Algal Research 58 (2021) 102380

Contents lists available at ScienceDirect

Algal Research

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

Comparative and phylogenomic analysis of nuclear and organelle genes in cryptic *Coelastrella vacuolata* MACC-549 green algae

Prateek Shetty [a], Attila Farkas [a], Bernadett Pap [a], Bettina Hupp [a], Vince Ördög [b,c], Tibor Bíró [d], Torda Varga [e], Gergely Maróti [a,d,*]

[a] Institute of Plant Biology, Biological Research Centre, Eötvös Loránd Research Network (ELKH), Szeged 6726, Hungary
[b] Department of Plant Sciences, Faculty of Agricultural and Food Sciences, Széchenyi István University, Mosonmagyaróvár 9200, Hungary
[c] Research Centre for Plant Growth and Development, School of Life Sciences, University of KwaZulu-Natal Pietermaritzburg, Scottsville 3209, South Africa
[d] Faculty of Water Sciences, National University of Public Service, Baja 6500, Hungary
[e] Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Center, Eötvös Loránd Research Network (ELKH), Szeged 6726, Hungary

ARTICLE INFO

Keywords:
Green algae
Phylogeny
Whole genome
Chloroplast genes
Taxonomy
Coelastrella

ABSTRACT

The nuclear, chloroplast and mitochondrial genomes of a unicellular green algal species of the *Coelastrella* genus was sequenced, assembled and annotated. The strain was previously classified as *Chlamydomonas* sp. MACC-549 based on morphology and partial 18S rDNA analysis. However, the proposed multi-loci phylogenomic approach described in this paper placed this strain within the *Coelastrella* genus, therefore it was re-named to *Coelastrella vacuolata* MACC-549. The strain was selected for de novo sequencing based on its potential value in biohydrogen production as revealed in earlier studies. This is the first thorough report and characterization for green algae from the *Coelastrella* genus. The whole genome annotation of *Coelastrella vacuolata* MACC-549 (including nuclear, chloroplast and mitochondrial genomes) shed light on interesting metabolic and sexual breeding features of this algae and served as a basis to taxonomically classify this strain.

1. Introduction

The exploitability of green algae in diverse fields has led to an increase in molecular studies of these unicellular eukaryotes. Algal biomass serves as a sustainable source of bioenergy and biofuels and algal products find extensive use in agriculture, food and pharmaceutical industry and wastewater treatment [41,43,47]. The metabolic versatility of unicellular green algae makes them attractive eukaryotic model organisms in several fields of basic and applied science, including the research for clean, green energy carriers [17,47,60].

Previously, we investigated the hydrogen production capability of several eukaryotic green algae maintained at Mosonmagyaróvár Algal Culture Collection (MACC) [33,34]. The collection center has catalogued more than 600 eukaryotic green algae and close to 400 prokaryotic cyanobacterial strains [41]. Initial taxonomic classification of the strains was carried out by morphological characterization using light microscopy and partial 18S rDNA sequence homology [28]. *Chlamydomonas* sp. MACC-549 exhibited highly interesting mixotrophic hydrogen evolution characteristics when co-cultivated with various native and non-native bacterial partners [33,34]. The species was shown to grow exceedingly well when associated with different bacterial symbionts. This symbiont-specific association increased algal biomass, leading to a remarkable surge in photolytic hydrogen production [33]. The potential exploitability of this strain in green energy generation justified the need for detailed genome-level characterization. Whole genome sequencing based analysis as presented in this paper placed this species as a member of the genus *Coelastrella*.

The genus of *Coelastrella* green algae has had a turbulent phylogenetic history with several taxonomic revisions depending on availability of data. Initially, many species of the genus were placed with *Scotiellopsis* under Scotielloideae along with other Chlorococcales [25,30]. The development of NGS based technologies led to a substantial increase of algal molecular data, providing genomic evidence to reclassify members of this species by utilizing phylogenomics approaches. Based on 18S and ITS2 marker sequences, the *Scotiellopsis* genus was moved into the genus of Scenedesmoideae [20,21,22]. More recent studies have moved some members into a separate subfamily of Coelastroideae based on ITS2 sequence-structure studies [23]. Today, members of this genus are divided into two main groups of "core" *Coelastrella* and *Coelastrella* sensu lato; the latter group consists of diverse *Coelastrella* species and members

* Corresponding author at: Institute of Plant Biology, Biological Research Centre, Eötvös Loránd Research Network (ELKH), Szeged 6726, Hungary.
E-mail address: maroti.gergely@brc.hu (G. Maróti).

https://doi.org/10.1016/j.algal.2021.102380
Received 23 December 2020; Received in revised form 27 April 2021; Accepted 27 May 2021
Available online 30 June 2021
2211-9264/© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

of a distinct genus of *Asterarcys* [57]. Diverse specimens and genomic data are essential to further probe the relationship between *Asterarcys* sp. and members of Coelastroideae.

This paper reveals the first annotated draft nuclear genome of a *Coelastrella* algae along with complete chloroplast and mitochondrial genomes. Morphological and phylogenomic analysis with multi-gene alignment using 10 different sequences was used to identify *C. vacuolata* MACC-549. This study also identifies industrially relevant genes involved in gametogenesis, hydrogen production and carotenoid biosynthesis.

2. Materials and methods

2.1. Morphological and growth curve studies

C. vacuolata MACC-549 cells were cultured for a period of 4 days in TAP medium (TRIS-acetate-phosphate) at 25 °C and initial OD_{750} standardized to 0.7, shaken at 200 rpm and incubated under 50 μmol m^{-2} s^{-1} light intensity. The algae cells were observed using a Zeiss observer Z1 optical and a scanning electron microscope (JEOL JSM-7100F/LV). 50 cells per algal strain were measured to calculate average area and diameter data from optical microscopy images.

Growth curve studies were carried out in 24-well plates. Two *Chlamydomonas* and one *Chlorella* strains were selected from the MACC and cultivated with *C. vacuolata* MACC-549 strain in TAP for a period of 7 days in 2 replicates. The growth of the strains was followed using a HIDEX Sense microplate reader. On the first day, all cultures were standardized to $OD_{750} = 0.2$ measured in a plate reader. Plates were measured in every 12 h for the first 48 h and once every day following that. Average OD_{750}values were plotted and visualized with Rstudio [44].

2.2. Algal DNA purification and amplification of 18S rDNA gene

2 ml alga cells were collected by centrifugation and the pellet was homogenized in 500 μl TEN buffer (10 mM Tris-HCl (pH 8.0); 10 mM EDTA (pH 8.0); 150 mM NaCl). 1 μl proteinase K was added to the suspension and was kept at 37 °C for 15 min. Then 150 μl 5 M NaCl and 80 μl 10% CTAB/700 mM NaCl were added followed incubation at 65 °C for 20 min. Extraction was carried out with phenol:chloroform:isoamylalcohol (25/24/1) and with chloroform:isoamylalcohol (24/1). Genomic DNA was precipitated with 0.7 volume isopropanol and then washed with ethanol. The DNA pellet was resuspended in Tris-EDTA (TE) buffer. Total genomic DNA was extracted and the 1.5-kb fragment of the 18S rDNA gene was amplified by PCR in a PCR cycle of initial denaturation of 1 min at 95 °C, followed by 35 cycles of 30 s at 95 °C, 30 s at 55 °C, 1 min at 72 °C and a final extension of 5 min at 72 °C with the oligonucleotide primers (UNI7F 5′-ACCTGGTTGATCCTGCCAG-3′, UNI1534R 5′-TGATCCTTCYGC AGGTTCAC-3′). Amplification product was purified with a NucleoSpin Gel and PCR cleanup kit (Macherey-Nagel).

2.3. Scanning electron microscopy

For electron microscopy, 7-day old cultures were spotted in a volume of 8 μl onto a silicon disc coated with 0.01% poly-L-lysine (Merck Millipore, Billerica, MA, USA). Cells were then fixed with 2.5% (*v*/v) glutaraldehyde and 0.05 M cacodylate buffer (pH 7.2 in PBS) overnight at 4 °C. Thereafter, the discs were gently washed three times with PBS and dehydrated with a graded ethanol series (30%, 50%, 70%, 80%, and 100% ethanol, each for overnight at 4 °C). The samples were washed one more time with 100% ethanol before being dried with a critical point dryer, followed by 12 nm gold coating (Quorum Technologies) and observed under a field emission scanning electron microscope (JEOL JSM-7100F/LV).

2.4. Whole genome sequencing

Coelastrella vacuolata MACC-549 was grown in TAP medium at 25 °C to OD_{750} of 0.7, shaken at 200 rpm and incubated under 50 μmol m^{-2} s^{-1} light intensity. Total DNA (nuclear, mitochondrial and chloroplast DNA) from the pelleted algae cells was isolated by the DNeasy Plant Mini Kit (Qiagen, Germany) according to the instructions of the manufacturer. Mate-paired in vitro fragment library was generated using Illumina Nextera Mate-Pair Kit (Cat.Num.: FC-132-1001) with insert sizes ranging between 7 and 11 kb. All quality measurements were performed on a TapeStation 2200 instrument (Agilent, USA). DNA sequencing was carried out on an Illumina MiSeq machine using MiSeq Reagent Kit v2 (500 cycles) sequencing chemistry resulting in 2 × 250 nt reads. 51,891,944 mate-paired reads were generated.

2.5. Transcriptome sequencing

The transcriptome of *C. vacuolata* MACC-549 was investigated using two different growth conditions. The eukaryotic algae were grown under axenic and bacterial-associated conditions, respectively. Total RNA was extracted using Trizol reagent (Invitrogen). RNA was cleaned using the RNEasy Mini Kit (Qiagen) and analyzed using Tapestation 2200 (Agilent). Paired-end libraries were prepared using the NEBNext Ultra II RNA Library Prep Kit for Illumina. An Illumina NextSeq instrument was used to generate 2 × 150 nt reads by using NextSeq 500/550 High Output Kit v2.5 (300 Cycles) for sequencing.

2.6. Sequence processing, genome assembly and annotation

The sequenced raw reads were processed with NxTrim [40] to produce true mate-paired reads and paired end reads. These processed reads were filtered to remove reads smaller than 35 bases. All the filtered reads were assembled with Spades (v 3.9.1) [3]. The quality of the assembly was validated with BUSCO (v4.0.6) [46] using Chlorophyta and Viridiplantae databases.

Annotation steps began with repeat identification. Repeats in the assembled scaffolds were identified using RepeatMasker [48], RepeatModeler [49], MITE-Hunter [19] and LTRHarvest [13]. Structural gene models were predicted using MAKER [5], SNAP [31] and Augustus [51] gene prediction tools. The transcriptome assembly along with a total of 339,990 algal proteins, downloaded from UniProt using "*taxonomy:"-Chlorophyta [3041]*"" tag for functional annotations. Two successive rounds of Maker annotation were carried out for structural annotation following repeat identification. The second round of annotation was carried out on models generated during the first round.

Organelle genomes were downloaded from NCBI and assembled scaffolds were blasted against this database to identify chloroplast and mitochondrial scaffolds. Identifying the scaffolds is an important first step as the organelle assembler needs a "seed" sequence to start the alignment with. Chloroplast genome was assembled using NOVOplasty (v3.8.2) [10]. Circularized genome was annotated on GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html) [53] along with reference chloroplasts from the order Sphaeropleales. Chloroplast genomes for other species such as *Monoraphidium neglectum* also had to be annotated on GeSeq as only the FASTA sequence was available without any annotation. Chloroplast genomes for all other Scenedesmaceae members were also annotated on GeSeq to identify inverted repeat regions. Reference algal chloroplast genomes were downloaded from NCBI using *"(chloroplast, complete genome) AND 'green algae'[porgn:_txid3041]*" tag. Specific genes encoding enzymes for carotenoid biosynthesis and hydrogen production were downloaded from UniProt. Additional mating locus genes for *Chlamydomonas reinhardtii* were downloaded from NCBI. Orthologues for all these genes were identified using OrthoFinder (v2.4.0)[14], with MUSCLE (v3.8.31) [11] as the alignment program.

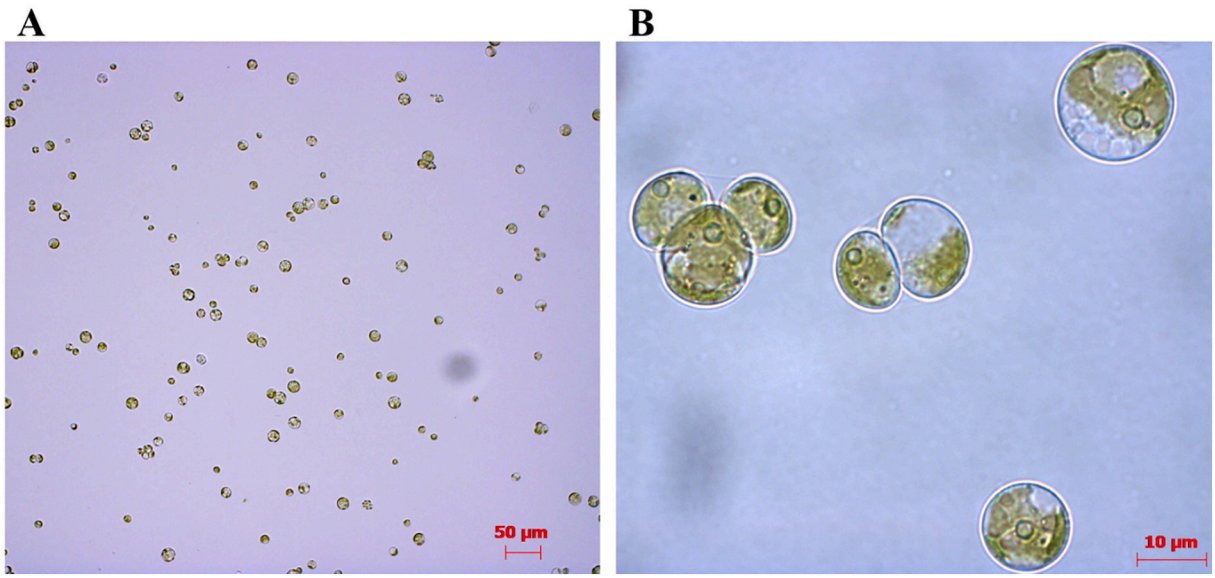


Fig. 1a. Optical microscopy images of *C. vacuolata* MACC-549. Panel A shows unicellular cell distribution. Panel B is a magnified image of the cells. Scale bars are indicated in the lower right corner.

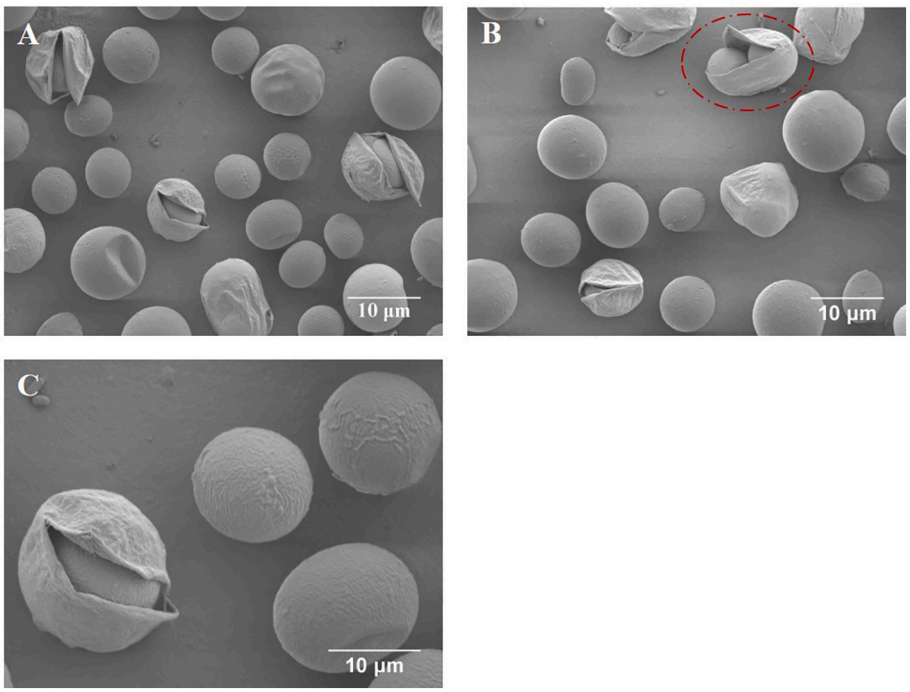


Fig. 2. Scanning electron microscopy images. Panel A represents SEM image of *C. vacuolata* MACC-549. Panel B shows 2 autospores, as they leave the autosporangium. Most cells show spherical or slightly ellipsoid shape.

2.7. Phylogenetic analysis

Phylogenetic analysis was carried out by using 3 nucleotide (18S, ITS2 and *tufA*) and 7 protein sequences (PetA, PsbD, PsaC, PsbB, RbcL, AtpA, and PsaB proteins coded on the chloroplast genome) (Supplementary Table 2) The bulk of all the sequences used were downloaded from a previous study [57]. Multiple alignments of separate loci were created by using the L-INS-i Iterative refinement method in MAFFT v.7.427 [27]. The multiple alignments were manually curated to correct false base mismatches introduced by the automatic multiple alignment algorithm. Highly variable and non-informative sites from the alignments were removed using GBlocks (v.91b) [6]. Minimum number of sequences for a conserved position and for a flank position were set to 50% of the number of sequences plus one, maximum number of contiguous non-conserved positions was set to 20, minimum length of a block was set to 2 and we allowed a gap position in all sequences.

After removing gaps, the 3 nucleotide sequences were concatenated into a single aligned sequence, referred to as the "Conventional loci alignment". We also assembled a 10 loci alignment by concatenating the 3 nucleotide and 7 protein sequences into one alignment. Overall, two phylogenetic trees were inferred by iqTree (v2.1.3) [38] using different substitution models for each locus [7] automatically identified by ModelFinder [26] applying 1000 bootstrap (BS) replicates and SH-aLRT test. Generated tree along with bootstrap values were visualized in iTOL

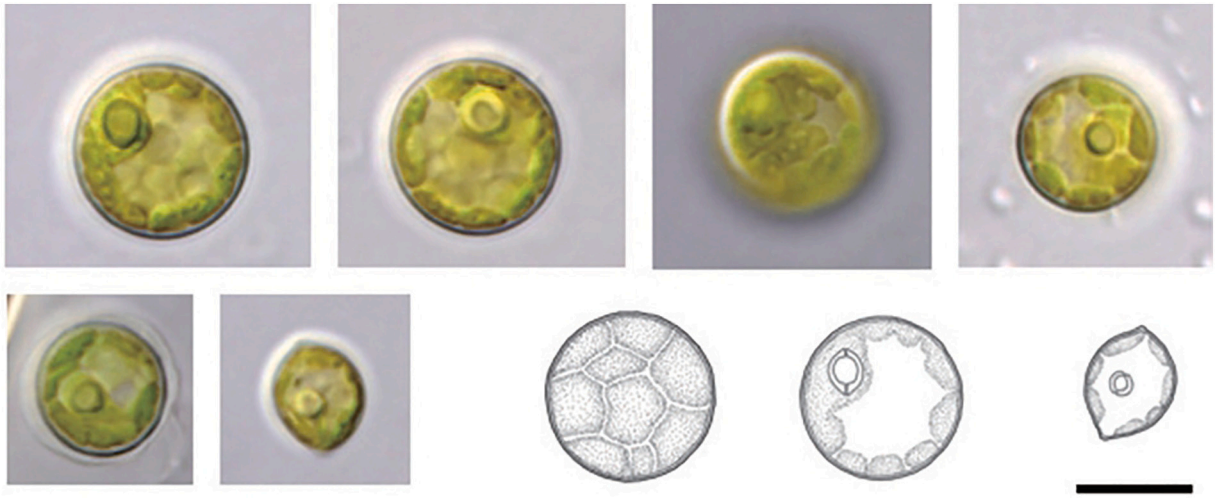


Fig. 1b. Optical microscopy images of *C. vacuolata*. Scale bar is 10 μm [50].

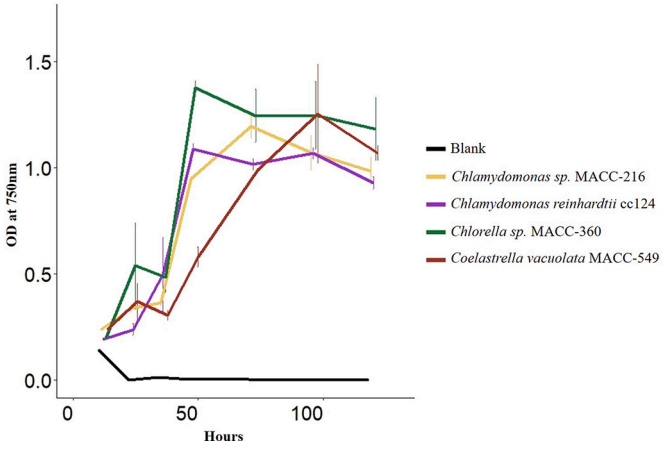


Fig. 3a. Growth curves of axenic *C. vacuolata* MACC-549 and *Chlamydomonas* and *Chlorella* species measured by optical density (750 nm).

[36].

3. Results

3.1. Morphological characterization

C. vacuolata MACC-549 was grown axenically in TAP medium and observed under both an optical microscope (Fig. 1a) and a scanning electron microscope (Fig. 2). A large degree of variation in cell sizes was observed, with a mean area of 196.21 ± 85.37 μm^2 and a mean diameter of 15.45 ± 3.36 μm. The average cell size was 2–5 times bigger than that of various *Chlamydomonas* and *Chlorella* cells (Fig. 3b). The cell walls were smooth and free of ridges, very similar to the cell surface morphology of known *Coelastrella vacuolata* isolates [Supplementary Table 1] (Fig. 1b). No flagella of any kind was detected on *C. vacuolata* MACC-549 cells. Asexual reproduction was observed with 2–8 cells autospores (Fig. 2). No sexual reproduction was observed.

3.2. Basic growth curve studies

Axenic *C. vacuolata* MACC-549 propagates well in TAP medium at 25 °C, shaken at 200 rpm and incubated under 50 μmol m^{-2} s^{-1} light intensity. However, the growth is rather slow compared to that of *Chlorella* and *Chlamydomonas* species under the same conditions (Fig. 3a). The peak OD_{750} for *C. vacuolata* MACC-549 propagated in axenic batch culture is on the 4th day of growth, while the investigated

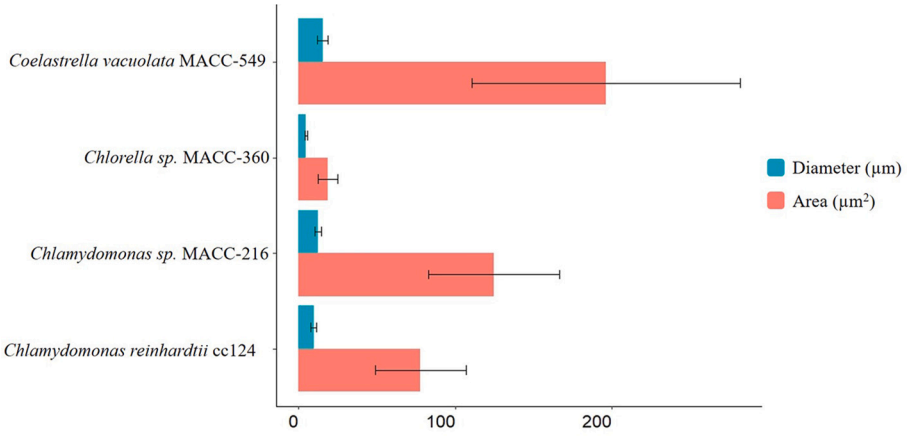


Fig. 3b. Average cell size of axenic *C. vacuolata* MACC-549 measured on optical microscopy images and compared against *Chlamydomonas* and *Chlorella* species.

Table 1
Read and genome assembly statistics.

Run1 raw reads	18,450,568
Run2 raw reads	33,441,376
Run1 True Mate paired reads	9,410,856
Run1 Paired end reads	6,308,476
Run2 True Mate paired reads	17,547,594
Run2 Paired end reads	11,178,240
Run1 Filtered True Mate paired reads	8,705,646
Run1 Filtered Paired end reads	6,304,362
Run2 Filtered True Mate paired reads	16,317,840
Run2 Filtered Paired end reads	11,174,650
Total scaffolds assembled	9136
Total number of filtered scaffolds	3744
Scaffolds without organelle genomes	3741
GC% of nuclear genome	51.30
N50 value of nuclear genome assembly	127,171
Total number of bases in nuclear genome	75,853,206

Chlorella and *Chlamydomonas* species reach their peak between the 2nd and 3rd days, respectively.

3.3. Whole genome assembly and annotation

The described nuclear, chloroplast and mitochondrial genomes of *C. vacuolata* MACC-549 represent the first annotated whole genome for *C. vacuolata*. About 42 million filtered reads were assembled with Spades (v 3.9.1), generating a total of 9136 scaffolds. These were filtered to retain scaffolds greater than 350 bp and to separate the organelle genomes (Table 1). The final nuclear genome had 3741 scaffolds comprising of 75.83 Mbp. The GC percentage of *C. vacuolata* MACC-549 nuclear genome was 51.3%, which is considerably lower than that of *Ch. reinhardtii* (62%), *Scenedesmus quadricauda* (63.2%) or *M. neglectum* (64.74%). All sequencing data, including raw reads and genome assemblies were uploaded to NCBI (BioProject: PRJNA629831, BioSample: SRR11665730, WGS: JABEVS000000000). The quality of the assembly was validated with BUSCO (v4.0.6) using Chlorophyta and Viridiplantae databases (Fig. 4). The assembled dataset comprised of 95.65% of BUSCO genes in the Chlorophyta database and 84% of BUSCO genes in the Viridiplantae dataset. Furthermore, only 2.69% and 8.7% of genes were missing from the Chlorophyta and Viridiplantae dataset, respectively. Considering that only a few genes were missing, it is likely that this assembled dataset was robust enough and achieved a good degree of completeness.

To separate the organelle genomes, the assembled scaffolds were compared against chloroplast and mitochondrial genomes downloaded from NCBI using specific search tags. Scaffolds belonging to chloroplast and mitochondrial genomes were removed prior to nuclear genome annotation which was performed by Maker (v3.01). A total of 517 repeat elements were identified and used as repeat libraries for Maker annotation. A total of 11,162 nuclear genes were predicted using two successive runs of Maker and predictive models based on SNAP and Augustus. The 11,162 nuclear genes of *C. vacuolata* MACC-549 were also compared against a complete set of algal proteins downloaded from UniProt. To improve annotation, a transcriptome assembly under two contrasting conditions; axenic and bacterial associated cultivation, was also used along with the protein data. These two contrasting conditions allowed to capture rare transcripts and genes that might be expressed under stressful, axenic conditions. A total of 10,285*C. vacuolata* MACC-549 proteins found matches against the UniProt database. We also looked at which organisms these results matched against and found that 8471 proteins (82.36%) matched those of *T. obliquus*. The bulk of these (4770) are annotated as "Uncharacterized protein" within *T. obliquus* and provided no functional information. After *T. obliquus,* the other organisms *Raphidocelis subcapitata* (4.01%) and *M. neglectum* (2.75%) were the top reference organisms (Supplementary Fig. 1).

Interpro was used to assign GO terms to all the genes. *Ch. reinhardtii* genome was also assigned GO terms during this run so that comparisons could be drawn between the two species (Fig. 6). A total of 8119 transcripts (48.07) were assigned to GO terms for *Ch. reinhardtii* while 5965 transcripts (53.5%) were assigned to GO terms for *C. vacuolata* MACC-549. GO terms for most categories were similar across both species, with higher proportion of GO terms identified in the genome of *C. vacuolata* MACC-549 compared to the genome of *Chlamydomonas reinhardtii*. Interestingly, genes with pigment production GO terms, specifically carotenoid biosynthesis were more abundant in *C. vacuolata* MACC-549. The whole genome and transcriptome reads were deposited in NCBI under the Bioproject accession; PRJNA629831. The assembled nuclear and organelle genomes were also deposited under the same Bioproject accession. The 1.5 kb long 18S rDNA gene was uploaded to Genbank (MW979811).

3.4. Chloroplast genome analysis

The chloroplast genome was fragmented into two scaffolds, while the mitochondrial genome of *C. vacuolata* MACC-549 was assembled into one single scaffold. In order to recover the complete chloroplast genome, a separate assembly was carried out using an organelle specific assembler NOVOplasty. This resulted in a completely circularized chloroplast genome of 204,380 bp (Fig. 5a, Table 2a). The mitochondrial genome was 34,063 bp in size (Fig. 5b, Table 2b). Both the chloroplast and mitochondrial genomes (uploaded along with nuclear genome; WGS: JABEVS000000000) were separately annotated using GeSeq. *C. vacuolata* MACC-549 has a longer inverted repeat section of 22,011 bp compared to chloroplast genomes of *C. saipanensis* and other members of Scenedesmaceae and is closer in size to the repeat region observed in *Ch. reinhardtii* (Table 2a).

A total of 87 CDS regions were annotated within the chloroplast genome of *C. vacuolata* MACC-549 (75 protein coding genes and 12 non-coding genes, Supplementary Table 3). The non-coding genes were made up of 7 free standing ORFs, 2 fragmented ORFs and 3 fragmented protein genes. Comparison of protein coding genes revealed that

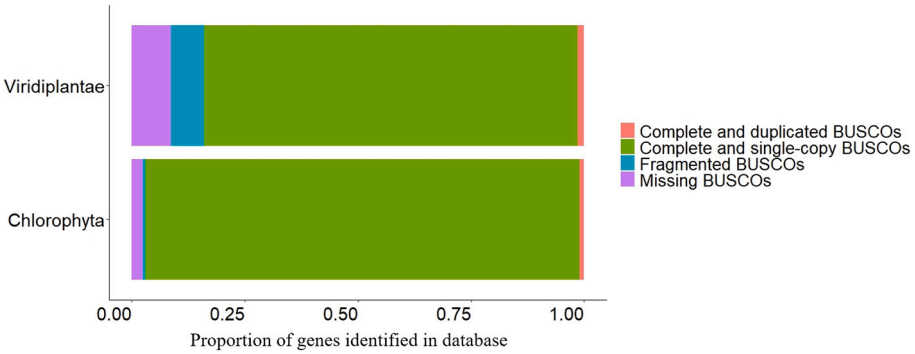


Fig. 4. BUSCO analysis of the assembled *C. vacuolata* MACC-549 scaffolds.

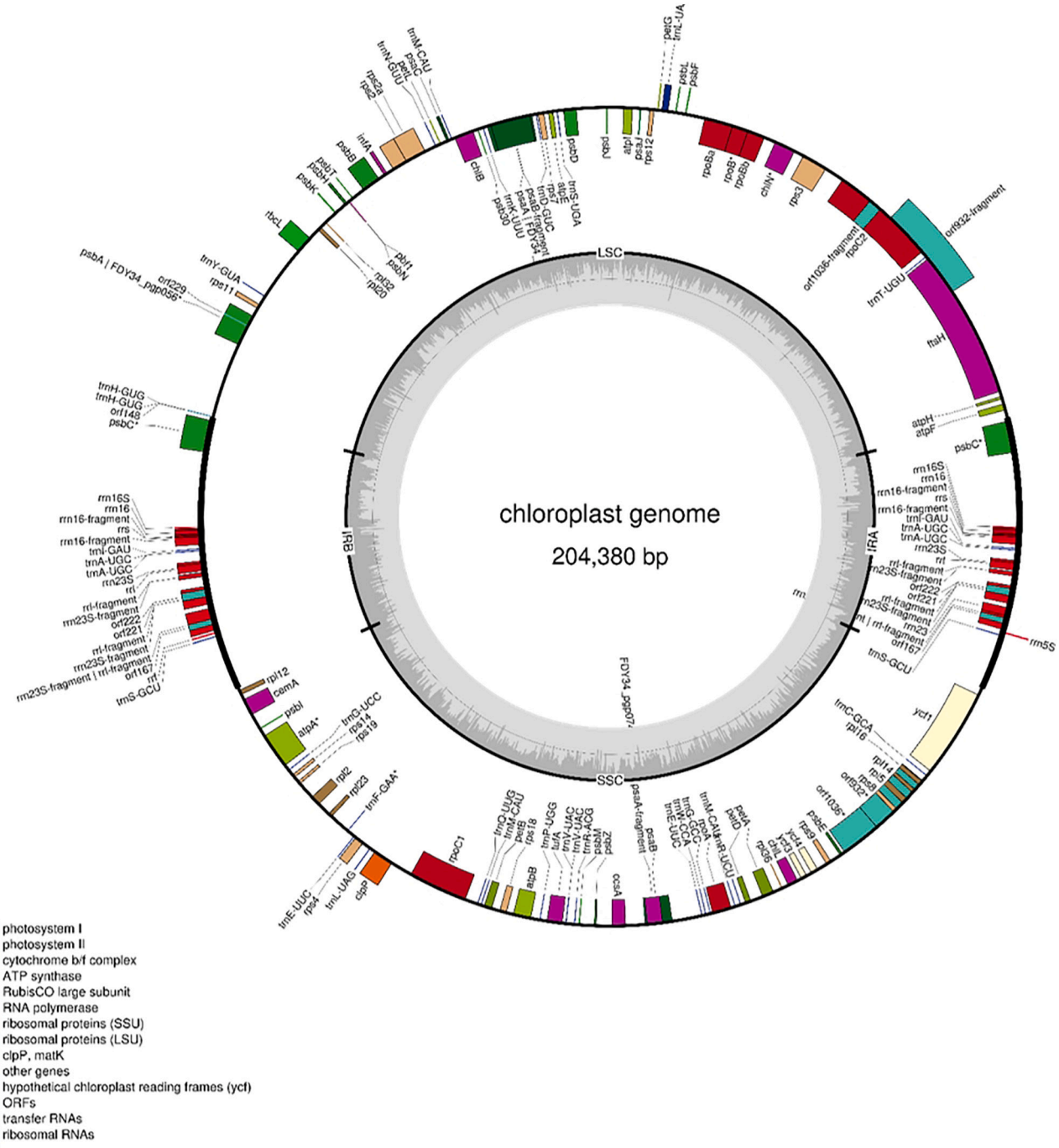


Fig. 5a. Chloroplast genome of *C. vacuolata* MACC-549. Genes are color-coded according to the functional categories listed in the index below the map. The GC content and inverted repeats (IR-A and IR-B) which separate the chloroplast genome into two single copy regions are indicated on the inner circle. The protein-coding genes as well as rRNA and tRNA genes are labelled inside and outside the circle. Genes on the inside of the map are transcribed in a clockwise direction; those on the outside of the map are transcribed counter-clockwise.

C. vacuolata MACC-549 and *C. saipanensis* shared all 75 protein coding genes between each other. However, the chloroplast genome of *C. saipanensis* had several intronic ORFs that were missing in *C. vacuolata* MACC-549. Free standing ORFs such as ORF1036, were shared among all Sphaeropleales. ORF1036 encodes a group II intron reverse transcriptase/maturase. *C. saipanensis* had several unique intronic ORFs that are completely absent in all other chloroplast genomes. Finally, *Ch. reinhardtii* had multiple missing protein coding genes (*rpl12*, *rpl32*, *ycf2*, *infA*) and non-coding genes (Fig. 7d).

3.5. Phylogenetic analysis

We inferred a phylogenetic tree by using conventional loci (the nuclear 18S and ITS2 genes as well as the chloroplast gene *tufA* with 1752, 1086 and 899 nucleotide characters, respectively) of 73 species. Seven additional chloroplast proteins (AtpA, PetA, PsaB, PsaC, PsbB, PsbD and RbcL with 501, 289, 735, 81, 508, 353 and 476 amino acid characters, respectively) downloaded from the chloroplast genome of 12 species were aligned, and a phylogenetic tree based on a combined data of the conventional loci and the protein sequences was built. The tree was rooted at *D. salina* CCAP19/18. The 3 nucleotide sequences were

Table 2a
Comparison of chloroplast genomes.

	C. vacuolata MACC-549	*Coelastrella saipanensis*	*Tetradesmus obliquus*	*Pectinodesmus pectinatus*	*Chlamydomonas reinhardtii*
Accession number		NC_042181.1	NC_008101.1	NC_036668.1	NC_005353.1
Chloroplast genome size	204,380	196,140	161,452	196,809	203,828
Size of Large single copy (LSC)	86,261	104,949	72,438	99,156	81,306
Size of Small single copy (SSC)	74,096	67,397	64,966	70,664	125,730
Size of inverter repeat	22,012	11,897	12,022	13,494	22,210
Inverted Repeat region (IR-A)	190,076–7707	184,244–196,140	149,430–161,452	183,316–196,809	34,238–56,448
Inverted Repeat region (IR-B)	93,968–115,979	104,950–116,846	72,440–84,462	99,157–112,650	137,755–159,965
No. of tRNA	33	31	32	30	28
No. of rRNA	14	12	14	11	10
No. of CDS	87	117	95	95	75

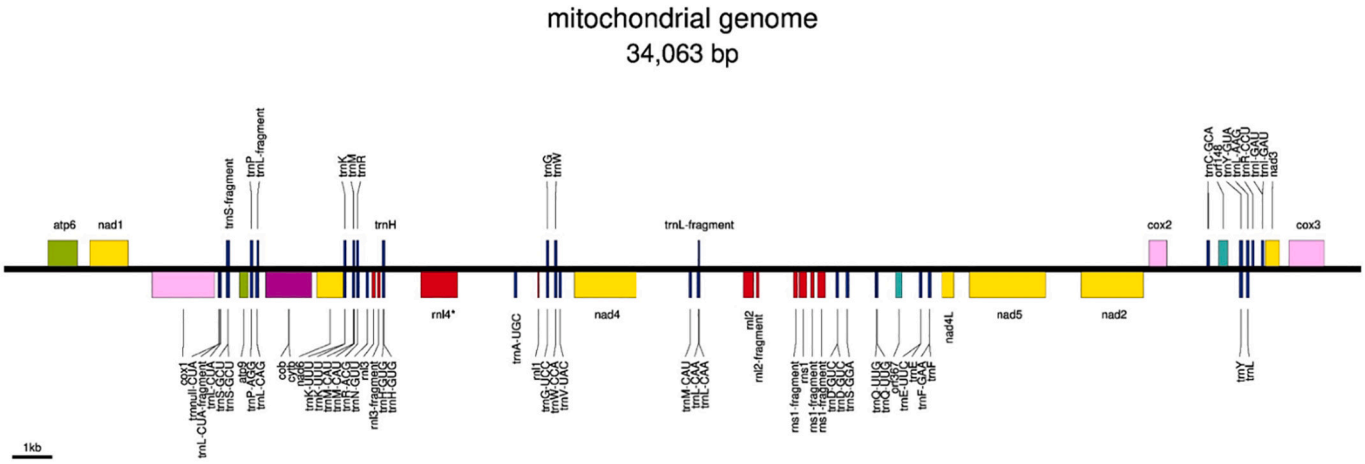


Fig. 5b. Mitochondrial genome of *C. vacuolata* MACC-549.

Table 2b
Comparison of mitochondrial genomes.

	C. vacuolata MACC-549	*Tetradesmus obliquus*	*Chlamydomonas reinhardtii*	*Pectinodesmus pectinatus*
Accession number		CM007918.1	NC_001638.1	NC_036659.1
Mitochondrial genome size	34,063 bp	41,704 bp	15,758 bp	32,195 bp
No. of tRNA	45	46	3	61
No. of rRNA	10	15	0	6
No. of CDS	16	33	6	16

selected as each of them has universal primers available and chloroplast genes were selected based on their high degree of conservation and because some of these have also been used in previous phylogenetic studies [16] (Supplementary Table 2).

Inclusion of protein sequences led to a general improvement in the bootstrap values and improved the phylogenetic resolutions in some cases (Fig. 7a). The phylogenetic tree inferred by using conventional loci (Fig. 7b) showed well-supported (BS > 80) algae relationships in most of the cases. *C. vacuolata* MACC-549 formed a monophyletic group with other *C. vacuolata* strains (BS = 99). *C. vacuolata* species form a monophyletic group with *C. tenuitheca* (BS = 99), which together is the sister group of all other *Coelastrella* sensu *lato* species (BS = 100). The position of *Asterarcys* species among *Coelastrella* lineages is uncertain (BS = 15) based on the three conventional loci, but the support of its position was highly improved after including the selected chloroplast protein sequences into our analysis (BS = 67). Overall, the combined conventional loci can resolve Sphaeropleales phylogeny, but in ambiguous cases, additional chloroplast proteins clearly helped in clarifying algal relationships. The strength and significance of this support will further increase with new chloroplast genomes available in the future.

3.6. Mating type locus genes

Ch. reinhardtii has a plus and a minus mating type locus, and these genes control sexual reproduction in *Ch. reinhardtii*. The proteins from both these types were compared against MAKER annotated *C. vacuolata* MACC-549 proteins to identify genes involved in sexual breeding. The *Ch. reinhardtii* minus type contains a total of 41 genes related to sexual breeding, while the plus type has only 40 genes. *C. vacuolata* MACC-549 has 30 orthologous genes present on the minus type and 28 orthologous genes present on the plus type, while *T. obliquus* had 29 orthologous genes on the minus strand and 29 orthologous genes present on the plus strand (Fig. 8).

3.7. Analysis of industrially relevant genes

Carotenoid biosynthesis proteins were downloaded from UniProt and compared against Maker annotated proteins of *C. vacuolata* MACC-549. Enzymes involved in the biosynthesis of phytoene, lycopene, β-carotene, zeaxantin, astaxanthin and α-carotene were identified in the genome and some enzymes in the pathway were present as multiple copies (Fig. 9). Genes coding for the Fe-only hydrogenase (HydA) along with genes encoding the accessory hydrogenase proteins HydG and

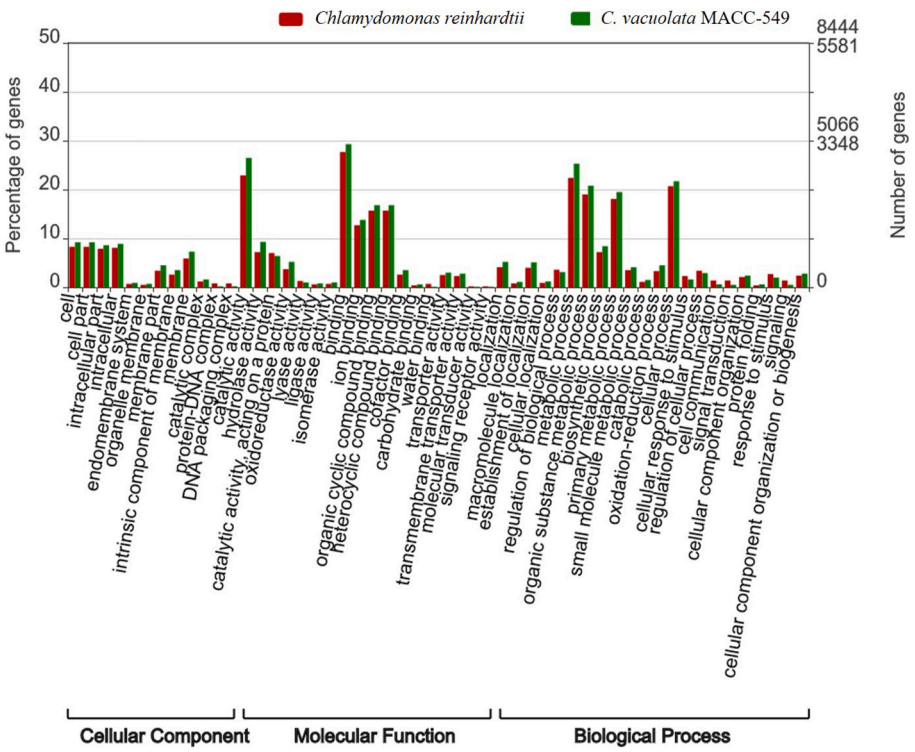


Fig. 6. GO annotation comparison between *Ch. reinhardtii* and *C. vacuolata* MACC-549.

HydEF were identified in *C. vacuolata* MACC-549 (Table 3). All genes had full length matches against either *Chlorella fusca* or *Chlorella variabilis* hydrogenase genes. All hydrogenase genes were present as single copies.

4. Discussion

4.1. Morphology and genome of C. vacuolata MACC-549 green algae

Optical microscopy images showed that the fresh *C. vacuolata* MACC-549 cells were mostly unicellular coccoid-shaped cells, with a low rate of aggregations. Electron microscopy images showed that all cells had a smooth cell wall, devoid of ridges. The only members of the genus *Coelastrella* to have no ridges on their cell walls are *C. vacuolata* (Fig. 1b). However, members of *Coelastrella* genus exhibit a great degree of variation in cell morphology. While *C. vacuolata* have completely smooth cell walls, other members such as *C. yingshanensis* have a narrow, spindle shaped morphology [57]. Conventionally, most members of the *Coelastrella* genus are coccoid shaped with meridional ribs and polar thickenings. It is important to note that studies on *Chlorophyta* have shown that the morphological similarities are not representative of the phylogenetic relationships between different species [32]. Finally, the lower growth rate of *C. vacuolata* MACC-549 is likely due to the large cell size of this specific algae. There is well-documented evidence that cell size is inversely proportional to growth rate [45]. Thus, considering the large cell size of this species, its slower growth rate is not surprising.

The nuclear genome *C. vacuolata* MACC-549 had a total size of ~75.8 Mbp and a completion percentage of ~96%. The assembled genome also had partially assembled organelle genome fragments, and these were removed to obtain scaffolds for the nuclear genome only. Genome annotation revealed that most proteins found matches against the species of *T. obliquus*. Interestingly, most of these matches against *T. obliquus* had a sequence similarity ranging between 45 and 80%. This indicates that functional studies need to be carried out on *C. vacuolata* MACC-549 genes to identify whether functions are also conserved. Another reason why so many proteins match back to *T. obliquus* is that algal proteins downloaded from UniProt have 18,769 proteins derived from *T. obliquus* and this is the only representative from the *Scenedesmaceae* subfamily. Thus, while *C. vacuolata* MACC-549 shares homology with proteins from *T. obliquus*, there is sufficient evidence that some of these genes might have different or accessory functions. Finally, 877 *C. vacuolata* MACC-549 proteins found no matches in the database, and were defined as "Protein of unknown function". Thus, *C. vacuolata* MACC-549 has several novel proteins that could be highly beneficial to understanding the repertoire of functions that eukaryotic microalgae are capable of.

Organelle genomes typically have a high read coverage across most whole genome sequence data. This - coupled with their prokaryotic genome structure - makes it relatively easy to assemble compared to the complex nuclear genomes of eukaryotic microalgae. Most algal whole genome sequencing projects also reveal plastid genomes. Some studies have separated the assembled nuclear scaffolds from the assembled plastid scaffolds using information such as read depth or by finding matches against available chloroplast genomes. In this study, the complete chloroplast genome was assembled using an organelle specific assembler along with a seed sequence from the partial chloroplast scaffold. This led to a completely assembled and circularized chloroplast genome. Such an approach has been used in the past but only in a few studies [37]. Here, we show the importance or generating a high quality chloroplast assembly and its utility in identifying plastid structures such as IR (inverted repeat), LSC (large single copy) and SSC (small single copy) regions. The structure and characteristics of the *C. vacuolata* MACC-549 chloroplast genome (size of chloroplast genome, size of inverted repeats, etc.) were similar to those of other members of the Sphaeropleales order (Tables 2a & 2b). However, it had a smaller LSC region compared to the reference *C. saipanensis*. Recent studies have shown that chloroplast genome in *C. saipanensis* is expanded and contains more intronic ORFs compared to *T. obliquus* and *P. pectinatus* [58]. This pattern held true, as comparative analysis showed that the chloroplast genome of *C. vacuolata* MACC-549 also lacked these intronic ORFs. Progressive Mauve alignments clearly showed the high degree of conservation shared between the two *Coelastrella* species (Fig. 7c).

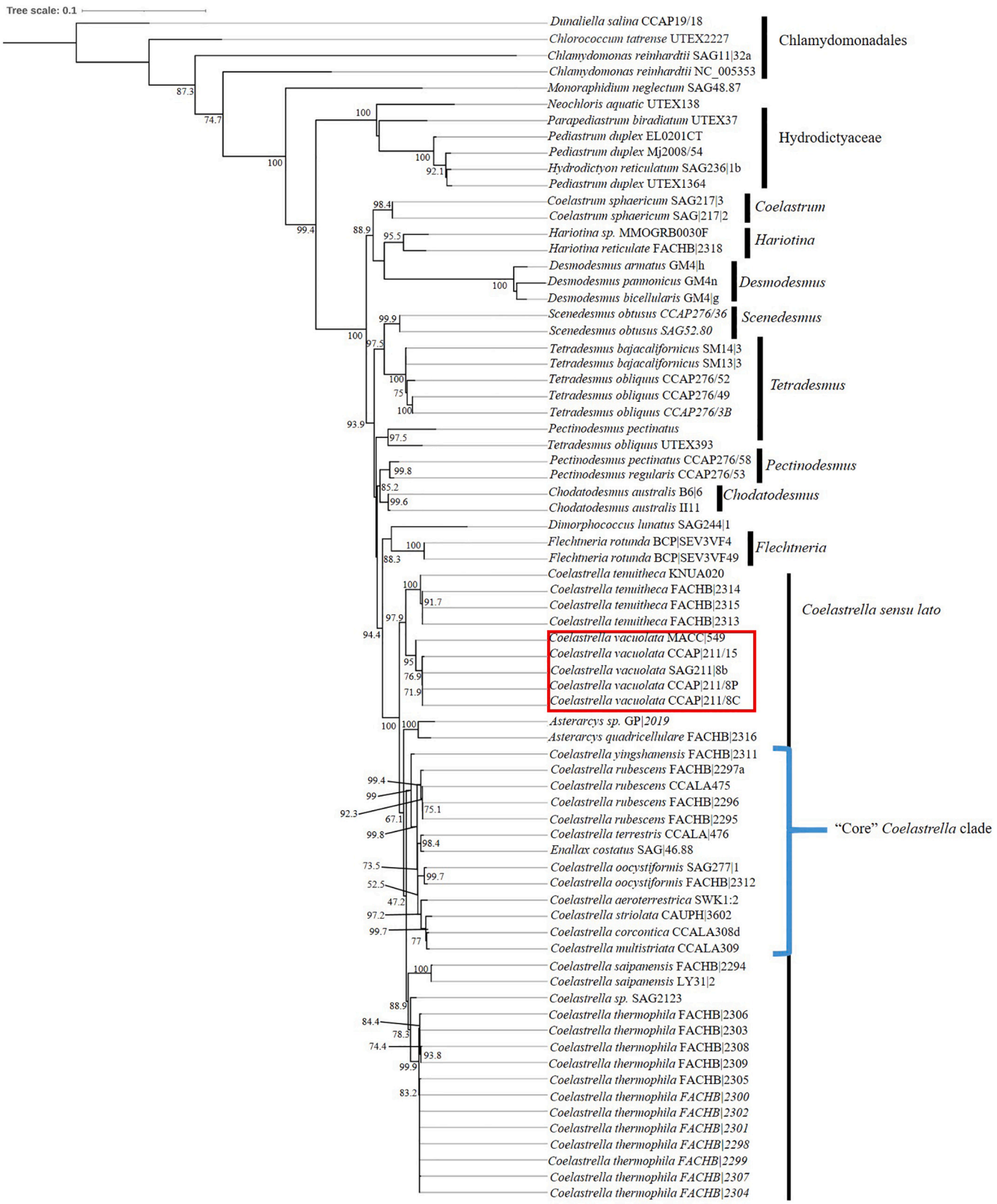


Fig. 7a. Phylogenetic tree constructed by iqTree using concatenated alignment of 3 nucleotide sequences (18S, ITS2 nuclear genes and *tufA* chloroplast gene) and 7 protein sequences (AtpA, PetA, PsaB, PsaC, PsbB, PsbD and RbcL proteins coded by chloroplast genes). Numbers next to clades represent percentage of bootstrap values. Red box marks the placement of *C. vacuolata* MACC-549. Tree is rooted at *Dunaliella salina*. Bootstrap values greater than 65 are shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

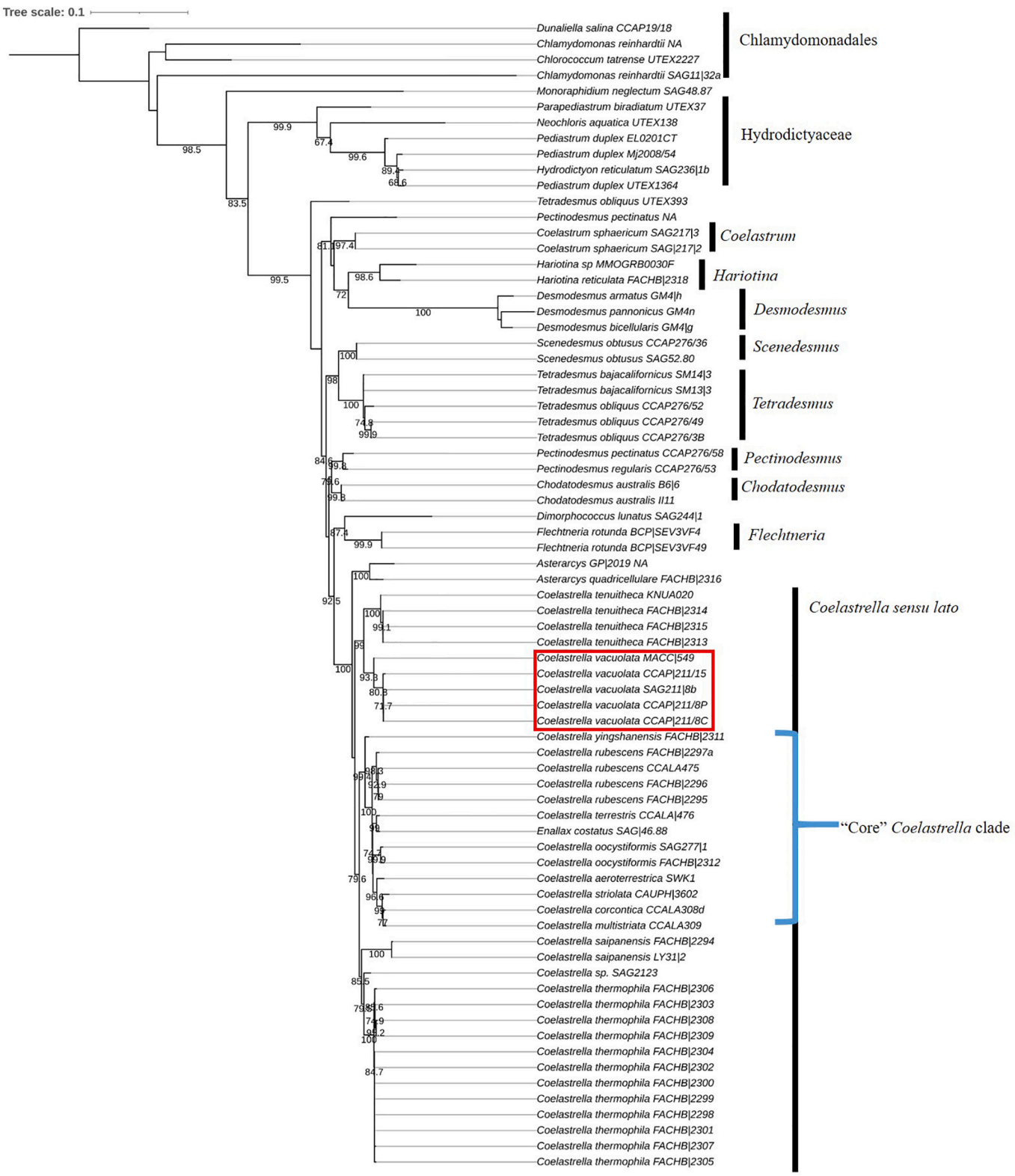


Fig. 7b. Phylogenetic tree constructed by iqTree using concatenated alignment of 3 gene sequences (18S, ITS2, *tufA*). Numbers next to clades represent percentage of bootstrap values. Red box marks the placement of *C. vacuolata* MACC-549. Tree is rooted at *Dunaliella salina*. Bootstrap values greater than 65 are shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Interestingly, less genes were identified in the chloroplast genome of *C. vacuolata* MACC-549 compared to the chloroplast genome of *C. saipanensis*.

4.2. Phylogenomic analysis

Conventionally, most algal studies use 18S rDNA or ITS sequences to investigate phylogenetic relationships. However, studies using only these sequences have led to unresolved phylogenies in Chlorophyta. The

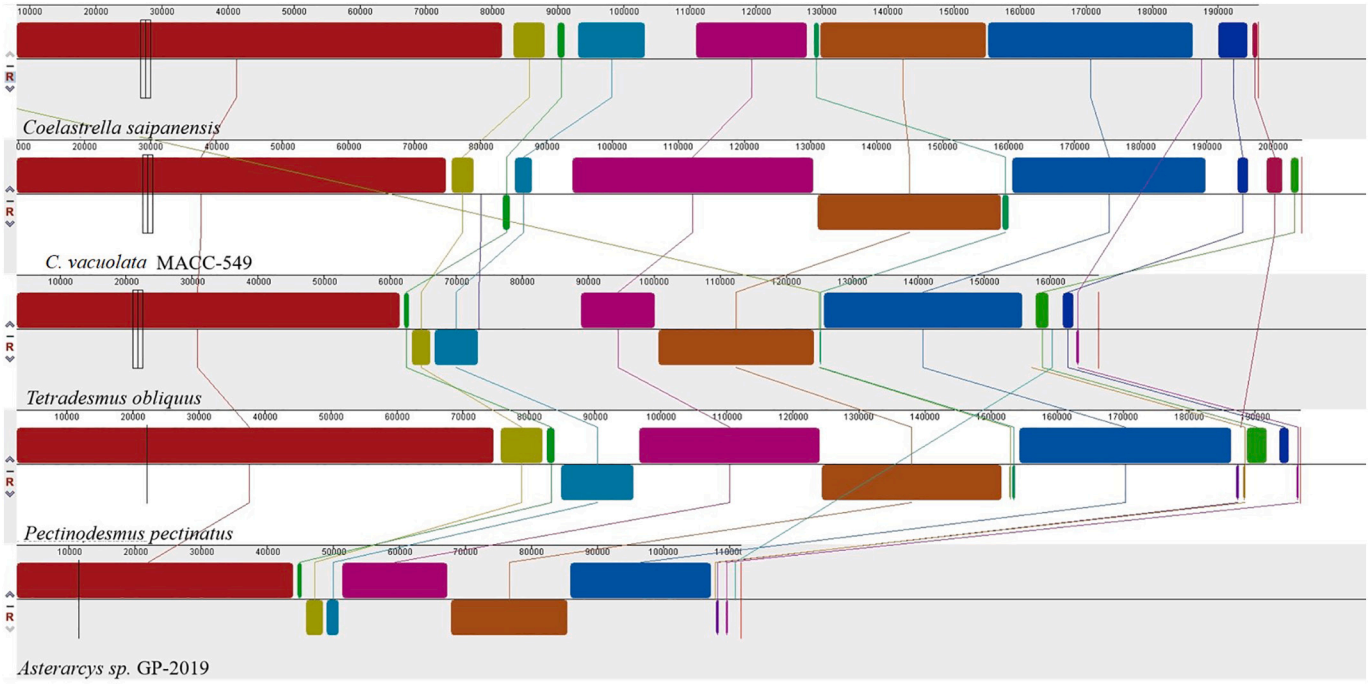


Fig. 7c. Multicollinearity of four chloroplast genomes across different *Scenedesmaceae* species. These genomes were aligned using ProgressiveMauve 2.3.1. Each colored block indicates a region of synteny between the genomes. Lines connecting blocks indicate putative homology. Boxes oriented in the same relative direction as *Coelastrella saipanensis* are shown above the line while those on the relative opposite strand are shown beneath the line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

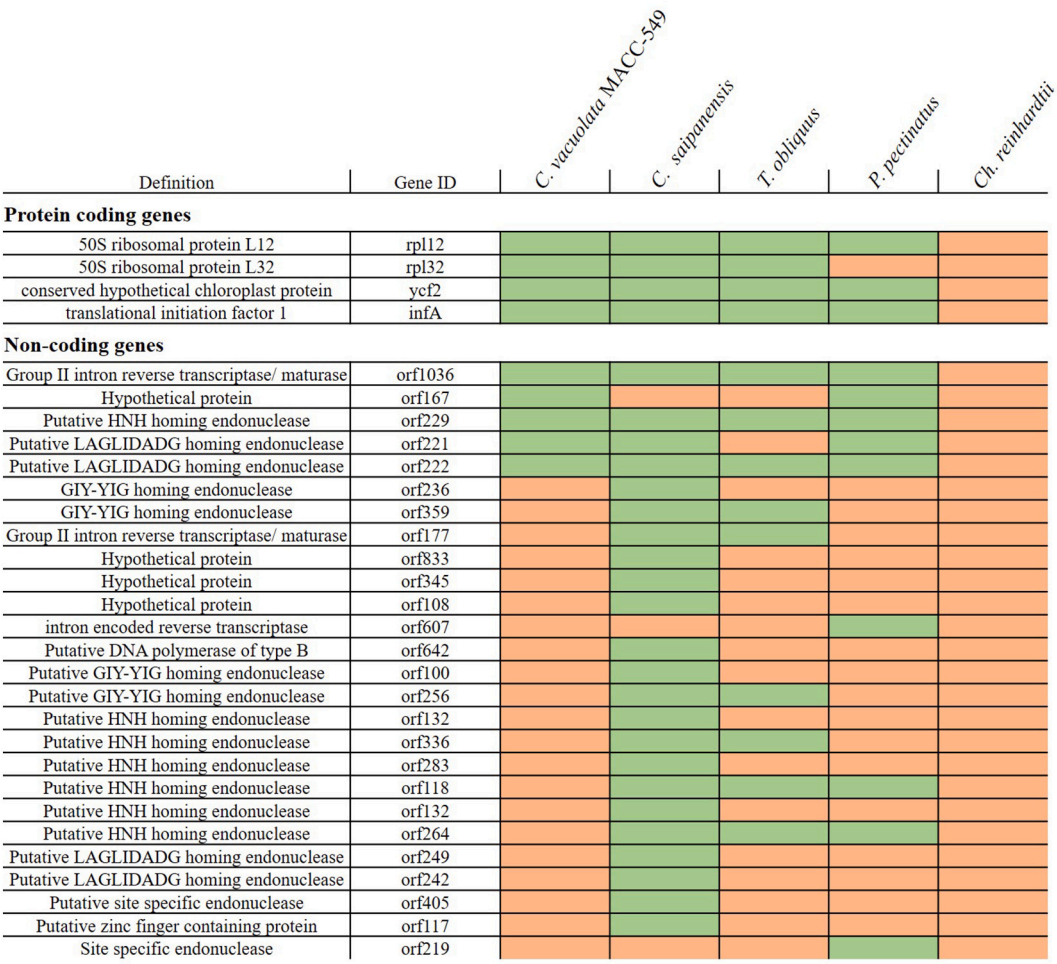
limited resolution of 18S based phylogenies have made it difficult to assess relationships in clades such as Scenedesmaceae [15,12,23], Ulvophyceae and Trebouxiophyceae [42,35]. In this study, a phylogenomic approach using 3 nucleotide (the nuclear 18S and ITS2 genes as well as the chloroplast gene *tufA*) and 7 protein sequences (PetA, PsbD, PsaC, PsbB, RbcL, AtpA, and PsaB proteins coded on the chloroplast genome) was used to build a phylogenetic tree and placed our model *C. vacuolata* MACC-549 species as close relative of *C. vacuolata*.

Utilizing more genes and diverse algal species allows for improved resolution and species-specific phylogeny. This was previously observed in the phylogenetic study of *Coelastrella* genus [57] where trees built using 18S rDNA data alone were too conserved for species-specific identification. Furthermore, phylogenetic relationships between *C. vacuolata*, *C. tenuitheca* sp. and *A. quadricellulare* weren't resolved with high bootstrap values using the 18S + ITS2 phylogeny. We were able to overcome this hurdle and obtain good resolution between the different *Coelastrella* and *Asterarcys* species with high bootstrap values (> 90%) using the proposed combined phylogenetic method including the chloroplast proteins.

However, multi-gene phylogenies can be hard to build without complete chloroplast genomes. Thus, we propose that future studies use universal primers for additional chloroplast genes such as *rbcL* and *tufA* to resolve complex phylogenies [56]. Unfortunately, most genera only have 18S rDNA sequence data and almost no sequence data on maker genes such as *tufA* and *rbcL*, thus limiting the number of studies that have used this approach [29,39]. A quick search on NCBI shows that only 2 species of the genus *Coelastrella* (search terms used: "txid75800 [Organism:exp] *tufA/rbcL*") and *Asterarcys* (search terms used: "txid183308[Organism:exp] *tufA/rbcL*") have sequence data for *tufA* and *rbcL* marker genes. On the other hand, 18S rDNA sequences are available for at least 10 different species of *Coelastrella* and 7 different species of *Asterarcys*. Amplifying additional marker genes such as *tufA* and *rbcL* along with the 18S rDNA fragment is a scalable and affordable way to develop highly resolved species-specific phylogenies, while waiting for whole chloroplast genomes to be available. This result is supported in our analysis where a novel, mixed loci phylogenetic tree had greater support compared to conventional loci trees alone. The alignments used in this paper is also publically available and can be used to place other candidate members of Sphaeropleales in the future (https://zenodo.org/record/4720668). Expanding the number of genes - as proposed here - would greatly aid phylogenies in complex clades such as *Coelastrella*, as has been observed in studies where multi-gene datasets successfully clarified the phylogenetic relationships in previously unresolved phylogenies [34,26]. This is the primary reason to include *tufA* in the nucleotide alignment, along with the other conventionally used genes (18S rDNA and ITS2).

4.3. Mating type locus analysis

The majority of genes present on the *Ch. reinhardtii* mating type locus were identified in *C. vacuolata* MACC-549 and *T. obliquus*. Sexual breeding in *Ch. reinhardtii* can be induced by nitrogen deprivation. Members of the Sphaeropleales order such as *T. obliquus* (*Scenedesmus obliquus*) have well documented sexual breeding behavior which can be induced by nitrogen deprivation [54] or chromium exposure [9]. An important gene: *MIDm* (Minus Dominance) was identified in the genome of *C. vacuolata* MACC-549 and *T. obliquus*. *MIDm* is the master regulator in the determination of mating type, and is upregulated in response to nitrogen deprivation, thus linking nitrogen deprivation and gametogenesis. However, other mating type locus genes are absent in both genomes of *T. obliquus* and *C. vacuolata* MACC-549. Genes essential for gametogenesis like *SAD1* (sexual adhesion) and *FUS1* (membrane bound protein) are missing in both *T. obliquus* and *C. vacuolata* MACC-549. Interestingly, studies in colonial Volvocales showed a similar result; the *MIDm* gene was identified by sequence similarity but other gametogenesis genes such as *FUS1* and *SAD1* were absent in *Volvox carteri* and *Gonium pectorale* [18]. Recent molecular studies have shown that these genes might be evolving too rapidly, and only share a very small region with high sequence similarity – as observed with the *FUS1* gene of *Gonium pectorale* [18].



Definition	Gene ID
Protein coding genes	
50S ribosomal protein L12	rpl12
50S ribosomal protein L32	rpl32
conserved hypothetical chloroplast protein	ycf2
translational initiation factor 1	infA
Non-coding genes	
Group II intron reverse transcriptase/ maturase	orf1036
Hypothetical protein	orf167
Putative HNH homing endonuclease	orf229
Putative LAGLIDADG homing endonuclease	orf221
Putative LAGLIDADG homing endonuclease	orf222
GIY-YIG homing endonuclease	orf236
GIY-YIG homing endonuclease	orf359
Group II intron reverse transcriptase/ maturase	orf177
Hypothetical protein	orf833
Hypothetical protein	orf345
Hypothetical protein	orf108
intron encoded reverse transcriptase	orf607
Putative DNA polymerase of type B	orf642
Putative GIY-YIG homing endonuclease	orf100
Putative GIY-YIG homing endonuclease	orf256
Putative HNH homing endonuclease	orf132
Putative HNH homing endonuclease	orf336
Putative HNH homing endonuclease	orf283
Putative HNH homing endonuclease	orf118
Putative HNH homing endonuclease	orf132
Putative HNH homing endonuclease	orf264
Putative LAGLIDADG homing endonuclease	orf249
Putative LAGLIDADG homing endonuclease	orf242
Putative site specific endonuclease	orf405
Putative zinc finger containing protein	orf117
Site specific endonuclease	orf219

Fig. 7d. Comparison of shared and unique chloroplast genes across different members of Sphaeropleales and *Ch. reinhardtii*. Genes shared across all species are not shown. Red boxes indicated missing genes, and green boxes indicate shared genes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

There currently has been no documented evidence of sexual breeding in the *Coelastrella* genus. However, the shared presence of mating locus genes with *T. obliquus* lends credibility to the idea that members of the genus *Coelastrella* might be able to sexually reproduce. It is important to note that strains can modulate their sexual competence and efficiency. This was empirically demonstrated in a study where the most zygotes were obtained from cultures that were maintained under sexual selection and the fewest from those maintained asexually. Curiously, over long term sexual selection, zygote production became spontaneous, and no nitrogen starvation was required to induce the process [4]. No similar studies have been carried out in members of the order Sphaeropleales.

4.4. Analysis of industrially relevant genes

Members of the *Coelastrella* genus are considered to be potent carotenoid producers [1,2,24]. However, carotenogenesis pathways and corresponding enzymes have primarily been studied in cyanobacteria [52]. Identification of genes involved in these pathways can provide important molecular resources to be utilized for qPCR-based transcript analysis in other *Coelastrella* and green algae species. In this study, enzymes required for the production of phytoene, lycopene, β-carotene, zeaxanthin, astaxanthin and α-carotene were identified. Multiple homologs were identified for important enzymes like GGPS; involved in the synthesis of phytoene, PDS; involved in the synthesis of lycopene, BKT; involved in the synthesis of astaxanthin and ZEP; involved in the synthesis of violaxanthin. Previous studies have shown that multiple isoforms exist for these enzymes and these isoforms are regulated differentially by light [8]. Interestingly, the enzyme LYC-E; involved in the synthesis of α-carotene was also identified. α-carotene is a precursor for lutein and their derivatives. The genes and pathways for carotenoid production in *C. vacuolata* MACC-549 was identified and will be an important resource for those working in the field of carotenoid production.

The hydrogenase genes in *C. vacuolata* MACC-549 share high similarity with the hydrogenase genes extracted from *Chlorella fusca* [59]. The Fe-hydrogenase enzyme is a monomeric enzyme with ferrodoxin as the electron donor. Apart from the hydrogenase structural gene, genes encoding the maturation proteins were also identified. The proteins coded by the *hydEF* and *hydG* genes are responsible for the incorporation of the cofactor into the apoenzyme. Earlier studies showed, that *C. vacuolata* MACC-549 is a potent hydrogen producer, especially under bacterial-associated conditions [34].

5. Conclusions

The past two decades has ushered in copious amounts of genomic data due to the development of affordable next generation sequencing technologies. Unfortunately, the benefits of these advances are unequally distributed, with model organisms having a large amount of molecular and genetic resources, while others are largely ignored. Such a phenomenon is also observed in plant studies and crops with underdeveloped molecular resources are now being referred to as orphans of

Fig. 8. Presence/absence heat-map of shared mating locus genes. Genes that are present in the specific alga are shown in blue and genes that are absent are shown in white. 8.a) Genes present/absent in minus type. 8.b) Genes present/absent in plus type. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the genomic revolution or orphan crops [55]. So far only a few model algae species have received the benefit of the genomic data revolution. Genera like *Chlamydomonas* and *Chlorella* now have a few complete and multiple draft genomes available while other industrially viable eukaryotic algal species have not received similar attention.

This situation is also exemplified in this study, where even nuclear genomes of reference species are largely uncharacterized. In this paper the genome of a cryptic algae species *C. vacuolata* MACC-549 was sequenced, assembled and annotated. *C. vacuolata* MACC-549 was morphologically identified by its distinct lack of longitudinal ridges along the cell surface. Phylogenomic analysis placed the species as a close branch of other *C. vacuolata* strains. This study marks the first description of the chloroplast genome for any *C. vacuolata* strain. Multiple intronic ORFs were identified in the chloroplast genome of *C. saipanensis*. Comparative genomic analysis revealed the absence of these ORFs in *C. vacuolata* and other Scenedesmaceae such as *T. obliquus* and *P. pectinatus*, indicating that the presence of intronic ORFs could be distinctiveness of the *C. saipanensis* chloroplast genome.

CRediT authorship contribution statement

PS composed the manuscript and performed all bioinformatics analyses; AF performed the light and electron microscopy analyses, BP and BH cultivated the MACC strains and participated in the genome sequencing, TV designed and developed the multi-loci phylogeny, VÖ provided the algae strains and discussed the data, TB provided useful practical hints and participated in the critical discussions and GM designed the study, composed the manuscript and thoroughly discussed the relevant literature. All authors read and approved the final manuscript.

Funding

This work was supported by the following international and domestic funds: NKFI-FK-123899 (GM), GINOP-2.2.1-15-2017-00042, Lendület-Programme (GM) of the Hungarian Academy of Sciences (LP2020-5/2020).

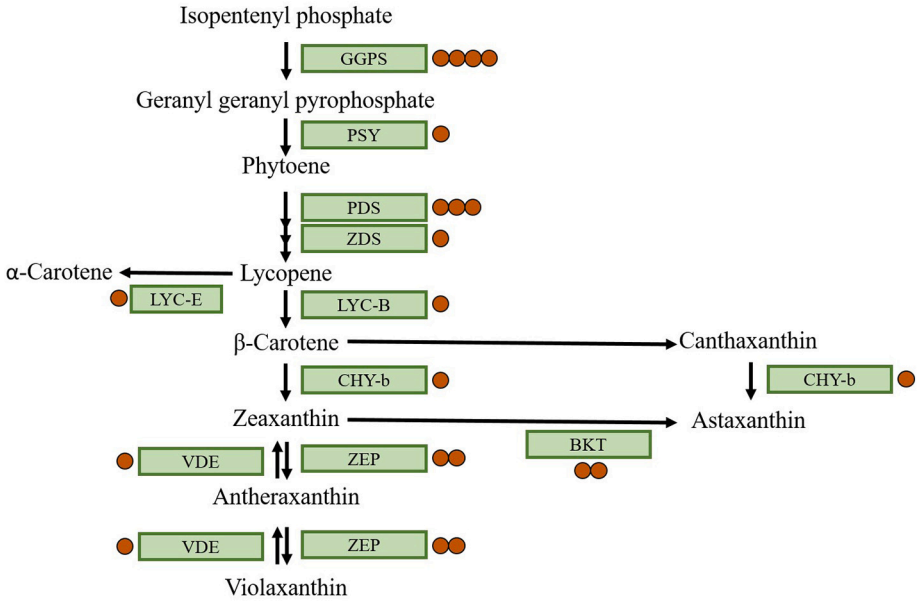


Fig. 9. Carotenoid biosynthesis pathway in *C. vacuolata* MACC-549. Enzymes that have orthologous matches are represented in green boxes. The filled brown circles show the number of homologs identified for the specific enzyme. GGPS (geranylgeranyl pyrophosphate synthase), PSY (phytoene synthase), PDS (phytoene desaturase), ZDS (zeta-carotene desaturase), LYC-B (lycopene beta cyclase), CHY-B (beta carotenoid hydroxylase), ZEP (zeaxanthin epoxidase), VDE (violaxanthin de-epoxidase), BKT (beta-carotene ketolase), LYC-E (lycopene epsilon cyclase). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Hydrogenase genes.

Gene name	Protein ID	Database organism name.
HYDA/Fe Hydrogenase	*Coelastrella_vacuolata* _sp.549_004415-RA	*Chlorella fusca*
HYDG	*Coelastrella_vacuolata* _sp.549_008760-RA	*Chlorella variabilis*
HYDEF	*Coelastrella_vacuolata* _sp.549_008761-RA	*Chlorella variabilis*

Declaration of competing interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.algal.2021.102380.

References

[1] Katsuya Abe, Hiroaki Hattori, Morio Hirano, Accumulation and antioxidant activity of secondary carotenoids in the aerial microalga *Coelastrella striolata var. multistriata*, Food Chem. 100 (2) (2007) 656–661.

[2] Nobuhiro Aburai, Satoshi Ohkubo, Hideaki Miyashita, Katsuya Abe, Composition of carotenoids and identification of aerial microalgae isolated from the surface of rocks in mountainous districts of Japan, Algal Res. 2 (3) (2013) 237–243.

[3] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (5) (2012) 455–477.

[4] G. Bell, Experimental sexual selection in Chlamydomonas, J. Evol. Biol. 18 (3) (2005) 722–734.

[5] Brandi L. Cantarel, Ian Korf, Sofia M.C. Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, Mark Yandell, MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes, Genome Res. 18 (1) (2008) 188–196.

[6] Jose Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, Mol. Biol. Evol. 17 (4) (2000) 540–552.

[7] Olga Chernomor, Arndt Von Haeseler, Bui Quang Minh, Terrace aware data structure for phylogenomic inference from supermatrices, Syst. Biol. 65 (6) (2016) 997–1008.

[8] Sacha Coesel, Miroslav Oborník, Joao Varela, Angela Falciatore, Chris Bowler, Evolutionary origins and functions of the carotenoid biosynthetic pathway in marine diatoms, PLoS One 3 (8) (2008), e2896.

[9] M. Grazia Corradi, Gessica Gorbi, Ada Ricci, Anna Torelli, Maria Bassi, Chromium-induced sexual reproduction gives rise to a Cr-tolerant progeny in *Scenedesmus acutus*, Ecotoxicol. Environ. Saf. 32 (1) (1995) 12–18.

[10] Dierckxsens, Nicolas, Patrick Mardulyn, and Guillaume Smits. "NOVOPlasty: de novo assembly of organelle genomes from whole genome data." Nucleic Acids Res. 45, no. 4 (2017): e18-e18.

[11] Robert C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (5) (2004) 1792–1797.

[12] Marek Eliaš, Yvonne Němcová, Pavel Škaloud, Jiří Neustupa, Veronika Kaufnerova, Lenka Šejnohová, *Hylodesmus singaporensis* gen. et sp. nov., a new autosporic subaerial green alga (Scenedesmaceae, Chlorophyta) from Singapore, Int. J. Syst. Evol. Microbiol. 60 (5) (2010) 1224–1235.

[13] David Ellinghaus, Stefan Kurtz, Ute Willhoeft, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons, BMC bioinformatics 9 (1) (2008) 18.

[14] David M. Emms, Steven Kelly, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, Genome Biol. 16 (1) (2015) 157.

[15] Thomas Friedl, Nataliya Rybalka, Systematics of the green algae: a brief introduction to the current status, in: Progress in Botany 73, Springer, Berlin, Heidelberg, 2012, pp. 259–280.

[16] Karolina Fučíková, Frederik Leliaert, Endymion D. Cooper, Pavel Škaloud, Sofie D'hondt, Olivier De Clerck, Carlos F.D. Gurgel, et al., New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data, Front. Ecol. Evol. 2 (2014) 63.

[17] Maria Lucia Ghirardi, Alexandra Dubini, Jianping Yu, Pin-Ching Maness, Photobiological hydrogen-producing systems, Chem. Soc. Rev. 38 (1) (2009) 52–61.

[18] Takashi Hamaji, Yuko Mogi, Patrick J. Ferris, Toshiyuki Mori, Shinya Miyagishima, Yukihiro Kabeya, Yoshiki Nishimura, et al., Sequence of the *Gonium pectorale* mating locus reveals a complex and dynamic history of changes in volvocine algal mating haplotypes, G3: Genes, Genomes, Genetics 6 (5) (2016) 1179–1189.

[19] Han, Yujun, and Susan R. Wessler. "MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences." Nucleic Acids Res. 38, no. 22 (2010): e199-e199.

[20] Nobutaka Hanagata, Phylogeny of the subfamily Scotiellocystoideae (Chlorophyceae, Chlorophyta) and related taxa inferred from 18S ribosomal RNA gene sequence data, J. Phycol. 34 (6) (1998) 1049–1054.

[21] Hegewald, Eberhard, and Nobutaka Hanagaza. "Validation of the new combinations of *Coelastrella* and *Neodesmus* and the description of the new subfamily Desmodesmoideae of the Scenedesmaceae (Chlorophyta)." Algological Studies/Archiv für Hydrobiologie, Supplement Volumes (2002): 7–9.

[22] Hegewald, Eberhard, and Nobutaka Hangata. "Phylogenetic studies on Scenedesmaceae (Chlorophyta)." Algological Studies/Archiv für Hydrobiologie, Supplement Volumes (2000): 29–49.

[23] Eberhard Hegewald, Matthias Wolf, Alexander Keller, Thomas Friedl, Lothar Krienitz, ITS2 sequence-structure phylogeny in the Scenedesmaceae with special reference to *Coelastrum* (Chlorophyta, Chlorophyceae), including the new genera *Comasiella* and *Pectinodesmus*, Phycologia 49 (4) (2010) 325–335.

[24] Che-Wei Hu, Lu-Te Chuang, Po-Chien Yu, Ching-Nen Nathan Chen, Pigment production by a new thermotolerant microalga *Coelastrella sp. F50*, Food Chem. 138 (4) (2013) 2071–2078.

[25] Tomáš Kalina, M. Puncochárová, Taxonomy of the subfamily Scotiellocystoideae Fott 1976 (Chlorellaceae, Chlorophyceae), Archiv für Hydrobiologie. Supplementband. Monographische Beiträge 73 (4) (1987) 473–521.
[26] Subha Kalyaanamoorthy, Bui Quang Minh, Thomas K.F. Wong, Arndt Von Haeseler, Lars S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates, Nat. Methods 14 (6) (2017) 587–589.
[27] Kazutaka Katoh, Daron M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (4) (2013) 772–780.
[28] Szabina Katona, D.E. Berthold Nándor Horváth, Zoltán Molnár, Péter Bálint, Vince Ördög, Bernadett Pap, et al., Phylogenetic re-evaluation of previously identified Chlamydomonas (Chlorophyta, Chlamydomonadaceae) strains from The Mosonmagyaróvár Algal Culture Collection, Hungary, using molecular data, S. Afr. J. Bot. 125 (2019) 16–23.
[29] Mudassar Anisoddin Kazi, C.R.K. Reddy, Bhavanath Jha, Molecular phylogeny and barcoding of Caulerpa (Bryopsidales) based on the tufA, rbcL, 18S rDNA and ITS rDNA genes, PLoS One 8 (12) (2013).
[30] Jiří Komárek, Bohuslav Fott, Chlorophyceae (Grünalgen), Ordnung: Chlorococcales, Das Phytoplankton des Süsswassers: Systematik und Biologie 7 (1983).
[31] Ian Korf, Gene finding in novel genomes, BMC bioinformatics 5 (1) (2004) 59.
[32] L. Krienitz, C. Bock, Present state of the systematics of planktonic coccoid green algae of inland waters, Hydrobiologia 698 (2012) 295–326, https://doi.org/10.1007/s10750-012-1079-z.
[33] Gergely Lakatos, Daniella Balogh, Attila Farkas, Vince Ördög, Péter Tamás Nagy, Tibor Bíró, Gergely Maróti, Factors influencing algal photobiohydrogen production in algal-bacterial co-cultures, Algal Res. 28 (2017) 161–171.
[34] Gergely Lakatos, Zsuzsanna Deák, Imre Vass, Tamás Rétfalvi, Szabolcs Rozgonyi, Gábor Rákhely, Vince Ördög, Éva Kondorosi Vince, Gergely Maróti, Bacterial symbionts enhance photo-fermentative hydrogen evolution of Chlamydomonas algae, Green Chem. 16 (11) (2014) 4716–4727.
[35] Claude Lemieux, Christian Otis, Monique Turmel, Chloroplast phylogenomic analysis resolves deep-level relationships within the green algal class Trebouxiophyceae, BMC Evol. Biol. 14 (1) (2014) 211.
[36] Ivica Letunic, Peer Bork, Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation, Bioinformatics 23 (1) (2007) 127–128.
[37] Linzhou Li, Haoyuan Li, Sunil Kumar Sahu, Yan Xu, Hongli Wang, Hongping Liang, Sibo Wang, The complete chloroplast genome of *Tetraselmis desikacharyi* (Chlorodendrophyceae) and phylogenetic analysis, Mitochondrial DNA Part B 4 (1) (2019) 1692–1693.
[38] Bui Quang Minh, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt Von Haeseler, Robert Lanfear, IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era, Mol. Biol. Evol. 37 (5) (2020) 1530–1534.
[39] Mónica B.J. Moniz, Michael D. Guiry, Fabio Rindi, tuf A phylogeny and species boundaries in the green algal order Prasiolales (Trebouxiophyceae, Chlorophyta), Phycologia 53 (4) (2014) 396–406.
[40] Jared O'Connell, Ole Schulz-Trieglaff, Emma Carlson, Matthew M. Hims, Niall A. Gormley, Anthony J. Cox, NxTrim: optimized trimming of Illumina mate pair reads, Bioinformatics 31 (12) (2015) 2035–2037.
[41] V. Ordog, Biotechnological application of microalgae in crop and plant protection, Doctoral thesis. Mosonmagyaróvár. (2015) 1–174.
[42] T. Proschold, Frederik Leliaert, Systematics of the green algae: conflict of classic and modern approaches, Systematics Association Special 75 (2007) 123.
[43] Otto Pulz, Wolfgang Gross, Valuable products from biotechnology of microalgae, Appl. Microbiol. Biotechnol. 65 (6) (2004) 635–648.
[44] RStudio Team, RStudio: Integrated development for R, in: URL, RStudio, Inc, Boston, MA, 2015. http://www.rstudio.com/.
[45] D.A. Schlesinger, L.A. Molot, B.J. Shuter, Specific growth rates of freshwater algae in relation to cell size and light intensity, Can. J. Fish. Aquat. Sci. 38 (9) (1981) 1052–1058.
[46] Mathieu Seppey, Mosè Manni, Evgeny M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness, in: Gene Prediction, Humana, New York, NY, 2019, pp. 227–245.
[47] Shetty, Prateek, Iulian Z. Boboescu, Bernadett Pap, Roland Wirth, Kornél Lajos Kovács, Tibor Bíró, Zoltán Futó, Richard A. White, and Gergely Maróti. "Exploitation of algal-bacterial consortia in combined biohydrogen generation and wastewater treatment." Frontiers in Energy Research 7 (2019): Terjedelem-13.
[48] Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013–2015.
[49] Smit, AFA, Hubley, R. RepeatModeler Open-1.0. 2008–2015.
[50] Mi Song, Ok-Min Lee, A study of newly recorded genera and species of aerial algae in the order Chlorococcales (Chlorophyta) from the Hongcheon-river, Korea, Journal of Ecology and Environment 37 (4) (2014) 315–325.
[51] Stanke, Mario, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. "AUGUSTUS: ab initio prediction of alternative transcripts." Nucleic Acids Res. 34, no. suppl_2 (2006): W435.
[52] S. Takaichi, M. Mochimaru, Carotenoids and carotenogenesis in cyanobacteria: unique ketocarotenoids and carotenoid glycosides, Cell. Mol. Life Sci. 64, no. 19-20 (2007) 2607.
[53] Michael Tillich, Pascal Lehwark, Tommaso Pellizzer, Elena S. Ulbricht-Jones, Axel Fischer, Ralph Bock, Stephan Greiner, GeSeq–versatile and accurate annotation of organelle genomes, Nucleic Acids Res. 45 (W1) (2017) W6–W11.
[54] Trainor, Francis R. "Reproduction in Scenedesmus." Algae 11, no. 2 (1996): 183–183.
[55] Rajeev K. Varshney, Timothy J. Close, Nagendra K. Singh, David A. Hoisington, Douglas R. Cook, Orphan legume crops enter the genomics era!, Curr. Opin. Plant Biol. 12 (2) (2009) 202–210.
[56] Helena Henriques Vieira, Inessa Lacativa Bagatini, Carla Marques Guinart, Armando Augusto Henriques Vieira, tufA gene as molecular marker for freshwater Chlorophyceae, Algae 31 (2) (2016) 155–165.
[57] Qinghua Wang, Huiyin Song, Xudong Liu, Benwen Liu, Zhengyu Hu, Guoxiang Liu, Morphology and molecular phylogeny of coccoid green algae *Coelastrella* sensu lato (Scenedesmaceae, Sphaeropeales), including the description of three new species and two new varieties, J. Phycol. 55 (6) (2019) 1290–1305.
[58] Qinghua Wang, Huiyin Song, Xudong Liu, Huan Zhu, Zhengyu Hu, Guoxiang Liu, Deep genomic analysis of *Coelastrella saipanensis* (Scenedesmaceae, Chlorophyta): comparative chloroplast genomics of Scenedesmaceae, Eur. J. Phycol. 54 (1) (2019) 52–65.
[59] Winkler, Martin, Burkhard Heil, Bettina Heil, and Thomas Happe. "Isolation and molecular characterization of the [Fe]-hydrogenase from the unicellular green alga *Chlorella fusca*." Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression 1576, no. 3 (2002): 330–334.
[60] Wenguang Zhou, Yecong Li, Min Min, Bing Hu, Paul Chen, Roger Ruan, Local bioprospecting for high-lipid producing microalgal strains to be grown on concentrated municipal wastewater for biofuel production, Bioresour. Technol. 102 (13) (2011) 6909–6919.