

Data Quality, Data Sets and New Directions: Plotting IMG's Next 10 Years

At the recent 10th Annual Genomics of Energy & Environment meeting hosted by the U.S. Department of Energy Joint Genome Institute (DOE JGI), a DOE Office of Science User Facility, Nikos Kyrpides (below), head of the DOE JGI Prokaryote Super Program, received the van Niel International Prize in Bacterial Systematics. The Van Niel Prize was established in 1986 in honor of microbiologist Cornelis Van Niel's contribution to scholarship in the field of microbiology, and is awarded every three years by the University of Queensland in Australia on the recommendation of a panel of experts of the International Committee on Systematics of Prokaryotes. Phil Hugenholtz, Director of the Center for Ecogenomics at the University of Queensland and a former DOE JGI colleague of Kyrpides, was on hand at the Meeting to present the award. Watch the ceremony at <http://bit.ly/JGI15KyrpidesVanNiel>.



An example of Kyrpides' efforts to systematically describe and classify microbes in action can be seen in the Integrated Microbial Genomes (IMG) data management system that his program developed and maintains in partnership with the Biosciences Computing Group of Berkeley Lab's Computational Research Division. IMG is the leading data analysis system of the DOE JGI's Prokaryote Super Program, and Kyrpides has been pushing the developments as the scientific lead of the project from its first working prototype in 2005 to its current incarnation. On the IMG system's 10th year anniversary, he took time to reflect on the milestones achieved thus far and future directions.

What are the highlights of the last 10 years to you?

In a period of 10 years, IMG has broken several records and has been established as one of the premier data management systems in the community for comparative analysis of microbial genomes and metagenomes. Its data size has grown 70-fold in terms of number of data sets and 22,000-fold in number of genes. We have currently almost 50,000 genomes in our system, containing 90 million genes. It's taken 20 years to sequence all of those genomes; I anticipate we will easily double that number in the next two years. We have 6,000 metagenome data sets, which contain 29 billion genes. As far as I know, this represents the largest publicly available database of metagenomics genes and therefore this is one more of IMG's records. We've grown from a few hundred to about 12,000 registered users in more than 90 countries. We *continued on page 2*

in this issue

A Decade Using IMG.	2
Tapping Microbial Communities in Colorado	4
Highlights from the Annual DOE JGI Meeting	6
Concerned about Melting Permafrost	8

Growing the Interest in Genomics

With a capacity crowd in attendance, the DOE JGI hosted the 10th Annual Genomics of Energy & Environment Meeting. To mark the occasion, instead of a single opening keynote address, the DOE JGI invited representatives from the three Bioenergy Research Centers to give a series of short talks that highlighted their collaborations with the DOE JGI, and featured applications of the basic science provided by the Institute. Blake Simmons from the Joint Bioenergy Institute (JBEI), Shawn Kaeppler of the Great Lakes Bioenergy Research Center (GLBRC), and Jerry Tuskan of the Bioenergy Science Center (BESC) all spoke briefly, while the closing keynote was delivered by Ed DeLong of the University of Hawaii at Manoa.

The themes of their talks echoed in presentations given over the three-day meeting held March 24–26, 2015 in Walnut Creek, Calif. Videos of these keynote talks, and of other presentations from the annual meeting, can be viewed on the DOE JGI YouTube channel at <http://bit.ly/JGIUM2015videos>. Images from the meeting are online at <http://bit.ly/JGI15UMphotos>. *continued on page 6*

Data and new directions

continued from page 1

provide an alternative source of data, particularly for metagenomes, and we add significant value through the integration of various data types, as well as with curation and annotation.

In terms of data integration, we've managed to integrate several different data types including one of the largest collections of curated metadata from the GOLD database, as well as several omics types including transcriptomics, metatranscriptomics, proteomics, and methylomics. In an effort to connect to our DNA synthesis program at the JGI, we have integrated a large collection of known natural products and connected them to their biosynthetic gene clusters, creating one of the largest resources in the field. We are currently working towards the integration of metabolomics and transposomics data produced at the JGI. Adding all of these means a completely different operation from the straightforward comparison of genes and genomes. With transcriptomes, for example, you're now talking about the expression of genes you already have, and expression levels vary under varying conditions. In transposomics, you look at the genes that are essential or have different fitness under varying conditions. So the original IMG's three-dimensional model of genes, genomes and functions has become more multidimensional as you add each of the different data types.

What do you think has helped IMG grow over the past 10 years?

One of the critical things is that it was a joint development between a group of engineers under the leadership of Victor Markowitz (<http://bit.ly/LBNL-BCG>), long experience in genomic data, and a group of biologists that had very strong genomics and bioinformatics backgrounds. Biologists provided the requirements

on how the data analysis tools and workflows should be organized, and the developers implemented exactly what the biologists wanted. It's clear there was a grand vision upfront to handle this much growth in the past 10 years. We can continue another 10 years on this current system, although we also need to start exploring new solutions for more efficient handling of the data deluge ahead.

One more of our early choices that I believed proved to be critical both for the growth and the success of the system was to offer only a single data processing option for all datasets submitted into our system. We do the annotation for the users, and we process the datasets the way we know best. Maintaining a huge system such as IMG gives you great power, and with great power comes great responsibility. I believe we're obliged to figure out and apply the best annotation practice at any time rather than allowing users to figure out what to use and which one choose as some other systems do. Providing an environment where all the data are uniformly processed and annotated is of paramount value and importance.

Looking forward to the next 10 years, what are some of the challenges the IMG system will need to tackle?

Our data sets are thousands of terabytes in size and we'll be going to petabytes soon. We need to scale at the level of hundreds of thousands of data sets and hundred of billions of genes. Right now our user interface can support the comparison of a few hundred datasets but what we need and what researchers are asking for is to compare thousands against thousands. No one is doing something like that now. Everyone is currently comparing a metagenome against isolate genomes, but no system can

efficiently provide a comparison of a metagenome against other metagenomes. Given the size of the data involved, that would take weeks and you can't do this efficiently on a production scale (i.e. on a weekly basis) even with high performance computing (HPC) right now.

The National Energy Research Scientific Computing Center (NERSC) is a vital partner in succeeding in the era of big data. We're already operating at the scale where processing of our data requires a HPC environment and we are very fortunate that at the JGI this is provided by NERSC. We need a bigger database and bigger computer clusters to support the growing community demand, but we also need to have the right computational environment to run our pipelines.

Another big challenge is how to support big data, without sacrificing data quality. For example, annotating the metadata in the Genomes OnLine Database (GOLD) is heavily manual, but it adds tremendous value to the sequence data. Manual annotation certainly contradicts with scaling, but the availability of metadata is critical information in order to interpret the data we have.

How do you see IMG integrating with KBase? What are the challenges here?

The two systems have different scientific goals and overall mission and because of that they also have fundamentally different design commitments, and follow different principles in data organization and user support. For example, while IMG's focus is on the comparative analysis of microbial genomes and metagenomes with emphasis on the interface between the two, KBase's focus seems to be more on the isolate genome side and metabolic modeling, at least for now. System integration



Former JGler Phil Hugenholtz (right) presented the Van Niel award to Nikos Kyrpidis (left). (Image by David Gilbert, JGI)

doesn't seem to be the right path here. Our primary goal instead is to enable users to review and analyze their data as well as move easily across the two systems. In order to achieve that we need to develop a seamless data transfer/exchange between JGI and KBase and this is currently the direction of our joint efforts.

What do you hope IMG will look like and be able to do for users in 10 years?

The exponential growth of sequence data is having already a dramatic

effect on the available solutions in data management. Some of the most frequently adopted solutions improvise on "cutting corners" and invariably select data partitioning instead of integration, and speed over accuracy, with detrimental effects on the quality and precision of the results.

My hope for the next 10 years is that IMG will persevere with its current course in supporting the JGI user community and JGI Science through its emphasis on high quality, and will maintain its position as a premier comparative analysis system for microbial genomes and metage-

nomes, worldwide. In terms of new directions, my expectation is that in the next decade, the biggest overhaul in the landscape of microbial genomics and metagenomics will be at the interface of the two, and therefore this is where a large part of future IMG developments will focus. In keeping with its goal of supporting the analysis of both the parts and the whole, I would like to see IMG playing a central role in enabling the identification and analysis of individual populations from environmental communities, as well as facilitating the elucidation of their role within the community.

What should the user community know about IMG as the data management system embarks on the next 10 years?

There's a huge amount of functionality in IMG already, but we certainly need to continue adding more. The two main directions in the near future include adding more functionality and efficiently supporting data/size growth. New functionality will include expanding the system to support the new data types produced from JGI functional genomics efforts (e.g. metabolomics and transposomics), but also creating specialized datamarts such as the IMG-ABC (integrating Natural Products and their corresponding Biosynthetic gene clusters).

We are also expanding our coverage of eukaryotic genomes to include more plant and fungal genomes into IMG. Our goal is to achieve a more holistic approach in data integration and analysis, in order to study complex biological systems, such as the plant microbiome. Of course, getting the large isolate genomes in the system will mean substantial increases in the comparison times and computational resources investments. But that's the obvious way to go, you need to have all the data integrated. If you have missing parts, discovery is missed.

Reconstructing Environmental Microbial Communities

Though microbes are critically important to environmental processes, accurately characterizing them is difficult because many cannot be cultivated in a laboratory setting. One workaround is to study DNA extracted from the metagenome, but studying a population rather than an individual raises different obstacles on the path to knowledge. The challenges of assembling genes and genomic fragments into meaningful sequence information for an unknown microbe has been likened to putting together a jigsaw puzzle without knowing what the final picture should look like, or even if you have all the pieces.

“For metagenomics,” said Jillian Banfield of the University of California, Berkeley and Berkeley Lab’s Earth Sciences Division, a longtime collaborator of the DOE JGI, “it is like reconstructing puzzles from a mixture of pieces from many different puzzles—and not knowing what any of them look like.”

Part of the problem lies in the fact that the more commonly used sequencing machines generate data in short lengths or fragments, on the order of a few hundred base pairs of DNA. Additionally, short-read assemblers may not be able to distinguish among multiple occurrences of the same or similar sequences and will therefore either fail to place them in the correct context, or eliminate them entirely from the final assembly, in the same way that putting together a jigsaw puzzle with many small pieces that look the same, is difficult. In a study published on the cover of the April 2015 edition of *Genome Research*, a team including DOE JGI and Berkeley Lab researchers compared two ways of using the next generation Illumina sequencing machines, one of which—TruSeq Synthetic Long-Reads—produced significantly longer reads than the other.



Cover image courtesy of *Genome Research* (Image by Zosia Rostomian, Berkeley Lab)

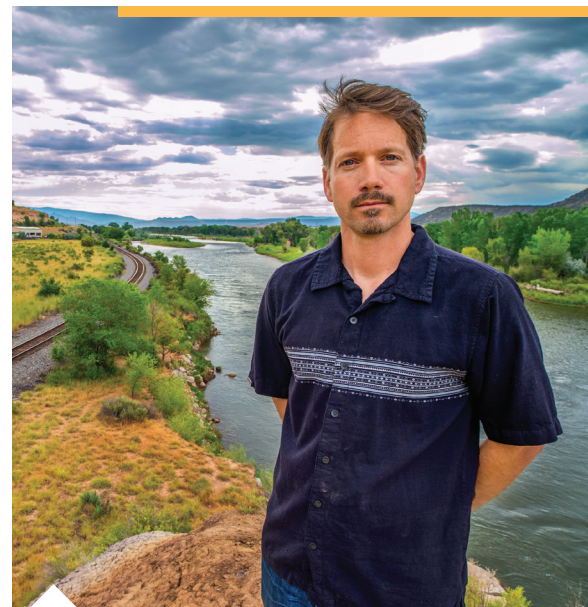
Metagenome data were generated from the Berkeley Lab-led DOE subsurface biogeochemistry field study site in Rifle, Colorado by a Banfield-led team. They evaluated the accuracy of the genomes reconstructed from the sequences produced by the two Illumina technologies to learn more about the microbes present in lower amounts than others and better determine the species richness of the metagenome samples. The project is part of the Berkeley Lab Genomes-to-Watershed Scientific Focus Area (SFA), which has a goal of developing an approach for gaining a predictive understanding of complex, biologically based system interactions from the genome to the watershed scale.

The team found that the longer reads captured more of the community’s diverse species. “Extending the analysis further to species with a lower abundance suggests that at least ... 2,100 different species are present,” they reported. “The true number of species is therefore expected to be much higher—

probably at the range of several thousands or tens of thousands of different species.”

DOE JGI Metagenome Program head Susannah Tringe noted that while the Rifle studies came out of the Community Science Project (CSP), the longer-read analyses conducted and reported in this study were motivated in part by the DOE JGI’s Emerging Technologies Opportunity Program (ETOP). Launched in 2013, the program seeks to develop and support selected new technologies that the DOE JGI could establish to add value to the high-throughput sequencing it currently carries out for its users.

Itai Sharon spoke at the 2014 DOE JGI Genomics of Energy & Environment Meeting on the benefits of multi-Kb Illumina reads. Watch his talk on the DOE JGI’s YouTube channel at http://bit.ly/JGIUM9_Sharon.



Berkeley Lab earth scientist Kenneth Hurst Williams describes the Genomes-to-Watershed Scientific Focus Area, and how the DOE JGI is contributing to the scientific effort. Watch the video at <http://bit.ly/JGI15WilliamsSFA>.

Expanding the Archaeal Tree of Life

Archaea, a domain of single-celled microorganisms, represent a significant fraction of the earth's biodiversity, yet they remain much less understood than bacteria. One reason for this lack of knowledge is relatively poor genome sampling, which has limited accuracy of the Archaeal phylogenetic tree. In a recent study published March 16, 2015 in *Current Biology*, researchers approximately doubled the genomic diversity sampled from this domain and reconstructed the first complete genomes for Archaea using cultivation-independent methods resulting in an extensive revision of the Archaeal tree of life.

Researchers from institutions including the University of California, Berkeley, DOE JGI, the Environmental Molecular Sciences Laboratory (EMSL), Pacific Northwest National Laboratory and Lawrence Berkeley National Laboratory used genome-

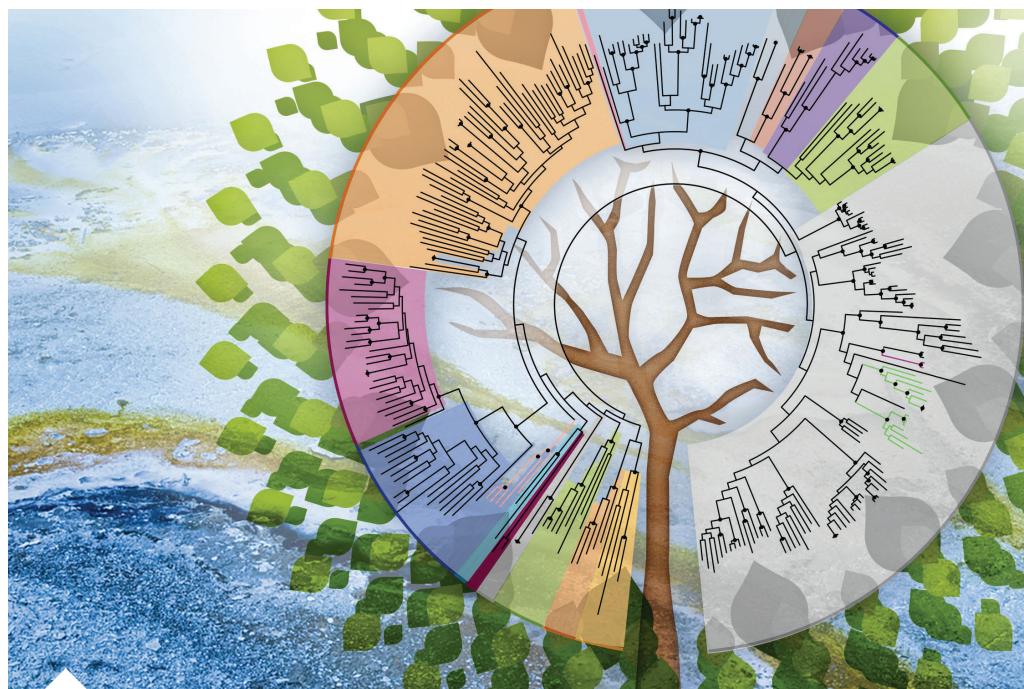


Image by Nathan Johnson, EMSL

resolved metagenomic analyses to investigate the diversity, genomes sizes, metabolic capacities and potential roles of Archaea in terrestrial subsurface biogeochemical cycles. They sequenced DNA in sediment and groundwater samples from a uranium-contaminated aquifer at DOE's Integrated Field Research Challenge site near Rifle, Colo. This is a former uranium mill and the primary site for DOE's Subsurface Systems Scientific Focus Area.

By sampling genomes of 100 different Archaea, researchers identified two novel phyla—named Woesearchaeota and Pacearchaeota—within the recently proposed superphylum comprised of 5 archaeal phylum-level groups abbreviated to DPANN. The unprecedented reconstruction of two complete genomes for members of this major superphylum showed these organisms have small genomes and limited metabolic capacities. Detailed metabolic analy-

ses of DPANN representatives revealed their primary contributions to the earth's biogeochemical cycles involve carbon and hydrogen metabolism. The data suggest these organisms may be involved in processing the sizeable reservoir of buried organic carbon, a finding that can be immediately implemented within genome-resolved ecosystem models to more accurately reflect the key role played by Archaea in the global carbon cycle.

Strikingly, the key features of DPANN Archaea closely parallel those of a putative bacterial superphylum. Their members are also predicted to have small genomes and to lack core metabolic pathways. Taken together, findings suggest these organisms depend on other members of the microbial community to survive and similar conditions have shaped two of the three major branches of the tree of life.

This piece was edited from the EMSL highlight, "Archaeal Tree of Life."



Jill Banfield, senior author of the study, discusses how the field of metagenomics has changed in the decade since her pioneering collaboration with the DOE JGI and what that partnership means moving forward with ongoing research in Rifle, Colorado. Watch her talk at <http://bit.ly/JGI15BanfieldSFA>.

Growing interest

continued from page 1

Putting Processes In Place

Blake Simmons from the Joint BioEnergy Institute (JBEI) delivered the first of the opening keynotes. He reminded the audience about the need to move to sustainable, renewable fuels in the transportation sector, where fossil fuels currently provide 97 percent of the fuels. “Beyond biofuels,” he added, “how do we displace the whole barrel of oil? What we derive from oil is basically everything.”

Speaking about improving the processes involved in biofuels production, he focused on ionic liquids that can help break down plant biomass and more efficiently liberate sugars. For this process, he said, JBEI researchers have been working with enzymes sourced from microbes whose genomes have been sequenced and analyzed by the DOE JGI.

Picking up on Simmons’ theme of improving pathways and processes for biofuels production in her presentation, Michelle Chang of the University of California, Berkeley talked about designing synthetic pathways for biofuels. “Fuels are the hardest product to make,” she said, as she

pointed out that studying and synthesizing pathways weren’t just for applications and chemicals production, but for also understanding very basic questions such as metabolisms in basic living systems production.

Building Up Biomass

Shawn Kaeppler from Great Lakes Bioenergy Research Center (GLBRC) talked about his team’s work on maize diversity, and their focus on the plant for its relationship to several candidate bioenergy grasses, including sorghum, switchgrass and miscanthus. He covered examples of genome-wide associated studies (GWAS), RNA sequencing and genetic mapping to help researchers discover genes and pathways associated with improving biofuels crops.

Similarly-focused plant talks came from DOE JGI collaborators Laura Bartley from the University of Oklahoma, and Tom Brutnell at the Danforth Center. Bartley gave an update on the progress of her switchgrass genome projects, among those selected for the DOE JGI Community Science Program. “We’re using switchgrass as

a place to generate hypothesis for understanding the accumulation of biomass. What we want is a large healthy plant that doesn’t compete with food crops,” she said, reminding the audience of the goals of bioenergy crop researchers. “Switchgrass has many of the features we want for one of these dedicated bioenergy crops.” So far, the switchgrass resequencing project led by Bartley has found 31 million DNA sequence variations (SNPs) across all of the genotypes, though nearly two-thirds of them are unique to a single genotype. She added that the switchgrass genome itself is still being improved, and some of the data are being validated by the resequencing findings.

Brutnell’s talk wasn’t on a bioenergy crop, but rather on a plant model for candidate bioenergy grasses. He works on green foxtail (*Setaria viridis*), another DOE JGI Community Science Program project and model for panicoid grasses, which include crops such as maize, sugarcane, miscanthus, and sorghum. “This is a really exciting place now in plant science,” he said. “Many plants are now sequenced by JGI, and provide us a way for moving across lineages and discovering new genes in photosynthesis.” He discussed the use of CRISPR/Cas 9 editing tools on *Setaria* plants for various studies. He also referenced the work done by Oak Ridge National Laboratory’s Jerry Tuskan on poplar. “What Jerry has done for poplar, we’d like to emulate for *Setaria*,” he said.

A Mutual Appreciation

Jerry Tuskan from Oak Ridge National Laboratory (ORNL) and the Bioenergy Science Center (BESC) delivered the final opening keynote talk. He focused on work being done to understand the mutualistic relationship between the poplar tree and



Science program heads (left to right) Susannah Tringe, Igor Grigoriev, Tanja Woyke, Nikos Kyrpidis and Jeremy Schmutz field user suggestions while DOE JGI Director Eddy Rubin moderates the session. (Image by Roy Kaltschmidt, Berkeley Lab)

the fungus *Laccaria bicolor*, both of which were sequenced by the DOE JGI. Knowing how plant health can be helped by beneficial microbes, as well as how to maintain that association, he said, can help bioenergy researchers cultivate agriculture-quality land for biofuels crop production without encroaching on farms. One of the studies he highlighted featured the search for a poplar protein that moves into *Laccaria* and appears to change its behavior. “We’re exploring this as a way to control behavior of host and microbiome,” he said. “There’s a lot more signaling than we were aware of. A broad genomics approach allows us to tease apart some of these communication mechanisms and then we can favorably influence how individual fungi interact with individual poplar.”

It’s hard to reference the first mutualistic fungus to have its genome sequenced without mentioning the project head, longtime DOE JGI collaborator Francis Martin of the French national research institute INRA. He described the interactions between plants and fungi in forest ecosystems as “a 400-million year old affair that shaped the biosphere via colonization of the terrestrial environment.” His own talk at the Meeting focused on the use of comparative genomics to decipher ecological traits to understand the interrelationships between four types of fungi: white rots, brown rots, leaf-decayers and ectomycorrhizal (ECM) fungi. One of the findings he shared is that on average, ECM lineages have reduced the complement of plant cell wall degrading enzymes compared to the number of genes in ancestral white rot wood decayers.

Sophien Kamoun of The Sainsbury Laboratory presented the flip side of the relationship between plants and microbes. He has previously collaborated with the DOE JGI on projects to sequence *Phytophthora* fungi, which can damage legumes. At the Meeting, he focused on filamentous plant

pathogens and how they are evolving to perturb the plant, even as they are under pressure to thrive. “As evolution and natural selection drives us to adapt, they specialize,” he said, adding that lineages with adaptable genomes are less likely to go extinct.

Finding a Microbial Community’s Rhythms

Ed DeLong from the University of Hawaii at Manoa delivered the closing keynote. He has collaborated with the DOE JGI for more than a decade, and his projects have focused on various marine microbes, ranging from deep-sea plankton to methane-oxidizing archaeon and Antarctic bacterioplankton.

DeLong’s talk focused on the application of omics techniques to discern the similarities between the daily rhythms of marine microbial communities located oceans apart. As he and his colleagues at MIT, the University of Hawaii, and other institutions reported in a recent paper published in *Proceedings of the National Academy of Sciences*, despite tremendous differences between their habitats, there are strikingly similar temporal patterns that trigger metabolic functions in

other members of the microbial community.

“I would definitely not have predicted that,” he said.

The team made use of the transcriptomes—the collection of RNA sequences in a cell that can tell researchers when and where genes are turned off—of microbial communities off the coast of Monterey, California and in the open ocean off Oahu, Hawaii. These communities contained species such as *Prochlorococcus*, *Synechococcus*, and *Ostrococcus*, all among the most productive and abundant photosynthetic microorganisms.

DeLong and his colleagues then compared these findings with data obtained at Station Aloha near Oahu, Hawaii, where 25 years of data have been collected as part of the Hawaii Ocean Time Series. Here they found once more that there were multiple species dominating the communities at various points throughout the day.

“There’s a finely tuned biological orchestra happening everyday that ripples through the whole community of these ocean waters,” he concluded. “We’re close to being able to map microbes in 4 dimensions,” he said, “and if you’d said that 10 years ago they’d have sent you to the hospital.”



Metabolomics group lead Trent Northern leads a tour for users. (Image by Roy Kaltschmidt, Berkeley Lab)

Microbial Activity in the Melting Arctic

The frozen soils embedded in the Arctic store roughly 1.5 billion tons of carbon. Rising global temperatures concern climate researchers because the permafrost soils may thaw completely. This event could potentially lead to the release of potent greenhouse gases carbon dioxide (CO₂) and methane in what would be the largest contribution of carbon transferred to the atmosphere by a single terrestrial process.

A team of scientists from institutions including the DOE JGI, Berkeley Lab, Pacific Northwest National Laboratory (PNNL) and the United States Geological Survey sought to determine the composition of microbial communities and their role in degrading permafrost organic carbon and the subsequent production of CO₂ and methane. A better understanding of these processes is necessary, they maintained, for generating more accurate models and thus predictions of the environmental consequences. The team reported on the application of multiple molecular technologies collectively referred to as “omics” to better characterize microbial activities in a paper published online March 4, 2015 in the journal *Nature*.

Microbial ecologist Janet Jansson from PNNL led the team that investigated three types of Alaskan soils,

ranging from completely thawed to completely frozen. Metagenomics (MG), or environmental genomics, enabled the researchers to identify the phylogeny—i.e., history of organismal lineages—of the communities’ microbial members, and the functional gene composition. Metatranscriptomics (MT) allowed the team to determine which genes were being expressed. Finally, metaproteomics (MP) provided insights on which proteins were actually produced.

For the study, researchers relied on soil cores collected in Alaska, focusing on their bacteria and archaea. Comparison of the MG, MT and MP data from the three soils provided insight into the linkages between omics data and elemental cycling pathways. In the thermokarst bog they found the highest rates of methane production and identified several microbes involved in this pathway. Additionally, several genes involved in methanogenesis were detected in both the MG and MT data sets and corresponding proteins in the MP data sets. Three draft methanogen genomes were identified, and comparisons with sequenced methane producers suggest these are previously undescribed microbes.

The work done by Jansson and her colleagues is just one of the ecosystem studies being conducted by the Department of Energy in Alaska. Through the Next-Generation Ecosystem Experiments (NGEE Arctic) project in Barrow, Alaska, a consortium of academic institutions and national laboratories is developing a process-driven ecosystem model that will allow researchers to better predict the evolution of Arctic ecosystems in a changing climate. Jansson is currently leading a Community Science Program project at the DOE JGI for sequencing of samples collected for the NGEE Arctic project.



Study first author Jenni Hultman prepping permafrost samples. (Janet Jansson, PNNL)

Science Programs Deputy Jim Bristow Retires



(left to right: DOE JGI Director Eddy Rubin, Science Deputy Jim Bristow, Operations Deputy Ray Turner, and Bristow’s wife Willow Dean) After a decade with the DOE JGI and 35 years with the University of California, Jim Bristow has retired. “When Jim came to the JGI, we were at the end of the Human Genome era and we were focused on one project and one tube of DNA,” said Rubin. “Jim figured out how to convert us to a user facility, Jim launched all the science programs and the project management office. Now we have user reviews and user committees and user meetings. He was such a foundational force.” As a parting gift, Turner had a sign made to mark Bristow’s next possible undertaking: a bike shop on Vashon Island in the Puget Sound. A video recapping Bristow’s time at the DOE JGI is at <http://bit.ly/JGIBristowRetirement>.

Contact The Primer

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

