AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

OMICS DATA SETS

Terabase Metagenome Sequencing of Grassland Soil Microbiomes

William C. Nelson,[a] Lindsey N. Anderson,[a] Ruonan Wu,[a] Jason E. McDermott,[a] Sheryl L. Bell,[a] Ari Jumpponen,[c] Sarah J. Fansler,[a] Kimberly J. Tyrrell,[a] Yuliya Farris,[a] Kirsten S. Hofmockel,[a,b] Janet K. Jansson[a]

[a]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington, USA
[b]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA
[c]Division of Biology, Kansas State University, Manhattan, Kansas, USA

William C. Nelson and Lindsey N. Anderson contributed equally to this work. Author order was determined by drawing straws.

ABSTRACT To enable an in-depth survey of the metabolic potential of complex soil microbiomes, we performed ultra-deep metagenome sequencing, collecting >1 Tb of sequence data from three grassland soils representing different precipitation regimes.

As part of the Pacific Northwest National Laboratory (PNNL) Science Focus Area program (1, 2), we are investigating the impact of environmental change on microbial community function in grassland soils. Three grassland soils, representing different moisture regimes, were selected for ultra-deep metagenome sequencing, resulting in >1 Tb of sequence data per location. This data set serves as a resource for deep analysis of soil microbiome composition and metabolic potential.

Soils were collected from three grassland field site locations. Arid regime soil (irrigated agriculture), characterized as a coarse silty loam, was collected from the Washington State University Irrigated Agriculture Research and Extension Center (IAREC) (46.25N, 119.73W). Intermediate precipitation regime soil (rain-fed and irrigated agriculture), characterized as a fine clay loam, was collected from the Konza Prairie Biological Station (KPBS) (39.10N, 96.61W) (3, 4). Frequent precipitation regime soil (rain-fed and tile-drained agriculture), characterized as a fine silty clay loam, was collected from the Iowa State University Comparison of Biofuel Systems (COBS) (41.92N, 93.75W) (5).

Surface soil samples (2 cm by 0 to 20 cm) were collected from three randomly selected field site block locations using a push corer (3 subsamples per block, 3 replicates per subsample). Replicate subsamples were sieved together, resulting in 9 independent samples per site. Samples were flash frozen and stored at −80°C until further processing.

DNA was extracted from 3 × 0.25 g soil for each of the 9 field samples per site using the PowerSoil DNA extraction kit (Qiagen), with bead beating, and quantified. The extracted DNA samples from each site were combined to generate a pooled sample from each location (IAREC, COBS, and KPBS) for sequencing. Metagenomic libraries were prepared using the TruSeq PCR-free kit (Illumina) and a starting material of 1 μg DNA from the pooled DNA. Sequencing was performed on an Illumina HiSeq X system at Fulgent Genetics (Los Angeles, CA), generating 150-nucleotide paired-end reads to a final effort of at least 1 Tb of sequence per site (Table 1). BBDuk (BBTools package v38.38) (6) was used to trim adapter sequences from raw reads (adapters_no_transposase database), to perform quality filtering (parameters: int, ow; k, 27; hdist, 1; qtrim, f; minlen, 35), and to remove contaminants (sequencing_artifacts and phix174_ill reference database). Assembly was performed using the metaHipMer assembler (see MIMS metadata files for the specific developmental version used for each site) with

Citation Nelson WC, Anderson LN, Wu R, McDermott JE, Bell SL, Jumpponen A, Fansler SJ, Tyrrell KJ, Farris Y, Hofmockel KS, Jansson JK. 2020. Terabase metagenome sequencing of grassland soil microbiomes. Microbiol Resour Announc 9:e00718-20. https://doi.org/10.1128/MRA.00718-20.

Editor Frank J. Stewart, Georgia Institute of Technology

Address correspondence to William C. Nelson, william.nelson@pnnl.gov, or Janet K. Jansson, janet.jansson@pnnl.gov.

This work is a contribution of the Scientific Focus Area Phenotypic Response of the Soil Microbiome to Environmental Perturbations project at the Pacific Northwest National Laboratory, which is operated for the Department of Energy by Battelle Memorial Institute.

Received 19 June 2020
Accepted 29 June 2020
Published 6 August 2020

TABLE 1 Metagenome statistics for grassland soil microbiomes

Site	Total no. of reads	No. of quality bases (Tb)	Total no. of scaffolds	N_{50} (bp)	No. of scaffolds ≥2,500 bp	Total length of scaffolds ≥2,500 bp (bp)	No. of predicted proteins	PNNL DataHub accession no.
IAREC	7,536,393,634	1.123	83,651,096	1,404	241,472	989,234,018	1,255,684	WA-TmG.1.0
KPBS	7,343,389,182	1.088	68,100,771	1,111	304,736	1,283,171,244	1,388,888	KS-TmG.1.0
COBS	7,723,367,404	1.152	77,470,427	1,194	289,845	1,152,748,070	1,255,684	IA-TmG.1.0

kmer lengths of 21, 31, 55, and 71 (7) on the NERSC Cori platform (https://docs.nersc.gov/systems/cori). Scaffolds <2,500 bp long were omitted from further analysis. Quality-screened reads were mapped to scaffolds using the Burrows-Wheeler Aligner (v0.7.12) (8), and depth of coverage was determined across each scaffold using SAMtools (v1.9) (9).

Prodigal (v2.6.3) (10) was used to predict coding regions. Predicted protein sequences were searched using hmmsearch (v.3.1b2) (11) against the eggNOG (v4.5) (12), Pfam (v32.0) (13), and Nucleo-Cytoplasmic Virus Orthologous Group (NCVOG) (release date, 9 June 2014) (14) databases. Annotation assignments were given based on best bit scores (E-value cutoff, 1.0e−05).

These metagenomes are intended as a resource for the scientific community and should facilitate understanding of the highly diverse and complex metabolic potential that is encoded in soil microbial genomes.

Data availability. Metagenomic sequence data have been deposited in the PNNL DataHub repository and are available for download under project doi numbers WA-TmG.1.0, KS-TmG.1.0, and IA-TmG.1.0. The versions described in this paper are the first versions. Packages contain raw reads, assemblies, functional annotations, field site plot maps, MIMS.me.soil.5.0 metadata information, and package "read me" files.

ACKNOWLEDGMENTS

This research was supported by the Department of Energy (DOE) Office of Biological and Environmental Research. This research is a contribution of the Scientific Focus Area Phenotypic Response of the Soil Microbiome to Environmental Perturbations project and the EMSL/JGI FICUS award (award 50978). The PNNL is operated for the DOE by Battelle Memorial Institute under contract DE-AC05-76RLO1830. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. DOE Office of Science User Facility operated under contract DE-AC02-05CH11231.

We thank Robert S. Egan, Leonid Oliker, and Katherine A. Yelick for access to NERSC resources, developmental metaHipMer code, and expert advice in running the assembly process.

REFERENCES

1. Jansson JK, Hofmockel KS. 2018. The soil microbiome—from metagenomics to metaphenomics. Curr Opin Microbiol 43:162–168. https://doi.org/10.1016/j.mib.2018.01.013.
2. Jansson JK, Hofmockel KS. 2020. Soil microbiomes and climate change. Nat Rev Microbiol 18:35–46. https://doi.org/10.1038/s41579-019-0265-7.
3. Fay PA, Carlisle JD, Knapp AK, Blair JM, Collins SL. 2000. Altering rainfall timing and quantity in a mesic grassland ecosystem: design and performance of rainfall manipulation shelters. Ecosystems 3:308–319. https://doi.org/10.1007/s100210000028.
4. Fay PA, Carlisle JD, Danner BT, Lett MS, McCarron JK, Stewart C, Knapp AK, Blair JM, Collins SL. 2002. Altered rainfall patterns, gas exchange, and growth in grasses and forbs. Int J Plant Sci 163:549–557. https://doi.org/10.1086/339718.
5. Jarchow ME, Liebman M. 2013. Nitrogen fertilization increases diversity and productivity of prairie communities used for bioenergy. Glob Change Biol Bioenergy 5:281–289. https://doi.org/10.1111/j.1757-1707.2012.01186.x.
6. Bushnell B. 2018. BBTools package. https://jgi.doe.gov/data-and-tools/bbtools.
7. Georganas E, Egan R, Hofmeyr S, Goltsman E, Arndt B, Tritt A, Buluç A, Oliker L, Yelick K. 2018. Extreme scale de novo metagenome assembly, p 122–134. *In* SC18: International Conference for High Performance Computing, Networking, Storage and Analysis, Dallas, TX. https://doi.org/10.1109/SC.2018.00013.
8. Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595. https://doi.org/10.1093/bioinformatics/btp698.
9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
10. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
11. Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
12. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC,

Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293. https://doi.org/10.1093/nar/gkv1248.

13. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. https://doi.org/10.1093/nar/gkt1223.
14. Yutin N, Wolf YI, Raoult D, Koonin EV. 2009. Eukaryotic large nucleocytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 6:223. https://doi.org/10.1186/1743-422X-6-223.