AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology® Resource Announcements

GENOME SEQUENCES

Phased Diploid Genome Sequence for the Fast-Growing Microalga *Picochlorum celeri*

Scott A. Becker,[a] Roberto Spreafico,[a] Jennie L. Kit,[a] Rob Brown,[a] Maria Likhogrud,[b] Wei Fang,[c] Matthew C. Posewitz,[c] Joseph C. Weissman,[b] Randor Radakovits[a]

[a]Synthetic Genomics, Inc., La Jolla, California, USA
[b]ExxonMobil Research and Engineering Company, Annandale, New Jersey, USA
[c]Department of Chemistry and Geochemistry, Colorado School of Mines, Golden, Colorado, USA

ABSTRACT *Picochlorum celeri* is a fast-growing marine microalga with high biomass productivity. Here, we report the use of PacBio sequencing to assemble the phased diploid genome of *P. celeri*.

Picochlorum celeri (*Chlorophyceae*) is an algal species that is of commercial interest due to its high photoautotrophic reproductive rates and biomass productivity (1). Various *Picochlorum* species have been studied for potential application in biomass production (1–5), aquaculture feedstock (6, 7), and wastewater remediation (8). In recent years, several *Picochlorum* genome assemblies have been published (2, 9–11) and some of these are proposed to be diploid (10). Here, we report the first fully phased diploid *Picochlorum* genome assembly published to date (the organism has two copies of each chromosome; we represent the linked differences between them consistently along each scaffold).

P. celeri was isolated from the Gulf Coast of Texas in June 2015 and grown in enriched Instant Ocean seawater medium (1). For PacBio and Illumina sequencing, cell lysis of duplicate biological samples was accomplished through bead beating for 3 min in a Mini-BeadBeater (Biospec Products, Inc.) with 1-mm beads from OPS Diagnostics (PFAW 1000-100-21). Following lysis, total DNA was extracted using the Qiagen DNeasy PowerPlant Pro kit according to the manufacturer's instructions. PacBio libraries were made using the SMRTbell template preparation kit with a molecular size cutoff of 10 kb. Illumina libraries were prepared using the TruSeq DNA LT sample preparation kit with a standard molecular size of 350 bp.

We obtained long reads from one single-molecule real-time (SMRT) cell on the Sequel instrument (Pacific Biosciences, Menlo Park, CA, USA) and short reads from the Illumina NextSeq system. The short reads were subsampled to 100× coverage and used only with GenomeScope (version 1.0.0) (12) for ploidy estimation; two kmer lengths (21 and 27) and two samples of reads gave heterozygosity estimates between 0.91% and 0.96%, indicating a diploid genome. The long reads were assembled with FALCON-Unzip (version 1.1.4, included in the pb-assembly conda recipe downloaded in December 2018) (13). In all, we gathered 273,487 long reads, with a mean insert length of 6,407 bp and an N_{50} of 9,250 bp. We worked with Phase Genomics (Seattle, WA, USA) to prepare a Hi-C library using a Phase Genomics Proximo Hi-C Plant kit, which is a commercially available version of the Hi-C protocol (14). Following the manufacturer's instructions for the kit, intact cells from two samples were cross-linked using a formaldehyde solution, digested using the Sau3AI restriction enzyme, and proximity ligated with biotinylated nucleotides to create chimeric molecules composed of fragments from different regions of the genome that were physically proximal *in vivo* but not necessarily genomically proximal. Continuing with the manufacturer's protocol, mole-

Citation Becker SA, Spreafico R, Kit JL, Brown R, Likhogrud M, Fang W, Posewitz MC, Weissman JC, Radakovits R. 2020. Phased diploid genome sequence for the fast-growing microalga *Picochlorum celeri*. Microbiol Resour Announc 9:e00087-20. https://doi.org/10.1128/MRA.00087-20.

Editor Antonis Rokas, Vanderbilt University

Address correspondence to Randor Radakovits, rradakovits@syntheticgenomics.com.

Received 27 February 2020
Accepted 17 April 2020
Published 14 May 2020

cules were precipitated with streptavidin beads and processed into an Illumina-compatible sequencing library. Quality control for the library was performed by sequencing a small number of read pairs (556,109 read pairs) on an Illumina iSeq system and then aligning the reads using BWA-MEM (version 0.7.17) (15) with the -5SP and -t 8 options specified. The alignment was assessed for true Hi-C pairs in which forward and reverse reads were not found genetically proximal. Notably, the percentage of high-quality read pairs that aligned $>$10 kb apart on contigs longer than 10 kb was 13.93% (expected, 1 to 15%), the percentage of intercontig high-quality read pairs on contigs longer than 10 kb was 35.79% (expected, 10 to 60%), and the percentage of same-strand high-quality read pairs was 9.78% (expected, 2 to 50%). Sequencing was performed on an Illumina HiSeq 4000 system, generating a total of 164,537,658 PE150 read pairs.

FALCON-Phase (version 2) (16) was run using default parameters to correct likely phase-switching errors in the primary contigs and alternate haplotigs from FALCON-Unzip and output its results in pseudohap format, creating one complete set of contigs for each phase. Hi-C reads were then aligned to phase 0 contigs following the manufacturer's recommendations (17). Briefly, reads were aligned using BWA-MEM with the -5SP and -t 8 options specified (all other options, default). SAMBLASTER (version 0.1.24) (18) was used to flag PCR duplicates, which were later excluded from analysis. Alignments were then filtered with SAMtools (version 1.9) (19) using the -F 2304 filtering flag to remove nonprimary and secondary alignments.

The Phase Genomics Proximo (version hash d33cacdd) Hi-C genome scaffolding platform was used to create chromosome-scale scaffolds from the FALCON-Phase phase 0 assembly, following the same single-phase scaffolding procedure described by Bickhart et al. (20). As in the LACHESIS method (21), this process computes a contact frequency matrix from the aligned Hi-C read pairs, normalized to the number of Sau3AI restriction sites (GATC) on each contig, and constructs scaffolds in such a way as to optimize expected contact frequency and other statistical patterns in the Hi-C data. Approximately 120,000 separate Proximo runs were performed to optimize the number of scaffolds and scaffold construction in order to make the scaffolds as concordant with the observed Hi-C data as possible. This process resulted in a set of 5 preliminary chromosome-scale scaffolds containing 13.5 Mbp of sequence (98.5% of the input assembly).

Juicebox (version 1.9.8) (22, 23) was then used to correct scaffolding errors, resulting in a total of 12 chromosome-scale scaffolds. FALCON-Phase was run a second time to detect and correct phase-switching errors that were not detectable at the contig level but were detectable at the chromosome-scale scaffold level. Metadata generated by FALCON-Phase for scaffold phasing were used to generate matching .assembly files (a file format used by Juicebox) for each phase and subsequently used to produce a diploid, fully phased, chromosome-scale set of scaffolds using a purpose-built script (https://github.com/phasegenomics/juicebox_scripts).

We polished the diploid assembly twice sequentially with the long reads, aligning with pbmm2 (version 1.0.0) and polishing with Arrow (version 2.3.3) (both available at https://github.com/PacificBiosciences/pbbioconda). Long-read polishing was stopped after two rounds when the consensus quality was estimated to be better than Q40 (a third round with Arrow suggested fewer than 1 in 10,000 changes). Short-read polishing was not done due to the risk of incorrectly merging the two phases of the genome (Shawn Sullivan, Phase Genomics, personal communication). The final phased and scaffolded genome consists of two phases totaling 27.43 Mbp spread across two pairs of 15 scaffolds each, with a scaffold N_{50} of 1.151 Mbp and an overall G+C content of 46%. The two phases of the genome were aligned and analyzed with the nucmer and dnadiff programs from the MUMmer4 suite (version 4.0.0.beta2) (24), finding a total of 112,875 single-nucleotide polymorphisms (SNPs) between the two phases (SNP heterozygosity, 0.8%). Assemblytics software (version available on 15 August 2016, git commit hash c937e96d) (25) was used to analyze the structural variation of the two phases of the genome and found 16,794 indels and larger structural variants affecting

558.74 kb. Default parameters were used for all software unless otherwise specified. This phased assembly will enable future studies to better understand the photosynthetic efficiency of *P. celeri*.

Data availability. This whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number JAACMV000000000. The raw data are available under the accession number PRJNA598876.

ACKNOWLEDGMENTS

This research was supported by ExxonMobil and Synthetic Genomics, Inc.

We acknowledge excellent technical work and discussions with Shawn Sullivan and Hayley Mangleson at Phase Genomics.

REFERENCES

1. Weissman JC, Likhogrud M, Thomas DC, Fang W, Karns DA, Chung JW, Nielsen R, Posewitz MC. 2018. High-light selection produces a fast-growing *Picochlorum celeri*. Algal Res 36:17–28. https://doi.org/10.1016/j.algal.2018.09.024.
2. Dahlin LR, Gerritsen AT, Henard CA, Van Wychen S, Linger JG, Kunde Y, Hovde BT, Starkenburg SR, Posewitz MC, Guarnieri MT. 2019. Development of a high-productivity, halophilic, thermotolerant microalga *Picochlorum renovo*. Commun Biol 2:388. https://doi.org/10.1038/s42003-019-0620-2.
3. de la Vega M, Diaz E, Vila M, León R. 2011. Isolation of a new strain of *Picochlorum* sp and characterization of its potential biotechnological applications. Biotechnol Prog 27:1535–1543. https://doi.org/10.1002/btpr.686.
4. Zhu Y, Dunford NT. 2013. Growth and biomass characteristics of *Picochlorum oklahomensis* and *Nannochloropsis oculata*. J Am Oil Chem Soc 90:841–849. https://doi.org/10.1007/s11746-013-2225-0.
5. Watanabe K, Fujii K. 2016. Isolation of high-level-CO_2-preferring *Picochlorum* sp. strains and their biotechnological potential. Algal Res 18:135–143. https://doi.org/10.1016/j.algal.2016.06.013.
6. Chen TY, Lin HY, Lin CC, Lu CK, Chen YM. 2012. *Picochlorum* as an alternative to *Nannochloropsis* for grouper larval rearing. Aquaculture 338–341:82–88. https://doi.org/10.1016/j.aquaculture.2012.01.011.
7. Kumar SD, Santhanam P, Ananth S, Kaviyarasan M, Nithya P, Dhanalakshmi B, Park MS, Kim MK. 2017. Evaluation of suitability of wastewater-grown microalgae (*Picochlorum maculatum*) and copepod (*Oithona rigida*) as live feed for white leg shrimp *Litopenaeus vannamei* post-larvae. Aquacult Int 25:393–411. https://doi.org/10.1007/s10499-016-0037-6.
8. Wang S, Shi X, Palenik B. 2016. Characterization of *Picochlorum* sp. use of wastewater generated from hydrothermal liquefaction as a nitrogen source. Algal Res 13:311–317. https://doi.org/10.1016/j.algal.2015.11.015.
9. Gonzalez-Esquer CR, Twary SN, Hovde BT, Starkenburg SR. 2018. Nuclear, chloroplast, and mitochondrial genome sequences of the prospective microalgal biofuel strain *Picochlorum soloecismus*. Genome Announc 6:e01498-17. https://doi.org/10.1128/genomeA.01498-17.
10. Foflonker F, Mollegard D, Ong M, Yoon HS, Bhattacharya D. 2018. Genomic analysis of *Picochlorum* species reveals how microalgae may adapt to variable environments. Mol Biol Evol 35:2702–2711. https://doi.org/10.1093/molbev/msy167.
11. Krasovec M, Vancaester E, Rombauts S, Bucchini F, Yau S, Hemon C, Lebredonchel H, Grimsley N, Moreau H, Sanchez-Brosseau S, Vandepoele K, Piganeau G. 2018. Genome analyses of the microalga *Picochlorum* provide insights into the evolution of thermotolerance in the green lineage. Genome Biol Evol 10:2347–2365. https://doi.org/10.1093/gbe/evy167.
12. 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. https://doi.org/10.1093/bioinformatics/btx153.
13. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13:1050–1054. https://doi.org/10.1038/nmeth.4035.
14. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–293. https://doi.org/10.1126/science.1181369.
15. Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595. https://doi.org/10.1093/bioinformatics/btp698.
16. Kronenberg ZN, Hall RJ, Hiendleder S, Smith TPL, Sullivan ST, Williams JL, Kingan SB. 2018. FALCON-Phase: integrating PacBio and Hi-C data for phased diploid genomes. BioRxiv 327064. https://doi.org/10.1101/327064.
17. Phase Genomics. 2019. Aligning and QCing Phase Genomics Hi-C data. https://phasegenomics.github.io/2019/09/19/hic-alignment-and-qc.html.
18. Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 30:2503–2505. https://doi.org/10.1093/bioinformatics/btu314.
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
20. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, Burton JN, Huson HJ, Nystrom JC, Kelley CM, Hutchison JL, Zhou Y, Sun J, Crisà A, Ponce de León FA, Schwartz JC, Hammond JA, Waldbieser GC, Schroeder SG, Liu GE, Dunham MJ, Shendure J, Sonstegard TS, Phillippy AM, Van Tassell CP, Smith TPL. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet 49:643–650. https://doi.org/10.1038/ng.3802.
21. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 31:1119–1125. https://doi.org/10.1038/nbt.2727.
22. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst 3:99–101. https://doi.org/10.1016/j.cels.2015.07.012.
23. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159:1665–1680. https://doi.org/10.1016/j.cell.2014.11.021.
24. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol 14:e1005944. https://doi.org/10.1371/journal.pcbi.1005944.
25. Nattestad M, Schatz MC. 2016. Assemblytics: a Web analytics tool for the detection of variants from an assembly. Bioinformatics 19:3021–3023. https://doi.org/10.1093/bioinformatics/btw369.