AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

PROKARYOTES

Complete Genome Sequence of *Escherichia coli* ML35

Angeline Casale,[a] Stephanie Clark,[a] Melissa Grasso,[a] Marta Kryschuk,[a] Lukas Ritzer,[a] Madyson Trudeau,[a] Laura E. Williams[a]

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

ABSTRACT We report here the complete genome sequence of *Escherichia coli* strain ML35. We assembled PacBio reads into a single closed contig with 169× mean coverage and then polished this contig using Illumina MiSeq reads, yielding a 4,918,774-bp sequence with 50.8% GC content.

Escherichia coli strain ML35 was isolated during studies of *lac* operon gene expression in the 1950s (1). ML35 does not synthesize lactose permease, but it constitutively expresses β-galactosidase (2). Since its isolation, ML35 has been used in a variety of experiments, including the investigation of interactions between *E. coli* and predatory bacteria (3). Williams and coworkers (4) are using ML35 and other *E. coli* strains to test the prey range of predatory bacteria. Comparative genomics will help us understand how genome variation within a prey species impacts variation in predation phenotypes.

We extracted genomic DNA from 3 ml of overnight culture grown in Trypticase soy broth at 37°C using the Wizard genomic DNA purification kit (Promega). Aliquots were used by the University of Maryland Institute for Genome Sciences to construct a PacBio library and by the University of Rhode Island Genomics and Sequencing Center to construct an Illumina library. Sequencing on a PacBio RS II instrument using P6-C4 chemistry yielded 93,133 subreads, with an N_{50} value of 12,583 bp, from two single-molecule real-time (SMRT) cells. For *de novo* assembly, we launched an Amazon EC2 instance of SMRT Portal version 2.3.0 and used the Hierarchical Genome Assembly Process version 3 (HGAP3) (5) with an estimated genome size 4.5 Mb and a target coverage of 30×. This generated contigs of 4,964,530 bp and 18,915 bp, with 169× and 18× mean coverages, respectively. The small contig is highly similar to regions of the large contig. Combined with its low coverage, this suggests that the small contig is an assembly artifact; therefore, we discarded it. To circularize the large contig, we used Gepard (6) to visualize overlap between the ends of the contig and BLAST (7) and EMBOSS extractseq (8) to specify coordinates and trim overlap, thereby generating a closed 4,918,091-bp contig.

To polish the closed contig, we processed 2 × 250-bp Illumina MiSeq reads using SolexaQA++ version 3.1.4 (9). We removed bases that had a quality score of <13 with DynamicTrim and then discarded reads that had <100 bp with LengthSort. This yielded 5,366,007 read pairs. Using the Burrows-Wheeler aligner "mem" (BWA-mem) algorithm version 0.7.13 (10), we mapped 94.8% of these reads to the closed contig. We sorted and indexed the alignment file with SAMtools (11) and then used Pilon version 1.22 (12) to identify and correct 717 small indels, yielding a corrected 4,918,780-bp contig. To confirm this sequence, we used the same Illumina MiSeq reads and DynamicTrim quality score cutoff but adjusted the LengthSort cutoff to 75 bp. After aligning these reads to the corrected contig, Pilon identified eight discrepancies, which we manually examined and corrected to generate the final genome sequence of 4,918,774 bp with 50.8% GC content.

Annotation with the Prokaryotic Genome Annotation Pipeline (PGAP) predicted 4,782 protein-coding sequences, 757 of which are annotated as hypothetical proteins, along with 95 tRNAs and 7 rRNA operons. By comparing the ML35 genome to that of

Received 10 January 2018 **Accepted** 16 January 2018 **Published** 15 February 2018

Citation Casale A, Clark S, Grasso M, Kryschuk M, Ritzer L, Trudeau M, Williams LE. 2018. Complete genome sequence of *Escherichia coli* ML35. Genome Announc 6:e00034-18. https://doi.org/10.1128/genomeA.00034-18.

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

E. coli MG1655 (GenBank accession no. NC_000913), we identified an 11-bp insertion in ML35's *lacY* gene that causes a frameshift and a nonsynonymous substitution in ML35's *lacI* gene that causes a V24E replacement, which is reported to impact the repressor protein function (13). These mutations may explain the Lac phenotype observed for ML35.

Accession number(s). This complete genome sequence has been deposited in GenBank under the accession no. CP025747. The version described in this paper is the first version, CP025747.1.

ACKNOWLEDGMENTS

This research was conducted as part of an undergraduate course in genomics during the fall 2017 semester at Providence College. All authors (with the exception of L.E.W.) were undergraduate students in the course and contributed equally to the project.

We thank Mark O. Martin for providing ML35. 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. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. Buttin G, Cohen GN, Monod J, Rickenberg HV. 1956. Galactoside-permease of *Escherichia coli*. Ann Inst Pasteur 91:829–857.
2. Zabin I, Kepes A, Monod J. 1959. On the enzymic acetylation of isopropyl-β-D-thiogalactoside and its association with galactoside-permease. Biochem Biophys Res Commun 1:289–292. https://doi.org/10.1016/0006-291X(59)90040-3.
3. Rittenberg SC, Shilo M. 1970. Early host damage in the infection cycle of *Bdellovibrio bacteriovorus*. J Bacteriol 102:149–160.
4. Enos BG, Anthony MK, DeGiorgis JA, Williams LE. 2018. Prey range and genome evolution of *Halobacteriovorax marinus* predatory bacteria from an estuary. mSphere 3:e00508-17. https://doi.org/10.1128/mSphere.00508-17.
5. 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.
6. Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23:1026–1028. https://doi.org/10.1093/bioinformatics/btm039.
7. 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.
8. 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.
9. Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485. https://doi.org/10.1186/1471-2105-11-485.
10. 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.
11. 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.
12. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal.pone.0112963.
13. Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH. 1994. Genetic studies of the *lac* repressor. XIV. Analysis of 4000 altered *Escherichia coli lac* repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. J Mol Biol 240:421–433. https://doi.org/10.1006/jmbi.1994.1458.