

Genome assembly and annotation of *Scenedesmus bijugus* var. *obtusiusculus* AT-UAM from Cuatro Ciénegas, México

Freddy Castillo-Alfonso,¹ Juan Carlos Sigala Alanis,² Marcia Morales-Ibarría,² Cristal Zúñiga³

AUTHOR AFFILIATIONS See affiliation list on p. 2.

ABSTRACT We report the draft genome sequence of *Scenedesmus bijugus* var. *obtusiusculus* AT-UAM, a CO₂-tolerant and lipid-producing microalga isolated from Cuatro Ciénegas, México. The strain shows potential for biotechnological applications due to its adaptability and high lipid accumulation.

KEYWORDS nanopore sequencing, chlorophyte genomics, integrated gene annotation, bioinformatics

Scenedesmus bijugus var. *obtusiusculus* AT-UAM is a promising microalga for sustainable energy and environmental applications. It exhibits high photosynthetic rates, adaptability to temperatures from below 10°C to above 40°C, and tolerance to elevated CO₂ levels (up to 10% [vol/vol]) (1). The species captures CO₂ from flue gases and shows optimal growth at pH 5.3, reaching 55% lipid content under nitrogen starvation (2). This strain was isolated from Poza Churince, 19.5 km southwest of Cuatro Ciénegas, Coahuila, México (3).

The strain was cultured in BG-11 medium for 72 h, under illumination (96 μmol photons m⁻² s⁻¹) and 28°C (1). The genomic DNA was extracted using an Ultraclean Plant DNA Isolation kit (MoBio Laboratories, Carlsbad, USA). The ITS and 18S rDNA sequences were obtained as previously described (3) and deposited in GenBank (<https://www.ncbi.nlm.nih.gov/nuccore/KP318981.1>).

Whole-genome sequencing was performed on two MinION flow cells using the Rapid Barcoding Kit 24 V14 (SQK-RBK114.24; Oxford Nanopore Technologies, Oxford, UK), employing transposase-based barcode attachment without additional shearing or size selection. Libraries were multiplexed with four barcodes from the same *Scenedesmus* genome. Read quality was assessed with FastQC v0.12.1 (4). The first run produced 719,863 reads, of which 682,872 (94.9%) were retained after filtering with Fastp v1.0.1 (5), and the second run generated 967,498 reads, retaining 849,870 (87.9%) high-quality reads, with average read length increasing from 2.0 kb to 2.3 kb. Genome k-mer histograms were generated with Jellyfish v1.0.1 (6) and analyzed with GenomeScope v2.0 (7), which estimated a haploid genome size of 117.36 Mbp, 81.6% unique, 1.23% error rate, 5.78× coverage, and 0.078% of duplication. Long-read assembly using Flye v2.9.6 (8), which includes both mitochondrial and chloroplast genomes, consists of 275 contigs spanning 100.2 Mb in total, with an N50 of 866,524 bp and a maximum contig of 3.71 Mb. GC content was 56.88%, and prior to submission to NCBI (9), contaminant sequences were removed using NCBI FCS GX v0.5.5 (10).

Genome completeness was assessed using BUSCO v5 (11) with the “chlorophyta_odb10” ortholog data set, which contains 1,519 universal single-copy orthologs. The *Scenedesmus bijugus* var. *obtusiusculus* assembly was 92.0% complete (90.9% single-copy and 1.1% duplicated), with 2.6% fragmented and 5.4% missing genes, indicating a nearly complete genome with an average of 1.01 orthologs per gene.

Editor Jason E. Stajich, University of California Riverside, Riverside, California, USA

Address correspondence to Cristal Zúñiga, czuniga2@sdsu.edu.

The authors declare no conflict of interest.

See the funding table on p. 2.

Received 15 August 2025

Accepted 9 January 2026

Published 4 February 2026

Copyright © 2026 Castillo-Alfonso et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Gene prediction was performed with Augustus v3.4.0 (12) using a custom training set from *Scenedesmus obliquus* coding sequences reported in NCBI (9). Prediction was performed on both strands using the genemodel=complete setting. A total of 10,412 protein-coding genes were identified. Functional annotation was performed using BLASTp ($e\text{-value} \leq 1e-3$) against a custom NCBI *Scenedesmaceae* protein database (9), yielding 10,384 significant matches. Additionally, the eggNOG-mapper web server (13) and the KEGG Automatic Annotation Server (14) were used to assign orthologs and metabolic functions. Carbohydrate-active enzymes (CAZymes) were predicted using the dbCAN3 (15) server using default parameters. A total of 301 CAZyme-encoding genes were identified in consensus for these two tools. Furthermore, SignalP v6.0 (16) was used to predict signal peptides, revealing that a subset of CAZymes is potentially secreted, indicating a role in extracellular carbohydrate degradation. No plasmids were detected using PlasmidFinder (17).

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (Grant No. 2313313) and the United States Department of Agriculture (USDA-NIFA-AFRI 2024-67017-42676). Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Freddy Castillo-Alfonso: Conceptualization, Formal analysis, Investigation, Methodology, and Writing—original draft. Juan Carlos Sigala Alanis: Methodology and Supervision. Marcia Morales-Ibarria: Resources and Methodology. Cristal Zúñiga: Funding acquisition, Methodology, and Writing—review and editing, Supervision.

AUTHOR AFFILIATIONS

¹Posgrado en Ciencias Naturales e Ingeniería, Universidad Autónoma Metropolitana Unidad Cuajimalpa, Ciudad de México, México

²Departamento de Procesos y Tecnología, Universidad Autónoma Metropolitana-Cuajimalpa, Ciudad de México, México

³Department of Biology, San Diego State University, San Diego, California, USA

AUTHOR ORCIDs

Freddy Castillo-Alfonso  <http://orcid.org/0000-0001-6043-5776>

Juan Carlos Sigala Alanis  <http://orcid.org/0000-0002-1665-5680>

Marcia Morales-Ibarria  <http://orcid.org/0000-0001-9504-0651>

Cristal Zúñiga  <http://orcid.org/0000-0002-0135-7429>

FUNDING

Funder	Grant(s)	Author(s)
National Science Foundation	Grant No. 2313313	Cristal Zúñiga
U.S. Department of Agriculture	USDA-NIFA-AFRI 2024-67017-42676	Cristal Zúñiga

AUTHOR CONTRIBUTIONS

Freddy Castillo-Alfonso, Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft | Juan Carlos Sigala Alanis, Methodology, Supervision | Marcia Morales-Ibarria, Methodology, Resources | Cristal Zúñiga, Funding acquisition, Methodology, Supervision, Writing – review and editing

DATA AVAILABILITY

The draft genome assembly of *Scenedesmus bijugus* var. *obtusiusculus* has been deposited in GenBank under the accession number [GCA_051107615.1](https://doi.org/10.1016/j.procbio.2024.08.014). The raw sequencing reads supporting this assembly have been deposited in the Sequence Read Archive (SRA) under accession numbers [SRX30009262](https://doi.org/10.1016/j.procbio.2024.08.014) and [SRX30009261](https://doi.org/10.1016/j.procbio.2024.08.014). Additional supplementary information related to genome annotation is available at Zenodo under DOI: <https://doi.org/10.5281/zenodo.17283899>.

REFERENCES

- Estrada-Graf AA, Sigala JC, Sánchez-Pérez M, Morales-Ibarria M. 2024. Induction of lipid production through controlled acidification: a transcriptional insight into the metabolism of *Scenedesmus obtusiusculus* AT-UAM. *Process Biochem* 147:175–185. <https://doi.org/10.1016/j.procbio.2024.08.014>
- Sánchez-García L, Cabello J, Jiménez-García LF, Revah S, Morales-Ibarria M. 2020. Enhancing the lipid content of *Scenedesmus obtusiusculus* AT-UAM by controlled acidification under indoor and outdoor conditions. *Algal Res* 51:102024. <https://doi.org/10.1016/j.algal.2020.102024>
- Toledo-Cervantes A, Garduño Solórzano G, Campos JE, Martínez-García M, Morales M. 2018. Characterization of *Scenedesmus obtusiusculus* AT-UAM for high-energy molecules accumulation: deeper insight into biotechnological potential of strains of the same species. *Biotechnol Rep (Amst)* 17:16–23. <https://doi.org/10.1016/j.btre.2017.11.009>
- Brown J, Pirrung M, McCue LA. 2017. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* 33:3137–3139. <https://doi.org/10.1093/bioinformatics/btx373>
- Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
- Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764–770. <https://doi.org/10.1093/bioinformatics/btr011>
- Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* 11:1432. <https://doi.org/10.1038/s41467-020-14998-3>
- Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA. 2016. Assembly of long error-prone reads using de Bruijn graphs. *Proc Natl Acad Sci U S A* 113:E8396–E8405. <https://doi.org/10.1073/pnas.1604560113>
- Schoch CL, Ciufo S, Domrachev M, Hottot CL, Kannan S, Khovanskaya R, Leipe D, McVeigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. 2020. NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020:baaa062. <https://doi.org/10.1093/database/baaa062>
- Astashyn A, Tvedte ES, Sweeney D, Sapojnikov V, Bouk N, Joukov V, Mozes E, Strobe PK, Sylla PM, Wagner L, Bidwell SL, Clark K, Davis EW, Smith-White B, Hlavina W, Pruitt KD, Schneider VA, Murphy TD. 2023. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *bioRxiv*. <https://doi.org/10.1101/2023.06.02.543519>
- Manni M, Berkeley MR, Seppely M, Zdobnov EM. 2021. BUSCO: assessing genomic data quality and beyond. *Curr Protoc* 1:e323. <https://doi.org/10.1002/cpz1.323>
- Hoff KJ, Stanke M. 2019. Predicting genes in single genomes with Augustus. *Curr Protoc Bioinformatics* 65:e57. <https://doi.org/10.1002/cpb.57>
- Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* 38:5825–5829. <https://doi.org/10.1093/molbev/msab293>
- Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35:W182–W185. <https://doi.org/10.1093/nar/gkm321>
- Zheng J, Ge Q, Yan Y, Zhang X, Huang L, Yin Y. 2023. dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res* 51:W115–W121. <https://doi.org/10.1093/nar/gkad328>
- Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H. 2022. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol* 40:1023–1025. <https://doi.org/10.1038/s41587-021-01156-3>
- Carattoli A, Hasman H. 2020. PlasmidFinder and *in silico* pMLST: identification and typing of plasmid replicons in whole-genome sequencing (WGS). *Methods Mol Biol* 2075:285–294. https://doi.org/10.1007/978-1-4939-9877-7_20