lou, T. Mu, A. Ke, Z. Ran, J. Xu, J. Chen, C. Zhou, X. Yan, Q. Xu, Y. Tan, Sphingolipids in marine microalgae: Development and application of a mass spectrometric method for global structural characterization of ceramides and glycosphingolipids in three major phyla, Analytica Chimica Acta, 986 (2017) 82-94.
[69] A. De Bigault Du Granrut, J.-L. Cacas, How Very-Long-Chain Fatty Acids Could Signal Stressful Conditions in Plants?, Frontiers in plant science, 7 (2016) 1490-1490.
[70] H.C. Resemann, C. Herrfurth, K. Feussner, E. Hornung, A.K. Ostendorf, J. Gömann, J. Mittag, N. Van Gessel, J.D. Vries, J. Ludwig-Müller, J. Markham, R. Reski, I. Feussner, Convergence of sphingolipid desaturation across over 500 million years of plant evolution, Nature Plants, 7 (2021) 219-232.
[71] J.N. Kong, K. Hardin, M. Dinkins, G. Wang, Q. He, T. Mujadzic, G. Zhu, J. Bielawski, S. Spassieva, E. Bieberich, Regulation of Chlamydomonas flagella and ependymal cell motile cilia by ceramide-mediated translocation of GSK3, Molecular Biology of the Cell, 26 (2015) 4451-4465.
[72] J.C. Kendrew, G. Bodo, H.M. Dintzis, R.G. Parrish, H. Wyckoff, D.C. Phillips, A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis, Nature, 181 (1958) 662-666.
[73] D.E. Bauer, G. Hatzivassiliou, F. Zhao, C. Andreadis, C.B. Thompson, ATP citrate lyase is an important component of cell growth and transformation, Oncogene, 24 (2005) 6314-6322.
[74] M.J. Hynes, S.L. Murray, ATP-Citrate Lyase Is Required for Production of Cytosolic Acetyl Coenzyme A and Development in Aspergillus nidulans, Eukaryotic Cell, 9 (2010) 1039-1048.
[75] M.-H. Liang, J.-G. Jiang, Characterization and nitrogen deficiency response of ATP-citrate lyase from unicellular alga Dunaliella tertiolecta, Algal Research, 20 (2016) 77-86.
[76] J. Fan, K. Ning, X. Zeng, Y. Luo, D. Wang, J. Hu, J. Li, H. Xu, J. Huang, M. Wan, W. Wang, D. Zhang, G. Shen, C. Run, J. Liao, L. Fang, S. Huang, X. Jing, X. Su, A. Wang, L. Bai, Z.M. Hu, J. Xu, Y. Li, Genomic Foundation of Starch to Lipid Switch in Oleaginous Chlorella, Plant Physiology, (2015) pp.01174.02015.
[77] E.C. Goncalves, A.C. Wilkie, M. Kirst, B. Rathinasabapathi, Metabolic regulation of triacylglycerol accumulation in the green algae: identification of potential targets for engineering to improve oil yield, Plant Biotechnology Journal, 14 (2016) 1649-1660.
[78] O. Avidan, U. Pick, Acetyl-CoA synthetase is activated as part of the PDH-bypass in the oleaginous green algaChlorella desiccata, Journal of Experimental Botany, 66 (2015) 7287-7298.

[79] S. Bellou, G. Aggelis, Biochemical activities in Chlorella sp. and Nannochloropsis salina during lipid and sugar synthesis in a lab-scale open pond simulating reactor, Journal of Biotechnology, 164 (2013) 318-329.
[80] K.W.M. Tan, Y.K. Lee, The dilemma for lipid productivity in green microalgae: importance of substrate provision in improving oil yield without sacrificing growth, Biotechnology for Biofuels, 9 (2016).
[81] F. Aid, Plant Lipid Metabolism, IntechOpen2020.
[82] B.B. Cael, A.J. Heathcote, D.A. Seekell, The volume and mean depth of Earth's lakes, Geophysical Research Letters, 44 (2017) 209-218.
[83] B.W. Eakins, G.F. Sharman, Volumes of the World's Oceans from ETOPO1, NOAA National Geophysical Data Center, Boulder, CO, 2010.
[84] A. Vadlamani, B. Pendyala, S. Viamajala, S. Varanasi, High Productivity Cultivation of Microalgae without Concentrated CO2 Input, ACS Sustainable Chemistry & Engineering, 7 (2019) 1933-1943.
[85] M. Hanifzadeh, E.C. Garcia, S. Viamajala, Production of lipid and carbohydrate from microalgae without compromising biomass productivities: Role of Ca and Mg, Renewable Energy, 127 (2018) 989-997.
[86] M.-M. Hanifzadeh, Approaches for Sustainable Production of Microalgae with High Productivity and Flexible Composition, Chemical Engineering, The University of Toledo, 2018.