Data in Brief xxx (xxxx) xxx

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Draft genome assembly and sequencing dataset of the marine diatom *Skeletonema costatum RCC75*

Maria Sorokina[a,*], Emanuel Barth[b], Mahnoor Zulfiqar[a], Michiel Kwantes[a], Georg Pohnert[a], Christoph Steinbeck[a,*]

[a] *Institute for Inorganic and Analytical Chemistry, Friedrich Schiller University, Lessingstrasse 8, Jena, Germany*
[b] *Bioinformatics Core Facility, Friedrich Schiller University, Leutragraben 1, Jena, Germany*

ARTICLE INFO

Article history:
Received 15 November 2021
Revised 20 January 2022
Accepted 3 February 2022
Available online xxx

Keywords:
Genome sequencing
Diatoms
Bacillariophyceae
PacBio sequencing
Illumina sequencing
Skeletonema costatum
Algal genome

ABSTRACT

Diatoms (Bacillariophyceae) are a major constituent of the phytoplankton and have a universally recognized ecological importance. Between 1,000 and 1,300 diatom genera have been described in the literature, but only 10 nuclear genomes have been published and made available to the public up to date. *Skeletonema costatum* is a cosmopolitan marine diatom, principally occurring in coastal regions, and is one of the most abundant members of the *Skeletonema* genus. Here we present a draft assembly of the *Skeletonema costatum RCC75* genome, obtained from PacBio and Illumina NovaSeq data. This dataset will expand the knowledge of the Bacillariophyceae genetics and contribute to the global understanding of phytoplankton's physiological, ecological, and environmental functioning.

* Corresponding authors.
E-mail addresses: maria.sorokina@uni-jena.de (M. Sorokina), christoph.steinbeck@uni-jena.de (C. Steinbeck).

https://doi.org/10.1016/j.dib.2022.107931

Please cite this article as: M. Sorokina, E. Barth and M. Zulfiqar et al., Draft genome assembly and sequencing dataset of the marine diatom Skeletonema costatum RCC75, Data in Brief, https://doi.org/10.1016/j.dib.2022.107931

Specifications Table

Subject	Omics
Specific Subject Area	Genomics
Type of Data	Table, Raw data, genome sequences in Fasta format
How the data was acquired	Genome sequence was acquired using Pacbio Sequel I and Illumina NovaSeq PE150
Data Format	Raw, analysed and filtered data
Description of Data Collection	The strain RCC75 was grown in a seawater medium for 10 days. Later it was split into four samples which were used for DNA Extraction and sequencing.
Data Source Location	Institute: Roscoff Culture Collection Town: Roscoff Country: France
Data Accessibility	This Whole Genome Sequencing project has been deposited at DDBJ/ENA/GenBank under the accession number JAHBBA000000000. The version described in this paper is version JAHBBA010000000. The raw data is available on NCBI SRA with the accession number PRJNA647329 at https://www.ncbi.nlm.nih.gov/bioproject/647329.

Value of the Data

- The Genome assembly data of *Skeletonema costatum* RCC75 is an addition to the only 10 published nuclear genomes from the Bacillariophyceae class.
- The algal research community will benefit from this data with its descriptive side of the species genome and how it relates to other *Skeletonema sp.*. It will allow exploring the similarities and differences between the different species within the *Skeletonema* genus, and the *Skeletonema costatum* species.
- This resource will improve the comprehension of metabolic pathways and lead to more marine natural products identification.

1. Data Description

Members of the Bacillariophyceae, commonly called diatoms, are unicellular siliceous algae of the complex phytoplankton community accounting for major primary production in aquatic ecosystems [1]. Diatoms have a large impact on marine silicon biogeochemical cycling as the gross production of biogenic silica exceeds the net oceanic floor silica deposition by a factor of 40 [2]. Because of their abundance and ability to fix carbon, they are also the major producers of oceanic, organic carbon and are hence large determinants of the global carbon cycle [3]. Currently, between 1,000 and 1,300 diatom genera are described, but only 10 nuclear genomes within the Bacillariophyceae have been published until now.

The genus *Skeletonema* comprises unicellular photosynthetic species with distinctive elliptical cells longitudinally stacked to form a colony of up to 24 cells [4]. The colony formation provides optimal survival in unstable and turbulent marine environments [5]. The cells within these chains (or colonies) are connected via long tubular projections called intercalary fultoportula processes (IFPPs). As with most diatoms, the cells take up silicic acid to produce biogenic silica that biomineralizes into a rigid silicified structure, known as frustule [6].

Skeletonema costatum (Fig. 1) is one of the most cosmopolitan and abundant species of genus *Skeletonema* [7] and is principally distributed in the coastal regions [4]. Due to their genetic variability and ecological diversity, these diatoms are well adapted to different environmental conditions and levels of salinity [8]. They are also an excellent paleoenvironmental indicator [9]. *S. costatum* can form algal blooms under optimum conditions. These blooms lead to an increased phytoplankton concentration in the oceans and are promoted by environmental factors such as changes in nutritional content, temperature, and atmospheric deposition [10]. Previously, to dis-

Please cite this article as: M. Sorokina, E. Barth and M. Zulfiqar et al., Draft genome assembly and sequencing dataset of the marine diatom Skeletonema costatum RCC75, Data in Brief, https://doi.org/10.1016/j.dib.2022.107931

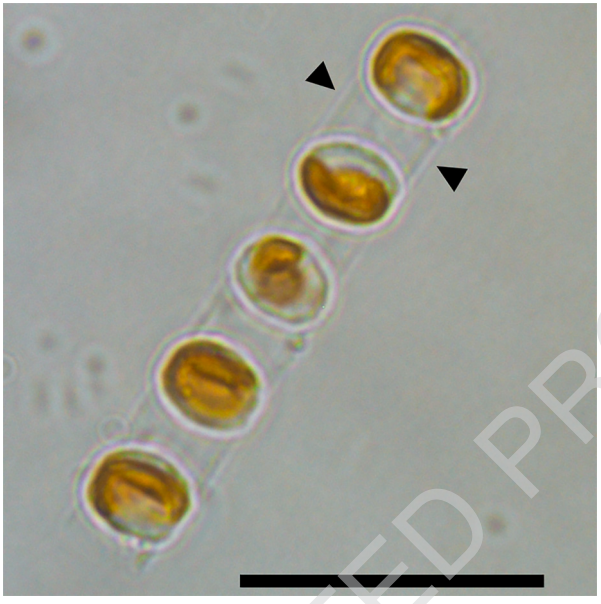

Fig. 1. Bright-field light microscopy image of an S. costatum RCC75 filament consisting of five cells. For the upper pair of cells, the connecting processes are indicated by triangles. Scale bar, 20 µm.

cover putative genes associated with an algal bloom, Ogura *et al*. sequenced and described the genome of *S. costatum* [11] During the same study, a transcriptome analysis under varying light conditions, temperature, and nutrients was performed and described, and the RNA sequence data was released on DDBJ (DRA007346).

The presented genome assembly of *S. costatum* and the raw sequencing data are openly and freely available within the BioProject PRJNA647329 in open FASTA format.

2. Experimental Design, Materials and Methods

2.1. Sample culture and DNA extraction

Here, we report the genome sequence of *Skeletonema costatum* RCC75, which was obtained from the Roscoff Culture Collection (Roscoff, France). The strain was grown in F/2 medium under a 14/10 h light/dark regime with an illumination of 15–24 µmol photons m−2 s−1 for 10 days as standing cultures at 18°C, without additional nutrients supplementation. On day 10, the culture was dense enough to be clearly visible with the naked eye and was then harvested in four samples of 50mL using a needleless syringe. Each sample was then filtered with Durapore 5.0 µm filters, which eliminated most of the obligatory culture microbiome. The filters with diatom cells on them were then inserted in 2 mL microtubes without scraping off the cells. The microtubes were flash-frozen with liquid nitrogen and stored until DNA extraction at −80°C.

DNA was extracted from all four samples using the DNeasy® Plant Mini Kit (Qiagen). Silicon carbide beads (1 mm, BioSpec) were added to each Eppendorf Tube. The cells were then lysed by the 1 mm beads on a beating mill (Qiagen TissueLyser II, 3 × 1 min at frequency 30 Hz, with 1 min at room temperature between each run). The manufacturer's instructions were followed from there, with the exception of the final elution step where the provided elution solution was replaced by an EDTA-free one, following the recommendations of the sequencing facility. The genomic DNA concentration was determined with a Qubit 3.0 (Thermofisher) and a SpeedVac

was used to concentrate the DNA. The DNA samples were then frozen at -80°C until the sequencing.

2.2. Genomic DNA sequencing

The genome sequencing was then performed by the commercial company Novogene (Cambridge, United Kingdom), using two parallel approaches, long reads with Pacbio Sequel I and a fine map with Illumina NovaSeq PE150.

According to the protocol provided by Novogene, the first step in the library construction for the Illumina fine-map sequencing and quality control consisted in the random fragmentation by sonication of the genomic DNA. The DNA fragments were then end-polished, A-tailed, and ligated with the full-length adapters of Illumina sequencing, and followed by further PCR amplification with P5 and indexed P7 oligos. The PCR products as the final construction of the libraries were purified with the AMPure XP system. Then libraries were checked for size distribution by Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA), and quantified by real-time PCR. The qualified libraries were then fed into Illumina sequencers, producing 2Gb of raw data.

For the PacBio sequencing, the first step in the generation of the SMRTbell library, required for this sequencing technology, was the generation of double-stranded 20k DNA fragments, by random DNA shearing. The SMRTbell library itself was produced by ligating universal hairpin adapters onto double-stranded DNA fragments. The hairpin dimers formed during this process were removed at the end of the protocol using a magnetic bead purification step with size-selective conditions. Adapter dimers were also removed using the PacBio MagBead kit. The final step of the library preparation protocol was to remove failed ligation products through the use of exonucleases. After the exonuclease and AMPure PB purification steps, the sequencing primer was annealed to the SMRTbell templates, followed by binding of the sequencing polymerase to the annealed templates. The sample was then sequenced on the PacBio Sequel platform, producing 25Gb of raw data.

2.3. Genome assembly

The genome assembly was performed by the Bioinformatics Core Facility Jena (BiC). The sequencing qualities of the PacBio long reads and the Illumina short reads were monitored using *LongQC* [12] (version 1.2.0) and *FastQC* [13] (version 0.11.9). Before assembly, all raw reads were checked for possible contamination with *Kraken 2* [14] (version 2.1.1). In addition to the standard *Kraken 2* libraries (archaea, bacteria, plasmid, viral, and human), we created and added three additional libraries based on the three available diatom genome assemblies of *Thalassiosira pseudonana* (GCF_000149415.2), *Thalassiosira oceanica* (GCA_000296205.1), and *Skeletonema costatum*[11] to provide a higher read classification resolution. Only reads that were classified as *T. pseudonana*, *T. oceanica*, *S. costatum*, or that could not be classified were kept for assembly. The genome assembly was performed with *Flye* [15] (version 2.8.1) using the parameters *–pacbio-raw* and *-g 30m*. For polishing the genome assembly, the filtered Illumina short reads were aligned to the draft assembly obtained from *Flye* using *Hisat2* [16] (version 2.2.1) with default parameters but not allowing reads to be spliced. Based on the short alignments, the genome assembly sequence was polished using *Pilon* [17] (version 1.23.2). A final assembly report was created utilizing *Quast* [18] (version 5.0.2), and the genome assembly statistics are shown in Table 1. Further re-sequencing will be needed to close the gaps in the draft genome sequence presented in this note and improve the overall genome quality.

Please cite this article as: M. Sorokina, E. Barth and M. Zulfiqar et al., Draft genome assembly and sequencing dataset of the marine diatom Skeletonema costatum RCC75, Data in Brief, https://doi.org/10.1016/j.dib.2022.107931

Table 1
Genome assembly statistics from Quast analysis.

# contigs	1282
# contigs (>=1,000 bp)	1,242
# contigs (>=50,000 bp)	304
Total length	51,134,913
Total length (>=1,000 bp)	51,104,503
Total length (>= 5000 bp)	50,448,718
Total length (>= 25000 bp)	43,834,615
Total length (>= 50000 bp)	36,634,768
Largest contig	756,974
N50	97,960
N75	42,259
L50	147
L75	342
GC (%)	45.13
Mismatches	
# N's	2,800
# N's per 100 kbp	5.48
Predicted genes	
# predicted genes (unique)	27,770
# predicted genes (>= 0 bp)	28,308 + 79 part
# predicted genes (>= 300 bp)	24,999 + 75 part
# predicted genes (>= 1500 bp)	7,002 + 18 part
# predicted genes (>= 3000 bp)	1,487 + 6 part

2.4. Code availability

The code containing the genome assembly workflow is available at Zotero [19].

Ethics Statements

Not applicable.

CRediT Author Statement

Maria Sorokina: Project Coordination, DNA extractions and writing the manuscript; **Emanuel Barth:** Genome Assembly; **Christoph Steinbeck:** Project supervision and obtaining the funds. **Georg Pohnert:** Project supervision and obtaining the funds, Samples provision; **Mahnoor Zulfiqar:** draft writing; **Michiel Kwantes:** DNA extractions. All authors reviewed the manuscript.
These authors contributed equally: Maria Sorokina, Emanuel Barth.
These authors jointly supervised this work: Christoph Steinbeck, Georg Pohnert.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

Support MS, MK, GP and CS: funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–Project-ID 239748522–SFB 1127.

Support MZ, GP, CS: Cluster of Excellence (EXS 2051) "Balance of the Microverse" funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–Project-ID 390713860.

References

[1] D.M. Nelson, P. Tréguer, M.A. Brzezinski, A. Leynaert, B. Quéguiner, Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation, Glob Biogeochem Cycles 9 (1995) 359–372, doi:10.1029/95GB01070.
[2] P. Tréguer, D.M. Nelson, A.J. Van Bennekom, D.J. Demaster, A. Leynaert, B. Quéguiner, The silica balance in the world ocean: a reestimate, Science 268 (1995) 375–379, doi:10.1126/science.268.5209.375.
[3] D.M. Nelson, D.J. DeMaster, R.B. Dunbar, W. Smith, Cycling of organic carbon and biogenic silica in the Southern Ocean: estimates of water-column and sedimentary fluxes on the Ross Sea continental shelf, J Geophys Res C Oceans 101 (1996) 18519–18532, doi:10.1029/96JC01573.
[4] S.W. Jung, S.M. Yun, S.D. Lee, Y.-O. Kim, J. Lee, Morphological characteristics of four species in the genus skeletonema in coastal waters of South Korea, ALGAE 24 (2009) 195–203, doi:10.4490/ALGAE.2009.24.4.195.
[5] R. Margalef, Life-forms of phytoplankton as survival alternatives in an unstable environment, Oceanol Acta 1 (1978) 493–509.
[6] B.E. Reimann, J.C. Lewin, B.E. Volcani, Studies on the biochemistry and fine structure of silica shell formation in diatoms. I. The structure of the cell wall of cylindrotheca fusiformis Reimann and Lewin, J Cell Biol 24 (1965) 39–55, doi:10.1083/jcb.24.1.39.
[7] J. Cheng, Y. Li, J. Liang, Y. Gao, P. Wang, H. Kin-Chung, et al., Morphological variability and genetic diversity in five species of Skeletonema (Bacillariophyta), Prog Nat Sci 18 (2008) 1345–1355, doi:10.1016/j.pnsc.2008.05.002.
[8] S. Balzano, D. Sarno, W.H.C.F. Kooistra, Effects of salinity on the growth rate and morphology of ten Skeletonema strains, J Plankton Res 33 (2011) 937–945, doi:10.1093/plankt/fbq150.
[9] Vossel H, Roeser P, Litt T, Reed JM. Lake Kinneret (Israel): New insights into Holocene regional palaeoclimate variability based on high-resolution multi-proxy analysis 2018;28. https://doi.org/10.1177/0959683618777071.
[10] R. Tian, Q. Lin, D. Li, W. Zhang, X. Zhao, Atmospheric transport of nutrients during a harmful algal bloom event, Reg Stud Mar Sci 34 (2020) 101007, doi:10.1016/j.rsma.2019.101007.
[11] A. Ogura, Y. Akizuki, H. Imoda, K. Mineta, T. Gojobori, S. Nagai, Comparative genome and transcriptome analysis of diatom, Skeletonema costatum, reveals evolution of genes for harmful algal bloom, BMC Genomics 19 (2018) 765, doi:10.1186/s12864-018-5144-5.
[12] Y. Fukasawa, L. Ermini, H. Wang, K. Carty, M.-S. Cheung, LongQC: a quality control tool for third generation sequencing long read data, G3 GenesGenomesGenetics 10 (2020) 1193–1196, doi:10.1534/g3.119.400864.
[13] Andrews S. FastQC a quality control tool for high throughput sequence data. 2010.
[14] D.E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2, Genome Biol 20 (2019) 257, doi:10.1186/s13059-019-1891-0.
[15] M. Kolmogorov, J. Yuan, Y. Lin, P.A. Pevzner, Assembly of long, error-prone reads using repeat graphs, Nat Biotechnol 37 (2019) 540–546, doi:10.1038/s41587-019-0072-8.
[16] D. Kim, J.M. Paggi, C. Park, C. Bennett, S.L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype, Nat Biotechnol 37 (2019) 907–915, doi:10.1038/s41587-019-0201-4.
[17] B.J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, et al., Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, PLOS ONE 9 (2014) e112963, doi:10.1371/journal.pone.0112963.
[18] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics 29 (2013) 1072–1075, doi:10.1093/bioinformatics/btt086.
[19] B.C.F. Jena, Bioinformatics-core-facility-Jena/SE20200226_16: citable release, Zenodo, 2022, doi:10.5281/zenodo.5862116.