

Draft genome of *Thermoanaerobacter thermohydrosulfuricus* strain AK152, a novel thermophilic and anaerobic bacterium isolated from a hot spring in Iceland

Clay A. Abraham,^{1,2} Kevin M. Bradley,¹ Sean Michael Scully,³ Johann Orlygsson,³ Derek Dube,² Steven A. Benner¹

AUTHOR AFFILIATIONS See affiliation list on p. 2.

ABSTRACT We present the draft genome of the bacterium *Thermoanaerobacter thermohydrosulfuricus* strain AK152, a thermophilic, endospore–spore-forming, anaerobe isolated from a hot spring in Grendalur, in Southwestern Iceland. This assembled genome will lay the foundation for identifying the carboxylic and amino acid fermentation pathways, suggesting biotechnological applications for this strain.

KEYWORDS *Thermoanaerobacter*, genomes, *Thermophilic anaerobe*

The *Thermoanaerobacter* genus is comprised of strict anaerobic thermophilic heterotrophic bacteria that have been detected in multiple extremophilic environments (1–5). *Thermoanaerobacter thermohydrosulfuricus* strain AK152 (AK152) was isolated by serial dilutions of sediment from a hot spring in Grendalur, Iceland. AK152 is ethanologenic and is exceptional in its ability to reduce multiple carboxylic acids to their corresponding alcohols and degrade a broad range of carbohydrates, yielding ethanol as the dominant end product (6).

AK152 was cultivated in basal mineral media with a 1.0 liquid-to-gas phase ratio and N₂ gas (<5% ppm O₂) for 48 h at 65°C (7, 8). Isolated cell pellets were stored at –80°C. Sodium dodecyl sulfate (SDS) and phenol:chloroform:iso-amyl alcohol were used to lyse the cell pellet (9). To remove residual SDS, two rounds of ethanol precipitation were performed. Extracted DNA was resuspended in sterile dH₂O and quantitated by a nanodrop spectrophotometer.

Library preparation and genomic sequencing of AK152's DNA was performed by Beijing Genomic Institute (BGI) using DNBseq. A total of 0.24 µg of DNA was sent to BGI for library preparation with BGI's Optimal DNA Library Prep Kit, which used PCR amplification. The MGISEQ-2000 platform with paired ends and 150 base pair (bp) reads sequenced the sample. BGI used the program SOAPnuke to obtain clean reads by removing adapter sequences, reads with 1% N content, reads below a quality of 20, and a minimum length of 150 bp (10). The cleaned raw data are listed in Table 1. The raw data from BGI were analyzed with FastQC (11). FastQC analysis found essentially no N reads, an average read quality of 39, and little sequence duplication.

The genome was assembled with two pipelines. The first was conducted with SPAdes version 3.15.4 using the isolate model (12). The second assembly method used the Read Assembly and Annotation Pipeline Tool (RAPT) version 0.5.4 suite of bioinformatics programs. RAPT includes the SKESA assembler version 2.5.0, ANI taxonomy identifier, NCBI Prokaryotic Genome Annotation Pipeline (PGAP), and CheckM (13–16).

The SPAdes and SKESA assemblies were both cross-referenced with QUAST version 5.2.0 for analysis (17). Additionally, each assembly was annotated with PGAP and the open reading frame finder GeneMarkS version 4.28 (18). Open reading frame locations (ORFs) were determined with GeneMarkS. Further, annotation of the assemblies was

Editor Irene L. G. Newton, Indiana University—
Bloomington, Bloomington, Indiana, USA

Address correspondence to Clay A. Abraham,
cabraham@fame.org.

The authors declare no conflict of interest.

See the funding table on p. 2.

Received 27 February 2024

Accepted 13 August 2024

Published 28 August 2024

Copyright © 2024 Abraham et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

TABLE 1 Analysis of *Thermoanaerobacter thermohydrosulfuricus* strain AK152 NGS and genome assembly data

Sequencing	Raw data	Assembly	SPAdes	SKESA
Clean Reads	4,055,209	Contigs	62	58
Clean Bases	1,216,562,700	Largest contig	229,593	181,799
Q20 (%)	96.62	N50	113,744	102,254
Q30 (%)	90.75	L50	8	9
GC (%)	35.09	GC (%)	33.83	33.78
		GeneMarkS (ORFs)	2,631	2,732
		PGAP (number of annotations)	811	2,154

performed with PGAP. The metrics from the programs used to assemble and assess the genome are listed in Table 1.

Displayed in Table 1, the SPAdes assembly generated the largest contig and more contigs overall when compared to SKESA. However, after annotation with GeneMarkS and PGAP, the SPAdes assembly produced fewer overall protein sequences and ORFs, with substantially fewer annotations made by PGAP. The PGAP annotation of the SKESA assembly identified over $2.5 \times$ more protein sequences than the SPAdes assembly as observed in Table 1. From this, the SKESA assembly, while finding fewer contigs overall, produced a more comprehensive annotation of the AK152 genome.

AUTHOR AFFILIATIONS

¹Foundation for Applied Molecular Evolution, Alachua, Florida, USA

²Department of Biology, University of Saint Joseph, West Hartford, Connecticut, USA

³Department of Natural Resource Science, University of Akureyri, Akureyri, Iceland

AUTHOR ORCID*s*

Clay A. Abraham  <http://orcid.org/0000-0002-7126-5896>

FUNDING

Funder	Grant(s)	Author(s)
Dynamic Combinatorial Chemistry		Clay A. Abraham

AUTHOR CONTRIBUTIONS

Clay A. Abraham, Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing | Kevin M. Bradley, Data curation, Methodology, Resources, Software | Sean Michael Scully, Investigation, Resources | Johann Orlygsson, Investigation, Resources | Derek Dube, Conceptualization, project administration, Resources, Supervision, Writing – review and editing | Steven A. Benner, Conceptualization, Funding acquisition, Resources, Supervision, Writing – review and editing

DATA AVAILABILITY

The assemblies in this article are deposited at DDBJ/ENA/GenBank under Biosample [SAMN39271361](https://www.ncbi.nlm.nih.gov/biosample/SAMN39271361) and under accession nos. [JAZE000000000](https://www.ncbi.nlm.nih.gov/nuccore/JAZE000000000) and [JBAMAR000000000](https://www.ncbi.nlm.nih.gov/nuccore/JBAMAR000000000) for the SKESA and SPAdes assemblies, respectively. The raw data reads are deposited at SRA under Bioproject accession no. [PRJNA1061494](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1061494).

REFERENCES

1. Imarsdottir A, Tarazewicz A, Gunnarsson I, Orlygsson J. 2010. Hydrogen production from sugars and complex biomass by *Clostridium* species, AK14, isolated from Icelandic hot spring. *Icel Agric Sci* 23:61–71.
2. Alves JI, Alves MM, Plugge CM, Stams AJM, Sousa DZ. 2016. Comparative analysis of carbon monoxide tolerance among *Thermoanaerobacter*

- species. *Front Microbiol* 7:1330. <https://doi.org/10.3389/fmicb.2016.01330>
3. Scully SM, Orlygsson J. 2015. Amino acid metabolism of *Thermoanaerobacter* strain AK90: the role of electron-scavenging systems in end product formation. *J Amino Acids* 2015:410492. <https://doi.org/10.1155/2015/410492>
 4. Li D, Greenfield P, Rosewarne CP, Midgley DJ. 2013. Draft genome sequence of *Thermoanaerobacter* sp. strain A7A, reconstructed from a metagenome obtained from a high-temperature hydrocarbon reservoir in the Bass Strait, Australia. *Genome Announc* 1:e00701-13. <https://doi.org/10.1128/genomeA.00701-13>
 5. Bhattacharya P, Barnebey A, Zemla M, Goodwin L, Auer M, Yannone SM. 2015. Complete genome sequence of the chromate-reducing bacterium *Thermoanaerobacter thermohydrosulfuricus* strain BSB-33. *Stand Genomic Sci* 10:74. <https://doi.org/10.1186/s40793-015-0028-7>
 6. Scully SM, Orlygsson J. 2020. Biotransformation of carboxylic acids to alcohols: characterization of *Thermoanaerobacter* strain AK152 and 1-propanol production via propionate reduction. *Microorganisms* 8:945. <https://doi.org/10.3390/microorganisms8060945>
 7. Hungate RE. 1950. The anaerobic mesophilic cellulolytic bacteria. *Bacteriol Rev* 14:1–49. <https://doi.org/10.1128/br.14.1.1-49.1950>
 8. Chades T, Scully SM, Ingvadottir EM, Orlygsson J. 2018. Fermentation of mannitol extracts from brown macro algae by *Thermophilic clostridia*. *Front Microbiol* 9:1931. <https://doi.org/10.3389/fmicb.2018.01931>
 9. Rio DC, Ares M, Hannon GJ, Nilsen TW. 2010. Purification of RNA by SDS solubilization and phenol extraction. *Cold Spring Harb Protoc* 2010:pdb.prot5438. <https://doi.org/10.1101/pdb.prot5438>
 10. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, Li Y, Ye J, Yu C, Li Z, Zhang X, Wang J, Yang H, Fang L, Chen Q. 2018. SOAPnuke: a mapreduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* 7:1–6. <https://doi.org/10.1093/gigascience/gix120>
 11. Wingett SW, Andrews S. 2018. FastQ screen: a tool for multi-genome mapping and quality control. *F1000Research* 7:1338. <https://doi.org/10.12688/f1000research.15931.2>
 12. Pribjelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes *de novo* assembler. *Curr Protoc Bioinformatics* 70:e102. <https://doi.org/10.1002/cpbi.102>
 13. RAPD - NCBI. Available from: https://www.ncbi.nlm.nih.gov/rapt/documentation/?utm_source=blog&utm_medium=referral&utm_campaign=rapt&utm_term=web&utm_content=20210512link3. Accessed November 1, 2023
 14. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS, Gonzales NR, Gwadz M, Lanczycki CJ, Song JS, Thanki N, Wang J, Yamashita RA, Yang M, Zheng C, Marchler-Bauer A, Thibaud-Nissen F. 2021. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res* 49:D1020–D1028. <https://doi.org/10.1093/nar/gkaa1105>
 15. Souvorov A, Agarwala R, Lipman DJ. 2018. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol* 19:153. <https://doi.org/10.1186/s13059-018-1540-z>
 16. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25:1043–1055. <https://doi.org/10.1101/gr.186072.114>
 17. Mikheenko A, Pribjelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 34:i142–i150. <https://doi.org/10.1093/bioinformatics/bty266>
 18. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 29:2607–2618. <https://doi.org/10.1093/nar/29.12.2607>