Next-Generation Sequencing of *Haematococcus lacustris* Reveals an Extremely Large 1.35-Megabase Chloroplast Genome

Nicholas Bauman,ᵃ Srividya Akella,ᵃ Elizabeth Hann,ᵃ* Robert Morey,ᵃ* Ariel S. Schwartz,ᵃ Rob Brown,ᵃ Toby H. Richardsonᵃ

ᵃSynthetic Genomics, Inc., La Jolla, California, USA

ABSTRACT *Haematococcus lacustris* is an industrially relevant microalga that is used for the production of the carotenoid astaxanthin. Here, we report the use of PacBio long-read sequencing to assemble the chloroplast genome of *H. lacustris* strain UTEX:2505. At 1.35 Mb, this is the largest assembled chloroplast of any plant or alga known to date.

Haematococcus lacustris (*Chlorophyceae*) is an algal species of commercial interest due to is ability to accumulate high levels of the red carotenoid astaxanthin, which is used in fish feed, cosmetics, and nutraceuticals (1–3). Efforts to increase *Haematococcus* productivity by targeting transformation to both nuclear and chloroplast genomes (4, 5) have been limited by the lack of high-quality genome assemblies. The use of PacBio single-molecule real-time (SMRT) sequencing has gained traction recently for improving genome assemblies, including for algal chloroplasts, due to the long reads that are generated (6, 7). *Floydiella* and *Volvox* species have some of the largest chloroplasts known among members of the *Viridiplantae* at ~500 kb (8, 9); previously, the largest known chloroplast (1.13 Mb) of any lineage belonged to the red alga *Corynoplastis japonica* (10).

H. lacustris UTEX:2505 was obtained from the Culture Collection of Algae at the University of Texas and grown in optimal *Haematococcus* medium (11). Total DNA was extracted using a modified cetyltrimethylammonium bromide method (12). Purified DNA was converted to a SMRTbell library according to the manufacturer's instructions (PacBio, Menlo Park, CA, USA) and size selected with a 10-kb cutoff using Blue Pippin (Sage Science, Beverly MA, USA). Twenty-four SMRT cells were sequenced using the Pacific Biosciences RS II platform, providing 13.6 Gb of mapped reads. The chloroplast genome was assembled *de novo* using the Hierarchical Genome Assembly Process (HGAP) version 2 algorithm (13), yielding a single contig. Over 91,000 subreads, with a mean subread length of 8,900 bp, went into the assembly of the chloroplast genome. GeneMarkS version 4.17 (14) was used to generate gene predictions *ab initio*. Gene annotations were generated through the Synthetic Genomics, Inc. (La Jolla, CA, USA) proprietary Archetype annotation pipeline, as previously described (15).

The circular closed chloroplast genome was assembled into 1.352 Mb at >500× coverage, containing the small single-copy (SSC) and long single-copy (LSC) regions, as well as two inverted repeat (IR) regions, and having an overall G+C content of 50%. While the total size of the chloroplast genome is very large, the combined size of its coding regions is ~110 kb, including 125 protein-coding genes and 12 tRNAs. Evidence for both *cis*- and *trans*-splicing events, features seen in other algae (16), is found throughout the chloroplast genome. A total of 139,006 repeats were found and classified into 34,710 families within the chloroplast. The repeats ranged from 42 to 104 bp, with the two most prominent repeats having a length of 43 bp. The copy numbers of each repeat ranged from 5 to 270 and were highly skewed toward

Address correspondence to Nicholas Bauman, nbauman@syntheticgenomics.com.

* Present address: Elizabeth Hann, University of California, Riverside, California, USA; Robert Morey, University of California, San Diego, La Jolla, California, USA.

intergenic regions throughout the chloroplast. Another highly complex and large (>525 kb) chloroplast genome belonging to *Volvox caterii* (9) contains large numbers of short repetitive DNA sequences, an issue that has hindered the completion of a full chloroplast genome assembly of that species to date. We believe that this work shows that long-read sequencing with PacBio SMRT technology (or another such platform) can be beneficial to assembling complex and repetitive organellar genomes in addition to complex nuclear genomes. Additionally, the size and abundance (~50% of total DNA) of this plastid genome mean that more coverage is required to obtain a good nuclear genome for the genus *Haematococcus*.

Accession number(s). The assembled *H. lacustris* UTEX:2505 chloroplast genome has been deposited at GenBank under the accession number MG677935.

ACKNOWLEDGMENT

REFERENCES

1. Lorenz RT, Cysewski GR. 2000. Commercial potential for *Haematococcus* microalgae as a natural source of astaxanthin. Trends Biotechnol 18:160–167. https://doi.org/10.1016/S0167-7799(00)01433-5.
2. Ambati RR, Phang SM, Ravi S, Aswathanarayana RG. 2014. Astaxanthin: sources, extraction, stability, biological activities and its commercial applications—a review. Mar Drugs 12:128–152. https://doi.org/10.3390/md12010128.
3. Olaizola M. 2003. Commercial development of microalgal biotechnology: from the test tube to the marketplace. Biomol Eng 20:459–466. https://doi.org/10.1016/S1389-0344(03)00076-5.
4. Sharon-Gojman R, Maimon E, Leu S, Zarka A, Boussiba S. 2015. Advanced methods for genetic engineering of *Haematococcus pluvialis* (Chlorophyceae, Volvocales). Algal Res 10:8–15. https://doi.org/10.1016/j.algal.2015.03.022.
5. Gutierrez CL, Gimpel J, Escobar C, Marshall SH, Henríquez V. 2012. Chloroplast genetic tool for the green microalgae *Haematococcus pluvialis* (Chlorophyceae, Volvocales). J Phycol 48:976–983. https://doi.org/10.1111/j.1529-8817.2012.01178.x.
6. Wu Z, Gui S, Quan Z, Pan L, Wang S, Ke W, Liang D, Ding Y. 2014. A precise chloroplast genome of *Nelumbo nucifera* (Nelumbonaceae) evaluated with Sanger, Illumina MiSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. BMC Plant Biol 14:289. https://doi.org/10.1186/s12870-014-0289-0.
7. Starkenburg SR, Polle JEW, Hovde B, Daligault HE, Davenport KW, Huang A, Neofotis P, McKie-Krisberg Z. 2017. Draft nuclear genome, complete chloroplast genome, and complete mitochondrial genome for the biofuel/bioproduct feedstock species *Scenedesmus obliquus* strain DOE0152z. Genome Announc 5(32):00617-17. https://doi.org/10.1128/genomeA.00617-17.
8. Brouard JS, Otis C, Lemieux C, Turmel M. 2010. The exceptionally large chloroplast genome of the green alga *Floydiella terrestris* illuminates the evolutionary history of the *Chlorophyceae*. Genome. Biol Evol 2:240–256. https://doi.org/10.1093/gbe/evq014.
9. Smith DR, Lee RW. 2010. Low nucleotide diversity for the expanded organelle and nuclear genomes of *Volvox carteri* supports the mutational-hazard hypothesis. Mol Biol Evol 27:2244–2256. https://doi.org/10.1093/molbev/msq110.
10. Muñoz-Gómez SA, Mejía-Franco FG, Durnin K, Colp M, Grisdale CJ, Archibald JM, Slamovits CH. 2017. The new red algal subphylum proteorhodophytina comprises the largest and most divergent plastid genomes known. Curr Biol 27:1677–1684. https://doi.org/10.1016/j.cub.2017.04.054.
11. Fábregas J, Domínguez A, Regueiro M, Maseda A, Otero A. 2000. Optimization of culture medium for the continuous cultivation of the microalga *Haematococcus pluvialis*. Appl Microbiol Biotechnol 53:530–535. https://doi.org/10.1007/s002530051652.
12. Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, Thompson WF. 2006. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat Protoc 1:2320. https://doi.org/10.1038/nprot.2006.384.
13. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10:563–569. https://doi.org/10.1038/nmeth.2474.
14. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29:2607–2618. https://doi.org/10.1093/nar/29.12.2607.
15. Gallone B, Steensels J, Prahl T, Soriaga L, Saels V, Herrera-Malaver B, Merlevede A, Roncoroni M, Voordeckers K, Miraglia L, Teiling C, Steffy B, Taylor M, Schwartz A, Richardson T, White C, Baele G, Maere S, Verstrepen KJ. 2016. Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. Cell 166:1397–1410. https://doi.org/10.1016/j.cell.2016.08.020.
16. Lefebvre-Legendre L, Reifschneider O, Kollipara L, Sickmann A, Wolters D, Kück U, Goldschmidt-Clermont M. 2016. A pioneer protein is part of a large complex involved in trans-splicing of a group II intron in the chloroplast of *Chlamydomonas reinhardtii*. Plant J 85:57–69. https://doi.org/10.1111/tpj.13089.