THE PRIMER

JGI
JOINT GENOME INSTITUTE
DEPARTMENT OF ENERGY
fall/winter
2019
volume 15
issue 3

in this issue

One of the accepted proposals focuses on microbial community structure and nutrient cycling in Chile's Atacama Desert. (Courtesy of A. Airo)

Preparing for a Sequence Data Deluge

Through the most recent Community Science Program proposal call for U.S. Department of Energy (DOE) Joint Genome Institute (JGI), a DOE Office of Science User Facility, 23 proposals were selected from 84 full submissions based on 105 letters of intent. Additionally, just over half of the accepted proposals come from primary investigators who have never led any previously accepted JGI proposal.

"These new CSP projects not only expand on and improve our collection of reference genomes, but deeply explore the structure and regulation of their genes," said Susannah Tringe, JGI User Programs Deputy. "They leverage new capabilities and higher throughput in DNA sequencing, synthesis and metabolomics to advance DOE mission science."

Among this year's accepted proposals:

- JGI collaborator and University of Wisconsin-Madison microbial ecologist Trina McMahon has previously made the case for conducting long-term time series studies (watch her talk at bit.ly/JGI2017TMcMahon). Her proposal calls for sequencing up to 1,000 metagenomes from samples collected over nearly 20 years in two freshwater lakes to understand how microbial communities in these waters change in composition, function and diversity. The proposal makes use of the JGI's very high sequencing throughput, enabled by the Illumina NovaSeq.
- To better understand how fungi orchestrate gene expression in response to key environmental conditions and vital substrates, Laszlo Nagy of Hungary's Biological Research Center, along with 39 collaborators around the globe, plans to generate broad comparative epigenomic data sets for five widely used fungal model organisms.
- Seagrasses store huge amounts of carbon, and the global loss of seagrass due to direct human activities is estimated at nearly a third of the total area, which has a significant

continued on page 2

Single Cell Genomics of Fungi

More than a million species of fungi are estimated to live on this planet, but most of that diversity remains unknown because the fungi have avoided detection and have not been cultured for study in laboratories.

Through projects such as JGI's 1000 Fungal Genomes, researchers aim to expand the known fraction of fungal diversity with representative genome sequences for various lineages. Even with such efforts, the majority of available genomes belong to just two major lineages, Ascomycota and Basidiomycota. The early-diverging lineages that are closer to the base of the Fungal Tree of Life have few representative genomes.

Reported in *Nature Microbiology*, a team led by JGI researchers has developed a pipeline to generate genomes from single cells of uncultivated fungi. The approach was tested on several uncultivated fungal species representing early diverging fungi, the earliest evolutionary branches in the fungal genealogy that provide a repertoire of important and valuable gene products.

"Conceptually, this is a pilot project," said JGI data scientist and first author Steven Ahrendt, who began working on

continued on page 2

Sequence Data Deluge

continued from page 1

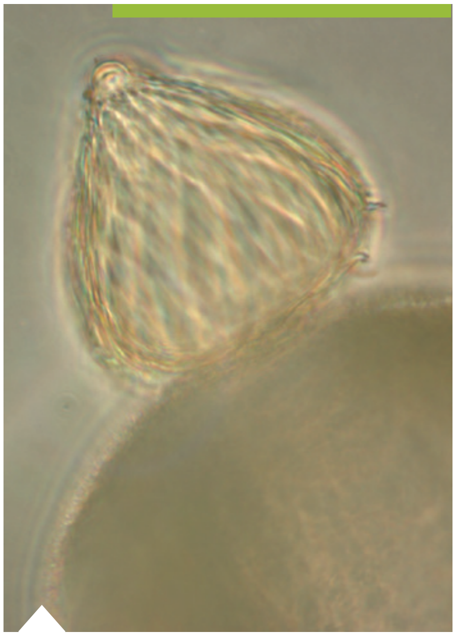
influence on the global carbon cycle. In 2016, JGI reported the first seagrass genome, *Zostera marina*. Yves Van de Peer of Belgium's Ghent University aims to build upon that effort with the Marine Angiosperm Genome Initiative to sequence and annotate five seagrass genomes.

- More than a decade ago, JGI launched the first effort toward developing the Genomic Encyclopedia of Bacteria and Archaea (GEBA). By 2017, the genomes of more than 1,000 phylogenetically diverse bacteria and archaea had been sequenced by the JGI. The latest GEBA proposal, led by William Whitman of Georgia Tech, is the largest project to date, with a goal of sequencing 10,000 microbes, half of them type strains, and coupling the sequences with transcriptomics and metabolomics.

See the full list of accepted proposals and learn more about each of them at jgi.doe.gov/preparing-for-sequence-data-deluge.

Fungal Single Cells

continued from page 1



***Blyttiomyces helicus*, on spruce pollen grain. (Joyce Longcore)**

the project as a postdoctoral fellow. "This is a similar idea to the approach JGI has taken with microbial dark matter — that the species are out there, but they don't show up in plate-based culturing."

The single-cell genomics approach was applied to eight fungi, seven of which belong to the early diverging lineages Cryptomycota, Chytridiomycota and Zoopagomycota. "Most of the phylogenetic diversity represents early diverging fungi," said JGI collaborator and study co-senior author Tim James of the University of Michigan. "We know from environmental DNA surveys that they're common in many habitats, but they're presumably microscopic so you really have to look for them. We don't know what they look like and we know we can't culture them, since what you can culture is not representative of what you see in environmental DNA. We would love to be able to look at a given sample and identify what the cells might look like, but we also want to look at the genomes of the organisms and infer what they're like. That's where single-cell genomics comes in."

What the study really highlights, James added, is that the single-cell approach is feasible for what he calls "fungal dark matter." The genomes of the species referenced in this paper are available in the JGI's Fungal Portal, MycoCosm, as well as in GenBank. Read the full story in jgi.doe.gov/expanding-fungal-diversity-one-cell.

Mapping Heat Resistance in Yeast

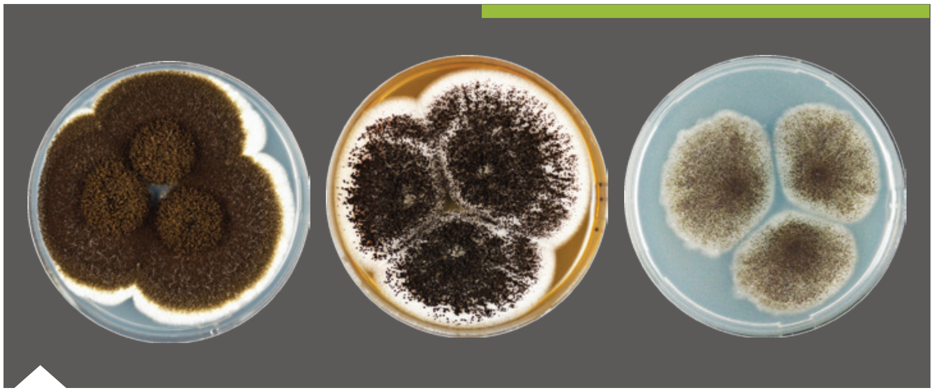
While efforts such as the JGI's 1000 Fungal Genomes Project aim to give researchers a better understanding of fungal diversity, they also highlight the corresponding need to be able to assign functions to unknown genes and improve functional annotation to harness the genomic data.

One of the challenges in understanding wild populations is that tests to find associations between genotype and phenotype don't allow researchers to compare across very distant species that no longer mate to produce progeny. The ability to pinpoint which genes underlie unique traits in far-flung species would speed up efforts to harness species diversity in traits relevant to DOE missions in energy and environment.

In a proof-of-concept study, researchers demonstrated that a new genetic mapping strategy called reciprocal hemizygosity analysis via sequencing (RH-Seq) can map associations between genotype and phenotype. Reported in *Nature Genetics*, the team led by researchers at the University of California, Berkeley and the Buck Institute for Research on Aging identified genes that promote heat resistance in the brewer's/baker's yeast *Saccharomyces cerevisiae*, allowing this species to grow better than its closest relative *S. paradoxus* at high temperatures (39°C/102°F).

The study, enabled in part by the JGI's Community Science Program, highlights the genes that could be used to make other fungal species besides *S. cerevisiae* more thermotolerant, and it paves the way for dissection of additional traits, in fungi and beyond.

SPOTLIGHT: *Aspergillus* Section Nigri



From left to right: *A. heteromorphus, A. ibericus, A. aculeatinus.* All images by Ellen Kirstine Lyhne, DTU.

In *Nature Genetics*, a team led by scientists at the Technical University of Denmark, the JGI and the Joint Bioenergy Institute (JBEI), a DOE Bioenergy Research Center, present the first large analysis of an *Aspergillus* fungal subgroup, section Nigri.

"*Aspergillus* is such an important genus, not just for biotechnology, but for all the roles it plays in agriculture, human health, enzyme production and food fermentation," said study co-author Scott Baker, Science Area Lead at the Environmental Molecular Sciences Laboratory, a national user facility at Pacific Northwest National Laboratory and a member of JBEI's Fungal Biotechnology group. "It's the Swiss Army knife of fungal genera as far as its impact on things that are important to us."

The results are part of a long-term project to sequence, annotate and analyze the genomes of more than 300 *Aspergillus* fungi. In choosing to focus on section Nigri, the researchers recognized the myriad of industrial applications for the fungi within this subgroup. They serve as production organisms for 49 of 260 industrial enzymes and are also very efficient producers of enzymes and secondary metabolites. JGI Fungal Program head Igor Grigoriev said that one of the aims in sequencing the genomes within the *Aspergillus* genus is to increase the catalog of carbohydrate active enzymes (CAZymes), which can degrade plant cell walls and thus have applications in the DOE mission to develop industrial processes for producing sustainable alternative fuels using candidate bioenergy feedstock crops.

The resulting compendium allows researchers to conduct intraspecies comparisons within section Nigri, as well as compare the fungal genomes within the subgroup to other genomes within the *Aspergillus* genus. All of the published Aspergillus genomes are publicly available from the JGI fungal genomes portal Mycocosm at genome.jgi.doe.gov/Aspergillus. Full story at jgi.doe.gov/spotlighting-differences-closely-related-fungi-species.

Genetic Code Changes in Yeast

Yeasts are some of the most important microbes used in biotechnology, but only a fraction of these yeasts has been harnessed for biotechnological applications. However, researchers studying various "non-conventional" yeast species aim to capitalize on yeast physiology and genetic features to drive biotechnology.

In nuclear genomes, a genetic code change where the amino acid assignment of a sense codon is swapped for a different amino acid is very rare.

Until 2016, the only such example in eukaryotes was the reassignment of CUG from leucine (CUG-Leu, its usual meaning) to serine (CUG-Ser) in a group of budding yeasts. Then, JGI researchers discovered a similar switch in *Pachysolen tannophilus*. Building on this information, another group of scientists followed-up, investigating the phylogenetic relationships in yeasts with standard and non-standard genetic codes. Reported in *Nature Communications*, this study looked at the genomes of 52 yeast species, including seven newly sequenced, using whole-genome data and mass spectrometry to determine phylogeny and genetic codes, respectively.

Within this data set, researchers observed all three CUG codon reassignments: CUG-Ser, CUG-Ala, and CUG-Leu. Why is the CUG codon unstable in yeasts? The authors hypothesize that the genetic code changes represent a profound defense mechanism. Full story at jgi.doe.gov/evolutionary-changes-genetic-code-of-yeasts.

—CINDI HOOVER

Maize Rhizosphere Microbiota Surveyed

In the *Proceedings of the National Academy of Sciences* (PNAS), researchers reported on the results of a large-scale field study that partially replicated earlier trials to identify soil microbes that colonize plants and which can be associated with particular traits. The work was conducted by an international team led by scientists at the Max Planck Institute for Developmental Biology, the Howard Hughes Medical Institute at the University of North Carolina at Chapel Hill, and Cornell University, and included JGI researchers.

"This is an extremely large and thorough temporal survey of the maize rhizosphere microbiota," noted study senior author Ruth Ley of MPI. "The dataset constitutes a rich resource for soil microbiologists and potentially plant breeders, once we zero in on the microbial traits that we'd like to breed for, to reduce dependence on fossil fuels in agriculture."

The work was conducted as part of the JGI's Rhizosphere Grand Challenge pilot projects involving maize and the model plant *Arabidopsis*. These projects highlighted JGI's capabilities and expertise that could be harnessed for scientific grand challenges assigned to the DOE's Office of Biological and Environmental Research in 2010.

Full story at jgi.doe.gov/reproducibility-matters-maize-rhizosphere.

From Great Prairie to Gigabytes

Excerpted from the DOE Office of Science feature by Shannon Brescher Shea

The American Midwest's Great Prairie is home to one of the planet's most complex systems of organisms that can't be seen with the naked eye. Researchers supported by the Department of Energy's Office of Science charted the genes of this community of microorganisms, an effort that could unlock new approaches to genetics, ecology, and agriculture.

To better understand the tiny organisms that support this ecosystem's fertility, researchers working with the JGI tackled the Great Prairie Metagenome Grand Challenge. Just a half a teaspoon of soil houses billions of microbial cells, which carry out essential processes such as carbon and nitrogen cycling. But scientists don't fully understand how they function or react to environmental changes.

Because most of these microbes are difficult to cultivate in the lab, researchers went straight to the source the fields and prairies of Iowa, Wisconsin, and Kansas.

A single gram of Great Prairie soil — about a small pinch of salt — contains more than one trillion bases of DNA (300 times more than the human genome), from more than 100,000 different species. In fact, the Great Prairie project generated so much complex data that when it started in 2009, JGI devoted more sequencing resources to it than it ever had to a single project.

Before the Challenge, scientists thought it might be impossible to piece together the genomes of individual microbes. But these investments resulted in tools and platforms that opened up new areas of research.

"If you can achieve genome binning in Kansas native prairie, you can do it anywhere," said Janet Jansson, chief scientist for biology at Pacific Northwest National Laboratory. From an analysis point of view, she says, "This is the worst case scenario in terms of high diversity and complexity."

Full story at science.energy.gov/news/featured-articles/2018/08-22-18.

Defining a Pan-Genome for Antarctic Archaea

Conditions such as extreme cold, high levels of salinity and the sheer distance from other parts of the world have kept microbial populations on Antarctica distinct and unique. In collaboration with the JGI, microbial ecologist Rick Cavicchioli and his team at the University of New South Wales (UNSW)-Sydney have been studying the haloarchaea that thrive in the very salty waters of Deep Lake to better understand how they contribute to global biogeochemical cycles, and possibly identify cold-temperature enzymes for biotechnological use.

The team compared two strains of *Halorubrum lacusprofundi* from Deep Lake and Rauer 1 Lake, one of the lakes in the nearby Rauer Islands. Additionally, they compared metagenome data generated from samples collected at four Rauer lakes to assess population level genomic variation. The genome and metagenome data are available through JGI's Integrated Microbial Genomes & Microbiomes (IMG/M) platform.

These data allowed researchers to define a haloarchaea "pan-genome." Described as the total pool of genetic material comprised by all members of a species, the pan-genome of haloarchaea contains the parts of the core genome shared across all haloarchaea, along with the pool of genes considered the flexible genome content that was likely acquired due to gene transfer events. Such events can occur in response to viral infections and are important for virus-host interactions. The team found that much of the strain variation appeared related to defense against viruses. Experiments infecting *H. lacusprofundi* with an Antarctic halovirus demonstrated differences in resistance between two strains. But perhaps the largest implication of this study is the evidence that *H. litchfieldiae* and *H. lacusprofundi* are found across all six lakes, suggesting that these species are endemic to Antarctica and distinct from other hypersaline environments. *—CH*

Antarctica's Deep Lake. (Rick Cavicchioli)

Mining IMG/M for Cas Proteins

The CRISPR-Cas system is an immune mechanism in bacteria that confers resistance to foreign genetic elements by incorporating short sequences from infecting viruses and phages. In the event of a new infection, the microbes use the genetic information encoded in CRISPR sequences to target the virus and release attack enzymes in the form of Cas enzymes to cut the DNA and disable the virus. In *Science*, a team led by Jennifer Doudna's group at the University of California, Berkeley, report the identification of active Cas enzymes — dubbed Cas14 — that target single-stranded DNA. In contrast, the seminal Cas9 proteins cleave double-stranded DNA.

JGI data scientist David Paez-Espino in Nikos Kyrpides' Microbiome Data Science group mined the IMG/M system with its large collection of publicly accessible metagenomic datasets from a wide variety of ecosystems around the world, conducting iterative searches using statistical analyses to continuously refine and improve the process. The results yielded several CRISPR-Cas systems, and based on several experiments conducted by the Doudna lab, close to 40 CRISPR-Cas14 systems belonging to eight subtypes were identified. Additionally, using Cas14a, the team was able to develop a Cas14-DETECTR that allows for CRISPR-based detection of ssDNA pathogens.

With few exceptions, the Cas14 proteins identified were found within the archaeal superphylum DPANN, named by a JGI-led team for the first five groups discovered: Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaea. The discovery of Cas proteins that can target single-stranded DNA molecules broadens the range of applications for CRISPR-Cas systems. While showcasing the unique capabilities provided from the IMG/M database, it also underscores the untapped potential waiting to be unearthed in sequencing and analyzing uncultivated microbes.

Corymbia citriodora subspecies *citriodora*. (Mervyn Shepherd)

Corymbia *Genome Expands Terpene Synthesis Knowledge*

Eucalyptus and *Corymbia* are closely-related eucalypts, but they inhabit different environments. Previous analysis of the *E. grandis* reference genome, an international effort by a team that included JGI researchers, revealed the largest number of terpene synthase (TPS) genes of any currently sequenced plant. Terpenes are economically important because of their use in industrial materials, pharmaceutical products, and as biofuel precursors. Collectively, hundreds of terpene compounds have been characterized from eucalypts.

As part of a proposal by DOE's Joint BioEnergy Institute (JBEI), the JGI worked on resequencing several eucalypt genomes to establish the feasibility of genome wide association studies for genetic traits that are desirable from a biofuels production perspective. By using genomic database alignment tools, researchers searched for TPS genes in *C. citriodora*. They then compared the list of putative genes from *C. citriodora* to known TPS gene sequences from *Eucalyptus* species and other plants. The locations of TPS genes and gene clusters were mapped against those of *E. grandis* to find differences in genome organization between the two species.

The annotation of two *C. citriodora* subspecies, reported in the journal *Heredity*, provides an excellent opportunity to investigate the conservation and evolution of this important gene family across eucalypt lineages. Improved knowledge of the evolution and selection of this gene family may help researchers manipulate TPS genes to increase terpene production for biofuel development. —*CH*

DAS Tool for Genome Recovery from Metagenomes

Developing a scalable computational approach that provides researchers with the ability to recover and reconstruct individual genomes, particularly from incredibly complex soil microbial communities, is crucial to understanding how these microbes respond and adapt to environmental changes. Since the inception of the Emerging Technologies Opportunity Program (ETOP) in 2013, a team of researchers including Jill Banfield at UC Berkeley has been working on improving methods for isolating and characterizing entire microbial genomes from sequences generated from environmental samples.

In *Nature Microbiology*, the team reported on the development and validation of a dereplication, aggregation and scoring tool (DAS Tool). To validate DAS Tool, the team applied it to assemblies from both simulated and real microbial communities and examined the bins obtained.

Using DAS Tool on data from soil samples, researchers were able to reconstruct 79 minimally contaminated (<5%) draft genomes to >70% completeness. Of those 79 genomes, 26 were high-quality draft genomes with >90% completeness. These results, noted JGI User Programs Deputy Susannah Tringe, suggest that extracting high-quality genomes from soil metagenome data is no longer "nearly impossible."

First Monoploid Reference Sequence of Sugarcane

Sugarcane is a cultivated crop that very efficiently converts solar energy into plant biomass, which can then be crushed to extract the sugar-laden juice that can be further purified to produce alternative biofuels. A C4 plant like sorghum and maize, sugarcane can also grow on nutrient-poorer soils and does not need to be planted annually. The United States produces more than half of the world's ethanol, mostly from corn. Brazil produces a quarter of the world's ethanol using sugarcane as the primary feedstock.

Improving sugarcane breeding methods using molecular biology techniques has been hampered by the crop's highly polyploid genome, which makes sequencing and assembly of the genome extremely challenging. The modern sugarcane cultivars are hybrids, each with more than 100 chromosomes; assembling a sugarcane genome is incredibly complex given its estimated size of 10 billion bases. As part of a proposal by JBEI, the JGI was part of an international team led by researchers from the French Agricultural Research Centre for International Development (CIRAD), who worked on sequencing and assembling fragments of sugarcane chromosomes into the first monoploid reference of the gene-rich part of the sugarcane genome. Reported in *Nature Communications*, their approach relied in part on having a sequence for sugarcane's relative sorghum, a JGI Plant Flagship Genome sequence, and knowing that

Sugarcane stalks. (Rufino Uribe, Flickr CC-SA 2.0)

there was a high level of colinearity between the two crops, which meant most genes in sorghum occurred roughly in the same order in sugarcane. The final sugarcane sequence generated is 382 Mb; 25,316 protein-coding gene models were predicted with more than 80 percent found to be colinear to comparable regions in sorghum.

Based in part on these positive results, in 2017, JGI approved a CSP proposal from the international team to complete the first draft genome sequence of the polyploid sugarcane variety R570. Among the numerous benefits expected from a genome sequence are better understanding the roles of genes in traits such as sucrose accumulation and disease, as well as targeting genes to improve biomass and sugar yield for biofuel production.

Archaeal Clues to Early Earth Conditions

For decades, longtime JGI collaborator Bill Inskeep of Montana State University has been conducting microbial field studies at Yellowstone. In this latest report, published in *Nature Microbiology*, he and his team describe a candidate phylum-level lineage of aerobic archaea found in iron-oxide microbial mats. The reddish hues caused by the presence of iron led Inskeep and his colleagues to name the archaeal lineage for the planet Mars.

To help determine where Marsarchaeota might fall amidst other known archaeal lineages, and through the JGI's CSP, the team used metagenome assemblies, transcriptomes and single amplified genomes from samples collected from several locations to thoroughly characterize the archaeal lineage, information that they believe will lend insights into discussions on the origin of archaea. They report that the Marsarchaeota are a sister group to the archaeal lineage named Geoarchaeota that Inskeep's team had previously identified and characterized, also with JGI's help. Additionally, the Marsarchaeota comprise 20-50 percent of the iron-oxide microbial mat communities in the 60-80°C temperature range.

Iron cycling is thought to have played a key role in the development of life on Earth; the iron-oxide microbial mats at Yellowstone researchers offer an analog for studying early Earth conditions. Thermophilic microbes are also of interest for their potential use in number of biotechnological applications.

Learn more about these and other JGI Science Highlights at jgi.doe.gov/category/science-highlights**.**

The iron (Fe)-oxide terraces at Echinus Geyser. (Bill Inskeep)

The 14th Annual DOE Joint Genome Institute

Genomics of Energy & Environment Meeting

New Lineages of Life Symposium

APRIL 2 – 5, 2019

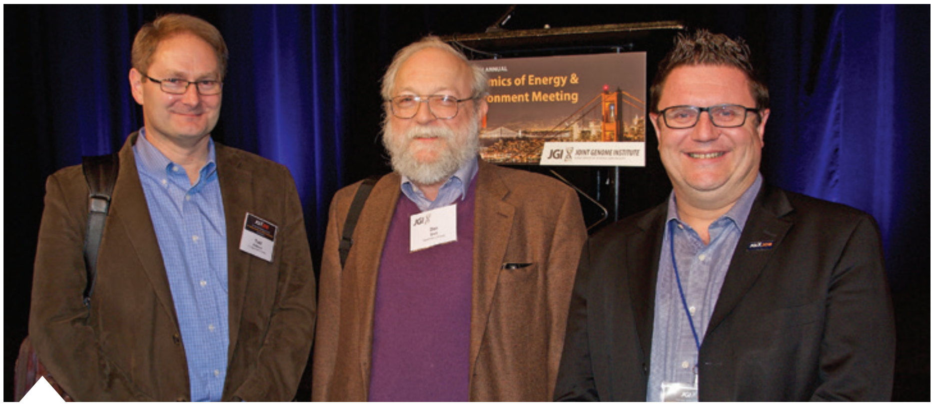
The annual meeting brings together members of the global scientific community pursuing grand challenges in energy and environmental genomics research, data science, and tech development.

Keynotes:
Arturo Casadevall, Johns Hopkins University
Mary Firestone, University of California, Berkeley
Michael Purugganan, New York University

A series of workshops precede the Meeting, including isolate and microbiome analyses using KBase, and a “Mini Colloquium” covering JGI strategic technologies. The Meeting will also be preceded by “NeLLi 2019 Symposium: From New Lineages of Life to New Functions.”

Learn more and register at usermeeting.jgi.doe.gov.

JGI Program Manager Dan Drell Retires



Dan Drell (center) at the 2018 JGI Genomics of Energy & Environment Meeting with BSSD Director Todd Anderson (left) and JGI Director Nigel Mouncey (right).

For more than a dozen years, Dr. Daniel Drell served as the JGI program manager at the Office of Biological and Environmental Research (BER), in the Biological Systems Science Division (BSSD). After 30 years in government service, most of that time with the BER, he retired in September. Before he stepped down, he reflected on his tenure with the JGI in a brief Q&A available at jgi.doe.gov/0918-jgi-program-manager-daniel-drell-retires.

New JGI Program Manager

Dr. Ramana Madupu is the JGI Program Manager as of October 1, 2018. Dr. Madupu is the BER Computational Biosciences Program Manager, and currently also manages the DOE’s Systems Biology Knowledgebase (KBase) program. The JGI and KBase will be co-located at the Integrative Genomics Building at Lawrence Berkeley National Laboratory in 2019.

Register Now

Microbial Genomics & Metagenomics (MGM) Workshop
February 25 – March 1, 2019
Walnut Creek, CA

Register at mgm.jgi.doe.gov

See us at these conferences:

AGU 2018 Fall Meeting
December 10 – 14, 2018
Washington, D.C.

Stop by Booth 847 and chat with JGI, KBase and EMSL representatives! Attend the December 9 EMSL/KBase workshop WS30!

PAG XXVII: Plant & Animal Genome Conference
January 12 – 16, 2019
San Diego, CA

Stop by Booth 509 and chat with JGI and KBase representatives!

Contact The Primer

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

19-JGI-5502