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

the
PRIMER
Winter 2010 Volume 7 Issue 1

Revealing the Soybean Sequence: A Series of Firsts

The soybean, one of the most important global sources of protein and oil, has become the first legume to have a published complete draft genome sequence. In the January 14 issue of the journal *Nature,* a team of researchers from the U.S. Department of Energy Joint Genome Institute (DOE JGI), the U.S. Department of Agriculture-Agricultural Research Service (USDA-ARS), the National Science Foundation (NSF), the University of Missouri, Purdue University, and a dozen other institutions described the sequence and how the information might be applied to agricultural strategies and biodiesel production.

"The soybean genome's billion-plus nucleotides afford us a better understanding of the plant's capacity to turn sunlight, carbon dioxide, nitrogen and water, into concentrated energy, protein, and nutrients for human and animal use," said Anna Palmisano, Associate Director of the DOE Office of Science's Office of Biological and Environmental Research. "This opens the door to crop improvements that are sorely needed for energy production, sustainable human and animal food production, and a healthy environmental balance in agriculture worldwide."

Briefly setting aside the list of potential applications to be derived from the sequence, Jeremy Schmutz, the study's first author and a DOE JGI scientist at the HudsonAlpha Institute for Biotechnology in Alabama, noted two other key points about the complete draft genome. "The soybean sequence project is the largest plant project done to date at the Joint Genome Institute," he said. "It *cont. on page 4*

also in this issue

Photo by Roy Kaltschmidt, LBNL

Now Available: Microbial Genomes Encyclopedia, Vol. 1

Nearly 2,000 microbes have been sequenced out of the estimated nonillion (10^{30}) in, on and around the Earth. And while the information is significantly impacting almost all aspects of microbiology, said DOE JGI Phylogenomics Program Head and University of California, Davis professor Jonathan Eisen, it is bypassing the ribosomal RNA Tree of Life, which allows researchers to track and understand how organisms are related to each other.

"We've done a very poor job of sampling across the tree in microbial studies," said Eisen. "If you look at phylogenetic diversity in the bacterial kingdom, most of the available genomes come from just 3 of the 40 major phyla. The same trend holds for archaea, eukaryotes and viruses. The solution is to use the tree to guide us, going through phylogenetic diversity to explicitly fill in missing branches of the tree with actual data."

To remedy the problem of insufficient phylogenetic diversity, Eisen and his colleagues sequenced 100 bacterial and archaeal genomes that represent little-studied branches of the Tree of Life. The work, considered the first "volume" of a Genomic Encyclopedia of Bacteria and Archaea or GEBA, was published in the December 24, 2009 issue of the journal *Nature.*

The GEBA *cont. on page 6*

IMG 3.0 Goes Live

The Integrated Microbial Genomes (IMG) system, featured in a recent edition of *Nucleic Acids Research** serves as a community resource for comparative analysis and annotation of all publicly available genomes from three domains of life in a uniquely integrated context. IMG 3.0, the 18th release, went live on December 21, 2009.

The content of IMG 3.0 has been updated with new microbial genomes available in RefSeq version 37 (June 02, 2009) and contains a total of 5,558 genomes consisting of 1,748 bacterial, 77 archaeal, 76 eukaryotic genomes, 2,606 viruses (including bacterial phages), and 1,051 plasmids that did not come from a specific microbial genome-sequencing project. Among these genomes, 4,650 are finished genomes, and 904 are draft genomes, and four are permanent draft (i.e., will never be finished) genomes. Twenty-seven new fungal genomes have been also included in IMG 3.0. Compared with IMG 2.9, IMG 3.0 contains 7,540,500 genes, an increase of 1,026,256 genes.

MetaCyc and KEGG pathways in IMG 3.0 have been updated with MetaCyc version 13.5 and KEGG version 52.0 respectively. The Pfam collection of protein domain families has been updated based on Pfam version 24.0, and Pfam clans have been added as an additional classification of Pfam domain families.

In addition, chromosomal gene cassettes† have been recomputed together with estimates of their conservation across IMG genomes.

Genes in IMG involved in regulatory interaction experiments controlling their expression are now linked to RegTransBase (http://regtransbase.lbl.gov). RegTransBase is a database of regulatory sequences and regulatory interactions on the transcriptional and posttranscriptional levels in prokaryotic genomes, that contains experimental data and predicted sites published in scientific journals.

IMG 3.0 also contains proteomic data from recent *Arthrobacter chlorophenolicus, Cryptobacterium curtum*, and *Brachybacterium faecium* studies.

The User Interface for IMG 3.0 has been extended with Scaffold Cart tools that facilitate the analysis of genomes at the level of individual scaffolds and contigs, such as individual chromosomes and plasmids. ACT (Artemis Comparison Tool), a viewer based on Artemis for pair-wise genome DNA sequence comparisons, has also been added to IMG's suite of synteny viewers.

For additional information: see *What's New* and *Using IMG:* http://img.jgi.doe.gov.

**Nucleic Acids Research*, 2010, Vol.38: http://nar.oxfordjournals.org/cgi/reprint/gkp887v1

†*PLoS ONE* 4(11): http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0007979

news

Cassava draft genome sequence spurs Gates Foundation funding

Not long after the first draft of the annotated cassava genome sequence was made available on the DOE JGI's Phytozome.net in November, the Bill & Melinda Gates Foundation announced a $1.3 million grant to fund the development of a genome variation database that will help farmers grow more disease-resistant and nutritious varieties of the root crop in less time.

A third of the cassava harvest in Africa is lost because of pathogens such as cassava brown streak disease. A staple food for more than 750 million people around the world, cassava was sequenced by the DOE JGI as part of CSP 2007 to help develop a variety with improved resistance to CBSD and other diseases. Claude Fauquet, chair and co-founder of the Global Cassava Partnership and a researcher at the Danforth Center, said having the genome sequence of cassava will benefit the international food security situation as well as help improve the farmers' health and economic growth.

Several varieties of cassava will be sampled by an international consortium that includes the University of Arizona, the DOE JGI, the University of Maryland and 454 Life Sciences, in collaboration with researchers in Kenya, Uganda and Tanzania, to identify genes that correspond to important traits and develop a genetic markers database.

the PRIMER / 3
Winter 2010 Volume 7, Issue 1

Studying life in a "Dead Zone"

For researchers at the University of British Columbia (UBC), the Saanich Inlet off the coast of British Columbia, Canada is an ideal "living lab" to study the microbial communities in low oxygen waters. As these so-called "dead zones" expands in oceans worldwide, so does interest in understanding how the microorganisms that thrive in these regions affect and are impacted by the changes to their ecosystems.

In the October 23, 2009 issue of the journal *Science,* a team of UBC and DOE JGI researchers described the results of a study conducted over several seasons on the microbial communities of Saanich Inlet, which led to the identification of the most abundant organism called SUP05. Study senior author and UBC professor Steven Hallam noted that the team obtained enough sequence coverage to assemble what they called "the SUP05 metagenome, a composite of the entire SUP05 population spanning the various environmental samples that we sequenced."

The project is part of the DOE JGI's Community Sequencing Program established in 2004 to tackle mission-relevant genomics projects that support the goals of the U.S. Department of Energy to develop clean, sustainable bioenergy sources and characterize biological and environmental processes such as biogeochemistry and carbon cycling.

Susannah Tringe, a metagenomics scientist at the DOE JGI, said that oxygen minimum zones (OMZs) are sinks for an essential nutrient that marine organisms need to survive—nitrogen—as well as sources for the greenhouse gases methane and nitrous oxide. "By studying the genomes of the uncultivated microbes found in OMZs, we can better understand how they participate in global geochemical cycles such as the carbon and nitrogen cycles," she said.

Hallam described SUP05 as a paradoxical organism, one that fixes carbon dioxide and removes toxic sulfides, but which might also be producing nitrous oxide, a more potent greenhouse gas than either carbon dioxide or methane. The researchers found that SUP05 is closely related to sulfur-eating gill symbionts of deep sea clams and mussels, though it utilizes nitrate rather than oxygen in its energy metabolism. Additionally, a comparative analysis revealed that 35 percent of the SUP05 genome is unique and is involved in helping the bacteria adapt to changing environmental conditions such as the seasonal increase and decrease of oxygen levels in Saanich Inlet, and the shifting balance of the nitrate and sulfide levels that are its key energy resources.

"Just as cyanobacteria play an essential role in producing atmospheric oxygen; in future oceans this could be one of those organisms that play similarly integral roles, albeit with different ecological outcomes," said Hallam. He noted that the SUP05 microorganism and its relatives will become increasingly important as OMZs continue to expand, providing researchers with a biological indicator useful in monitoring the changing state of the global ocean.

"Global warming is changing the chemistry of the oceans and one of the byproducts of change is that the ocean pH is becoming acidic," Hallam said. "Blooming SUP05 populations have the potential to help offset rising carbon dioxide levels that ultimately lead to ocean acidification."

Hallam and his team intend to use their time-resolved studies as a basis for comparison in the context of another CSP project of Hallam's approved earlier this year which focuses on an extensive OMZ in the eastern North Pacific Ocean. The team also plans to eventually compare their work in Saanich Inlet to data collected from other dead zones around the world.

Study first author David Walsh (left) and study second author Elena Zaikova (below), both at the University of British Columbia, on a water sampling trip at Saanich Inlet (right). Courtesy of the Hallam Lab.

Winter 2010 Volume 7, Issue 1

Image courtesy of the United Soybean Board

Jeremy Schmutz, a DOE JGI scientist at the HudsonAlpha Institute for Biotechnology in Alabama

Soybean Sequence

cont. from page 1

also happens to be the largest whole genome shotgun plant that's ever been sequenced. We took the approximately 1.2 Gigabase genome, broke it apart and reassembled it like a puzzle."

One major significant application of the soybean genome sequence Schmutz mentioned is for biodiesel production. Right now the plant doesn't produce enough oil to compete with petroleum products, though he noted the legume is the major source of biodiesel worldwide.

Tom Clemente, a professor with appointments at the Center for Biotechnology and Center for Plant Science Innovation at the University of Nebraska, Lincoln, said the soybean genome sequence could offer solutions to the production problem. "We can now zero in on the control points governing carbon flow towards protein and oil," he said. "With the combination of informatics, biochemistry and genetics we can target the development of a soybean with greater than 40 percent oil content."

University of Missouri professor Gary Stacey, Director of the Center for Sustainable Energy and Associate Director of the National Center for Soybean Biotechnology, said he and his colleagues have also identified more than 46,000 genes from the sequence analysis, of which 1,110 are involved in lipid metabolism. "These genes and their associated pathways are the building blocks for soybean oil content and represent targets that can be modified to bolster output and lead to the increase of the use of soybean oil for biodiesel production," he said.

Schmutz said another major application of the soybean genome sequence would be to provide a reference sequence for more than 20,000 legume species, helping agricultural researchers boost soybean yields and learn more about the nitrogen-fixing symbiosis so critical to successful agricultural crop rotation strategies.

In the past, he said, farmers picked plants in the field and bred them together to improve crop yields. "In recent years we've kind of tapped out on traditional soybean breeding and can't seem to increase the yield anymore. Using genomics allows us to breed specific genes, identify specific traits such as drought tolerance, pathogen resistance and more seed production, and breed them back into soybean lines," he said.

For example, the soybean genome sequence has already allowed researchers to identify first resistance gene for Asian Soybean Rust (ASR), a disease that can reduce yields by as much as 80 percent in some countries. Another discovery effort spurred by the sequence has pinpointed a mutation that researchers can use to find soybean lines with lower levels of the sugar stachyose, which will improve the ability of animals and humans to digest soybeans.

the PRIMER / 5
Winter 2010 Volume 7, Issue 1

Photo by Kim Closser; Studio3, Columbia, MO

Image courtesy of the United Soybean Board

JGI collaborator Gary Stacey, Associate Director of the National Center for Soybean Biotechnology at the University of Missouri

"This is a milestone for soybean research and promises to usher in a new era in soybean agronomic improvement," said Stacey. "The genome provides a parts list of what it takes to make a soybean plant and, more importantly, helps to identify those genes that are essential for such important agronomic traits as protein and oil content."

A third research project that compared the genomes of soybean and corn has also led to the discovery of a single-base pair mutation that reduces phytate production in soybean, which could in turn reduce the environmental runoff from livestock waste. Phytate is the form in which phosphorous is stored in plant tissue and it isn't absorbed by animals that eat feeds with soybean mixed in. This can lead to higher phosphorus levels in manure, which can become a major contaminant.

Photo by Roy Kaltschmidt, LBNL

Winter 2010 Volume 7, Issue 1

Image by Roy Kaltschmidt, LBNL

GEBA *cont. from page 1*

project was launched in May 2007 with the goal of first identifying unrepresented branches from the phylogenetic Tree and then identifying organisms from these branches that could provide DNA samples for sequencing. The DOE JGI team collaborated with researchers at the non-profit German Collection of Microorganisms and Cell Cultures, DSMZ (http://www.dsmz.de/), to sequence 100 bacterial and archaeal genomes for the pilot project, though Eisen said approximately 170 genomes have been sequenced and many of them are being finished.

The findings reveal that using a guide such as the rRNA Tree of Life to phylogenetically select organisms, especially uncultured ones, allows diverse genomes to be sequenced, and in turn provides them for use in annotating other genome sequences. "You might not care about the genomes sequenced in this study," said Eisen, "but they provide the ability to study other genomes you might care about more."

He also noted that the GEBA project emphasizes the vastness of microbial diversity, pointing out that to even sample half of the known phylogenetic diversity, researchers would need to sequence another 10,000 genomes of what are still mostly uncultured organisms and then understand the biology of the organisms.

"Many people have talked about doing something like this for the last 10 years; no one tried it until JGI tried it," said Eisen. "What this project is is an example of the type of project that a large genome center like JGI can do in the new generation where sequencing is a lot cheaper and a lot faster, and ho5w a place like JGI can set itself apart from all the other labs around the country. This has been an amazing project."

Videos of Eisen discussing the GEBA project can be viewed at: http://www.scivee. tv/user/7476.

For a list of the GEBA pilot project targets, see: http://www.jgi.doe.gov/sequencing/GEBAseqplans.html.

More information available at http://www.jgi.doe.gov/News/news_09_12_23.html.

the PRIMER / 7
Winter 2010 Volume 7, Issue 1

DOE JGI Developing Integrated Tracking System

Software Project Manager Steven Wilson

To keep up with changes to sequencing strategies and the resulting plethora of projects, the DOE JGI is developing a system that allows project managers and collaborators to check on the status of a project.

"For the decade JGI has existed, we used a single process for sequencing. In 2008, we added two other platforms and we increased production by 20-fold," said DOE JGI Director Eddy Rubin while discussing the need for an Integrated Tracking System (ITS). "We used to be an in and out burger place; now we're a Chinese restaurant with a million different dishes. We've changed what we're doing and the variety of types of projects. We need to develop a tracking system to move forward."

The ITS project is being led by Software Project Manager Steven Wilson, who noted that while several systems are in place to handle everything from proposal submission to post-production sequencing and annotation, the DOE JGI doesn't as yet have a single-strategy scheduling component that can track a project's status without using custom queries.

"What the system will be able to do is change the way information is presented," said Wilson. "ITS expected to improve the efficiency of the scheduling process with a better visualization and cost-tracking, improving the flexibility involved in bringing new types of projects and sequencing platforms into play."

DOE JGI's own Jim Bristow (center left) and Susannah Tringe (center right) appeared on a panel with Jay Keasling on September 28, 2009 at the Berkeley Repertory Theatre's Roda Stage. KTVU Channel 2 health and science editor John Fowler (left) moderated the talk entitled: "From the sun to your gas tank: A new breed of biofuels may help solve the global energy challenge and reduce the impact of fossil fuels on global warming," discussing ways to convert the solar energy stored in plants into liquid fuels. A video of the talk can be viewed at http://www.youtube.com/watch?v=mRTwuxVurIE

Rector Magnificus Designatus Karel Luyben (on the right), Chancellor of the Delft University of Technology in the Netherlands, visited the DOE JGI on December 8, 2009 and was given a tour of the sequencing facilities by Sanger Coordinator Simon Roberts.

The second week of December 2009 brought a cold storm down from the north, dropping several inches of snow on the ridges surrounding Walnut Creek, Calif., including on the slopes of the 3,864-foot Mount Diablo, as seen from the front steps of the DOE JGI.

Winter 2010 Volume 7, Issue 1

Genomics of Energy & Environment

March 24-26
Walnut Creek, California

The 2010 Department of Energy Joint Genome Institute (DOE JGI) Genomics of Energy & Environment meeting will be held March 24-26 in Walnut Creek, California and specifically emphasize the genomics of renewable energy strategies, biomass conversion to biofuels, environmental gene discovery, and engineering of fuel-producing organisms.

Confirmed keynote speakers include:

Rita Colwell
Distinguished Professor, University of Maryland and Johns Hopkins University Bloomberg School of Public Health

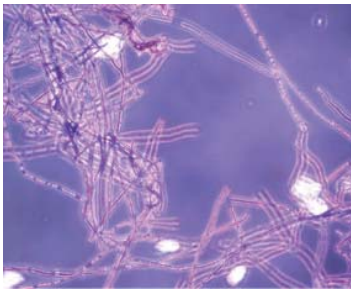
Jay Keasling
CEO, DOE Joint BioEnergy Institute

Scheduled speakers include:

Cristina Cuomo, Broad Institute; **Evan DeLucia,** University of Illinois at Urbana-Champaign; **Richard Flavell,** Ceres; Steven Hallam, University of British Columbia; **Dennis Hedgecock,** University of Southern California; **Madhu Khanna,** University of Illinois at Urbana-Champaign; **Steve Knapp,** University of Georgia; **Tom Mitchell-Olds,** Duke University; **Steve Moose,** University of Illinois at Urbana-Champaign; **Joseph Noel,** Salk Institute for Biological Studies; **Forest Rohwer,** San Diego State University; **Steven Savage,** Cirrus Partners; **Gary Stacey,** University of Missouri; **Jim Tiedje,** Michigan State University; **Adrian Tsang,** Concordia University; **Detlef Weigel,** Max Planck Institute for Developmental Biology; **Alexandra Worden,** Monterey Bay Aquarium Research Institute

Under the leadership of Fungal Genomics Program head Igor Grigoriev, DOE JGI hosted the Comparative Genomics of Thermophilic Fungi Jamboree December 2-4, 2009. This Jamboree brought together 21 collaborators to explore the genomics and biology of thermophilic fungi by comparative analysis of three genomes: *Thielavia terrestris* (below), *Sporotrichum thermophilum* and *Chaetomium globosum*—two thermophiles and a mesophile.

Image courtesy of Andrew Tsang, Concordia University



Contact The Primer
David Gilbert, Editor / DEGilbert@lbl.gov / (925) 296-5643

CSO 18246