JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

the PRIMER
Spring 2011 Volume 8 Issue 2

Genomics Revolution Revisited: 6th DOE JGI User Meeting

BY MASSIE SANTOS BALLON

When Rob Knight from the University of Colorado took the stage at the 6th Genomics of Energy & Environment Meeting in Walnut Creek, Calif., he likened the decreasing costs of sequencing and resulting democratization of the process to the fall of the Berlin Wall.

"Don't be a bystander – join the revolution," he urged the 420-plus members of the audience gathered for the DOE Joint Genome Institute's annual User Meeting on March 22-24, 2011. Though his talk focused on what he called "microbial biogeography" of microbial communities on various parts of the human face as well as those transferred from fingertips to computer keyboards, he also noted that the vast microbial diversity can all ultimately be mapped to a single tree.

"There is one universal tree of life, which is why projects such as GEBA are so critical," he said, referencing the DOE JGI Genomic Encyclopedia for Bacteria and Archaea project being led by Jonathan Eisen to fill in the microbial gaps on the tree of life.

Given the numerous comments regarding the significant changes in sequencing costs and capacities, several talks discussed the DOE JGI's future growth (see page 3). Persis Drell, Director of the SLAC National Accelerator Laboratory, delivered the User Meeting's opening keynote, offering attendees a look at how another user facility is making the transition from a single service to offering multiple applications (see page 2).

The role microbes play in ecosystems was highlighted by Lawrence Berkeley National Laboratory's Terry Hazen, who followed Drell with a keynote *continued on page 4*

DOE JGI's Susan Lucas (foreground, left) and others look at oil samples collected by keynote speaker Terry Hazen (right). (Roy Kaltschmidt, LBNL)

also in this issue

Rumen-ating on Approximately 30,000 Novel Enzymes

Among the images associated with farm life is that of the cow contemplatively chewing on grass. Shown the same image, biofuels researchers have wondered how microbial communities in the cow's forestomach or rumen help break down the cellulose and hemicellulose in the plant cell walls to extract nutrients, and how they can harness that information to work toward commercial-scale biofuel production.

In the January 28, 2011 issue of *Science*, a team of DOE JGI researchers and members of the Energy Biosciences Institute answered the latter question, reporting the discovery of nearly 30,000 novel enzymes in the microbial community of a cow rumen that can break down complex sugars such as cellulose into small sugars. The project stands out for its sheer scale—270 billion bases of sequencing—making it the largest published data set generated from sequencing a single sample. In addition, the resources deployed in sequencing: single-cell genomics, data analysis and computational capabilities, and all the intra-departmental collaborations entailed within the DOE JGI and with its collaborators, made the project and publication a reality.

A fragment of switchgrass decomposing in contact with cow rumen microbes. (Damon Tighe, DOE JGI)

"Our study demonstrates *continued on page 10*

Spring 2011 Volume 8 Issue 2

SCALING SCIENTIFIC CHALLENGES

As the DOE JGI genome sequencing portfolio and the data sets continue to expand, the keynote speakers selected for the DOE JGI 6th Annual User Meeting on March 22, 2011 discussed how their respective laboratories have addressed similar challenges and retooled to offer new resources and applications for a diversifying community of collaborators.

When Persis Drell, director of the SLAC National Accelerator Laboratory, delivered the opening keynote, she began with the message she received late one night in April 2009: "We have a laser."

Persis Drell (Roy Kaltschmidt, LBNL)

For decades SLAC has been a center for cutting-edge physics but when the Linac Coherent Light Source (LCLS) was turned on, the facility expanded its services to include astrophysicists and structural biologists as well.

Drell said the key is the intense x-ray beams — the LCLS shoots six billion electrons into a 30-micron space — that allow researchers to see structures on an atomic scale. One potential use for beams this powerful is in the field of structural biology, where producing crystals from molecules of interest and making them large enough to diffract well has been a challenge. Drell noted that one of the early studies has demonstrated that structures can be imaged using the LCLS — which can provide information at the 8.5-angstrom resolution level — from small crystals. One of the long-term goals is to make a similar process work for a single protein.

She cautioned that the techniques are still evolving. "I think we'll have very few partnering biologists until we have some techniques worked out."

Though the work done these past two years has been largely proof of principle, Drell noted that among the near-term goals are increasing capacity to meet the demand for time with the LCLS. She also noted that competition is coming in the form of other lasers being turned on in Japan and Germany. "We will have fun while we hold this frontier," she noted. "A new scientific frontier is being opened, and the biggest surprises are yet to come."

After Drell, microbial ecologist Terry Hazen then took the stage to describe the research done by his team after an explosion on the Deepwater Horizon oil drilling rig on April 20, 2010 killed 11 men and led to nearly five million barrels of oil being spilled into the Gulf of Mexico. The resulting efforts to cap the rig and clean up the oil spill ultimately involved nearly 32,000 people and 7,000 ships.

"What we wanted to do was apply a systems biology approach to the oil spill," Hazen said, "looking at exactly what happens at all levels and developing models for cellular properties, communities and ecosystems."

In a paper published in *Science* four months after the first explosion, Hazen and his team reported that "at 10 kilometers beneath the ocean, the oil had completely disappeared."

"Where's all the oil?" Hazen asked. "We took 170 samples from July to August and we were looking for that crude, and we couldn't find it. We went to same depths and it wasn't there. We took samples back to lab and chemically we couldn't detect it by any technique that we had."

Terry Hazen (Roy Kaltschmidt, LBNL)

To explain the mystery of the missing oil, the team used plume samples taken from May to October of 2010, tracked the thousands of bacterial and archaeal species using a DNA-based array developed by Lawrence Berkeley National Laboratory, and took advantage of the DOE JGI's single-cell genomics expertise to identify a new "oil-seeking" species related to *Oceanospirillales*.

"If I had to look for oil-degrading bacteria, the Gulf of Mexico would be the place," Hazen said, citing National Academy of Sciences figures that estimate as much as a million barrels of oil go into these waters annually. "These bugs have to be adapted to the oil." The *Oceanospirilalles* bacteria, Hazen added, have genes that can degrade C6 to C20 alkanes and exhibits a rapid chemotaxic response to the presence of oil.

Hazen's team is continuing its studies of the bacteria from the oil spill, and he noted that "though it was certainly an ecological disaster, the systems biology approach explains exactly what was going on."

Watch the keynote speeches from Drell and Hazen on the JGI's YouTube channel.

the PRIMER / 3
Spring 2011 Volume 8 Issue 2

Laying the Foundation for a Genomic Foundry

At the beginning of his talk at the 6th Annual DOE JGI User Meeting, Daniel Distel, director of the Ocean Genome Legacy Foundation, acknowledged the Institute's Community Sequencing Program for making it possible for small labs like his to have genomic data (see page 8).

In the decade since the DOE JGI was founded however, sequencing costs and technologies have changed drastically, raising questions about the Institute's continued usefulness as a user facility. Taking advantage of the audience gathered for the meeting, DOE JGI Director Eddy Rubin posed several questions to the user community in attendance, running ideas by them regarding the DOE JGI's future directions.

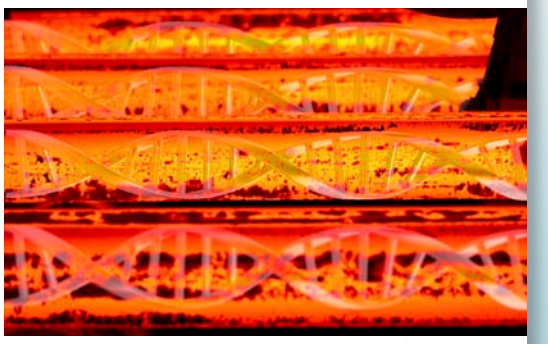
"We're here to enable DOE-relevant science," Rubin stressed while reiterating the call for the user community to acknowledge and notify the Institute of publications and new grants funded based on genomic information and analyses provided by the DOE JGI. The information, he noted, would enable the DOE JGI to better measure its positive impact in the scientific community and demonstrate value for the taxpayers' investment.

"We view ourselves becoming a Genomic Foundry modeled after DOE Nanotechnology Centers," said Rubin. While the Institute would still be user-driven and focus on solving energy and environmental problems, the DOE JGI is positioning itself to move beyond sequencing, providing resources to help researchers turn the sequence data into useful biology.

Part of that repositioning involved teaming with the National Energy Research Scientific Computing Center (NERSC) facility at Lawrence Berkeley National Laboratory last year to gain access to large-scale computational resources. Another Institute goal Rubin outlined was to have researchers and postdoctoral fellows in residence at the DOE JGI instead of just receiving their samples for study. The Visiting Scientist Program (VSP) would offer researchers access to experimental, computational, and personnel resources for genomic research, with length of stay determined by project complexity. (http://www.jgi.doe.gov/whoweare/visiting-scientist-program.html)

To illustrate the point about the changes in sequencing technologies, Len Pennacchio, head of the DOE JGI Genome Technologies Department, compared the performance of the sequencers used at the Institute, from the recently phased-out Sanger machines to the Roche 454 sequencers to the Illumina Genome Analyzer and HiSeq machines. "I think it's clear technology has been changing rapidly," he said. "At the JGI, we've gone from producing one megabase of sequence a day with Sanger to producing 20 gigabases a day with HiSeq. However, one of the tradeoffs with this massive output increase is read length dropping from 700 base pair reads to 100 and this presents new challenges in data analysis."

Pennacchio also described the results from pilot studies using the Pacific Biosciences single-molecule real-time (SMRT) sequencer. Last year the company selected the DOE JGI as one of 10 facilities to do early testing on their machine. Both strengths and weaknesses were presented and included impressively long reads (1,500 base pairs, on average), faster run times, and less genome coverage biases. On the flip side, initial throughput is less and error rates are more than existing short-read technologies. Continued R&D is expected on this platform. The machine will not replace the other sequencing platforms, he said, and while there are plans to use it for projects such as single cell profiling and metagenome and eukaryote genome assembly validation, the real question is what the



DOE JGI user community would like to do with the sequencing technology.

"We know sequencing analysis is going to be our bread and butter, but we want to add front end and back end capabilities, such as custom analysis," he said. "We want to challenge you all to think about how you could use this technology," Pennacchio added, reminding the audience that the Community Sequencing Program is now accepting proposals for the 2012 Portfolio.

DOE JGI Deputy Director Jim Bristow carried on the emphasis on the CSP program in his talk. Briefly addressing the crowd, he outlined the goals behind the current Program call, and the emphasis on community research.

"We have tremendous power to build communities around sequencing projects," he noted, using the *Daphnia pulex* genome project as an example (see page 7). "We want to get a collection of proposals that are potentially synergistic and consider getting PIs together beforehand to maximize value of data being generated."

He said the proposals should be for projects involving plant-microbe interactions, microbial emission/capture of greenhouse gases, and the metagenomics of biogeochemistry. For more information about the DOE JGI Community Sequencing Program and details about the proposal requirements, go to http://www.jgi.doe.gov/programs.

Genomics Revolution *continued from page 1*

speech (see page 2) on the research he and his team conducted in the months immediately after the April 2010 Deepwater Horizon oil spill in the Gulf of Mexico. With a bit of show and tell, Hazen held up small vials containing oil from the now-capped wells, noting, "This is worth more than platinum." He spoke about the work involved in understanding why some of the oil from the undersea wells never made it to the surface, and how microbes in the Gulf contributed to the clean-up.

Several talks emphasized marine microbial studies (see page 11). Peer Bork from the European Molecular Biology Laboratory used the Global Ocean Sampling project led by J. Craig Venter several years ago as an introduction to environmental metagenomic studies, before turning to the ongoing Tara Oceans Expeditions, which is the basis for a handful of DOE JGI Community Sequencing Program (CSP) projects. He also compared the challenge of analyzing the flood of sequence data generated from these projects with the ongoing Human Gut Microbiome project, the data catalog that DOE JGI researchers are involved in maintaining.

Mary Ann Moran from the University of Georgia looked at better ways to collect –omics data given "now there's a whole suite of tools and a nice emergence of model organisms." Christopher Scholin from the Monterey Bay Aquarium Research Institute highlighted techniques that his team has developed to do real-time probe array analysis using an ecogenomics sensing system they call ESP, for Environmental Sample Processor.

"Most of the analytical work we do in the ocean, we can reference it immediately to the prevailing conditions," he said. "The ability to carry out interactive experiments and test hypotheses remotely from a molecular biology perspective is within reach."

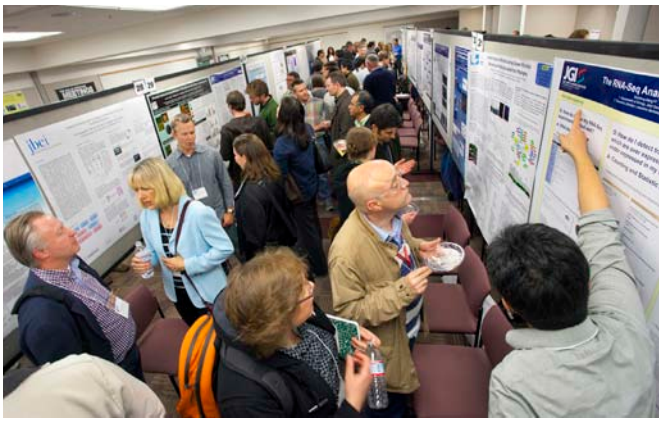
Another DOE JGI Program area, Plants, was highlighted during the Meeting. Collaborator Ruth Ley from Cornell University referenced the talks delivered by Bork and Knight, noting "the pattern emerging is that there's a core microbiome shared by all of us, and there's a variable microbiome influenced by other factors." She then shifted gears to introduce the DOE JGI Grand Challenge project focused on the rhizosphere microbiome, drawing parallels between how microbes influence plants and the way the gut microbiome influences health in humans. Another Grand Challenge project involving the Great Prairie soil metagenome was later discussed by Michigan State University's Jim Tiedje (see page 9).

Cornell University researcher Ed Buckler followed Ley's talk with a presentation on how the diversity of maize plants can be used to understand complex structures. Millions of years of domestication have led to what he described as an exquisitely fine-tuned system where several dozen genes are involved in determining the angle of a leaf in relationship to the sun.

Buckler ended his talk by pushing for genomics-enabled breeding, which would speed the process of selecting the right varieties to cross and cost less than planting crops in the field for standing breeding methods. Tom Juenger from the University of Texas at Austin, who spoke on crop modeling predictions for switchgrass in anticipation of climate changes, echoed his final message. Though the grass doesn't fare too well on marginal lands in the models, he pointed out that the computer programs fail to account for genetic variation in this potential bioenergy feedstock.

Scott Hodges from the University of California, Santa Barbara discussed columbine flowers, noting that the genome of Colorado's state flower *Aquilegia coerulea* is now available in the DOE JGI's plant portal Phytozome. Sequenced under the aegis of the CSP, *Aquilegia* is positioned on

DOE JGI's Kanwar Singh described how the PacBio SMRT sequencer works to User Meeting attendees.



One of the two User Meeting poster sessions.

Spring 2011 Volume 8 Issue 2

DOE JGI Director Eddy Rubin (left) in conversation with Michigan State University's Director for Center of Microbial Ecology Jim Tiedje (right).

NERSC Director Kathy Yelick discussed the increasing computing demands for genomic studies.

the phylogenetic tree nearly equidistant between the current model systems *Arabidopsis* and rice. The genus is shedding light on how plants adapt, at the molecular level, to environmental changes. Magnus Nordborg from the Gregor Mendel Institute compared *Arabidopsis thaliana* to *Arabidopsis lyrata*, another CSP project, to understand why the genomes of these two species differ so tremendously in size.

Mike Thomashow of Michigan State University also discussed *Arabidopsis*, but focused on the regulatory networks involved in its ability to tolerate cold. By understanding the pathways associated with this model system's ability to deal with stressful environmental conditions, he said, the information could be used to broaden regions for crops and improve their yields.

Other speakers reminded the audience that genome sequencing has broad environmental applications. Penn State University's Stephan Schuster talked about extracting DNA and assembling genomes from extinct and endangered animals such as the mammoth and the Tasmanian devil. "Looking back and studying populations from a long time ago gives you the ability to work with populations and make predictions on what we could learn for endangered species in the future," he said. Gene Robinson from the University of Illinois discussed the use of genomics to understand social behavior in insects such as paper wasps and bumblebees.

Pam Silver from Harvard University talked about studies done on carboxysomes in cyanobacteria and how they regulate carbon fixation before introducing some ideas that she called "really blue sky." One of them, funded by the DOE Advanced Research Projects Agency — Energy or ARPA-E, is focused on using bacteria such as *Synechococcus* (previously sequenced by DOE JGI) to produce octanol for cars.

Several talks featured bioenergy projects funded by the CSP. Jerry Tuskan from Oak Ridge National Laboratory and the BioEnergy Science Center gave an update on the association genetics studies being done with poplar to make it a viable candidate feedstock for biofuel production. Starting with 1,400 genotypes collected from the wild, he and his team have grown 1,100 individuals and are in the process of phenotyping all of them.

Daniel Distel, director of Ocean Genome Legacy, discussed the shipworm, a wormlike marine mollusk also referred to as the termite of the sea (see page 8). Like other metagenomic projects at the DOE JGI involving gut microbiomes, the shipworm's genome is being studied to identify enzymes that could be used to break down cellulose for cellulosic ethanol production.

DOE JGI's Zhong Wang closed the User Meeting by contrasting Distel's presentation of "CAZyme champs that encode the most enzymes for degrading cellulose" in shipworm and DOE JGI projects such as termite hindgut, with the story of the cow rumen metagenomic project that was published in *Science* earlier this year.

Finally, with all the emphasis on sequencing, Kathy Yelick, director of the National Energy Research Scientific Computing Division at LBNL, discussed the increasing computing requirements for genomic studies without increasing energy input. Based on current projections, she said, the DOE JGI will require a Petascale machine just for its own use. "Moore's Law is sort of pathetic next to genomic growth," she said. "Soon it will be much more expensive to pay for computers to analyze sequencers than to pay for sequencers themselves."

Videos from the 6th Annual User Meeting are available on the DOE JGI YouTube channel.

Spring 2011 Volume 8 Issue 2

Battling Brown Tide with Genomics

Previously known collectively as "red tide," the term "harmful algal blooms" (HABs) was introduced two decades ago to describe the accumulation of algal biomass that can sometimes also turn the ocean waters brown or green and reduce the amount of light and oxygen available in an ecosystem.

First published online February 23, 2011 in the *Proceedings of the National Academy of Sciences*, a team of researchers including DOE JGI scientists led by Igor Grigoriev, reported the first complete and annotated genome sequence of a HAB species: *Aureococcus anophagefferens*.

Algae sequester significant amounts of carbon, and *Aureococcus*, so tiny that 50 of them side by side span the width of a single human hair, is no exception to the rule. "It's a photosynthetic organism that plays a big role in carbon cycling, particularly in coastal ecosystems, and can degrade organic carbon," said first author Christopher Gobler of Stony Brook University.

When billions of *Aureococcus* cells come together, however, they outcompete the other marine phytoplankton in the area, damaging the food web in the marine ecosystem as well as economically impacting the shellfish industry. Economic losses attributed to this marine phytoplankton and other HAB phenomena in the United States over the course of the last decade have been estimated at one billion dollars. "When one of these blooms occurs and you get a billion cells per liter, it represents milligrams of carbon per liter, which is much higher than you typically see in coastal ecosystems," Gobler said.

The 56-million base pair genome of *Aureococcus* was sequenced by the DOE JGI from an isolated sample collected from the shores of Long Island, NY, one of the areas most affected by the microalga when it first appeared 25 years ago on the east coast of the United States. By sequencing its genome, scientists hope to learn more about *Aureococcus*' ability to capture CO_2, survive in varying marine environments, and outgrow many of its competitors.

Aerial view of Great South Bay, NY during a brown tide bloom in June 2008. Billions of *A. anophagefferens* cells per liter crowded into the coastline and turned the water brown. (Suffolk County Department of Health Services)

"In the decade since publishing the draft of the human genome, JGI has pioneered the exploration of marine algal genomics with sequences of the first diatoms, *Ostreococcus* and cyanobacteria," said Grigoriev. "Compared to these phytoplankton, which can inhabit the same estuaries, *Aureococcus*, which outcompetes them, shows genome-encoded advantages to benefit from alternative nutrients, survive under variable light conditions, and encode the largest number of selenoproteins (which use the trace element selenium to perform essential cell functions) known to date."

Gobler elaborated on how *Aureococcus* can outcompete the other phytoplankton in a coastal estuary, noting that "the surprise was the concordance between the genome and the ecosystem where it's blooming."

Don Anderson, a senior scientist at Woods Hole Oceanographic Institution who has studied harmful algal blooms for decades and is a tireless promoter of research efforts in this field, said that the *Aureococcus* genome is "a great advance, a great resource for our community. For decades, scientists have been trying to understand why this species blooms, when it blooms, how it is able to dominate when there are so many other competing species in the water with it. With this new genomic data we have a new approach. We're getting answers based on the genes, though we still need other approaches that collect relevant oceanographic and chemical data to go along with the inferences drawn from the presence and absence of genes," he said. "As we learn about *Aureococcus* with this approach, that knowledge will help us make similar advances with other HAB species."

Grigoriev also noted that the multidisciplinary approach of combining genome sequencing with other techniques allows researchers to explore a new area of ecogenomics, which is closely connected to the DOE mission in biogeochemistry.

"Aligning the physico-chemical parameters of an ecosystem with the genomic potential of its inhabitants enables us to monitor changes in the environment. Harmful algal blooms are a recently reported phenomenon and could be connected to the growing human population along coastlines, which created conditions for *Aureococcus* to thrive, in turn adversely affecting estuaries. On the other hand, massive algal blooms can reduce carbon dioxide in the atmosphere. So by employing the ecogenomics approach we can start building balanced models of targeted environments."

Watch a slideshow narrated by Chris Gobler regarding the significance of the *Aureococcus* genome sequence on the DOE JGI's SciVee channel at http://www.scivee.tv/node/27510.

the PRIMER / 7
Spring 2011 Volume 8 Issue 2

Sentinel of Change: Water Flea Genome Improve Environmental Monitoring Capabilities

The water flea, *Daphnia pulex*, is roughly the size of the equal sign on a keyboard but it has enormous influence in environmental monitoring studies. For decades, researchers have used the tiny crustacean to develop and monitor environmental regulations.

"*Daphnia* is one of the most widely used model systems for environmental protection agencies around the world," said project leader and Indiana University Center for Genomics and Bioinformatics genomics director John Colbourne. "The costly challenge of evaluating conditions in the environment and of our water supplies may be overcome by *Daphnia's* potential use as a high-tech and modern version of the mineshaft canary."

When the project was begun, environmental agencies and toxicology researchers were looking into aquatic model systems that acted as sentinel species in order to diagnose the presence of problematic chemicals in fresh water and extrapolate their effects. In the February 4, 2011 issue of *Science*, DOE JGI researchers and the *Daphnia* Genomics Consortium marked nearly a decade of collaboration with the release of the water flea's 200-million base genome sequence, containing the most genes of any animal characterized to date.

At the recently-concluded Genomics of Energy and Environment Meeting, DOE JGI Deputy Director Jim Bristow charted the growth of the Consortium along with the progress on the genome sequence as he spoke of the DOE JGI's ability to build communities around sequencing projects.

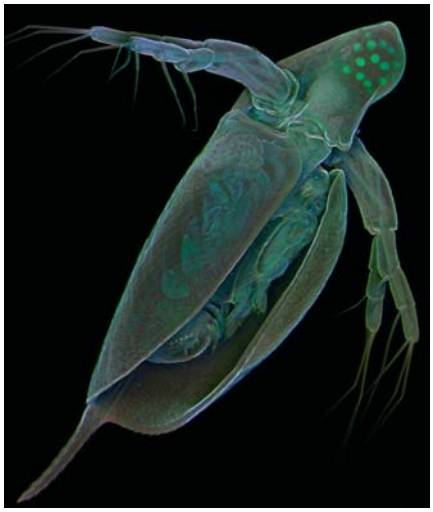
"When the *Daphnia* project was approved in 2003, the Consortium consisted of 26 researchers and the literature consisted of 50 papers," he said. "When the genome sequence was released in 2007, there were 150 researchers working on *Daphnia* and 18 publications. When the annotated genome was published in *Science* this year, it was one of 50 papers published simultaneously on the water flea by the now 475-strong *Daphnia* Consortium."

The genomic data from this keystone freshwater species could be used to help researchers develop and conduct real-time monitoring systems of the effects of environmental remediation efforts, said Colbourne. Study coauthor Michael Pfrender of the University of Notre Dame noted another application of the *Daphnia* genome with regard to developing commercial biofuels. "When you grow algae in large open-air tanks to select for biofuel production, they're invaded by *Daphnia* that graze down the algae," he said. "You're faced with either learning how to control *Daphnia* or learning how to use it to harvest the hydrocarbons."

The *Daphnia* genome was sequenced and annotated at the DOE JGI. Jeffrey Boore, who led the portion of this project done at the DOE JGI and is now CEO of Genome Project Solutions, which also made contributions to this study, noted that the water flea genome will help researchers conducting comparative genomic studies involving insects and crustaceans.

"Crustaceans are the closest living relatives to insects, and *Daphnia pulex* is the first animal from this group to have its genome completely sequenced," Boore said.

Igor Grigoriev, head of the DOE JGI Eukaryotic Genomics group, noted that the *Daphnia* genome, which was sequenced using the Sanger method, is the most compact with over 31,000 genes, a third of which are of unknown function. He attributed the large number of genes to expansion by tandem duplication, adding that the number is a conservative estimate. "There are more genes in *Daphnia* than there are in the human genome, more than any animal," he said.



***Daphnia pulex,* commonly called water flea (Dr. Jan Michels, Christian-Albrechts-University, Kiel)**

Annotating the *Daphnia* genome also afforded researchers an alternate way of determining gene functions. "This goes beyond having the first crustacean genome," said Grigoriev. "By combining functional genomics assays we can see that the genes of unknown function – the new genes resulting from adaptations, not the ones conserved because they're key for housekeeping – are the most responsive to environmental changes. In that sense, the organism's environment can direct us to areas of interest; if you pick a genome and would like to annotate all the genes, go to the environment."

Colbourne added that the ability to use *Daphnia* for environmental studies focused on identifying associations between gene function and disease has also contributed to its recent inclusion as a model system for biomedical research by the National Institutes of Health, joining established models like the fruit fly *Drosophila* and the worm *C. elegans*. He added that the *Daphnia* Genomics Consortium is working on improving techniques to do large-scale genome population assays to better track and manage remediation applications.

Spring 2011 Volume 8 Issue 2

Shipping out CAZymes

BY JYOTI MADHUSOONAN

"Interest in cellulosic ethanol as an alternative fuel is perhaps not very surprising, since fermentation is one of the oldest technologies known to man, and probably the best loved," said Dan Distel, executive director of the Ocean Genome Legacy Foundation, at the DOE JGI User Meeting. Cellulosic ethanol is part of broader set of strategies geared to wean the U.S. off fossil fuels, and Distel collaborates with the DOE JGI to study shipworms, marine mollusks so specialized they're known as the 'termites of the sea,' which explained why he titled his talk "How to eat a wooden ship."

Despite the notorious nickname, shipworms and their microbial symbionts could hold one of the keys to a biofuel revolution in their dietary preferences. The microbes in shipworms may prove to be an unusually rich source of carbohydrate-active enzymes (CAZymes) suitable for the industrial production of cellulosic ethanol, a biofuel the National Renewable Energy Laboratory estimates could produce just one-tenth the carbon emissions of petroleum.

Distel presented a metagenomic view of the shipworm symbionts at the User Meeting, along with comparisons of their CAZyme activity to that of ruminants and termites. Markedly different from terrestrial cellulose consumers, the shipworm has evolved unique anatomical modifications and an unusual microbial community to facilitate wood fermentation. Though the absence of microbes in the shipworm digestive tract (called the caecum—pronounced "SEE-cum") is conspicuous, the mollusks instead harbor dense populations of intracellular bacteria in a specialized organ near their gills, the Gland of Deshayes.

A single Gram-negative species, *Teredinibacter turnerae*, was long believed to be the sole symbiont in the shipworm. However, sequencing the symbiont metagenome of the giant Pacific shipworm *Bankia setacea* revealed not just one species but several different ribotypes (distinguished by 16S ribosomal RNA fingerprinting) of a microbial consortium closely related to one another, yet phylogenetically distinct from microbes in termites and ruminants.

The shipworm microbiome consists of about 10 different intracellular symbionts types rich in CAZyme-encoding genes, which account for about 2.5 percent of all their protein-coding genes. "When compared to the bacterial champions of CAZyme production that live in cows and termites, shipworm isolates still stand out," said Distel. Genomic comparisons of the shipworm symbionts reveal that the smaller number of species is amply compensated for by the higher proportions of specialized enzymes they produce.

Nearly all animals that consume cellulose rely on microbial fermentation processes to digest it. Termite hindguts and the cow rumen (see page 1) harbor dense and diverse microbial populations that secrete several thousand kinds of enzymes to degrade cellulose.

The enzymes produced by ruminant microbes outnumber those made by shipworms, but the latter compensate with a significantly higher proportion of cellulose and hemicellulose depolymerases. Distel noted that although the shipworm metagenome only has about 500 CAZyme genes, half of them are active against cellulose and hemicellulose, as opposed to 20 percent of the enzymes produced in the cow rumen that have activity against pretreated biomass.

Using liquid chromatography and mass spectrometry to identify which of these CAZyme proteins were physiologically relevant and functional, Distel's group found that shipworms specifically transport a cocktail of CAZymes from their gill-sac to the caecum to process lignocellulose. The presence of active gill-encoded CAZymes in the caecum and the absence of lignocellulose active enzymes in the gills together highlight the selective transport of a specific cocktail of CAZymes from the gills to the digestive tract, a process unknown in any other species.

L. pedicellatus **(Dr. Ruth Turner, Harvard University (deceased))**

The list of differences between shipworms and other cellulose-digesters highlights not only the uniqueness of the mollusk, but also the potential it holds for industrialization. As a model system, shipworms are simple, uniquely rich in secreted CAZymes, and highly amenable to genomic and proteomic analysis. Their mechanisms of lignocellulose degradation place shipworms and their intracellular symbionts high among the most promising sources for systems that could yield cellulosic ethanol. As Distel said, "They're an awesome place to look for enzyme mixes that would be useful for industrial processes."

Jyoti Madhusoodanan is a Bay Area-based freelance science writer with a Ph.D in microbiology.

the PRIMER / 9
Spring 2011 Volume 8 Issue 2

Grand Challenge Spotlight: The Great Prairie Project

When DOE JGI Director Eddy Rubin speaks of scaling up sequencing projects lately, one of the examples he uses is a pilot study launched this past year as a DOE Grand Challenge endeavor.

Speaking at the DOE JGI User Meeting, Jim Tiedje, Director of the Center for Microbial Ecology at Michigan State University, and one of the project leads along with Janet Jansson of Lawrence Berkeley National Laboratory, described the scale and complexity of the project to sequence the prairies of the Midwestern United States.

"One of the largest ecosystems in the United States, the Midwest prairie represents the largest expanse of the world's most fertile soils," said Tiedje. "It parallels the ocean gyre as the most important ecosystem for primary productivity and biogeochemical cycling."

As prairie soils account for nearly 30 percent of the continental United States' land surface and as much as 40 percent of soil organic carbon stocks, they are an important component of the carbon cycle. The project involves comparing the soil metagenomes from prairie land that has been converted and cultivated as farmland for several decades with the soil metagenomes of untouched prairie lands.

Among the questions Tiedje and his colleagues hope to answer through this work are how microbial diversity has been impacted by land use, and what influences these conditions might have had on genetic diversity and the expression of ecofunctional genes. Another question to be considered is how these lands might impact ecosystem and genomic diversity.

To better understand this ecosystem, Tiedje and his colleagues took soil samples from cultivated and uncultivated prairie in Wisconsin, Iowa and Kansas, and then had the metagenome sequenced at the DOE JGI, yielding more than a terabase of data.

University of Wisconsin's Randy Jackson sampling Wisconsin native prairie (Jim Tiedje, Michigan State University)

Tiedje credited Michigan State University colleague C. Titus Brown, who handled the computational aspect of the project, using data partitioning and data reduction methods to handle the large sequence dataset. Working with 219 gigabases (Gb) of Iowa corn field soil metagenome sequence, Brown conducted abundance filtering, using exact data reduction techniques to eliminate unconnected reads and then prefiltered for assembly, without removing low-abundance reads, which might have useful information.

Tiedje noted that the more than 200 Gb of sequence had an estimated 0.73X coverage; to reach the cow rumen depth of sampling, he said, referencing the DOE JGI's project which appeared in Science earlier this year, one to two terabases of sequence would have been generated.

Metagenomic assembly was done using reference genomes sequenced by the DOE JGI, and preliminary annotation was done using the CAMERA pipeline. Working on the question of ecofunctional genes, Tiedje and his team looked for groups such as members of the *nif* family, which are involved in nitrogen fixation. Among their preliminary findings was noting 109 taxonomically diverse genomes containing the gene *nifH*. The work continues, but among the obstacles the team is still trying to overcome are soil diversity, non-uniform coverage and the lack of reference genomes.

Tiedje ended his talk with a look at another large-scale metagenomics project he's involved in: studying land-use change in the Amazon forest. The collaboration between American and Brazilian researchers also includes University of Texas at Arlington's Jorge Rodrigues, who has a DOE JGI Community Sequencing Program project involving microbes in the termite hindgut.

"It's the world's largest carbon dioxide-sequestering terrestrial system, it holds a fifth of the world's freshwater system, it controls flux of atmospheric gases at a global scale," Tiedje ticked off the reasons to study this region. "Deforestation in the Amazon releases 1.6 petagrams — that's 10 to the 15th power — of carbon annually."

Rumen-ating on Enzymes

continued from page 1

the potential of deep sequencing of a complex community to accurately reveal genes of interest at a massive scale, and to generate draft genomes of uncultured novel organisms involved in biomass deconstruction," wrote the first author and former DOE JGI postdoctoral fellow Matthias Hess, now an assistant professor at Washington State University Tri-Cities in Richland, Wash. "The general approach presented here will be applicable to other environmental microbial communities."

Speaking at the 6th Annual Genomics of Energy & Environment Meeting, study co-author Zhong Wang noted that the cellulases used to break down plant mass in industrial biofuel processes are currently predominantly sourced from a single fungal species and more diversity is needed.

"Industry is seeking better ways to break down biomass to use as the starting material for a new generation of renewable biofuels," said DOE JGI Director and project lead Eddy Rubin. "Together with our collaborators, we are examining the molecular machinery used by microbes in the cow to break down plant material."

Hess worked with colleagues at the University of Illinois in using the fistulated cow model that allows direct access through a tube into the foregut. Along with Rubin's other postdoctoral fellow Alex Sczyrba, Hess used the candidate bioenergy feedstock switchgrass *(Panicum virgatum)* to determine which microbes in the cow rumen were involved in digesting the plant mass. The switchgrass samples were placed in nylon bags and then inserted into the cow rumen, where they were left to be digested for 72 hours. When the bags were removed after three days, the DNA from the microbes that were adherent to the switchgrass were isolated and then sequenced.

"Microbes have evolved over millions of years to efficiently degrade recalcitrant biomass," said Rubin. "Communities of these organisms can be found in diverse ecosystems, such as in the rumen of cows, the guts of termites, in compost piles, as well as covering the forest floor. Microbes have solved this challenge, overcoming the plant's protective armor to secure nutrients, the rich energy source that enables them and the cow to thrive."

While the sequence produced during the study was enormous — nearly 100 times greater than the sequence of a single human genome — Hess said that generating the data was not the most challenging part of this project. "The real challenge was to analyze the vast amount of data for which no reference genome was available and to identify and produce full-size functional enzymes based solely on information obtained from billions and billions of short snippets of DNA sequences," he said.

To analyze the information, the researchers developed a genome assembly strategy that could handle the vast amount of data while making sure to avoid misassemblies that would have led to chimeras — artificial genes not present in the microbial community.

Through employing different filters, Sczyrba whittled down the number of the more than two million predicted genes to 27,755 candidate genes that encoded a specific category of enzymes called carbohydrate-active enzymes (CAZymes) that can break down plant polysaccharides (e.g. cellulose) into small sugars. Hess then identified the most promising candidates, tested a subset of 90 candidate genes for functionality and found that more than 50 percent of the tested candidates had cellulose-degrading activity, with almost 20 percent able to break down the "real-world" biofuel crop switchgrass. The finding indicated that a significant fraction of the 30,000 genes identified are indeed active against plant material and would be a treasure trove of novel enzymes for biofuel researchers.

The fistulated cow system allows direct access into the animal's foregut, enabling researchers to incubate biomass-containing nylon bags to isolate rumen microbes associated with a defined plant substrate to identify genes and genomes participating in biomass deconstruction. (Jonas Løvaas Gjerstad)

Hess said that the discovery of these novel enzymes from this one study significantly increased the number of enzymes believed to act on carbohydrates by nearly a third compared to numerous previous studies carried out over decades.

Besides the identification of genes encoding enzymes that might play a major role in future processes for the industrial production of lignocellulosic biofuel, Hess and Sczyrba wanted to assemble the entire genomes of organisms involved in biomass breakdown from the rumen. Using various computational puzzle-solving approaches, they were able to build 15 genomes for the rumen microbes, none of which matched anything that had previously been described.

To confirm their computational results, the team then turned to DOE JGI Microbial Genome Program *continued on page 12*

the PRIMER / 11
Spring 2011 Volume 8 Issue 2

Watching the ocean's watchmen

BY JYOTI MADHUSOODANAN

Marine microbial communities account for most of the ocean's biomass and metabolic activity and sequester as much carbon as land sources. In addition, their genomic and transcriptomic responses to miniscule environmental changes prove that they might be the best sentinels of climate change as well. Three speakers at the DOE JGI User Meeting—Peer Bork, Mary Ann Moran and Chris Scholin—talked about different projects that assess the interactions of the ocean microbiome with physicochemical changes in their surroundings.

Sailing the world's oceans to sample the genomic blueprints of microbial communities from 30 different sites, DOE JGI user Peer Bork from the European Molecular Biology Laboratory combines data from physical parameters like temperature and nutritional factors with diverse genomic information to address the question, "Can we identify genes that capture complex environmental properties like biomass production?"

The group's objective is to identify biomarkers that reveal community metabolic adaptations to environmental gradients. Their findings show that climatic influences have much greater impact than nutrient factors on community functional genetic composition.

"We find that the gene repertoire in terms of diversity and richness correlates with primary production in a community," said Bork. As part of the Tara Oceans Expedition, the group plans to extend their findings to encompass about a hundred sample points and correlate genomic data to oceanographic, imaging and chemical information.

Mary Ann Moran, from the University of Georgia, addressed a similar question from a different angle- the community microbial transcriptome response to fluxes in environmental chemicals. Monitoring microbial community responses to sulfur aerosols in the Sargasso Sea, Moran's metatranscriptomic approach connects the current deluge of -omics data to the ecogenomic implications of environmental change, allowing them to address ecological community models rather than individual species.

As Moran said, "We don't need to sequence every single transcript of RNA from every species to get a sense of how important it is."

Starting with a pool of community transcripts from seawater samples, her group used 454 and Illumina sequencing to identify several bacterioplankton transcripts involved in the transmembrane transport of organic compounds. In model experiments, they used microarray analysis to compare and validate the responses of phytoplankton and marsh vascular plants to increased organic carbon content in their environment. Extending these models into a metatranscriptomic analysis of seawater microflora, Moran's research identified the metabolic pathways that microbial communities use to degrade DMSP, an organic compound that is the single most important component of atmospheric sulfur aerosols.

Finally, Chris Scholin, president of the Monterey Bay Aquarium Research Institute (MBARI), uses ESP to 'take the pulse of the ocean." Not quite as far-fetched as extra-sensory perception, the Environmental Sample Processor (ESP) remotely senses marine microbial dynamics in response to atmospheric wind forcing, methane spurts and more.

"ESP can be run in all kinds of bizarre places where we can't be for extended periods, like being deployed for months at deep sea hydrothermal vents," said Scholin. The applications of ecogenomic sensors such as ESP are immense, especially to marine metagenomic researchers. Moored at sea, the ecogenomic sensor is capable of sampling seawater, processing and archiving samples of DNA, RNA and microbial metabolites for extended periods of time. Two-way transmission allows scientists to continuously monitor incoming data and request additional samples or processing when necessary. Scholin cited a case when ESP was deployed to monitor harmful algal blooms off the Southern California coast and he could monitor the data even as he himself was vacationing in Death Valley.

Though he isn't a DOE JGI User (yet), Scholin's ESP tool could provide vital information in terms of monitoring water quality and pollutant levels, sampling bacterial communities in remote ocean locations and assessing microbial metabolic activity, such as the sulfur aerosol pathways that Moran's research addresses.

Drifting Environmental Sample Processor off Central California coast (Phil Sammet © 2010 MBARI)

Rumen-ating on Enzymes

continued from page 10

head Tanja Woyke for help. Using single cell genomics, they were able to examine the microbial genomes without requiring that they be cultured, a useful tool as only about one percent of the planet's microbial species can be readily grown in the laboratory. They were able to isolate a single rumen microbe using a cell sorter, and without having it grow, they generated the genome of this uncultured microbe using single cell sequencing technology.

When the DNA sequences derived from the single genome were mapped to the 15 computationally assembled genomes, the researchers found that more than 98 percent of the data matched to one single genome that had been assembled *in silico.*

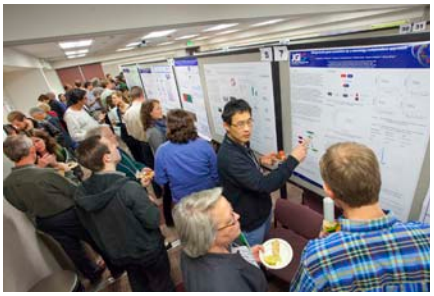
"The single cell data made us confident that what we saw was real," Hess said. "Otherwise we'd have computational data only, which would have made our work much, much less convincing."

6th Annual Sequencing, Finishing and Analysis in the Future Meeting

Santa Fe, NM | June 1-3, 2011

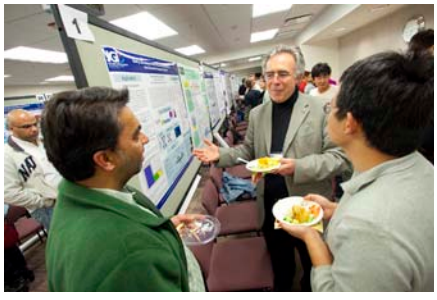
The SFAF meeting focuses on laboratory methods and computational tools, including new sequencing technologies used to help sequence, assemble and finish genomes. For more information, contact Chris Detter at cdetter@lanl.gov or go to http://www.lanl.gov/finishinginthefuture.

SLAC National Laboratory Director Persis Drell with brother and DOE Office of Biological and Environmental Research program manager Dan Drell.

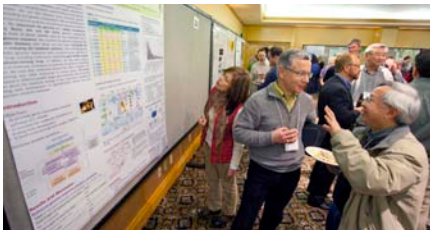


DOE JGI's Changbin Du discusses his poster with ORNL's Miriam Land, who works on the annotations for genomes sequenced at the DOE JGI.

DOE JGI's Matt Zane shows off the Illumina sequencers to User Meeting attendees.



Bob Cottingham of Oak Ridge National Laboratory in converation with DOE JGI's Karan Bhatia (left) and Zhong Wang (right).



Francis Martin of INRA at a poster session for the DOE JGI User Meeting.

Contact The Primer
David Gilbert, Editor / DEGilbert@lbl.gov