genomeA Journals.ASM.org

Complete Genome Sequence of the Extreme Thermophile *Dictyoglomus thermophilum* H-6-12

David A. Coil,[a] Jonathan H. Badger,[b] Heather C. Forberger,[c] Florenta Riggs,[d] Ramana Madupu,[e] Nadia Fedorova,[e] Naomi Ward,[f] Frank T. Robb,[g] Jonathan A. Eisen[a,h]

University of California Davis Genome Center, Davis, California, USA[a]; J. Craig Venter Institute, La Jolla, California, USA[b]; AstraZeneca Pharmaceuticals, Wilmington, Delaware, USA[c]; Silver Spring, Maryland, USA[d]; J. Craig Venter Institute, Rockville, Maryland, USA[e]; University of Wyoming, Laramie, Wyoming, USA[f]; University of Maryland School of Medicine, Baltimore, Maryland, USA[g]; University of California Davis, Department of Evolution and Ecology, Department of Medical Microbiology and Immunology, Davis, California, USA[h]

Here, we present the complete genome of the extreme thermophile, *Dictyoglomus thermophilum* H-6-12 (phylum *Dictyoglomi*), which consists of 1,959,987 bp.

Received 30 January 2014 **Accepted** 3 February 2014 **Published** 20 February 2014

Citation Coil DA, Badger JH, Forberger HC, Riggs F, Madupu R, Fedorova N, Ward N, Robb FT, Eisen JA. 2014. Complete genome sequence of the extreme thermophile *Dictyoglomus thermophilum* H-6-12. Genome Announc. 2(1):e00109-14. doi:10.1128/genomeA.00109-14.

Address correspondence to Jonathan A. Eisen, jaeisen@ucdavis.edu.

Dictyoglomus thermophilum is an extremely thermophilic, chemo-organotrophic, cellulolytic, obligate anaerobe originally isolated from a hot spring in Japan (1). The cells are rod-shaped, non-spore-forming, nonmotile, and form unusual spherical bodies of up to 100 cells. *D. thermophilum* and *Dictyoglomus turgidum* comprise the only two species in the *Dictyoglomi* phylum. The genome of *D. turgidum* has been sequenced, and the strain is unable to utilize cellulose (2). *D. thermophilum* was selected in 2002 as part of a National Science Foundation-funded "Assembling the Tree of Life" project at The Institute for Genomic Research (TIGR) to sequence the genomes of representatives of the seven phyla of bacteria that at the time had cultured representatives but no available genome sequence.

D. thermophilum was obtained from the ATCC, grown, and its DNA was extracted using standard techniques. Sanger sequencing and genome assembly were performed as previously described for genomes sequenced by TIGR (3–5). Small- and large-insert plasmid libraries were constructed in pUC-derived vectors after random mechanical shearing (nebulization) of genomic DNA. Sequencing resulted in 23,127 reads, with an average read length of 790 base pairs. The sequences were assembled using the Celera Assembler (6). The coverage criteria were that every position required at least double-clone coverage (or sequence from a PCR product amplified from genomic DNA) and either sequence from both strands or two different sequencing chemistries. The sequence was edited manually, and additional PCR and sequencing reactions were done to close gaps, improve coverage, and resolve sequence ambiguities (7). All repeated DNA regions were verified by PCR amplification across the repeat and sequencing of the product. The final assembly contains 1,959,987 bp, a G+C content of 34%, and an estimated coverage of ~9×.

The replication origin was determined by the colocalization of genes (*dnaA*, *dnaN*, *recF*, and *gyrA*) often found near the origin in prokaryotic genomes and GC nucleotide skew (G·C/G+C) analysis (8). Completeness of the genome was assessed using the PhyloSift software (9) to sequence for 40 highly conserved single-copy marker genes (10). Thirty-nine of these 40 markers were found in this assembly, and the missing marker (porphobilinogen deaminase) was found in only 80% of the original 1,000 genomes used to generate the markers.

An initial set of open reading frames likely to encode proteins (coding sequences [CDSs]) were predicted as previously described (7). All predicted proteins >30 amino acids were searched against a nonredundant protein database, as previously described (7). Protein membrane-spanning domains were identified by TopPred (11). The 5′ regions of each CDS were inspected to define the initiation codons using similarity searches, as well as the positions of ribosomal binding sites and transcriptional terminators. Two sets of hidden Markov models, Pfam version 11.0 (12) and TIGRFAMs 3.0 (13), were used to determine CDS membership in families and superfamilies. Pfam version 11.0 hidden Markov models were also used, with a constraint of a minimum of two hits to find repeated domains within proteins and mask them.

This resulted in 1,912 predicted protein coding sequences for *D. thermophilum* H-6-12 at the time of submission to Genbank (2008).

Nucleotide sequence accession numbers. The genome sequence has been deposited at GenBank under the accession no. CP001146. The version described in this paper is version CP001146.1.

ACKNOWLEDGMENTS

Sanger sequencing was performed at the Institute for Genomic Research (TIGR) in Rockville, MD.

We thank the many people who contributed to this project, including Martin Wu, Kevin Penn, Julie Enticknap, Liz O'Connor, Hoda Khouri, Jan Weidman, Yasmin Mohamoud, Grace Pai, Shannon Smith, Tamara Feldblum, Terry Utterback, Jason Inman, and Mihai Pop.

This work was funded by the National Science Foundation "Assembling the Tree of Life" grant no. 0228651, overseen by Jonathan A. Eisen and Naomi Ward.

REFERENCES

1. **Saiki T, Kobayashi Y, Kawagoe K, Beppu T.** 1986. *Dictyoglomus thermophilum* gen. nov., sp. nov., a chemoorganotrophic, anaerobic, thermophilic Bacterium. Int. J. Syst. Bacteriol. **35:**253–259.
2. **Brumm P, Hermanson S, Hochstein B, Boyum J, Hermersmann N, Gowda K, Mead D.** 2011. Mining *Dictyoglomus turgidum* for enzymatically active carbohydrases. Appl. Biochem. Biotechnol. **163:**205–214. http://dx.doi.org/10.1007/s12010-010-9029-6.
3. **Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM, Dunbar HE, Tran PL, Moran NA, Eisen JA.** 2006. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol. **4:**e188. http://dx.doi.org/10.1371/journal.pbio.0040188.
4. **Heidelberg JF, Seshadri R, Haveman SA, Hemme CL, Paulsen IT, Kolonay JF, Eisen JA, Ward N, Methe B, Brinkac LM, Daugherty SC, Deboy RT, Dodson RJ, Durkin AS, Madupu R, Nelson WC, Sullivan SA, Fouts D, Haft DH, Selengut J, Peterson JD, Davidsen TM, Zafar N, Zhou L, Radune D, Dimitrov G, Hance M, Tran K, Khouri H, Gill J, Utterback TR, Feldblyum TV, Wall JD, Voordouw G, Fraser CM.** 2004. The genome sequence of the anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. Nat. Biotechnol. **22:**554–559. http://dx.doi.org/10.1038/nbt959.
5. **Heidelberg JF, Paulsen IT, Nelson KE, Gaidos EJ, Nelson WC, Read TD, Eisen JA, Seshadri R, Ward N, Methe B, Clayton RA, Meyer T, Tsapin A, Scott J, Beanan M, Brinkac L, Daugherty S, DeBoy RT, Dodson RJ, Durkin AS, Haft DH, Kolonay JF, Madupu R, Peterson JD, Umayam LA, White O, Wolf AM, Vamathevan J, Weidman J, Impraim M, Lee K, Berry K, Lee C, Mueller J, Khouri H, Gill J, Utterback TR, McDonald LA, Feldblyum TV, Smith HO, Venter JC, Nealson KH, Fraser CM.** 2002. Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. Nat. Biotechnol. **20:**1118–1123. http://dx.doi.org/10.1038/nbt749.
6. **Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC.** 2000. A whole-genome assembly of *Drosophila*. Science **287:**2196–2204. http://dx.doi.org/10.1126/science.287.5461.2196.
7. **Tettelin H, Radune D, Kasif S, Khouri H, Salzberg SL.** 1999. Optimized multiplex PCR: efficiently closing a whole-genome shotgun sequencing project. Genomics **62:**500–507. http://dx.doi.org/10.1006/geno.1999.6048.
8. **Lobry JR.** 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13:**660–665. http://dx.doi.org/10.1093/oxfordjournals.molbev.a025626.
9. **Daarling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA.** 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ **2:**e243. https://peerj.com/articles/243/.
10. **Wu D, Jospin G, Eisen JA.** Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PLoS One **8:**e77033. http://dx.doi.org/10.1371/journal.pone.0077033.
11. **Claros MG, von Heijne G.** 1994. TopPred II: an improved software for membrane protein structure predictions. Comput. Appl. Biosci. **10:**685–686.
12. **Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL.** 2000. The Pfam protein families database. Nucleic Acids Res. **28:**263–266. http://dx.doi.org/10.1093/nar/28.1.263.
13. **Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O.** 2001. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. **29:**41–43. http://dx.doi.org/10.1093/nar/29.1.41.