AMERICAN SOCIETY FOR MICROBIOLOGY

Microbiology® Resource Announcements

GENOME SEQUENCES

Chromosome and Plasmid Sequences of *Pantoea* sp. Strain SOD02 Isolated from an Urban Freshwater Stream

Ibrahim Abaherah,[a] Geraldine Almonte,[a] Kathryn Donohue,[a] Daniel Furey,[a] Kathleen Garvey,[a] Jacqueline Joyal,[a] Devyn Luden,[a] Kathryn Mulvey,[a] Sean O'Donnell,[a] Nina Pitre,[a] Aura Rexach,[a] Kiley Sheehan,[a] Laura E. Williams[a]

[a]Department of Biology, Providence College, Providence, Rhode Island, USA

ABSTRACT After isolating *Pantoea* sp. strain SOD02 from an urban freshwater stream in Providence, RI, we used PacBio RSII data for *de novo* assembly and Illumina MiSeq data for polishing. This yielded complete circular sequences for a 4,227,027-bp chromosome with 54.7% GC and a 926,844-bp plasmid with 54.0% GC.

To obtain bacteria for use as prey in our studies of predatory *Bdellovibrio* (1), we swabbed water from an urban stream in Providence, RI (41.835°N, 71.443°W), onto Trypticase soy agar and incubated at 28°C. After three rounds of picking and streaking, we grew pure culture overnight in Trypticase soy broth (TSB) at 28°C and then combined the culture 1:1 with 50% glycerol to establish freezer stocks.

We used the Wizard genomic DNA purification kit (Promega, Madison, WI) to extract DNA from separate overnight cultures grown from freezer stock in TSB at 28°C. The University of Maryland Institute for Genome Sciences used one extraction for long-read sequencing, which involved shearing DNA using a g-TUBE at 3,400 rpm, size selection on a Blue Pippin instrument with an 11,000-bp cutoff, library preparation using SMRTbell template prep kit 1.0, and sequencing one SMRT (single-molecule real-time) cell on PacBio RS II with P6-C4 chemistry. The University of Rhode Island Genomics and Sequencing Center used another extraction for short-read sequencing, which involved shearing DNA using a Covaris S220 focused ultrasonicator, library preparation using PrepX DNA library kit, visualization on high-sensitivity BioAnalyzer chips, quantification with the KAPA Illumina quantification kit, and sequencing on Illumina MiSeq to obtain 2× 250-bp paired-end reads.

Unless otherwise noted, default parameters were used for all software. We compared Hierarchical Genome Assembly Process v3 (HGAP3) (2) and Canu 2.2 (3) for *de novo* assembly of PacBio data (119,255 subreads at an N_{50} of 13,141 bp). We tested HGAP3 with estimated genome size 4.5 Mbp, which generated two contigs. After identifying and trimming overlap between contig ends with BLASTN (4) and EMBOSS 6.6.0.0 (5) extractseq, the circularized contigs were 4,226,901 and 926,802 bp. We tested Canu with an estimated genome size of 4.25 Mbp, which generated two contigs. After trimming overlap with extractseq based on Canu information, the circularized contigs were 4,226,861 and 926,798 bp. Assemblies were almost identical by dnadiff (6), with 198 indel bp between chromosome contigs and one single nucleotide variant (SNV) and 84 indel bp between plasmid contigs. To proceed, we used the circularized contigs from Canu and rotated them to start at *dnaA* for the chromosome and *repB* for the plasmid.

For polishing, we separately aligned raw Illumina MiSeq read 1 (R1) and read 2 (R2) datasets (5,181,180 reads each) to the circularized and rotated contigs using the Burrows-Wheeler aligner "mem" (BWA-mem) 0.7.17 (7) with the option to report all possible alignments for each read. Samtools (8) analysis showed 99.3% R1 and 97.8% R2 reads aligned to at least one location. Using Polypolish v0.5.0 (9), we removed

Editor David Rasko, University of Maryland School of Medicine

Address correspondence to Laura E. Williams, lwillia7@providence.edu.

The authors declare no conflict of interest.

Received 18 August 2022
Accepted 12 September 2022
Published 21 September 2022

alignments based on insert size and then corrected the sequence. For the chromosome, Polypolish reported 437× coverage and corrected 168 indel bp. For the plasmid, Polypolish reported 375× coverage and corrected 48 indel bp. To confirm, we aligned MiSeq reads to corrected contigs using POLCA within MaSuRCA 4.0.8 (10), which made no additional corrections. The chromosome is 4,227,027 bp (54.7% GC) with 3,816 protein-coding genes, 78 tRNAs, and 22 rRNAs predicted by annotation with PGAP version 6.2 (11). The plasmid is 926,844 bp (54.0% GC) with 796 protein-coding genes. Digital DNA:DNA hybridization analysis using the Type Strain Genome Server (12) classified SOD02 in the bacterial genus *Pantoea*.

Data availability. *Pantoea* sp. strain SOD02 sequences have been deposited in GenBank under chromosome no. CP102604 and plasmid no. CP102605. PacBio and MiSeq reads have been deposited in the SRA under BioProject no. PRJNA866139 and SRA no. SRX16926094 and SRX16926095, respectively.

ACKNOWLEDGMENTS

This research was conducted as part of an undergraduate course in genomics during the spring 2020 and spring 2022 semesters at Providence College. L.E.W. was the course instructor for both semesters. Independently of the course, S.O. isolated the strain. All other authors were undergraduate students in the course and contributed equally to the project.

We thank Lisa Sadzewicz and Luke Tallon at the Institute for Genome Sciences at the University of Maryland Baltimore for PacBio sequencing services and Janet Atoyan at the Genomics and Sequencing Center at the University of Rhode Island for Illumina sequencing services.

This research was supported by an Institutional Development Award (IdeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant no. P20GM103430 and by funding from Providence College. This research is based upon work conducted at a Rhode Island NSF EPSCoR research facility, the Genomics and Sequencing Center, supported in part by the National Science Foundation EPSCoR Cooperative Agreement EPS-1004057. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

We declare no conflict of interest.

REFERENCES

1. Williams LE, Cullen N, DeGiorgis JA, Martinez KJ, Mellone J, Oser M, Wang J, Zhang Y. 2019. Variation in genome content and predatory phenotypes between *Bdellovibrio* sp. NC01 isolated from soil and *B. bacteriovorus* type strain HD100. Microbiology (Reading) 165:1315–1330. https://doi.org/10.1099/mic.0.000861.
2. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10:563–569. https://doi.org/10.1038/nmeth.2474.
3. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116.
4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
5. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.
6. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12. https://doi.org/10.1186/gb-2004-5-2-r12.
7. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
8. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
9. Wick RR, Holt KE. 2022. Polypolish: short-read polishing of long-read bacterial genome assemblies. PLoS Comput Biol 18:e1009802. https://doi.org/10.1371/journal.pcbi.1009802.
10. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. Bioinformatics 29:2669–2677. https://doi.org/10.1093/bioinformatics/btt476.
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. 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.