AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Complete Genome Sequence of *Escherichia coli* BL21-AI

Niketa Bhawsinghka,[a] Katie F. Glenn,[a] Roel M. Schaaper[a]

[a]Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA

ABSTRACT *Escherichia coli* BL21-AI is a commercially available strain possessing a phage T7-based protein-expression system. A combination of tight regulation and high yield makes it widely used for high-level expression of toxic recombinant proteins. Here, we present the complete genome sequence of BL21-AI and provide insights on its genome.

Escherichia coli B is widely used for the induced expression of recombinant proteins. It lacks certain proteases and has an additional type II secretion system and different cell wall and outer membrane composition, making it more favorable for protein secretion (1). Examples of commonly used *E. coli* B strains are BL21 and BL21(DE3) (2). However, the leaky nature of these expression systems makes them less suitable for the induction of deleterious or toxic proteins. One strain developed specifically to combat leaky expression is *E. coli* BL21-AI. In this strain, expression of plasmid-borne genes is driven by a chromosomally encoded bacteriophage T7 RNA polymerase (T7-RNAP), which in turn is controlled by the tightly regulated arabinose-inducible P_{BAD} promoter (3, 4). The strain is produced by Invitrogen (Thermo Fisher Scientific), but its precise construction has not been disclosed due to proprietary considerations. Nevertheless, it is derived from strain BL21, and the two genome sequences are expected to be similar except for *araB*, the insertion locus of the T7-RNAP cassette and linked tetracycline-resistance marker. To aid in future investigations using BL21-AI, including the construction of related strains, we deemed it necessary to determine its exact genome sequence. Here, we describe the complete genome sequence of BL21-AI and compare it with that of BL21 (GenBank accession number CP010816).

BL21-AI competent cells were purchased from Invitrogen (catalogue number C607003, lot number 2067410) and streaked onto Luria agar plates containing tetracycline (15 μg/ml). Following incubation at 37°C for 16 h, two single colonies were used to inoculate LB liquid and were grown to saturation. Genomic DNA was extracted using Qiagen Genomic-tip 100/G, and whole-genome sequencing was performed at the National Human Genome Research Institute of the NIH. PacBio barcoded libraries were generated (SMRTbell Express template prep kit v2.0, Sequel II 16-overhang adapter barcode kit), pooled, and run on a PacBio Sequel II single-molecule real-time (SMRT) cell for 20 h using the continuous long-read sequencing mode. Demultiplexed subreads (825,284 and 931,154 reads with mean lengths of 7,908 bp and 8,354 bp, respectively, for two isolates) were assembled into a single contig with the PacBio SMRTLink v7.0 HGAP4 pipeline (5). Quality control of the sequence reads was performed using the default input filtering parameters in the pipeline. The contig was circularized using Circlator v1.5.3 (6) ("all" mode, "threads 8") followed by two rounds of consensus polishing with the SMRTLink v7.0 resequencing pipeline (https://github.com/PacificBiosciences/GenomicConsensus). Two independent colonies of BL21-AI were processed in parallel, yielding identical genome sequences with average coverage depths of 1,341× and 1,600×, respectively. Genome annotation was carried out using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v4.10 (7). Default parameters were used for all software unless otherwise specified.

Citation Bhawsinghka N, Glenn KF, Schaaper RM. 2020. Complete genome sequence of *Escherichia coli* BL21-AI. Microbiol Resour Announc 9:e00009-20. https://doi.org/10.1128/MRA.00009-20.

Editor David Rasko, University of Maryland School of Medicine

Address correspondence to Roel M. Schaaper, schaaper@niehs.nih.gov.

Received 10 January 2020
Accepted 12 February 2020
Published 5 March 2020

The genome of BL21-AI is 4,530,564 bp long with a G+C content of 50.8%. Aligning the sequence (using the progressiveMauve algorithm) (8) with BL21 showed one large block of correspondence, confirming the overall relatedness of the two genomes, with the exception of a 4,084-bp deletion/insertion segment at BL21-AI coordinates 71213 to 75296. This region of nonconformity results from deletion of the BL21 *araB* gene (BL21 coordinates 71213 to 72853) and replacement by the T7-RNAP gene along with a *tetA* cassette, accounting for the observed tetracycline resistance of BL21-AI. In addition, the Mauve alignment revealed a total of 231 additional single-base substitutions (BL21-AI 68837 to 88634) flanking either side of the *ara-tet*-T7*pol* insert. These substitutions match the corresponding sequence of *E. coli* K-12 strain MG1655 (GenBank accession number U00096.3) spanning a total of 14 genes from *araD* to *leuA*, with the exception of a G⟶A mutation in *sgrR* (coordinate 81500) changing a Gln codon (CAA) to a termination codon (TAA). Overall, we infer that to generate BL21-AI from BL21, a Δ*araB*::T7RNA*pol-tetA* cassette was transferred from a K-12 strain by P1 transduction, resulting in the transfer of a surrounding ~20-kb chromosomal donor DNA segment.

Data availability. The complete genome sequence of *Escherichia coli* BL21-AI has been deposited at GenBank under accession number CP047231. The raw sequence data are available under SRA accession number SRR10851446.

ACKNOWLEDGMENTS

This research was supported by project number Z01 ES101905 of the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

We acknowledge Greg Solomon of the NIEHS Epigenomics and DNA Sequencing Core Laboratory, Adam Burkholder of the Office of Environmental Science Cyberinfrastructure, and Zhixiong Zhou of the NIEHS Genome Integrity and Structural Biology Laboratory for their DNA sequencing and bioinformatics assistance and review of this paper.

REFERENCES

1. Yoon SH, Han M-J, Jeong H, Lee CH, Xia X-X, Lee D-H, Shim JH, Lee SY, Oh TK, Kim JF. 2012. Comparative multi-omics systems analysis of Escherichia coli strains B and K-12. Genome Biol 13:R37. https://doi.org/10.1186/gb-2012-13-5-r37.
2. Studier FW, Moffatt BA. 1986. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol 189:113–130. https://doi.org/10.1016/0022-2836(86)90385-2.
3. Lee N, Francklyn C, Hamilton EP. 1987. Arabinose-induced binding of AraC protein to aral2 activates the araBAD operon promoter. Proc Natl Acad Sci U S A 84:8814–8818. https://doi.org/10.1073/pnas.84.24.8814.
4. Miyada CG, Stoltzfus L, Wilcox G. 1984. Regulation of the araC gene of Escherichia coli: catabolite repression, autoregulation, and effect on araBAD expression. Proc Natl Acad Sci U S A 81:4120–4124. https://doi.org/10.1073/pnas.81.13.4120.
5. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13:1050–1054. https://doi.org/10.1038/nmeth.4035.
6. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol 16:294–294. https://doi.org/10.1186/s13059-015-0849-0.
7. 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.
8. Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147. https://doi.org/10.1371/journal.pone.0011147.