Protist 172 (2021) 125845
http://www.elsevier.de/protis
Published online date 16 November 2021

Protist

ORIGINAL PAPER

*Protist Genome Reports*

# The Draft Genome of the Centric Diatom *Conticribra weissflogii* (Coscinodiscophyceae, Ochrophyta)

**Linzhou Li**[a,b,1], **Hongli Wang**[b,c,1], **Sibo Wang**[b], **Yan Xu**[b,c], **Hongping Liang**[b,c], **Huan Liu**[b], **and Eva C. Sonnenschein**[a,2]

[a]Department of Biotechnology and Biomedicine, Technical University of Denmark, Søltofts Plads 221, 2800 Kgs. Lyngby, Denmark
[b]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China
[c]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

Submitted September 6, 2021; Accepted November 10, 2021

**Here, we present a 231 Mb draft genome of the centric diatom *Conticribra weissflogii* CCMP1336. Comparative genomics of *C. weissflogii* and other Ochrophyta support the existence of unique carbon-concentrating mechanisms and chitin metabolic processes in diatoms. The whole-genome project is available at CNSA (https://db.cngb.org/search/project/CNP0001903/).**

**Key words:** Diatom; genomics; photosynthesis; carbon-concentrating mechanism; chitin synthases.

The single-celled eukaryotic diatoms belonging to the Ochrophyta are crucial for global carbon fixation similar to green plants (Falkowski et al. 1998). While the high photosynthetic capacity of $C_4$ $CO_2$-concentrating mechanisms is widespread among plants, and was initially believed to be absent from unicellular algae (Hopkinson et al. 2016), it was identified in some diatoms, such as *Thalassiosira pseudonana* (also referred to *Cyclotella pseudonana*) (Kustka et al. 2014) and *Conticribra weissflogii* (Reinfelder 2011) (previously known as *Thalassiosira weissflogii*) (Stachura-Suchoples and Williams 2009). The diatom-specific $C_4$ pathway appears unconventional in comparison to land plants, but the precise mechanisms remain to be resolved (Hopkinson et al. 2016).

Another unique characteristic of diatoms in comparison to other Ochrophyta and land plants are their silicified cell walls (Martin-Jézéquel et al. 2000). Accordingly, they are central to global biosili-

[2]Corresponding author;
e-mail evaso@bio.dtu.dk (E.C. Sonnenschein).
[1]These authors contributed equally.

https://doi.org/10.1016/j.protis.2021.125845
1434-4610/© 2021 The Author(s). Published by Elsevier GmbH.

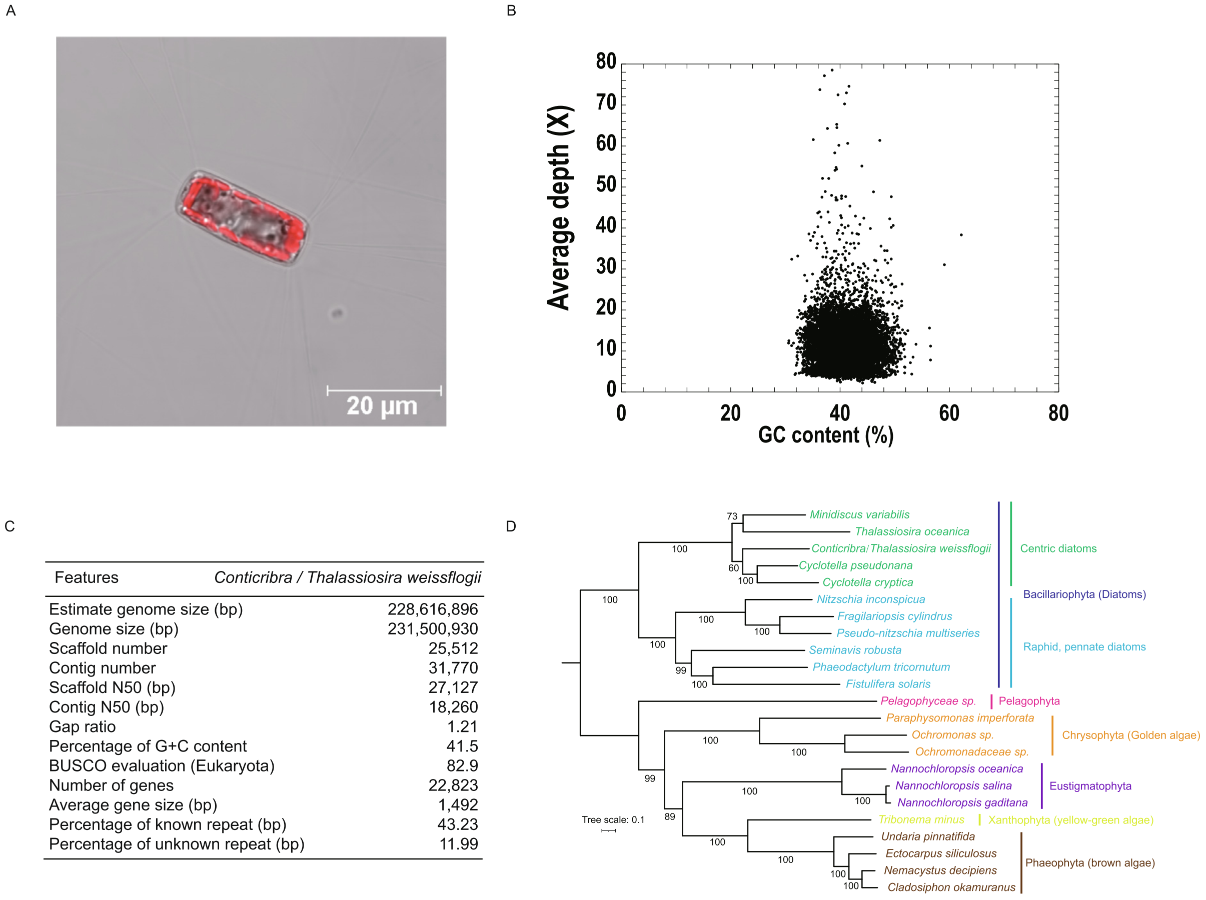
cification (Mock et al. 2008) and the diatom frustules have gained attention for industrial applications (Rabiee et al. 2021). However, the exact composition of the diatom cell wall remains difficult to decipher. In fungi, chitin is a key component of the cell wall, synthesized by chitin synthases (CHSs). It has been suggested that chitin also plays an important role in the cell wall of diatoms such as *T. pseudonana* and *C. weissflogii* (Cheng et al. 2021; Durkin et al. 2009). Interestingly, chitin is generally considered only to be present in centric rather than pennate diatoms (Durkin et al. 2009), however, CHSs can be found in both groups (Shao et al. 2019).

The centric diatom *C. weissflogii* serves as model system to study these environmentally important processes (Sonnenschein et al. 2021) and is a nutrient-rich and robust resource of renewable biomass for energy, food and feed. Although a full-length transcriptome of *C. weissflogii* has recently been released (Cheng et al. 2021), genomic data are still lacking. Here, we report the draft genome sequence of *C. weissflogii* strain CCMP1336 and take a glance at the unique features of the *C. weissflogii* genome in comparison to other Ochrophyta genomes including the $CO_2$-concentrating mechanism and the evolutionary origin of diatom CHSs.

An axenic culture of *C. weissflogii* CCMP1336 (Fig. 1A) was obtained from NCMA (National Center for Marine Algae and Microbiota, East Boothbay, Maine, USA). *C. weissflogii* was cultivated in 1.5 L f/2 (Guillard and Ryther 1962, 1975) culture medium prepared with 3% Instant Ocean (IO, Aquarium Systems Inc., Sarrebourg, France) in a 2 L Erlenmeyer flask with sterile aeration and incubation at 80 µmol photons $m^{-2}$ $s^{-1}$ and 18 °C for 8 days. The cells were pelleted by centrifugation (4000 × *g*, 10 min and 4 °C), frozen in liquid nitrogen and stored at −80 °C until freeze-drying.

Long fragment DNA was extracted using a modified cetyltrimethylammonium bromide protocol followed by stLFR (single-tube long fragment reads) library construction. 150 bp paired-end reads were generated using MGI-SEQ. Data noise reduction was performed to remove low-quality and duplicated reads using SOAPfilter (v2.2) with default parameters. Then, the newly constructed reads were assembled into scaffolds using Supernova (v2.1.1) according to the manufacturer's protocol. The 231 Mb draft genome assembly of *C. weissflogii* CCMP1336 (accession number CNP0001903 at CNSA https://db.cngb.org/cnsa/) comprised 31,770 scaffolds with a scaffold N50 of 27,127 bp (Fig. 1C). The final, assembled genome size was similar to the estimated genome size (228 Mb) as predicted by GenomeScope 2.0 (Ranallo-Benavidez et al. 2020). Although the assembly is fragmented, it is more complete (contig N50 = 18.3 kb) than previously published genomes of centric diatoms, for example, *Cyclotella cryptica* (contig N50 = 12 kb) and *Thalassiosira oceanica* (contig N50 = 3.6 kb) (Lommer et al. 2012; Traller et al. 2016). BUSCO evaluation (Waterhouse et al. 2018) showed that the assembly shared 76% core eukaryote genes, which was higher than the genomic assemblies of other centric diatoms (Supplementary Material Table S1). The evaluation of the GC-depth distribution demonstrated a good quality of the genome assembly based on completeness and accuracy (Fig. 1B).

Transposon elements were identified by employing two strategies. For novel repeats, PILER (Edgar and Myers 2005) and RepeatScout (Price et al. 2005) were used to search for DNA transposons and LTR_finder (Xu and Wang 2007) was used to search for retrotransposons. For known repeats, RepeatMasker (www. RepeatMasker.org) and ProteinMask were used to search repeats based on the existing database. All candidates were combined into a custom library, and the assembly was examined again based on the custom library. Finally, a total of 54% transposon elements (TEs) were identified, which is similar to the TE content of *C. cryptica* (53% of the genome). LTRs (long terminal repeats) and DNA transposon elements comprised approximately 38% and 5% of the genome, respectively. The most common LTR was of the Ty1-*copia* group, which accounted for 80% of the total LTR. For gene annotation, the BRAKER2 (Brůna et al. 2021) pipeline was used to automatically train and annotate gene models. The pipeline combined evidence of self-training models with homologous protein information to obtain more accurate gene models. A total of 22,823 gene models were predicted (Fig. 1C). The gene models of *C. weissflogii* showed similar lengths and numbers as in other centric diatoms, which implies a reliable annotation (Supplementary Material Table S2). To further evaluate the gene models, BLASTP was used to align them against several protein databases (SwissProt, KEGG, KOG and TrEMBL). InterProScan was performed to identify protein domains. As a result, 80% of the gene models could be aligned to the known protein and domain databases.



| Features | *Conticribra / Thalassiosira weissflogii* |
|---|---:|
| Estimate genome size (bp) | 228,616,896 |
| Genome size (bp) | 231,500,930 |
| Scaffold number | 25,512 |
| Contig number | 31,770 |
| Scaffold N50 (bp) | 27,127 |
| Contig N50 (bp) | 18,260 |
| Gap ratio | 1.21 |
| Percentage of G+C content | 41.5 |
| BUSCO evaluation (Eukaryota) | 82.9 |
| Number of genes | 22,823 |
| Average gene size (bp) | 1,492 |
| Percentage of known repeat (bp) | 43.23 |
| Percentage of unknown repeat (bp) | 11.99 |

**Figure 1.** The morphology and genome assembly statistics of *C. weissflogii.* (**A**) Light micrograph of *C. weissflogii* imaged with a Nikon Inverted Fluorescence Microscope-EclipseTi2. The chloroplasts are indicated by the red color. **(B)** GC-depth plot showing the distribution of the GC content and average reads mapping. **(C)** Statistics of genome assembly and gene models of *C. weissflogii*. **(D)** The phylogenetic tree of Ochrophyta genomes constructed using the maximum-likelihood method by RAxML based on a concatenated method of 201 single-copy genes with 500 bootstrap replicates.

To improve the evolutionary understanding of Thalassiosirales, OrthoFinder was applied on published genomes of Ochrophyta and *C. weissflogii* and a phylogenetic tree was constructed based on single copy orthogroups (i.e. gene clusters with only one gene of each species). Supporting previous work on the re-classification of former *Thalassiosira* species (Alverson et al. 2011; Stachura-Suchoples and Williams 2009), our phylogenetic analysis showed the distinct placement of *C. weissflogii* (formerly *Thalassiosira weissflogii*) from *Cyclotella pseudonana* (also *Thalassiosira pseudonana*) and *Thalassiosira oceanica* (Fig. 1D).

To identify differences between diatoms and other Ochrophyta, a heatmap was generated based on the gene numbers in the orthogroups among Ochrophyta (Fig. 2A). The enrichment of the gene ontology (GO) category of photosynthesis-light harvesting (Fig. 2B and Supplementary Material Table S3) could indicate that diatoms may share a unique photosynthetic pathway or $CO_2$-concentrating mechanisms (CCMs). This could support the previous hypothesis that CCMs are more common in diatoms than in other ochrophytes (Kustka et al. 2014; Raven and Giordano 2017; (Reinfelder 2011). Furthermore, the orthogroups of centric diatoms were enriched in chitin metabolic processes supporting the general notion that occurs only in centric, not pennate diatoms (Shao et al. 2019). The *C. weissflogii*-specific orthogroups were enriched in similar GO categories, as well as carbohydrate metabolic processes and signal transduction.

To analyse the phylogenetic relatedness of chitin synthases (CHSs) in *C. weissflogii* and diatoms, CHSs from representative fungi and all ochrophytes

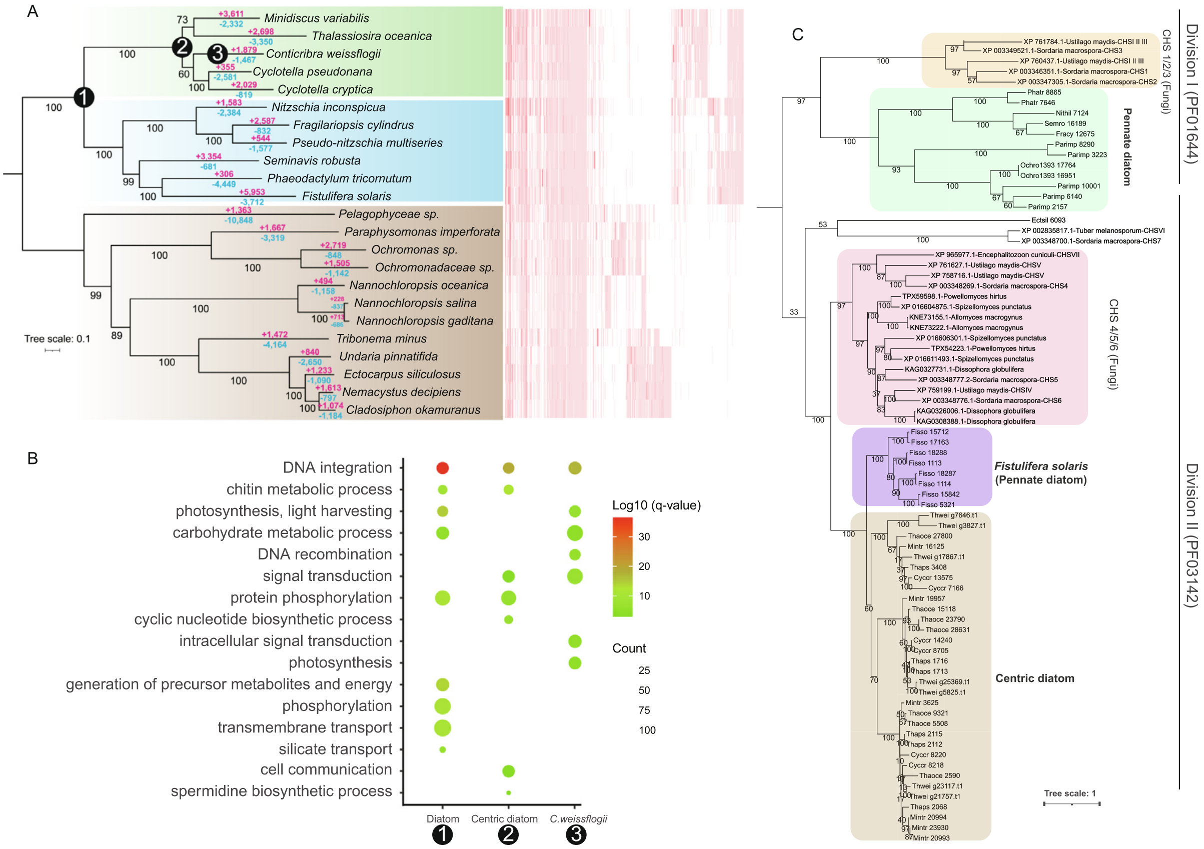


**Figure 2.** Comparative genomic analysis of ochrophytes. **(A)** Heatmap of the proteins of each orthogroup of Ochrophyta sorted according to the phylogenetic tree. The intensity of red indicates the copy number of proteins. The pink numbers in the phylogenetic tree represent the expansion orthogroups while the blue numbers represent the contraction orthogroups. **(B)** The gene ontology enrichment of the genes in diatoms (including centric and pennate diatoms), centric diatoms and *C. weissflogii* extracted from the nodes highlighted in Fig. 2A. **(C)** Phylogenetic tree of chitin synthases. The protein sequences were retrieved from all ochrophytes and several representative fungi (Torruella et al. 2015).

were retrieved based on the common conserved domains, including Chitin_synth_1 (PF01644) and Chitin_synth_2 (PF03142) (Supplementary Material Table S4). The phylogenetic tree was divided into two main clades based on the division I and II domains (Torruella et al. 2015; Fig. 2C). As previously reported (Torruella et al. 2015), the division I clade comprised the fungal CHSs 1/2/3 and CHSs from pennate diatoms. The division II included the fungal CHSs 4/5/6, CHSs from centric diatoms including *C. weissflogii*, and those from the pennate diatom *Fistulifera solaris*. While chitin is generally considered to be present in centric, but not in pennate diatoms, CHS genes have previously been detected in pennate diatoms and their function has been investigated (Shao et al. 2019). Why the CHSs of the pennate diatom *F. solaris* are phylogenetically closer to those of centric rather than pennate diatoms remains to be investigated.

In conclusion, this is the first study on the nuclear genome of the model diatom *C. weissflogii.* Comparative genomic analyses between diatoms and other ochrophytes indicated the existence of unique CCMs and chitin metabolic processes in diatoms. Mining the draft genome for biosynthetic genes using antiSMASH (Blin et al. 2021) identified five terpene, two NRPS-like genes and one Type III PKS gene (Supplementary Material Table S5) suggesting a bioactive potential of *C. weissflogii*, which could support its use as a sustainable, biotechnological production system.

## Declaration of Competing Interest

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

## Acknowledgements

This work was supported by the Danish National Research Foundation to the Center for Microbial Secondary Metabolites (CeMiSt) (DNRF137) and by the Novo Nordisk Foundation (grants NNF20OC0064249 and NNFOC0055625). This work is part of the 10KP project (https://db.cngb.org/10kp/). This work is also supported by China National GeneBank (CNGB; https://www.cngb.org/).

## Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.protis.2021.125845.

## References

**Alverson AJ, Beszteri B, Julius ML, Theriot EC** (2011) The model marine diatom *Thalassiosira pseudonana* likely descended from a freshwater ancestor in the genus *Cyclotella*. BMC Evol Biol **11**:125

**Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, Van Wezel GP, Medema MH, Weber T** (2021) AntiSMASH 6.0: Improving cluster detection and comparison capabilities. Nucleic Acids Res **49**:W29–W35

**Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M** (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinform **3(lqaa108)**

**Cheng H, Bowler C, Xing X, Bulone V, Shao Z, Duan D** (2021) Full-length transcriptome of *Thalassiosira weissflogii* as a reference resource and mining of chitin-related genes. Mar Drugs **19**:392

**Durkin CA, Mock T, Armbrust EV** (2009) Chitin in diatoms and its association with the cell wall. Eukaryot Cell **8**:1038–1050

**Edgar RC, Myers EW** (2005) PILER: Identification and classification of genomic repeats. Bioinformatics **21**:152–158

**Falkowski PG, Barber RT, Smetacek V** (1998) Biogeochemical controls and feedbacks on ocean primary production. Science **281**:200–206

**Guillard RRL, Ryther JH** (1962) Studies of marine planktonic diatoms. I. Cyclotella nana Hustedt and Detonula confervacea Cleve. Can. J. Microbiol. **8**:229–239

**Guillard RRL** (1975) Culture of phytoplankton for feeding marine invertebrates. In Smith WL, Chanley MH (Eds.) Culture of Marine Invertebrate Animals. Plenum Press; New York, USA, pp. 26–60

**Hopkinson BM, Dupont CL, Matsuda Y** (2016) The physiology and genetics of $CO_2$ concentrating mechanisms in model diatoms. Curr Opin Plant Biol **31**:51–57

**Kustka AB, Milligan AJ, Zheng H, New AM, Gates C, Bidle KD, Reinfelder JR** (2014) Low $CO_2$ results in a rearrangement of carbon metabolism to support C4 photosynthetic carbon assimilation in *Thalassiosira pseudonana*. New Phytol **204**:507–520

**Lommer M, Specht M, Roy A-S, Kraemer L, Andreson R, Gutowska MA, Wolf J, Bergner SV, Schilhabel MB, Klostermeier UC, et al.** (2012) Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome Biol **13**:R66

**Martin-Jézéquel V, Hildebrand M, Brzezinski MA** (2000) Silicon metabolism in diatoms: Implications for growth. J Phycol **36**:821–840

**Mock T, Samanta MP, Iverson V, Berthiaume C, Robison M, Holtermann K, Durkin C, Bondurant SS, Richmond K, Rodesch M, et al.** (2008) Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses. Proc Natl Acad Sci USA **105**:1579–1584

**Price AL, Jones NC, Pevzner PA** (2005) De novo identification of repeat families in large genomes. Bioinformatics **21**:351–358

**Rabiee N, Khatami M, Jamalipour SG, Fatahi Y, Iravani S, Varma RS** (2021) Diatoms with invaluable applications in nanotechnology, biotechnology, and biomedicine: Recent advances. ACS Biomater Sci Eng **7**:3053–3068

**Ranallo-Benavidez TR, Jaron KS, Schatz MC** (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun 2020:11:1432

**Raven JA, Giordano M** (2017) Acquisition and metabolism of carbon in the ochrophyta other than diatoms. Philos Trans R Soc B Biol Sci **372**:20160400

**Reinfelder JR** (2011) Carbon concentrating mechanisms in eukaryotic marine phytoplankton. Annu Rev Mar Sci **3**:291–315

**Shao Z, Thomas Y, Hembach L, Xing X, Duan D, Moerschbacher BM, Bulone V, Tirichine L, Bowler C** (2019) Comparative characterization of putative chitin deacetylases from *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* highlights the potential for distinct chitin-based metabolic processes in diatoms. New Phytol **221**:1890–1905

**Sonnenschein EC, Syit DA, Grossart H-P, Ullrich MS** (2021) Chemotaxis of *Marinobacter adhaerens* and its impact on attachment to the diatom *Thalassiosira weissflogii*. Appl Environ Microbiol **78**:6900–6907

**Stachura-Suchoples K, Williams DM** (2009) Description of *Conticribra tricircularis*, a new genus and species of *Thalassiosirales*, with a discussion on its relationship to other continuous cribra species of *Thalassiosira* Cleve (Bacillariophyta) and its freshwater origin. Eur J Phycol **44**:477–486

**Torruella G, De Mendoza A, Grau-Bové X, Antó M, Chaplin MA, Del Campo J, Eme L, Pérez-Cordón G, Whipps CM, Nichols KM, et al.** (2015) Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi. Curr Biol **25**:2404–2410

**Traller JC, Cokus SJ, Lopez DA, Gaidarenko O, Smith SR, McCrow JP, Gallaher SD, Podell S, Thompson M, Cook O, et al.** (2016) Genome and methylome of the oleaginous diatom *Cyclotella cryptica* reveal genetic flexibility toward a high lipid phenotype. Biotechnol Biofuels **9**:258

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

**Xu Z, Wang H** (2007) LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res **35**:265–268