Protist, Vol. 170, 125684, December 2019
http://www.elsevier.de/protis
Published online date 19 October 2019

Protist

ORIGINAL PAPER

Protist Genome Reports

The Draft Genome of *Hariotina reticulata* (Sphaeropleales, Chlorophyta) Provides Insight into the Evolution of Scenedesmaceae

Yan Xu[a,b,2], Linzhou Li[c,d,2], Hongping Liang[b], Barbara Melkonian[e], Maike Lorenz[f], Thomas Friedl[f], Morten Petersen[g], Huan Liu[a,g], Michael Melkonian[e,1], and Sibo Wang[a,g,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]Department 'Experimentelle Phykologie und Sammlung von Algenkulturen' (EPSAG), University of Göttingen, Nikolausberger Weg 18, 37073 Göttingen, Germany
[g]Department of Biology, University of Copenhagen, Copenhagen, Denmark

Submitted October 9, 2019; Accepted October 13, 2019

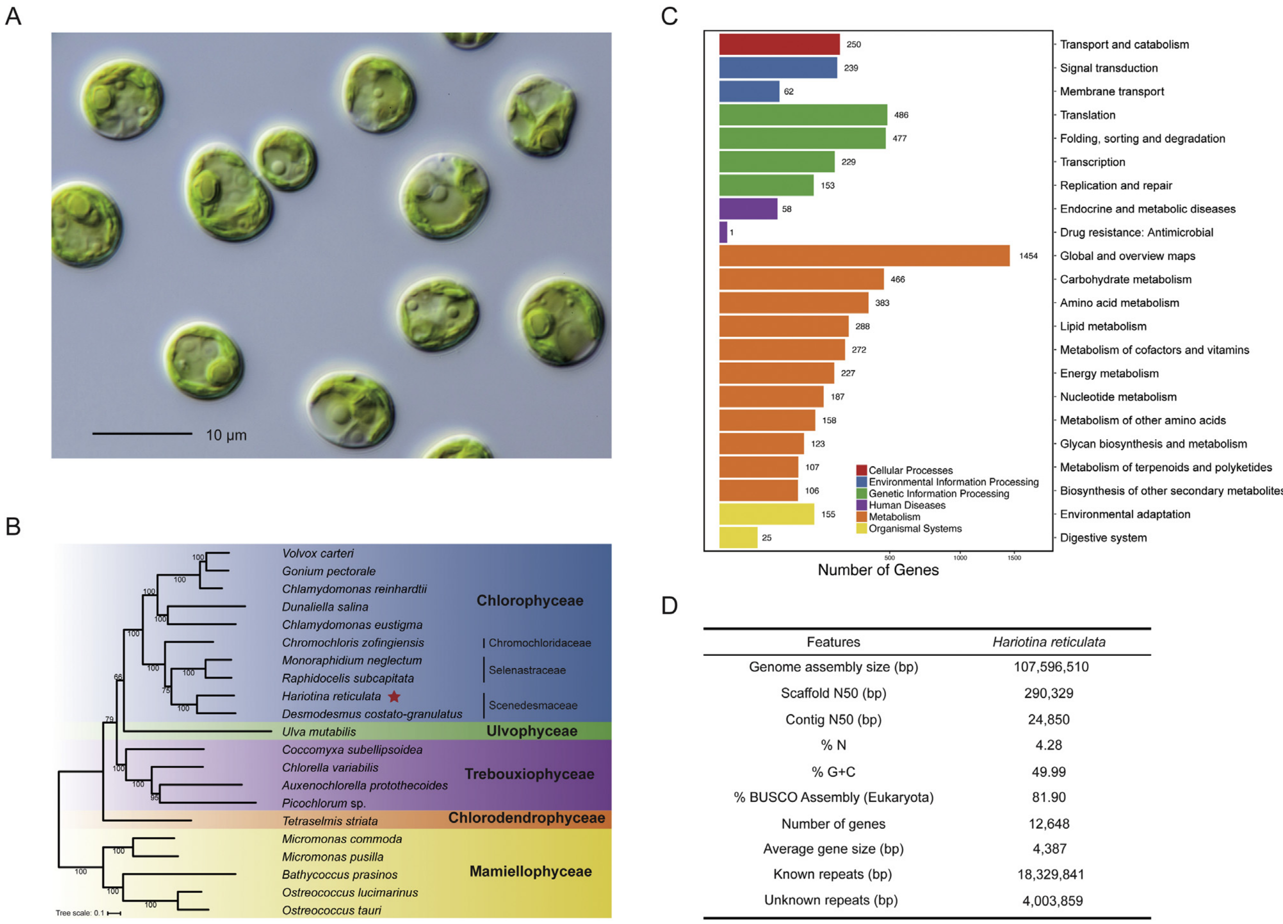
Hariotina reticulata **P. A. Dangeard 1889 (Sphaeropleales, Chlorophyta) is a common member of the summer phytoplankton of meso- to highly eutrophic water bodies with a worldwide distribution. Here, we report the draft whole-genome shotgun sequencing of *H. reticulata* strain SAG 8.81. The final assembly comprises 107,596,510 bp with over 15,219 scaffolds (>100 bp). This whole-genome project is publicly available in the CNSA (https://db.cngb.org/cnsa/) of CNGBdb under the accession number CNP0000705.**

Key words: Scenedesmaceae; genome; algae; comparative genomics.

The Scenedesmaceae Oltmanns 1904 is the largest family in the order Sphaeropleales (Chlorophyta) with over 300 described species (Guiry and Guiry 2019) and contains such well-known genera as *Coelastrum*, *Desmodesmus* and *Tetradesmus*, the latter two formerly in genus *Scenedesmus* (Krienitz and Bock 2012). The Scenedesmaceae are common constituents of freshwater phytoplankton, and because of their rapid growth and high lipid contents they are intensively studied as potential sources of biofuels and in wastewater treatment

[1]Corresponding authors;
[2]Equal contribution.
e-mails michael.melkonian@uni-koeln.de (M. Melkonian), wangsibo1@genomics.cn (S. Wang).

https://doi.org/10.1016/j.protis.2019.125684



Features	*Hariotina reticulata*
Genome assembly size (bp)	107,596,510
Scaffold N50 (bp)	290,329
Contig N50 (bp)	24,850
% N	4.28
% G+C	49.99
% BUSCO Assembly (Eukaryota)	81.90
Number of genes	12,648
Average gene size (bp)	4,387
Known repeats (bp)	18,329,841
Unknown repeats (bp)	4,003,859

Figure 1. Genome Features and phylogenomic analysis of *H. reticulata*. (**A**) Light micrograph (Nomarski Interference Contrast) of *H. reticulata*. Please note that no coenobia (only single cells and occasionally sporangia) were observed in this axenic strain under the culture conditions used in this study. (**B**) The phylogenetic tree was constructed by using the maximum-likelihood method in RAxML based on a concatenated sequence alignment of 142 single-copy genes (numbers on branches refer to % bootstrap values; in total 500 bootstrap iterations). (**C**) KEGG pathway mapping of annotated *H. reticulata* genes. (**D**) The table displays the summary statistics of the *H. reticulata* genome.

(Brennan and Owende 2010; Pittman et al. 2011; Shuba and Kifle 2018). Despite their ecological and biotechnological importance, only scarce genomic resources have been provided in the Scenedesmaceae (five genomes from *Desmodesmus* and *Tetradesmus* spp.). Originally confined to flat or curved coenobia, consisting of ovate- to spindle-shaped cells, Komárek and Fott (1983) expanded the family to include also genera with three-dimensional coenobia. Although several of these genera have subsequently been transferred to the class Trebouxiophyceae, molecular phylogenetic analyses presented 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 (Coelastroidea). This clade consists of the morphologically diverse genera *Coelastrum*, *Hariotina*, *Asterarcys*, *Coelastrella*, and *Dimorphococcus*. Although *Hariotina* P.A. Dangeard, because of superficial resemblance of the coenobia with those of *Coelastrum*, had been transferred to *Coelastrum* (as *C. reticulatum*) by Senn (1899), it was not monophyletic with *Coelastrum* and its type species *C. sphaericum* (Hegewald et al. 2010). This finding corroborated earlier observations reported by Hegewald et al. (2002) and Krienitz et al. (2003) who had also argued for the separation of genus *Hariotina* from *Coelastrum*. *Hariotina reticulata* (Fig. 1A) is a common member of the summer phytoplankton in eutrophic water bodies with a worldwide distribution (e.g. Munawar et al. 2017; Zohary et al. 2017). The draft nuclear genome of *H. reticulata* (strain SAG 8.81) represents the first nuclear genome sequence from subfamily Coelastroidea and thus from a Scenedesmaceae with three-dimensional coenobia. A chloroplast genome sequence from another strain of the same species has recently been reported (He et al. 2018). The draft nuclear genome sequence of *H. reticulata* has been established in the frame of the 10 KP project (Cheng et al. 2018), a phylodiverse genome sequencing plan.

An axenic culture of *H. reticulata* (SAG 8.81; 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 hr L/D cycle up to a volume of 1,000 mL. The dense culture was harvested by centrifugation (1,000 × *g*, 10 min), the pellet rapidly transferred into liquid nitrogen and stored at −80 °C until freeze-drying. DNA extraction was done on the freeze-dried samples using the CTAB method (Sahu et al. 2012). During all steps of cultivation until nucleic acid extraction, 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.

The genomic DNA was used to construct a 10X genomics library, and sequenced by the BGISEQ-500 platform. Paired-end reads of 100bp in length were generated using the BGISEQ-500 sequencer. For assembly of the 10X Genomics Chromium library data, the raw reads were directly used for genome assembly by using the Supernova software (v2.1.0) according to the manufacturer's protocol (Weisenfeld et al. 2017).

Three types of repeats (DNA transposon elements, retrotransposon elements, tandem repeats) were identified in the genome of *H. reticulata*: DNA transposons and retrotransposon elements were identified using LTRharvest. RepeatModeler (version 1.0.8; http://www.repeatmasker.org/RepeatModeler/) was used to search for other repeats using the *denovo* approach. Alternatively, RepeatMasker (http://www.repeatmasker.org) was applied to using a custom library comprising a combination of Repbase and a de novo predicted repetitive element library.

For gene prediction, we used de novo gene prediction methods. Briefly, the program GeneMark-ES (V4.32) was used to analyze the genome by utilizing unsupervised training and the genes having an incomplete gene structure were removed from the predicted gene sets. The final gene set was evaluated with two approaches: The BUSCO (V. 3; Waterhouse et al. 2018) core eukaryotic gene-mapping approach was used to determine the gene set completeness. Gene functional annotation was performed by BLASTP (1e-5) against several known databases, such as SwissProt, KEGG (plant database), COG, and NR (plant database). InterProScan (using data from Pfam, PRINTS, SMART, ProDom and PROSITE) was used to identify protein motifs and protein domains of the predicted gene set. Gene Ontology information was obtained through Blast2go (version 2.5.0).

The genomes of *H. reticulata* were compared with selected algal genomes. These algal genomes were used to define orthogroups (using OrthoFinder, version 1.1.8). Single copy orthogroups (i.e. gene families with only one gene member per species) were used to construct phylogenetic trees based on maximum likelihood. RAxML was used for phylogenetic tree construction with the CAT + GTR amino acid substitution model. The online tool iTOL was applied to edit and display the final phylogenetic tree.

We performed whole-genome shotgun sequencing of genomic DNA using a 10X genomics library and sequenced by the BGISEQ-500 platform, resulting in a 1,564 X genome coverage sequencing data for *H. reticulata*. The *H. reticulata* genome sequence was assembled into 15,219 scaffolds (>100 bp). The assembled genome size was 107,596,510 bp, and covered 100% of the predicted genome size (107,326,711 bp) with a scaffold N50 reaching 291 kb (Fig. 1D). Analyses of genome completeness indicated that the genome assembly captured at least 81.9% of the eukaryotic BUSCO dataset. Repetitive elements accounted for 21.88% of the genome. Among the identified repeats, 17.1% were classified as known or reported repeat families, with long terminal repeats (LTRs) and long interspersed elements (LINEs) being predominant, representing 15.5 Mb and 4.9 Mb, respectively (Supplementary Material Table S1).

We predicted 12,648 protein-coding genes, of which 75.77% (9,583) were assigned putative functions with the NR plant database and 55.43% (6,477) with the Swissprot database. The *H. reticulata* genome size is similar to that of *Desmodesmus armatus* (115Mb;

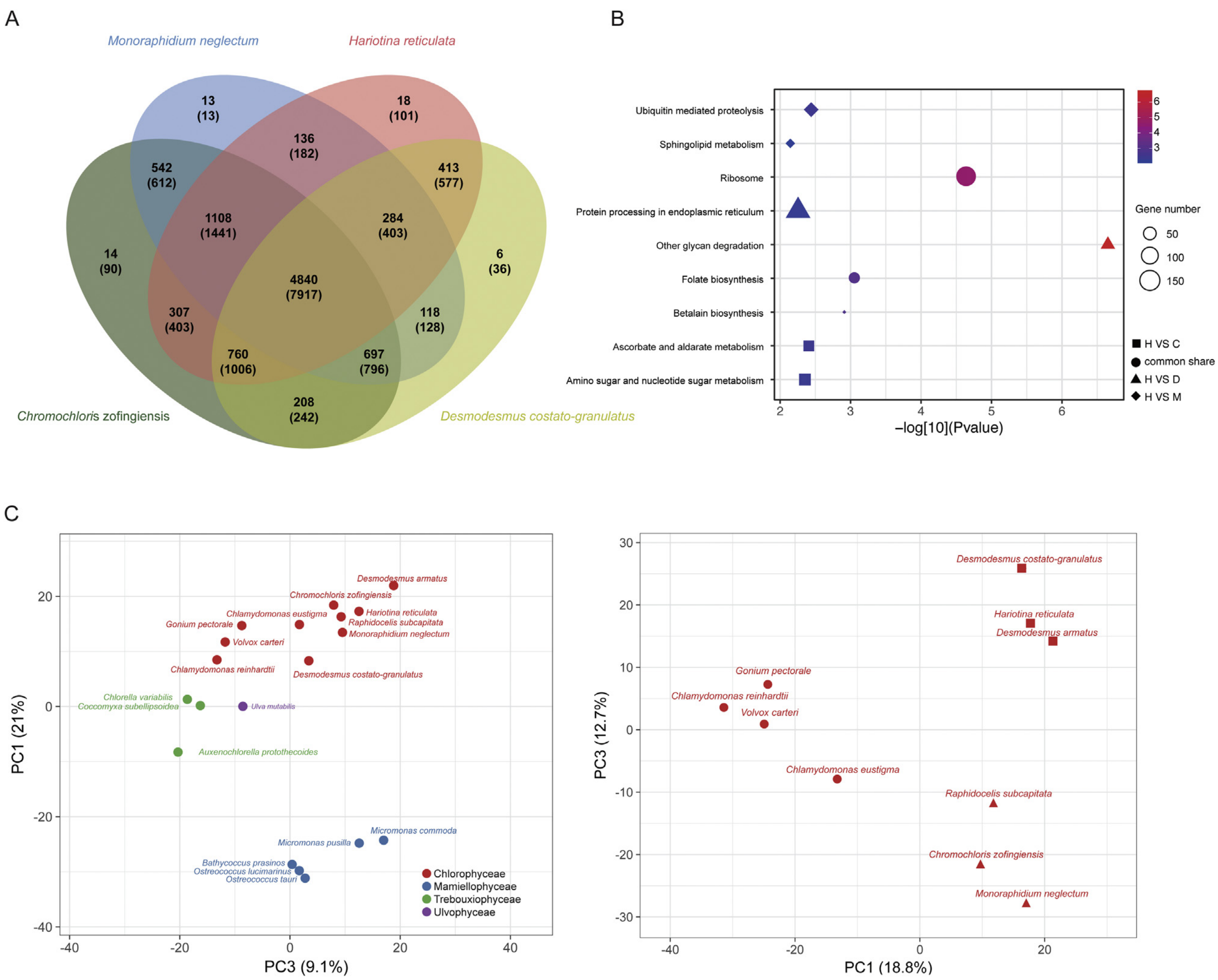


Figure 2. Comparative analysis of the genomes of *H. reticulata* and other Sphaeropleales. (**A**) Venn diagram showing unique and shared orthogroups among *H. reticulata*, *Chromochloris zofingiensis*, *Monoraphidium neglectum* and *Desmodesmus costato-granulatus*. Gene numbers are given in parentheses. (**B**) The KEGG pathways enriched in genes shared among four Sphaeropleales or unique in the *H. reticulata* genome. (H VS C: genes unique in *H. reticulata* compared to *C. zofingiensis*; common share: genes in *H. reticulata* shared among four Sphaeropleales; H VS D: genes unique in *H. reticulata* compared to *D. costato-granulatus*; H VS M: genes unique in *H. reticulata* compared to *M. neglectum*.) (**C**) Principal components analysis (PCA) of the type and number of Pfam domains of all genes across selected Chlorophyta. (**D**) PCA of the type and number of Pfam domains of all genes across selected Chlorophyceae.

https://www.ncbi.nlm.nih.gov/assembly/GCA_007449985.1) and *Desmodesmus obliquus* (108Mb; Carreres et al. 2017) but considerably larger than that of the small spineless *Desmodesmus costato-granulatus* (Wang et al. 2019), a "*Scenedesmus quadricauda*" LWG002611, presumably *D. communis* (65.35 Mb; Dasgupta et al. 2018), and some sequenced Selenastraceae (*Chromochloris zofingiensis, Monoraphidium neglectum,* and *Raphidocelis subcapitata*).

A phylogenomic tree inferred from a concatenated alignment of 142 nuclear-encoded, single copy genes supported the position of *H. reticulata* in the Sphaeropleales and family Scenedesmaceae as sister to *D. costato-granulatus* (Fig. 1B). Based on KEGG pathway mapping, we annotated and mapped 237 level-3 pathways for 5,906 genes which belong to 22 level-2 metabolic pathways. A summary of the findings is presented in Figure 1C and Supplementary Material Table

S2. The largest number of sequences were those associated with Global and overview maps (1,454, 18.29%), followed by sequences that were involved in translation (486, 6.11%), "Folding, sorting and degradation" (477, 6.00%), and carbohydrate metabolism (466, 5.86%). About 4,840 gene families were commonly shared among *H. reticulata* and three other Sphaeropleales used in our genome comparisons (Fig. 2A). Of the 101 genes unique to *H. reticulata*, most are involved in RNA degradation, phenylpropanoid biosynthesis and ascorbate and aldarate metabolism. To further compare the gene-complement between *H. reticulata* and three Sphaeropleales, we performed KEGG enrichment of unique genes between *H. reticulata* and each Sphaeropleales (Fig. 2B). 2,942 genes were unique to *H. reticulata* when performing the gene family clustering between *H. reticulata* and *D. costato-granulatus*, and these unique genes were enriched in biological process categories such as "other glycan degradation" and "protein processing in endoplasmic reticulum". 2,235 unique genes could be identified when comparing the gene family clustering between *H. reticulata* and *M. neglectum*, and according to the KEGG enrichment display, these unique genes mainly participate in betalain biosynthesis, sphingolipid metabolism and ubiquitin mediated proteolysis. Gene family clustering between *H. reticulata* and *C. zofingiensis* shows 1,122 genes unique to *H. reticulata*, which are highly enriched in biological process categories such as ascorbate and aldarate metabolism, and amino sugar and nucleotide sugar metabolism (Fig. 2B).

To characterize the patterns of functional diversification in gene domains among Sphaeropleales and other Viridiplantae, principal component analysis (PCA) was used to analyze the number of Pfam domains in genomes among selected Viridiplantae (Fig. 2C). The results show that the type and number of conserved Pfam domains of all genes in Sphaeropleales are close to other Chlorophyceae. Interestingly, further PCA analysis among Chlorophyceae revealed that Scenedesmaceae, Selenastraceae, and Volvocaceae form three separate clusters, which suggests that although Scenedesmaceae, Selenastraceae and Volvocaceae all belong to class Chlorophyceae, they may display separate functional diversification in gene domains (Fig. 2D).

Our draft genome sequence of *H. reticulata* strain SAG 8.81 provides the first insight into genome features of a member of subfamily Coelastroidea, a separate lineage within Scenedesmaceae.

Data Availability

The whole genome assemblies for *H. reticulata* established in this study are available on CNGBdb and were deposited in CNSA (https://db.cngb.org/cnsa/) under the accession number CNP0000705.

Acknowledgements

We thank Gerd Günther (http://www.mikroskopia.de/index.html), who took microscopic images of *Hariotina reticulata* strain SAG 8.81. This work is part of the 10KP project led by BGI-Shenzhen and China National Gene Bank (CNGB).

Appendix A. Supplementary Data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.protis.2019.125684.

References

Brennan L, Owende P (2010) Biofuels from microalgae—A review of technologies for production, processing, and extractions of biofuels and co-products. Renewable Sustain Energy Rev **14**:557–577

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, 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)** (2018)

Guiry MD, Guiry GM (2019) AlgaeBase. World-wide electronic publication, National University of Ireland, Galway, searched on 23 September 2019 http://www.algaebase.org

He LJ, Wang ZK, Lou SL, Lin XZ, Hu F (2018) The complete chloroplast genome of the green algae *Hariotina reticulata* (Scenedesmaceae, Sphaeropleales, Chlorophyta). Genes Genomics **40**:543–552

Hegewald E, Coesel PFM, Hegewald P (2002) A phytoplankton collection from Bali, with the description of a new

Desmodesmus species (Chlorophyta, Scenedesmaceae). Arch Hydrobiol Suppl Algol Stud **105**:51–78

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

Komárek J, Fott B (1983) Chlorophyceae (Grünalgen). Ordnung: Chlorococcales. In Huber-Pestalozzi (ed) Das Phytoplankton des Süßwassers. Systematik und Biologie 7. Teil, 1. Hälfte. E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart, 1044 p

Krienitz L, Bock C (2012) Present state of the systematics of planktonic coccoid green algae of inland waters. Hydrobiologia **698**:295–326

Krienitz L, Hegewald E, Hepperle D, Wolf M (2003) The systematics of coccoid green algae: 18S rRNA gene sequence data versus morphology. Biologia **58**:437–446

Munawar M, Fitzpatrick M, Niblock H, Kling H, Rozon R, Lorimer J (2017) Phytoplankton ecology of a culturally eutrophic embayment: Hamilton Harbour, Lake Ontario. Aquat Ecosyst Health Management **20**:201–213

Pittman JK, Dean AP, Osundeko O (2011) The potential of sustainable algal biofuel production using wastewater resources. Bioresour Technol **102**:17–25

Senn G (1899) Über einige coloniebildende einzellige Algen. Bot Zeitung **57**:39–105

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

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

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

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

Zohary T, Fishbein T, Shlichter M, Naselli-Flores L (2017) Larger cell or colony size in winter, smaller in summer – a pattern shared by many species of Lake Kinneret phytoplankton. Inland Waters **7**:200–209