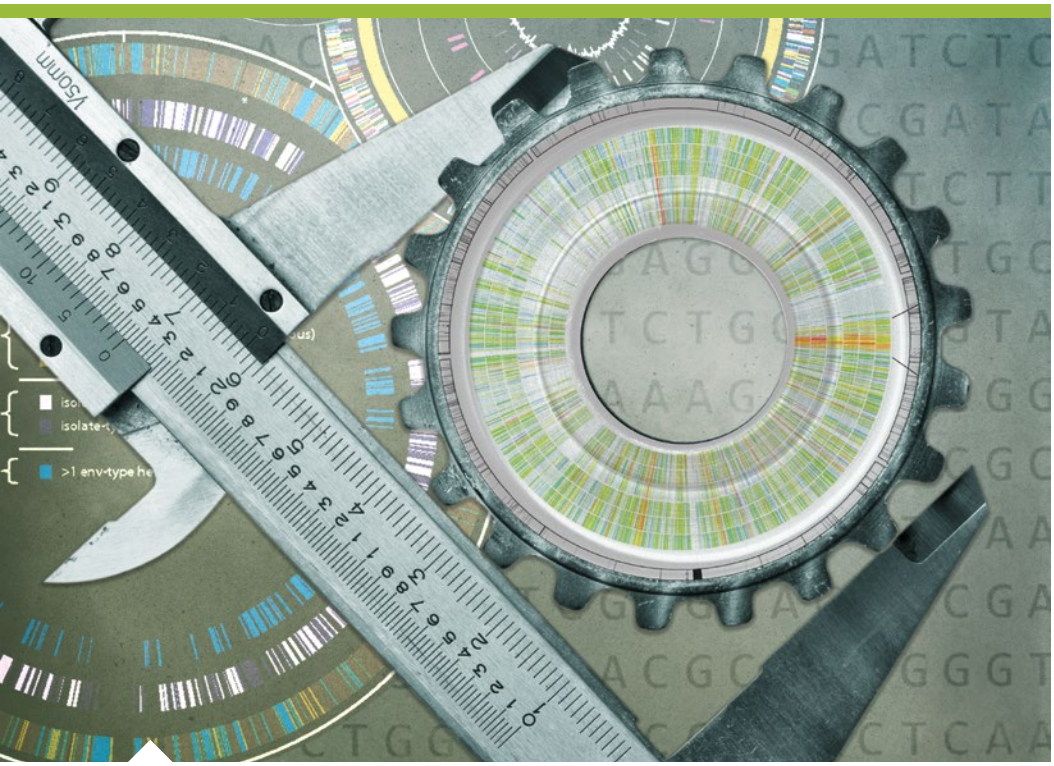
THE PRIMER

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

Summer 2017
volume 14
issue 3

Expanding Minimum Information Genome Standards



(Artistic rendering by Zosia Rostomian, Berkeley Lab Creative Services)

During the Industrial Revolution, the establishment of standards—e.g., sizes of nuts and bolts, etc.—allowed builders to produce supplies in bulk, maintain production quality and fuel interstate commerce. The importance of standards is dramatically illustrated when they don't exist or are not commonly accepted.

More than a century after the Industrial Revolution, advances in DNA sequencing technologies have caused similarly dramatic shifts in scientific research, and one aspect is studying the planet's biodiversity.

Published August 8, 2017 in *Nature Biotechnology*, an international team led by researchers at the DOE Joint Genome Institute (JGI), a DOE Office of Science User Facility, has developed standards for the minimum metadata to be supplied with single amplified genomes (SAGs) and metagenome-assembled genomes (MAGs) submitted to public databases.

Microbes play crucial roles in regulating global cycles involving carbon, nitrogen, and phosphorus among others, but many of them remain uncultured and unknown. Learning more about this so-called "microbial dark matter" involves sequencing metagenomes or amplified DNA of single cells, then bioinformatically extracting genomes. As genomic data production has ramped up over the past two decades, using various platforms around the world, scientists have *continued on page 4*

in this issue

Uncovered: 1000 New Microbial Genomes

"Bacteria and archaea comprise the largest amount of biodiversity of free-living organisms on Earth," said JGI's Prokaryote Super Program head Nikos Kyrpides. "They have already conquered nearly every environment on the planet, so they have found ways to survive under the harshest of conditions with different enzymes and with different biochemistry."

JGI scientists have taken a decisive step forward in uncovering the planet's microbial diversity. In a paper published June 12, 2017 in *Nature Biotechnology*, Kyrpides and his team of researchers report the release of 1,003 phylogenetically diverse bacterial and archaeal reference genomes—the single largest release to date.

The effort is part of the JGI's Genomic Encyclopedia of Bacteria and Archaea (GEBA) initiative that aims to sequence thousands of bacterial and archaeal genomes to fill in unexplored branches of the tree of life. Microbes play important roles in regulating Earth's biogeochemical cycles— *continued on page 2*

New Microbial Genomes

continued from page 1

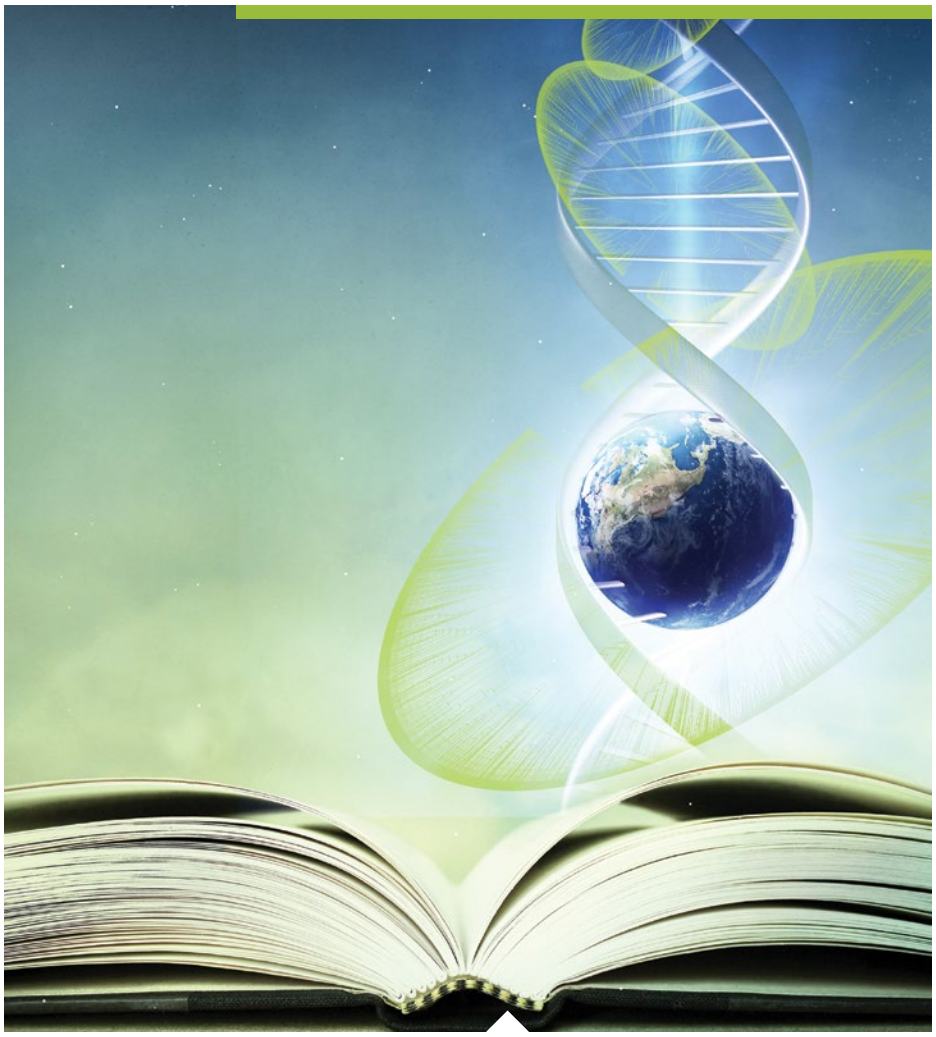
processes that govern nutrient circulation in terrestrial and marine environments, for example. Uncovering the functions of genes, enzymes and metabolic pathways through genome sequencing and analysis has wide applications in the fields of bioenergy, biomedicine, agriculture and environmental sciences.

"In addition to identifying over half a million new protein families, this effort has more than doubled the coverage of phylogenetic diversity of all type strains with genome sequences," said Supratim Mukherjee, a JGI computational biologist and co-first author of the paper.

Since a great portion of research in microbial genomics has been focused on human pathogens or biotechnological work horses, GEBA is the main effort worldwide attempting to address the phylogenetic coverage knowledge gap by sequencing a diverse set of cultured but poorly characterized microbial type strains.

"It was recognized that we weren't sampling many parts of the tree of life," said Rekha Seshadri, a JGI computational biologist and co-first author of the paper. "And if we sampled some of those parts of the tree, we'd discover new functions, which could be an important resource for new applications."

The release of these genomes is the culmination of almost a decade's worth of work, with the first 56 GEBA genomes published in 2009. The microorganisms were isolated from environments ranging from sea water and soil, to plants, and to cow rumen and termite guts. Genome sequencing and analysis was done at the JGI through the Community Science Program, and the 1,003 genomes are publicly available through the Integrated Microbial Genomes with Microbiomes (IMG/M) system, with all associated metadata in compliance with the Genomics Standards Consortium available through the Genomes OnLine Database.



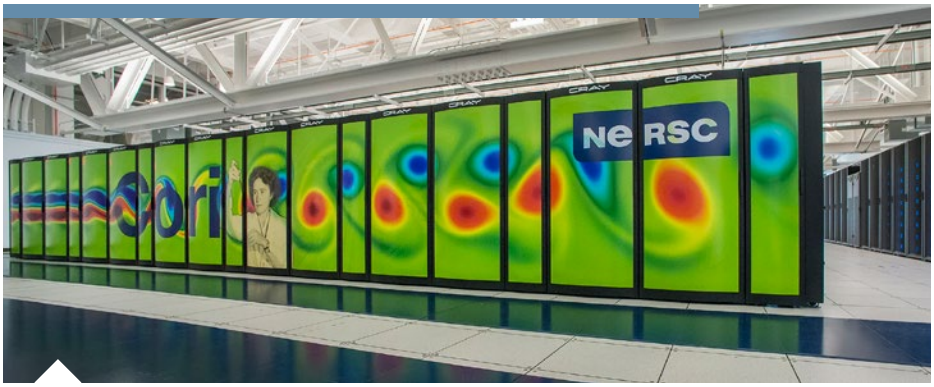
(Artistic rendering by Zosia Rostomian, Berkeley Lab Creative Services)

Seshadri said that with the release of high quality genomic information, JGI is providing a wealth of new sequences that will be invaluable to scientists interested in experiments such as characterizing biotechnologically relevant secondary metabolites or studying enzymes that work under specific conditions. And because Kyrpides' research team sequenced type strains that are readily available from culture collections, scientists can perform follow-up experiments with them in the lab, she added.

"The partnership with culture collection centers such as the Leibniz Institute DSMZ in Germany and the ATCC Global Bioresource Center in the U.S. was critical in accomplishing this endeavor," said Kyrpides. "At a time when throughput can come at the cost of quality, resulting in highly fragmented and chimeric or contaminated genomes, the significance of genomes from the type strains as invaluable taxonomic signposts cannot be overstated."

Full story at http://jgi.doe.gov/uncovered-1000-new-microbial-genomes/.

Six Proposals Meld Genomics, Supercomputing in First JGI-NERSC Call



NERSC's supercomputer Cori. (Roy Kaltschmidt, Berkeley Lab)

Six proposals were selected to participate in a new partnership between the JGI and the National Energy Research Scientific Computing Center (NERSC) through the "Facilities Integrating Collaborations for User Science" (FICUS) initiative.

"This FICUS program represents a major milestone for the JGI, as for the first time we provide a service to the community, not based on products (i.e. DNA sequencing or synthesis) but on data alone," said Prokaryote Super Program head Nikos Kyrpides. "As Microbiome research is rapidly moving towards Data Science we anticipate that the demand for this program will also increase."

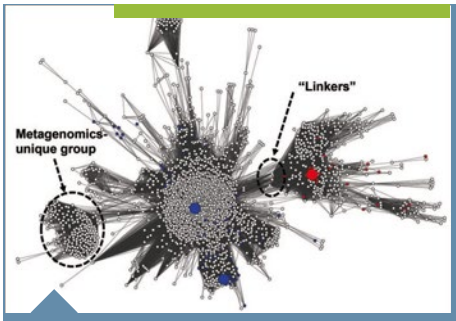
The JGI-NERSC FICUS call is the latest partnership since the collaborative science initiative was formed in 2014 by the Office of Biological and Environmental Research (BER) to harness the combined expertise and resources of two of the national user facilities stewarded by the DOE Office of Science in support of DOE's energy, environment, and basic research missions. The expertise and capabilities available at these national user facilities, both at the Lawrence Berkeley National Laboratory (Berkeley Lab), will help researchers explore the wealth of genomic and metagenomic data generated worldwide through access to supercomputing resources and computational science experts to accelerate discoveries. Through the JGI-NERSC FICUS call, users can query across all available data to look for patterns across data sets in the JGI's Integrated Microbial Genomes and Microbiomes (IMG/M) database with the help of NERSC's supercomputer Cori, resulting in a more powerful analysis with increased capacity for novel discoveries.

The accepted proposals come from:

- Patricia (Patsy) Babbitt of the University of California (UC), San Francisco (*See sidebar*)
- David Baker at the University of Washington
- Phillip Brooks of UC Davis
- Ed DeLong of the University of Hawaii at Manoa
- Steve Hallam of Canada's University of British Columbia
- Kostas Konstantinidis of Georgia Institute of Technology

The full list of approved proposals is available at http://jgi.doe.gov/our-projects/csp-plans/fy-2017-csp-plans/#jgi-nersc.

"I really believe the future of computing is going to be dominated by biology. The volumes of biological data that need to be synthesized, aggregated and interrogated will require supercomputers," said Kjiersten Fagnan, both JGI's Chief Informatics Officer and NERSC's Data Science Engagement Group Lead.

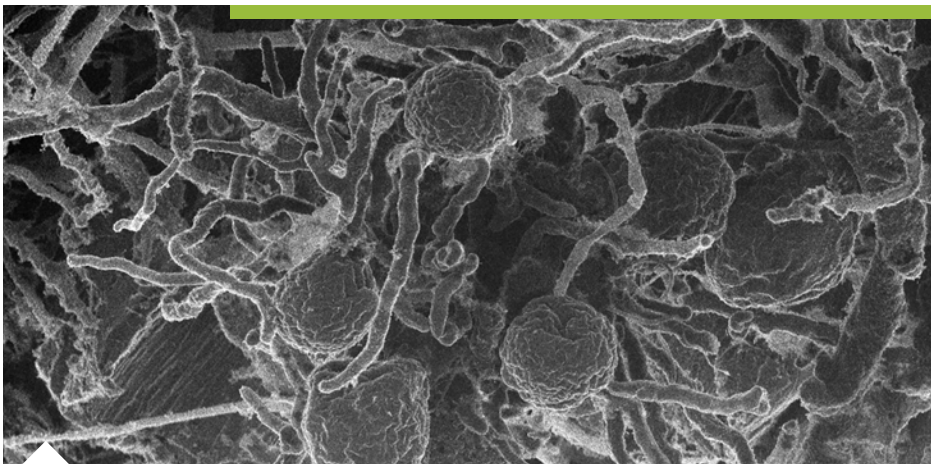


A sequence similarity network of a family of enzymes from the nitroreductase superfamily (some nitroreductases can reduce TNT or trinitrotoluene, a significant soil contaminant). Nodes represent enzyme sequences, while edges represent pairwise similarities more significant than 1e-42 (BLAST E-value). Red and blue nodes represent enzymes found in public sequence databases and belong to two sub-families, and white nodes represent sequences found only in the JGI's metagenomic database (IMG/M). Large nodes represent experimentally-characterized enzymes of diverse functions. Notably, a significant expansion of the sequence space is observed (from 300 enzymes to >10,000), revealing a new potential group of enzymes found only in IMG/M. "Linkers" that are also unique to metagenomes display sequence similarity to experimentally-characterized enzymes of diverse functions and serve as attractive targets for synthesis and biochemical assays for intermediate function. (Eyal Akiva and Patsy Babbitt)

"If you look at the data sets being generated and the questions that people have, you can see that researchers are going to have to combine different datasets—like genomics, metabolomics, protein crystal structures and potentially even brain scans and more—to find answers. This work cannot be done on a laptop or small cluster."

Full story at http://jgi.doe.gov/doe-user-facilities-ficus-join-forces-to-tackle-biology-big-data/.

Fungal Enzyme Complexes Formed to Break Down Cellulose



Scanning electron micrograph (SEM) of *Neocallimastix californiae*. (Chuck Smallwood, PNNL)

Cost-effectively breaking down bioenergy crops into sugars that can then be converted into fuel is one of the biggest barriers in the commercial production of sustainable biofuels. To reduce this barrier, bioenergy researchers are looking to nature and the estimated 1.5 million species of fungi that, collectively, can break down almost any substance on earth, including plant biomass. Now a team led by researchers at the University of California (UC), Santa Barbara has found for the first time that early lineages of fungi can form complexes of enzymes capable of degrading plant biomass. By consolidating these enzymes, in effect into protein assembly lines, they can team up to work more efficiently than they would as individuals.

"There are protein complexes in bacteria called cellulosomes that pack together the enzymes to break down plant biomass," said UC Santa Barbara's Michelle O'Malley, senior author of the study published May 26, 2017 in *Nature Microbiology*. "The idea is that these clusters are better at attacking biomass because they are keeping the different enzymes in place with plugs called dockerins so they work more efficiently. This has been detailed in bacteria for more than 20 years, but now seen for the first time in fungi."

The study involved a comparative genomics analyses of five fungi that belong to the Neocallimastigomycetes, a clade of the early-diverging lineages that are not well-studied.

Read the full story at http://jgi.doe.gov/fungal-enzymes-team-up-efficiently-break-down-cellulose/.

Genome Standards

continued from page 1

worked together to establish definitions for terms such as "draft assembly" and data collection standards that apply across the board. One critical item that needs standardization is "metadata," defined simply as "data about other data."

"Over the last several years, single-cell genomics has become a popular tool to complement metagenomics," said study senior author Tanja Woyke, head of the JGI Microbial Program. "Starting in 2007, the first single-cell genomes from environmental cells appeared in public databases and they are draft assemblies with fluctuations in the data quality. Metagenome-assembled genomes have similar quality challenges. For researchers who want to conduct comparative analyses, it's really important to know what goes into the analysis. Robust comparative genomics relies on extensive and correct metadata."

In their paper, Woyke and her colleagues proposed four categories of genome quality. Low-Quality Drafts would be less than 50 percent complete, with minimal review of the assembled fragments, but would still be required to be less than 10 percent contaminated with non-target sequence. Medium-Quality Drafts would be at least 50 percent complete, with minimal review of the assembled fragments and less than 10 percent contamination. High-Quality Drafts would be more than 90 percent complete with the presence of the 23S, 16S and 5S rRNA genes, as well as at least 18 tRNAs, and with less than 5 percent contamination. The Finished Quality category is reserved for single contiguous sequences without gaps and less than 1 error per 100,000 base pairs, the same standard applied to isolate genomes.

The JGI has generated approximately 80 percent of the over 2,800 SAGs and more than 4,500 MAGs registered on the JGI's Genomes OnLine Database (GOLD). JGI scientist and study first author Bob Bowers said many of the SAGs already in GOLD would be considered Low-Quality or Medium-Quality Drafts. While these are highly valuable datasets, for some purposes, researchers might prefer to use High-Quality or Finished datasets.

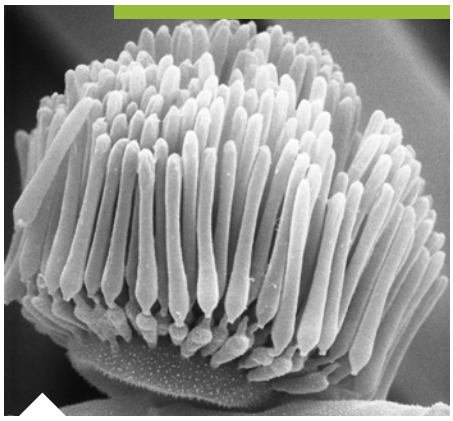
Finding a Major Gene Expression Regulator in Fungi

Though fungi have been around for a billion years and collectively are capable of degrading nearly all naturally-occurring polymers and even some human-made ones, most of the species that have been studied belong to just two phyla, the Ascomycota and Basidiomycota. The remaining 6 groups of fungi are classified as "early diverging lineages," the earliest branches in fungal genealogy.

"By and large, early-diverging fungi are very poorly understood compared to other lineages. However, many of these fungi turn out to be important in a variety of ways," said JGI analyst Stephen Mondo (See page 4). In the May 8, 2017 issue of *Nature Genetics*, a team led by Mondo and his JGI colleagues reported the prevalence of a marker for functionally important genes in early diverging fungi.

Changing a single letter, or base, in an organism's genetic code can lead to changes in protein structures and functions, impacting its traits. In addition, though, subtler changes can and do happen, involving modifications of the DNA bases themselves. The best-known example of this kind of change is a methylation of the base cytosine at the 5th position on its aromatic ring (5mC). In eukaryotes, a less-well known modification involves adding a methyl group at position 6 of adenine (6mA).

"This is one of the first direct comparisons of 6mA and 5mC in eukaryotes, and the first 6mA study across the fungal kingdom," said JGI Fungal Genomics head and senior author Igor Grigoriev. "6mA has been shown to have different functions depending on the organism. For example, in animals it is involved in suppressing transposon activity, while in algae it is positively associated with gene expression. Our analysis has shown that 6mA modifications are associated with expressed genes and is preferentially positioned based on gene function and conservation."



Linderina pennispora (ZyGoLife Research Consortium, Flickr, CC BY-SA 2.0)

Watch Stephen Mondo on the proposed role of 6mA in early diverging fungi at the 2017 Genomics of Energy & Environment Meeting at http://bit.ly/JGI2017Mondo.

Read the full story at http://jgi.doe.gov/finding-major-gene-expression-regulator-fungi/.

continued from page 4

Moving from a proposal in print to implementation requires community buy-in. Woyke and Bowers conceived of the minimum metadata requirements for SAGs and MAGs as extensions to existing metadata standards for sequence data, referred to as "MIxS," developed and implemented by the Genomic Standards Consortium (GSC) in 2011. The GSC is an open-membership working body that ensures the research community is engaged in the standards development process and includes representatives from the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI). This is important since these are the main data repositories where the minimum metadata requirements are implemented. By working directly with the data providers, the GSC can assist both large-scale data submitters and databases to align with the MIxS standard and submit compliant data.

Nikos Kyrpides, head of the JGI Prokaryote Super Program and GSC Board member, noted that JGI has been involved in organizing the community to develop genomic standards as part of its core mission. "The GSC has been instrumental in bringing the community together to develop and implement a growing body of relevant standards. In fact, the need to expand MIxS to uncultivated organisms was identified in one of the recent GSC meetings at the JGI."

"These extensions complement the MIxS suite of metadata standards by defining the key data elements pertinent for describing the sampling and sequencing of single-cell genomes and genomes from metagenomes," said GSC President and study co-author Lynn Schriml of the Institute of Genome Sciences at University of Maryland School of Medicine. "These standards open up a whole new area of metadata data exploration as the vast majority of microbes, referred to as microbial dark matter, are currently not described within the MIxS standard."
Read the full story at http://jgi.doe.gov/defining-standards-genomes-uncultivated-microorganisms/.

Hot spring at Yellowstone National Park. (Paul Blainey, Christina Mork and Geoffrey Schiebinger)

New Technology to Access Microbial Dark Matter

The majority of the planet's microbial diversity remains uncultivated, and their genes and metabolic functions could have potential applications in fields ranging from bioenergy to biotechnology to environmental research. There are thousands of microbial datasets in the JGI's IMG/M system, and many of them have been uncovered through the use of metagenomic sequencing and single-cell genomics. Despite their utility, these techniques have limitations: single-cell genome amplifications are time-consuming, and often incomplete, and shotgun metagenomics sequencing generally works best if the environmental sample is not too complex.

In *eLife*, Stanford University researchers led by Stephen Quake reported the development of a microfluidics-based, mini-metagenomics approach to mitigate these challenges. They extracted 29 novel microbial genomes from Yellowstone hot spring samples while still preserving single-cell resolution to enable accurate analysis of genome function and abundance. Applying this new technology to additional sample sites will add to the range of hitherto uncharacterized microbial features with potential DOE mission applicability. The work was enabled by the JGI's Emerging Technologies Opportunity Program (ETOP).

Tracking Microbial Succession in Petroleum Wells

While shifting toward more sustainable, alternative energy sources, people still rely on fossil fuels for energy and transportation fuels. Understanding how microbial communities in subsurface petroleum reservoirs are impacted by human activity, and how their responses can sour oil production, can influence oil industry practices. Microbes are invisible to the naked eye, but play key roles in maintaining the planet's biogeochemical cycles.

In the North Sea is an offshore petroleum reservoir managed by a joint venture of which Shell, one of the world's largest oil companies, is a partner. Over the past 15 years, 32 oil wells have been drilled, reaching depths over 2 kilometers below the seafloor to extract petroleum from this deep subsurface reservoir. To learn more about how some of these subsurface microbial populations respond to disruptions in their environment, researchers in the petroleum industry conducted a comparative genetic analysis of the microbial communities in multiple oil wells within an offshore oil field. During the analyses, reported in *The ISME Journal*, tools developed at the JGI and made available through the IMG/M system were utilized. The data generated provides insights into just how deep subsurface microbial communities are perturbed by active oil wells injecting foreign substances into these previously isolated populations, and helps industry researchers develop new techniques for managing microbiological problems.

Lessons from Simulating a Deep Ocean Oil Spill

The 2010 Deepwater Horizon oil spill released 4.1 million barrels of oil into the Gulf of Mexico and was the first major release of oil and natural gases into the deep ocean (1,500 meters). Due to the depth of the spill, vast plumes of small oil droplets remained trapped deep in the ocean (900-1,300 meters) where they underwent biodegradation by the local microbial community. Until now, researchers have been puzzled over the metabolic capabilities driving the shifts between microbial communities to degrade the crude oil. In the *Proceedings of the National Academy of Science*, a team led by Berkeley Lab researchers have been able to present the first complete picture of how successive waves of microbial populations degraded the released oil. They were also able to recover high-quality genomes of the key microbial players, and determine the metabolic factors driving the shifts between microbial communities.

Identifying the microbes involved in degrading hydrocarbons (the chief components of petroleum and natural gas), as well as the drivers that trigger successive waves of microbial responders, allows researchers to better understand how the microbial community adapted to the events of seven years ago. The expertise and resources used to reconstruct the microbial genomes demonstrates how new technology development through ETOP at the JGI is enabling energy and environment research. They discovered that the microbial community transformed with chemical changes in residual oil. Additionally, their lab-based method allowed for the first time the successful resolution of high-quality genomes and the characterization of functional capabilities for all the key microbes.

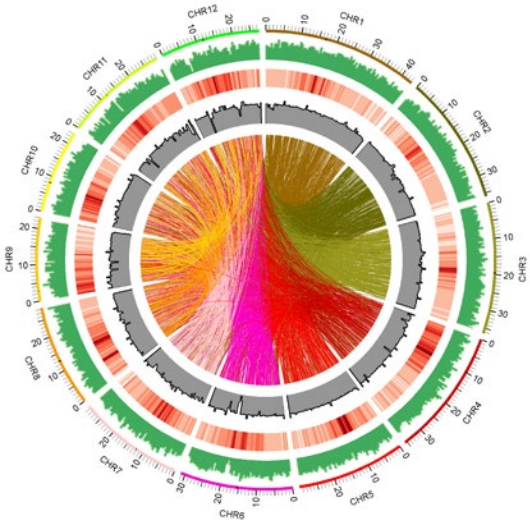
A Gene that Influences Grain Yields in Grasses

Setaria species are related to several candidate bioenergy grasses including switchgrass and *Miscanthus*. As model systems to study grasses that photosynthetically fix carbon from CO_2 through a water-conserving (C_4) pathway, the genomes of both green foxtail (*Setaria viridis*) and foxtail millet (*S. italica*) have been sequenced and annotated through the JGI's Community Science Program. A team led by Tom Brutnell at the Donald Danforth Plant Science Center and including JGI researchers reported in *Nature Plants* that they had identified genes that may play a role in flower development on the panicle of green foxtail, highlighting the utility of *S. viridis* as a model crop. Panicle development is critical for determining grain yield which in turn is crucial to food crops as well as candidate crops for producing renewable and sustainable fuels. A homologous gene in maize was identified as playing a similar role, illustrating the value of model systems in finding genes involved in important properties in potential bioenergy-relevant plants.

Field of *Setaria viridis* growing in western Nebraska. (Pu Huang)

Mutant Rice Database for Bioenergy Research

For more than half of the world's population, rice is the primary staple crop. As a grass, it is a close relative of the candidate bioenergy feedstock switchgrass. A team led by University of California, Davis, and including researchers at the JGI and the Joint BioEnergy Institute (JBEI), a DOE Bioenergy Research Center, have assembled the first major large-scale collection of mutations for grass models. They used the model rice cultivar Kitaake (*Oryza sativa* L. ssp. *japonica*), and compared the genes against the reference rice genome of another japonica subspecies called Nipponbare available on the JGI Plant Portal Phytozome. Boosting yields of bioenergy feedstock crops such as grasses requires a better understanding of how enzymes and proteins synthesize plant cell walls in order to modify the processes and the composition. The team's goal is to have a functional genomics resource for grass models involved in plant cell wall biosynthesis studies. Until now, mutant collections for grass models have lagged behind those available for the *Arabidopsis* model system.



Genome-wide distribution of fast neutron-induced mutations in the Kitaake rice mutant population. (Guotian Li and Rashmi Jain)

Fast-neutron irradiation or exposure to high energy neutrons, induces a wide variety of mutations by making changes in DNA. Using this approach, rice researchers were able to create the first major, large-scale collection of mutations for grass models. Resequencing the 1,504 mutants has allowed researchers to identify structural variants and mutations, providing an invaluable resource for grass models being used to improve candidate bioenergy feedstock crops such as switchgrass. Information on this new, large-scale collection of more than 90,000 mutations affecting nearly 60 percent of all rice genes is available on a publicly accessible database called KitBase. This comprehensive resource could help identify rice lines with mutations in specific genes and to characterize gene function.

To learn more about each of these stories, go to the JGI's Science Highlights page: http://jgi.doe.gov/category/science-highlights/.

Join us in San Francisco...

FOR THE 13TH ANNUAL
Genomics of Energy & Environment Meeting
HOSTED BY THE
U.S. Department of Energy Joint Genome Institute
Hilton San Francisco Union Square

MARCH 13–16, 2018

REGISTER NOW
http://usermeeting.jgi.doe.gov/

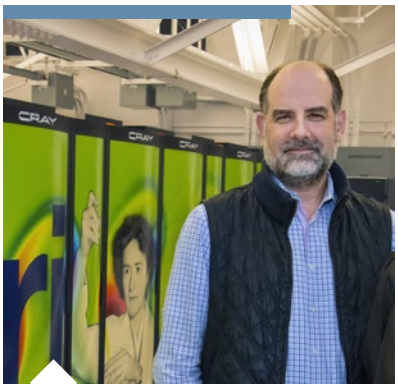
We welcome any and all researchers and students pursuing frontier energy and environmental genomics research. Also all current JGI Community Science Program (CSP) users, as well as investigators considering an application for future CSP calls.

AGENDA: Workshops to be held Tuesday, March 13–Wednesday, March 14. The Meeting opening keynote address is at 5pm on Wednesday evening, followed by the first poster session. The main sessions take place all day Thursday, March 15 through Friday afternoon, March 16.

KEYNOTE SPEAKERS: John Ioannidis, Stanford; Victoria Orphan, Caltech; Kristala Prather, MIT.

TOPICS: Microbial genomics, fungal & algal genomics, metagenomics, and plant genomics; genome editing, secondary metabolites, pathway engineering, synthetic biology, high-throughput functional genomics, high-performance computing applications and societal impact of technological advances.

Nikos Kyrpides Named ASM's 2018 USFCC/J. Roger Porter Awardee



Dr. Kyrpides will receive his award at the 2018 ASM Microbe Meeting in Atlanta, Ga.

Nikos Kyrpides, JGI's Prokaryote Super Program head, has been selected as the 2018 USFCC/J. Roger Porter Award recipient by the American Society for Microbiology (ASM). This award recognizes outstanding efforts by a scientist who has demonstrated the importance of microbial biodiversity through sustained curatorial or stewardship activities for a major resource by the scientific community.

For more than a decade, Dr. Kyrpides and his JGI colleagues have been working toward developing a comprehensive genomic catalog by sequencing the genome of at least one representative of every bacterial and archaeal species. (More on page 1)

VEGA

In parallel with the JGI User Meeting on March 14–15 will be the first ever **Viral EcoGenomics and Applications (VEGA) 2018 Symposium**: Big data approaches to help characterize Earth's Virome.

This Symposium brings together the "viral ecogenomics" community to foster discussions about how to best capture and characterize uncultivated viruses, understand the role of viruses in natural ecosystems, and functionally explore viral genetic diversity.

Contact The Primer
David Gilbert, Managing Editor
DEGilbert@lbl.gov
Massie Santos Ballon, Editor

17-JGI-4277