






Metagenome-Assembled Genomes from a Microbiome Converting Xylose to Medium-Chain Carboxylic Acids

 Matthew J. Scarborough,^a  Kevin S. Myers,^{b,c} Nathaniel W. Fortney,^{b,c}  Abel T. Ingle,^{b,c,d}  Timothy J. Donohue,^{b,c,e}
Daniel R. Noguera^{b,c,d}

^aDepartment of Civil and Environmental Engineering, University of Vermont, Burlington, Vermont, USA

^bGreat Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, Wisconsin, USA

^cWisconsin Energy Institute, University of Wisconsin-Madison, Madison, Wisconsin, USA

^dDepartment of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, Wisconsin, USA

^eDepartment of Bacteriology, University of Wisconsin-Madison, Madison, Wisconsin, USA

ABSTRACT There is growing interest in producing beneficial products from wastes using microbiomes. We previously performed multiomic analyses of a bioreactor microbiome that converted carbohydrate-rich lignocellulosic residues to medium-chain carboxylic acids. Here, we present draft metagenome-assembled genomes from this microbiome, obtained from reactors in which xylose was the primary carbon source.

Using sludge from a well-characterized bioreactor producing hexanoic and octanoic acids from lignocellulosic residues (1–3), we sought to enrich for different community members by operating a new bioreactor with a synthetic medium that contained xylose as the primary carbon source at a concentration of 18,700 mg L⁻¹. Biomass samples were collected after 96 (sample 16X), 102 (sample 17X), and 108 (sample 18X) days of bioreactor operation. DNA was extracted using a phenol-chloroform extraction method described previously (3). DNA aliquots of ~3,000 ng for each of the three samples were shipped to the Joint Genome Institute (JGI, Berkeley, CA, USA; jgi.doe.gov) for sequencing.

The HiSeq 2500 system (Illumina, San Diego, CA) was used to generate 150-bp paired-end reads. Assembly was performed using JGI's metagenome workflow (4) with `jgi_mga_meta_rq.py` version 2.1.0. Trimmed, screened, and paired-end Illumina reads were corrected using BFC (5) version r181. Reads were assembled from single samples with SPAdes (6) version 3.11.1 with “—meta” mode enabled, and binning was performed using MetaBAT (7) version 0.32.5 with default settings to create metagenome-assembled genomes (MAGs). CheckM (8) version 1.0.11 was used with default settings to assess MAG quality, and GTDB-Tk (9) version 0.1.6 was used for taxonomic assignment using the GTDB (10) database release 202. GTDB-Tk was run using the following commands: `gtdbtk identify --genome_dir ./--out_dir ./GTDBTk/--extension fasta --cpus 16; gtdbtk align --identify_dir ./GTDBTk/--out_dir ./GTDBTk/align --cpus 16`. The software program dRep (11) version 3.2.2 was used to select representative MAGs from 31 MAGs obtained in this work plus 10 MAGs from prior analyses (1, 3). Default dRep settings were used except that the completeness threshold was set to 60%, the contamination weight was set to 0.5, and the N_{50} weight was set to 5. The phylogenetic tree was created using RAxML-NG version 0.9.0 with the following sequential commands: `raxml-ng --parse --msa gtdbtk.bac120.user_msa.fasta --model LG+G8+F --prefix T1; raxml-ng --all --msa T1.raxml.rba --model LG+G8+F --prefix RAxML_ --threads 13 --seed 2`.

The 16X assembly had an N_{50} value of 13.3 kb, an L_{50} value of 288, and an average coverage of 453× and produced 10 MAGs from 118,660,326 filtered reads. The 17X assembly had an N_{50} value of 5.6 kb, an L_{50} value of 722, and an average coverage of 402× and produced 9 MAGs from 111,508,430 filtered reads. The 18X assembly had an

Editor Alejandro Sanchez-Flores, Universidad Nacional Autónoma de México

Copyright © 2022 Scarborough 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/).

Address correspondence to Matthew J. Scarborough, mscarbor@uvm.edu.

The authors declare no conflict of interest.

Received 24 November 2021

Accepted 9 March 2022

Published 28 March 2022

TABLE 1 Summary of recovered metagenome-assembled genomes^f

MAG ID	MAG name	Taxonomic classification (GTDB) ^a	Taxonomic classification (NCBI)	Comp. (%) ^b	Cont. (%) ^c	Size (Mb)	No. of scaffolds	Nearest GTDB ^a genome (NCBI assembly no.)	ANI (%) ^d	AF (%) ^e	NCBI genome accession no.
ACET1	UW_Xyl_ACET1	s__Acetobacter_senegalensis	Acetobacter sp.	99.5	0	3.1	36	GCF_001580995.1	97.56	89	JAJGAA0000000000
COR1.1	UW_Xyl_COR1.1	s__UBA7748_sp900314535	Atopobium sp.	99.19	3.23	2.5	130	GCA_900314535.1	97.3	80	JAJGAB0000000000
COR4	UW_Xyl_COR4	s__Olegusella_sp002407925	Olegusella sp.	78.65	2.82	1.4	165	GCA_002407925.1	98.53	72	JAJGAC0000000000
COR5	UW_Xyl_COR5	g__Olsenella	Olsenella sp.	100	6.45	2.7	46	GCF_009695875.1	92.91	84	JAJGAD0000000000
COR6	UW_Xyl_COR6	g__RUG013	Denitrobacterium sp.	93.03	2.49	1.9	133	GCF_001486445.1	88.71	85	JAJGAE0000000000
LAC5.1	UW_Xyl_LAC5.1	s__Limosilactobacillus_mucosae	Limosilactobacillus sp.	99.18	0	2.0	17	GCF_001436025.1	96.49	88	JAJGAF0000000000
LAC6	UW_Xyl_LAC6	s__Lentilactobacillus_buchneri	Lactobacillus sp.	99.06	0	2.5	22	GCF_001434735.1	97.34	94	JAJGAG0000000000
LCO2	UW_Xyl_LCO2	g__Eubacterium_H	Eubacterium sp.	77.14	2.12	2.5	384	NA	NA	NA	JAJGAH0000000000

^aGTDB, Genome Taxonomy Database (10); g., genus name; s., species name.

^bComp., completeness percentage estimated with CheckM (8).

^cCont., contamination percentage estimated with CheckM (8).

^dANI, average nucleotide identity with nearest GTDB genome estimated with GTDB-Tk (9).

^eAF, alignment fraction with nearest GTDB genome estimated with GTDB-Tk (9).

^fNA, not applicable. GTDB-Tk did not report a nearest genome because the AF with near genomes was below the default threshold of 65%.

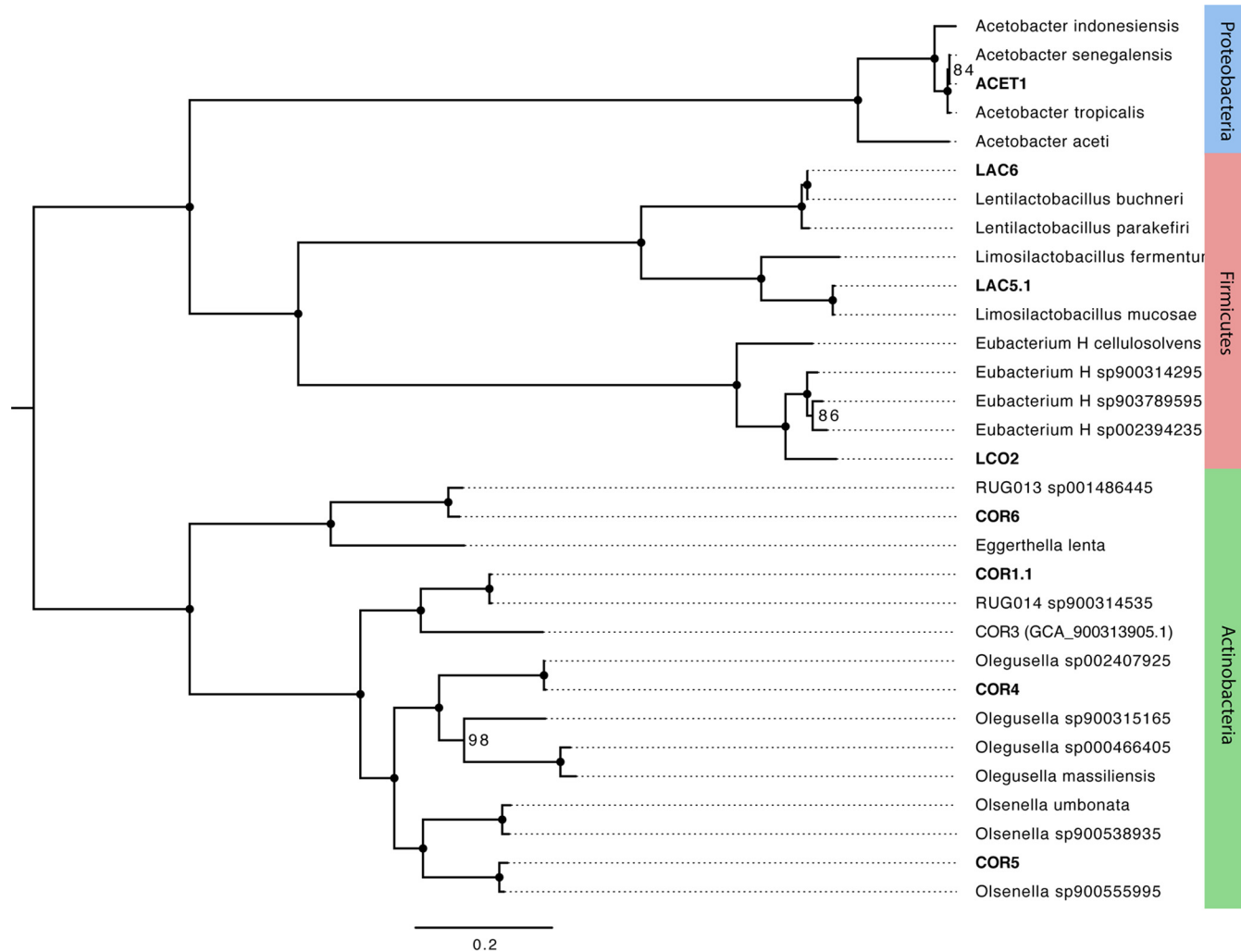


FIG 1 Phylogenetic tree constructed from recovered MAGs and related genomes. For each recovered MAG, three to five of the most closely related genomes were selected based on GTDB-Tk results (e.g., genus, species, average nucleotide identity [ANI], and alignment fraction [AF]). Genomes retrieved from NCBI were processed using GTDB-Tk to create concatenated sequences of 120 bacterial single-copy marker genes. The tree was constructed from the sequences using RAxML-NG (13) with 1,000 bootstraps. Filled circles indicate bootstrap values of 100, and all others are labeled as numbers. ACET, *Acetobacterales*; COR, *Coriobacteriales*; LAC, *Lactobacillales*; LCO2, *Lachnospirales*.

N_{50} value of 9.2 kb, an L_{50} value of 490, and an average coverage of 394 \times and produced 12 MAGs from 117,409,984 filtered reads. From the three assemblies, we obtained a total of 31 MAGs (completeness, >75%; contamination, <10%).

After dereplication, eight MAGs were selected as representative MAGs (Table 1), two of which are improvements to previously published MAGs (1, 3). COR1.1 is an update of COR1 (NCBI BioSample [SAMN09651346](#)), and LAC5.1 is an update of LAC5 (NCBI BioSample [SAMN09651352](#)). The other six MAGs represent new draft genome sequences of organisms within the GTDB (10) taxonomic orders *Acetobacterales* (ACET1), *Coriobacteriales* (COR4, COR5, and COR6), *Lactobacillales* (LAC6), and *Lachnospirales* (LCO2) (Fig. 1). Organisms related to *Acetobacterales*, *Coriobacteriales*, *Lactobacillales*, and *Lachnospirales* are regularly found via 16S rRNA gene amplicon sequencing (12) as abundant organisms during chain elongation. We provide the draft genome sequences of these organisms as a resource to further elucidate metabolic processes in this emerging biotechnological field.

Data availability. The raw metagenomic sequencing data are available under NCBI BioProject [PRJNA518398](#) (sample 16X), BioProject [PRJNA518399](#) (sample 17X), and BioProject [PRJNA518400](#) (sample 18X). MAGs are available under NCBI BioProject [PRJNA771338](#). Individual genome accession numbers for the MAGs are provided in Table 1.

ACKNOWLEDGMENTS

The work described here was supported by the Great Lakes Bioenergy Research Center, U.S. Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research under award number DE-SC0018409. 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 number DE-AC02-05CH11231.

REFERENCES

1. Scarborough MJ, Myers KS, Donohue TJ, Noguera DR. 2020. Medium-chain fatty acid synthesis by "*Candidatus Weimeria bifida*" gen. nov., sp. nov., and "*Candidatus Pseudoramibacter fermentans*" sp. nov. *Appl Environ Microbiol* 86:e02242-19. <https://doi.org/10.1128/AEM.02242-19>.
2. Scarborough MJ, Hamilton JJ, Erb EA, Donohue TJ, Noguera DR. 2020. Diagnosing and predicting mixed-culture fermentations with unicellular and guild-based metabolic models. *mSystems* 5:e00755-20. <https://doi.org/10.1128/mSystems.00755-20>.
3. Scarborough MJ, Lawson CE, Hamilton JJ, Donohue TJ, Noguera DR. 2018. Metatranscriptomic and thermodynamic insights into medium-chain fatty acid production using an anaerobic microbiome. *mSystems* 3:e00221-18. <https://doi.org/10.1128/mSystems.00221-18>.
4. Clum A, Huntemann M, Bushnell B, Foster B, Foster B, Roux S, Hajek PP, Varghese N, Mukherjee S, Reddy TBK, Daum C, Yoshinaga Y, O'Malley R, Seshadri R, Kyrpides NC, Eloe-Fadrosh EA, Chen I-MA, Copeland A, Ivanova NN. 2021. DOE JGI Metagenome Workflow. *mSystems* 6:e00804-20. <https://doi.org/10.1128/mSystems.00804-20>.
5. Li H. 2015. BFC: correcting Illumina sequencing errors. *Bioinformatics* 31:2885–2887. <https://doi.org/10.1093/bioinformatics/btv290>.
6. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 27:824–834. <https://doi.org/10.1101/gr.213959.116>.
7. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7:e7359. <https://doi.org/10.7717/peerj.7359>.
8. 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>.
9. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 36:1925–1927. <https://doi.org/10.1093/bioinformatics/btz848>.
10. Parks DH, Chuvpochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* 38:1079–1086. <https://doi.org/10.1038/s41587-020-0501-8>.
11. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 11:2864–2868. <https://doi.org/10.1038/ismej.2017.126>.
12. Candry P, Ganigué R. 2021. Chain elongators, friends, and foes. *Curr Opin Biotechnol* 67:99–110. <https://doi.org/10.1016/j.copbio.2021.01.005>.
13. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35:4453–4455. <https://doi.org/10.1093/bioinformatics/btz305>.