2011

Complete Genome Sequence of the Cellulolytic Thermophile Clostridium thermocellum DSM1313

Lawrence Feinberg
Mascoma Corporation, lfeinberg@mascoma.com

Justine Foden
Mascoma Corporation

Trisha Barrett
Mascoma Corporation

Karen Walston Davenport
Los Alamos National Laboratory

David Bruce
Los Alamos National Laboratory

See next page for additional authors

JOURNAL OF BACTERIOLOGY, June 2011, p. 2906–2907 Vol. 193, No. 11
0021-9193/11/$12.00 doi:10.1128/JB.00322-11

Complete Genome Sequence of the Cellulolytic Thermophile *Clostridium thermocellum* DSM1313[▽]

Lawrence Feinberg,[1]* Justine Foden,[1] Trisha Barrett,[1] Karen Walston Davenport,[2] David Bruce,[2] Chris Detter,[3] Roxanne Tapia,[2] Cliff Han,[2] Alla Lapidus,[4] Susan Lucas,[3] Jan-Fang Cheng,[3] Samuel Pitluck,[3] Tanja Woyke,[3] Natalia Ivanova,[3] Natalia Mikhailova,[3] Miriam Land,[5] Loren Hauser,[5] D. Aaron Argyros,[1] Lynne Goodwin,[2] David Hogsett,[1] and Nicky Caiazza[1]

Mascoma Corporation, 67 Etna Rd., Suite 300, Lebanon, New Hampshire 03755[1]; Los Alamos National Laboratory, Los Alamos, New Mexico 87545[2]; DOE Joint Genome Institute, Walnut Creek, California 94598[3]; Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, Pennsylvania 19111[4]; and Oak Ridge National Laboratory, Computational Biology and Bioinformatics Group, P.O. Box 2008, MS 6420, Oak Ridge, Tennessee 37831[5]

Received 7 March 2011/Accepted 25 March 2011

***Clostridium thermocellum* DSM1313 is a thermophilic, anaerobic bacterium with some of the highest rates of cellulose hydrolysis reported. The complete genome sequence reveals a suite of carbohydrate-active enzymes and demonstrates a level of diversity at the species level distinguishing it from the type strain ATCC 27405.**

Clostridium thermocellum is a thermophilic, anaerobic bacterium of both fundamental and applied significance. The bacterium is highly cellulolytic, degrading cellulose through the action of a tethered, multienzyme complex called a cellulosome. The cellulosome of *C. thermocellum* is the best-characterized cellulase complex and serves as a paradigm for cellulolytic microorganisms (6, 7). In fermenting cellulose, *C. thermocellum* produces ethanol and organic acids as primary products. The fermentative capability of the bacterium has been the focus of several decades of research, owing to its potential to be used commercially in the consolidated bioprocessing of lignocellulosic material to ethanol and other products (6, 8). Recent advances in the development of genetic tools for *C. thermocellum* DSM1313 have enabled scientists to modify the cellulosome and primary metabolism of the bacterium (10, 11). The complete genome sequence of *Clostridium thermocellum* DSM 1313 expands the knowledge base for thermophilic, cellulolytic bacteria; enables comprehensive comparisons of closely related microbes; and facilitates further genetic engineering of the microbe.

The genome of *Clostridium thermocellum* was sequenced at the Joint Genome Institute (JGI) using a combination of Illumina (1) and 454 (9) technologies. An Illumina GAii shotgun library with reads of 36 bases, a 454 Titanium draft library with an average read length of 324.1 ± 200.9 bases, and a paired-end 454 library with an average insert size of 7.2 kb were generated for this genome. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov/. Illumina sequencing data were assembled with Velvet (12), and the consensus sequences were shredded into 1.5-kb overlapped fake reads and assembled together with the 454 data. Draft assemblies were based on 179.3 Mb of 454 draft data and all of the 454 paired-end data. Newbler parameters are -consed-a 50-l 350-g-m-ml 20.

The initial Newbler assembly contained 84 contigs in two scaffolds. We converted the initial 454 assembly into a Phrap assembly by making fake reads from the consensus, collecting the read pairs in the 454 paired-end library. The Phred/Phrap/Consed software package was used for sequence assembly and quality assessment (2, 3, 4) in the following finishing process. Illumina data were used to correct potential base errors and increase consensus quality using Polisher software developed at JGI (A. Lapidus, unpublished data). After the shotgun stage, reads were assembled with parallel Phrap (High Performance Software, LLC). Possible misassemblies were corrected with gapResolution (C. Han, unpublished data) or Dupfinisher (5) or by sequencing cloned bridging PCR fragments with subcloning. Gaps between contigs were closed by editing in Consed, by PCR, and by bubble PCR primer walks. A total of 227 additional reactions and 4 shatter libraries were necessary to close gaps and to raise the quality of the finished sequence.

Nucleotide sequence accession number. The final annotated genome of *C. thermocellum* DSM1313 has been deposited in GenBank under accession number CP002416.

This work was supported by the Bioenergy Science Center (BESC), Oak Ridge National Laboratory, a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research, under contract DE-PS02-06ER64304.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231.

REFERENCES

1. **Bennett, S.** 2004. Solexa Ltd. Pharmacogenomics **5:**433–438.
2. **Ewing, B., and P. Green.** 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:**186–194.
3. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:**175–185.

* Corresponding author. Mailing address: Mascoma Corporation, 67 Etna Rd., Suite 300, Lebanon, NH 03755. Phone: (603) 676-3320. Fax: (603) 676-3321. E-mail: lfeinberg@mascoma.com.

▽ Published ahead of print on 1 April 2011.

4. **Gordon, D., C. Abajian, and P. Green.** 1998. Consed: a graphical tool for sequence finishing. Genome Res. **8:**195–202.
5. **Han, C., and P. Chain.** 2006. Finishing repeat regions automatically with Dupfinisher, p. 141–146. *In* H. R. Arabnia and H. Valafar (ed.), Proceedings of the 2006 International Conference on Bioinformatics and Computational Biology. CSREA Press, Las Vegas, NV.
6. **Hogsett, D. A.** 1995. Cellulose hydrolysis and fermentation by Clostridium thermocellum for the production of ethanol. Ph.D. thesis. Dartmouth College, Hanover, NH.
7. **Lynd, L. R., P. J. Weimer, W. H. van Zyl, and I. S. Pretorius.** 2002. Microbial cellulose utilization: fundamentals and biotechnology. Microbiol. Mol. Biol. Rev. **66:**506–577.
8. **Lynd, L. R. et al.** 2008. How biotech can transform biofuels. Nat. Biotechnol. **26:**169–172.
9. **Margulies, M., et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376–380.
10. **Olson, D. G., et al.** 2010. Deletion of the Cel48S cellulase from *Clostridium thermocellum*. Proc. Natl. Acad. Sci. U. S. A. **107:**17727–17732.
11. **Tripathi, S. A., et al.** 2010. Development of pyrF-based genetic system for targeted gene deletion in *Clostridium thermocellum* and creation of a pta mutant. Appl. Environ. Microbiol. **76:**6591–6599.
12. **Zerbino, D., and E. Birney.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18:**821–829.