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

the PRIMER
Spring 2013 Volume 10 Issue 2

Highlighting functionality in genomic research

DOE JGI Metagenome Program head Susannah Tringe at the 8th Annual Genomics of Energy & Environment Meeting. (David Gilbert, DOE JGI)

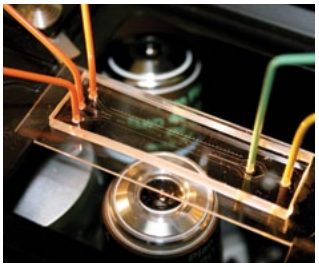
One of the recurring themes at the 8th Annual Genomics of Energy & Environment Meeting held March 26-28, 2013 was the application of the technologies to enable groundbreaking science. Many of the talks from the DOE Joint Genome Institute's annual meeting at the Walnut Creek Marriott centered on analyses and functional studies being carried out on publicly available genomes — primarily sequenced by the DOE JGI — with demonstrable applications in the fields of energy and environment research.

Chris Voigt from MIT opened the meeting with a look at how the DOE JGI might play a key role in the field of synthetic biology, from the view of a prospective user and co-director of a synthetic biology program (see page 2 for more details). The opportunities for deploying synthetic biology were *(continued on page 4)*

also in this issue

Expanding capabilities with new partnerships

In line with the vision outlined in its strategic plan, "Forging the Future — A Ten-Year Strategic Vision" (http://bit.ly/JGI-Vision) the DOE JGI has positioned itself to provide the most current technology and expertise to their users in order to address pressing energy and environmental scientific challenges. An important early step in this process is the launch of the Emerging Technologies Opportunity Program (ETOP) (http://www.jgi.doe.gov/programs/ETOP/).



Microfluidic device involved in high-throughput sorting of microbial cells. (Roman Stocker)

The ETOP's primary purpose is 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. The program was one of several recommendations that emerged from the DOE JGI's strategic planning as well as a complementary process carried out by DOE's Office of Biological and Environmental Research (http://bit.ly/JGI_GEspw). Now, a new set of partnerships is taking shape in response to the ETOP's first call for proposals. These span the development of new scalable DNA synthesis technologies to the latest approaches to high throughput sequencing and characterization of single microbial cells from complex environmental samples.

"A core philosophy of the DOE JGI is that our suite of technical and analytical capabilities needs to evolve continuously so that the scientific achievements of our users can be maximized," said Jim Bristow, who oversees the ETOP as DOE JGI's Deputy Director of Science Programs. "The DOE JGI strives to integrate these expanded activities in innovative and effective ways. This is critical if the biological sciences are to realize the full benefits and promise of genome sequencing. While state-of-the art massive-scale sequencing remains a critical *(continued on page 7*

Systematically scaling up DNA synthesis

In his keynote address opening the 8th Annual Genomics of Energy & Environment Meeting, Chris Voigt, co-director of the Synthetic Biology Center at the Massachusetts Institute of Technology focused on two big ideas: the intersection of sequencing, synthesis and design; and optimizing the cycle of designing, debugging and correcting.

Much of his talk focused on the work his lab had done involving synthesizing a 25Kbp DNA sequence with more than a dozen genes involved in nitrogen fixation, a key process in agriculture because it ensures that plants get the needed nutrients to grow. One goal of biotechnology researchers is to be able to take the organism's ability to fix nitrogen and introduce it directly into a plant or another organism that associates directly with the plant.

Chris Voigt (Roy Kaltschmidt, LBNL)

"Sequencing is not enough; synthesis is not enough. You need design on top of both functions."

The Voigt lab focused on the model bacteria *Klebsiella oxytoca*, which contains a cluster of 16 genes that need to be moved. The gene cluster project took six years of his lab's attention as they worked from the ground up.

Voigt walked the audience through the basics of synthetic biology terms and how a simple project might not require much in the way of design if the goal is to simply find a single element that provides the needed level of activity. Scaling up, however, from working on single genes toward multi-part functions requires a combination of design, sequencing, and synthesis.

"You need sequencing and informatics to identify the element and figure out what the function might be, but that's not enough," he said. "You need to be able to synthesize the DNA, but that's not enough because you're not putting it back in the organism from which it came but into another organism. You need synthetic biology and design."

Applying these ideas to the nitrogen fixation project, Voigt said that when the project began in 2004, they had to do much of the work from scratch. A few years into the project, they found out that they had been basing the synthesis on an error-riddled DNA sequence in the database. After the team resequenced the DNA, they found the errors and corrected them, then re-synthesized the gene cluster.

Another part of the process involved "refactoring," a term borrowed from the software industry and used to refer to rewriting the underlying code while providing the same desired functionality. For synthetic biologists, Voigt said, the term referred to rewriting the DNA sequence in a way the researchers understood while the synthesized sequence produced the same desired functionality.

The work paid off in the production of a version that had 7.3 percent of the activity levels seen in the wild type. After the six-year effort resulted in this proof of principle, Voigt looked for ways to help his team accelerate the optimization process and shorten the cycle of designing, building and testing. He says the team is still working on the nitrogen fixation gene cluster, with two genes left to optimize.

One of the solutions has led to a collaboration with Boston University researchers who developed a programming language named Eugene. Described as a constrained language for synthetic biology, the system allowed researchers to clearly specify design rules in a way that humans could easily read and understand.

"One of the first things we realized would be a problem was just the very simple aspects of design," he said. "If we're going to scale up we need to take a fundamentally different approach to design."

The full video of Voigt's talk is available online at http://bit.ly/JGI13_Voigt.

A voyage of science discovery and outreach

Eric Karsenti (David Gilbert, DOE JGI)

"There is a problem with science: we enjoy it a lot, we do it in the labs, but we don't communicate it to the public."

Many years ago, Eric Karsenti of the European Molecular Biology Laboratory in Heidelberg, Germany was inspired by Charles Darwin's account of his voyages on the *HMS Beagle*. A sailor himself, he thought that sharing such accounts was a good way of talking about science to the general public.

"There is a problem with science," he told the audience attending his closing keynote speech at the DOE JGI's 8th Annual Genomics of Energy & Environment Meeting. "We enjoy it a lot, we do it in the labs, but we don't communicate it to the public." Five years ago, after discussing the matter over with friends, he came up with the idea of leading a scientific expedition related to ocean studies that would also double as an outreach mission.

In September 2009, after 18 months of assembling funds for the trip, a team of researchers and sailors and finding a boat, the TARA Oceans Expedition left Lorient, France aboard a 36-meter schooner on a three-year journey. The stated scientific goal was to learn more about the plankton gene network by sampling the major oceans in order to characterize the environments.

To make the project more useful to a wider research community, Karsenti and his colleagues planned to use the samples collected as a way to provide a pipeline of data for multiple ocean models. They also looked for a correlation between the global ocean circulation and the geographic distribution of ecosystems.

The scientific consortium Karsenti assembled was comprised primarily of researchers studying plankton and corals, and a team of data handlers — sequencing data, environmental data, imaging data and a group working on bioinformatics and modeling. Marine plankton range in size from miniscule like viruses to larger organisms like zooplankton. Over the course of the voyage, which took 2.5 years and spanned 60,000 miles and 153 research stations, the team collected 27,000 biological samples ranging in size from less than one micron to more than 1,000 microns in waters ranging from photic zones, mesopelagic zones and oxygen minimum zones. To collect samples, they had tools ranging from nets to pumps and an instrument that allows water samples to be collected at many different depths throughout the water column.

Even though the trip is over, and the analysis of the environmental data is almost done, he said, the imaging data is still a work in progress. "My whole career I carefully avoided genomics and large-scale projects," Karsenti commented during his keynote as he talked about the status of the sequencing being done with the help of Genoscope in France. They are working with both DNA and RNA, and looking at single amplified genomes, as well, he said. Eventually, he hopes to have all three datasets of sequence from the Atlantic, Indian and Southern Oceans available to the research community at large.

Aside from the scientific data collected, Karsenti said the voyage also successfully met the second half of its stated objective to bring science to the general public. At every port, people boarded the schooner to talk to the scientists and learn about the work being done.

"We saw a lot of kids and each time the boat stopped in a different town, we had a lot of kids come and we explained to them what we're doing," Karsenti recalled. "Not only the scientists but even the crew, the professional sailors understood what we were doing — we taught them — and then they were talking to the kids. I'm surprised at how kids get interested in this; they asked a lot of questions. This was a fabulous human experience."

According to the expedition statistics, 5,000 children boarded the boat over the course of the voyage. The full video of Karsenti's talk is available online at http://bit.ly/JGI13_Karsenti.

Highlighting functionality *(continued from page 1)*

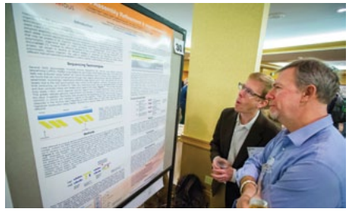
revisited, first by Adam Guss from Oak Ridge National Laboratory (ORNL), who talked about how the bacterium *Clostridium thermocellum* could be engineered to produce biofuels from cellulosic substrates. Later on, Jane Lau from the Joint BioEnergy Institute talked about the focus of JBEI's Feedstocks Division on xylan, a component of hemicellulose, to build better biofuel feedstocks.

Talks spanned from handling microscope volumes to amplifying the impact of research information in the public databases. Paul Blainey of Broad Institute discussed a microfluidics technique to shrink reaction volumes as a way to reduce contamination in single-cell genomics. David Weston of ORNL spoke about applying plant genomics toward developing more accurate climate models with the help of the DOE systems biology project KBase. The goal, he added, is to use modern molecular biology techniques to predict ecosystem responses. Jose Pruneda-Paz of UC San Diego on genetic studies being done to learn more about the transcription factors in the model plant *Arabidopsis* that are being regulated by the circadian clock.

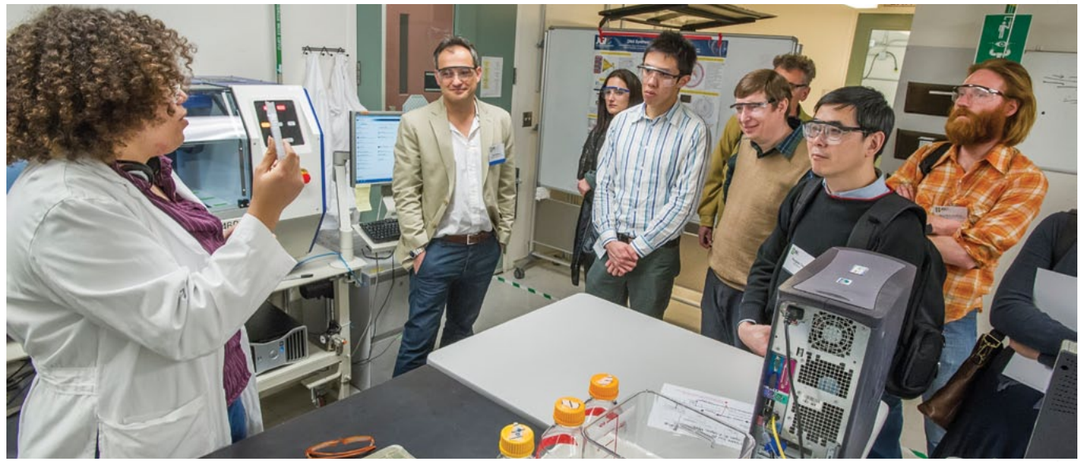
Another recurring theme during the meeting was the emphasis on plant model system and biomass feedstock studies. Zander Myburg from the University of Pretoria gave an update on how the eucalyptus genome sequence is helping researchers study the genomics of wood formation. Sam Hazen University of Massachusetts compared the fine points of determining cookie quality (from a previous job) to the rules applied toward feedstock quality. Rick Amasino from the University of Wisconsin talked about a project funded by the Great Lakes Bioenergy Research Center that focused on the genes and conditions that affected flowering in plants such as the model grass *Brachypodium distachyon*. His talk was followed by Sean Gordon of the U.S. Department of Agriculture's Agriculture Research Service. Gordon also talked about *Brachypodium* but focused on its tolerance to abiotic stress, particularly during what he called "the large field trial known as climate change."

Greg Bell, Director of the Energy Sciences Network (ESnet), a high-performance, unclassified national network built to support scientific research talked about a science demilitarized zone (DMZ) and how research campuses could optimize their networks to share data and not constrain discoveries within geographic barriers. Several speakers talked about projects that relied on collaborations from around the world. Tanja Woyke, the DOE JGI Microbial Program head, talked about the "skewed" view of the microbial world based on the inability to cultivate most microorganisms in the lab, and a project underway to tap the vast microbial "dark matter" to fill in the uncharted branches on the tree of life. Wayne Reeve from Murdoch University in Australia gave an update on an ongoing project to sequence 100 root-nodulating bacteria from around the world as part of the GEBA-RNB project being done in collaboration with the DOE JGI, and reminded attendees of the importance of nitrogen fixation for crops.

Aside from functionality, collaborative efforts in understanding plant-microbe interactions were covered. Chris Schadt of ORNL gave an update on the poplar project and discussed plans to study the root microbiome to better understand the interactions between the microbes found in and around poplar roots. Peter Larsen of Argonne National Laboratory talked about molecular interactions within *in vitro* microbial plant communities. Sarah Lebeis from the University of North Carolina touched on a collaboration between the Dangl lab and the DOE JGI to study the *Arabidopsis* root microbiome (research featured on the August 2, 2012 cover of the journal *Nature*) and the ongoing project to study the plant's immune system.



One session was devoted to research being conducted on microbial communities in a wide range of places. Eric Allen of UC San Diego discussed his team's long-term study being conducted at Lake Tyrrell in Australia to reconstruct the genomes for all species in the microbial community.



Poster session and tours. (Roy Kaltschmidt, LBNL)

Eoin Brodie of Berkeley Lab discussed work that is part of the DOE JGI's Community Sequencing Program to develop functional genomics studies involving microbial communities in the Mediterranean grassland in Loma Ridge, Calif. Laura Hug from UC Berkeley talked about microbial communities in aquifer sediment at a former uranium tailing mining site in Colorado that is now part of the DOE's Integrated Field-Scale Subsurface Research Challenge.

Several DOE JGI researchers gave updates on their respective programs. Noting that sequencing is becoming less of a bottleneck, Deputy Director of Genomic Technologies Len Pennacchio talked about other large-scale capabilities, particularly those that will help link sequence to functional information and can be provided to its user community. Sam Deutsch gave an update on the state of the DNA synthesis pipeline. DOE JGI Plant Program head Jeremy Schmutz of the HudsonAlpha Institute for Biotechnology talked about the future of the plant program, focusing on new approaches toward sequencing and assembling *de novo* plant genomes.

Looking beyond lab results, two speakers touched upon industrial-level research and applications in the energy field. Bob Schmidt of SG Biofuels discussed the challenge of making jatropha a viable biofuels feedstock by determining the plant's genome sequence. Jack Newman of Amyris talked about the company's focus on using synthetic biology to make microbes produce the antimalarial drug artemisenin, and how the techniques were adapted to make other biochemical products such as Biofene®, the company's brand of renewable farnesene.

Eldredge Bermingham from the Smithsonian Tropical Research Institute veered away from the emphasis on plants and bioenergy with a presentation on about Panama as an ideal site for what he termed "the genomics of resilience to any form of human perturbation," studying climate change over decades as it impacted coral formations and weather conditions. Other researchers focused on applications of genomics to fields other than energy and environment. Joe DeRisi from UC San Francisco told the story of how he and his lab ended up studying inclusion body disease, a virus killing boa constrictors. Tom Gilbert from the Natural History Museum of Denmark talked about how techniques to study microbes and microbial communities in soil and water could be applied to help conservationists monitor biodiversity, particularly for endangered species.

Eric Karsenti from the EMBL Heidelberg closed the annual meeting with a look back at the three-year TARA Oceans expedition he mounted (see page 3 for details) with a two-fold goal of studying the plankton ecosystem in the oceans while also serving as a floating science education and outreach center to the general public.

Videos from the meeting have been posted online on the DOE JGI's YouTube channel at http://bit.ly/JGI13_UM8 and on the SciVee channel at http://www.scivee.tv/node/58289. Images of the speakers are on the DOE JGI's Flickr page at http://www.flickr.com/photos/doe_jgi/.

"Take the tools and developments on the genomics side of DOE and apply them to environmental concerns, particularly climate."
– David Weston, ORNL on KBase project (http://bit.ly/JGI13_Weston)

Peachy insights into breeding biofuels crops

Several plants sequenced by the DOE Joint Genome Institute have been considered "flagship genomes" due to their importance to DOE mission and plant science. Among these plants are poplar, the first tree sequenced and a candidate bioenergy feedstock, and soybean, the primary source of biodiesel in the United States. Other plant genomes are important for their role as a small reference model for other plants; for example, *Brachypodium distachyon* is a reference grass related to switchgrass and the moss *Physcomitrella* is a comparator for land plants.

Building on the value of comparator genomes, the relationship between a peach and a poplar may not be obvious at first glance, but to botanists both trees are part of the rosid superfamily, which includes not only fruit crops like apples, strawberries, cherries, and almonds, but many other plants as well, including rose that gives the superfamily its name.

"The close relationship between peach and poplar trees is evident from their DNA sequence," said Jeremy Schmutz, head of the DOE JGI Plant Program and a faculty investigator at the HudsonAlpha Institute for Biotechnology. In the March 24, 2013 issue of *Nature Genetics*, Schmutz and several colleagues were part of the International Peach Genome Initiative (IPGI) that published the 265-million base genome of the Lovell variety of *Prunus persica*.

The publication comes three years after the International Peach Genome Consortium publicly released the draft assembly of the annotated peach genome on the DOE JGI Plant portal Phytozome.net and on other websites.

The team compared 141 peach gene families to those of six other fully sequenced diverse plant species to unravel unique metabolic pathways, for instance, those that lead to lignin biosynthesis — the molecular "glue" that holds the plant cells together — and a key barrier to deconstructing biomass into fuels.

Rapidly growing trees like poplars and willows are candidate "biofuel crops" from which it is expected that cellulosic ethanol and higher energy content fuels can be efficiently derived. Domesticating these crops requires a deep understanding of the physiology and genetics of trees, and scientists are turning to long-domesticated fruit trees for hints.

"Using comparative genomics approaches, characterization of the peach sequence can be exploited not only for the improvement and sustain-ability of peach and other important tree species, but also to enhance our understanding of the basic biology of trees," the team wrote.

For bioenergy researchers, the size of the peach genome makes it ideal to serve as a plant model for studying genes found in related genomes, such as poplar, one of the DOE JGI's original Flagship Genomes (http://bit.ly/JGI-Plants), and develop methods for improving plant biomass yield for biofuels.

"One gene we're interested in is the so-called "evergreen" locus in peaches, which extends the growing season," said DOE JGI's Daniel Rokhsar, under whose leadership sequencing of the peach genome began back in 2007. "In theory, it could be manipulated in poplar to increase the accumulation of biomass."

Learn more about poplar and DOE JGI Plant Flagship Genomes at http://genome.jgi.doe.gov/programs/plants/flagship_genomes.jsf.

Peach blossom (Jonathan Eisen)

A trace element's central role in harmful algal blooms

Four years after it first appeared and devastated the scallop industry, the algal masses of *Aureococcus anophagefferens* that turned the bays of Long Island, NY brown disappeared. The first species of harmful algal bloom to have its genome sequenced in 2011, the team that did the work on *A. anophagefferens* included DOE JGI researchers.

The tiny phytoplankton can outcompete diatoms in marine ecosystems and can have a significant impact on the carbon cycle in coastal ecosystems. The genome sequence indicated the presence of genes that allowed billions of algal cells to form a harmful algal bloom.

In the 2011 article, the team noted that there were many proteins that appeared to require the trace element selenium (Se). Following up on this comment, researchers focused on these proteins and their roles in harmful algal blooms. In a study published March 7, 2013 in *The ISME Journal*, the team grew algal cultures with varying concentrations of selenium under conditions mimicking the summer environmental conditions in Long Island estuaries. They found that higher levels of selenium collected in shallow estuaries, allowing the alga to form blooms. Such events were less likely in deeper waters.

"As *A. anophagefferens* relies on selenoproteins for growth and as a scarcity of Se may prohibit bloom formation in off-shore waters, Se is likely to have a key role in shaping the niche space and bloom occurrences of this species," they wrote. "Moreover, as some phytoplankton do not require Se, Se availability is likely to shape the succession and composition of phytoplankton communities in general."

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)

Expanding capabilities *(continued from page 1)*

component of the DOE JGI, other large-scale capabilities particularly those that will help link sequence to function will be provided to JGI users in the future."

For the first cycle of the ETOP, the DOE JGI has selected six new partnerships and expects to commit approximately $3 million over the next two years:

- **"Single cell approaches to metagenomics,"** proposed by Stephen Quake of Stanford University. This technology could make it easier to isolate single cells of interest from complex environmental samples.
- **"Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules,"** proposed by Jay Shendure of the University of Washington. This technology could streamline the DNA synthesis pipeline and increase the output.
- **"High-throughput sorting of microbial cells with specific functional traits for single cell genomics by combining labeling with heavy water, Raman microspectroscopy, microfluidics and flow cytometry,"** proposed by Roman Stocker of MIT and Michael Wagner of the University of Vienna (Austria). This technology could accelerate the functional characterization of genes from metagenomic sequencing experiments, one of DOE JGI's highest priorities.
- **"Generation of high-quality genomic DNA from plants and other organisms, large insert libraries and high-quality physical maps for improved physical map and sequence level-assemblies,"** proposed by Rod Wing of the Arizona Genomics Institute (AGI). This technology could make considerably easier the isolation of DNA from plants in amounts and quality that can be more effectively sequenced by the DOE JGI.
- **"Development of a pipeline for high-throughput recovery of near-complete and complete microbial genomes from complex metagenomic datasets,"** proposed by Jill Banfield of the University of California, Berkeley and Lawrence Berkeley National Laboratory, Chongle Pan of Oak Ridge National Laboratory, and Brian Thomas of the University of California, Berkeley. This technology could result in better methods for isolating and characterizing entire microbial genomes from the fragmentary sequences typical of environmental samples.
- **"Development and Implementation of High Throughput Methods for Fungal Culturing and Nucleic Acid Isolation,"** proposed by Jon Magnuson of Pacific Northwest National Laboratory. Similar to the AGI proposal above, this technology could make the isolation of DNA from fungi in amounts and quality that can be more effectively sequenced by DOE JGI considerably easier.

Spring 2013 Volume 10 Issue 2

A cost-effective process for assembling microbial genomes

Ranked among the world leaders in sequencing microbial genomes, the DOE JGI focuses on their potential applications in the fields of bioenergy and environment. Despite tremendous advances in cost reduction and throughput of DNA sequencing, significant challenges remain in the process of efficiently reconstructing these genomes. Existing technologies are good at cranking out short reads that are then assembled into longer pieces so that the order of those letters can be determined and the function of the target sequence discerned. However, genome assembly remains challenging due to the very large number of very small pieces generated by sequencers, which must then be assembled using current approaches.

As a national user facility, the DOE JGI is also focused on developing tools that more cost-effectively enable the assembly and analysis of the sequence that it generates. In a collaboration with Pacific Biosciences (PacBio) and the University of Washington, DOE JGI researchers helped develop an improved workflow for genome assembly that the team describes as "a fully automated process from DNA sample preparation to the determination of the finished genome." The work was reported in the May 5, 2013 issue of *Nature Methods*.

"We are always on the lookout for new approaches that will improve upon the efficient delivery of high-quality data to our growing community of researchers," said Len Pennacchio, DOE JGI's Deputy Director of Genomic Technologies. "This technique is one of many improvements that we are pursuing in parallel to achieve additional economies of scale."

The technique known as HGAP (Hierarchical Genome Assembly Process) uses PacBio's single molecule, real-time DNA sequencing platform, which produces reads that can be even longer than those provided by the workhorse technology of the Human Genome Project era, the Sanger sequencing technology, which produced reads of about 700 bases. The Sanger process involved creating multiple DNA libraries, conducting multiple runs, and combining the data, so that gaps in the code were covered and accuracies of a DNA base assignment were very high. Post-Sanger methods still typically require multiple libraries and often a mix of technologies to produce optimal results.

DOE JGI researchers are part of a team that has developed what is described as "a fully automated process from DNA sample preparation to the determination of the finished genome." (Roy Kaltschmidt, LBNL)

With HGAP, the team reported, "only a single, long-insert shotgun DNA library is prepared and subjected to automated continuous long-read SMRT sequencing, and the assembly is performed without the need for circular consensus sequencing."

The *de novo* assembly method was tested using three microbes previously sequenced by the DOE JGI. The data collected were compared against the reference sequences for these microbes and the team found that the HGAP method produced final assemblies with >99.999% accuracy.

The team will now seek to extend the utility of this new assembly method beyond microbes to the genomes of more complex organisms.

Save the Date

The 2014 Department of Energy Joint Genome Institute (DOE JGI)

GENOMICS *of* ENERGY & ENVIRONMENT MEETING

March 18-20, 2014

Walnut Creek, CA

Scientists interested in learning about state of the art genome sciences and associated technologies as well as their potential applications to challenges in bioenergy and environmental issues are invited to participate in the 9th Annual DOE JGI Joint Genome Institute Genomics of Energy and Environment Meeting.

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

CSO26342