The Draft Genome of Coelastrum proboscideum (Sphaeropleales, Chlorophyta)

Liang, Hongping; Wang, Hongli; Xu, Yan; Li, Linzhou; Melkonian, Barbara; Lorenz, Maike; Friedl, Thomas; Sahu, Sunil Kumar; Yu, Jin; Liu, Huan

Total number of authors:
12

Published in:
Protist

Link to article, DOI:
10.1016/j.protis.2020.125758

Publication date:
2020

Document Version
Peer reviewed version

Citation (APA):
Liang, H., Wang, H., Xu, Y., Li, L., Melkonian, B., Lorenz, M., Friedl, T., Sahu, S. K., Yu, J., Liu, H., Melkonian, M., & Wang, S. (2020). The Draft Genome of Coelastrum proboscideum (Sphaeropleales, Chlorophyta). *Protist*, *171*(5), Article 125758. https://doi.org/10.1016/j.protis.2020.125758

Protist Genome Reports

The Draft Genome of *Coelastrum proboscideum* (Sphaeropleales, Chlorophyta)

Hongping Liang[a,b,2], Hongli Wang[a,b,2], Yan Xu[a,b], Linzhou Li[c,d], Barbara Melkonian[e,f], Maike Lorenz[g], Thomas Friedl[g], Sunil Kumar Sahu[a], Jin Yu[a,b], Huan Liu[a,h], Michael Melkonian[e,f,1], and Sibo Wang[a,h,1]

[a]BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China
[b]BGI Education Center, University of Chinese Academy of Sciences, Beijing, China
[c]China National GeneBank, BGI-Shenzhen, Jinsha Road, Shenzhen 518120, China
[d]Department of Biotechnology and Biomedicine, Technical University of Denmark, Copenhagen, Denmark
[e]University of Duisburg-Essen, Campus Essen, Faculty of Biology, Universitätsstr. 5, 45141 Essen, Germany
[f]Max Planck Institute for Plant Breeding Research, Carl-von_Linne-Weg 10, 50829 Cologne, Germany
[g]Department 'Experimentelle Phykologie und Sammlung von Algenkulturen' (EPSAG), University of Göttingen, Nikolausberger Weg 18, 37073 Göttingen, Germany
[h]Department of Biology, University of Copenhagen, Copenhagen, Denmark

Running title: Draft Genome of *Coelastrum proboscideum*

Coelastrum proboscideum **Bohlin 1896 (Sphaeropleales, Scenedesmaceae, Chlorophyta) is a coenobial species with cosmopolitan distribution in diverse freshwater habitats.** ***Coelastrum*** **spp. are widely tested for biotechnological applications such as carotenoid and lipid production, and in bioremediation of wastewater. Here, we report the draft genome of** ***Coelastrum proboscideum*** **var.** ***dilatatum*** **strain SAG 217-2. The final assembly comprised 125,935,854 bp with**

over 8,357 scaffolds. The whole-genome data is publicly available in the Nucleotide Sequence Archive (CNSA) of China National GeneBank (CNGB) (https://db.cngb.org/cnsa/) under the accession number CNA0014153.

Key words: Scenedesmaceae; Coelastroideae; genome; algae.

[1]Corresponding authors; e-mails michael.melkonian@uni-koeln.de; wangsibo1@genomics.cn

[2]These authors contributed equally.

The Scenedesmaceae Oltmanns, 1904 is the largest family in the order Sphaeropleales (Chlorophyceae) with over 300 described species containing some well-known genera such as *Coelastrum*, *Desmodesmus* and *Tetradesmu*s (Guiry and Guiry 2020). Alga of the Scenedesmaceae family are common constituents of freshwater phytoplankton, and because of their rapid growth and high lipid contents are intensively studied as potential sources of biofuels (Arora et al. 2019; Neofotis et al. 2016; Shuba and Kifle 2018). Previously, draft genomes have been obtained from species of *Desmodesmus* and *Tetradesmus* genera (Carreres et al. 2017; Starkenburg et al. 2017; Wang et al. 2019) but not from *Coelastrum*. Molecular phylogenetic analyses by Hegewald et al. (2010) concluded that taxa with spherical coenobia, that were previously placed in a separate family (Coelastraceae Wille, 1909), were part of the Scenedesmaceae forming a separate clade that the authors recognized at the subfamily level (Coelastroideae). Within Coelastroideae, the draft nuclear genome sequence of *Hariotina reticulata* was recently reported (Xu et al. 2019). Genus *Coelastrum* is the most species-rich genus in the subfamily with 30 taxonomically accepted species (Guiry and Guiry 2020). It has a worldwide distribution in the plankton of freshwater habitats from arctic to tropical environments and is often abundant under eutrophic conditions (Guiry and Guiry 2020). As such, non-pollen palynomorphs (NPPs) of *Coelastrum* spp. act as eutrophication markers in paleoecology (Stivrins et al. 2018). This is true also for *C. proboscideum* Bohlin, 1896. Strain SAG 217-2 (http://sagdb.uni-goettingen.de/detailedList.php?str_number=217-2) of *C. proboscideum* var. *dilatatum* is an authenic strain isolated by W. Vischer in 1924 from a small pond in Switzerland, the variety is currently regarded as a synonym of the type species *C. sphaericum* Nägeli (Guiry and Guiry 2020). *Coelastrum* spp. have been found to be morphologically highly polymorphic in culture and *C. proboscideum* SAG 217-2 is no exception (Fig. 1A; see also Fenwick et al. 1966; Großmann 1920; Hajdu et al. 1976). Strains of *Coelastrum* spp. are widely used in applied research, e.g. the production of secondary carotenoids (astaxanthin) or of lipids for biofuels as well as in bioremediation of wastewater (Del Campo et al. 2000; Mousavi et al. 2018; Rauytanapanit et al. 2019; Ribeiro et al. 2019; Soares et al. 2019; Úbeda et al. 2017), although the taxonomic identity of the (sometimes local) strains employed, is often not clear. A mitochondrial genome sequence from *Coelastrum* sp. F187 has recently been reported (Wang et al. 2017). The draft nuclear genome of *C. proboscideum* (strain SAG 217-2) represents the second nuclear genome sequence from a Scenedesmaceae with three-dimensional coenobia; it has been established in the frame of the 10 KP project, a phylodiverse genome sequencing plan (Cheng et al. 2018).

An axenic culture of *C. proboscideum* (SAG 217-2) (Sammlung von Algenkulturen, University of Göttingen, Germany) was grown in 3N BBM +V culture medium (https://www.ccap.ac.uk/media/documents/3N_BBM_V.pdf) in aerated Erlenmeyer flasks at 40 μmol photons m^{-2} s^{-1} in a 14:10 h L/D cycle up to a volume of 1,000 mL. The culture was harvested by centrifugation (300 *g*, 10 min), and then the pellet was immediately stored at -80 °C until freeze-drying. During all the steps of cultivationthe axenicity was monitored by sterility tests as well as light microscopy. Light microscopy was performed with a Leica DMLB light microscope using a PL-APO 100/1.40 objective, an immersed condenser N.A. 1.4 and a Metz Mecablitz 32 Ct3 flash system.

Total DNA was extracted by using a modified CTAB protocol (Sahu et al 2012). The extracted DNA of *C. proboscideum* was used to construct 10X Genomics Chromium library using the manufacturer's recommended protocols to obtain Linked-Reads. The library was sequenced by the BGISEQ-500 150bp pair-end platform. A total of 126G (~1128X) Linked-Reads were obtained (Supplementary Material Table S1). The genome size was estimated by Jellyfish (version 2.2.10) with 21-mer (Guillaume and Carl 2011), and the K-mer distribution diagram drawn by GenomeScope (Gregory et al. 2017). The raw data was assembled using Supernova (version 2.1.1) with default parameters (Weisenfeld et al. 2017).

For detecting the repetitive elements, we used both *de-novo* and homolog-based method to find DNA transposon elements, retrotransposon elements, and tandem repeats. For *ab initio* prediction we used Piler-DF, RepeatScout, MITE-hunter, LTR_FINDER, and RepeatModeler (version 1.0.8; http://www.repeatmasker.org/RepeatModeler/). Among them, Piler (http://www.drive5.com/piler) detected repeat elements such as satellites and transposons, RepeatScout (https://bix.ucsd.edu/repeatscout/) identified all repeat classes, MITE-hunter (Han et al. 2010) discovered miniature inverted repeat transposable elements (MITEs) from the genomic sequence, while LTR-FINDER (Ellinghaus et al. 2008) predicted the location and structure of full-length LTR retrotransposons. All results from *ab initio* prediction were merged as homolog database to identified repetitive sequences by RepeatMasker (Chen et al. 2004).

We used automated BRAKER2 (Hoff et al. 2016) to obtain accurate gene models of *C. proboscideum*, which combined de novo and homology-based predictions with GeneMark-ES/ET (Besemer and Borodovsky 2005) and AUGUSTUS (Stanke et al. 2006). For training GeneMark-TP and AUGUSTUS, we selected all Chlorophyta proteins from the NR database (non-redundant protein database). To assess genome completeness, we used BUSCO (Waterhouse et al. 2018) core eukaryotic proteins with E-values < $1e^{-5}$. For the functional annotation of genes, the *C. proboscideum* genes were searched against several databases, including NR, SwissProt, KEGG, COG, InterProScan and GO by blastp (E-value < $1e^{-5}$).

A phylogenetic analysis was performed using 24 previously published Chlorophyta genomes including 13 Chlorophyceae, 1 Ulvophyceae, 4 Trebouxiophyceae, 1 Chlorodendrophyceae, and 5 Mamiellophyceae. We selected 111 single-copy gene families to construct a concatenated phylogenetic tree which performed by OrthoFinder version 2.3.3 (Emms and Kelly 2019). The amino acid alignments were generated by MAFFT version 7.310 (Katoh et al. 2002)). The genes were concatenated for each species, and were used for maximum likelihood phylogenetic analyses by RAxML version 8.2.4 (Stamatakis 2014) with the CAT+GTR amino acid substitution model, and 500 repetitions. Carbohydrate active enzymes (CAZymes) were searched in the Carbohydrate-active enzyme database by dbCAN2 meta server (http://bcb.unl.edu/dbCAN2/blast.php). Next, CAZymes were annotated using HMMER (E-Value < 1e^{-15}, coverage > 0.35), DIAMOND (E-Value < 1e^{-102}) and Hotpep (Frequency > 2.6, Hits > 6), respectively.

The estimated and assembled genome size was 130,685,110 bp and 142,407,839 bp, respectively (Supplementary Material Figure S1). After manual filtration, the finally obtained 125,935,854 bp genome contained 8,357 scaffolds (>100bp) having scaffold N50 of 60,253 bp (Fig. 1B). The assembled genome size was close to the estimated size (nearly 96%). Compared with published genomes of other Sphaeropleales, the genome size of *C. proboscideum* was within their range (48.9M~208Mb) (Supplementary Material Table S2). The *C. proboscideum* genome size is somewhat larger than that of the second member of the subfamily

Coelastroideae, *H. reticulata,* whose draft genome was recently assembled (Xu et al. 2019). Using the Benchmarking Universal Single-Copy Orthologs (BUSCO) eukaryote database, the genome was identified to be 81.9% complete with 3.6% fragments, while 14.5 % were missing (Figure 1B). Besides, the sequencing quality and potential contaminations were also checked by analyzing GC content in10 kb sliding window (Fig. 1C). The assembly contained 40,916,197 bp known repeats and 9,868,354 bp unknown repeats, accounting for a total of 35.6% repeats in the *C. proboscideum* genome, dominated by long interspersed elements (LINE) 34,443,187 bp (24%).

Finally, we predicted a total of 16,196 protein-coding genes with an average gene length of 2,205 bp (Fig. 1B). About 71% (11,428 genes) of the gene set was aligned to the NR database, while 47% (7,526 genes), 47% (7,527 genes), 31% (8,332 genes), and 51% (11,843 genes) were aligned by KEGG, Swissprot, COG, and InterPro respectively. In the KEGG database, 7,527 genes were mapped including Cellular Processes, Environmental Information Processing, Genetic Information Processing, Human Diseases, Metabolism, and Organismal Systems. The global and overview maps mapped almost 1,786 genes, mainly corresponding to carbohydrate metabolism (631 genes), and 547 genes were found to be involved in translation (Fig. 1D).

A phylogenomic tree inferred from a concatenated alignment of 111 nuclear-encoded, single copy genes supported the position of *C. proboscideum* in the family Scenedesmaceae as sister to *H. reticulata*, both in subfamily Coelastroideae (Fig. 2A).

To further compare *C. proboscideum* with other algae, we generated five species gene family clustering including two Scenedesmaceae (*Desmodesmus costato-granulatus*, *H. reticulata*), one Selenastraceae (*Monoraphidium neglectum*), and one Chromochloridaceae (*Chromochloris zofingiensis*) (Fig. 2B). There were 4,316 gene families commonly shared among the five algae, and 6,950 gene families were commonly shared between *C. proboscideum* and *C. zofingiensis*. With respect to the other three algae, 6,073 gene families were commonly shared between *C. proboscideum* and *D. costato-granulatus*, *H. reticulata* shared 7056, and *M. neglectum* shared 6,450 gene families (Fig. 2B). In the cluster, 4,350 genes were unique in *C. proboscideum*, most of them involved in metabolic pathways (244 genes) and biosynthesis of secondary metabolites (112 genes). The top 30 highly enriched genes in the KEGG pathway are shown in Supplementary Material Figure S2. Cell walls are key components for many algae and are important for many essential processes including development, defense against pathogens and the acclimation to environmental changes. Synthesis and degradation of cell wall oligo- and polysaccharides is facilitated by carbohydrate-active enzymes (CAZymes). In total, 158 CAZymes were identified in *C. proboscideum*, including glycoside hydrolases (GH) 63 (40%), glycosyltransferases (GT) 63 (40%), carbohydrate-binding molecules (CBM) 15 (8%), auxiliary activities (AA) 10 (6.3%), carbohydrate esterases (CE) 9 (5.7%), whereas no polysaccharide lyases (PL) were detected (Fig. 2C). The number of CAZymes was fewer than in other Scenedesmaceae: *H. reticulata* (319; Xu et al. 2019), and *D. costato-granulatus* (246; Wang et al. 2019). The CAZymes of GT (63) and GH (63), which are involved in starch and sucrose metabolism, were the most abundant CAZymes in *C. proboscideum* (Fig. 2C).

Our draft genome sequence of *C. proboscideum* strain SAG 217-2 provides insight into genomic features of a second member of subfamily Coelastroideae, a separate lineage within Scenedesmaceae (Sphaeropleales, Chlorophyceae).

Data Availability

The whole genome assemblies for *C. proboscideum* in this study are available on CNGBdb and were deposited in CNSA (https://db.cngb.org/cnsa/) under the accession number CNA0014153. Additional information of raw data and some genome information is given in Supplementary Material Table S1.

Acknowledgements

This work was supported by the Shenzhen Municipal Government of China (grant numbers No. JCYJ20151015162041454) and the Guangdong Provincial Key Laboratory of Genome Read and Write (Grant No. 2017B030301011). We thank Gerd Günther (http://www.mikroskopia.de/index.html), who took microscopic images of *Coelastrum proboscideum* strain SAG 217-2. This work is part of the 10KP project led by BGI-Shenzhen and China National GeneBank.

Author Contributions

These authors contributed equally: Hongping Liang, Yan Xu. *e-mail: michael.melkonian@uni-koeln.de; wangsibo1@genomics.cn

Declaration of Interests

The authors declare no competing interests.

References

Arora N, Tripathi S, Pruthi V, Poluri KM (2019) An Integrated Approach of Wastewater Mitigation and Biomass Production for Biodiesel Using *Scenedesmus* sp. In Gupta S, Bux F (eds) Application of Microalgae in Wastewater Treatment. Springer, Cham, pp 467-494

Carreres BM, de Jaeger L, Springer J, Barbosa MJ, Breuer G, van den End EJ, Kleinegris DMM, Schäffers I, Wolbert EJH, Zhang H, Lamers PP, Draaisma, RB, Martins dos Santos VAP, Wijffels, RH, Eggink G, Schaap PJ, Martens DE (2017) Draft genome sequence of the oleaginous green alga *Tetradesmus obliquus* UTEX 393. Genome Announc **5**:e01449-16

Cheng SF, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux, P-M, Li, F-W, Melkonian B, Mavrodiev EV, Fu, Sun WJ, Fu Y, Yang HM, Soltis DE, Graham SW, Soltis PS, Liu X, Xu X, Wong GK-S (2018) 10KP: A phylodiverse genome sequencing plan. GigaScience **7**:1-9

Dasgupta CN, Nayaka S, Toppo K, Singh AK, Deshpande U, Mohapatra A (2018) Draft genome sequence and detailed characterization of biofuel production by oleaginous microalga *Scenedesmus quadricauda* LWG002611. Biotechnol Biofuels **11**: 308

Del Campo JA, Moreno J, Rodrıíguez H, Vargas MA, Rivas J, Guerrero MG (2000) Carotenoid content of chlorophycean microalgae: factors determining lutein accumulation in *Muriellopsis* sp. (Chlorophyta). J Biotechnol **76**:51–59

Fenwick MG, Hansen LO, Lynch DL (1966) Polymorphic forms of *Coelastrum proboscideum* Bohn. Trans Am Microsc Soc **85**:579-581

Großmann E (1920) Zellvermehrung und Koloniebildung bei einigen Scenedesmaceen. Ont Rev Ges Hydrobiol Hydrogr **9**:371-394

Guiry MD, Guiry GM (2020) AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. http://www.algaebase.org; searched on 13 June 2020

Hajdu L, Hegewald E, Cronberg G (1976) Beiträge zur Taxonomie der Gattung *Coelastrum* (Chlorophyta, Chlorococeales). Ann Hist-nat Mus Nat Hung 68:31-38

Hegewald E, Wolf M, Keller A, Friedl T, Krienitz L (2010) ITS2 sequence-structure phylogeny in the Scenedesmaceae with special reference to *Coelastrum* (Chlorophyta, Chlorophyceae), including the new genera *Comasiella* and *Pectinodesmus*. Phycologia **49**:325-335

Mousavi S, Najafpour GD, Mohammadi M, Seifi MH (2018) Cultivation of newly isolated microalgae *Coelastrum* sp. in wastewater for simultaneous CO2 fixation, lipid production and wastewater treatment. Bioprocess Biosystems Eng **41**:519–530

Neofotis P, Huang A, Sury K, Chang W, Joseph F, Gabr A, Twary S, Qiu W, Holguin O, Polle JEW (2016) Characterization and classification of highly productive microalgae strains discovered for biofuel and bioproduct generation. Algal Res **15**:164-178

Rauytanapanit M, Janchot K, Kusolkumbot P, Sirisattha S, Waditee-Sirisattha R, Praneenararat T (2019) Nutrient deprivation-associated changes in green microalga

Coelastrum sp. TISTR 9501RE enhanced potent antioxidant carotenoids. Mar Drugs **17**:328

Ribeiro DM, Minillo A, Silva CAA, Fonseca GG (2019) Characterization of different microalgae cultivated in open ponds. Acta Scientiarum Technol **41**:e37723

Sahu SK, Thangaraj M, Kathiresan K (2012) DNA extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. ISRN Mol Biol **2012**:205049

Shuba ES, Kifle D (2018) Microalgae to biofuels: 'Promising' alternative and renewable energy, review. Renewable Sustain Energy Rev **81**:743–755

Soares AT, da Costa DC, Vieira AAH, Antoniosi Filho NR (2019) Analysis of major carotenoids and fatty acid composition of freshwater microalgae. Heliyon **5**:e01529

Starkenburg SR, Polle JEW, Hovde B, Daligault HE, Davenport KW, Huang A, Neofotis P, McKie-Krisberg Z (2017). Draft nuclear genome, complete chloroplast genome, and complete mitochondrial genome for the biofuel/bioproduct feedstock species *Scenedesmus obliquus* strain DOE0152z. Genome Announc **5**: e00617-17

Stivrins N, Soininen J, Tõnnod I, Freiberg R, Veskie S, Kisand V (2019) Towards understanding the abundance of non-pollen palynomorphs: A comparison of fossil algae, algal pigments and sedaDNA from temperate lake sediments. Rev Paleobot Palynol **249**:9-15

Úbeda B, Gálvez JA, Michel M, Bartual A (2017) Microalgae cultivation in urban wastewater: *Coelastrum* cf. *pseudomicroporum* as a novel carotenoid source and a potential microalgae harvesting tool. Bioresour Technol **228**:210–217

Wang S, Li L, Xu Y, Melkonian B, Lorenz M, Friedl T, Sonnenschein E, Liu H, Melkonian M (2019) The draft genome of the small, spineless green alga *Desmodesmus costato-granulatus* (Sphaeropleales, Chlorophyta)**.** Protist **170**:125697

Wang ZK, He LJ, Hu F, Lin XZ (2017) Characterization of the complete mitochondrial genome of *Coelastrum* sp. F187. Mitochondrial DNA Part B **2**:455-456

Han Y, Wessler S R (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids **38**:199–199

Ellinghaus D, Kurtz S, Willhoef U (2008) LTRharvest, an efcient and fexible sofware for de novo detection of LTR retrotransposons. BMC Bioinform 9:18

Chen N (2004) Using repeatmasker to identify repetitive elements in genomic sequences Curr Protoc Bioinformatics **5**:4–10

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:435–439

Hoff K J, Lange S, Lomsadze A, Borodovsky M, Stanke M (2016) BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinfom **32：**767-769

Besemer J, Borodovsky M (2005) GeneMark: web sofware for gene fnding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res **33**:451–454

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol **35**:543-548

Guillaume M and Carl K (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**:764-770

Katoh K, Misawa K, Kuma K, Miyataa T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res **30**:3059–3066

Emms D. M and Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol **20**:1-14

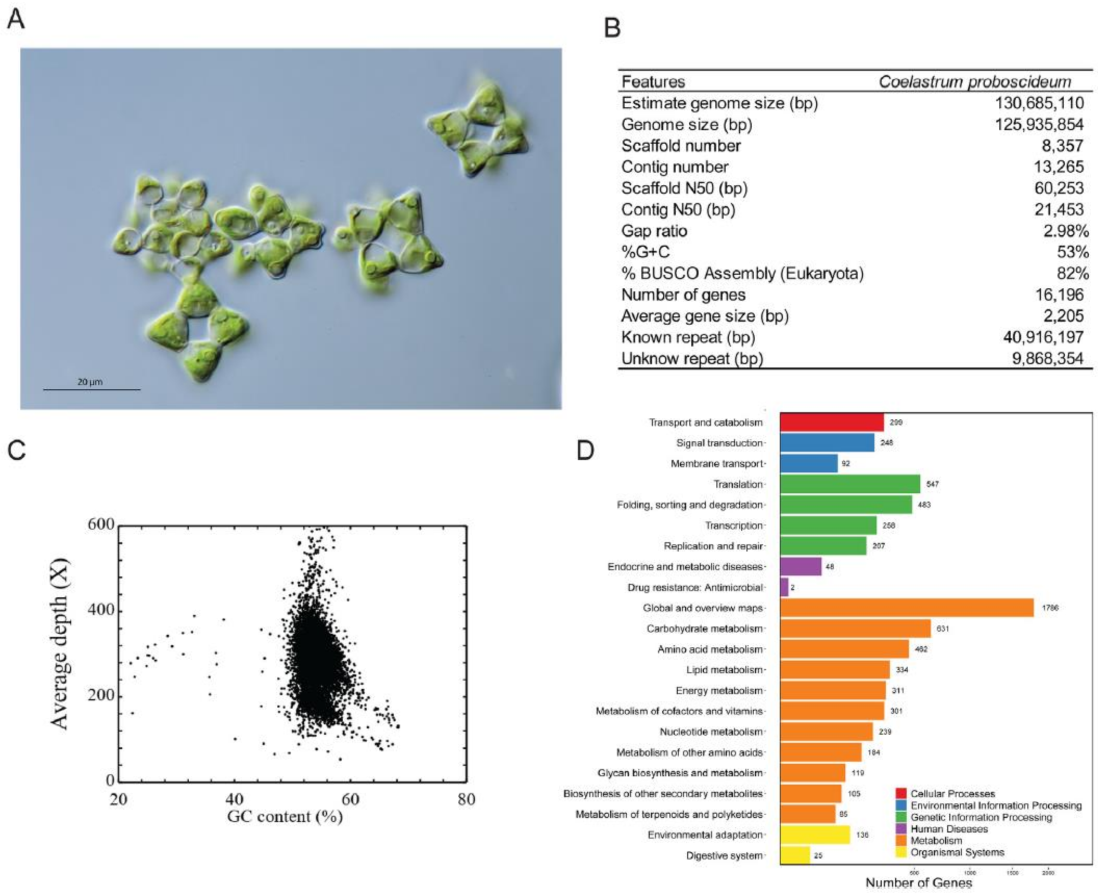
Stamatakis, A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**:1312-1313

Vurture, GW, Sedlazeck, FJ, Nattestad, M, Underwood, CJ, Fang, H, Gurtowski, J, Schatz, MC (2017) GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics **33**:2202–2204

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB (2017) Direct determination of diploid genome sequences. Genome Res **27**:757-767

Xu Y, Li L, Liang H, Melkonian B, Lorenz M, Friedl T, Petersen M, Liu H, Melkonian M, Wang S (2019) The draft genome of *Hariotina reticulata* (Sphaeropleales, Chlorophyta) provides insight into the evolution of Scenedesmaceae. Protist **170**:125684

Figure Legends



Features	*Coelastrum proboscideum*
Estimate genome size (bp)	130,685,110
Genome size (bp)	125,935,854
Scaffold number	8,357
Contig number	13,265
Scaffold N50 (bp)	60,253
Contig N50 (bp)	21,453
Gap ratio	2.98%
%G+C	53%
% BUSCO Assembly (Eukaryota)	82%
Number of genes	16,196
Average gene size (bp)	2,205
Known repeat (bp)	40,916,197
Unknow repeat (bp)	9,868,354

Figure 1. *C. proboscideum* morphology and genome assembly. (**A**) Light micrograph (Nomarski Interference Contrast) of *C. proboscideum* SAG 217.2 (**B**) Statistics of the *C. proboscideum* genome assembly and annotations. (**C**) GC-depth plot showing the distribution between the GC content and the average reads mapping depth. (**D**) KEGG pathway mapping of *C. proboscideum* coding-proteins.

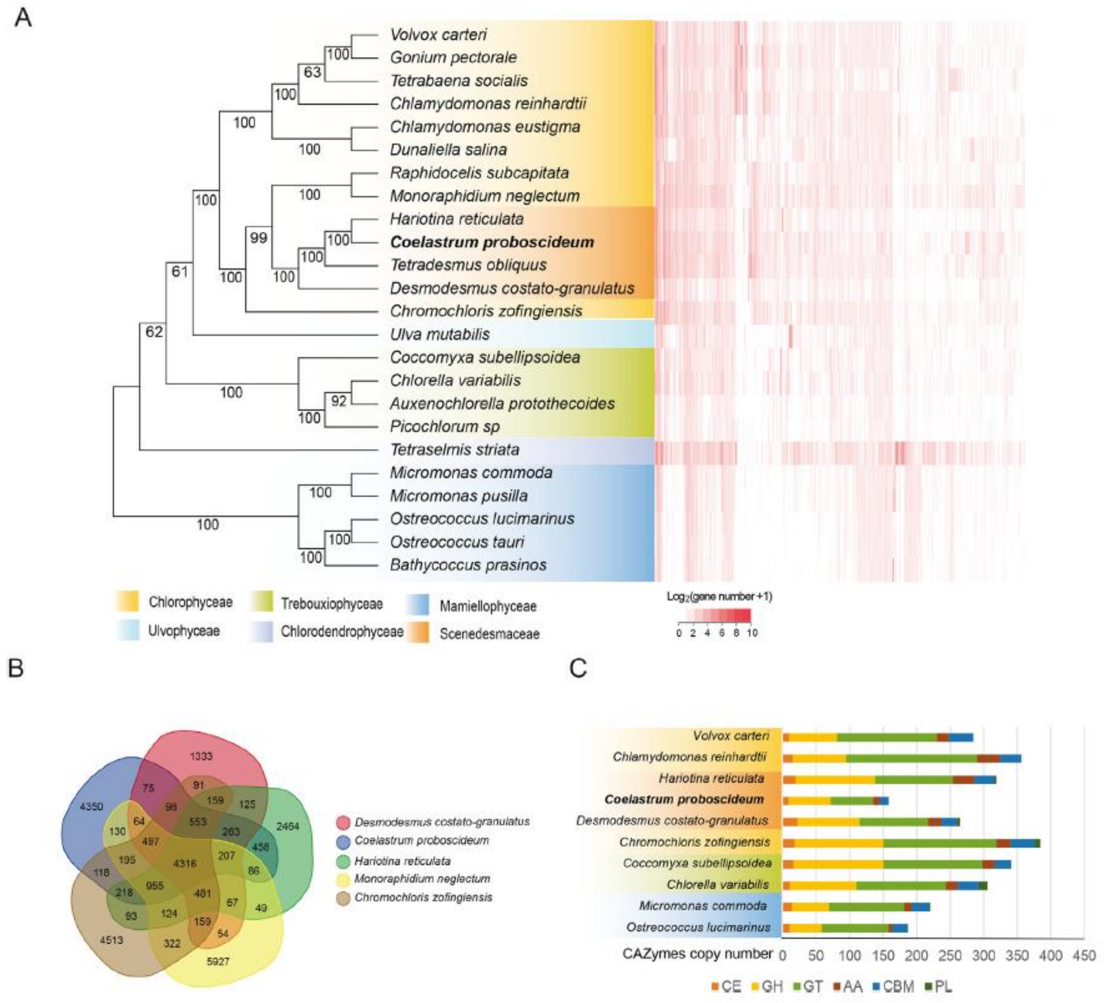


Figure 2. Evolutionary analysis of *C. proboscideum* in comparison with other selected green algae. (**A** The phylogenetic tree was constructed using the maximum-likelihood method by RAxML based on a concatenated sequence alignment of 111 single-copy genes with 500 bootstrap iterations. The *C. proboscideum* was in bold. The bootstraps were show in each branch, while ignored branch length. A k-means clustering of gene families based on the gene abundance of each species is shown in the right panel; each column represents the copy number of families and each row represents one species. (**B**) Venn diagrams showing the number of gene families shared among 5 algae, including *Coelastrum proboscideum*, *Desmodesmus costato-granulatus, Hariotina reticulata, Monoraphidium neglectum* and *Chromochloris zofingiensis*. (**C**) CAZymes distribution in different algae: GTs (glycosyltransferases), GHs (glycoside hydrolases), PLs (polysaccharide lyases), CEs (carbohydrate esterases), AAs (enzymes of the auxiliary activities), and CBMs (carbohydrate-binding modules).

Legends to Supplementary Material Figures and Tables

Figure S1. The kmer distribution of *C. proboscideum* in the genome size estimate.

The K-mer distribution diagram of BGI-500 paired-end reads using GenomeScope based on k value of 21. K-mer coverage (x axis) was plotted against each frequency (y axis).

Figure S2. KEGG enrichment scatter plot of *C. proboscideum* unique genes.

The x axis represents the Q-value, and y axis represents the name of the pathway. Dot sizes represent the copy number of different genes and the color indicates the Q-value.

Supplementary Material Table S1:

Information of raw Linked-Reads.

Supplementary Material Table S2:

Information on genome sizes and gene set of algal species used in this study.