Towards “Tera-Terra”: Terabase Sequencing of Terrestrial Metagenomes

Microbial ecologists are taking a metagenomics approach to analyze complex and diverse soil microbial communities

Janet Jansson, Earth Sciences Division, Lawrence Berkeley National Laboratory
(partially supported by the U.S. Department of Energy and LBNL under Contract# DE-AC02-05CH11231)

From a microbiological perspective, soil is largely unexplored even though we know it has a rich diversity of microbial life. Depending on its physical and chemical properties, soil can contain 109-1010 microbial cells per gram, including tens of thousands of different bacterial, archaeal, and fungal species, plus viruses and protists.

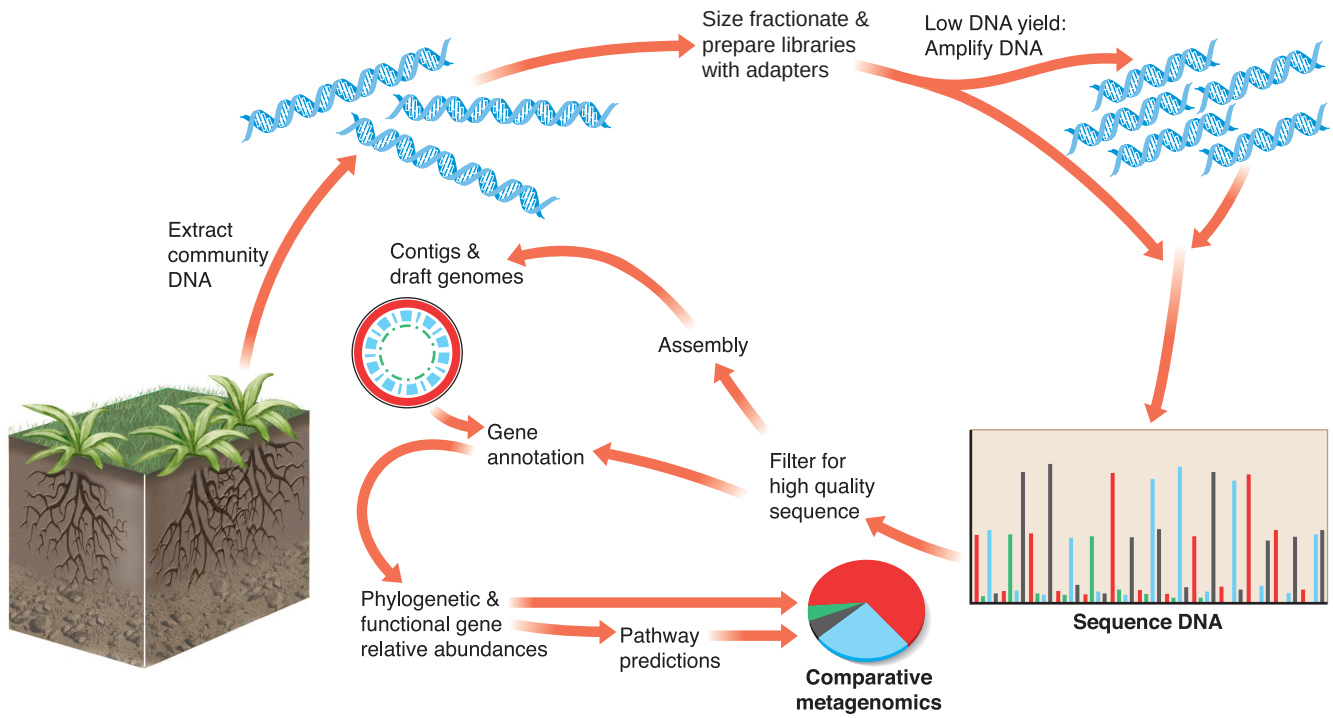
Soil microbes carry out life-sustaining functions for our planet, including cycling of nutrients and promoting plant growth. Respiring soil microorganisms, producing metabolic byproducts such as carbon dioxide and methane, cycle enormous volumes of carbon-containing gases from the terrestrial ecosystem into the atmosphere. Perturbing this ecosystem—for example, when global warming raises temperatures—potentially alters the flux of these gases. Despite the immensity of the carbon reservoir in soil (approximately 2,300 gigatons), its fate in the face of climate change is not known.

The high diversity of soil microbial communities makes them difficult to study. Exacerbating this difficulty, few soil microorganisms are amenable to isolation and cultivation, steps that in conventional terms are crucial for elucidating microbial physiology and biochemistry. Although most isolated soil bacteria are “weeds” such as pseudomonads and actinobacteria, 16S rRNA gene surveys reveal that few species from the great bulk of soil bacteria have cultured representatives. For example, members of the Acidobacteria phylum are widespread in soil but are notoriously difficult to cultivate. To address these issues, microbial ecologists are now using a metagenomics approach, sequencing soil DNA to develop a better understanding of the microbial identities and their potential functions in soils.

Defining, Then Refining Metagenomics

Initially, investigators who used metagenomics to study soil microbial communities cloned DNA extracted from soil into bacterial artificial chromosome (BAC) vectors or fosmid vectors. One advantage of this cloning-based approach is that it permits investigators to screen for expression of particular phenotypes, including production of antibiotics and enzymes. This strategy is

FIGURE 1



Soil metagenomics methodology.

being employed by members of a European initiative, MetaExplore, who are screening fosmid clones from a variety of environmental samples to access enzymes of interest to industry, including chitinases and dehalogenases: (http://cordis.europa.eu/fetch?CALLER=FP7_PROJ_EN&ACTION=D&DOC=1&CAT=PROJ&RCN=90382).

Another approach is to sequence subsets of the metagenome, such as collections of ribosomal RNA (rRNA) signature sequences. For example, based on 16S rRNA gene sequence data, we have developed a relatively good understanding of the species diversity and distribution of specific bacterial and archaeal phyla in different soils. Further, based on work from Noah Fierer, Rob Knight, and their colleagues at the University of Colorado, Boulder, we know that pH and salinity are major drivers of microbial biogeography. From these and other studies, we also know that soils contain high abundances of Acidobacteria, whose 26 subgroups vary in abundance from one soil type to another. Also, some phyla are more prevalent in a given soil type than in others. More generally, databases of 16S sequences are yielding insights into how chemical and physical parameters correlate with microbial distributions in soils.

Here, I use the term metagenome to refer to sequencing of total community DNA, including both phylogenetic and functional genes, while taking a shotgun-sequencing approach. Although few shotgun soil metagenome studies are published, more are anticipated during the next year (Table 1) as investigators take advantage of recent advances in sequencing instruments, for example, using 454 pyrosequencing and Illumina technologies. These 2nd-generation sequencing approaches generate megabases to gigabases of sequence data, respectively, in single runs with relatively short read lengths of approximately 400 to 100 bp, respectively. Other sequencing technologies recently developed, including the Pacific Biosciences platform for sequencing single molecules of DNA, hold

promise for generating longer sequencing read lengths.

Soil Metagenome Projects: Some Examples

In a project involving my group at the Lawrence Berkeley National Laboratory, James Tiedje and his colleagues at Michigan State University, and the Joint Genome Institute (JGI), we are using a combination of second-generation platforms to sequence DNA from microbes in soil samples from the Great Prairie of the United States, including native prairie and adjacent cultivated soils from Wisconsin, Iowa, and Kansas (Table 1). This project aims to determine the impact of land management (tillage, fertilization, etc.) on soil microbial communities and their functions, including cycling of carbon and nitrogen. One of the sites, Kansas native prairie, is also the focus of another project that is specifically addressing the impact of altered rainfall patterns due to climate change on carbon cycling processes in the Great Prairie (DOE-Biological Systems Research, under Contract No. DE-SC0004953). The Kansas prairie metagenome that was sequenced at JGI currently has the largest amount of sequence data of any soil metagenome to date, approaching 400 Gb of Illumina sequence, and will serve as a resource for this project.

Also in collaboration with JGI, we sequenced DNA extracted from Alaskan permafrost soil samples collected by Mark Waldrop from the U.S. Geological Survey (USGS). The aim is to use metagenomics to gain an understanding of the impact of climate warming-induced thaw on the microbial degradation of carbon reserves that have been trapped in permafrost for thousands of years and that have potential to contribute large amounts of greenhouse gases to the atmosphere.

Other ongoing soil metagenome sequencing projects include several that focus on field sites for which there is substantial temporal environmental and climate data. For example, the UK Rothamsted Field Station is one of the longest-running field stations in the world and has served as the site for several metagenome sequencing projects. One of these projects, "DeepSoil" (Table 1), is sequencing DNA from a long-term grassland and an adjacent fallow site at Rothamsted. The overarching goal of this sequencing effort is to establish the long-term impact of plants on the soil microbiota. Another project at Rothamsted is a French metagenome sequencing project, Metasoil, coordinated by Tim Vogel and Pascal Simonet of the Ecole Centrale de Lyon, France. The Metasoil project is sequencing DNA from the Park Grass site at Rothamsted that was established in 1856. Their strategy relies on constructing and sequencing a fosmid library in addition to shotgun metagenome sequencing.

Cheryl Kuske and coworkers at the Department of Energy (DOE) Los Alamos National Laboratory, in collaboration with JGI, are sequencing soils from selected free air-carbon dioxide enrichment (FACE) sites in the United States. These sites were established to determine the influence of increases in atmospheric CO_2 levels due to climate change on terrestrial ecosystems. In addition, Folker Meyer and coworkers at the DOE Argonne National Laboratory are sequencing metagenomes from several different U.S. soils that were collected across a range of habitats to determine which microorganisms and functional processes predominate in different soil ecosystems. Together these soil metagenomics projects will be a tremendous resource to the scientific community and will provide a much greater understanding of microbial diversity and functions in soil.

Data Handling and Analysis Challenges

Although the sequencing of DNA is no longer a bottleneck, large amounts of sequence data generated from analyzing highly diverse soil communities are proving a challenge to accommodate. This issue is exacerbated by the need to cope with short reads—for example, 75–125 bp—that arise from analyses using the Illumina instrument. Thus, better algorithms, new bioinformatics tools, and "terabytes" of computer storage are required.

Increased access to supercomputers, such as the National Energy Research Scientific Computing Center (NERSC) at the Lawrence Berkeley National Laboratory, can help. For instance, we used NERSC to perform BLASTX of our permafrost metagenome data (190 million reads, with approximately 50 gigabases of Illumina 113 bp x 2 paired end sequences). This analysis took approximately 800,000 core hours, or the equivalent of more than 85 com-

Jansson's Terroir: Soil Microbes, Wild Blueberries in Sweden, Wines in California

"If you plan to have children, it's important to have a husband willing and able to share responsibility with their upbringing, so you don't suffer by being out of touch with science for an extended period," Janet Jansson told an interviewer who asked her years ago what it takes to be a successful woman in science. "Of course, I might have been a little extreme," Jansson says now. "When my first daughter was being born, I called [my lab] from the delivery room . . . before I called my parents."

After living in Sweden for 20 years, Jansson and her husband Christer now live in California, where they are senior staff scientists at the Lawrence Berkeley National Laboratory. Their three children—a daughter, 22, and twins, a son and daughter, 21—were born in Sweden, and now attend universities in California. "Sweden was a great place to raise children and for me to develop a research career," she says. "When my twins were born, I got a year of parental leave that I shared 50% with my husband. This allowed both of us to have time with the kids and time to keep up with the research."

Jansson, 52, uses molecular tools to study microorganisms in complex environments. Earlier, she focused mainly on soil communities, but now is also studying microbial communities in other types of sites, including the human gut and oil-contaminated sediments from the Deepwater Horizon spill. "These may seem like rather disparate environments, but, to me, as a microbial ecologist, they are all interesting, complex microbial habitats," she says. For soil, "we have a major focus in understanding the impact of climate change on microbial processing of soil carbon. Our studies of the human gut are focused on inflammatory bowel diseases and the impact that microorganisms inhabiting the gut have on human health."

Jansson grew up in Albuquerque, N.M., where her father worked as an electrical engineer at Sandia National Laboratory. All of her relatives have Norwegian backgrounds, and come from Minnesota. When her parents retired—her mother was an elementary school teacher—they moved back to Minnesota. "That is really the place I call home," she says.

In Albuquerque, "I grew up in a national lab environment," she says. "As a child I loved animals and nature. I lived right by the Sandia Mountains, and loved to go hiking and camping with my family. I had a pet rabbit that followed me around like a dog. As I grew up, I became more concerned about the way humans were impacting the environment. However, I had no clear career plans before I started college."

Her husband Christer, a specialist in biochemistry and plant molecular biology, is Swedish. She holds dual citizenship in Sweden, after having lived there for so long, and she and her children are fluent in the language. "It's kind of funny being married to a Swede, because when I grew up,

puter years, which lasted 2 weeks using the NERSC supercomputer and nodes at JGI. Cloud computing will further help to reduce this bottleneck.

Another challenge is the large numbers of errors that different sequencing platforms generate. How can we differentiate sequencing errors from microheterogeneity within DNA samples from soil microbial communities? Also, there can be difficulties with different steps in sample processing. For example, each DNA extraction procedure can introduce its own bias with respect to sample loss or preferential lysis of some members of the microbial community over others. The most commonly used extraction procedures rely on beating with microscopic beads to lyse cells, although pressure lysis is another attractive option. Ideally, different laboratories should each use the same extraction protocol. However, despite the availability of commercial kits, laboratories typically follow their own favorite DNA extraction methods.

Another problem lies with soil samples that have low biomass or high levels of contaminants such as humic acids that result in low DNA yields. For example, permafrost soils yield relatively little DNA in our experience. However, amplifying DNA before preparing a library might help. Two DNA-amplifying methods are used: multiple displacement amplification (MDA) and emulsion PCR (emPCR). Of the

my Grandpa Martin used to tease all the time about Swedes, and how they didn't measure up to Norwegians," she says. "I don't know how he would have felt about me being married to one!"

Jansson earned her B.S. degree in biology and soil science in 1980 from New Mexico State University, where she started in chemical engineering, but soon learned that she preferred biology. She received an M.S. in soil microbiology in 1983 from Colorado State University, and a doctorate in microbial ecology in 1988 from Michigan State University, after which she and Christer moved to Sweden. She was pregnant with their first child, and he had a job waiting for him at Stockholm University.

"I thought it would be a great adventure to live there for a few years, but I didn't have any position lined up," she says. "Therefore, I was happy to get a postdoc position with Stefan Nordlund to study regulation of nitrogen fixation in *Rhodospirillum rubrum* in the same biochemistry department at Stockholm University where my husband had a position." While in Sweden, she became professor of microbiology at Södertörn University College, then professor and chair of environmental microbiology at the Swedish University of Agricultural Sciences, where she also served as vice dean of the faculty for natural resources and agricultural sciences.

"I was in Sweden at the right time because there was a growing concern about environmental consequences of genetically modified microorganisms," she says. "I had the tools in molecular microbiology and expertise in soil microbiology that were required for tracking specific microbes in the environment." One major boost for her career during that time was coordinating a large concerted action project for the European Union on marker genes and reporter genes for monitoring specific microbes "MAREP." It involved 40 partners from 11 different countries, "and many of these partners are still close colleagues and friends to this day," she says.

She credits James Tiedje at Michigan State University with encouraging her to develop molecular tools to study soil microorganisms. Her thesis, in fact, involved developing a method for extracting DNA from soil and using DNA probes to detect specific bacteria. Currently, she and Tiedje are collaborating on soil metagenomics research, looking at the Great Prairie of the continental United States.

Jansson enjoys gardening, and running with her husband and dog in the hills of Mount Diablo, near Walnut Creek, where they live, or in Berkeley, with members of her lab group. She also loves downhill skiing and—while in Sweden—was a figure skater. She misses one pastime from Sweden, picking wild blueberries and wild mushrooms in the forests near their home. But she is discovering California's singular pleasures. "There are always the vineyards," she says, "and I love tasting the California wines."

Marlene Cimons
Marlene Cimons lives and writes in Bethesda, Md.

two, the MDA approach is subject to considerable bias, whereas emPCR should be less biased because each template is separately amplified. However, to my knowledge, no one has directly compared the two methods.

Sometimes the volume of data falls short for conducting a metagenome analysis. For instance, when Susannah Tringe and coworkers at JGI first assembled soil metagenome data, their efforts failed because the 100 Mbp of sequence data that they collected proved insufficient. They estimated that they would need 2–5 Gbp to obtain draft genome assemblies of the most dominant organisms in soil, and current estimates from analysis of the Great Prairie metagenome data suggest that probably closer to 2 Tbp of data are needed! However, even a relatively low level of coverage was sufficient for some initial comparisons of the soil metagenome from a Minnesota farm to other available metagenome sequence datasets.

Recently Etienne Yergeau and colleagues at the National Research Council of Canada produced 1 Gbp (from Yergeau et al 2010: 853 Mb raw data, 533 Mb after filtering) of sequence data from permafrost soil after amplifying their sample via MDA, which introduced considerable bias. Nevertheless, when these data were compared to other metagenome data, DNA extracts from Minnesota farm soil—but not data from marine or other habitats—proved to be most closely related to the permafrost sample.

Table 1. Examples of ongoing soil metagenome sequencing projects

Sequencing project	Strategy	Status (2011)	Principal Investigators	Funding
"Metasoil": Rothamsted Park Grass permanent grassland, started 1856	454–FLX and Titanium (fosmids and shotgun metagenome)	24 Gigabases	Pascal Simonet & Tim Vogel (Ecole Centrale de Lyon, France)	France
"DeepSoil": Rothamsted Highfield permanent grassland and permanent bare-fallow plots, started 1959	Illumina paired end sequencing	80 Gigabases	Dirk Evers (Illumina), Tim Vogel, (Folker Meyer (Argonne National Lab), Janet Jansson (LBNL/JGI), James Tiedje (MSU)	Illumina
Great Prairie Grand Challenge pilot study. Native prairie and adjacent cultivated corn sites in Wisconsin, Iowa and Kansas	454 Titanium and Illumina paired end sequencing	>1.6 Terabase	James Tiedje (MSU); Janet Jansson, Susannah Tringe & Eddy Rubin (LBNL/JGI)	DOE-JGI
Alaskan permafrost, thermokarst bog and active layer samples (3 sites)	Illumina paired end sequencing	>80 Gigabases	Janet Jansson, Eddy Rubin, Rachel Mackelprang & Jenni Hultman (LBNL/JGI), Mark Waldrop (USGS)	DOE-JGI
24 Sites across continental US	454-FLX & Titanium	1.2 Gigabases	Folker Meyer	DOE-Argonne laboratory
DOE FACE sites- impact of elevated CO2–5 sites	454–FLX & Titanium	5 Gigabases	Cheryl Kuske (Las Alamos National Laboratory)	DOE-JGI

Organizing, Setting Standards for Metagenome Data

Although we are learning a great deal about dominant bacteria and archaea in soils based on 16S rRNA gene sequence data, many of the dominant operational taxonomic units (OTUs) that we detect in soil have no close representatives in culture collections. Researchers are addressing these deficiencies through initiatives such as the Genomic Encyclopedia for Bacteria and Archaea (GEBA) project that Jonathan Eisen of the University of California, Davis (UCD) and JGI coordinates. The long-term goal is to fill in the phylogenetic tree of life by sequencing genomes from underrepresented phyla. Another project, "Microbial Earth," being coordinated by Nikos Kyrpides at JGI, calls for sequencing microbial type strains in culture collections.

Meanwhile, the Earth Microbiome Project (EMP) is an initiative that aims to sequence what some call the "dark matter" of biology, the full microbial diversity on Earth (www.earthmicrobiome.org/). The EMP will begin sequencing 10,000 metagenomes from various collections and habitats, and eventually will cover hundreds of thousands of such samples, pending dedicated support.

The soil microbial ecology research community has established an international consortium, the International TerraGenome Consortium (www.terragenome.org). The consortium recognizes the high complexity of the soil environment and is focused on determining "the soil metagenome." TerraGenome is a clearinghouse for information about funding for soil metagenomics research, for development and provision of bioinformatics tools, for metadata standards, and for workshops and meetings on these topics. For example, TerraGenome set forth criteria for metadata obtained from analyzing soil samples that researchers must meet before their sequence data may be deposited into centrally held databases. This effort to set the minimum information about an environmental marker sequence (MIMARKS) was coordinated through the Genome Standards Consortium (GSC) (http://gensc.org/gc_wiki/index.php/Main_Page).

Harnessing Metagenomics To Study Microbial Ecology in Soils

Through soil metagenomics research, we can address fundamental questions about soil microbial ecology. For example, is there functional microbial redundancy in soil? Soil microbial community compositions differ in different soils in terms of dominant populations, according to 16S rRNA gene surveys. Although soil pH is a

key driver of soil community composition, biogeography also plays a role.

As an illustration, we can compare soil microbial diversity to the diversity of microbial communities in the human gut. The gut microbiota from one individual to another differs at the 16S rRNA gene level, but at the broad functional level the communities are rather homogenous in healthy individuals. This pattern suggests that several different bacterial species can carry out the same functional roles in the human intestine. The situation in soil might be similar, but we have yet to explore and compare many soil metagenomes in depth to determine whether that possibility holds.

Metagenomics can help us determine whether microorganisms in soils embody a specialized cache of gene functions. Available metagenome sequence datasets are already providing clues as to what functions are predominant in soils. For example, genes for cellobiose phosphorylase, an enzyme that degrades plant carbohydrates, were identified in a Minnesota farm soil metagenome, but not in one from the Sargasso Sea. When we screened permafrost for other functional genes specifically involved in cycling carbon and nitrogen, the samples included several genes that were more or less prevalent after thaw.

Metagenomics can also help to address whether rare species play an important functional role in soils. For example, although methanogens may not be numerically dominant in permafrost, they play a key role in producing methane, which is 21 times more potent as a greenhouse gas than carbon dioxide. With deep sequencing, it should be possible to obtain genomes of some of the dominant species in soil and even some species of relatively low abundance, provided that they do not have large amounts of strain heterogeneities.

As we collect soil metagenome sequence data, we need to improve how we mine such datasets. For example, the way we conduct BLAST searches might overlook valuable information, while the unassembled reads might be too short for annotating genes with confidence. Thus, we might well need to develop new assembly and annotation algorithms.

Another challenge is how to integrate different kinds of omics data, including metatranscriptomics and metaproteomics, to better understand functional processes of soil microbial communities. Metagenome sequence data, while informative, provides information about genes with the potential for being expressed, but cannot determine which ones are functional. Also, because we sequence total DNA, it is not possible to distinguish genes from actively growing cells from those in dormant or dead cells. Perhaps some analyses should be reserved for that fraction of DNA from active community members—for example, by extracting DNA that is allowed to incorporate stable isotopes or bromodeoxyuridine during replication. Ultimately, combining these approaches should enable us to gain a better understanding of which microbes are alive and active, and which enzymes and pathways function in soil microbial communities under different conditions. Then we can begin to truly comprehend soil microbial communities from the microscopic to the global scale.

SUGGESTED READING

Fierer, N., and R. B. Jackson. 2006. The diversity and biogeography of soil bacterial communities. Proc. Natl. Acad. Sci. USA **103:**626–631.

Gilbert, J. A., and C. L. Dupont. 2011. Microbial metagenomics: beyond the genome. Annu. Rev. Mar. Sci. **3:**357–371.

Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. 2005. Comparative metagenomics of microbial communities. Science **308:**554–557.

Vogel, T. M., P. Simonet, J. K. Jansson, P. R. Hirsch, J. M. Tiedje, J. D. Van Elsas, M. J. Bailey, R. Nalin, and L. Philippot. 2009. TerraGenome: a consortium for the sequencing of a soil metagenome. Nature Rev. Microbiol. **7:**1.

Yergeau, E., H. Hogues, L. G. Whyte, and C. W. Greer. 2010. The functional potential of high arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses. ISME J. **4:**1206–1214.

Yilmaz, P., R. Kottmann, d. Field, R. Knight, J.R. Cole, et al. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotech. 29:415–419.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.