genomeA
Journals.ASM.org

A Gapless, Unambiguous Genome Sequence of the Enterohemorrhagic *Escherichia coli* O157:H7 Strain EDL933

Haythem Latif,[a] Howard J. Li,[a] Pep Charusanti,[a] Bernhard Ø. Palsson,[a] Ramy K. Aziz[a,b]

Bioengineering Department, University of California San Diego, La Jolla, California, USA[a]; Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, Egypt[b]

***Escherichia coli* EDL933 is the prototypic strain for enterohemorrhagic *E. coli* serotype O157:H7, associated with deadly food-borne outbreaks. Because the publicly available sequence of the EDL933 genome has gaps and >6,000 ambiguous base calls, we here present an updated high-quality, unambiguous genome sequence with no assembly gaps.**

Received 23 July 2014 **Accepted** 28 July 2014 **Published** 14 August 2014

Citation Latif H, Li HJ, Charusanti P, Palsson BØ, Aziz RK. 2014. A gapless, unambiguous genome sequence of the enterohemorrhagic *Escherichia coli* O157:H7 strain EDL933. Genome Announc. 2(4):e00821-14. doi:10.1128/genomeA.00821-14.

Address correspondence to Ramy K. Aziz, ramy.aziz@gmail.com.

Escherichia coli serotype O157: H7, a causative agent in food poisoning outbreaks leading to hemorrhagic colitis or hemolytic uremic syndrome, gained public attention following its association with an outbreak in 1993 related to the U.S. fast-food chain Jack-in-the-Box (1, 2) and another large outbreak among schoolchildren in Sakai, Japan in 1996 (3). Moreover, *E. coli* O157:H7 strains are known for their prophage-rich genomes and are currently considered the bacterial genomes with the largest number of integrated phages (4, 5). Strain EDL933 (ATCC 43895), isolated from ground beef linked to a massive hamburger outbreak in Michigan, USA in 1982 (6), is the prototypic reference strain representing this pathotype.

Although the full genome of EDL933 was sequenced and published in 2001 (5), the deposited assembled genome has >6,000 ambiguous base calls and a chromosomal gap of 4,000 bp. While the utility of this reference genome, cited in >3,200 publications, is indisputable, several analyses reliant on a pristine reference (e.g., single nucleotide polymorphism studies) are hindered by those ambiguities and gaps. EDL933 has long phage-associated repeat regions >7 kb. Microbial genomes with these characteristics are the most complex to assemble (7), so we resorted to single-molecule sequencing using PacBio followed by polishing using Illumina short-reads to complete the EDL933 sequence. This produced a gapless genome assembly, with no ambiguous base calls, and an updated genome annotation.

Genomic DNA from the EDL933 strain was prepared for PacBio and Illumina sequencing. PacBio libraries were prepared according to standard library preparation procedures with Blue Pippen size selection for >20-kb fragments and sequenced using P5/C3 chemistry and 3 h movies on the RS II system at the UCSD Genomics Core, San Diego, CA. Illumina libraries were prepared according to the TruSeq DNA PCR-Free sample preparation kit protocol (Illumina) and paired-end sequenced (2×250) on a MiSeq. SMRTAnalaysis 2.2.0 HGAP v2 assembly of PacBio reads (66,927) produced three polished contigs: two corresponding to the chromosome and one the plasmid. When compared with the reference, NC_002655, a region of high read density within the two chromosomal contigs was shown to be a large duplication that unites the two contigs. After the plasmid and chromosome were circularized, reads were mapped back to the assembled sequences to check for variants by first using Bridge Mapper (RS_BridgeMapper.1) with PacBio reads and then Breseq v0.24rc6 (8) with Illumina short reads. Coverage was ~100× for PacBio data and ~300 for Illumina data. The final assembled genome was automatically annotated, then manually corrected, through the RAST server using SEED annotation tools (9, 10).

The updated EDL933 genome consists of a 5,547,323-bp chromosome and a 92,076-bp plasmid compared with 5,528,445 and 92,077 bp in the current EDL933 assembly. This gapless assembly eliminates 6,641 ambiguous base calls in the current EDL933 chromosome including 2,413 non-N ambiguous bases and 4,000 N's belonging to a chromosomal gap. The updated genome has 5,675 and 97 annotated coding sequences (CDSs) compared with 5,286 and 99 CDSs found in the current reference chromosome and plasmid, respectively.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited in GenBank under the accession numbers CP008957 and CP008958. The RAST version is available at http://rast.nmpdr.org under job 157998.

ACKNOWLEDGMENTS

This work was supported by the National Institute of Health under grant GM098105. Haythem Latif is supported through the National Science Foundation Graduate Research Fellowship under grant DGE1144086.

REFERENCES

1. **Chen J, Griffiths MW.** 1999. Cloning and sequencing of the gene encoding universal stress protein from *Escherichia coli* O157:H7 isolated from Jack-in-a-Box outbreak. Lett. Appl. Microbiol. **29:**103–107. http://dx.doi.org/10.1046/j.1365-2672.1999.00594.x.
2. **Pennington H.** 2010. *Escherichia coli* O157. Lancet **376:**1428–1435. http://dx.doi.org/10.1016/S0140-6736(10)60963-4.
3. **Michino H, Araki K, Minami S, Takaya S, Sakai N, Miyazaki M, Ono A, Yanagawa H.** 1999. Massive outbreak of *Escherichia coli* O157:H7 infection in schoolchildren in Sakai City, Japan, associated with consump-

tion of white radish sprouts. Am. J. Epidemiol. **150:**787–796. http://dx.doi.org/10.1093/oxfordjournals.aje.a010082.

4. **Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H.** 2003. Prophage genomics. Microbiol. Mol. Biol. Rev. **67:**238–276. http://dx.doi.org/10.1128/MMBR.67.2.238-276.2003.
5. **Perna NT, Plunkett G III, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR.** 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157 H7. Nature **409:**529–533.
6. **Riley LW, Remis RS, Helgerson SD, McGee HB, Wells JG, Davis BR, Hebert RJ, Olcott ES, Johnson LM, Hargrett NT, Blake PA, Cohen ML.** 1983. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. N. Engl. J. Med. **308:**681–685. http://dx.doi.org/10.1056/NEJM198303243081203.
7. **Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, McVey SD, Radune D, Bergman NH, Phillippy AM.** 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol. **14:**R101. http://dx.doi.org/10.1186/gb-2013-14-9-r101.
8. **Deatherage DE, Barrick JE.** 2014. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol. Biol. **1151:**165–188. http://dx.doi.org/10.1007/978-1-4939-0554-6_12.
9. **Aziz RK, Devoid S, Disz T, Edwards RA, Henry CS, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Stevens RL, Vonstein V, Xia F.** 2012. SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. PLoS One **7:**e48053. http://dx.doi.org/10.1371/journal.pone.0048053.
10. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9:**75. http://dx.doi.org/10.1186/1471-2164-9-75.