Algal Research 35 (2018) 449–461

Contents lists available at ScienceDirect

Algal Research

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

Genomic characterization reveals significant divergence within *Chlorella sorokiniana* (Chlorellales, Trebouxiophyceae)

Blake T. Hovde[a,1], Erik R. Hanschen[a,1], Christina R. Steadman Tyler[a], Chien-Chi Lo[a], Yuliya Kunde[a], Karen Davenport[a], Hajnalka Daligault[a], Joseph Msanne[b,c], Stephanie Canny[d], Seong-il Eyun[g], Jean-Jack M. Riethoven[d,e], Juergen Polle[f], Shawn R. Starkenburg[a,*]

[a] *Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87544, United States of America*
[b] *Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE 68588, United States of America*
[c] *School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE 68588, United States of America*
[d] *Center for Biotechnology, University of Nebraska-Lincoln, Lincoln, NE 68588, United States of America*
[e] *School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, NE 68588, United States of America*
[f] *Brooklyn College, City University of New York, New York, NY, 11210, United States of America*
[g] *Department of Life Science, Chung-Ang University, Seoul 06974, Korea*

ARTICLE INFO

Keywords:
Chlorella sorokiniana
Genome analysis
Meiosis
DNA methylation
Polyketide synthetase (PKS)

ABSTRACT

Selection of highly productive algal strains is crucial for establishing economically viable biomass and bioproduct cultivation systems. Characterization of algal genomes, including understanding strain-specific differences in genome content and architecture is a critical step in this process. Using genomic analyses, we demonstrate significant differences between three strains of *Chlorella sorokiniana* (strain 1228, UTEX 1230, and DOE1412). We found that unique, strain-specific genes comprise a substantial proportion of each genome, and genomic regions with > 80% local nucleotide identity constitute < 15% of each genome among the strains, indicating substantial strain specific evolution. Furthermore, cataloging of meiosis and other sex-related genes in *C. sorokiniana* strains suggests strategic breeding could be utilized to improve biomass and bioproduct yields if a sexual cycle can be characterized. Finally, preliminary investigation of epigenetic machinery suggests the presence of potentially unique transcriptional regulation in each strain. Our data demonstrate that these three *C. sorokiniana* strains represent significantly different genomic content. Based on these findings, we propose individualized assessment of each strain for potential performance in cultivation systems.

1. Introduction

Development and deployment of a productive, stable, and economically viable algal cultivation system requires detailed genetic and phenotypic knowledge of the platform strain(s) [1,2]. This knowledge is gained initially through sequencing and characterization of the genomic content, enabling the formation of testable hypotheses to accelerate algal strain improvement. Identification of both conserved and strain specific pathways will facilitate strain improvement through targeted genetic modification or selective breeding. However, high quality genome assemblies from microalgae production strain candidates are not widely available and many algal genomes are sequenced with short read sequence data, resulting in highly fragmented assemblies, thus impeding accurate gene annotation, transcriptomic analyses, and *in silico* metabolic modeling.

Nearly finished genomes of microalgae production strains inform many other biological functions relevant to bioproduct and biofuel production. Five functions of particular interest include (1) conservation and divergence in energy capture, (2) metabolism and carbon storage, (3) the capacity for sexual reproduction (thereby facilitating artificial selection of desirable traits), (4) the capacity for epigenetic modifications, enabling more nuanced regulation of metabolic and energy storage pathways, and (5) the production of antibiotic compounds, which may assist in crop defense or design of synthetic antibiotics.

First, given the importance of energy capture through photosynthesis in algal growth and production, the genes underlying photosynthesis are predicted to be highly conserved. Protein functionality in core

* Corresponding author at: PO Box 1663, Los Alamos, NM 87544, United States of America.
E-mail address: shawns@lanl.gov (S.R. Starkenburg).
[1] Equal contributors.

https://doi.org/10.1016/j.algal.2018.09.012
Received 3 May 2018; Received in revised form 10 August 2018; Accepted 16 September 2018
Available online 01 October 2018

photosynthetic processes includes carbohydrate metabolism; capturing electron excitation energy; and the anabolism of pigments, lipids, and amino acids. Previous phylogenomic analyses have characterized the set of genes restricted to photosynthetic organisms [3,4]. Loss of these genes may indicate a relative specialist photosynthetic strategy [4] as a result of adaptation to environmental niches. Such specialization may have subsequent consequences on production capabilities.

Second, once energy has been conserved, it may be utilized during metabolism or stored in different ways, which may have profound effects on product value. Understanding, and subsequently targeting, specific biochemical pathways *via* genome editing techniques can also lead to accumulation of additional lipids or other high value products [5].

Third, the ability to reproduce sexually has numerous consequences on both long-term evolution and laboratory selection. Recombination during sexual reproduction results in the purging of deleterious alleles and creation of novel gene combinations [6,7]. Accordingly, sexual reproduction has been used to accelerate artificial or experimental selection throughout eukaryotes [8–15]. The capacity for sexual reproduction in candidate production strains may be exploited through strategic breeding to improve biomass and bioproduct yields as used in traditional food cultivation.

Fourth, while genomic information helps inform functionality of an organism, understanding the factors that regulate the accessibility of the genome is necessary to analyze the phenotypic productivity of a given algae species. These factors collectively constitute chromatin remodeling mechanisms that, when inherited after mitotic activity, are deemed epigenetic in nature. Epigenetic machinery is responsible for posttranslational modification of amino acids in histone proteins or nucleic acids in RNA and DNA. DNA methylation, particularly on cytosine residues, is important for genome protection from opportunistic genetic elements, gene expression, and genomic stability. Two seminal reports of silencing mechanisms employed by microalgae species suggest the existence of DNA modification machinery, including RNA-mediated DNA methylation, and DNA modifications that are present in plants, but not in mammals [16,17]. However, for many algal species, these modifications are uncharacterized and only a handful of genetic signatures of epigenetic machinery have been identified in select species [16,18–22].

Finally, algae produce an array of defensive compounds, including some products of polyketide synthase (PKS) enzymes. PKS enzymes are large, multi-domain genes that encode a variety of naturally-occurring biotoxins and defensive compounds. Polyketide secondary metabolites include antimicrobial, antifungal, insecticide, and immunosuppressive chemicals [23]. Thus, we are interested in understanding polyketide diversity, evolution, and synthesis to determine the value of PKS products as a potential algal bioproduct. Synthesized from acetyl- or malonyl-CoA, polyketides are produced by polyketide synthase genes that possess a constrained set of canonical functional domains, including ketoacyl synthase (KS), acyl transferase (AT), ketoacyl reductase (KR), dehydratase (DH), enoyl reductase (ER), acyl carrier protein (ACP; also known as a phosphopantetheine attachment site), and a thioreductase (TE). PKS genes are categorized into three major structural groups. Type I PKSs are large, modular proteins which are found throughout bacteria, fungi, and algae [24,25]. Each module elongates and modifies the polyketide. Type II PKSs are smaller, aggregative proteins present in green algae and bacteria that iteratively act on polyketide chains [24,26]. Type III PKSs are restricted to streptophyte green algae and bacteria [27,28] and operate as homodimers. Phylogenomic investigations of polyketide synthases in green algae have found three PKS genes in green algae *Ostreococcus lucimarinus* and *Ostreococcus tauri* [25,29]. This relatively high abundance of PKS genes (1.5% of the genome length) suggests important, but unknown function in green algae [25,30].

Chlorella sorokiniana, a freshwater chlorophyte, is being evaluated for utilization as a feedstock for biofuels and bioproducts given its high degree of productivity during short periods of cultivation [1,31,32]. Although a few phenotypic comparisons between multiple *C. sorokiniana* strains have been performed [33–36], the genetic basis for the varied phenotypes remains unknown. Here, we present the genome sequences and gene annotations of three strains of *C. sorokiniana* and results of a comparative analysis of gene content between these strains (DOE1412, 1228, and UTEX 1230). The use of long read technologies and optical mapping generated high-quality, chromosome-level, genome assemblies of *C. sorokiniana*. We report a significant disparity of gene content, with each strain containing a large complement of unique genes and high genomic divergence. Defining the genomic variation among *C. sorokiniana* strains is a necessary step for realizing the potential of *C. sorokiniana* as a commodity feedstock and will inform differences in growth patterns and growth conditions between strains. While genomic differences may underlie differences in growth patterns, the basis for sexual reproduction, PKS defense, and photosynthesis are conserved among *C. sorokiniana* strains. These results highlight the potential to develop and improve *C. sorokiniana* for use in industrial applications through epigenetic modification, sexual reproduction, and bioengineering of different genomic elements.

2. Methods

2.1. Strain information

C. sorokiniana UTEX 1230 (hereafter 1230) is one of the most productive strains identified and is being evaluated for utilization as a biofuel feedstock [31]. *C. sorokiniana* has an optimal growth temperature of 37 °C and is able to grow heterotrophically on a variety of sugars that enhance oil accumulation [37]. Growth in an optimized mixotrophic and heterotrophic bioreactor supplemented with glucose enabled *C. sorokiniana* 1230 to accumulate 30–40% of its cell mass as lipids [38]. Furthermore, while growing in the absence of nitrogen (following pre-growth with ammonia at dry weight production rates equivalent to growth in the presence of ammonia), the energy content of the algae increased by nearly 50% on a dry weight basis (Dr. Sangeeta Negi, unpublished results).

C. sorokiniana strain 1228 (hereafter 1228) was first studied by Phycal, Inc., and was described to be a clonal isolate from a UTEX 1230 cultivation sample. *C. sorokiniana* 1228 has genomic content distinct from the other *C. sorokiniana* strains [37] and serves as a possible reservoir of genomic material for genetic engineering applications.

C. sorokiniana DOE1412 (equivalent to UTEX B 3016; hereafter 1412) was isolated through prospecting efforts by Dr. Juergen Polle's laboratory and demonstrates excellent growth characteristics and lipid accumulation potential [33]. Initial phylogenetic analysis of the rDNA 18S gene identified this strain as *C. sorokiniana*. However, another molecular marker for phylogenetic analysis, the rDNA internal transcribed spacer region 2 (ITS2), revealed that strain 1412 falls into the Chlorellales order, although the family, genus, and species was not resolved. Productivity data on the indoor and outdoor performance of this strain can be found in the final NAABB report [31,32].

2.2. Genome sequencing and assembly

2.2.1. DNA preparation and sequencing

For the short-read assembly of *C. sorokiniana* 1230, genomic DNA was purified following a standard protocol. Briefly, cells were resuspended in SDS-EB buffer (2% SDS, 400 mM NaCl, 50 mM EDTA, 100 mM Tris-HCl [pH 8]) by vortexing and extracted twice with phenol/chloroform/isoamyl alcohol (25:24:1 by volume). Nucleic acids were precipitated by adding two volumes of 100% ethanol. Resuspended nucleic acids were then treated with RNase A for 1 h and genomic DNA precipitated with CTAB (10% *w/v*) in 0.7 M NaCl, to remove polysaccharides. The DNA pellet was finally resuspended in TE buffer.

For all other assemblies, high molecular weight algal gDNA was extracted from cells imbedded in agarose, purified and concentrated using AMPure PB beads. The DNA was then fragmented using Covaris g-Tubes. Fragmented and purified DNA was processed for 20 kb SMRT bell library prep. The long insert libraries were size selected using a Blue Pippin instrument (Sage Sciences, Beverly, MA). The sequencing primer was annealed to the selected SMRT bell templates. The libraries were bound to DNA polymerase and loaded on the PacBio RSII for sequencing. Sequencing was completed using either C2/P4 or C3/P5 chemistry and 3-h movies.

2.2.1.1. C. sorokiniana 1228. The 1228 draft genome was generated by utilizing PacBio sequencing [39] and OpGen optical mapping (OpGen, Gaithersburg, MD) to align contigs at the chromosome level (Genbank accession number PQAU00000000). HGAP version 2.2.0 [40] was used to assemble the genome. Forty-one SMRT cells of data were used with each preparation using the 20 kbp prep and C3P5 chemistry with Blue Pippin size selection. The 41 SMRT cells of PacBio sequencing generated 6.050 Gbp of data, and coverage of the genome with PacBio data is 99.19×. The consensus sequences were shredded into 20 kbp overlapping pieces and assembled with Phrap (SPS-4.24) [41,42]. Some editing in Consed [43] was done to create the final assembly. The scaffolding is based on alignment of the contigs to the OpGen maptigs which were generated using *Bam*HI. For optical mapping, the DNA from *C. sorokiniana* 1228 was prepared according to the described methods in the Opgen technical bulletin [44].

2.2.1.2. C. sorokiniana 1230. The 1230 genome was assembled by merging of long-read and short-read assemblies (Genbank accession number PKFC00000000). The long-read PacBio assembly was generated with HGAP version 2.3.0 [40]. Approximately 5.840 Gbp of PacBio data was sequenced [39], providing 99.8× coverage of the genome. The short-read Illumina assembly was generated by preparing four libraries of average insert size of 300 bp, 500 bp, 2000 bp and 5000 bp and sequencing using an Illumina HiSeq 2500 instrument by Cofactor Genomics (St. Louis, MO). Approximately 25.816 Gbp of Illumina short reads were assembled by Cofactor Genomics and computationally annotated and further manually curated, removing organellar genome contigs, at the University of Nebraska-Lincoln. All consensus sequences from these assemblies were computationally shredded and reassembled with Phrap, version SPS-4.24 [41,42] to allow for manual editing and curation with Consed [43].

2.2.1.3. C. sorokiniana 1412. The 1412 genome assembly was generated using a combination of Illumina [45] and PacBio [39] technologies (Genbank accession number PKFD00000000). Illumina short-insert paired-end libraries were constructed and sequenced on the HiSeq instrument generating 33.624 Gbp of data. Illumina data were assembled with Velvet, version 1.2.08 [46] and with Newbler, version 2.6 (from 454 Life Sciences). Additionally, a PacBio long read library was created and sequenced on the RS II instrument generating 10.433 Gbp of draft data. These data were assembled with HGAP, version 2.3.0 [40]. All consensus sequences were computationally shredded and reassembled with Phrap, version SPS-4.24 [41,42] and manually edited with Consed [43]. The final assembly includes 65 contigs greater that 20 Kbp with an estimated genome size of 57.88 Mbp. The estimated fold coverage of the genome is 476× and 180× for Illumina and PacBio data, respectively. Illumina reads were mapped to the 1412 assembly using Bowtie 2 [47], implemented in EDGE [48] with default parameters.

2.2.2. Genome annotation and statistics

For each genome assembly, an in-house custom MAKER2 pipeline [49] was used for structural gene annotation. For the 1412 annotation, Trinity (release 2013-2-25 [50,51]) was used to assemble approximately 24 Gb of paired end HiSeq Illumina (2 × 101 bp reads) RNA-seq data which resulted in 23,329 assembled transcripts. These assembled transcripts were fed into MAKER2 to improve structural annotation. Functional annotation of genes was performed with InterProScan version 5.21 [52]. Basic genome statistics were calculated with the PERL script "assembleaton_stats.pl" [53], and GAG: Genome Annotation Generator [54].

2.3. Comparative analyses

2.3.1. Chlorellales chloroplast phylogenetic tree

We generated a concatenated phylogeny using Bayesian Markov chain Monte Carlo, implemented in MrBayes version 3.2.2 [55], and maximum-likelihood analyses, implemented in RAxML version 8.2.10 [56]. The data matrix included sequences for 10 Chlorellales terminal taxa, including six species and strains of *Chlorella*. The outgroup taxa represented three non-*Chlorella* species [57]. The sequence data consisted of 27 chloroplast protein sequences (Supplementary Table 1); mitochondrial and nuclear genes were not used due to lack of available data from other genera within the Chlorellales. Accordingly, we did not perform multi-locus species-tree analyses since the chloroplast genes effectively belong to the same locus. Therefore, these genes should be less influenced by incomplete lineage sorting due to the reduced effective population size of the chloroplast genome [58]. Genes were aligned using Muscle version 3.8.31 [59] before concatenation. The best-fitting combination of partitioning scheme and protein substitution models was determined using PartitionFinder version 2.1.1 [60] using AICc and a greedy search algorithm with branch lengths linked across partitions. A total of 27 possible partitions were initially defined (one for each protein) and the best-fitting strategy included 13 data blocks (Supplementary Table 1). Four independent Bayesian runs of four chains each (three heated chains and one cold chain) were run for 2×10^7 generations with a burn-in of 5×10^6 generations. Trees were sampled every 100 generations. We considered the runs to have adequately sampled the solution space when the standard deviation of split frequencies was below 5×10^{-3}. The tree was independently constructed using maximum likelihood (ML) methods with the rapid bootstrap analysis and the same partition scheme. Fifty ML replicate trees were used to estimate bootstrap support.

2.3.2. Full genome level comparisons

For the nucleotide identity analysis, queries of nucleotide identities were performed using a custom Perl script which calculates the percentage of a query genome that matches a reference genome at a user specified nucleotide identity level. Once the percentage nucleotide identity is determined by the user (80%) then Nucmer, from the MUMmer package [61] was utilized to identify homologous regions of the reference and query genomes that share a nucleotide identity at the specified level.

2.3.3. Gene level analysis to determine non-homologous genes within the three C. sorokiniana genomes

A three-way comparison of annotated genes within the three *Chlorella* strains reported here was carried out with BLASTP and TBLASTN [62] to identify potentially unique genes within each strain. For each *Chlorella* strain, predicted translated amino acid sequences for each gene were queried with TBLASTN and BLASTP against the other two strains. Genes were considered homologous to one or both of the other genomes when a) > 50% of the query gene product was accounted for in the full alignment length, b) the amino acid identity of the alignment was > 40%, and c) the blast expect value (*E*-value) was $< 1 \times 10^{-10}$ for BLASTP or $< 1 \times 10^{-5}$ for TBLASTN. Genes were considered unique within a genome (lacking a homolog in the other two *C. sorokiniana* genomes) when the previous criteria were not met. Non-unique gene counts are calculated as average number of genes shared based on all reciprocal analyses performed.

2.3.4. ViridiCut2, sex-related genes, and flagella-related genes

The presence of ViridiCut2 genes in *C. sorokiniana* was investigated by searching the genomes of *C. sorokiniana* 1228, *C. sorokiniana* 1230, and *C. sorokiniana* 1412 with the collection of ViridiCut2 *Chlamydomonas reinhardtii* proteins [4]. BLASTP with an *E*-value of 1×10^{-5} was used to search *C. sorokiniana* proteins [62]. Then, the functional characterizations of both *C. reinhardtii* and *C. sorokiniana* were obtained via standalone InterProScan version 5.21–60.0 [52]. Functional annotations included Gene3D version 3.5 [63], PANTHER version 10.0 [64], and Pfam version 30.0 [65]. Functional annotations with an *E*-value above 1×10^{-5} were ignored. *C. sorokiniana* proteins were determined to be orthologous if any of the functional annotations matched InterProScan output functional annotations from Gene3D, PANTHER, or Pfam. The match with the lowest *E*-value was automatically determined as orthologous if the unordered list of unique annotations were identical. Gene duplication in *C. sorokiniana* was not considered. If neither the *C. reinhardtii* nor the *C. sorokiniana* genes had any functional annotations, genes were considered orthologous if the E-value of the match was below 1×10^{-15}. For matches that were not automatically determined to be orthologous, MUSCLE (version 3.8.31) [59] alignments were created and manually inspected and compared to the functional annotations. Often, one gene appeared to be a functional subset of the other gene (either *C. reinhardtii* or *C. sorokiniana* was substantially longer, including additional functional domains), possibly due to inaccurate gene modeling. In these rare cases (4.3–8.3% of ViridiCut2 genes), matches were deemed orthologous. As we are evaluating the presence of specific sets of genes, the decision to accept rare cases of functional subsets results in a conservative estimate for the number of genes absent or lost in *C. sorokiniana*. Additionally, automatically assigning orthology based on unordered and unique annotations (as opposed to ordered and accounting for repeated domains) results in a conservative estimate for the number of ViridiCut2 and sex-related genes absent or lost in *C. sorokiniana* strains. We evaluated the number of genes present in the homologous recombination and meiosis (yeast) KEGG maps using the KEGG Automatic Annotation Server [66].

In order to ensure that genes annotated as missing from *Chlorella* genomes are indeed absent, all genes in the *Chlorella* genome assemblies were masked. The remaining intergenic regions were searched using TBLASTN [62] with an *E*-value of 1×10^{-6}. Significant matches were considered to indicate the presence of that gene.

2.3.5. PKS genes

The genomes of *C. sorokiniana* 1228, 1230, and 1412 were searched using a previously annotated PKS gene from *Chrysochromulina tobin* [67] as a query. This initial search identified a PKS gene from each *C. sorokiniana* strain. These three genes were used to re-query the genomes of *C. sorokiniana*. When significant hits to the genome were found outside of annotated genes, these regions were extracted and Augustus version 3.0 [68] was used to predict gene models in those genomic regions. Significant tblastn hits (*E*-value = 1×10^{-5}) were treated as manually identified 'exonpart' hints, Augustus was trained on *C. reinhardtii*, and exactly one gene was predicted for each genomic region. This procedure was validated by annotating the genomic regions of the three initial PKS genes, which resulted in high similarity (> 95% identity) to the original maker annotations.

Proteins were annotated in two ways. First, the protein translations of predicted gene models were annotated using Pfam version 31.0 [65]. However; these models had several sequences encoding canonical PKS domains in predicted introns. These domains were absent from predicted protein annotations. Additionally, several sequences encoding domains were split across intron/exon boundaries. Therefore, a second, functional annotation was performed without gene predictions. A three frame translation of genomic regions were searched with hidden Markov models (HMMs) downloaded from Pfam version 31.0 [65] using HMMer version 3.0 [69]. An *E*-value threshold of 1×10^{-5} was used. Functional annotations for all three translations were sorted to produce the protein's final functional annotation.

The two methods of annotation produced broadly similar protein annotations (in terms of protein length and number of PKS modules). However, the first, model-based annotation often excluded dehydratase, enoyl reductase, and acyl carrier protein domains. In one case, this resulted in a protein without dehydratase and reductase domains, resulting in nine modules consisting of ketide synthase and ketoreductase domains. These two methods produced highly similar PKS gene models in 1412, suggesting that the models are accurate and that RNA expression data is highly valuable for PKS gene modeling.

Because Pfam does not identify families with the medium-chain dehydrogenase/reductase superfamily [70], protein domains annotated as zinc-dependent alcohol dehydrogenases (PF00107) were considered enoyl reductases, following Shelest et al. [26]. Given the challenges of producing manually curated annotations for PKS genes (which may be up to 82.9 kbp long) in the absence of RNA expression data, these functional annotations must be considered preliminary until models are further validated with additional RNA or protein expression data.

2.3.6. DNA methylation genes

The presence of epigenetic machinery was determined in the *C. sorokiniana* strains using queries of known *Arabidopsis thaliana*, *Homo sapiens*, *Mus musculus*, or *Zea mays* protein sequences against *C. sorokiniana* protein sequence data. Once sequences with similar homology were found for each protein for each strain, these sequences were queried using BLASTP [62] and for Pfam domains [65]. Protein sequences were confirmed by query of annotated genomes using Pfam and Interpro domains considered essential for epigenetic function in each protein for each strain.

3. Results and discussion

3.1. General genome characteristics

3.1.1. Genome assembly statistics and phylogeny

High quality genome assemblies were generated for all three strains using Single Molecule Real Time (SMRT) sequencing technology by Pacific Biosciences (PacBio) [39]. The three genome assemblies ranged from 57.9 Mb to 61.4 Mb with the highest quality assembly (*C. sorokiniana* 1230) consisting of 20 contigs (18 nuclear genome contigs, 1 mitochondrial and 1 chloroplast contig), which represents nearly chromosome resolution (12 chromosomes). Thus, these genomes build on the relatively few nearly-complete algal genome assemblies [71–74].

For each genome, the number of predicted genes is 12,166 (1228), 12,611 (1412), and 12,871 (1230) (Table 1). Genes in 1228 and 1230 were relatively coding rich with an average of 10 exons per gene (9 introns per gene), while *C. sorokiniana* 1412 averaged 13 exons per gene (12 introns per gene), which correlated to a substantially longer average gene length. Average intron length was not substantially longer in 1412 compared to 1228 and 1230 (Table 1). The three *Chlorella* genomes are very gene rich, ranging from 31.1% of the genome covered by CDS regions in 1228 to 42.1% in 1412. This corresponds to gene coverage of 72.9–92.5% of the genome respectively.

At the rDNA 18S level, the three *C. sorokiniana* strains described in this study appear nearly identical (Supplementary Table 2). The 18S rDNA gene coding region appears to be copied multiple times within a region of Chromosome 1, which confounded assembly efforts in each *C. sorokiniana* genome we assembled. Therefore, an alternative seemingly non-coding or truncated partial 18S sequence occurred multiple times in each *C. sorokiniana* assembly. In addition, repeat regions in the genome assemblies led to unplaced contigs representing 1.9% and 0.44% of the assembled genomes of 1228 and 1412 respectively (Table 1). Based on the analysis of the 18S repeat region, it is likely that some repeat regions are not entirely represented in the 1230 genome since the 18S region was collapsed into only two copies in the *C. sorokiniana* 1230 final assembly. Chromosome assignments for each contig

Table 1
Assembly and annotation statistics of *Chlorella sorokiniana* 1230, 1228, and 1412.

	C. sorokiniana 1230	*C. sorokiniana* 1228	*C. sorokiniana* 1412
Genome assembly:			
Assembled genome size	58.5 Mb	61.4 Mb	57.9 Mb
Contigs in assembly	20	64	65
Sequencing/assembly methods	PacBio + Illumina	PacBio + OpGen optical mapping	PacBio + Illumina
Average contig size	2660 kb	959 kb	890 kb
N50	3.82 Mb	2.41 Mb	2.20 Mb
GC content	63.80%	65.30%	64.10%
Read coverage (Pac-Bio)	100×	99×	476×
Read coverage (Illumina)	N/A	N/A	180×
Unscaffolded contigs	0	24	8
Length of unplaced contigs	–	1170 Kb	256 Kb
% of genome assembly	–	1.90%	0.44%
Genome annotation:			
Annotated protein coding genes	12,871	12,166	12,611
Average gene length	3754 bp	3681 bp	4474 bp
Average CDS length	1632 bp	1572 bp	1932 bp
Average exon length	152 bp	152 bp	156 bp
Average exons per gene	10.9	10.3	13
Average intron length	215 bp	227 bp	231 bp
% of genome assembly as genes	82.50%	72.90%	92.50%
% of genome as CDS	35.90%	31.10%	42.10%

are displayed in Supplementary Fig. 1.

In order to determine the relationships of sequenced representatives of *Chlorella sorokiniana*, we estimated a species tree using 27 concatenated chloroplast protein sequences. Based on the species included (Fig. 1), free-living *C. sorokiniana* forms a clade, sister to the symbiotic *C. variabilis*, *C. vulgaris*, and *Micractinium conductrix*. The recently sequenced *C. sorokiniana* UTEX 1602 [75] is closely related to *C. sorokiniana* DOE1412 (4 differences out of 6778 amino acids), demonstrating that the diversity of available *C. sorokiniana* genomes is well represented in our other analyses.

Genome completeness was analyzed using BUSCO (Benchmarking Universal Single Copy Orthologs) version 3.0 [76]; we queried both the genome assemblies and the protein annotations against the core eukaryotic BUSCO dataset (Supplementary Table 3). These results show that the three genomes are consistent in BUSCO content among each other (80.2–81.2% of BUSCO genes at nucleotide level; 92.4–94.7% BUSCO genes at protein level); however, between protein-based and genome-based analyses there is a large discrepancy between the number of genes identified, likely due to using a BUSCO gene set that is not specific to algae. Based on the divergent set of these BUSCO genes, the higher number of genes identified using protein-level analysis suggests more complete genome assemblies, consistent with our chromosome-level assembly and high N50 values.

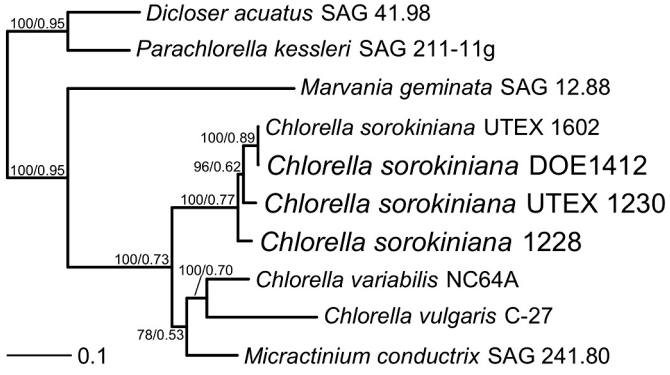

Fig. 1. Phylogenetic analysis of Chlorellales based on 27 concatenated chloroplast proteins (Supplementary Table 1). Species sequenced here are emphasized with larger font (center). Independent Bayesian and maximum likelihood (ML) analyses estimated the same tree topology. Numbers indicate ML bootstrap values and Bayesian posterior probabilities respectively. Branch lengths shown are from the ML estimation. The tree is rooted following Sun et al. [57].

3.1.2. Telomeres and centromeres

The telomere sequence $(TTTAGGG)_n$ has been identified and is consistent in all three assemblies. This sequence is the same telomere sequence as in *Chlorella variabilis* [77]. Assembly of the centromere was problematic in a number of chromosomes (Chromosome 1–4, Supplementary Fig. 1). No detailed analysis of repeat structure within the centromeric region has been performed to determine cause, though many of the regions likely associated with the centromere were identified as unplaced repeat regions in the *C. sorokiniana* 1228 assembly.

3.1.3. Genome scale rearrangement

C. sorokiniana 1228 has a large inversion present on Chromosome 4 when compared to *C. sorokiniana* 1230 and 1412 genomes (Supplementary Fig. 2). In contrast, the 1230 and 1412 do not appear to have large scale genome structural inversions; however, these genomes do show several small inversions (< 200 kbp in size) throughout their genomes (Supplementary Fig. 3). This observation is consistent with the phylogenetic tree, nucleotide identity, and gene comparison analyses (see sections below) that also suggest that the 1412 and 1230 genomes share more similarity to each other than to the 1228 strain.

3.1.4. Nucleotide identity

Despite nearly identical 18S sequences, we determined low nucleotide identity between the three strains of *C. sorokiniana*, though strain 1230 has a single SNP within the full 18S sequence compared to the other two strains (1113C > T; 1700 bp alignment). Over 87% of the genome in each of the three strains has < 80% nucleotide identity (Fig. 2, Supplementary Table 2).

3.2. Analysis of conserved and divergent gene content

3.2.1. Comparison of global gene inventory

To determine if a set of unique genes is present among the three *C. sorokiniana* strains compared, each predicted gene model was compared to the genes annotated in the two other genomes. Cutoffs of *E*-value of 1×10^{-5}, 40% minimum amino acid identity and 50% query coverage were used. A substantial number of proteins in each genome were designated unique by this metric and further characterized with Blast2GO [78] to assign additional function and Gene Ontology (GO) terms to these unique genes. *C. sorokiniana* 1228 had the fewest designated

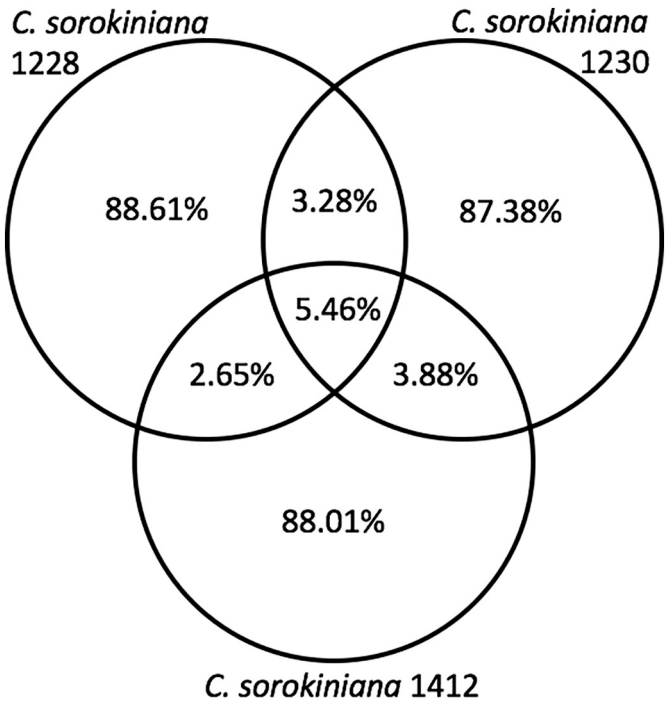

Fig. 2. Intraspecies nucleotide divergence in *Chlorella sorokiniana.* Nucmer was used to identify the regions of genomes that share nucleotide identity; the majority (> 87%) of each genome shares < 80% nucleotide identity when compared to the two other corresponding genomes.

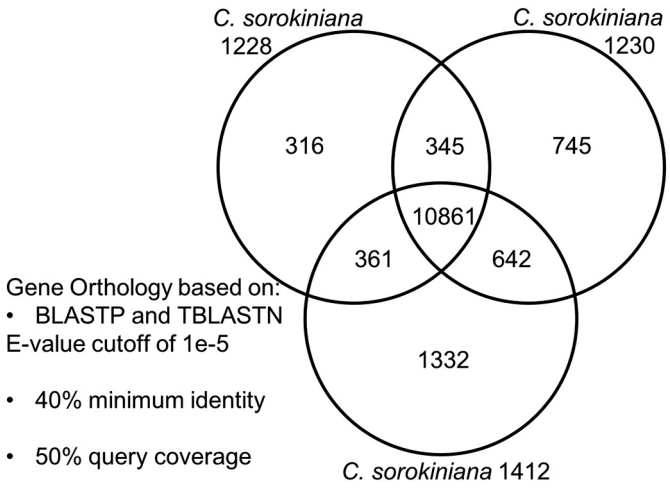

Fig. 3. Comparative analysis of protein coding genes in the three *Chlorella sorokiniana* strains. Number of proteins with an ortholog, as identified by the cutoffs listed, are shown above.

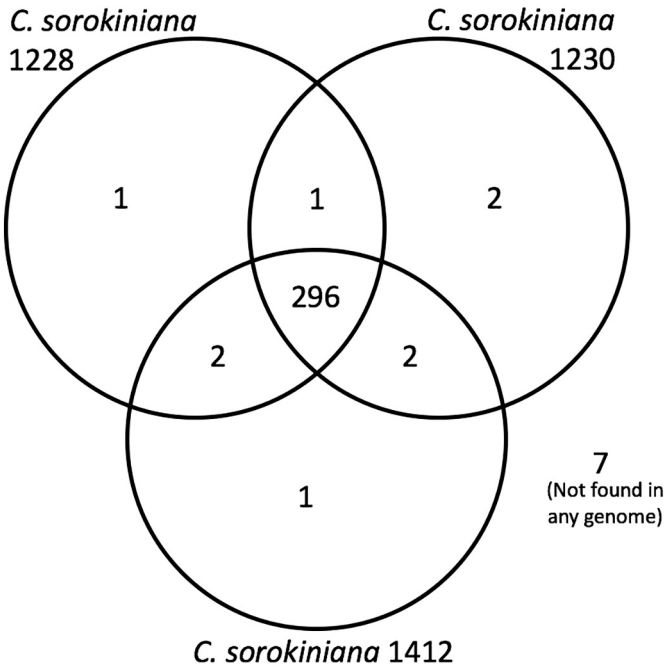

Fig. 4. Conservation of ViridiCut2 photosynthesis-related genes within the three *Chlorella sorokiniana* genomes.

unique genes with 316 (Fig. 3), while *C. sorokiniana* 1412 had the most at 1332. All unique protein sequences are available in Supplementary Datasets 1–3. The majority of these genes (60.5–82%) are of unknown function (Supplementary Table 4). Of interest, some of the unique genes in each genome were assigned Enzyme Commission (EC) numbers (Supplementary Datasets 4–6). Within *C. sorokiniana* 1412 genes, two unique genes (CSJ00004822-RA and CSJ00005316-RA) were designated with EC 3.1.1.4, which is a phospholipase involved in lineolic acid metabolism derived from lectin to linoleate conversion. EC 5.4.99.30 UDP-arabinopyranose mutase was also identified as unique and is utilized in sugar metabolism. Overall, the unique genes identified in each strain warrant further investigation, particularly those genes with unknown function. Beta-*N*-acetylhexosaminidase (CSI_122800011331-RA), potentially found in N-Glycan and Glycosphingolipid biosynthetic pathways, is one of only two potential enzymes with functional annotation identified in the 1228 unique protein list. The other is a potential ATPase (CSI_122800000435-RA) used in purine metabolism.

In *C. sorokiniana* 1230, there are a number of unique genes that were assigned an enzyme code by Blast2GO annotation. These include kynurenine formamidase (CSI2_123000004392-RA), a unique enzyme important in tryptophan synthesis that can convert N-formyl derivatives into formate. A gene of similar function is found in *C. sorokiniana* 1412 and *C. sorokiniana* 1602. Of particular interest is carbonate dehydratase (CSI2_123000007719-RA), an enzyme that catalyzes the hydration of gaseous CO_2 to carbonic acid. Although there are approximately 10 other genes with this function present in all three strains, this gene (CSI2_123000007719-RA) seems to be of unique origin.

3.2.2. ViridiCut2 genes

Given the interest in the core set of genes involved in photosynthesis and algal growth, we annotated Viridiplantae photosynthetic-related genes in *C. sorokiniana*. We found 296 of 312 previously annotated ViridiCut2 genes [4] in all three strains of *C. sorokiniana* (Fig. 4). Only nine ViridiCut2 genes were found in one or two strains of *C. sorokiniana* but not in all three strains. Seven ViridiCut2 genes were not found in any strain of *C. sorokiniana*. These genes are often not well characterized, though two genes are characterized as methyltransferases of unknown function (Supplementary Table 5). Similarly, of the ten ViridiCut2 genes lost in one or two strains of *C. sorokiniana*, two are annotated as bZIP or mTERF transcription factors.

Of the 312 previously annotated ViridiCut2 genes [4], seven were not found in any *C. sorokiniana* strain. This represents a similarly low proportion (2.2%), compared to the 4.0% absence of BUSCO genes (Supplementary Table 3). Similarly, *C. vulgaris* NC64A has lost 15 ViridiCut2 genes (Supplementary Table 5), nine of which are losses shared with *C. sorokiniana*. This low rate of ViridiCut2 gene loss may suggest a relatively low level of genomic photosynthesis specialization, indicating that *C. sorokiniana* may be a photosynthetic generalist and has not adapted to specific photosynthetic/environmental conditions [4]. The consequences of this possible generalist photosynthetic gene repertoire

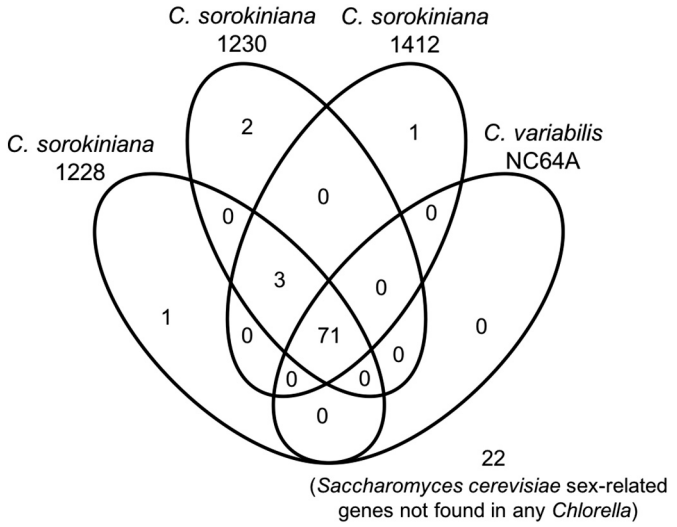

Fig. 5. Conservation and variation in sex-related genes between *Chlorella sorokiniana* and *Chlorella variabilis*. Genes identified by functional domain searches matched sex-related gene in *Saccharomyces cerevisiae* (see Methods section).

on evolutionary trajectories or laboratory selection is unknown.

3.2.3. Sex-related genes

Trebouxiophyceae green algae are often presumed to be asexual [79]; however, some genes involved in meiosis and sexual reproduction exist in several Trebouxiophyceae species [77,79]. This discrepancy suggests cryptic sex is prolific throughout the Trebouxiophyceae class. Building upon this knowledge, we determined the presence of sex-related genes underlying sexual reproduction in *C. sorokiniana* in two ways. First, we searched the *Chlorella* genomes (both *C. sorokiniana* and *C. variabilis*) with 100 annotated sex-related and meiosis-related genes from *Saccharomyces cerevisiae* [80]. We found 71 of these genes in all *Chlorella* genomes (Fig. 5). Three genes are found in all *C. sorokiniana* genomes and are absent in *C. variabilis* (Fig. 5). Of the 100 sex-related genes in *Saccharomyces cerevisiae*, 22 are missing in *Chlorella* genomes (Fig. 5). To determine if the lack of these 22 genes would prevent sexual reproduction in *Chlorella*, we performed a second analysis using the *C. reinhardtii* genome as a positive control. It is well established that *C. reinhardtii* has sexual reproduction [81–83]. Of the 22 genes absent in *Chlorella*, only a single gene was found in *C. reinhardtii*; this gene is involved in recovery from DNA damage/replication checkpoint arrest (Supplementary Fig. 4). Second, we used yeast KEGG annotations to identify *Chlorella* genes involved in homologous recombination and meiosis [66]. Of the 44 homologous recombination genes in the KEGG map, we found 25 genes in *C. sorokiniana* and 16 genes in *C. variabilis* (Fig. 6, Supplementary Table 6). Of the 99 meiosis-related genes in the KEGG map, we found 38–39 genes in *C. sorokiniana* and 36 in *C. variabilis* (Supplementary Fig. 5, Supplementary Table 6). *C. reinhardtii* has 24 genes in the homologous recombination KEGG map and 35 genes in the meiosis KEGG map (Fig. 6, Supplementary Figs. 5-6, Supplementary Table 6).

Based on the gene set used here [80], it is likely that all strains of *Chlorella*, including the endosymbiont *C. variabilis* NC64A [77], are capable of sexual reproduction or minimally, were capable of sexual reproduction recently in their evolutionary history. However, the high degree of contiguity in the assemblies and modest level of SNP-level heterogeneity observed in the 1412 Illumina sequencing reads (Supplementary Fig. 7), suggest that source cultures contain haploid genomes. Therefore, this genomic data cannot conclusively demonstrate sexual reproduction by identifying signatures of recombination. In general, these results substantiate previous investigations into sex-related genes in green algae [77,79], though with an expanded set of sex-related genes (Figs. 5–6, Supplementary Fig. 5, Supplementary Fig. 6). Further experimentation is required to characterize any sexual cycle or demonstrate if a transient diploid state exists in these *C. sorokiniana* strains.

3.2.4. Flagella-related genes

The flagellar system is critical for initiating sexual reproduction in *Chlamydomonas*, through both flagellar agglutination to initiate mating and the use of basal bodies to coordinate mitosis and cytokinesis [84,85]. Given this critical role of flagella-related genes in sexual reproduction, we investigated the presence of *Chlamydomonas* flagella genes, identified by Prochnik et al. [86], to complement our analysis of sex-related genes. We found the complete or nearly complete gene sets for protein secretion, membrane trafficking, kinesin motor proteins, microtubule cytoskeleton, and actin cytoskeleton in all strains of *Chlorella* (Supplementary Table 7). There are fewer basal body proteins in *Chlorella* relative to *Chlamydomonas*, but it is uncertain whether this represents a reduction of basal body proteins in *Chlorella* or an expansion in *Chlamydomonas* (Supplementary Table 7). The set of flagellar-related genes we found in *Chlorella* suggests the capacity for sexual reproduction, consistent with our analysis of meiosis-related genes. A more detailed analysis of flagellar system and basal body proteins in green algae is necessary to understand the critical components for sexual reproduction.

3.2.5. Gag retrotransposons

Gag retrotransposon-like signatures were identified within each of the three *Chlorella sorokiniana* strains. These transposon elements are located in unique genomic regions in each strain (Supplementary Fig. 1); thus, viral intrusion into these genomes likely occurred after their divergence. If the transposition events took place pre-divergence, we would expect to see similarities between the location and sequence of the transposon signature, which is not observed. The unique sequences within each individual genome cluster by strain (Fig. 7), demonstrating that the transposon insertion events occurred after divergence of the strains.

3.2.6. Polyketide synthetases in C. sorokiniana

Given the importance of antibiotic PKS genes in green algae [25], we characterized the PKS repertoire in *C. sorokiniana* and compared this repertoire to *C. reinhardtii* and *Ostreococcus lucimarinus*. We found a diverse set of Type I PKS genes in all strains of *C. sorokiniana*; each strain of *C. sorokiniana* has four PKS genes, which have between six and 11 modules (Fig. 8, Supplementary Dataset 7). We found three additional, single-domain, Type II PKS genes in each strain. The Type I PKS genes form four functionally similar groups, two of which are functionally highly conserved among strains (groups 1 and 2, Fig. 8). Group 1 appears unique among described PKS genes given the nucleoporin autopeptidase domain at the C-terminus of the protein, rather than the more canonical thioesterase domain (Fig. 8). The prevalence and function of this unique nucleoporin autopeptidase domain may be underappreciated. Groups 3 and 4 are more variable between strains. Multiple modules have variable presence of dehydratase and enoyl reductase domains (Fig. 8). For example, the sixth module in group 4 PKS genes is functionally different in all three strains. This PKS innovation between *C. sorokiniana* strains is consistent with our analyses suggesting significant genetic differences and gene content between *C. sorokiniana* strains.

The PKS repertoire of *C. sorokiniana* is substantially larger than *C. reinhardtii*. Both have three, single domain, Type II PKS genes, which likely interact with other proteins to form iteratively functional polypeptides. However, *C. reinhardtii* only has a single Type 1 PKS gene [25,26]. This gene is in Group 4 (Fig. 8), which suggests strong evolutionary selection for the maintenance of this particular polyketide product. Similarly, the PKS repertoire of *C. sorokiniana* is highly divergent from *Ostreococcus lucimarinus*, which contains five single-

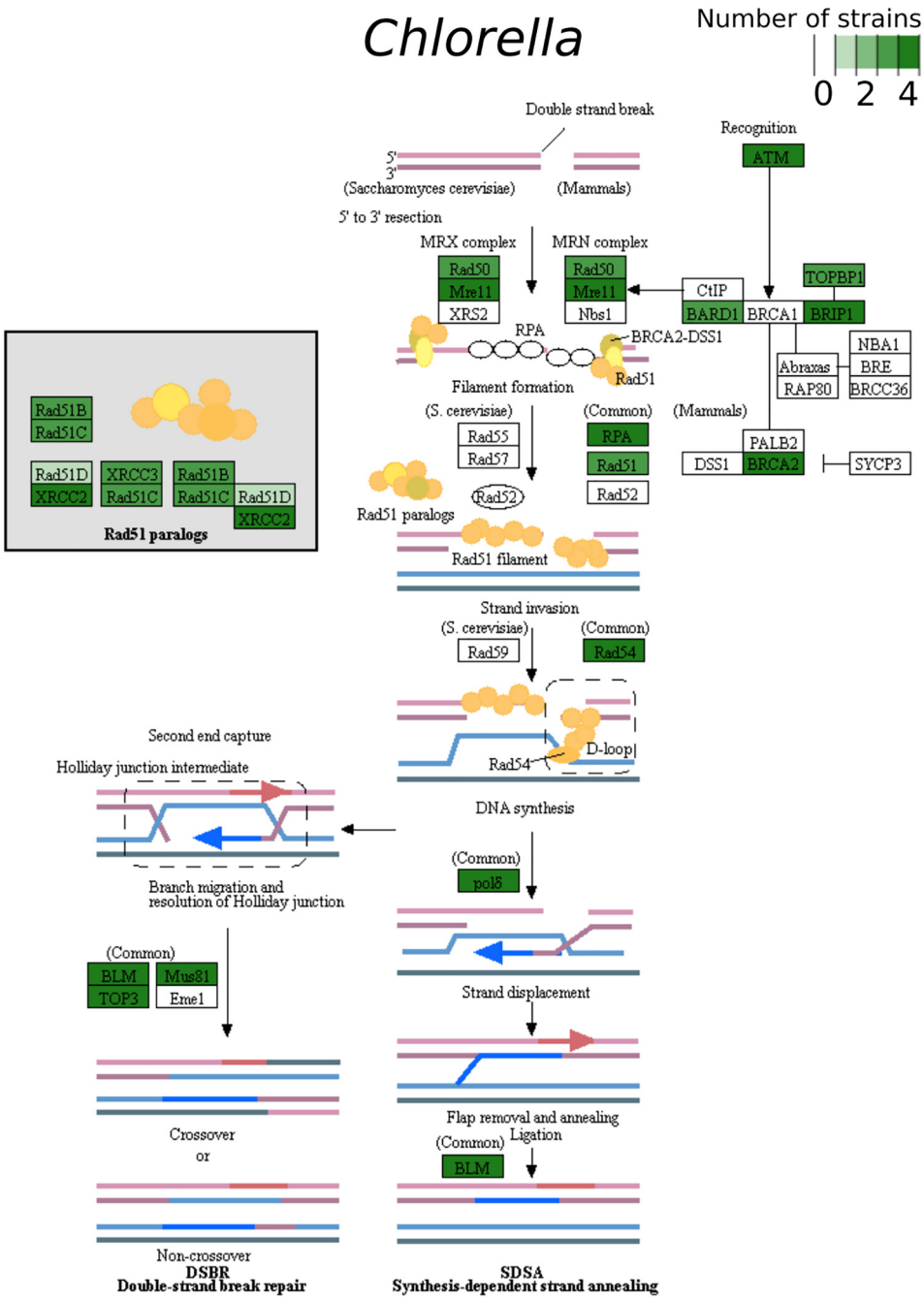

Fig. 6. Inventory of genes related to homologous recombination in four genomes of *Chlorella*. See Supplementary Fig. 6 for the abundance of genes related to homologous recombination in *Chlamydomonas*.

domain, Type II PKS genes and three modular Type I PKS genes. The structure of these three Type I PKS genes is not similar to *C. sorokiniana* PKS genes (Supplementary Fig. 8). Specifically, *O. lucimarinus* PKS genes are either shorter (four modules) or longer (14 modules) than *C. sorokiniana* and include alternate functional domains, including condensation and sulfotranferase domains (Supplementary Fig. 8).

We are currently unable to predict the products produced by these PKS genes, though it is likely that the products do vary based on how significantly different the domains are in content, as well as number of modules within each PKS complex. The diverse PKS repertoires of *C. sorokiniana*, *C. reinhardtii*, and *O. lucimarinus* emphasizes the variation in polyketide synthases across green algae. The structural diversity of polyketide synthase genes within *C. sorokiniana* strains and between green algae species suggests high levels of evolutionary PKS innovation. Given the potential roles of these PKS products in antibiotic production and cellular defense, further attention to the evolutionary diversity and function of PKS genes in green algae is warranted.

3.2.7. DNA methylation machinery

We determined the presence of epigenetic machinery responsible for DNA modifications and potential mechanisms of epigenetic transcriptional repression that *C. sorokiniana* strains utilize. Common DNA methylation enzymes responsible for 5-methylcytosine (5mC) are present in all *C. sorokiniana* genomes. These proteins contain specific domains for 5-methylcytosine methyltransferase activity recognizing hemi-methylated DNA for maintenance of DNA methylation. The presence of multiple DNA methyltransferase orthologs suggests that *C. sorokiniana* maintains the ability to methylate DNA in the same nucleic acid

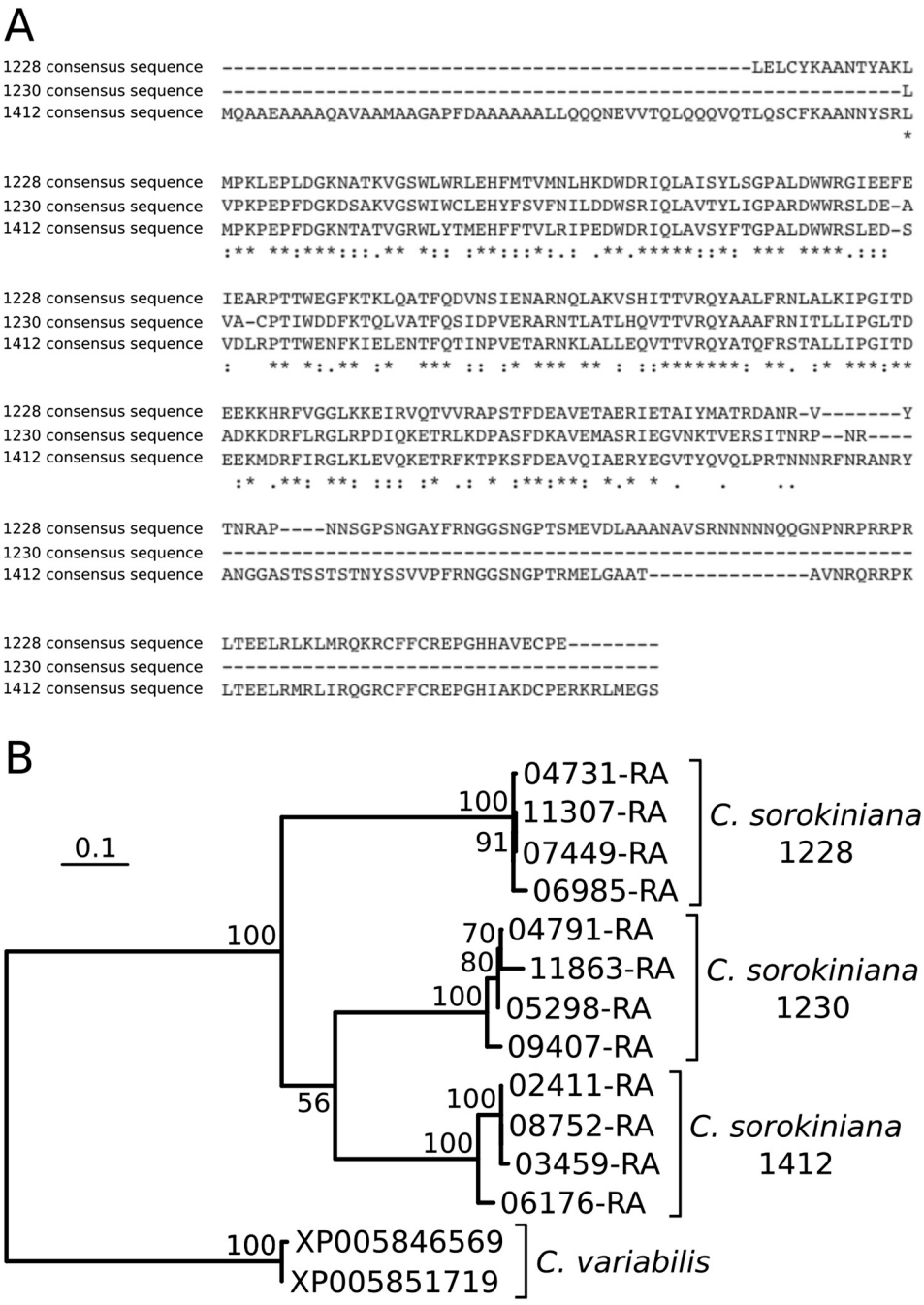

Fig. 7. (A) Alignment of the three-strain specific consensus gag-transposon sequences. (B) A maximum likelihood phylogenetic tree of the unique sequences identified within each strain.

contexts as plants; *i.e.* mCG, mCHH, mCHG, where H is any nucleotide except G (Table 2, Supplementary Table 8).

While the capacity for DNA methylation is similar among all *C. sorokiniana* strains investigated, the enzymes responsible for these modifications differ among strains. For example, *C. sorokiniana* 1228 has more sequence similarity to DNMT1 than MET1, though it does have appreciably identical sequences to DMT1, a MET1 ortholog. Clearly, there is redundancy among these sequences such that mCG likely occurs in *C. sorokiniana* species for heterochromatin formation.

Previous studies suggest the presence of DNA methylation proteins similar to *A. thaliana* in some microalgae species, including *Chlorella variabilis* NC64A and *C. sorokiniana* [18]. These proteins, called chromomethylases (CMTs) contain three functional domains: a CHROMO domain and two domains present in other DNA methyltransferases, C-5 cytosine methyltransferase (C-5MT) and BROMO adjacent homology (BAH) domains. We did not find the presence of these three essential CMT domains (CHROMO, BAH, C-5MT) in the same sequence context in any of the *C. sorokiniana* strains. Lack of CMT proteins would suggest the inability to perform methylation in the CHH context. Both CMT2 and DRM2, an RNA-directed DNA methylation (RdDM) protein, utilize a positive-feedback loop with methylation of lysine 9 on histone 3 (H3K9) for mCHH [87]. We were unable to identify any sequence homology to CMT2 or any RdDM proteins of the DRM family (Table 2). In plants, RdDM processes are responsible for *de novo* methylation, while in mammals, this type of methylation is conducted by DNMT3a and DNMT3b. Sequences analogous to proteins from these families were not found in any of the *C. sorokiniana* strains. However, one DNA methyltransferase, DMT1, found in C. *sorokiniana* 1228, has been

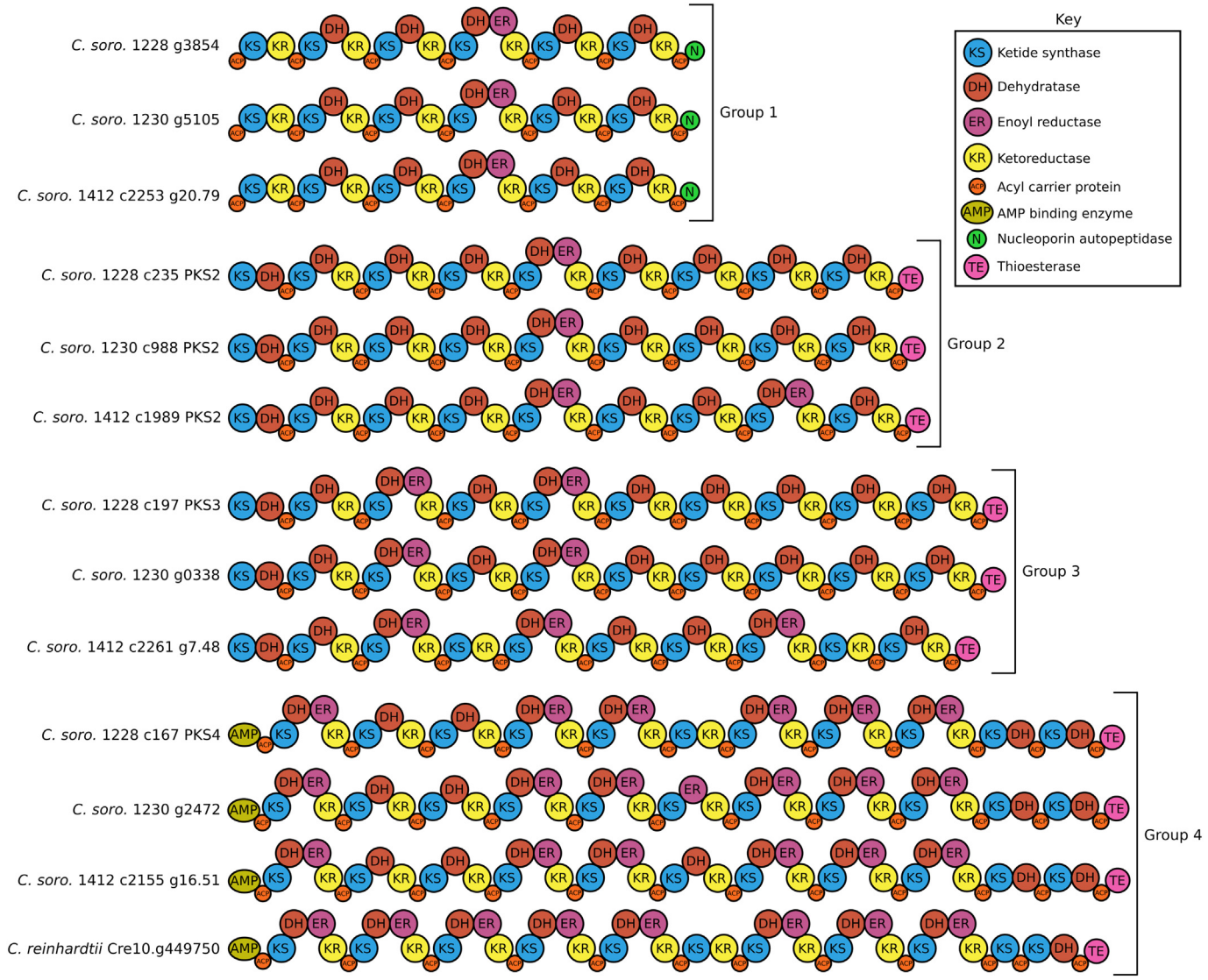

Fig. 8. Domain structure of Type I PKS genes found in *Chlorella sorokiniana* and *Chlamydomonas reinhardtii*. Note that the only Type I PKS gene identified in *C. reinhardtii* is very similar to one of the PKS genes in all three *C. sorokiniana* genomes (Group 4).

Table 2
Epigenetic-specific genes within three *Chlorella sorokiniana* strains.

Epigenetic machinery genes	*C. sorokiniana* 1228	*C. sorokiniana* 1230	*C. sorokiniana* 1412
DNA methyltransferase (maintenance) (*DNMT1*, *MET1*, *MET2*, *MET3*)	3	3	2
De novo DNA methyltransferase (*DNMT3a*, DNMT3b)	0	0	0
Chromomethylases (*CMT1*, *CMT2*, *CMT3*)	0	0	0
RNA directed DNA methylation (*DRM1*, *DRM2*)	0	0	0
TET (*TET1*, *TET2*)	0	0	0
DEMETER (*DML1*, *ROS1*)	1	1	1
Dicer (*DCR*, *DCL1*)	4	3	2
Argonaute (*AGO1*, *AGO2*)	1	1	1

shown to *de novo* methylate DNA *in vitro* [88]. It's unclear if these *C. sorokiniana* strains methylate DNA in the CHH context or have the capability to methylate DNA *de novo*; experimental validation is necessary to determine the distinction in methylome characteristics among the strains.

To fundamentally alter heterochromatin structure, cellular machinery must have the capacity for removal of DNA methylation. In all three *C. sorokiniana* species, we found DEMETER-like proteins, which are DNA glycosylases responsible for active demethylation of maternal alleles induced by RdDM processes (even though the strains lack essential RdDM proteins) [89]. The presence of DNA glycosylases in *C. sorokiniana* species implies that these microalgae do not require hydroxymethylation of cytosine in DNA (5hmC) found in animals. In mammals, 5hmC is catalyzed by TET family proteins for downstream removal of DNA methylation as mammals lack direct DNA demethylase capability [90]. As both 5mC and 5hmC respond to bisulfite treatment in the same manner, it is difficult to resolve the two base modifications when identifying the methylome, though recent improvements in epigenetic assays have been developed for this purpose. We did not find TET proteins in the *C. sorokiniana* species, suggesting that 5hmC is not present in the methylome. Indeed, the *A. thaliana* genome does not contain appreciable quantities of 5hmC, though it should be noted that DEMETER proteins are able to remove this modification [91,92]. While it is evident that plants do not require the 5hmC intermediary modification for inactivation of repression, it's still unclear if microalgae

possess both 5mC and 5hmC. Thus, we recommend specific delineation of these two modifications for experimental analysis of the microalgae methylome.

There is still much debate concerning which RNA-dependent mechanism can be classified as “epigenetic” in nature. However, RNA-directed modification of the genome and feedback loops involved snRNAs are common in plants and are therefore likely to occur in microalgae. We found many components of the RNA-induced silencing complex (RISC), including the essential proteins DICER and ARGONAUTE (Table 2, Supplementary Table 8). Additionally, we found many proteins involved in direct RNA modification in all three strains. Thus, *C. sorokiniana* strains likely employ RNAi machinery capability coupled with DNA maintenance methylation to inhibit transcriptionally active RNA and perpetuate the methylome during mitosis. These maintenance methyltransferases, while different among each strain, all perform the same function, and potentially have a dual role in *de novo* methylation as the presence of transposable elements would require some epigenetic protective mechanism for the genome.

This preliminary analysis suggests that all three strains contain enzymes for DNA methylation in the CG and CH contexts, though it is unclear if those enzymes can specifically modify mCHH. In contrast to another report [18], we found no evidence that any of the strains contain CMTs (chromomethyltransferases). Further, these strains contain genes for DNA glycosylases and lack genes for TET proteins, suggesting active DNA demethylation capability and no 5hmC formation. However, RNA-directed DNA methylation (RdDM) process machinery was not identified, thus it's unclear how these organisms perform *de novo* DNA methylation. Further biochemical analysis of the methylome and associated proteins will provide deeper understanding of these processes; however, our analysis suggests that these *Chlorella* strains likely utilize DNAme for genomic stability and potentially, transcriptional regulation.

4. Conclusions

Here we describe a comparative analysis of three unique *Chlorella sorokiniana* genomes. Analysis of nucleotide identity shows that < 15% of the genomes contain 80% nucleotide identity, which calls into question the species taxonomy of these *Chlorella*. Other aspects of these genomes are relatively conserved, including photosynthetic generalization and sex-related genes. Cataloging meiosis and other sex related genes in *C. sorokiniana* showed that this species likely maintains the ability to perform meiosis and homologous recombination. If sexual reproduction is observed, strategic breeding could be utilized to improve biomass and bioproduct yields in this lineage. Our preliminary investigation of epigenetic machinery in the *Chlorella* genomes indicates that DNA methylation machinery is present. Overall, future research may utilize the capacity for sexual reproduction and genomic differences in *C. sorokiniana* to realize the potential for *C. sorokiniana* bioproduct and biofuel production.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.algal.2018.09.012.

Acknowledgements

This work was supported by the Bioenergy Technology Office within the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy through contract DE-NL0029949. Special thanks to Cheryl Gleasner and Kim McMurry for technical assistance.

B.H. performed the genome annotation, global comparative analysis including genome rearrangement and assembly statistics, unique gene analysis, and retrotransposon analysis. E.R.H. performed the global comparative analysis including species phylogeny, ViridiCut2 gene, sex-related gene, and PKS gene annotations. C.R.S. performed the epigenetic analysis. C.L. completed the comparative nucleotide identity analysis. Y.K., K.D., and H.D. contributed to the sequencing and assembly of all three *Chlorella* genomes. J.M. purified genomic DNA and S.C., S.E., and J.J.R. contributed to the assembly and annotation of the UTEX 1230 genome. J.P. extracted DNA from DOE1412 and conceived the project. S.R.S. designed and conceived of this study, and with B.H. and E.R.H., performed the global comparative analysis. All authors except J.M., S.C., S.E., and J.J.R. contributed to the writing and/or editing of the article. The authors declare no potential financial or other interests that could be perceived to influence the outcomes of the research. No conflicts, informed consent, human or animal rights applicable. All authors declare agreement to authorship and submission of the manuscript for peer review.

References

[1] C.J. Unkefer, R.T. Sayre, J.K. Magnuson, D.B. Anderson, I. Baxter, I.K. Blaby, J.K. Brown, M. Carleton, R.A. Cattolico, T. Dale, T.P. Devarenne, C.M. Downes, S.K. Dutcher, D.T. Fox, U. Goodenough, J. Jaworski, J.E. Holladay, D.M. Kramer, A.T. Koppisch, M.S. Lipton, B.L. Marrone, M. McCormick, I. Molnár, J.B. Mott, K.L. Ogden, E.A. Panisko, M. Pellegrini, J. Polle, J.W. Richardson, M. Sabarsky, S.R. Starkenburg, G.D. Stormo, M. Teshima, S.N. Twary, P.J. Unkefer, J.S. Yuan, J.A. Olivares, Review of the algal biology program within the National Alliance for advanced biofuels and bioproducts, Algal Res. 22 (2017) 187–215, https://doi.org/10.1016/j.algal.2016.06.002.

[2] J.B. Shurin, M.D. Burkart, S.P. Mayfield, V.H. Smith, Recent progress and future challenges in algal biofuel production, F1000Research 5 (2016) 2434, https://doi.org/10.12688/f1000research.9217.1.

[3] S.S. Merchant, S.E. Prochnik, O. Vallon, E.H. Harris, J. Karpowicz, G.B. Witman, A. Terry, A. Salamov, L.K. Fritz-Laylin, L. Maréchal-Drouard, W.F. Marshall, L.-H. Qu, D.R. Nelson, A. Sanderfoot, M.H. Spalding, V.V. Kapitonov, Q. Ren, P. Cardol, H. Cerutti, G. Chanfreau, P. Ferris, H. Fukuzawa, D. González-Ballester, B. Mueller-Roeber, S. Rajamani, R.T. Sayre, I. Dubchak, D. Goodstein, L. Hornick, Y.W. Huang, Y. Luo, D. Martínez, W. Chi, A. Ngau, B. Otillar, A. Porter, L. Szajkowski, G. Werner, K. Zhou, V. Igor, D.S. Rokhsar, A.R. Grossman, The *Chlamydomonas* genome reveals the evolution of key animal and plant functions, Science 318 (2010) 245–250, https://doi.org/10.1126/science.1143609.

[4] S.J. Karpowicz, S.E. Prochnik, A.R. Grossman, S.S. Merchant, The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage, J. Biol. Chem. 286 (2011) 21427–21439, https://doi.org/10.1074/jbc.M111.233734.

[5] E.M. Trentacoste, R.P. Shrestha, S.R. Smith, C. Gle, A.C. Hartmann, M. Hildebrand, W.H. Gerwick, Metabolic engineering of lipid catabolism increases microalgal lipid accumulation without compromising growth, Proc. Natl. Acad. Sci. 110 (2013) 19748–19753, https://doi.org/10.1073/pnas.1309299110.

[6] R.A. Fisher, The Genetical Theory of Natural Selection, Oxford University Press, Oxford, 1930.

[7] H.J. Muller, The relation of recombination to mutational advance, Mutat. Res. 1 (1964) 2–9, https://doi.org/10.1016/0027-5107(64)90047-8.

[8] J. Bennnetzen, S. Hake, Handbook of Maize, Springer, New York, 2011, https://doi.org/10.1002/9783527674305.

[9] L.T. Morran, O.G. Schmidt, I.A. Gelarden, R.C. Parrish, C.M. Lively, Running with the red queen: host-parasite coevolution selects for biparental sex, Science 333 (2011) 216–218, https://doi.org/10.1126/science.1206360.

[10] J. Maynard Smith, The Evolution of Sex, Cambridge University Press, Cambridge, 1978.

[11] L. Trut, I. Oskina, A. Kharlamova, Animal evolution during domestication: the domesticated fox as a model, BioEssays 31 (2009) 349–360, https://doi.org/10.1002/bies.200800070.

[12] W.C. Ratcliff, M.D. Herron, K. Howell, J.T. Pentz, F. Rosenzweig, M. Travisano, Experimental evolution of an alternating uni- and multicellular life cycle in *Chlamydomonas reinhardtii*, Nat. Commun. 4 (2013) 2742, , https://doi.org/10.1038/ncomms3742.

[13] J.K. Conner, Artificial selection: a powerful tool for ecologists, Ecology 84 (2003) 1650–1660, https://doi.org/10.2307/3449986.

[14] C. Langdon, F. Evans, D. Jacobson, M. Blouin, Yields of cultured Pacific oysters *Crassostrea gigas* Thunberg improved after one generation of selection, Aquaculture 220 (2003) 227–244, https://doi.org/10.1016/S0044-8486(02)00621-X.

[15] C. Darwin, On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, John Murray, London, 1859.

[16] F. Maumus, A.E. Allen, C. Mhiri, H. Hu, K. Jabbari, A. Vardi, M.-A. Grandbastien, C. Bowler, Potential impact of stress activated retrotransposons on genome evolution in a marine diatom, BMC Genomics 10 (2009) 1–19, https://doi.org/10.1186/1471-2164-10-624.

[17] H. Cerutti, F. Ibrahim, Turnover of mature miRNAs and siRNAs in plants and algae, in: H. Großhans (Ed.), Regul. MicroRNAs. Adv. Exp. Med. Biol, Vol. 700 Springer, New York, 2010, pp. 124–139.

[18] E.J. Kim, X. Ma, H. Cerutti, Gene silencing in microalgae: mechanisms and biological roles, Bioresour. Technol. 184 (2015) 23–32, https://doi.org/10.1016/j.biortech.2014.10.119.

[19] J. Casas-Mollano, E. Zacarias, X. Ma, E.J. Kim, H. Cerutti, RNA-mediated silencing in eukaryotes: evolution of protein components and biological roles, in:

G. Hernandez, R. Jagus (Eds.), Evol. Protein Syntheis Mach. Its Regul, Springer, Cham, 2016, pp. 513–529.
[20] E.E. Jarvis, T.G. Dunahay, L.M. Brown, DNA nucleoside composition and methylation in several species of microalgae, J. Phycol. 28 (1992) 356–362, https://doi.org/10.1111/j.0022-3646.1992.00356.x.
[21] P. Babinger, I. Kobl, W. Mages, R. Schmitt, A link between DNA methylation and epigenetic silencing in transgenic *Volvox carteri*, Nucleic Acids Res. 29 (2001) 1261–1271, https://doi.org/10.1093/nar/29.6.1261.
[22] A. Veluchamy, A. Rastogi, X. Lin, B. Lombard, O. Murik, Y. Thomas, F. Dingli, M. Rivarola, S. Ott, X. Liu, Y. Sun, P.D. Rabinowicz, J. McCarthy, A.E. Allen, D. Loew, C. Bowler, L. Tirichine, An integrative analysis of post-translational histone modifications in the marine diatom *Phaeodactylum tricornutum*, Genome Biol. 16 (2015) 1–18, https://doi.org/10.1186/s13059-015-0671-8.
[23] J. Staunton, K.J. Weissman, Polyketide biosynthesis: a millennium review, Nat. Prod. Rep. 18 (2001) 380–416, https://doi.org/10.1039/a909079g.
[24] B. Shen, Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms, Curr. Opin. Chem. Biol. 7 (2003) 285–295, https://doi.org/10.1016/S1367-5931(03)00020-6.
[25] U. John, B. Beszteri, E. Derelle, Y. Van de Peer, B. Read, H. Moreau, A. Cembella, Novel insights into evolution of protistan polyketide synthases through phylogenomic analysis, Protist 159 (2008) 21–30, https://doi.org/10.1016/j.protis.2007.08.001.
[26] E. Shelest, N. Heimerl, M. Fichtner, S. Sasso, Multimodular type I polyketide synthases in algae evolve by module duplications and displacement of AT domains in trans, BMC Genomics 16 (2015) 1–15, https://doi.org/10.1186/s12864-015-2222-9.
[27] F. Gross, N. Luniak, O. Perlova, N. Gaitatzis, H. Jenke-Kodama, K. Gerth, D. Gottschalk, E. Dittmann, R. Müller, Bacterial type III polyketide synthases: phylogenetic analysis and potential for the production of novel secondary metabolites by heterologous expression in pseudomonads, Arch. Microbiol. 185 (2006) 28–38, https://doi.org/10.1007/s00203-005-0059-3.
[28] B.S. Moore, J.N. Hopke, Discovery of a new bacterial polyketide biosynthetic pathway, Chembiochem 2 (2001) 35–38, https://doi.org/10.1002/1439-7633(20010105)2:1<35::AID-CBIC35>3.0.CO;2-1.
[29] B. Palenik, J. Grimwood, A. Aerts, P. Rouze, A. Salamov, N. Putnam, C. Dupont, R. Jorgensen, E. Derelle, S. Rombauts, K. Zhou, R. Otillar, S.S. Merchant, S. Podell, T. Gaasterland, C. Napoli, K. Gendler, A. Manuell, V. Tai, O. Vallon, G. Piganeau, S. Jancek, M. Heijde, K. Jabbari, C. Bowler, M. Lohr, S. Robbens, G. Werner, I. Dubchak, G.J. Pazour, Q. Ren, I. Paulsen, C. Delwiche, J. Schmutz, D. Rokhsar, Y. Van de Peer, H. Moreau, I.V. Grigoriev, The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation, Proc. Natl. Acad. Sci. 104 (2007) 7705–7710, https://doi.org/10.1073/pnas.0611046104.
[30] P.J. Keeling, C.H. Slamovits, Causes and effects of nuclear genome reduction, Curr. Opin. Genet. Dev. 15 (2005) 601–608, https://doi.org/10.1016/j.gde.2005.09.003.
[31] P.J. Lammers, M. Huesemann, W. Boeing, D.B. Anderson, R.G. Arnold, X. Bai, M. Bhole, Y. Brhanavan, L. Brown, J. Brown, J.K. Brown, S. Chisholm, C. Meghan Downes, S. Fulbright, Y. Ge, J.E. Holladay, B. Ketheesan, A. Khopkar, A. Koushik, P. Laur, B.L. Marrone, J.B. Mott, N. Nirmalakhandan, K.L. Ogden, R.L. Parsons, J. Polle, R.D. Ryan, T. Samocha, R.T. Sayre, M. Seger, T. Selvaratnam, R. Sui, A. Thomasson, A. Unc, W. Van Voorhies, P. Waller, Y. Yao, J.A. Olivares, Review of the cultivation program within the National Alliance for advanced biofuels and bioproducts, Algal Res. 22 (2017) 166–186, https://doi.org/10.1016/j.algal.2016.11.021.
[32] NAABB, National Alliance for Advanced Biofuels and Bio-products (NAABB) Synopsis, http://www.energy.gov/eere/bioenergy/downloads/national-alli-ance-advanced-biofuels-and-bioproducts-synopsis-naabb-final, (2014).
[33] P. Neofotis, A. Huang, K. Sury, W. Chang, F. Joseph, A. Gabr, S. Twary, W. Qiu, O. Holguin, J.E.W. Polle, Characterization and classification of highly productive microalgae strains discovered for biofuel and bioproduct generation, Algal Res. 15 (2016) 164–178, https://doi.org/10.1016/j.algal.2016.01.007.
[34] M. Huesemann, A. Chavis, S. Edmundson, D. Rye, S. Hobbs, N. Sun, M. Wigmosta, Climate-simulated raceway pond culturing: quantifying the maximum achievable annual biomass productivity of Chlorella sorokiniana in the contiguous USA, J. Appl. Phycol. 30 (2018) 287–298, https://doi.org/10.1007/s10811-017-1256-6.
[35] M. Huesemann, T. Dale, A. Chavis, B. Crowe, S. Twary, A. Barry, D. Valentine, R. Yoshida, M. Wigmosta, V. Cullinan, Simulation of outdoor pond cultures using indoor LED-lighted and temperature-controlled raceway ponds and Phenometrics photobioreactors, Algal Res. 21 (2017) 178–190, https://doi.org/10.1016/j.algal.2016.11.016.
[36] M. Huesemann, B. Crowe, P. Waller, A. Chavis, S. Hobbs, S. Edmundson, M. Wigmosta, A validated model to predict microalgae growth in outdoor pond cultures subjected to fluctuating light intensities and water temperatures, Algal Res. 13 (2016) 195–206, https://doi.org/10.1016/j.algal.2015.11.008.
[37] A.N. Barry, S.R. Starkenburg, R.T. Sayre, Strategies for optimizing algal biology for enhanced biomass production, Front. Energy Res. 3 (2015) 1–5, https://doi.org/10.3389/fenrg.2015.00001.
[38] J.N. Rosenberg, N. Kobayashi, A. Barnes, E.A. Noel, M.J. Betenbaugh, G.A. Oyler, Comparative analyses of three *Chlorella* species in response to light and sugar reveal distinctive lipid accumulation patterns in the microalga *C. sorokiniana*, PLoS One 9 (2014), https://doi.org/10.1371/journal.pone.0092460.
[39] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, S. Turner, Real-time DNA sequencing from single polymerase molecules, Science 323 (2009) 133–138, https://doi.org/10.1126/science.1162986.
[40] C.S. Chin, D.H. Alexander, P. Marks, A.A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E.E. Eichler, S.W. Turner, J. Korlach, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, Nat. Methods 10 (2013) 563–569, https://doi.org/10.1038/nmeth.2474.
[41] B. Ewing, L. Hillier, M.C. Wendl, P. Green, Base-calling of automated sequencer traces using Phred I. accuracy assessment, Genome Res. 8 (1998) 175–185, https://doi.org/10.1101/gr.8.3.175.
[42] B. Ewing, P. Green, Base-calling of automated sequencer traces using Phred II. Error probabilities, Genome Res. 8 (1998) 186–194, https://doi.org/10.1101/gr.8.3.175.
[43] D. Gordon, C. Abajian, P. Green, Consed: a graphical tool for sequence finishing, Genome Res. 8 (1998) 195–202, https://doi.org/10.1101/gr.8.3.195.
[44] OpGen, Isolating High Molecular Weight DNA for Analysis of Animal and Plant Genome Samples, (2013).
[45] S. Bennett, Solexa Ltd, Pharmacogenomics 5 (2004) 433–438, https://doi.org/10.1517/14622416.5.4.433.
[46] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs, Genome Res. 18 (2008) 821–829, https://doi.org/10.1101/gr.074492.107.
[47] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–360, https://doi.org/10.1038/nmeth.1923.
[48] P.E. Li, C.C. Lo, J.J. Anderson, K.W. Davenport, K.A. Bishop-Lilly, Y. Xu, S. Ahmed, S. Feng, V.P. Mokashi, P.S.G. Chain, Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform, Nucleic Acids Res. 45 (2017) 67–80, https://doi.org/10.1093/nar/gkw1027.
[49] C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects, BMC Bioinformatics. 12 (2011) 491, https://doi.org/10.1186/1471-2105-12-491.
[50] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. Leduc, N. Friedman, A. Regev, *De novo* transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis, Nat. Protoc. 8 (2013) 1494–1512, https://doi.org/10.1038/nprot.2013.084.
[51] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (2011) 644–652, https://doi.org/10.1038/nbt.1883.
[52] P. Jones, D. Binns, H.Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A.F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.Y. Yong, R. Lopez, S. Hunter, InterProScan 5: genome-scale protein function classification, Bioinformatics 30 (2014) 1236–1240, https://doi.org/10.1093/bioinformatics/btu031.
[53] K. Bradnam, Assemblaton Software, Genome Center, UC Davis, 2011, http://korflab.ucdavis.edu/datasets/Assemblathon/Assemblathon2/Basic_metrics/assemblathon_stats.pl.
[54] B. Hall, T. Derego, S. Geib, GAG: The Genome Annotation Generator, (2014).
[55] F. Ronquist, M. Teslenko, P. Van Der Mark, D.L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M.A. Suchard, J.P. Huelsenbeck, MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space, Syst. Biol. 61 (2012) 539–542, https://doi.org/10.1093/sysbio/sys029.
[56] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313, https://doi.org/10.1093/bioinformatics/btu033.
[57] L. Sun, L. Fang, Z. Zhang, X. Chang, D. Penny, B. Zhong, Chloroplast phylogenomic inference of green algae relationships, Sci. Rep. 6 (2016) 1–6, https://doi.org/10.1038/srep20528.
[58] C.W. Birky, Relaxed cellular controls and organelle heredity, Science 222 (1983) 468–475, https://doi.org/10.1126/science.6353578.
[59] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797, https://doi.org/10.1093/nar/gkh340.
[60] R. Lanfear, P.B. Frandsen, A.M. Wright, T. Senfeld, B. Calcott, PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses, Mol. Biol. Evol. 34 (2016) 772–773, https://doi.org/10.1093/molbev/msw260.
[61] S. Kurtz, A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S.L. Salzberg, Versatile and open software for comparing large genomes, Genome Biol. 5 (2004) R12, https://doi.org/10.1186/gb-2004-5-2-r12.
[62] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2.
[63] J. Lees, C. Yeats, J. Perkins, I. Sillitoe, R. Rentzsch, B.H. Dessailly, C. Orengo, Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis, Nucleic Acids Res. 40 (2012) D465–D471, https://doi.org/10.1093/nar/gkr1181.
[64] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, P.D. Thomas, PANTHER version 11: expanded annotation data from gene ontology and Reactome pathways, and data analysis tool enhancements, Nucleic Acids Res. 45 (2017) D183–D189, https://doi.org/10.1093/nar/gkw1138.
[65] R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L.L. Sonnhammer, J. Tate, M. Punta, Pfam:

the protein families database, Nucleic Acids Res. 42 (2014) D222–D230, https://doi.org/10.1093/nar/gkt1223.
[66] M. Kanehisa, Toward pathway engineering: a new database of genetic and molecular pathways, Sci. Technol. Japan. 59 (1996) 34–38.
[67] B.T. Hovde, C.R. Deodato, H.M. Hunsperger, S.A. Ryken, W. Yost, R.K. Jha, J. Patterson, R.J. Monnat, S.B. Barlow, S.R. Starkenburg, R.A. Cattolico, Genome sequence and transcriptome analyses of *Chrysochromulina tobin*: metabolic tools for enhanced algal fitness in the prominent order Prymnesiales (Haptophyceae), PLoS Genet. 11 (2015) e1005469, , https://doi.org/10.1371/journal.pgen.1005469.
[68] M. Stanke, B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints, Nucleic Acids Res. 33 (2005) W465–W467, https://doi.org/10.1093/nar/gki458.
[69] S. Eddy, Profile hidden Markov models, Bioinformatics 14 (1998) 755–763 (doi:btb114 [pii]).
[70] H. Riveros-Rosas, A. Julián-Sánchez, R. Villalobos-Molina, J.P. Pardo, E. Piña, Diversity, taxonomy and evolution of medium-chain dehydrogenase/reductase superfamily, Eur. J. Biochem. 270 (2003) 3309–3334, https://doi.org/10.1046/j.1432-1033.2003.03704.x.
[71] M.S. Roth, S.J. Cokus, S.D. Gallaher, A. Walter, D. Lopez, E. Erickson, B. Endelman, D. Westcott, C.A. Larabell, S.S. Merchant, M. Pellegrini, K.K. Niyogi, Chromosome-level genome assembly and transcriptome of the green alga *Chromochloris zofingiensis* illuminates astaxanthin production, Proc. Natl. Acad. Sci. 114 (2017) E4296–E4305, https://doi.org/10.1073/pnas.1619928114.
[72] E. Corteggiani Carpinelli, A. Telatin, N. Vitulo, C. Forcato, M. D'Angelo, R. Schiavon, A. Vezzi, G.M. Giacometti, T. Morosinotto, G. Valle, Chromosome scale genome assembly and transcriptome profiling of *Nannochloropsis gaditana* in nitrogen depletion, Mol. Plant 7 (2014) 323–335, https://doi.org/10.1093/mp/sst120.
[73] H. Nozaki, H. Takano, O. Misumi, K. Terasawa, M. Matsuzaki, S. Maruyama, K. Nishida, F. Yagisawa, Y. Yoshida, T. Fujiwara, S. Takio, K. Tamura, S.J. Chung, S. Nakamura, H. Kuroiwa, K. Tanaka, N. Sato, T. Kuroiwa, A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*, BMC Biol. 5 (2007) 28, https://doi.org/10.1186/1741-7007-5-28.
[74] A.S. Schwartz, R. Brown, I. Ajjawi, J. McCarren, S. Atilla, N. Bauman, T.H. Richardson, Complete genome sequence of the model oleaginous alga *Nannochloropsis gaditana* CCMP1894, Genome Annoucements 6 (2018) (e01448-17).
[75] 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, Plant J. 93 (2018) 566–586, https://doi.org/10.1111/tpj.13789.
[76] 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, https://doi.org/10.1093/bioinformatics/btv351.
[77] G. Blanc, G. Duncan, I. Agarkova, M. Borodovsky, J. Gurnon, A. Kuo, E. Lindquist, S. Lucas, J. Pangilinan, J. Polle, A. Salamov, A. Terry, T. Yamada, D.D. Dunigan, I.V. Grigoriev, J.-M. Claverie, J.L. Van Etten, The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex, Plant Cell 22 (2010) 2943–2955, https://doi.org/10.1105/tpc.110.076406.
[78] S. Götz, J.M. García-Gómez, J. Terol, T.D. Williams, S.H. Nagaraj, M.J. Nueda, M. Robles, M. Talón, J. Dopazo, A. Conesa, High-throughput functional annotation and data mining with the Blast2GO suite, Nucleic Acids Res. 36 (2008) 3420–3435, https://doi.org/10.1093/nar/gkn176.
[79] K. Fučíková, M. Pažoutová, F. Rindi, Meiotic genes and sexual reproduction in the green algal class Trebouxiophyceae (Chlorophyta), J. Phycol. 51 (2015) 419–430, https://doi.org/10.1111/jpy.12293.
[80] C. Burns, J.E. Stajich, A. Rechtsteiner, L. Casselton, S.E. Hanlon, S.K. Wilke, O.P. Savytskyy, A.C. Gathman, W.W. Lilly, J.D. Lieb, M.E. Zolan, P.J. Pukkila, Analysis of the basidiomycete *Coprinopsis cinerea* reveals conservation of the core meiotic expression program over half a billion years of evolution, PLoS Genet. 6 (2010), https://doi.org/10.1371/journal.pgen.1001135.
[81] M.O.P. Iyengar, T.V. Desikachary, Volvocales, Indian Council of Agricultural Research, New Delhi, 1981.
[82] H. Ettl, Chlrophyta I — Phytomonidia, in: H. Ettl, J. Gerloff, H. Heynig, D. Mollenhauer (Eds.), Susswaserflora von Metteleropa, Gustav Fischer, Stuttgart, 1983, pp. 1–807.
[83] E.H. Harris, The *Chlamydomonas* Sourcebook (Volume 1), Elsevier, San Diego, California, 2009.
[84] U. Goodenough, H. Lin, J.H. Lee, Sex determination in *Chlamydomonas*, Semin. Cell Dev. Biol. 18 (2007) 350–361, https://doi.org/10.1016/j.semcdb.2007.02.006.
[85] F.R. Cross, J.G. Umen, The *Chlamydomonas* cell cycle, Plant J. 82 (2015) 370–392, https://doi.org/10.1111/tpj.12795.
[86] S.E. Prochnik, J. Umen, A.M. Nedelcu, A. Hallmann, M. Miller, I. Nishii, P. Ferris, A. Kuo, T. Mitros, L.K. Fritz-Laylin, U. Hellsten, J. Chapman, O. Simakov, S.A. Rensing, A. Terry, J. Pangilinan, V. Kapitonov, J. Jurka, A. Salamov, H. Shapiro, J. Schmutz, J. Grimwood, E. Lindquist, S. Lucas, I.V. Grigoriev, R. Schmitt, D. Kirk, D.S. Rokhsar, Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*, Science 329 (2011) 223–226, https://doi.org/10.1126/science.1188800.Genomic.
[87] X. Cao, W. Aufsatz, D. Zilberman, M.F. Mette, M.S. Huang, M. Matzke, S.E. Jacobsen, Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation, Curr. Biol. 13 (2003) 2212–2217, https://doi.org/10.1016/j.cub.2003.11.052.
[88] R. Nishiyama, Y. Wada, M. Mibu, Y. Yamaguchi, K. Shimogawara, H. Sano, Role of a nonselective *de novo* DNA methyltransferase in maternal inheritance of chloroplast genes in the green alga, *Chlamydomonas reinhardtii*, Genetics 168 (2004) 809–816, https://doi.org/10.1534/genetics.104.030775.
[89] M.A. Matzke, R.A. Mosher, RNA-directed DNA methylation: an epigenetic pathway of increasing complexity, Nat. Rev. Genet. 15 (2014) 394–408, https://doi.org/10.1038/nrg3683.
[90] H. Wu, Y. Zhang, Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation, Genes Dev. 25 (2011) 2436–2452, https://doi.org/10.1101/gad.179184.111.genomes.
[91] R.M. Erdmann, A.L. Souza, C.B. Clish, M. Gehring, 5-Hydroxymethylcytosine is not present in appreciable quantities in *Arabidopsis* DNA, G3, 5 (2015) 1–8, https://doi.org/10.1534/g3.114.014670.
[92] H. Jang, H. Shin, B.F. Eichman, J.H. Huh, Excision of 5-hydroxymetylcytosine by DEMETER family DNA glycosylases, Biochem. Biophys. Res. Commun. 446 (2014) 167–172, https://doi.org/10.1016/j.bbrc.2014.03.060.Excision.