AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

Application of Long Sequence Reads To Improve Genomes for *Clostridium thermocellum* AD2, *Clostridium thermocellum* LQRI, and *Pelosinus fermentans* R7

Sagar M. Utturkar,[a] Edward A. Bayer,[b] Ilya Borovok,[c] Raphael Lamed,[c] Richard A. Hurt,[d] Miriam L. Land,[d] Dawn M. Klingeman,[d,e] Dwayne Elias,[d] Jizhong Zhou,[f] Marcel Huntemann,[g] Alicia Clum,[g] Manoj Pillay,[g] Krishnaveni Palaniappan,[g] Neha Varghese,[g] Natalia Mikhailova,[g] Dimitrios Stamatis,[g] T. B. K. Reddy,[g] Chew Yee Ngan,[g] Chris Daum,[g] Nicole Shapiro,[g] Victor Markowitz,[g] Natalia Ivanova,[g] Nikos Kyrpides,[g] Tanja Woyke,[g] Steven D. Brown[a,d,e]

Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, Tennessee, USA[a]; Department of Biological Chemistry, The Weizmann Institute of Science, Rehovot, Israel[b]; Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Ramat Aviv, Israel[c]; Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA[d]; BioEnergy Science Center, Oak Ridge, Tennessee, USA[e]; Institute for Environmental Genomics, University of Oklahoma, Norman, Oklahoma, USA[f]; DOE Joint Genome Institute, Walnut Creek, California, USA[g]

We and others have shown the utility of long sequence reads to improve genome assembly quality. In this study, we generated PacBio DNA sequence data to improve the assemblies of draft genomes for *Clostridium thermocellum* AD2, *Clostridium thermocellum* LQRI, and *Pelosinus fermentans* R7.

Received 3 August 2016 **Accepted** 8 August 2016 **Published** 29 September 2016

Citation Utturkar SM, Bayer EA, Borovok I, Lamed R, Hurt RA, Land ML, Klingeman DM, Elias D, Zhou J, Huntemann M, Clum A, Pillay M, Palaniappan K, Varghese N, Mikhailova N, Stamatis D, Reddy TBK, Ngan CY, Daum C, Shapiro N, Markowitz V, Ivanova N, Kyrpides N, Woyke T, Brown SD. 2016. Application of long sequence reads to improve genomes for *Clostridium thermocellum* AD2, *Clostridium thermocellum* LQRI, and *Pelosinus fermentans* R7. Genome Announc 4(5):e01043-16. doi:10.1128/genomeA.01043-16.

Address correspondence to Steven D. Brown, brownsd@ornl.gov.

Draft genome sequences have been generated for *Clostridium thermocellum* AD 2 (1), *Clostridium thermocellum* LQRI (DSM 2360) (2), and *Pelosinus fermentans* R7 (3), which encompassed 131, 110, and 65 contigs, respectively.

Clostridium (*Ruminiclostridium*) *thermocellum* strains are known for their potent cellulolytic capabilities (4). *C. thermocellum* strain AD2 is derived from a cellulose adhesion-defective (AD) mutant strain and played a critical role in describing the original cellulosome concept (5). *C. thermocellum* strain LQRI is an LQ8 reisolate used in enzyme studies to propose electron flow paths (6), and has been mistakenly called LQR1 on occasion. A genome for LQ8 (DSM 1313) has been reported (7). *P. fermentans* type strain R7 belongs to the *Negativicutes* within the *Firmicutes* phylum and in the presence of a fermentable substrate it can reduce Fe(III) (8). Complete genomes for *Pelosinus fermentans* JBW45 (9) and *Pelosinus* sp. strain UFO1 (10) have recently been reported using only single-molecule DNA sequencing technology, and the utility of long read sequences has been shown to improve other microbial genome assemblies (11–15).

In this study, genomic DNA of all three strains underwent Pacific Biosciences RSII standard template preparation and sequencing. Raw reads were assembled using the HGAP (version: 2.3.0) protocol. The HGAP assembly of AD2 contained 10 contigs and the 3.55-Mb finished genome was generated by super-assembly using the Geneious (version 8.1.6) software combined with PCR and Sanger, as described previously (14). The HGAP assembly for LQRI contained two contigs, totaling 3.61 Mbp in size, with an input read coverage of 166.7. A small duplicated 12-kb LQRI artifact contig was removed from the assembly, leaving a circular chromosome of 3.57 Mb. The final HGAP assembly for the R7 genome contained two contigs totaling 5.02 Mb in genome size. Genes for all three genomes were identified using Prodigal (16) and annotations were performed as described previously (17). A total of 3,055, 3,071, and 4,631 protein-coding genes were identified in the AD2, LQRI, and R7 genomes, respectively.

A prior comparison of single-molecule sequencing-based genome assembly to short-read and hybrid assembly approaches showed that assemblies based on shorter-read technologies were confounded by a large number of longer repeats, in particular multiple copies of ~5-kb rRNA gene operons (11). Consistent with this observation, earlier drafts of AD2, LQRI, and R7 genome sequences contained single copies of the rRNA genes. The improved genomes for the *C. thermocellum* strains reported in this study each contain four copies of the 5S, 16S, and 23S rRNA genes. The improved *P. fermentans* R7 genome reported in this study contains 12, nine, and nine copies of the 5S, 16S, and 23S rRNA genes, respectively.

Finished or near-finished genome assemblies for these strains represent a significant improvement over earlier draft assemblies. We expect that the protein-coding potential will be superior (14) and that these genomes will facilitate comparative and functional genomic studies. Lastly, these new genome sequences will also facilitate a broader and more detailed comparison between sequencing and assembly technologies.

Accession number(s). This whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under accession numbers CP013828, CP016502, and AKVN00000000 for AD2, LQRI, and R7, respectively.

FUNDING INFORMATION

This work, including the efforts of Steven D. Brown, was funded by DOE | SC | Biological and Environmental Research (BER).

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. This research was supported by the Bioenergy Science Center (BESC), which is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. ORNL is managed by UT-Battelle, LLC, Oak Ridge, TN, for the DOE under contract DE-AC05-00OR22725.

REFERENCES

1. **Brown SD, Lamed R, Morag E, Borovok I, Shoham Y, Klingeman DM, Johnson CM, Yang Z, Land ML, Utturkar SM, Keller M, Bayer EA.** 2012. Draft genome sequences for *Clostridium thermocellum* wild-type strain YS and derived cellulose adhesion-defective mutant strain AD 2. J Bacteriol **194:**3290–3291. http://dx.doi.org/10.1128/JB.00473-12.
2. **Hemme CL, Mouttaki H, Lee Y-J, Zhang G, Goodwin L, Lucas S, Copeland A, Lapidus A, Glavina del Rio T, Tice H, Saunders E, Brettin T, Detter JC, Han CS, Pitluck S, Land ML, Hauser LJ, Kyrpides N, Mikhailova N, He Z, Wu L, Van Nostrand JD, Henrissat B, He Q, Lawson PA, Tanner RS, Lynd LR, Wiegel J, Fields MW, Arkin AP, Schadt CW, Stevenson BS, McInerney MJ, Yang Y, Dong H, Xing D, Ren N, Wang A, Huhnke RL, Mielenz JR, Ding S-Y, Himmel ME, Taghavi S, van der Lelie D, Rubin EM, Zhou J.** 2010. Sequencing of multiple clostridial genomes related to biomass conversion and biofuel production. J Bacteriol **192:**6494–6496. http://dx.doi.org/10.1128/JB.01064-10.
3. **Brown SD, Podar M, Klingeman DM, Johnson CM, Yang ZK, Utturkar SM, Land ML, Mosher JJ, Hurt RA, Phelps TJ, Palumbo AV, Arkin AP, Hazen TC, Elias DA.** 2012. Draft genome sequences for two metal-reducing *Pelosinus fermentans* strains isolated from a Cr(VI)-contaminated site and for type strain R7. J Bacteriol **194:**5147–5148. http://dx.doi.org/10.1128/JB.01174-12.
4. **Blumer-Schuette SE, Brown SD, Sander KB, Bayer EA, Kataeva I, Zurawski JV, Conway JM, Adams MW, Kelly RM.** 2014. Thermophilic lignocellulose deconstruction. FEMS Microbiol Rev **38:**393–448. http://dx.doi.org/10.1111/1574-6976.12044.
5. **Bayer EA, Kenig R, Lamed R.** 1983. Adherence of *Clostridium thermocellum* to cellulose. J Bacteriol **156:**818–827.
6. **Lamed R, Zeikus JG.** 1980. Ethanol production by thermophilic bacteria: relationship between fermentation product yields of and catabolic enzyme activities in *Clostridium thermocellum* and *Thermoanaerobium brockii*. J Bacteriol **144:**569–578.
7. **Feinberg L, Foden J, Barrett T, Davenport KW, Bruce D, Detter C, Tapia R, Han C, Lapidus A, Lucas S, Cheng J-F, Pitluck S, Woyke T, Ivanova N, Mikhailova N, Land M, Hauser L, Argyros DA, Goodwin L, Hogsett D, Caiazza N.** 2011. Complete genome sequence of the cellulolytic thermophile *Clostridium thermocellum* DSM1313. J Bacteriol **193:**2906–2907. http://dx.doi.org/10.1128/JB.00322-11.
8. **Shelobolina ES, Nevin KP, Blakeney-Hayward JD, Johnsen CV, Plaia TW, Krader P, Woodard T, Holmes DE, VanPraagh CG, Lovley DR.** 2007. *Geobacter pickeringii* sp. nov., *Geobacter argillaceus* sp. nov. and *Pelosinus fermentans* gen. nov., sp. nov., isolated from subsurface kaolin lenses. Int J Syst Evol Microbiol **57:**126–135. http://dx.doi.org/10.1099/ijs.0.64221-0.
9. **De León KB, Utturkar SM, Camilleri LB, Elias DA, Arkin AP, Fields MW, Brown SD, Wall JD.** 2015. Complete genome sequence of *Pelosinus fermentans* JBW45, a member of a remarkably competitive group of *Negativicutes* in the *Firmicutes* phylum. Genome Announc **3**(5):e01090-01015. http://dx.doi.org/10.1128/genomeA.01090-15.
10. **Brown SD, Utturkar SM, Magnuson TS, Ray AE, Poole FL, Lancaster WA, Thorgersen MP, Adams MW, Elias DA.** 2014. Complete genome sequence of *Pelosinus* sp. strain UFO1 assembled using single-molecule real-time DNA sequencing technology. Genome Announc **2**(5):e00881-00814. http://dx.doi.org/10.1128/genomeA.00881-14.
11. **Brown SD, Nagaraju S, Utturkar S, De Tissera S, Segovia S, Mitchell W, Land ML, Dassanayake A, Köpke M.** 2014. Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of *Clostridium autoethanogenum* and analysis of CRISPR systems in industrial relevant Clostridia. Biotechnol Biofuels **7:**40. http://dx.doi.org/10.1186/1754-6834-7-40.
12. **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.
13. **Utturkar SM, Klingeman DM, Bruno-Barcena JM, Chinn MS, Grunden AM, Köpke M, Brown SD.** 2015. Sequence data for *Clostridium autoethanogenum* using three generations of sequencing technologies. Scientific Data **2:**150014. http://dx.doi.org/10.1038/sdata.2015.14.
14. **Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, Brown SD.** 2014. Evaluation and validation of de novo and hybrid assembly techniques to derive high quality genome sequences. Bioinformatics **30:**2709–2716. http://dx.doi.org/10.1093/bioinformatics/btu391.
15. **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. http://dx.doi.org/10.1038/nmeth.2474.
16. **Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11:**119. http://dx.doi.org/10.1186/1471-2105-11-119.
17. **Woo HL, Utturkar S, Klingeman D, Simmons BA, DeAngelis KM, Brown SD, Hazen TC.** 2014. Draft genome sequence of the lignin-degrading *Burkholderia* sp. strain LIG30, isolated from wet tropical forest soil. Genome Announc **2**(3):e00637-14. http://dx.doi.org/10.1128/genomeA.00637-14.