Draft *de novo* genome construction of *Scytonema* sp. PRP1: identified from single-cell sequencing library preparation

Preston Parana,[1] Camille Mumm,[1] Michael J. McConnell,[2] Alan P. Boyle[1,3]

AUTHOR AFFILIATIONS See affiliation list on p. 2.

ABSTRACT We present the genome sequence of *Scytonema* sp. PRP1, a cyanobacterium identified during single-cell sequencing library preparation of de-identified human brain samples associated with the Brain Somatic Mosaicism Network. Our 8,266,022 bp genome comprises 140 contigs, containing 6,765 protein-coding sequences, 37 tRNA genes, and 11 rRNA genes.

Scytonema sp. PRP1 was identified as a contaminant during single-cell sequencing library preparation of de-identified human brain samples (1). A single brain nucleus preparation was amplified using multiple annealing and looping-based amplification cycles (MALBAC) at the University of Virginia School of Medicine per Burbulis et al. (2). Although the contamination source remains uncertain, shared buffers were consistently used during the preparation and sequencing of >100 nuclei from the same sample (3), making reagent contamination unlikely. Given MALBAC amplification's high sensitivity, circulating environmental spores are a likely source of this species.

As previously described in Burbulis et al. (2), cells were lysed using 2.5 µL buffer (50 mM Tris pH 8.8, 50 mM NaCl, 50 mM KCl, 1 µM EDTA, 1 µM DTT, 0.1% Tween-20, and 20 mg/mL Proteinase K [Qiagen]), digested at 55°C for 3 h, then inactivated at 75°C and 80°C. Lysates were supplemented with amplification buffer (3 mM dNTPs, 0.667 µM primers) and underwent 5 MALBAC heating/cooling cycles, with enzyme mix added at each cycle. After PCR amplification, DNA was precipitated with NaCl and ethanol (2), purified via centrifugation and QIAquick PCR columns, quantified using Qubit 2.0 HS DNA assay, and prepared for sequencing. Libraries of five single-cell MALBAC-amplified products were prepared with 300 ng barcoded-adapted sample (ONT Native Barcoding kit 24 V14, SQK-NBD114.24) and sequenced on a MinION R10.4.1 flow cell (Mk1B). Genomic DNA was neither fragmented nor size-selected. Data were basecalled using Guppy v6.2.11 (ONT, score filter 9). Chimeric reads were split using duplex_tools (ONT), and adapter trimmed using Porechop v0.2.4 (4). This sample yielded 1,226,304 passing reads; after removing reads that mapped to the human genome (3.3%), 1,185,629 remained for assembly (N50 = 1,414 bp), of which 86.43% mapped to the *Scytonema* genus using Blobtools v1.1.1 (5) and BLASTn v2.14.0 (6). Unicycler v0.5.0 (7) was used for hybrid genome assembly in bold mode; all software used default parameters unless otherwise noted. Contigs <400 bp were filtered (5), and Medaka v1.11.1 (8) polished the final assembly, producing a 101.7× coverage genome of 8.26 Mb with 140 contigs, an N50 of 107.9 KB, a completeness of 82.2%, and 43.5% GC content. QUAST v5.2.0 (9) and BUSCO v5.5.0 (10) assessed genome contiguity and completeness. Sequences were submitted to GenBank and annotated using PGAP v6.6 (11), predicting 6,765 protein-coding genes, 7 CRISPR genes, 11 rRNA genes, 37 tRNA genes, 395 pseudogenes, and 4 ncRNA genes. Proksee (https://proksee.ca) visualized the annotated genome (Fig. 1).
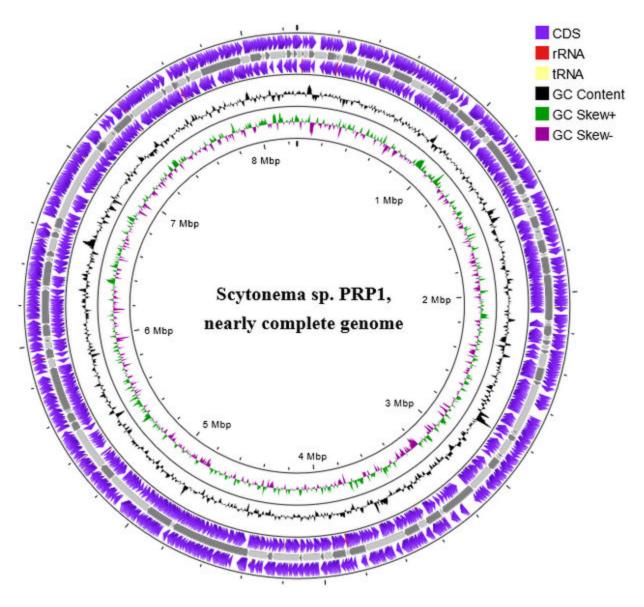
FIG 1 Circular genome of *Scytonema* sp. PRP1 created via Proksee, indicating (outer to inner) CDS on the forward and reverse strands (indigo), tRNAs (yellow), rRNAs (red), GC content (black), positive GC skew (green), and negative GC skew (purple).

GTDB-Tk v2.1.1 (12) assigned PRP1 to the *Scytonema* genus, identifying *Scytonema sp.* NIES-4073 (GenBank accession AP018268.1) as its closest relative (FastANI: 95.8%). This is supported by the whole genome phylogenetic tree generated by the Type (Strain) Genome Server (TYGS) (13), suggesting PRP1 and NIES-4073 are distinct species originating from the same common ancestor.

AUTHOR AFFILIATIONS

[1]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA
[2]Lieber Institute for Brain Development, Baltimore, Maryland, USA

³Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA

AUTHOR ORCIDs

Preston Parana http://orcid.org/0009-0000-4887-6560
Camille Mumm http://orcid.org/0000-0002-7585-9851
Michael J. McConnell http://orcid.org/0000-0001-6196-0791
Alan P. Boyle http://orcid.org/0000-0002-2081-1105

DATA AVAILABILITY

The whole genome shotgun project of *Scytonema* sp. PRP1 has been deposited in GenBank under the accession no. JBAHYJ000000000, and the raw data has been deposited under the accession no. SRR27967422.

REFERENCES

1. McConnell MJ, Moran JV, Abyzov A, Akbarian S, Bae T, Cortes-Ciriano I, Erwin JA, Fasching L, Flasch DA, Freed D, et al. 2017. Intersection of diverse neuronal genomes and neuropsychiatric disease: the brain somatic mosaicism network. Science 356:eaal1641. https://doi.org/10.1126/science.aal1641
2. Burbulis IE, Wierman MB, Wolpert M, Haakenson M, Lopes M-B, Schiff D, Hicks J, Loe J, Ratan A, McConnell MJ. 2018. Improved molecular karyotyping in glioblastoma. Mutat Res 811:16–26. https://doi.org/10.1016/j.mrfmmm.2018.06.002
3. Zhou W, Mumm C, Gan Y, Switzenberg JA, Wang J, De Oliveira P, Kathuria K, Losh SJ, McDonald TL, Bessell B, Van Deynze K, McConnell MJ, Boyle AP, Mills RE. 2024. A personalized multi-platform assessment of somatic mosaicism in the human frontal cortex. bioRxiv. https://doi.org/10.1101/2024.12.18.629274
4. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genom 3:e000132. https://doi.org/10.1099/mgen.0.000132
5. Laetsch DR, Blaxter ML. 2017. BlobTools: interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. F1000Res 6:1287. https://doi.org/10.12688/f1000research.12232.1
6. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421
7. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595
8. medaka: sequence correction provided by ONT Research. Available from: https://github.com/nanoporetech/medaka. Retrieved 26 Feb 2024.
9. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34:i142–i150. https://doi.org/10.1093/bioinformatics/bty266
10. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38:4647–4654. https://doi.org/10.1093/molbev/msab199
11. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569
12. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2022. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics 38:5315–5316. https://doi.org/10.1093/bioinformatics/btac672
13. Meier-Kolthoff JP, Göker M. 2019. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat Commun 10:2182. https://doi.org/10.1038/s41467-019-10210-3