The complete chloroplast genome of the marine microalgae *Chaetoceros muellerii* (Chaetoceroceae)

Yajun Li and Xiaodong Deng

Hainan Provincial Key Laboratory for Functional Components Research and Utilization of Marine Bio-resources, Institute of Tropical Bioscience and Biotechnology, Hainan Academy of Tropical Agricultural Resource, Chinese Academy of Tropical Agricultural Sciences, Haikou, P. R. China

**ABSTRACT**

In this study, the complete chloroplast genome of *Chaetoceros muellerii* was sequenced by using PacBio sequencing platform. The chloroplast genome was 116,284 bp in length, harboring a large single copy (LSC) region of 61,946 bp, a small single copy (SSC) region of 39,308 bp, and a pair of inverted repeats (IR) regions of 7515 bp. The overall GC content was 30.86%. The chloroplast genome of *C. muellerii* contained 131 genes in total, including 95 protein-coding genes, 30 transfer RNA (tRNA) genes and 6 ribosomal RNA (rRNA) genes. The phylogenetic analysis revealed that *C. muellerii* was closely related to *Chaetoceros simplex* with 100% bootstrap values. This study will contribute to the phylogenetic and taxonomic analysis of species in the family Chaetoceroceae.

Chaetoceros Ehrenberg 1844 is the unique genus in the Chaetoceroceae (Biddulphiales, Bacillariophyta) with more than 500 recorded species and infraspecific taxa (Guiry and Guiry 2020). As the largest genus of planktonic diatom, Chaetoceros species are wildly distributed from polar to tropical waters, and play an important role in global carbon cycle and aquatic ecosystems (Nelson et al. 1995). As reported, *Chaetoceros muellerii* contains the fruitful lipids and has the nutrition value for the growth requirements (Liang et al. 2020). To data, the growth requirements, nutrition value and lipid extraction of *C. muellerii* have been well documented (Naghdi et al. 2016; Kumaran et al. 2017; Remize et al. 2020). However, little information is available about genetic and genomic researches on *C. muellerii*. In this study, the complete chloroplast genome of *C. muellerii* was sequenced and assembled by using the third-generation sequencing technology, the genome sequence has been submitted to NCBI GenBank with an accession number of MW004650.

*C. muellerii* strain was obtained from Center for Collections of Marine Algae of Xiamen University (CCMA-187, N24.61", E118.32°). The Chloroplast DNA was isolated with Plant Chloroplast DNA column extraction kit (BioRab, Beijing) according to the instructions of the manufacturer, and sequenced by combining Illumina Hiseq sequencing (Naghdii et al. 2016; Kumaran et al. 2017; Remize et al. 2020). However, little information is available about genetic and genomic researches on *C. muellerii*. In this study, the complete chloroplast genome of *C. muellerii* was sequenced and assembled by using the third-generation sequencing technology, the genome sequence has been submitted to NCBI GenBank with an accession number of MW004650.

*C. muellerii* strain was obtained from Center for Collections of Marine Algae of Xiamen University (CCMA-187, N24.61", E118.32°). The Chloroplast DNA was isolated with Plant Chloroplast DNA column extraction kit (BioRab, Beijing) according to the instructions of the manufacturer, and sequenced by combining Illumina Hiseq4000 and PacBio sequencing platform at Nextomics Biosciences Co. Ltd (Wuhan, China). PacBio library construction was mainly based on the PacBio Sample Net-Shared Protocol (https://www.pacbio.com/wp-content/uploads/2015/09/Shared-Protocol-10-kb-to-20-Kb-Template-Preparation-with-Low-Input-DNA.pdf). Approximately 200 ng of cpDNA were sheared to 15–20 kb via a covaries g-tube device, and then the DNA fragments were end-repaired and ligated SMRTbell adapters. After annealing sequencing primer and binding polymerase to SMRTbell templates, the PacBio library was constructed and sequenced on the PacBio Sequel system, yielding a total of 7924.1 Mb of subreads. For Illumina Hiseq sequencing, approximately 1 µg cpDNA was sheared to 300–500 bp using Covaris M220 and the sequencing libraries were constructed using TruSeq™ Nano DNA Sample Prep Kit. After 8-cycle PCR amplification, purification, quantification and validation, the DNA libraries were sequenced on Illumina Hiseq system. Total 7384.5 Mb of raw data was generated and then filtered using Trimomatic 0.39 (Bolger et al. 2014), obtaining 6530.1 Mb of clean data. The resulting clean data of Illumina Hiseq and the PacBio data were used for De novo assembly of the chloroplast genome using NOVOPlasty v2.7.2 (Dierckxsens et al. 2016) with *Saccharum hildebrandtii* (GenBank: MF563371.1) as the reference. GapCloser V1.12 software (Luo et al. 2012) was used to perform vulnerability completion and base correction. The chloroplast genome annotation was conducted by online software GeSeq (Tillich et al. 2017).

The chloroplast genome length of *C. muellerii* was 116,284 bp with a typical quadripartite structure, which compose of 61,946 bp of a large single copy (LSC) region, 39,308 bp of a small single copy (SSC) region and 7515 bp of...
two inverted repeats (IR) regions. The overall GC content is 30.86%. A total of 131 genes (130 unique genes) were predicted in the chloroplast genome of \textit{C. muellerii}, including 95 protein-coding genes, 30 transfer RNA genes and 6 ribosomal RNA genes. Moreover, none of gene contains intron. Three rRNAs (rns, rnl and rrn5) and three tRNAs (trnp-UGG, trnl-GAU and trnA-UGC) were located in the repeats regions. The IR regions also contain photosystem II protein Y (psbY), acyl
carrier protein (acpP) and hypothetical chloroplast reading frames 89(ycf89). Ribosomal protein 32 (rpL32) was located in the border of IRA and SSC, and ycf45 was located in the border of IRA and LSC.

Phylogenetic analyses were performed using maximum likelihood (ML) in PhyloSuite using the concatenated coding sequences of 116 chloroplast coding genes for 37 species of Bacillariophyta (Zhang et al. 2020). Supports for nodes were calculated via 5000 ultrafast bootstrap replicates. The result of phylogenetic analysis revealed that *C. muellerii* was suggested more closely to *Chaetoceros simplex*, forming a clade with *Acanthoceras zachariasii* and *Cerataulina daemon* in this study (Figure 1).

In conclusion, the complete chloroplast genome sequence of *C. muellerii* can provide a reliable barcode for understanding the phylogeny and evolution of Chaetoceroceae species. In the next study, we will analyses the conserved and variable regions of *C. muellerii* cpDNA sequence, and develop effective DNA barcodes to accurate identification of Chaetoceroceae species.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This work was supported by the Key Research and Development Program of Hainan Province [ZDYF2019171], and Financial Fund of the Ministry of Agriculture and Rural Affairs, P.R.China [NFZX2018].

**Data availability statement**

The genome sequence data that support the findings of this study are openly available in GenBank of NCBI at [https://www.ncbi.nlm.nih.gov/](https://www.ncbi.nlm.nih.gov/) under the accession no. MW004650. The associated BioProject, SRA, and Bio-Sample numbers are PRJNA684148, SRS7869517, and SAMN17052596 respectively.

**References**

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114–2120.

Dierckxsens N, Mardulyn P, Smits G. 2016. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 45:e18.

Guiry MD, Guiry GM. 2020. AlgaeBase. World-wide electronic publication, National University of Ireland, Galway; [accessed 2020 May 22]. [http://www.algaebase.org](http://www.algaebase.org).

Kumar J, Jose B, Joseph V, Singh ISB. 2017. Optimization of growth requirements of marine diatom *Chaetoceros muellerii* using response surface methodology. Aquac Res. 48(4):1512–1513.

Liang C, Zhang Y, Wang L, Shi L, Xu D, Zhang X, Ye N. 2020. Features of metabolic regulation revealed by transcriptomic adaptions driven by long-term elevated pCO2 in *Chaetoceros muellerii*. Phycological Res. 68(3):236–248.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 1(1):18.

Naghdi FG, Bai X, Thomas-Hall SR, Sharma K, Schenk PM. 2016. Lipid extraction from wet *Chaetoceros muellerii* culture and evaluation of remaining defatted biomass. Algal Research. 20:205–212.

Nelson DM, Treguer P, Brzezinski MA, Leynaert A, Queguiner B. 1995. Production and dissolution of biogenic silica in ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global Biogeochem Cycles. 9(3):359–372.

Remize M, Planchnon F, Loh AN, Le Grand F, Bideau A, Le Goic N, Fleury E, Miner P, Corvaisier R, Volety AK, et al. 2020. Study of synthesis pathway of the essential polyunsaturated fatty acid 20:5n-3 in the diatom *Chaetoceros muellerii* using 13C-isotope labeling. Biomolecules. 10(5):797.

Tillich M, Lehward P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017. GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 45(W1):W6–W11.

Zhang D, Gao F, Jakovlić I, Zou H, Zhang J, Li WX, Wang GT. 2020. PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour. 20(1):348–355.