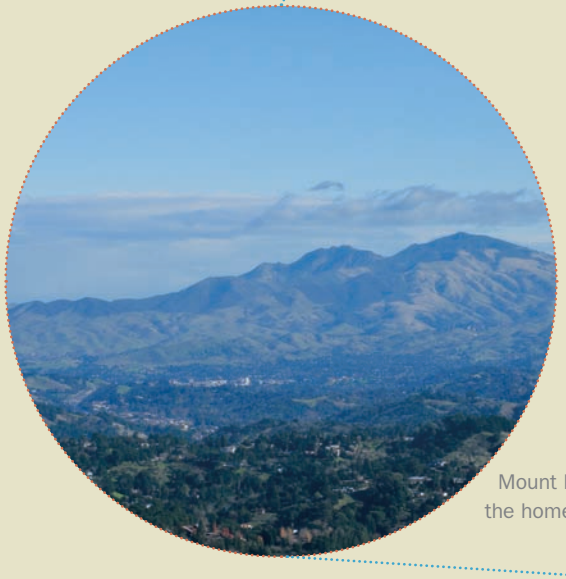
2010 U.S. Department of Energy Joint Genome Institute Progress Report

Sequencing the world of possibilities for energy and the environment

Table of Contents



Mount Diablo, looming over Walnut Creek, California,
the home of the DOE Joint Genome Institute

DOE JGI Mission

The mission of the U.S. Department of Energy Joint Genome Institute is to serve the diverse scientific community as a user facility, enabling the application of large-scale genomics and analysis of plants, microbes, and communities of microbes to address the DOE mission goals in bioenergy and the environment.

Director's Perspective

Eddy Rubin with Production head Susan Lucas at the DOE JGI Sanger phase-out ceremony

2010 marked a year of transitions for the U.S. Department of Energy Joint Genome Institute (DOE JGI). Changes occurred in DOE JGI's sequencing technologies, in the information technology infrastructure that underlies all of its processes, and in our perspective on the genome capabilities that the Institute should offer in the future, in addition to sequencing as it evolves as a DOE genomic user facility.

In the fall, DOE JGI marked the complete phasing out of the "workhorse" Sanger sequencing platform. This mature sequencing technology has been the mainstay of production sequencing for the Institute's first 10 years (we had over 105 such machines during the peak of the human genome sequencing efforts). With the steady improvements in next-generation sequencers, it simply no longer made sense to maintain the elaborate infrastructure and costs associated with the support of the Sanger instruments. Subsequently, DOE JGI has focused its production activities on the evolving Illumina platform. A concerted effort was made in the past year to provide computational and molecular biology support of the Illumina platform such that the full complement of the DOE JGI's products — plant, fungal, microbial, and metagenomic — could be generated with a data quality that our users have come to expect from the Sanger platform, at dramatically reduced cost and increased throughput. In addition, a third-generation sequencing technology has been brought online at DOE JGI, the Pacific Biosciences (PacBio) Single Cell Molecular Real-Time Detection System. This instrument has the potential to generate extremely long nucleotide base reads in short periods of time. It is expected that PacBio's modest throughput of long reads will complement the massive data output of the short-read Illumina technologies in the coming year to supply our users with improved genomic products.

With the DOE JGI's transition to more data-intensive technologies, we sought to maximize our in-house computational capabilities while also tapping into a more robust computational infrastructure. In 2010, we signed a formal Memorandum of Understanding and partnered with the Lawrence Berkeley National Laboratory's National Energy Research Scientific Computing Center (NERSC) Division to more closely integrate DOE JGI staff and computing capability with this world-class institution. This alliance represents a coming-of-age for the combined genomics and high performance computing enterprise. NERSC has a successful track record in responding to these kinds of massive-scale data challenges, even though the tasks DOE JGI presents are quite different from NERSC's more traditional High-Performance Computing workload. Data-centric computing is an important future direction for NERSC, and with genomics and subsequent levels of complex -omics data, we foresee exponential increases in computation and storage requirements, which NERSC is well positioned to handle for us.

Our ever-expanding primary user community — over 1,800 principal investigators, co-PIs, collaborators, and genome annotators on active DOE JGI Genome Projects in FY 2010 — continues to benefit from access to DOE JGI's data generation and analysis capabilities.

A particularly important metric of DOE JGI is the wealth of contributions to the scientific literature, with more than 150 journal papers connected to the DOE mission areas of energy and the environment—of which 22 were manuscripts published in the high-impact journals *Science*, *Nature,* and *PNAS*.

Here is a sampling of just a few of the publications from the past year:

- The comparative analysis, published in the journal *Science*, of the genome of the multicellular alga *Volvox carteri* with that of the unicellular alga *Chlamydomonas reinhardtii*, which was previously sequenced by the DOE JGI. The results have implications for how these aquatic photosynthetic organisms can be engineered and harnessed for potential biofuels.

- The publication in *Nature Biotechnology* of the genome analysis of the white rot fungus *Schizophyllum commune*, which joins another white rot and the first brown rot fungus as the third wood-decaying fungus completed by DOE JGI.
- The ability of leaf-cutter ants to tend a fungal "garden" to optimize their ability to break down cellulose for their nutrition — something that could be mimicked for biofuels production improvements — was described in a *PLoS Genetics* publication.
- The culmination of some Community Sequencing Program (CSP) research on characterizing the foregut microbiome of the Tammar wallaby. A metagenomic analysis, published in the *Proceedings of the National Academy of Sciences* (PNAS), revealed microbes found therein that could be considered for new strategies for breaking down plant biomass toward the production of cellulosic biofuels.
- The study, published in *The Plant Cell*, of the microalga *Chlorella variabilis* that highlighted the pathway through which the organism acquired the cell wall component chitin, as well as mechanisms for carbon capture and lipid production that may have implications for biotechnology applications.
- The first metagenomic analysis of a syntrophic community — where bacteria and archaea cooperatively break down organic matter — analyzed in the context of a plastics wastewater system and published in the *International Society for Microbial Ecology* (ISME) *Journal.*
- The development of a new computational tool for microbial genome annotation quality control known as "GenePRIMP," which highlights errors that need to be manually corrected — a feature that is critically important to the DOE JGI user community for their efficient use of these data, was published in *Nature Methods.*
- The validation of two ribosomal RNA removal methods have enabled the analysis of the subset of genes that, for example, are transcribed under certain environmental conditions, appeared in *Nature Methods.*
- The genome analysis of the Western clawed frog *Xenopus tropicalis* was reported in and appeared on the cover of *Science*, recalling DOE JGI's prior emphasis on understanding cell and evolutionary development through sequencing.
- As a footnote to the quality of DOE JGI science and scientists, the journal *Science*, in its final issue of 2010, had a list of the 10 most important "Insights of the Decade." DOE JGI investigators were key contributors to three of these insights:
  1) Pioneering contributions to the field of paleogenomics — the analysis of genomes from ancient samples;
  2) Debunking the notion of so-called "junk DNA," which had no immediately obvious associated functions, and showing that evolutionarily conserved DNA was frequently involved in gene regulation; and
  3) The development of the computational tools that were employed by others to help characterize nearly 1,000 microbial species found in the human body—the human "microbiome."

In continuing to fill DOE JGI's streamlined sequencing and data analysis pipelines, the CSP portfolio has grown and diversified. The 35 approved FY 2011 CSP targets include arctic algae, barley (whose genetic code is nearly equivalent to sequencing two full human genomes), and microbial communities in deep-sea hydrothermal vents. Also, the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project continues to fill in unrepresented branches of the Tree of Life with an analysis of an additional 61 bacteria and archaea that could lead to the identification of new proteins and subfamilies useful in addressing DOE missions.

DOE JGI also continues to devote a significant portion of its capacity to sequencing on behalf of the three U.S. Department of Energy Bioenergy Research Centers (BRCs). Our collaborations of the past year have

led to publications characterizing: a genus of heat-thriving, plant biomass-degrading bacteria (with the BioEnergy Science Center, Oak Ridge, TN); the analysis of leaf-cutter ant-fungal symbiosis (noted above, with the Great Lakes BioEnergy Research Center, Madison, WI); and microbial genes in arid grasslands (with the Joint BioEnergy Institute, Emeryville, CA).

Looking ahead, The Grand Challenges:

With the dramatic increases in sequencing productivity, we have positioned DOE JGI to work on Grand Challenge projects of tera- (trillion) and peta- (quadrillion) base scale. Transitioning in project scale, among the most notable achievements of the past year, published early in 2011 in the journal *Science*, is the massive analysis of the network of organisms isolated from the cow forestomach—the rumen. This work revealed the genomes and genes of previously uncharacterized plant-digesting microbes encoding tens of thousands of enzymes that may possess powerful capabilities for degrading biomass into simple sugars, the essential first step in cellulosic biofuel production.

Another Grand Challenge advanced in 2010 was the Great Prairie Soil Metagenome project. DOE JGI's is interested in the Great Prairie because this environment sequesters the most carbon of any soil system in the United States, yielding large amounts of biomass that may play a prominent role in future cellulosic biofuels strategies. Our collaborators are sampling pristine, never-tilled prairie soil, which we are comparing against Iowa cornfields that have been in continuous cultivation for over 100 years. The information coming to light will enable improved soil management, carbon sequestration, and ecosystem productivity. Other Grand Challenges include: A broad survey of rhizobial (nitrogen-fixing organism) genomes collected from distinct geographical regions, the first large-scale attempt to understand the genetics involved in plant-bacteria interactions; a large-scale comparative analysis of Brassica genomes (e.g., *Arabidopsis* species, etc.), and a broad sampling of the Earth's microbiome — a massive multidisciplinary, multi-institutional, and international effort to analyze microbial communities across the globe to produce a gene atlas describing (among other features), environmental metabolic models for different ecosystems.

As genome-enabled science advances further into the second decade of the 21st century, DOE JGI is involved in strategic planning to best position itself as the user facility of the future. In the past decade, DOE JGI has served largely as a DNA sequence generation facility and as it goes forward, what is taking shape is a facility that will provide users a variety of genomic and other "-omics" capabilities in addition to DNA sequencing and analysis. In refining our strategic planning we are engaging our users to inform us about the kinds of science they hope to be doing in the next five to 10 years and what sorts of capabilities will be required to enable their science. Among the capabilities being contemplated are: massively parallel culture of cells and their genomic monitoring under multiple conditions, large-scale functional assays, DNA synthesis, and the creation of large pieces of DNA for genome engineering.

DOE JGI remains committed to providing users with cutting-edge genomic technologies and the associated informatics support enabling the seamless data-to-knowledge conversion required to address challenges of energy and the environment to push forward the frontiers of discovery.

I look forward to apprising you of our progress along this exciting path.

Edward M. Rubin, MD, PhD
Director
DOE Joint Genome Institute

DOE Mission Areas

Bioenergy

The United States is the single largest consumer of petroleum-based energy, and imports more than half of the amount used domestically. Two-thirds of the energy consumed goes to transportation and industry, which partially explains the impetus behind the Department of Energy's focus on developing clean and sustainable alternative fuel sources such as cellulosic biofuels.

Sequencing projects at DOE JGI that contribute toward meeting this goal tend to focus on one of three categories: developing plants that can be used as feedstocks for biofuel production; identifying organisms (e.g., fungi and microbes) with enzymes and pathways that can break down the lignin and cellulose in plant cell walls; and identifying organisms with enzymes and pathways that can ferment sugars into biofuels.

The root of the Cassava plant contains as much as 40 percent starch, ranking among the top five crops in global starch production, making it a candidate biofuel feedstock.

The Great Prairie sequesters the most carbon of any soil system in the United States. A comparative metagenomic analysis could enable more efficient carbon sequestration techniques as well as improve annual yields of biomass for cellulosic biofuel production.

Carbon Cycle

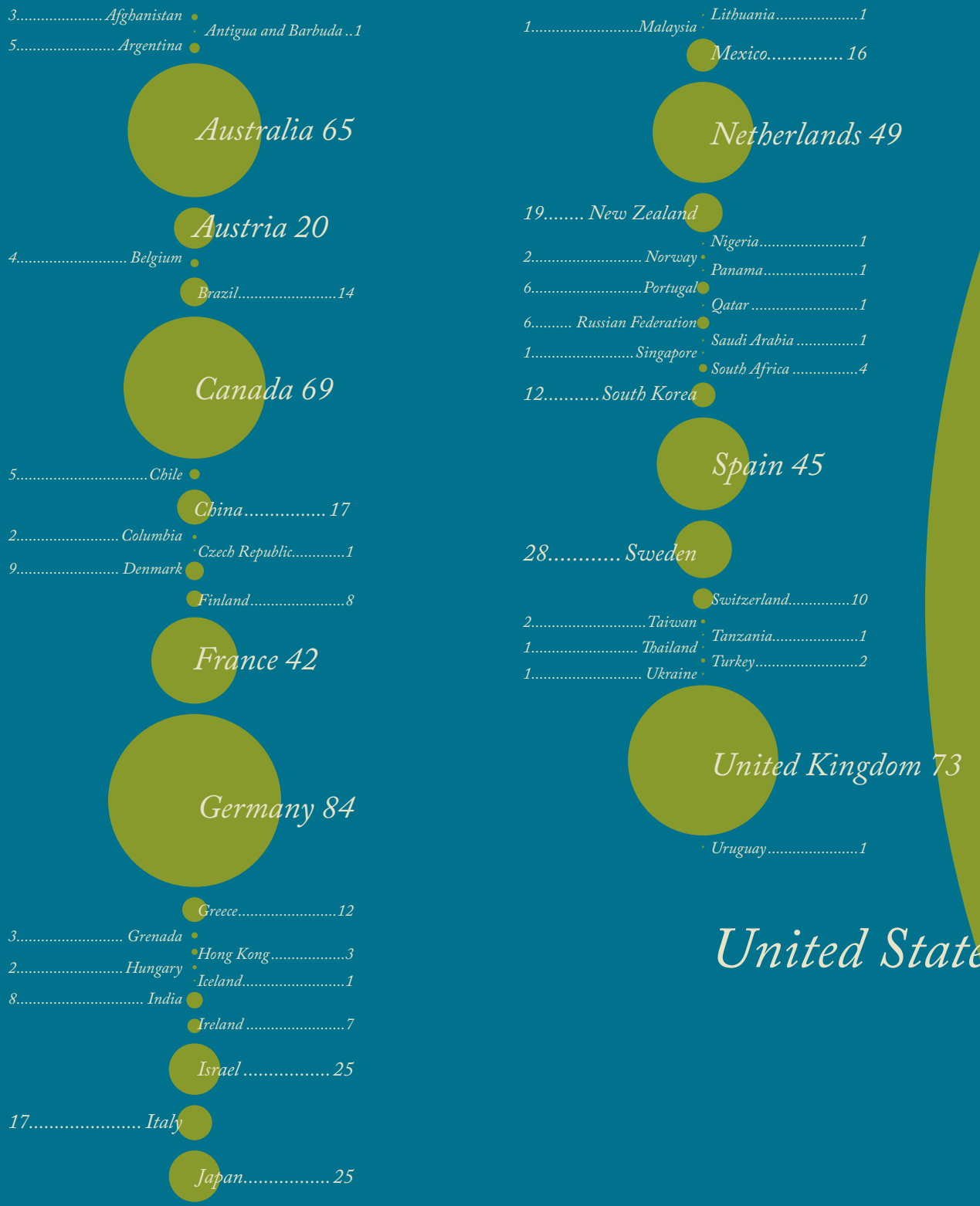
The global carbon cycle regulates the levels of atmospheric carbon dioxide and the Earth's climate. It is heavily dependent on the microbes that fix atmospheric carbon, promoting plant growth and degrading organic material. As microbes make up the largest component of the Earth's biodiversity, understanding how they metabolize carbon, and how environmental changes affect these processes, is crucial. DOE JGI is sequencing several microbes and microbial communities that influence carbon cycling. Improved predictive models based on these data will facilitate the development of effective strategies toward reducing the effects of increasing carbon dioxide emissions on the global climate.

Biogeochemistry

Beyond the carbon cycle, a wide spectrum of processes regulates our natural environment, and the field of biogeochemistry explores these biological, physical, geological, and chemical processes and the reactions involved. DOE JGI focuses on microbes and microbial communities that can degrade or otherwise transform environmental contaminants such as toxic chemicals or heavy metals in order to understand how they restore and maintain the environment.

Microbes make up the largest component of the Earth's biodiversity, and those that thrive in extreme environments, such as thermophiles found in hot springs, are of interest to biofuels researchers looking for enzymes that could be used to break down plant biomass for cellulosic biofuel production.

By the Numbers



Number of Users

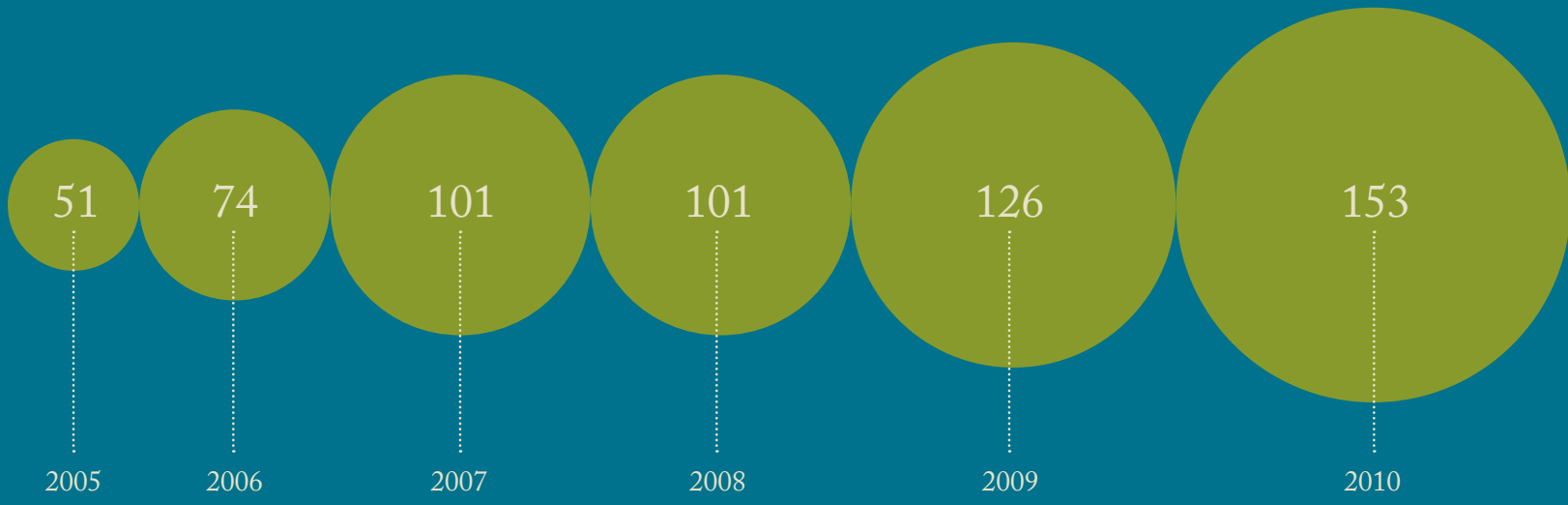
52

Countries

1,871

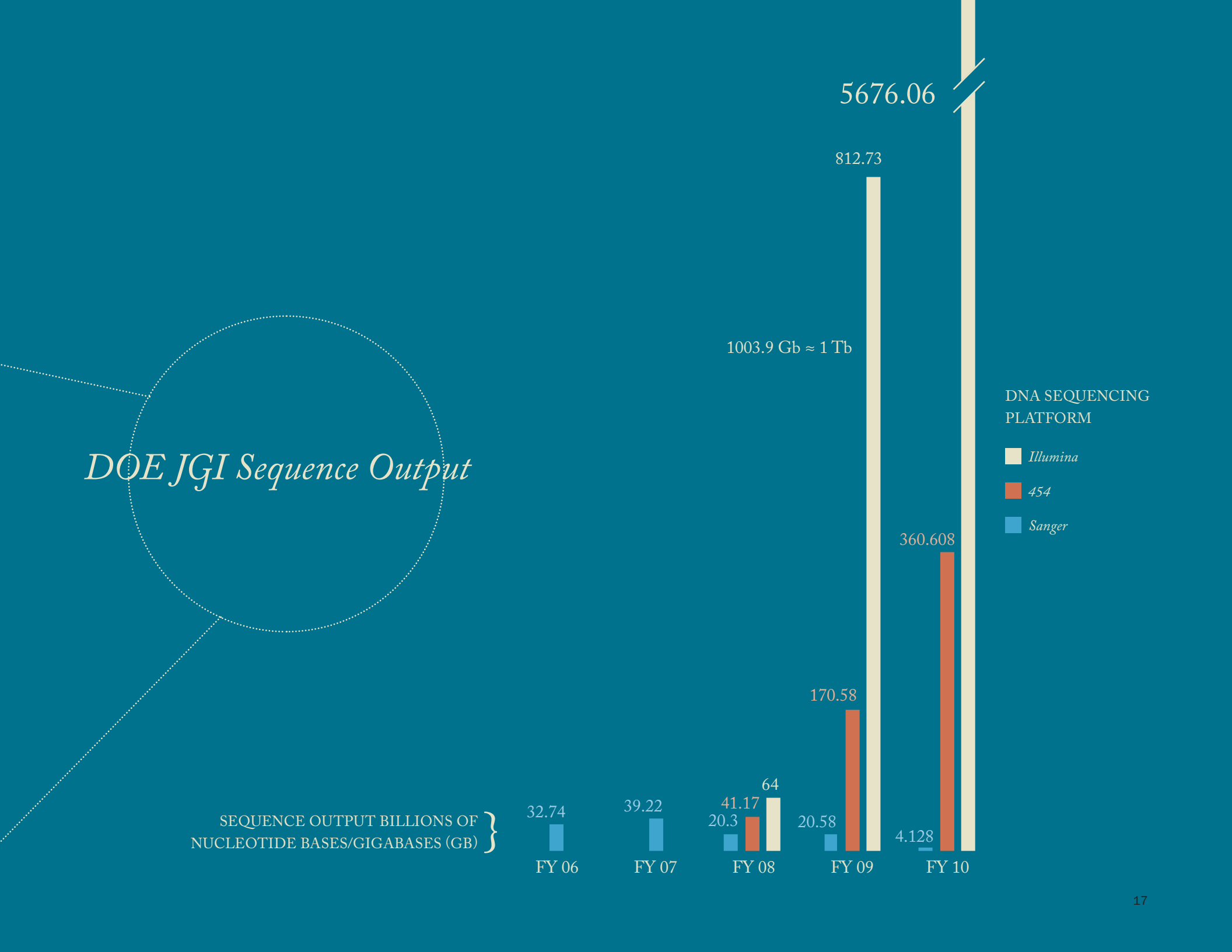
Total users worldwide

Contributing to the Scientific Literature/Impact



NUMBER OF PUBLICATIONS

DOE JGI Sequence Output



DOE JGI — Team Science Realized

Part of the Lawrence Berkeley National Laboratory (Berkeley Lab), the DOE JGI in Walnut Creek, California, carries out production-scale sequencing and analysis augmented by specialized tasks drawn from the specific capabilities provided by its partner laboratories: Lawrence Livermore National Laboratory (LLNL), Los Alamos National Laboratory (LANL), Oak Ridge National Laboratory (ORNL), Pacific Northwest National Laboratory (PNNL), and the HudsonAlpha Institute for Biotechnology (in Huntsville, Alabama). The DOE JGI's 80,000-square-foot headquarters in Walnut Creek has a staff of 285 employees, of which 250 are in Walnut Creek, which draws primarily from Berkeley Lab and LLNL employees, and 35 additional employees at its other partner locations.

The DOE JGI Director, working with the DOE JGI Joint Coordinating Committee (JCC), which includes members from each of the partner laboratories, coordinates the vision for the overall DOE JGI as well as the activities carried out at each of the partner labs. The JGI Director, representing the JCC, communicates DOE JGI's technological and strategic vision to the DOE Office of Biological and Environmental Research (OBER), and funding decisions are based upon these discussions.

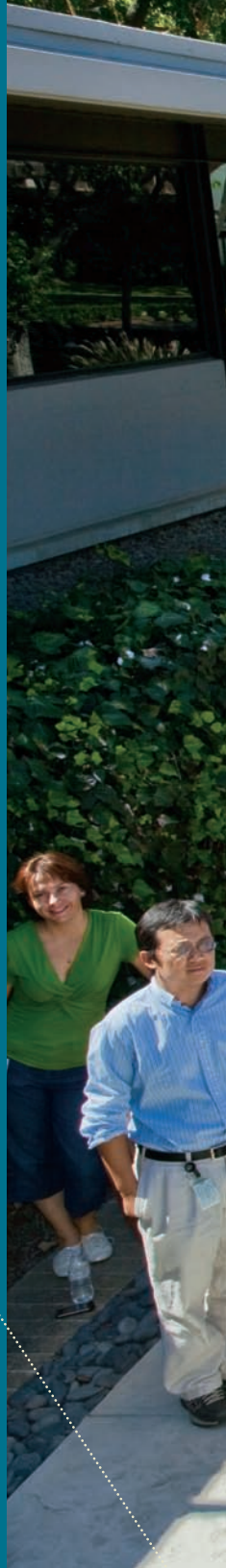
DOE JGI User Community

DOE JGI's user facility approach is based on the concept that by focusing the most advanced sequencing and analysis resources on the best peer-reviewed proposals drawn from a diverse community of scientists, DOE JGI will both catalyze creative approaches to addressing DOE mission challenges and advance genomic science. This strategy has clearly worked, only partially reflected in the fact that DOE JGI has played a major role in more than 50 papers published in the prestigious journals *Nature* and *Science* alone over the past three years. The involvement of a large and engaged community of users working on important energy and environmental problems has helped maximize the impact of DOE JGI science.

The Community Sequencing Program (CSP) was created to provide the scientific community at large with access to high-throughput sequencing at DOE JGI. Sequencing projects are chosen based on scientific merit — judged through independent peer review — and relevance to the DOE mission.

DOE JGI's largest customers are the DOE Bioenergy Research Centers (BRCs), launched in 2007 to accelerate basic research in the development of next-generation cellulosic and other biofuels through focused efforts on biomass improvement, biomass degradation, and strategies for fuels production. The three centers are the Joint BioEnergy Institute (JBEI) led by Berkeley Lab and located in Emeryville, California; the BioEnergy Science Center (BESC) at Oak Ridge National Laboratory; and the Great Lakes Bioenergy Research Center (GLBRC) at the University of Wisconsin, Madison. By agreement with DOE, the BRC projects are afforded top priority for sequencing and analysis at DOE JGI.

After the BRCs, the DOE JGI user community draws heavily from academic research institutions, along with the national laboratories, federal agencies, and a small number of companies. Worldwide, there are more than 1,800 unique collaborators on active projects. The broader user community — composed of investigators, collaborators, and annotators — is engaged in collaborations that are global in nature. In addition, DOE JGI's influence is extensive, considering that hundreds more researchers every year tap sequence data posted by DOE JGI on its numerous genome portals and at the National Center for Biotechnology Information's (NCBI) GenBank.




JGI
DOE JOINT GENOME INSTITUTE

User Meeting 5

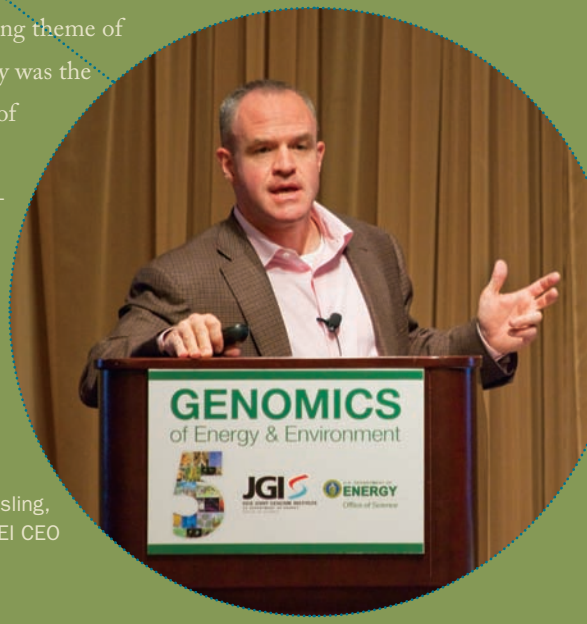
Several hundred researchers converged on the Walnut Creek Marriott March 24-26, 2010, to hear how genome sequencing can be applied to a number of DOE-relevant missions including the development of cellulosic biofuels and understanding and maintaining important biological and geological processes such as the carbon and nitrogen cycles for the environment.

During the first day of the Meeting, DOE JGI premiered a short video done in collaboration with the Ex'pression College for Digital Arts. The animated project explains how sequencing genomes of plants, fungi, microbes, and metagenomes can impact the quest to produce cellulosic biofuels.

The first keynote speaker was Jay Keasling, director of the Joint BioEnergy Institute, who compared the engineering of fuel-producing microbes to building a computer or synthesizing styrene. Using synthetic biology, he said, microbes can and are being developed to efficiently convert lignocellulosic biomass into replacement fuels for gasoline, diesel, and jet engines. He also stressed the need for a "systems biology" approach that can speed the development of advanced biofuels, integrating information from a multitude of different experiments on the assumption that in a system producing an advanced biofuel, every individual component of the system wields a certain degree of influence.

"Can we consolidate all the system components for production of an advanced biofuel into a bioprocessing microbe?" Keasling asked, answering immediately: "Yes, we can!"

The recurring theme of the second day was the development of biofuels, from feedstock production to the challenges faced in bringing



Jay Keasling, JBEI CEO

Rita Colwell, Stockholm Water Prize winner

Roger Pennell, Vice President, Ceres

the product to market. In the midst of this discussion, Roger Pennell, a vice president at the biotechnology company Ceres, delivered the day's keynote lecture on how genes and genomics can improve energy crops, and how these crops can in turn reduce the nation's dependence on fossil fuels.

Pennell noted that his colleagues have found that DNA function in one crop plant stays conserved across different crop species. They are working on harnessing genomic data to develop plants that can tolerate a variety of environmental stressors and produce high yields.

"There's no more fitting end to this meeting than having Rita Colwell talk on 'Solving Problems with Sequences,'" said DOE JGI's Len Pennacchio as he introduced the final keynote speaker. Colwell, announced as the 2010 Stockholm Water Prize recipient earlier in the week, closed the User Meeting with a talk about her decades of research on cholera. She spoke of the techniques applied to her work, ranging from simple, low-cost water purification methods, to sequencing the genome of the disease-causing bacterium *Vibrio cholerae* to better understand the disease-causing microorganism, and finally to the use of bioinformatics and global tracking systems to monitor the spread of these microorganisms in the oceans through influences such as various weather patterns and climate change.

Videos of the 2010 User Meeting were posted on the DOE JGI's SciVee channel: http://bit.ly/aOtKDZ.

A list of the 2010 DOE JGI Annual Genomics of Energy and Environment User Meeting speakers can be found in Appendix E.

CSP Portfolio heralds shift to large-scale, complex projects

For the 2011 Community Sequencing Program call, researchers were invited to submit proposals for projects that advance capabilities in fields such as large-scale resequencing, single-cell genomics, and metatranscriptomics, while remaining relevant to the DOE missions of bioenergy, carbon cycling, and biogeochemistry.

"Our mission hasn't changed; what is changing is the scale and complexity of the projects — which will increase," said DOE JGI Director Eddy Rubin. "In the past year alone, the JGI's sequence output has increased fivefold to 5 terabytes or 5 trillion nucleotides. Connected with the increased productivity, we're beginning to position the

JGI to work on projects of tera- and petabase scale. This highlights one of the directions genomics is going as data output begins to rival the output of the high-energy physics and astronomy communities."

A total of 35 new genomic sequencing projects, composed mostly of large-scale projects, was selected for DOE JGI's Community Sequencing Program (CSP) to be characterized for bioenergy and environmental applications.

Of the approved proposals, two involve plant genomes and two involve algal genomes; 10 are fungal projects; nine are microbial projects (six of them involve single-cell genomics); and 12 are either metagenome (microbial communities) or metatranscriptome projects. The projects make the most of DOE JGI's increased sequencing capacity, allocating 10 terabases, a 30-fold increase compared with last year's one-third of a trillion nucleotides.

Researchers at DOE JGI have been among the pioneers of the methodology known as single-cell genomics, in which the DNA isolated from a lone cell is amplified, allowing researchers to study the genomes of organisms that have not or cannot be cultured in a laboratory setting. This is a critically important capability, as it is well known among microbiologists that 99 percent of the microbial world is difficult to grow in culture and thus very difficult to study. Two-thirds of the approved CSP 2011 microbial projects involve the use of single-cell genomics to learn more about uncultured microbes found in ecosystems such as deep-sea hydrothermal vents and terrestrial subsurface aquifers.

Metatranscriptomics focuses on the complex region of the complete genetic code that is transcribed into RNA molecules and provides information on gene expression and gene function. Half of the metagenomic projects approved also involve plans to conduct metatranscriptomic studies.

Plant and Algal Projects

Barley (Dag Endresen)

Barley

The single largest project in the CSP 2011 portfolio, and DOE JGI's largest project to date, is the genome of barley, with an anticipated 5-billion base genome (67 percent larger than the human genome). Proposed by Gary Muehlbauer of the University of Minnesota, barley ranks fifth in the world among all crops cultivated and is grown on 4 million acres in the United States alone. From an agricultural standpoint, barley has been extensively used in crop rotation, so farmers have the infrastructure and cultural knowledge needed to sustain cereal crop production on an annual basis. From a bioenergy perspective, the crop can be used to produce ethanol from the grain or for cellulosic ethanol from the straw. The genomic data will be useful for researchers wishing to conduct comparative genomic studies with other grasses such as wheat, rice, sorghum, and *Brachypodium*.

Resequencing Brachypodium distachyon

In 2010, the genome sequence of the wild grass *Brachypodium distachyon* was published in the journal *Nature*, to help researchers develop grasses tailored to serve as feedstocks for biofuel production. Under CSP 2011, 50 inbred lines of *Brachypodium* will be sequenced and compared with the previously sequenced genome. Led by John Vogel of the US Department of Agriculture Agricultural Research Service, the project will further develop genomic resources for *Brachypodium* and enable the more rapid identification of candidate genes involved in traits such as drought tolerance that are of interest to DOE.

B. distachyon (John Vogel, USDA-ARS)

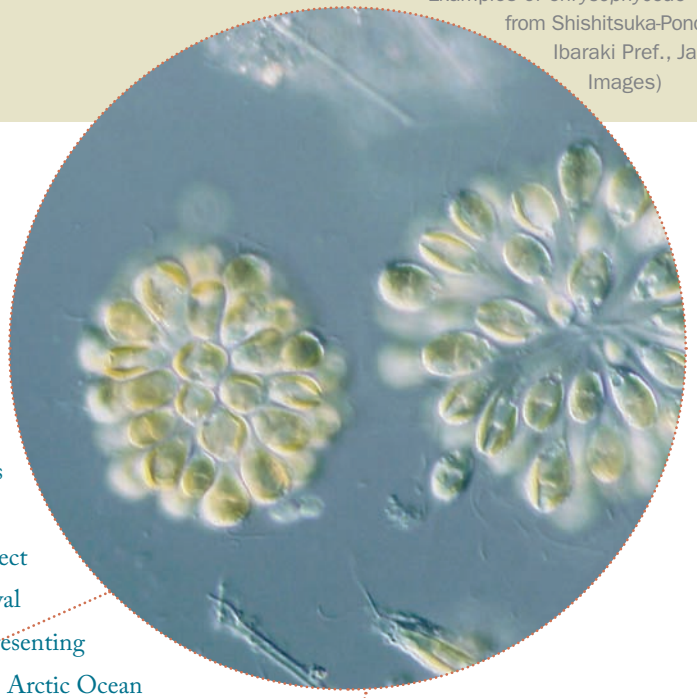
Analyses of Green Algal Strains

Algae contribute significantly toward the total global carbon sequestration, and they are a focus of attention as an alternate source of transportation fuels. The project involves conducting genomic and transcriptomic sequencing of green algal strains from isolated acidic waters to understand their role in the carbon cycle — specifically, how they can fix carbon. A noteworthy element is the project's plan for student participation, as a way to boost undergraduate education in genomics and informatics. The research team, which includes DOE JGI's Education head Cheryl Kerfeld, the 2011 recipient of the American Society for Biochemistry and Molecular Biology Education Award, intends to engage undergraduates in bioinformatics using the data collected from this project.

Acid water lake in abandoned mine

Planktonic Eukaryotes

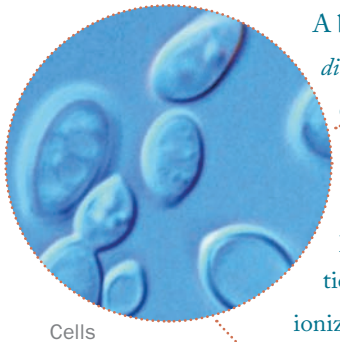
Marine phytoplankton constitute less than 1 percent of photosynthetic biomass but account for approximately half of the global photosynthesis. As sea levels in the Arctic Ocean are changing, researchers are working to understand the impact on biological productivity. In this project proposed by Connie Lovejoy from Laval University in Canada, five protists representing different algal classes isolated from the Arctic Ocean are being investigated for adaptation to perennial cold conditions and for identification of genes linked to mixotrophy — how certain microorganisms can assimilate organic compounds as carbon sources. A second goal seeks to characterize the horizontal gene transfer events over evolutionary time of donor algal genes to the original host protozoa. A third goal will be to link pigment and lipid signatures with coding genes and transcripts. This information will have downstream applicability for research on algal biofuels as alternative energy sources.



Examples of *Chrysophyceae–Synura* spp. from Shishitsuka-Pond, Tsuchiura, Ibaraki Pref., Japan (NEON/ Images)

Fungal Projects

Aureobasidium pullulans



Cells of *Aureobasidium pullulans* var. *pullulans*. (Nina Gunde-Cimerman and Cene Gostincar, University of Ljubljana, Slovenia)

A black, yeast-like fungus, *Aureobasidium pullulans* thrives in a variety of environments and has been found on plants, PVC pipes, and even on the walls of the Chernobyl nuclear power plant. Its genomic information could help researchers studying ionizing radiation to develop new methods of assessing radiation effects. Additionally, as recent studies indicate, a strain of this fungus has been found in Arctic glaciers, and researchers hope to learn more about its role in the carbon cycle, especially as these glacial habitats are being affected by climate change. The project proposed by Martina Turk and Nina Gunde-Cimerman of the University of Ljubljana in Slovenia calls for sequencing all four *A. pullulans* varieties. From a bioenergy perspective, the fungus may hold enzymes that can help break down biomass, underscoring the metabolic versatility of fungi and relevance to the DOE bioenergy mission. Identifying genes involved in allowing the fungus to tolerate a number of environmental conditions could also lead to the development of drought-tolerant and salt-tolerant crops.

Comparative Transcriptomics Fungal Pipeline

Fasciola hepatica (J. Berkovec/ Wikimedia Commons)

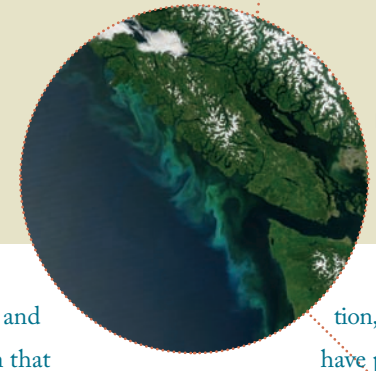
Two general strategies are employed by fungi in breaking down lignin. One used by white rot fungi involves breaking down the lignin so that other enzymes or organisms can break down the cellulose. A second process, used by brown rot fungi, modifies but does not remove the lignin, and focuses on degrading the cellulose. Genomic information is now available for several white rot and brown rot fungi but because of a lack of standard growth conditions, and the use of various technological platforms, researchers are having difficulty conducting intraspecific comparative analyses. To remedy this, a team led by Antonio Pisabarro of Spain's Public University of Navarre proposes developing a comparative transcriptomics pipeline. The pilot project involves a dozen brown rot and white rot fungi whose genomes either have been or are in the process of being sequenced by DOE JGI for their relevance to bioenergy and carbon cycling. Pisabarro and his collaborators want to improve the process by which gene expressions and gene functions of these fungi are compared.

Diversity of Mycorrhizal Fungi

Building on previous work conducted on the *Laccaria bicolor* genome and those of other symbiotic fungi, this project will sequence a phylogenetically and ecologically diverse suite of mycorrhizal (plant root-associated) fungi, which include the major clades of symbiotic species associated with trees and woody shrubs. Analyses of these genomes will provide insight into the diversity of mechanisms for the mycorrhizal symbiosis and those responsible for promoting growth and health of domesticated trees for bioenergy. Complementing another CSP 2011 project devoted to sequencing fungal wood decayers, research into these genomes will also illuminate the functional basis of transitions between decayer and symbiotic lifestyles. The taxa proposed here include many representatives of major clades for which no genome data are currently available.

Rozella allomycis (Timothey Jamey/Wikimedia Commons)

Microbial Projects



Phytoplankton bloom off Vancouver Island, Canada. (NASA, MODIS Rapid Response Team)

Sulfidic Aquifer Subsurface Biofilms

Microoxic zone in Italy's Frasassi cave system. (Copyright Macalady Group (PSU) and Coldigioco Café Science Consortium)

The terrestrial subsurface remains one of the least explored microbial habitats on Earth, and is critical for understanding pollutant migration and attenuation, subsurface processes such as limestone dissolution (affecting porosity), and the processing of carbon and nitrogen compounds relevant to global processes (e.g., climate). Sulfur is one of the 10 most abundant elements on Earth, and its complex redox chemistry and tendency to polymerize have lead to a comparably complex array of microbial sulfur metabolisms linking carbon, sulfur, nitrogen, and iron in biogeochemical cycles. The deep and sulfidic Frasassi aquifer (of Ancona, Italy) has emerged as a model system for studying sulfur cycling in the terrestrial subsurface. In addition, this sequencing project by Jennifer Macalady of Penn State University has relevance for understanding exoelectrogens (microorganisms that can deposit electrons onto solids or molecules outside their cell membranes). The survey may also yield applications for wastewater treatment and associated fuel cell bioengineering, as well as revealing capabilities relevant for radionuclide, metal, and organic pollutant remediation that can be applied to environments at DOE subsurface sites.

Microbial Systems Ecology of Expanding Oxygen Minimum Zones

Oxygen minimum zones (OMZs) are widespread oceanographic features expanding due to global warming. Although inhospitable to aerobic life forms, they support a thriving microbiota whose combined metabolic activity results in biological nitrogen loss and greenhouse gas production. There is increasing evidence that ocean warming trends will decrease dissolved oxygen concentrations, causing hypoxic boundary layer expansion. To properly diagnose or mitigate these transitions, this project from Steven Hallam of Canada's University of British Columbia launches a systems-level investigation of microbial community responses to OMZ expansion, charting the gene expression patterns of indigenous microbial communities found in coastal and open ocean OMZs in the eastern Subarctic Pacific Ocean as part of an ongoing time series program monitoring microbial community responses to changing levels of water column oxygen deficiency. The results will enable the development of *in situ* monitoring tools and probabilistic models for intrinsic ocean processes. In addition, the metabolic processes unfolding within these OMZs have potential application in the development of biological energy alternatives and wastewater treatment processes.

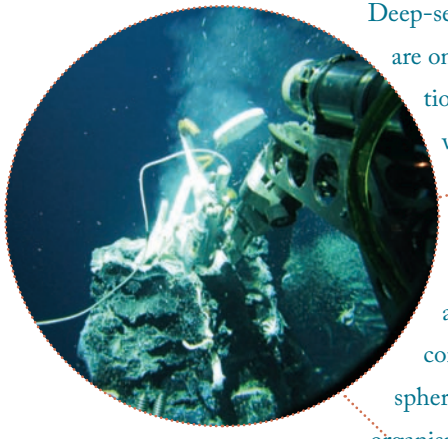
Microbial Diversity of Etoliko Lagoon

Located in a 1,700-hectare wetland area of western Greece, the Etoliko Lagoon essentially supports two ecosystems by having both an oxygen-rich area and a distinct oxygen-poor zone, which is noted for having increased levels of sulfides and salinity. The lagoon has high concentrations of methane and sulfides. This CSP project from George Tsiamis of Greece's University of Ioannina calls for sequencing nearly two-dozen microbes to learn more about the rich microbial diversity in the lagoon, and the genomic information collected will also enrich DOE JGI's ongoing GEBA project. Sequencing the microbes found in the marine sediment offers researchers the opportunity to learn more about anaerobic methane oxidation, a globally important process that plays a key role in reducing methane flux from marine sediments to the atmosphere, and how methane-oxidizing archaea and sulfate-reducing bacteria coexist.

Etoliko Lagoon, Greece. (Kakouris Spyridon/SpirosK photography, www.flickr.com/photos/spirosk)

Metagenomic Projects

Hydrothermal Vent Microbial Community



Sampling at the hydrothermal diffuse vent site 'Crab Spa' during a research cruise in October 2008. (Stefan Sievert, Woods Hole Oceanographic Institution)

Deep-sea hydrothermal vents are one of the premier locations for discovering a wide diversity of novel nonphotosynthetic microorganisms capable of fixing carbon dioxide at the most extreme conditions within our biosphere. Hydrothermal organisms utilize all six of the known biochemical pathways for carbon dioxide fixation, providing a wide array of mechanisms for inorganic carbon conversion. This project from Peter Girguis of Harvard University represents the first comprehensive metagenomic survey of hydrothermal vent microbial communities from three hydrothermal vent observatories and two sediment-covered hydrothermal systems. Among the goals is to develop an extensive metagenomic database to enable the first global assessment of community functional potential, ecology and evolution at deep-sea hydrothermal vents. This database will provide an inventory of novel gene sequences that will aid research in renewable energy, carbon and heavy metal sequestration, and environmental remediation. Enzymes sourced from these microorganisms may also serve to convert feedstocks — ranging from hydrogen and carbon dioxide gases to agricultural products — into energy dense transportation fuels.

Complete Shipworm Microbiome

As a wood-boring bivalve, the shipworm has two bacterial populations that can break down cellulose — one in the gut and the other in a specialized organ in the gills. DOE JGI sequenced the only shipworm species adapted to cold water — *Bankia setacea* — as part of the CSP 2009 portfolio to identify the enzymes in these microbes involved in breaking down wood for cellulosic biofuel production. Now DOE JGI collaborator Daniel Distel of the Ocean Genome Legacy Foundation has proposed studying both microbial communities in different species of this "termite of the sea," including *Lyrodus pedicellatus,* a species adapted to warm water. Wood-boring bivalves are the only marine animals known to sustain normal growth and reproduction feasting solely on wood.

L. pedicellatus shipworm (Damon Tighe, DOE JGI)

Microbial mats in Octopus Spring, Yellowstone National Park. (David Strong)

Yellowstone Hot Spring Metagenome Analysis

Chlorophototrophic organisms use photochemical reaction centers to convert light energy into chemical energy. The chlorophototrophic microbial mats of alkaline siliceous hot springs of Yellowstone National Park (YNP) have been studied for decades as models for understanding the composition, structure and function of microbial communities. As proposed by Don Bryant of Penn State University, a deep metagenomic sequencing of six microbial mat communities in YNP will provide

detailed analyses of community compositions and identification of major metabolic activities. These communities are of interest because of their ability to capture and efficiently convert solar energy into chemical/biological energy; the important biogeochemical cycling that occurs in these communities; their autotrophic carbon fixation and carbon recycling; and because they offer an opportunity to identify novel microorganisms that perform all of the above functions. They may also provide useful information on bioremediation because of the presence of toxic chemicals in these springs, and the mechanisms and evolutionary origins of photosynthesis.

Metagenomes of Soda Lakes and Salt Marsh Soils

The microbial diversity of soda lakes has been show to be high compared with their opposite counterparts, the acidophilic microbial communities, which consist of a few dominating populations. Proposed by Gerald Muyzer of the Netherlands' Delft University of Technology, sequencing the microbial communities that inhabit soda lakes and soda solonchak (salt marsh) soils will provide insight into the molecular mechanisms for adaptation to high pH and high salinity, the carbon dioxide-uptake mechanisms, and the pathways that generate energy from sulfur compounds. In addition, the results should inform whether there are enzymes present that can be used for biotechnological purposes, such as enzymes that can hydrolyze lignocellulose and other biopolymers (e.g., pectin, xylan, and chitin), which can be used as novel feedstocks for the production of biofuels or specialty chemicals.

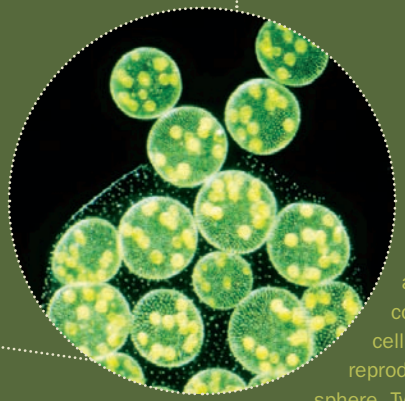
Soda Lake in Kenya's Rift Valley, East Africa

For the complete list of CSP 2011 sequencing projects, see: http://www.jgi.doe.gov/sequencing/cspseqplans2011.html

Switchgrass flower (Jeremy Schmutz, HudsonAlpha Institute for Biotechnology)

DOE JGI Programs

Plant Genomics Program



Juvenile spheroids of the green alga, *Volvox carteri*, hatching from a parental spheroid. Each juvenile contains about 2,000 tiny somatic cells at the surface, and 16 large reproductive cells (gonidia) deeper in the sphere. Two days after their birth, each of the juvenile gonidia will divide to produce a new juvenile spheroid containing the same two cell types. (David Kirk, WUSTL)

Genome sequences, assemblies, and gene annotations of key plant and algal species are generated at DOE JGI and interrogated to determine how the information can be applied toward accelerating biofuels development and understanding adaptation to climate change.

One of the genomes released in the past year was that of the soybean (*Glycine max*), which appeared in the January 14, 2010, issue of *Nature*. This first legume species to be sequenced was also the largest-to-date completed at DOE JGI. The soybean, whose genomic information is used to study nitrogen fixation, is also the primary source of biodiesel production in the United States. A video of Jeremy Schmutz — the study's first author and a DOE JGI scientist at the HudsonAlpha Institute for Biotechnology in Alabama — discussing the soybean genome project can be viewed online at http://bit.ly/fWUc82.

In the February 12, 2010, issue of *Nature*, the International Brachypodium Initiative, a consortium that includes researchers from DOE JGI, published the complete sequence of the wild grass *Brachypodium distachyon*. With this genome, researchers have been able to conduct comparative genomic analyses across all three major grass subfamilies to learn more about specific biological plant traits that could improve their usefulness as candidate energy feedstocks toward developing cellulosic biofuels. A video of study lead author and DOE JGI collaborator John Vogel of the U.S. Department of Agriculture (USDA) Agricultural Research Service (ARS) discussing the *Brachypodium* genome project is online at http://bit.ly/awNYWW.

Algae are another potential source of carbon-neutral transportation fuel and in the July 9, 2010, issue of *Science*, researchers led by DOE JGI and the Salk Institute reported the 138 million nucleotide genome of the multicellular alga *Volvox carteri.*

The work complements the genome of the unicellular alga *Chlamydomonas reinhardtii,* which was sequenced by DOE JGI in 2007 and is used by researchers working on algal biofuel generation. The *Volvox* genome could have similar applications.

DOE JGI collaborator and co-first author Jim Umen at the Salk Institute noted *Volvox* can shift its energies from producing food through photosynthesis to supporting the functions of its non-reproducing somatic cells.

"While we don't yet understand this trait well, it could factor into how photosynthetic organisms can be engineered to do what we want, such as make biofuels or other products, rather than what they typically do, which is grow and make more of themselves," he said.

DOE JGI bioinformaticist and co-first author Simon Prochnik added that having the genomes of both *Volvox* and *Chlamydomonas* has allowed researchers to identify roughly 1,800 protein families unique to both algae. The genomes also provide the opportunity to conduct comparative analyses to learn more about photosynthetic processes and the transition from unicellularity to multicellularity.

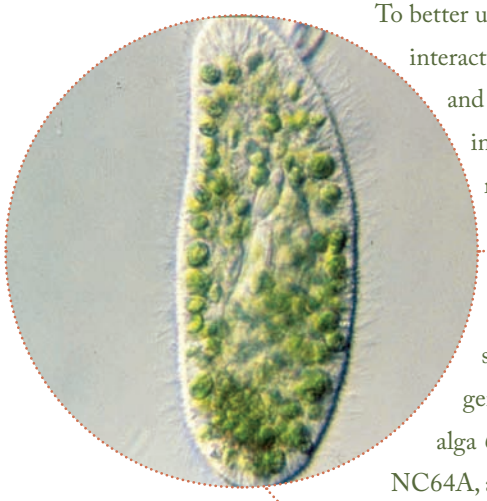
"Having the *Volvox* genome is a fantastic resource for directing further research towards our target areas of interest," he said. "With this pair of algal genomes in hand, it enables us to conduct much more detailed comparisons than would be possible if we only had one species."

An additional benefit to having two algal genomes for comparison, he added, is the ability to study how a single-cell ancestor evolved complex cellular processes in a short evolutionary period. What the team found, he said, is "an astonishing lack of innovation" in the *Volvox* genome when compared with *Chlamydomonas*, particularly given their completely different morphologies. "The notion

that 'if you're small, you're simple' is starting to unravel. The more unicellular organisms we sequence, the more we see this."

A video of Prochnik discussing the *Volvox* genome project is available on YouTube at http://bit.ly/ffXbTp and on SciVee at http://bit.ly/9kqePh.

Green algae are key components of the global carbon cycle and from a bioenergy perspective, algae are increasingly viewed as a viable feedstock for biofuel and biodiesel production because of their high lipid content. However, algal viruses can infect up to a fifth of all algae at any one time and impact their productivity.



Paramecium bursaria with symbiotic *Chlorella* cells. (Courtesy of Jim van Etten, University of Nebraska-Lincoln)

To better understand the interactions between algae and their viruses, an international team of researchers including DOE JGI Fungal Genomics head Igor Grigoriev has sequenced the genome of the microalga *Chlorella variabilis* NC64A, a model system for DNA virus/alga interactions.

"Like other microalgae, there is an increasing interest in using *Chlorella* in a variety of biotechnological applications," wrote the team in the study published in the September 2010 issue of *The Plant Cell*. "The sequence of the NC64A genome will help in the optimization of these various processes."

The 46-million base genome, sequenced using the whole genome shotgun Sanger approach, revealed that the alga, long thought to be asexual, might engage in sexual reproduction after all. Another interesting finding involves *Chlorella's* rigid cell walls, which are built with chitin, the same material used by crustaceans for their hard exoskeletons, and might have come from "the capture of metabolic genes by horizontal gene transfer from either algal viruses, prokaryotes or fungi," writes the team.

Aside from studying candidates for bioenergy production, DOE JGI is also sequencing plants and algae that can elucidate their roles in carbon sequestration and how they cope with toxic pollutants and environmental change. For example, in November 2010, after the annotated cassava genome was released on the portal, the Bill and Melinda Gates Foundation announced a $1.3 million grant to fund the development of a genome variation database to grow more disease-resistant and nutritious varieties of the plant. Cassava is a staple food for more than 750 million people around the world and a third of the annual harvest currently is lost due to pathogens such as cassava brown streak disease.

The cassava data set is one of more than two dozen plant genomes publicly available on Phytozome (http://phytozome.net), a portal that provides access to integrated, cross species, annotated genomic data as well as tools for analysis. Among the new additions are the first two citrus genomes, sweet orange (*Citrus sinensis*) and Clementine mandarin (*Citrus clementina*).

Fred Gmitter, Jr., head of the International Citrus Genomics Consortium (ICGC) and a citrus geneticist at the University of Florida, announced the availability of the annotated genome sequences during the Plant and Animal Genome conference held in January 2011.

"Citrus is the most economically significant and widely grown fruit crop in the world," said Gmitter, who added that in the United States alone, the citrus industry is worth $20 billion annually. "As more plant genomes find their way to Phytozome, it is becoming more widely recognized as a first source for genome researchers looking to exploit genome sequences in their research."

The citrus genomes join two dozen other plant genomes, many of which were sequenced at DOE JGI for their potential applications to DOE-relevant missions. Another late-breaking initial release on Phytozome is the

The *Citrus* genome analysis will facilitate comparative genomics research and be particularly valuable for comparative studies with poplar and other trees.

8x mapped *Eucalyptus grandis* BRASUZ1 genome assembly and a preliminary annotation. *Eucalyptus* is listed as one of the DOE's candidate biomass energy crops. Genome sequencing is essential for understanding the basis of its superior growth properties and to extend these attributes to other species. Genomics will also allow researchers to adapt *Eucalyptus* trees for green energy production in regions (such as the Southeastern United States) where it cannot currently be grown. The unique evolutionary history, keystone ecological status, and adaptation to marginal sites make *Eucalyptus* an excellent focus for expanding our knowledge of the evolution and adaptive biology of perennial trees.

Eucalyptus species are among the fastest-growing woody plants in the world, with mean annual increments up to 100 cubic meters per hectare. *Eucalyptus* is the most valuable and most widely planted genus of plantation forest trees in the world (approximately 18 million hectares) due to its wide adaptability, extremely fast growth rate, good form, and excellent wood and fiber properties.

Eucalyptus is one of DOE's candidate biomass energy crops, due to its wide adaptability, extremely fast growth rate, good form, and excellent wood and fiber properties. (Roy Kaltschmidt, Berkeley Lab)

Fungal Genomics Program

Turkey-tail polyplore *(Trametes versicolor)* is a prime wood decomposer with enzymes that could be useful for cellulosic biofuel production. It has also been used to break down organic pollutants.

Built on established expertise in genomics, new sequencing technologies, and strong connections with user communities, DOE JGI's Fungal Genomics Program aims to scale up the sequencing and analysis of fungal genomes for DOE mission areas and to develop the Genomic Encyclopedia of Fungi. Understanding molecular mechanisms of interactions between plants and fungi — both symbionts and pathogens — is essential for the health and productivity of biofuel feedstock crops. Reference genomes of mycorrhiza and other soil-inhabiting fungi will also facilitate comprehensive metagenomics studies of the rhizosphere. Combining new sequencing technologies and comparative genomics analysis allows the Fungal Program to address large and complex projects such as surveying the broad phylogenetic and ecological diversity of fungi and capturing genomic variation in natural populations and engineered industrial production strains.

In its first year, the DOE JGI Fungal Genomics Program doubled the number of newly sequenced and fungal genomes compared with the previous year's work. In addition, many previously released genomes were improved on both the assembly and annotation levels. To scale up sequencing and analysis of fungal diversity for DOE science and application, different sampling techniques were combined, from a broad sampling of the fungal Tree of Life to the resequencing of mutant strains of model fungus *Neurospora crassa*, with focused sampling of groups of lignocellulose degraders (*Agaricomycotina*), sugar fermenters (*Saccharomycotina*), and feedstock plant pathogens (Dothideomycetes). These are the focus areas of the Genomic Encyclopedia of Fungi initiated this year.

Among the notable projects from the Fungal Genomics Program in 2010 was the release of the second white rot fungal genome, that of *Schizophyllum commune*. Many sequencing projects at DOE JGI focus on identifying enzymes in organisms such as fungi that can break down cellulose in plant mass to help bring down the cost of cellulosic biofuel production.

As DOE JGI Fungal Genomics Program head Igor Grigoriev noted, "When we go into a forest, we don't see layers of dead branches because wood decay fungi take care of them. So when we think about bioenergy and degrading biomass and converting that into biofuel, we would like to learn the most efficient ways of doing that from fungi, which have invented many ways of doing that in nature."

In the September 2010 issue of *Nature Biotechnology,* Grigoriev and colleagues reported on the 38.5 million-base genome of the white rot fungus *Schizophyllum commune*, commonly known as the split-gill fungus. *S. commune* is the third wood-decaying fungus to be sequenced by DOE JGI; the first white rot fungus sequenced was *Phanerochaete chrysosporium* in 2004 and the first brown rot fungus, *Postia placenta*, was sequenced in 2009.

Found on every continent except Antarctica, *S. commune* breaks down cellulose and lignin by invading xylem tissue, and researchers hope that studying its

Schizophyllum commune
(Carla Wick)

genome will help them harness the most relevant set of enzymes for specific biofuel production strategies. Additionally, white rot fungi have potential bioremediation applications, as they have enzymes that can break down contaminants such as uranium and heavy metals.

Grigoriev said DOE JGI has 40 fungal genome sequencing projects in its queue, and more than a dozen of them involve wood-decaying fungi. "We think we're only touching the surface and we need to look at more genomes in order to understand the whole scope of diversity and mechanisms applied to degrading cellulose."

According to the Genomes OnLine Database, DOE JGI is responsible for more than a third of all fungal genomes sequenced or in queue to be sequenced worldwide.

A video of Grigoriev discussing the *S. commune* project is available on YouTube at http://bit.ly/htfVzc and on SciVee at http://bit.ly/huDRUf.

In the June 2010 issue of *Applied and Environmental Microbiology,* a team of DOE JGI researchers and the USDA Forest Service Forest Products Lab published a comparative analysis of transcript and protein expression of genes that code for secreted proteins from the first two sequenced wood-decaying fungi: the brown rot fungus *Postia placenta* and the white rot basidiomycete *Phanerochaete chrysosporium.*

And in the November 2009 issue of *BMC Genomics*, researchers from the Vienna University of Technology and DOE JGI reported on the mycoparasitic fungus *Trichoderma atroviride,* which could be used in place of chemical fungicides to counter plant pathogens.

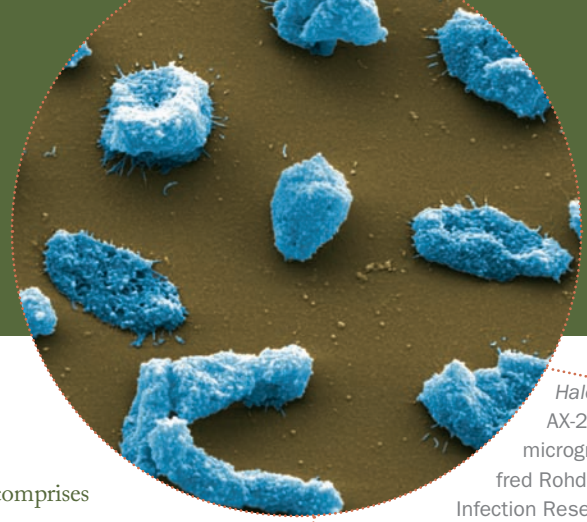
Moving forward, the Program is focused on comparative analysis of groups of similar fungi rather than individual genomes to elucidate characteristics of these groups. This is reflected in the 2011 Community Sequencing Program projects such as comparative genomics and transcriptomics of several species of brown rot and white rot fungi, yeasts and *Aspergilli* for biotechnology, exploring genome diversity of mycorrhizae, sequencing pathogenic Dothideomycetes, and others.

MycoCosm (jgi.doe.gov/fungi), a genomics resource for fungal biologists, was launched in March 2010 and offers researchers access to more than 50 newly sequenced and annotated fungal genomes based on work done at DOE JGI and elsewhere. The portal was developed in response to the fungal community's call for one-stop shopping place for fungal genomics data and tools.

MycoCosm offers Web-based genome analysis tools for fungal biologists to navigate through sequenced genomes, explore them in context of comparative genomics and genome-centric analysis, and integrate genomics data from JGI and its users. It also promotes user community participation in data submission, annotation, and analysis. The portal is a home for genomics and other -omics data for fungi that are important for energy and environment, and has become a community data hub equipped with tools for genome-centric analysis and comparative genomics.

These tools are promoted through JGI user training and jamborees, which also focus on genomic analysis. For example, one jamboree in FY10 focused on properties of thermophilic fungi important for production of cellulolytic enzymeswhile another explored the evolution of different Basidiomycete lifestyles (saprotrophs, pathogens, and symbionts) in order to gain a more thorough understanding of the fungal contribution toward sustainable growth of bioenergy crops.

Microbial Genomics Program



Halorhabdus utahensis AX-2T (Scanning electron micrograph provided by Manfred Rohde, Helmholtz Centre for Infection Research, Braunschweig)

During Fiscal Year 2010, the Program sequenced and finished 151 bacterial/archaeal genomes, 60 of which were represented in publications. With the release of these genomes into GenBank, the number of DOE JGI microbial finished genomes is rapidly approaching 500.

The Program has been expanding its product catalog beyond a finished microbial genome or genome resequencing and has adopted the generation of draft genomes (isolate draft genomes and single-cell genomes) as well as transcriptome sequence. This expansion and the increase in microbial genome products is going hand-in-hand with and has been stimulated by new high-throughput technologies and capabilities, such as *de novo* microbial Illumina assemblies and single-cell genomics. The increased throughput supports the user community by enabling DOE-relevant science at a grander scale.

As the vast majority of microbes are uncultured to date, single-cell genomics has been a Program focus to enable not only JGI science but also DOE user community proposed single-cell research. As published in the April 23, 2010, issue of *PLoS ONE*, the Program generated one complete genome of the uncultured symbiont Candidatus *Sulcia muelleri* DMIN of the green sharpshooter by amplifying and sequencing the genome from a single cell; the first complete single-cell genome from any organisms. Six highly DOE-relevant community sequencing projects involving single-cell genome sequencing are currently under way.

The Microbial Genomics Program comprises the Genome Biology Program, Phylogenomics, the Structural Genomics and Single Cell Genomics Groups at DOE JGI, as well as groups at Los Alamos National Laboratory and Oak Ridge National Laboratory. Showcasing these collaborations, a report published online October 1, 2010, in the *Journal of Bacteriology* reported that 20 species of *Clostridia* bacteria have been sequenced at DOE JGI. The bacteria contain enzyme complexes involved in breaking down biomass, which could be useful for commercializing cellulosic biofuel production. In keeping with DOE JGI's goals of providing users with services beyond sequencing, all draft genomes were annotated through the JGI-ORNL annotation pipeline and then analyzed using the JGI Integrated Microbial Genomes data management system.

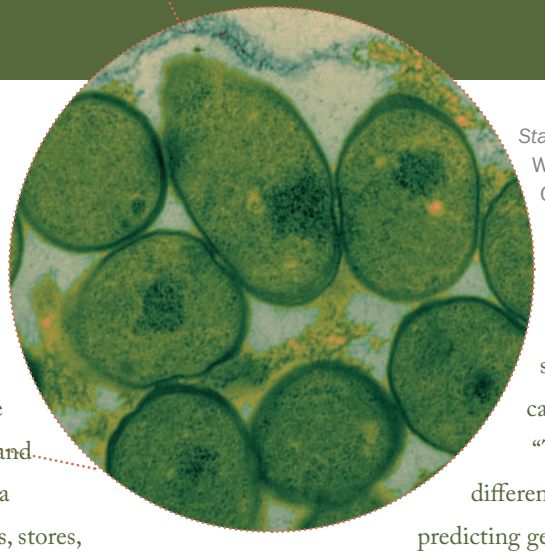
One of the ongoing MGP Initiatives is the continuation of the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project, which aspires to sequence thousands of bacterial and archaeal genomes from diverse branches of the Tree of Life. A pilot GEBA study released December 2009 in the Journal *Nature* reported that 53 bacterial and three archaeal novel and highly diverse genomes had been sequenced.

Another Microbial Program flagship project involves the sequencing of 100 strains of nitrogen-fixing bacteria. Agricultural productivity is heavily dependent on nitrogen, so the symbiotic interactions between nitrogen-fixing bacteria and crops are of interest to researchers around the world.

"The value here (in addition to the obvious scientific value)," said Metagenome Program lead Nikos Kyrpides, "is that we formed a partnership with a consortium of 50-plus scientists, coordinated by Wayne Reeve of Australia's Murdoch University, which will be sending the samples — we're expecting the first samples — and will help us in the analysis."

Rhizobial bacteria are found in the soil and interact with legumes such as peas and clover at the root nodules. The bacteria fix atmospheric nitrogen inside the nodules and contribute nearly two-thirds of the nitrogen used in agricultural production.

The team plans to look at rhizobial genomes collected from distinct geographical regions. Reeve noted that this project is the first large-scale attempt to understand the genetics involved in the plant-bacteria interactions.



Starkeya novella (Rick Webb, University of Queensland)

"I consider this project one of the Microbial Program flagship projects," said Microbial Genome Program lead Tanja Woyke, who also noted that the data could eventually be linked to the Rhizosphere Grand Challenge.

Several tools are available for visualizing and analyzing genomic data on the Microbial Genome Program site at http://genome.jgi-psf.org/programs/bacteria-archaea/index.jsf. One of these is the Integrated Microbial Genomes (IMG) system, which 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.

A recently developed and novel tool is GenePRIMP (Gene PRediction IMprovement Pipeline), a quality control pipeline that consists of a series of computational units that identify erroneous gene calls and missed genes and correct a subset of the identified defective features. More than 1,000 microbial genomes have been sequenced at various sequencing centers in the past 15 years, and the number is expected to increase at least tenfold within the next two years. While the data collected on microbial genomics is increasing, genomic standards have not caught up with the technological advances that have made the sequencing process faster and cheaper. As a result, the DNA sequences being released are of varying levels of quality, impacting researchers' ability to use this information.

As described in the June 7, 2010, issue of *Nature Methods*, the input to GenePRIMP needs to be a file of gene cells in GenBank or EMBL format.

"Consistent high-quality annotation on microbial genomes is key to their utility," said Owen White, director of bioinformatics at the University of Maryland School of Medicine and head of the Human Microbiome Project Data Analysis and Coordination Center that tracks, stores, analyzes, and distributes the data. "Software such as GenePRIMP is an important component in our quality control toolbox."

First author Amrita Pati, a software developer in the DOE JGI's Genome Biology Program, noted that GenePRIMP significantly reduces the amount of time scientists spend checking the whole genome by specifically highlighting errors that need to be manually corrected.

"With GenePRIMP, we have achieved a major breakthrough in the improvement of the quality of structural annotations such as gene predictions," said Genome Biology Program head and study senior author Nikos Kyrpides. He pointed out that using GenePRIMP offers researchers three major advantages: high-quality results with reduced errors; an approach that can be used regardless of the automated software originally used to check gene annotations; and a method to standardize gene calling.

"There are a lot of different tools used for predicting genes in prokaryotes," Kyrpides said. "The major problem we have is that they all produce very variable results. This impedes our ability to compare genomes sequenced and annotated from various sources, as they use different tools for gene prediction. GenePRIMP is not substituting for any of the available methods; a user can employ any available automatic gene prediction method, and then use GenePRIMP to correct the initial output. It will generate a much more standardized output, thus not only significantly improving quality, but also significantly facilitating comparative analysis."

GenePRIMP is available for use by researchers at http://geneprimp.jgi-psf.org/.

Study first author Pati is featured in a short video describing how the tool works on YouTube at http://bit.ly/bUG18f and on SciVee at http://bit.ly/awfUiY.

Metagenomics Program

Tammar wallaby (Mehgan Murphy, Smithsonian National Zoo)

A primary motivation for metagenomics is that most microbes found in nature exist in complex, interdependent communities and cannot readily be grown in isolation in the laboratory. One can, however, isolate DNA or RNA from the community as a whole, and studies of such communities have revealed a diversity of microbes far beyond those found in culture collections. In the DOE JGI's Metagenomics Program, the projects encompass multiple scientific goals, each based on sequencing a community of organisms, not a single isolate.

One such project focuses on the microbial communities in marsupials. Australia and New Zealand have been separate land masses for millennia, and the unique marsupials found there such as kangaroos and wallabies have adaptations and dietary preferences distinct from plant eaters anywhere else. Kangaroo and wallaby forestomachs are adapted to efficiently break down lignocellulosic plant mass to extract nutrients and the microbes in these forestomachs may be different from those found elsewhere.

To answer this question, in 2007 DOE JGI selected the foregut microbiome of the Tammar wallaby as one of the Community Sequencing Program projects.

In a metagenomic analysis published August 17, 2010, in the *Proceedings of the National Academy of Sciences*, a team that included DOE JGI's Susannah Tringe, Jan-Fang Cheng, and Phil Hugenholtz — now director of the Australian Center for Ecogenomics at the University of Queensland — looked at the plant biomass conversion process of the Tammar wallaby's foregut microbiome.

The wallaby microbiome DNA was sequenced using Sanger and 454 pyrosequencing, and the data was annotated using DOE JGI's Integrated Microbial Genomes for Metagenomes (IMG/M) system. Among their findings, the team identified unique bacterial lineages that break down drought-tolerant plants, which are common in Australia's native species. They said the data indicate that the enzymes in the Tammar wallaby and other Australian animals are unique and, in a nod to other DOE JGI metagenomic sequencing projects, added that "their repertoire … is distinct from those of the microbiomes of higher termites and the bovine rumen."

Another notable metagenomic publication in 2010 involved the microbiome of the leaf-cutter ant *Atta colombica*, done in collaboration with Cameron Currie and Garret Suen of the University of Wisconsin-Madison and the Great Lakes Bioenergy Research Center. Living in colonies of several millions, leaf-cutter ants are community gardeners on a very large scale. Each year, they harvest hundreds of pounds of leaves, using them to cultivate the fungal gardens that serve as their primary food source.

The fungal gardens, noted researchers including DOE JGI staff scientist Susannah Tringe in their paper published September 23, 2010, on *PLoS Genetics*, serve as the ants' external digestive system and have five separate layers to break down the plant biomass into nutrients. Currie's team took samples from the top and bottom layers of ant fungal gardens in Panama to better understand how the biomass is broken down at each stage. They knew that ants and the fungi they tend enjoy a mutualistic relationship, but that the fungi aren't capable of breaking down cellulose, which was found to decrease by 30 percent on average in these gardens.

The researchers identified a previously unknown microbial community involved in breaking down the

Leaf-cutter ant (Jarrod Scott, University of Wisconsin-Madison)

cellulose in these fungal gardens, and said the ants' ability to also cultivate these microbes helped establish these ants as "widespread, dominant insect herbivores" in the tropics of the Americas.

Another publication featured the first metagenomic analysis of syntrophic communities (wherein microbes live off the byproducts of others) in a wastewater stream. Terephthalate is the byproduct of a common compound used extensively by the plastics industry. The volume of terephthalate-containing wastewater generated is equivalent to the amount of wastewater generated by 20 million people.

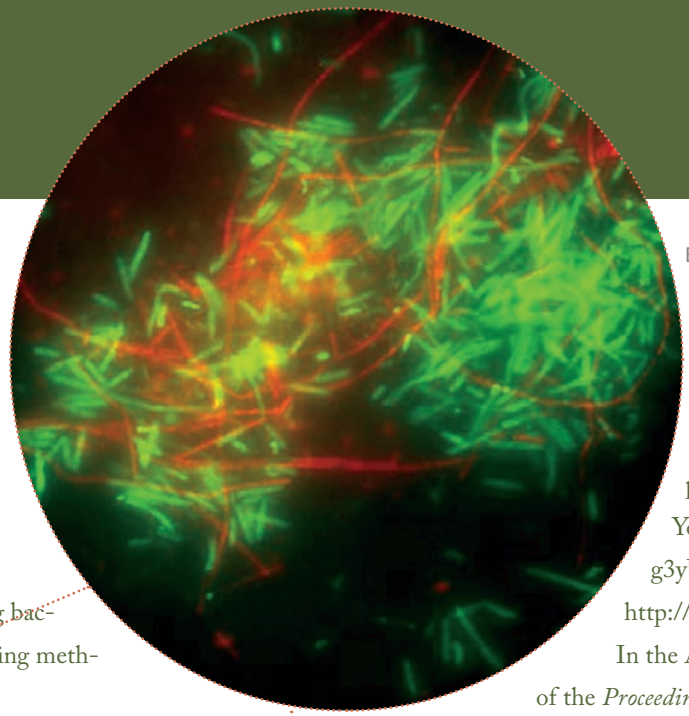
To remediate the wastewater, syntrophic communities composed of bacteria that help break down the organic matter, and methanogens that remove the hydrogen released so that this degradation process can continue, work at more than 100 terephthalate-degrading facilities worldwide.

DOE JGI selected a lab-scale, anaerobic, terephthalate-degrading bioreactor that operates at higher than normal temperatures as one of its 2006 Community Sequencing Program projects to better understand these communities. The project led to the publication of the first metagenomic analysis of the syntrophic communities in this methane-producing bioreactor, which was first published online August 5, 2010, in *The ISME Journal.*

"The whole process is more complicated than just having two populations," said DOE JGI research scientist and study first author Thanos Lykidis. "It reveals there are additional players in this phenomenon besides hydrogen-producing bacteria and hydrogen-consuming methanogens."

Lykidis said the researchers identified bacteria commonly found in sewage systems, and also gained insight into how the various populations interact with each other to allow the degradation process to proceed. The DOE JGI is involved in more than one-third of worldwide metagenomic sequencing projects, from terrestrial and aquatic to host-associated environments as they relate to the DOE's core interests.

Senior author Wen-Tso Liu of the University of Illinois said thermophilic bioreactors could help reduce the footprint of wastewater treatment facilities. "Since water is coming out at around 50° C, if we can have another microbial community that grows at 45-50° C and do the same job at the same degradation rate, why not do that? That's what this paper is about."



Biofilm in terephthalate wastewater (Liu Lab, University of Illinois)

A video of Lykidis and Liu discussing the project is available on YouTube at http://bit.ly/g3yUDc and on SciVee at http://bit.ly/dEK5wE.

In the August 17, 2010, issue of the *Proceedings of the National Academy of Sciences,* DOE JGI collaborator Alexandra Worden of the Monterey Bay Aquarium Research Institute and DOE JGI researchers reported the application of targeted metagenomics toward studying tiny members of one of the four major phytoplankton lineages. By understanding the role of the pico-prymnesiophytes, which make up a quarter of the global picophytoplankton biomass, researchers hope to get a better picture of marine photosynthesis and how the tiny eukaryotes play a role in the global carbon cycle.

Genomic studies of low-biomass environments are often limited by the amount of DNA available, and one solution has been to use a whole genome amplification technique that uses phi29 DNA polymerase known as multiple displacement amplification (MDA).

Microorganisms from compost are a promising source of novel carbohydrate-active enzymes discovery since many are able to degrade plant biomass.

When used in single-cell genomic studies, one noted drawback of the procedure has been amplification bias that impacts quantitative analyses. In a paper published in the December 2010 issue of *Nature Methods,* former DOE JGI researcher Phil Hugenholtz and his colleagues looked at whether similar amplification biases were detected in studying microbial communities.

Hugenholtz said the results indicated the bias should actually be predictable because the technical replicates were highly reproducible. This means that the possibility exists to correct for the bias so that MDA can be used for quantitative analysis of metagenomes.

MDA was first performed on environmental DNA samples extracted from garden compost, activated sludge, and termite hindgut, and biases were assessed using high-throughput small-subunit rRNA gene amplicon pyrosequencing. The team found that anywhere from 3 to 28 percent of the taxa detected in the samples were skewed, with relative abundance numbers unpredictably increased or reduced.

To complement the rapidly growing exploration of microbial communities, DOE JGI researchers have also called for a metagenome classification system. The sheer number of metagenomics data sets generated in the past few years through advances in sequencing technologies makes it difficult for researchers to track and access data.

In the July 2010 issue of the journal *Environmental Microbiology,* DOE JGI researchers noted that "if you were looking for metagenomes from organisms in the digestive tracts of various animals, they might be named gut but could also be 'rumen,' 'forestomach,' 'cecum,' or 'fecal'."

The current classification system only distinguishes between environmental or host-associated projects; the team proposed a five-tier system that starts with the same broad terms listed above and then goes on to narrower ecosystem definitions, allowing metagenomics researchers to conduct comparative analyses more efficiently, as well as better access the metagenomic data sets of interest.

Beyond metagenomic analyses, researchers interested in having metatranscriptomic analyses done have contacted DOE JGI to ask for recommendations on how to enrich the mRNA molecules. The transcriptome is the fraction of the complete genetic sequence that is copied (transcribed) into different types of RNA molecules that tell researchers what genes are turned on and off under various conditions, and hint at what the functions of these genes are.

While researchers focus their attentions on the messenger RNA (mRNA), which transfers the genetic information from the DNA to functional proteins, the bulk of cellular RNA is composed of ribosomal RNA (rRNA), which researchers prefer to remove before sequencing to make mRNA detection easier.

DOE JGI researchers led by Shaomei He, an Energy Biosciences Institute-funded postdoctoral researcher, validated two popular rRNA removal methods in the October 2010 issue of *Nature Methods* with the help of commercially available kits.

"When people use these methods, there's no systematic analysis to show if there's an associated bias," He said. "The two questions we wanted to answer in this project were: 1) how efficient are these methods at rRNA removal? and; 2) do they introduce any biases in mRNA?"

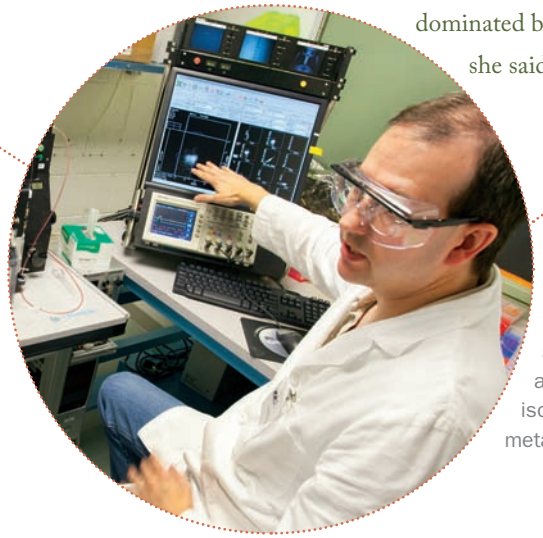
The team noted that both rRNA removal techniques have limitations. For example, said He, the hybridization kit uses generic probes that miss some bacteria and all archaea while the exonuclease kit also removes partially degraded mRNA. "Ultimately what you use depends on samples of interest and the biological question being

asked," she said.

The final outcome favors the use of the subtractive hybridization method by itself, which adequately preserved mRNA relative abundance for quantitative analyses, and He noted that researchers are already working on ways to improve the subtractive hybridization method. "The limitations of the commercial kit associated with its generic probes are likely to be overcome by a sample-specific hybridization approach recently developed by Ed DeLong's lab at MIT, especially if the community is dominated by archaea," she said.

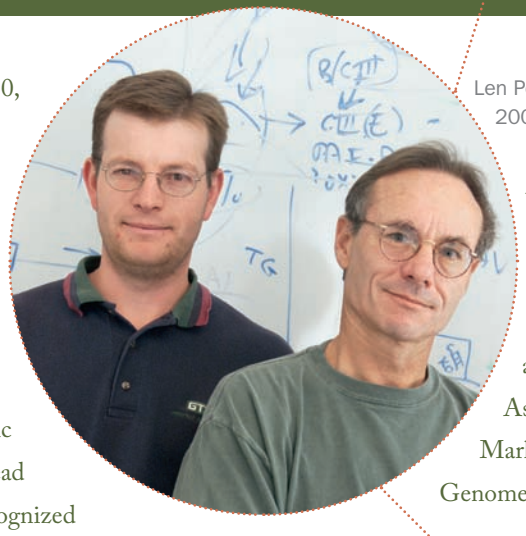
The IMG system is among the tools available on the Metagenomics Program site (http://genome.jgi-psf.org/programs/metagenomes/index.jsf). Also on the list is CLaMS, which stands for "Classifier for Metagenomic Sequences" and is a metagenome binning/clustering tool. A third tool available is called FAMeS, which provides access to simulated data sets and aims to facilitate standardized benchmarking of tools for metagenomic analysis.

Pacific Biosciences (PacBio) Single Cell Molecular Real-Time Detection System (Roy Kaltschmidt, Berkeley Lab)



Single cell genomics offers the ability to sort out one cell from a complex environmental sample, isolate the DNA, and generate a metabolic profile.

Legacy Project Highlights



Len Pennacchio (left) and Eddy Rubin (right) in 2001. (Roy Kaltschmidt, Berkeley Lab)

The December 17, 2010, issue of *Science* includes a special feature called "Insights of the Decade" and three of them feature work done by DOE JGI researchers.

DOE JGI's Director Eddy Rubin and Genomic Technologies Program head Len Pennacchio were recognized among the researchers involved in the study of so-called "junk DNA," conserved sequences that had no immediately obvious functions. Rubin and Pennacchio were instrumental in determining that these sequences actually contain regulatory DNA that functions while physically removed from the associated genes.

"The scope of this 'dark genome' became apparent in 2001, when the human genome was first published," noted *Science*, but understanding the sequences had to wait until the mouse genome was sequenced in 2002.

Rubin was among those recognized separately for contributions to advancing the field of paleogenomics through sequencing work on Neanderthal and Pleistocene cave bear samples.

Another key insight *Science* noted involves sequencing the genomes of nearly 1,000 microbial species found in the human body. The data collected from the Human Microbiome Project are being maintained in an online catalog overseen by DOE JGI's Associate Director and CIO Victor Markowitz and Metagenome Program and Genome Biology head Nikos Kyrpides.

From Prokaryote to Eukaryote

In the March 5, 2010, issue of *Cell* a team of researchers reported the sequence of a common soil organism called *Naegleria gruberi*, which has the ability to change its form. When stressed, the organism shifts from amoeboid to flagellate, swimming around with two sperm-like tails. When environmental conditions are unfavorable, it hibernates as a hard cyst. Biologists are interested in the *Naegleria* genome because it offers them a look at how organisms may have transitioned from single-celled prokaryotes to eukaryotes.

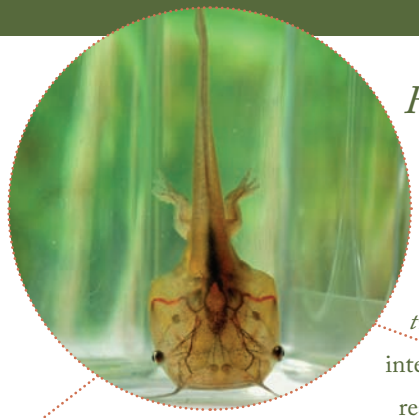
The 41 million-base genome was compared with the genomes of 16 other eukaryotes, providing researchers with what they think may be a core set of 4,000 genes that may have been part of the first and most primitive eukaryotes.

"Analyzing the *Naegleria* genomes shows us what it would be like to be on this planet more than a billion years ago, and what kind of organisms were around then, and what they might have looked like," said DOE JGI bioinformaticist Simon Prochnik, a co-author on the paper.

Additionally, the *Naegleria* genome indicated that the organism might produce hydrogen as a byproduct of adapting to muddy soils with lower oxygen levels. "Any kind of system that can make hydrogen has potential interest from a bioenergy point of view," Prochnik added.

A video of study first author Lillian Fritz-Laylin of the University of California, Berkeley, and Prochnik discussing the *Naegleria* project is online at http://bit.ly/fLVfAM.

Free-swimming form of *Naegleria* with two flagella. (Lillian Fritz-Laylin, UC Berkeley)



Late-stage *Xenopus. tropicalis* tadpole. (Siwei Zhang, Jingjing Li, Enrique Amaya/Univ. of Manchester, UK. Courtesy of *Science*)

First Amphibian Genome

The genome of the Western clawed frog *Xenopus tropicalis* was reported by an international team of researchers led by DOE JGI Plant Genome Program head Dan Rokhsar and UC Berkeley professor Richard Harland and appeared on the cover of the April 30, 2010, issue of *Science*.

With more than 1.7 billion bases across 10 chromosomes, the frog genome sequence provides scientists with a new tool to understand how genes work at the most basic level. Researchers can also use the information to better understand the factors causing the vast die-off of amphibians around the planet.

"The availability of the *Xenopus* genome also opens up the possibility of studying the effect of endocrine disruptors at the molecular and genomic level", said DOE JGI bioinformaticist and study first author Uffe Hellsten. "When you look at segments of the *Xenopus* genome, you literally are looking at structures that are 360 million years old and were part of the genome of the last common ancestor of all birds, frogs, dinosaurs, and mammals that ever roamed the earth."

A video on the project is available at http://bit.ly/eZm2lq.

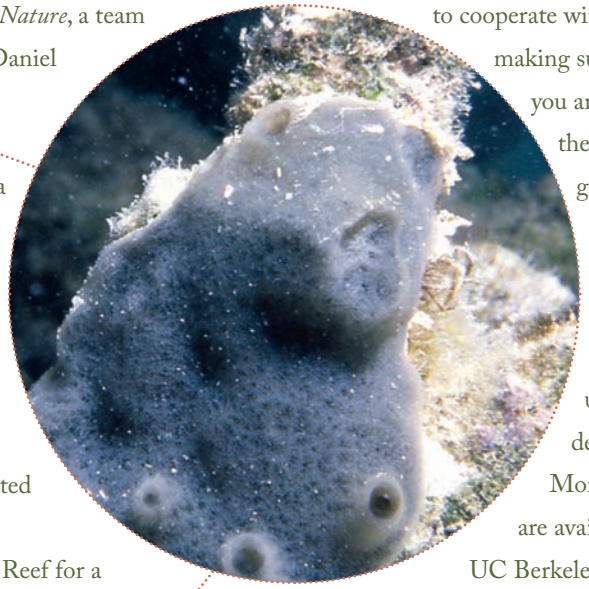
Sequencing the Simplest Animal — Sponge

In the August 5, 2010, issue of *Nature*, a team of researchers led by DOE JGI's Daniel Rokhsar and the University of Queensland's Bernard Degnan reported the draft genome of the sea sponge *Amphimedon queenslandica.*

"This incredibly old ancestor possessed the same core building blocks for multicellular form and function that still sits at the heart of all living animals, including humans," said Degnan, who collected the sponge whose genome was sequenced from the Great Barrier Reef for a 2005 Community Sequencing Program project. "It now appears that the evolution of these genes not only allowed the first animals to colonize the ancient oceans, but underpinned the evolution of the full biodiversity of animals we see today."

All living animals are descended from the common ancestor of sponges and humans, which lived more than 600 million years ago. A sponge-like creature may have been the first organism with more than one cell type and the ability to develop from a fertilized egg produced by the merger of sperm and egg cells — that is, an animal.

"If you are a cell in a multicellular organism, you have to cooperate with other cells in your body, making sure that you divide when you are supposed to as part of the team," said Rokhsar. "The genes that regulate this cooperation are also the ones whose disruption can cause cells to behave selfishly and grow in uncontrolled ways to the detriment of the organism." More details about the project are available in the release from UC Berkeley at http://bit.ly/ewU2Jh.



Adult *Amphimedon queenslandica* from the Great Barrier Reef. (Marie Gauthier)

Magellan system (this page and inset)
(Roy Kaltschmidt, Berkeley Lab)

Software & Analysis Tools

The foundation of all four programs at DOE JGI, as well as the processes ranging from project and sample tracking to analytical software tools, is informatics. Over time, informatics scientists and software engineers have expanded the tools and systems used both internally and by scientists worldwide. Projects range from supporting ongoing scientific studies and preparing for longer-term, larger-scale goals involving data management and analysis to developing methods for processing, organizing, and analyzing genome and metagenome datasets.

To help address DOE JGI's ever-escalating computational needs, the Institute has teamed up with Berkeley Lab's National Energy Research Scientific Computing Center (NERSC) Division. In 2010, DOE JGI generated approximately 5 terabases (5 trillion nucleotide letters of genetic code) from its plant, fungal, microbial, and metagenome (microbial community) user programs.

"This represents an exciting coming-of-age for the combined genomics and high-performance computing enterprise," said NERSC Division Director Kathy Yelick. "Datacentric computing is an important future direction for NERSC, and genomics is seeing exponential increases in computation and storage."

Among the resources brought to bear was a dedicated 10 gigabit per second (Gbps) link deployed by the engineers at the Energy Sciences Network (ESnet) between both institutions on its high-bandwidth Science Data Network (SDN) designed to haul massive scientific datasets — such as those from the Great Prairie Soil Metagenomes project that DOE JGI is piloting for the DOE Grand Challenge program. NERSC also offers environmental and energy-usage monitoring as well as an uninterruptible power supply to maintain data continuity and integrity.

In March 2010, DOE JGI became an early user of NERSC's Magellan Cloud Computing System based in Oakland, CA. Users still log on to the Institute's network and submit scientific computing jobs to batch queues managed by hardware located in Walnut Creek, but the job information then travels 20 miles on reserved SDN bandwidth directly to NERSC's Magellan system. Upon completion, the data results are sent back to Walnut Creek on the SDN, again within milliseconds.

In November 2010, the Magellan system received the "Best Use of High Performance Computing (HPC)" Award in New Orleans, LA, during the international supercomputing conference SC10, which focuses on high-performance computing, networking, storage, and data analysis. (For more information about the award, see

the original Argonne news release at http://bit.ly/b3N2m4.)

On December 14, 2010, the Energy Sciences Network (ESNet), which offers DOE researchers high-bandwidth, reliable connections, tweeted that DOE JGI had sent more than 50 terabits of genomics data in just the past 10 hours at the rate of nearly 10 gigabits per second. According to a December 10, 2010 blog post on ESNet's *Network Matters* site, the DOE JGI was also a significant contributor in helping ESNet traffic exceed the 10 petabytes barrier in November.

"Keep it coming," ESNet tweeted back to DOE JGI. "We can handle it."

"It does put things into perspective," noted DOE JGI systems infrastructure architect Jeremy Brand. "JGI is using 10 percent of what all DOE national labs via ESnet together are using. It is the single largest consumer, even larger than the Large Hadron Collider — the world's largest and highest-energy particle accelerator."

R&D Projects

Among the tasks performed by the Informatics group is research into and development of new methods or extending existing ones for processing, organizing, and analyzing genome and metagenome data sets. One recent R&D project is Meraculous, a software tool developed for Illumina-only genome projects to improve the quality of short-read DNA sequence assembly for small and large genomes. Another is CombEST, which allows annotation of fungal genomes using Illumina expressed sequence tag (EST) data, offering better performance, higher quality gene models generation, and more easily conducted computations in parallel. A third is ClaMS, or Classifier for Metagenomic Sequences, which was designed to help the "Binning" process for classification of contiguous sets of overlapping metagenomic DNA fragments according to their likely taxonomic origin, which enables analysis of genetic content of individual populations. An iterative version — iClaMS — is self-training and separates a set of sequences into clusters based on the similarity of sequence composition.

Many of the software tools developed at DOE JGI are freely available to download by academic institutions. One such tool released in 2009 was Gap Resolution, which expedites gap closure and assembly improvement in sequence data generated by the 454 Titanium platform and then assembled using the commercial platform Newbler. Since its release, 37 requests for the software have been processed.

Polisher

To improve genomic data after a draft genome has been sequenced and assembled, researchers rely on repeat resolution, gap closure, and polishing. One DOE JGI tool, Polisher, resolves base calling errors from sequences run on the 454 platform by aligning Illumina read data to the draft assembly.

The software package was designed to support correctly called bases that might be considered "below standard quality," facilitate the error correction of an assembled genome using Illumina read data by automatically correcting consensus errors, and suggesting primer walking reactions to improve the base quality.

The software requires the Arachne genome assembler developed by the Broad Institute and the graphical assembly viewer Consed. The Fasta sequence of the draft is aligned to Illumina reads using Arachne and the regions where 70 percent of the Illumina coverage disagree with

the sequence are targeted. Corrections such as insertions and deletions are made in these low-quality regions are made after they've been verified with the Illumina data.

The software was developed by Stephan Trong and his colleagues in DOE JGI's Finishing group and was made freely available to academic institutions in December 2010.

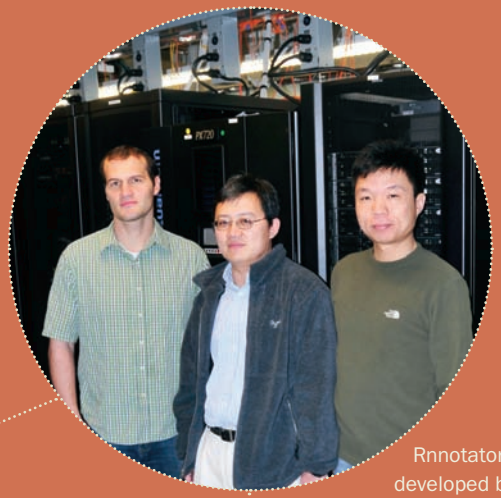
Rnnotator

High-throughput mRNA sequencing (RNA-Seq) involves isolating RNA, converting it to a library of cDNA fragments, and then sequencing the material. The data produced is useful for studying transcriptomes because it provides a comprehensive list of the transcripts and their expression levels under known conditions so that gene models can be developed or refined and gene expression levels determined.

One limitation of RNA-Seq is that the utility of the information generated is dependent on the quality of the reference genomes available. Many genome assemblies tend to be incomplete or unavailable, and sequencing RNA from microbial communities to study metatranscriptomes is challenging because the individual microbial genomes themselves may not exist.

To assist researchers, DOE JGI developed an automated software pipeline called Rnnotator that preprocesses RNA-Seq data and then conducts *de novo* assembly of RNA-Seq reads into transcriptomes. The team developed standards for evaluating the resulting transcriptome assemblies. Rnnotator also allows researchers to identify novel transcripts in sequences that were not in the reference genome.

The software was developed by Zhong Wang and his colleagues in DOE JGI's Genome Technologies group and was made freely available to academic institutions in November 2010.



Rnnotator developed by DOE JGI's (from left to right) Jeffrey Martin, Zhong Wang and Xiandong Meng, provides an automated software pipeline for the de novo assembly of transcriptomes — reflecting the genes actively expressed at a given time. (David Gilbert, DOE JGI)

Education & Outreach

Education

DOE JGI Education head Cheryl Kerfeld was the recipient of the 2011 American Society for Biochemistry and Molecular Biology (ASBMB) Award for Exemplary Contributions to Education. Kerfeld was recognized for "encouraging effective teaching and learning of biochemistry and molecular biology through her own teaching, leadership in education, writing, educational research, mentoring, and public enlightenment." (Learn more about the award and other ASBMB recipients at http://bit.ly/actTJG.)

To help fill a void in the life sciences training and help teaching institutions keep up with the rapid advances in sequencing technologies, DOE JGI's Education Program develops programs and tools to train the next generation of genomic scientists on large-scale DNA sequencing and bioinformatic analysis by integrating the use of these materials in their research experiences.

Cheryl Kerfeld, DOE JGI Education Head

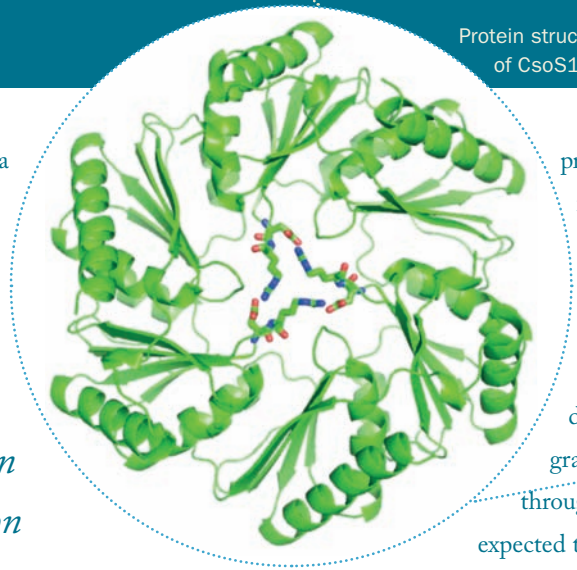
Undergraduate Research in Microbial Genome Annotation

The Interpret a Genome Program provides students in colleges and universities with access to recently sequenced bacterial genomes, such as those of organisms from little-known branches of the Tree of Life selected as part of the DOE JGI's Genomic Encyclopedia of Bacteria and Archaea (GEBA) project. The students analyze and annotate the genomes in the context of their own classwork, gaining hands-on knowledge of genomics and bioinformatics.

The research experience provides students with a "real-world" opportunity to study complete sequences of microorganisms and make novel discoveries that enrich the scientific community as a whole. Ultimately, the program's goal is to allow students nationwide to annotate GEBA genomes while learning about genomics and bioinformatics.

Fifteen institutions nationwide have been selected to participate in the 2011 microbial genome annotation program, joining the 65 institutions worldwide that already participate. Additional faculty training is now supported by an NSF grant. As their annotation platform, students will use the Integrated Microbial Genomes Annotation Collaboration Toolkit (IMG-ACT), a wiki/Web portal fusion that lets them work with existing genome data sets and record their discoveries. The platform is the result of a collaboration between DOE JGI's Education Program and faculty members from several universities around the country, and was described in the August 10, 2010, issue of *PLoS Biology*.

Aside from access to genomic data and "virtually endless" research possibilities, IMG-ACT offers students and their teachers access to bioinformatics databases, instructor course management, and student notebooks. Since IMG-ACT was launched in 2008, more than 100 faculty members and 1,600 students nationwide have participated in the program, noted Kerfeld and her colleagues. IMG-ACT is in turn linked to other databases used in microbial genome annotation, including IMG/EDU, the educational version of the Integrated Microbial Genomes



Protein structure (closed version) of CsoS1D (Cheryl Kerfeld)

database that is widely used by researchers in genome biology.

A complementary metagenomics annotation tool, IMG-ACTM has also been developed by the DOE JGI Education Program and is currently being used by the first test group of instructors.

The ASM/JGI Bioinformatics Institute

In collaboration with Hiram College, DOE JGI's Education Program also developed a series of workshops that provides college faculty who teach science, technology, engineering, and math (STEM) courses training in understanding and using bioinformatics tools.

The ASM/JGI Bioinformatics Institute, which is managed by the American Society of Microbiology, offers participants with little to no familiarity in the use of bioinformatics tools hands-on experience in accessing the Internet for databases, tools, and resources and identifying tools and materials for developing classroom activities and research projects back at their educational institutes.

To apply for the workshops, STEM faculty must be current, full-time teachers at community colleges, 4-year colleges, and research universities. After participating in the Institute workshops, participants are expected to demonstrate the effectiveness of the training they received not just in developing curricula for use in their classes but also sharing such modules with other STEM faculty in education publications, and in presenting a project at a national professional society meeting.

Undergraduate Research in Microbial Characterization

In collaboration with the University of Missouri, Columbia, and the University of South Florida, DOE JGI is developing hands-on tools to help undergraduates understand the process by which genomic knowledge is created and accumulated. The program offers students the opportunity to isolate, characterize, sequence, and annotate the genomes of novel organisms that form the foundation of the food chain and the energy chain found in deep-sea vents.

The research project affords undergraduates the opportunity to work in a variety of biological disciplines, from isolating these novel autotrophs and then sequencing their genomes, to using bioinformatics to analyze and interpret their data and annotate the information.

The modules developed focus on providing students with an understanding of processes such as gene evolution, protein structure and function, metabolic pathways, and ecological adaptation. These tools will be shared with faculty at diverse types of undergraduate institutions through workshops and are expected to be applicable to a variety of life science laboratory courses.

Undergraduate Research in Microbial Functional Genomics

To bring functional genomics research into the undergraduate experience, DOE JGI has teamed with the ASM, Hiram College in Ohio, and St. Cloud State University in Minnesota to incorporate concepts such as reverse and forward genetics, protein overexpression, and protein crystallization in undergraduate lab experiences. The program offers students experience in contributing to the global database of protein structures and functions in much the same way the Interpret a Genome Program allows students to both learn and contribute in the field of bioinformatics.

Berkeley Lab Director Paul Alivisatos at the Berkeley Lab Open House with Big Betty and one of the high school students who made the cow. (David Gilbert, DOE JGI)

Community Outreach

Building on a previous collaboration, DOE JGI teamed with students from Ex'pression College for Digital Arts to develop a short animated video describing the Institute's role in the search for clean alternative energy sources. The video was first shown at the 2010 User Meeting and was a finalist in the 2010 Multimedia Awards held by *The Scientist* magazine. The video is available in English and Mandarin Chinese on DOE JGI's YouTube Web site.

To help foster community relations, DOE JGI also collaborated with art students from a neighboring high school to build a papier mache model related to a prominent sequencing project with bioenergy applications. Big Betty the Energy Cow made her debut at the Berkeley Lab Open House held October 2, 2010, during which the DOE JGI assisted nearly 3,500 attendees in extracting DNA from strawberries. DOE JGI also participated in the inaugural Science & Technology Festival in Washington D.C. on October 23-25, 2010, and at the satellite event held at UC Berkeley on January 23, 2011. The latter events were geared toward inspiring the next generation of scientists and engineers through hands-on activities.

DOE JGI's own Jim Bristow and Susannah Tringe appeared on a panel with Joint BioEnergy Institute CEO Jay Keasling on January 18, 2011, at the Lesher Center for the Arts in Walnut Creek. KTVU Channel 2 health and science editor John Fowler moderated the talk titled: "The Future of Fuel: 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://bit.ly/fIVE9w.

In front of a full house in the Margaret Lesher Theatre, Jim Bristow (right), Susannah Tringe (center right) and Jay Keasling (center left) discussed biofuels with KTVU's John Fowler (left). (Roy Kaltschmidt, Berkeley Lab)

From left, DOE JGI's Christine Naca, Melanie Alexandre, Nicole Shapiro, Megan Kennedy, Angela Tarver, Bridget Swift, and Martin Pollard accept the 2010 Ergo Cup in Dallas, Texas.

Safety & Ergonomics

The DOE JGI emphasizes employee safety in the workplace and relies on the Integrated Safety Management (ISM) Plan to ensure that all work processes are executed with the health and safety of employees and guests, the public, and the environment in mind. The core components of the highly successful ISM program at DOE JGI are the employee-driven safety committees and Berkeley Lab's newly instituted Job Hazards Analysis process. The various employee safety groups focus on ergonomics, safety culture, and emergency response and provide valuable support to the professional safety staff at DOE JGI. The Job Hazards Analysis process was developed by Berkeley Lab's EH&S team to identify any work hazards and required training prior to starting work and is used by DOE JGI employees, supervisors and outside contractors to implement appropriate safety measures to conduct safe operations throughout the facility.

DOE JGI is widely recognized as a leader in the area of ergonomics, the science of designing equipment, and practices to reduce musculoskeletal disorders in the workplace. DOE JGI has an onsite ergonomist who works with the members of the Ergonomics Working Group to develop ergonomic solutions suited to the worker and the environment. DOE JGI employees routinely perform highly repetitive tasks in both laboratory and office environments.

At the Annual Applied Ergonomics Conference held March 22-25, 2010, in Dallas, Texas, DOE JGI won its second Ergo Cup for its entry in the Ergonomic Program Improvement Initiatives category: "Empowering Employees in Ergonomics."

The entry focused on employee-driven elements of the DOE JGI program that promote a safety culture and ergonomic awareness, highlighting key elements such as providing safety information in high-traffic areas, including hallways and restrooms, and employee-led safety walkthroughs.

DOE JGI previously won the 2007 Ergo Cup for its entry to the Team-driven Workplace Solutions category. The Shake 'N Plate instrument was designed to ease upper body fatigue for employees working on the Sanger sequencing production line were manually processing 22 cm x 22 cm plates of bacterial cultures.

"This award is arguably the most prestigious recognition in applying ergonomics and extremely relevant to reducing musculoskeletal disorder (MSD) injuries," said Andrew S. Imada, President of the International Ergonomics Association. "To win this award twice in such a short interval is truly remarkable."

Other employee-driven initiatives encouraged by management and designed to reduce the total injuries resulting from performing repetitive and detail-oriented tasks also led to the central installation of computer usage tracking software that reminds and instructs employees to take regular breaks from their work, and the designation of an on-site relaxation and rejuvenation room.

Appendices

Appendix A: Glossary

ANNOTATION: The process of identifying the locations of genes in a genome and determining what those genes do for the structure and functioning of the cell/organism from which they come.

ARCHAEA: One of the three domains of life that subsume single-celled microorganisms with RNA sequences that are functionally different from bacteria. (Eukaryotes comprise the third domain.)

ASSEMBLY: Compilation of overlapping DNA sequences (obtained from a given genome) that have been clustered together based on their degree of sequence identity or similarity. The purpose of assembly is to reconstruct longer portions of the genome (see Contigs.)

BARCODING: The practice of appending known, unique synthetic DNA sequences to sequencing libraries to allow pooling of libraries for next-generation sequencing, after which sequence data can be assigned to particular libraries or samples based on the barcode sequence.

BASE: A unit of DNA. There are four bases: adenine (A), guanine (G), thymine (T), and cytosine (C). The sequence of bases is the genetic code.

BASE PAIR: Two DNA bases complementary to each other (A and T or G and C) that join the complementary strands of DNA to form the double helix characteristic of DNA.

BIOREMEDIATION: Using microorganisms to break down and chemically transform contaminants and other unwanted substances in waste and other substances.

BIOGEOCHEMISTRY: A study of the biosphere's interactions with the Earth's chemical environment.

BIOINFORMATICS: The use of computers to collect, store, and analyze biological information.

BRIDGE AMPLIFICATION: A proprietary technique used by Illumina sequencing platforms to generate single-stranded clusters of template DNA.

CLONING: Using specialized DNA technology to produce multiple, exact copies of a single gene or other segment of DNA, to obtain enough material for further study.

CONTIG: Group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome.

COVERAGE: The number of times a region of the genome has been sequenced during whole genome shotgun sequencing.

CURATION: Analysis of genome annotations to improve function predictions and maintain data integrity.

DRAFT GENOME: The term for an incomplete genome sequence can be applied to a wide range of sequences, from those that have the minimum amount of information needed for submission to a public database, to assembled genomes that have undergone manual and automatic review but still have sequence errors that need to be corrected (see Finished Genome).

EUKARYOTES: The domain of life containing organisms that consist of one or more cells, each with a nucleus and other well-developed intracellular compartments. Eukaryotes include all animals, plants, and fungi.

FINISHED GENOME: In accordance with the 1996 Bermuda standard, this is a gapless sequence with a nucleotide error rate of one or less in 10,000 bases.

FLOW CELL: Resembles a microscopic slide only with eight channels on which DNA samples are loaded for analysis on the Illumina sequencing platforms.

FOSMID: A vector suitable for cloning genomic inserts approximately 40 kilobases in size.

GENBANK: Open-access, publicly available collection of annotated sequences submitted by individual laboratories and large-scale sequencing centers that is overseen by the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/).

LIBRARY: An unordered collection of clones containing DNA fragments from a particular organism or environment that together represent all the DNA present in the organism or environment.

MAPPING: Charting the location of genes on chromosomes.

METAGENOMICS (ALSO ENVIRONMENTAL GENOMICS OR COMMUNITY GENOMICS): The study of genomes recovered from environmental samples through direct methods that do not require isolating or growing the organisms present. This field of research allows the genomic study of organisms that are not easily cultured in a laboratory.

METATRANSCRIPTOMICS: The study of the region of the complete genetic code that is transcribed into RNA molecules and provides information on gene expression and gene function.

MICROBIOME: A defined environment within which a community of microbes exists and interacts with each other.

PAIRED-END READS: DNA library preparation technique that lets researchers look at both the forward and reverse template strands of a large DNA fragment and provides positional information.

PCR: Acronym for Polymerase Chain Reaction, a method of DNA amplification.

PHYLOGENY: The evolutionary history of a molecule such as gene or protein, or a species.

PROKARYOTES: Unlike eukaryotes, these organisms, (e.g., bacteria) are characterized by the absence of a nuclear membrane and by DNA that is not organized into chromosomes.

READ LENGTH: The number of nucleotides tabulated by the DNA analyzer per DNA reaction well.

SEQUENCE: Order of nucleotides (base sequence) in a nucleic acid molecule. In the case of DNA sequence, it is the precise ordering of the bases (A, T, G, C) from which the DNA is composed.

SEQUENCING BY SYNTHESIS: Proprietary sequencing technique used by Illumina systems in which four fluorescently labeled nucleotides determine the sequence of a DNA fragment one base at a time.

SUBCLONING: The process of transferring a cloned DNA fragment from one vector to another.

TRANSCRIPTOME: A collection of all the RNA transcripts in a given cell that serves as a snapshot of global gene expression.

Appendix B: DOE JGI Sequencing Process

DNA: Life's Code

Deoxyribonucleic acid or DNA is a molecule made up of four chemical components — the nucleotides Adenine (A), Thymine (T), Cytosine (C), and Guanine (G) — and is found in all living organisms. In the DNA molecule's double-helix structure, the A's always bind with T's, and C's always bind with G's.

What is DNA Sequencing?

To read a book, we must know the order in which letters appear, and how these are grouped to form words and sentences. To read and understand the book of life, researchers need to know the order, or sequence, in which the four letters that make up DNA appear.

Illumina Sequencing Technology

This platform is based on parallel sequencing of millions of fragments using proprietary technology that amplifies the template by bridge PCR onto a glass flow cell slide, and a reversible terminator-based sequencing-by-synthesis chemistry that allows detection of the sequence in real time. The sequencing workflow is at heart a three-step process: Libraries are prepared from for a nucleic acid sample, amplified, and then sequenced using massively parallel synthesis. Use of the Illumina HiSeq platform allows DOE JGI to generate hundreds of gigabases of data in less than a week.

The strategy is outlined below and opposite.

1 Prepare sample by randomly fragmenting genomic DNA and ligating adapters to both ends of the fragments.

2 Randomly bind single-stranded DNA fragments to the inside surface of the eight channels on the top and bottom surfaces of each flow cell.

3 DNA Amplification

- **a** Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
- **b** Fragments become double-stranded DNA bridges.
- **c** By completion of amplification, several million dense clusters of single-stranded DNA have been generated in each channel of the flow cell with a sequencing primer attached.

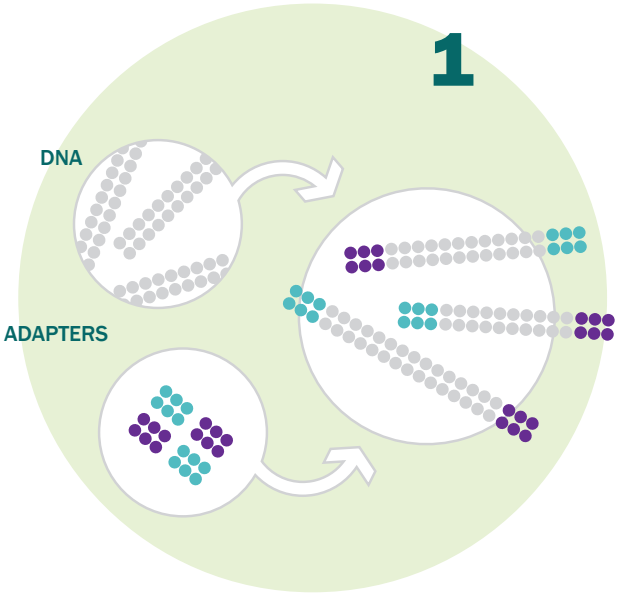
4 To initiate the first sequencing cycle and determine the first base, all four labeled reversible terminators and DNA polymerase enzyme are first added to the top surface of the flow cell slide.

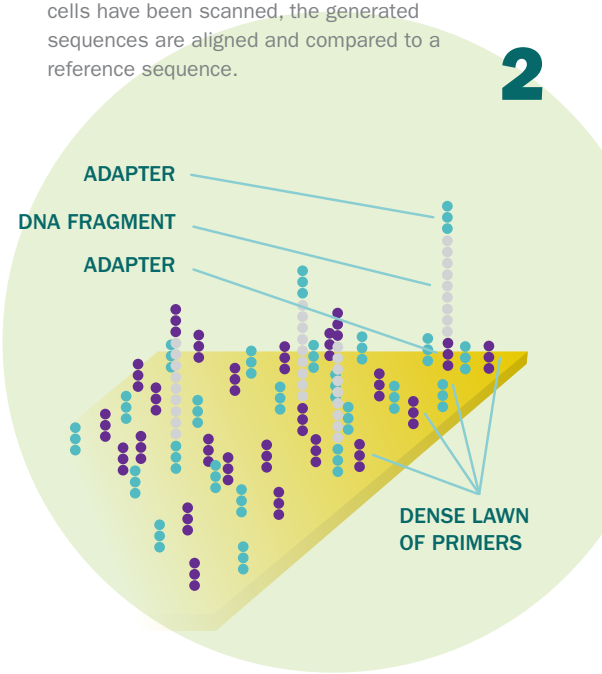
5 In the first cycle, the first base is incorporated. Its identity is determined by the signal given off and then recorded. In subsequent cycles, the process of adding sequencing reagents, removing unincorporated bases and capturing the signal of the next base to identify is repeated.

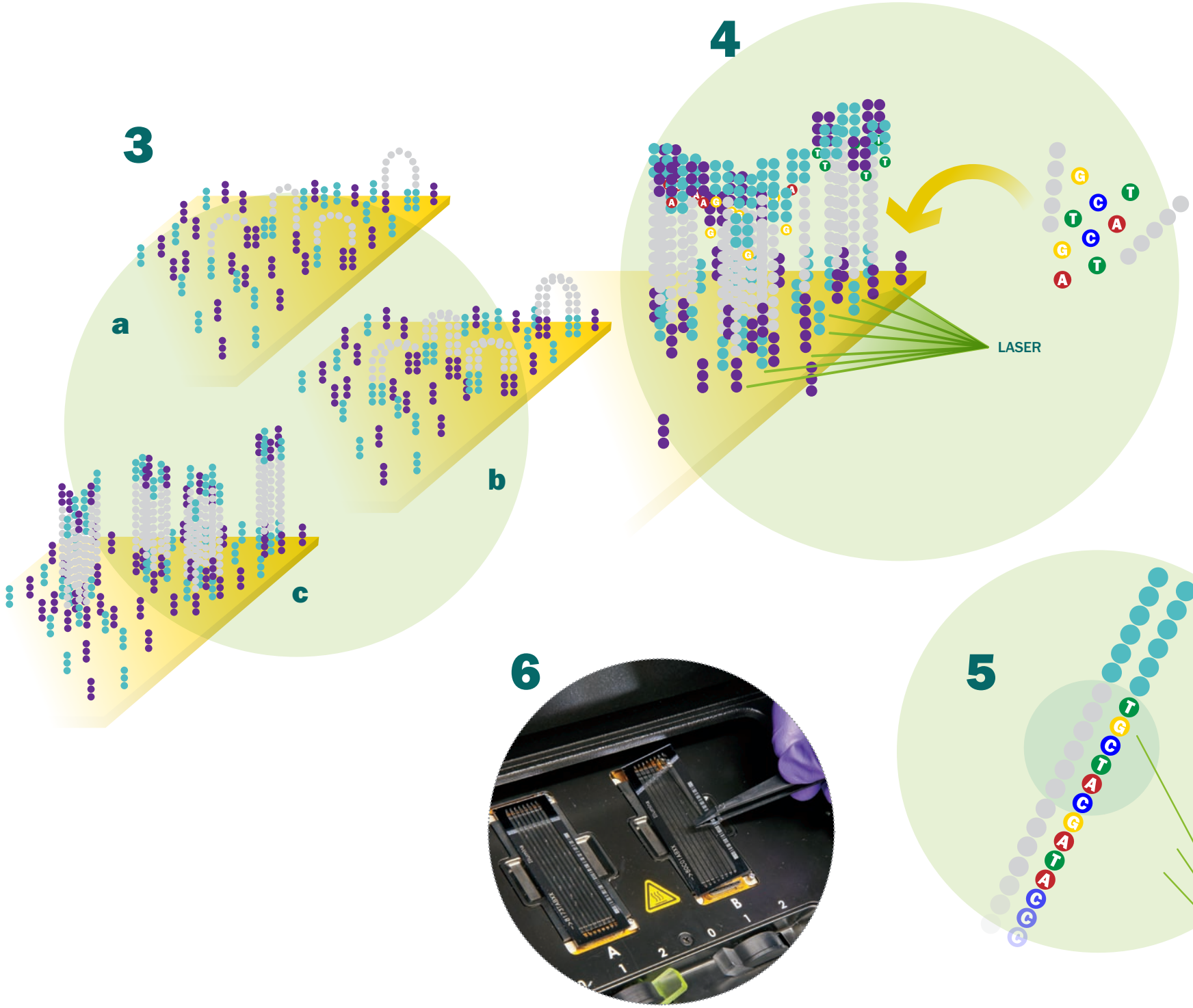
6 Once the top surface of the flow cell has been scanned, the imaging step is repeated on the bottom surface. When both flow cells have been scanned, the generated sequences are aligned and compared to a reference sequence.

Illumina HiSeq machines.

Photos, Roy Kaltschmidt, Berkeley Lab






3
a
b
c
4
LASER
5
6

Appendix C: FY 2011 CSP Projects

PROPOSER	AFFILIATION	PROJECT DESCRIPTION
ALGAE		
Kerfeld, Cheryl	DOE Joint Genome Institute	Genome and transcriptome analyses of two extremely acidophilic and one neutrophilic eukaryotic algal species with diverse mechanism for CO_2 acquisition
Lovejoy, Connie	Laval University, Canada	Small planktonic single celled eukaryotes from the Arctic Ocean
PLANTS		
Muehlbauer, Gary	University of Minnesota	Whole genome shotgun sequencing of the barley genome
Vogel, John	USDA-ARS	Surveying natural diversity of the model grass *Brachypodium distachyon*
FUNGI		
de Vries, Ronald	CBS-KNAW Fungal Biodiversity Centre, the Netherlands	Comparative analysis of *Aspergilli* to facilitate novel strategies in fungal biotechnology
Goodwin, Stephen	Purdue University	Sequencing of pathogens and extremophiles in the Dothideomycetes
Hibbett, David S.	Clark University	Community proposal to sequence a diverse assemblage of saprotrophic *Basidiomycota (Agaricomycotina)*
Jeffries, Thomas	Forest Products Laboratory	Yeasts of biotechnological, taxonomic and physiological interest
Martin, Francis	INRA, France	Exploring the genome diversity of mycorrhizal fungi to understand the evolution and functioning of symbiosis in woody shrubs and trees
Pisabarro, Antonio	Public University of Navarre, Spain	Comparative transcriptomics pipeline for saprophytic basidiomycota
Pringle, Anne	Harvard University	Comparative transcriptomics of closely related saprotrophic and ectomycorrhizal *Amanita* species
Spatafora, Joseph	Oregon State University	Phylogenomics and the origin and diversification of Kingdom Fungi
Turgeon, Gillian	Cornell University	Cochliobolus: expanded and deepened
Turk, Martina	University of Ljubljana, Slovenia	The varieties of the black yeast-like fungus *Aureobasidium pullulans*: evolution and use in biotechnology
BACTERIA/ARCHAEA		
Bertilsson, Stefan	Uppsala University, Sweden	Genome wide diversity and population genetics in uncultured aquatic bacteria: Single-cell genomics of the freshwater SAR11 group and the ubiquitous *Actinobacteria* ac1 lineage
Chistoserdova, Ludmila	University of Washington	Genomes of fifty methylotrophs isolated from Lake Washington
Eichorst, Stephanie	Los Alamos National Laboratory	Populating the branches of the Phylum *Acidobacteria* with relevant soil strains
Eisen, Jonathan	DOE Joint Genome Institute	Continuation of the Genomic Encyclopedia of Bacteria and Archaea pilot project

PROPOSER	AFFILIATION	PROJECT DESCRIPTION
Kalyuzhnaya, Marina	University of Washington	Coupling function to genomics via single-cell phenotyping and genome sequencing
Liu, Wen-Tso	University of Illinois	Single-cell genomics for uncultured Archaea dominating in a terrestrial subsurface aquifer abundantly containing methane
Sievert, Stefan	Woods Hole Oceanographic Institute	Shedding light on the dark: Single-cell genomics of uncultivated *epsilonproteobacteria* inhabiting the subseafloor biosphere at deep-sea hydrothermal vents
Stepanauskas, Ramunas	Bigelow Laboratory for Ocean Sciences	Generating reference genomes for marine ecosystem research: Single-cell sequencing of ubiquitous, uncultured bacterio-plankton clades
Tsiamis, George	University of Ioannina, Greece	Unraveling the unique microbial diversity of the Etoliko lagoon in Western Greece through a single-cell genomics approach
METAGENOMES		
Bryant, Donald	Penn State University	Metagenomic and metatranscriptomic analysis of anoxygenic, chlorophototrophic microbial mat communities in Yellow-stone National Park
Campbell, Barbara	University of Delaware	Metagenomic and metatranscriptomic analysis of carbon cycling in Delaware coastal waters
Dionisi, Hebe	Patagonian National Research Center, Argentina	Microbial community structure and metabolic potential of chronically polluted marine sediments from cold regions of the northern and southern hemispheres
Distel, Dan	Ocean Genome Legacy Foundation, Center for Marine Research	The complete shipworm microbiome: A comparative genomic and metagenomic analysis of lignocellulose-degrading microbial communities from multiple species of wood-boring bivalves
Girguis, Peter	Harvard University	Linking mantle to microbe: A community-wide effort to ally hydrothermal vent microbial identity and ecology to geochemi-cal cycles via metagenomics
Hallam, Steven	University of British Columbia, Canada	Microbial systems ecology of expanding oxygen minimum zones in the eastern subtropical north pacific ocean
Kerfeld, Cheryl	DOE Joint Genome Institute	Metagenomic sequencing for understanding microbial carbon cycling by biological soil crusts of arid lands
Macalady, Jennifer	Penn State University	Uncultivated and novel microbial lineages in terrestrial subsurface biofilms from a sulfidic aquifer
McMahon, Katherine	University of Wisconsin	High-resolution temporal and spatial dynamics of microbially mediated carbon processing revealed though time-series metagenomics in freshwater lakes
Moran, Mary Ann	University of Georgia	Transcriptional analysis of a marine bacteria-phytoplankton binary model system
Muyzer, Gerard	Delft University of Technology, the Netherlands	Metagenomics of microbial communities from soda lakes and soda solonchak soils
Tringe, Susannah	DOE Joint Genome Institute	Microbial community impact on carbon sequestration in managed wetland “carbon farming”

Appendix D: Review Committee and Advisory Committtee Members

CSP 2011 Reviewers

EUKARYOTIC PROPOSAL REVIEWERS

JODY BANKS Purdue University
DAREN BROWN USDA Agricultural Research Station
KENNETH BRUNO Pacific Northwest National Laboratory
DAN CULLEN USDA Forest Service Forest Products Laboratory
JEFFREY DEAN University of Georgia
DAVID GEISER Penn State University
TOM JEFFRIES University of Wisconsin, Madison
ANDREW KOPPISCH Los Alamos National Laboratory
LI-JUN MA Broad Institute
GEORGIANA MAY University of Minnesota
KEVIN MCCLUSKEY University of Missouri, Kansas City
SABEEHA MERCHANT University of California, Los Angeles
DON NATVIG University of New Mexico
DANIEL PETERSON Mississippi State University
GARRET SUEN University of Wisconsin, Madison
CHRIS TOWN J. Craig Venter Institute
TSVETAN BACHVAROFF (AD HOC) Smithsonian Institute
AMY BRUNNER (AD HOC) Virginia Polytechnic Institute and State University
GARY STACEY (AD HOC) University of Missouri
CHUNXIAN CHEN (AD HOC) University of Florida

PROKARYOTIC PROPOSAL REVIEWERS

ELIZABETH EDWARDS University of Toronto
HECTOR GARCIA MARTIN Joint BioEnergy Institute
KIRK HARRIS University of Colorado
WILLIAM INSKEEP Montana State University
JAVIER IZQUIERDO Dartmouth College/BioEnergy Science Center
DAVE KIRCHMAN University of Delaware
ROB KNIGHT University of Colorado at Boulder
WEN-TSO LIU University of Illinois
TODD LOWE University of California, Santa Cruz
WILLIAM MOHN University of British Columbia
NANCY MORAN University of Arizona
BISWARUP MUKHOPADHYAY Virginia Polytechnic Institute and State University
HOWARD OCHMAN University of Arizona
MIRCEA PODAR Oak Ridge National Laboratory
RUTH RICHARDSON Cornell University
DOUGLAS RUSCH J. Craig Venter Institute
MARK SCHELL University of Georgia
PATRICIA SOBECKY University of Alabama
JOHN SPEAR Colorado School of Mines
SERGEY STOLYAR Pacific Northwest National Laboratory
TATIANA TATUSOVA National Center for Biotechnology Information
TAMAS TOROK Lawrence Berkeley National Laboratory
PETER TURNBAUGH Washington University
DANIEL VAN DER LELIE Brookhaven National Laboratory
DAVID WARD Montana State University
BART WEIMER University of California, Davis

Program Advisory Committees

FUNGAL PROGRAM ADVISORY COMMITTEE

JOAN BENNETT Rutgers University
RANDY BERKA Novozymes
DANIEL CULLEN Forest Products Laboratory
DANIEL EASTWOOD Warwick University (UK)
STEPHEN GOODWIN Purdue University
DAVID HIBBETT Clark University
THOMAS JEFFRIES USDA Forest Service Forest Products Laboratory
GERT KEMA Plant Research International (Netherlands)
CHRISTIAN KUBICEK Vienna University of Technology (Austria)
CHERYL KUSKE Los Alamos National Laboratory
FRANCIS MARTIN INRA (France)
CONRAD SCHOCH NCBI
JOSEPH SPATAFORA Oregon State University
JASON STAJICH UC Riverside
JOHN TAYLOR UC Berkeley
ADRIAN TSANG Concordia University (Canada)
GILLIAN TURGEON Cornell University
RYTAS VILGALYS Duke University

MICROBIAL GENOMICS AND METAGENOMICS PROGRAMS ADVISORY COMMITTEE ("MG- MAC")

SWAPNIL CHHABRA Lawrence Berkeley National Laboratory/Joint BioEnergy Institute
CAMERON CURRIE University of Wisconsin
JED FUHRMAN University of Southern California
STEVE HALLAM University of British Columbia
JAVIER IZQUIERDO BioEnergy Science Center
BOB LANDICK Great Lakes Bioenergy Research Center
FOLKER MEYER Argonne National Laboratory
NANCY MORAN Yale University
MARY ANN MORAN University of Georgia
FOREST ROHWER San Diego State University
DOUG RUSCH J. Craig Venter Institute
RAMUNAS STEPANAUSKAS Bigelow Lab
MATT SULLIVAN University of Arizona
TATIANA TATUSOVA National Center for Biotechnology Information
NIELS VAN DER LELIE Brookhaven National Laboratory
PHIL HUGENHOLTZ University of Queensland

PLANT PROGRAM ADVISORY COMMITTEE

JOE ECKER The Salk Institute for Biological Studies
JEFF DANGL University of North Carolina
STEPHEN MOOSE University of Illinois
SABEEHA MERCHANT University of California, Los Angeles
GARY STACEY University of Missouri
THOMAS MITCHELL-OLDS Duke University
EVA HUALA Carnegie Institute/TAIR
CATHY RONNING US Department of Energy

Scientific Advisory Committee (SAC)

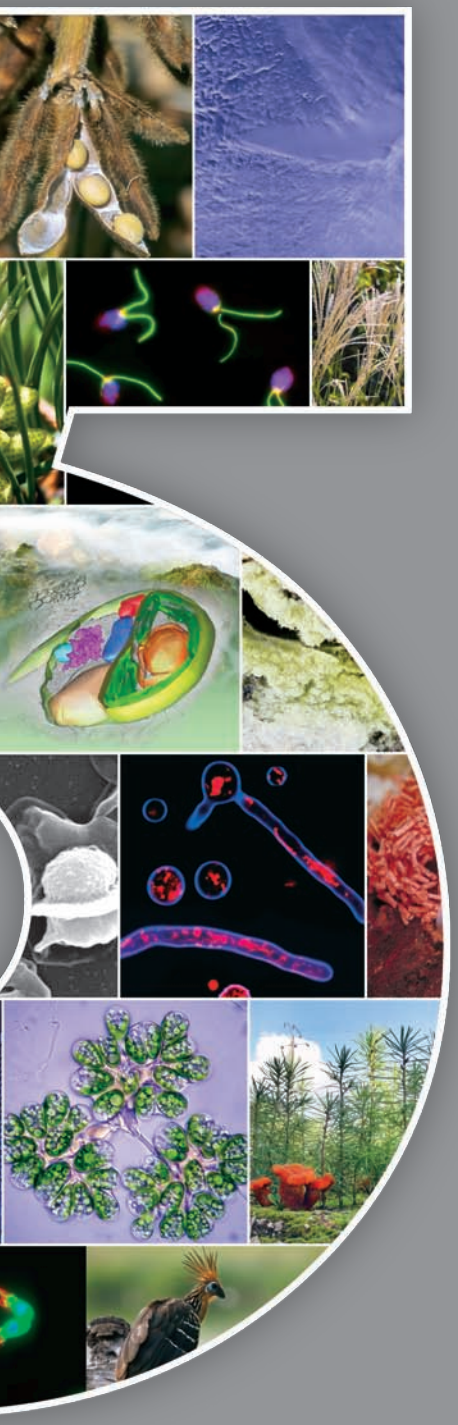
The Scientific Advisory Committee is a board convened by the JGI Director to provide a scientific and technical overview of the JGI. Among the board's responsibilities are: providing technical input on large-scale production sequencing, new sequencing technologies, and related informatics; overview of DOE JGI's scientific programs; and an overview of the Community Sequencing Program (CSP). One of the most important tasks of the SAC is to set the final sequence allocation for the CSP based on input from the CSP Proposal Study Panel on CSP project prioritization, and the concurrence of the DOE Office of Biological and Environmental Research.

MEMBERS

BRUCE BIRREN Broad Institute
TOBY BLOOM Broad Institute
JEFF DANGL University of North Carolina
JOE ECKER The Salk Institute for Biological Studies
JIM KRUPNICK Lawrence Berkeley National Laboratory
ERIC J. MATHUR SG Biofuels
NANCY MORAN Yale University
JULIAN PARKHILL The Sanger Institute
DOUG RAY Pacific Northwest National Lab

Appendix E: 2010 JGI User Meeting

Scenes from the User Meeting. Photographs by Roy Kaltschmidt, Berkeley Lab



The DOE JGI Fifth Annual User Meeting took place March 25-27, 2009, in Walnut Creek. The keynote speakers were Jay Keasling from the Joint BioEnergy Institute, Roger Pennell from Ceres, Inc., and Rita Colwell from the University of Maryland.

Other featured speakers were:

EDDY RUBIN Director, DOE Joint Genome Institute

DENNIS HEDGECOCK University of Southern California

STEVEN HALLAM University of British Columbia

ALEXANDRA WORDEN MBARI

GARY STACEY University of Missouri

CHRISTINA CUOMO Broad Institute, MIT

FRANCIS MARTIN INRA

STEVE KNAPP Monsanto

EVAN DELUCIA University of Illinois/EBI

MADHU KHANNA University of Illinois/EBI

TOM MITCHELL-OLDS Duke University

FOREST ROHWER San Diego State University

VICTORIA ORPHAN Cal Tech

DETLEF WEIGEL Max Planck Institute

STEVE MOOSE University of Illinois/EBI

STEVE SAVAGE Cirrus Partners

TANJA WOYKE DOE Joint Genome Institute

ADRIAN TSANG Concordia University

JOSEPH NOEL Salk Institute

Videos of the 2010 User Meeting talks are available on the DOE JGI's SciVee channel at: http://www.scivee.tv/node/17188

Appendix F: 2009-2010 Publications

Abt B et al. Complete genome sequence of *Cellulomonas flavigena* type strain (134T). *Standards in Genomic Sciences.* 2010:3(1): 15-25.

Aklujkar M et al. The genome of *Geobacter bemidjiensis,* exemplar for the subsurface clade of Geobacter species that predominate in Fe(III)-reducing subsurface environments. *BMC Genomics.* 2010 Sep 9;11:490.

Allgaier M et al. Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community. *PLoS One.* 2010 Jan 21;5(1):e8812.

Alverson AJ et al. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Mol Biol Evol.* 2010 Jun;27(6):1436-48.

Anderson IJ et al. Complete genome sequence of *Halorhabdus utahensis* type strain (AX-2T). *Standards in Genomic Sciences.* 2009:1(3):218-225.

Ayala-Del-Río HL et al. The genome sequence of Psychrobacter arcticus 273-4, a psychroactive Siberian permafrost bacterium reveals mechanisms for adaptation to low temperature growth. *Appl Environ Microbiol.* 2010 Apr;76(7):2304-12.

Baker SE. Selection to sequence: opportunities in fungal genomics. *Environ Microbiol.* 2009 Dec;11(12):2955-8.

Blanc G et al. The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell.* 2010 Sep;22(9):2943-55.

Blow MJ et al. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet.* 2010 Sep;42(9):806-10.

Burnum KE et al. Proteome insights into the symbiotic relationship between a captive colony of *Nasutitermes corniger* and its hindgut microbiome. *ISME J.* 2011 Jan;5(1):161-4.

Cannon GC et al. Carboxysomal carbonic anhydrases: Structure and role in microbial CO_2 fixation. *Biochim Biophys Acta.* 2010 Feb;1804(2):382-92. Epub 2009 Oct 8.

Chain PS et al. Genomics. Genome project standards in a new era of sequencing. *Science*. 2009 Oct 9;326(5950):236-7.

Chan YF et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science*. 2010 Jan 15;327(5963):302-5.

Chang YJ et al. Complete genome sequence of *Acidaminococcus fermentans* type strain (VR4T). *Standards in Genomic Sciences.* 2010:3(1): 1-14.

Chapman JA et al. The dynamic genome of *Hydra*. *Nature*. 2010 Mar 25;464(7288): 592-6.

Chertkov O et al. Complete genome sequence of *Aminobacterium colombiense* type strain (ALA-1T). *Standards in Genomic Sciences.* 2010:2(3): 280-289.

Chiang YM et al. Characterization of a polyketide synthase in *Aspergillus niger* whose product is a precursor for both dihydroxynaphthalene (DHN) melanin and naphtho-γ-pyrone. *Fungal Genet Biol.* 2010 Dec 19.

Chovatia M et al. Complete genome sequence of *Thermanaerovibrio acidaminovorans* type strain (Su883T). *Standards in Genomic Sciences.* 2009:1(3):254-261.

Clum A et al. Complete genome sequence of *Pirellula staleyi* type strain (ATCC 27377T). *Standards in Genomic Sciences.* 2009:1(3):308-316.

Coulbourne JK et al. The ecoresponsive genome of *Daphnia pulex. Science.* 4 Feb 2011; 331(6017):555-561.

Cuvelier ML et al. Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc Natl Acad Sci* U S A. 2010 Aug 17;107(33):14679-84.

DeAngelis KM et al. Strategies for enhancing the effectiveness of metagenomic-based enzyme discovery in lignocellulolytic microbial communities. *Bioenergy Res.* Jun 2010; 3(2):Sp. Iss. SI, 146-148.

Dehal PS et al. MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D396-400.

Deli A et al. LmbE proteins from *Bacillus cereus* are de-N-acetylases with broad substrate specificity and are highly similar to proteins in *Bacillus anthracis. FEBS J.* 2010 Jul;277(13):2740-53.

Ditty JL et al. (2010) Incorporating genomics and bioinformatics across the life sciences curriculum. *PLoS Biol* 8(8): e1000448.

Donaher N et al. The complete plastid genome sequence of the secondarily nonphotosynthetic alga *cryptomonas paramecium:* reduction, compaction, and accelerated evolutionary rate. *Genome Biol Evol.* 2009 Nov 13;2009:439-48.

Dover N et al. Novel structural elements within the non-proteolytic *clostridium botulinum* type f toxin gene cluster. *Appl Environ Microbiol.* 2010 Dec 23.

Du J et al. Bifurcation and enhancement of autonomous-nonautonomous retrotransposon partnership through LTR Swapping in soybean. *Plant Cell.* 2010 Jan;22(1):48-61.

Duncan KE. Biocorrosive thermophilic microbial communities in alaskan north oil slope facilities. *Environ Sci Technol.* 2009 Oct 15;43(20):7977-7984.

Elkins JG et al. Complete genome sequence of the cellulolytic thermophile *Caldicellulosiruptor obsidiansis* OB47T. *J Bacteriol.* 2010 Nov;192(22):6099-100.

Engelbrektson A et al. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J.* 2010 May;4(5):642-7.

Ferris P et al. Evolution of an expanded sex-determining locus in *Volvox. Science.* 2010 Apr 16;328(5976):351-4.

Foster B et al. Complete genome sequence of *Xylanimonas cellulosilytica* type strain (XIL07T). *Standards in Genomic Sciences.* 2010:2(1): 1-8.

Fritz-Laylin LK et al. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell.* 2010 Mar 5;140(5):631-42.

Glavina del Rio T et al. Complete genome sequence of *Chitinophaga pinensis* type strain (UQM 2034T). *Standards in Genomic Sciences.* 2010:2(1): 87-95.

Göker M et al. Complete genome sequence of *Ignisphaera aggregans* type strain (AQ1.S1T). *Standards in Genomic Sciences.* 2010:3(1): 66-75.

Göker M et al. Complete genome sequence of *Olsenella uli* type strain (VPI D76D-27CT). *Standards in Genomic Sciences.* 2010:3(1): 76-84.

Goode DL et al. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Res.* 2010 Mar;20(3):301-10. Epub 2010 Jan 12.

Gotea V et al. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.* 2010 May;20(5):565-77.

Gowda M et al. Genome-wide characterization of methylguanosine-capped and polyadenylated small RNAs in the rice blast fungus *Magnaporthe oryzae. Nucleic Acids Res.* 2010 Nov 1;38(21):7588-69.

Gronow S et al. Complete genome sequence of *Veillonella parvula* type strain (Te3T2010:2(1): 57-65.

Guisinger MM et al. Implications of the plastid genome sequence of Typha *(Typhaceae, Poales)* for understanding genome evolution in Poaceae. *J Mol Evol.* 2010 Jan 21.

Haley BJ et al. Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae. BMC Microbiol.* 2010 May 27;10:154.

Han C et al. Complete genome sequence of *Kangiella koreensis* type strain (SW-125T). *Standards in Genomic Sciences.* 2009:1(3):226-233.

Han JI et al. Complete genome sequence of the metabolically versatile plant growth-promoting endophyte, *Variovorax paradoxus* S110. *J Bacteriol.* 2010 Dec 23.

Harmon-Smith M et al. Complete genome sequence of *Sebaldella termitidis* type strain (NCTC 11300T). *Standards in Genomic Sciences.* 2010:2(2): 220-227.

Hasan NA et al. Comparative genomics of clinical and environmental *Vibrio mimicus. Proc Natl Acad Sci* U S A. 2010 Nov 15.

Hather GJ et al. The United States of America and scientific research. *PLoS One.* 2010 Aug 16;5(8). pii: e12203.

He S et al. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat Methods.* 2010 Oct;7(10):807-12.

He SM et al. Metatranscriptomic array analysis of *'Candidatus Accumulibacter phosphatis'*-enriched enhanced biological phosphorus removal sludge. *Environ Microbiol.* 2010 May;12(5):1205-17.

Hellsten U et al. The genome of the Western clawed frog *Xenopus tropicalis. Science.* 2010 Apr 30;328(5978):633-6.

Hemme CL et al. Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *ISME J.* 2010 May;4(5):660-72.

Hemme CL et al. Genome announcement: sequencing of multiple clostridia genomes related to biomass conversion and biofuels production. *J Bacteriol.* 2010 Dec;192(24):6494-6.

Hess M et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science.* 2011 Jan 28;331(6016):463-7.

Hill KK et al. Recombination and insertion events involving the botulinum neurotoxin complex genes in *Clostridium botulinum* types A, B, E and F and *Clostridium butyricum* type E strains. *BMC Biol.* 2009 Oct 5;7:66.

Hirschman L et al. Meeting Report: "Metagenomics, metadata and meta-analysis" (M3) workshop at the Pacific Symposium on Biocomputing 2010. *Standards in Genomic Sciences.* 2010:2(3): 357-360.

Hollister EB et al. Structure and dynamics of the microbial communities underlying the carboxylate platform for biofuel production. *Appl Microbiol Biotechnol.* 2010 Sep;88(1):389-99.

Hooper S et al. Estimating DNA coverage and abundance in metagenomes using a gamma approximation. *Bioinformatics.* 2010 Feb 1;26(3):295-301.

Hyatt D et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.* 2010 Mar 8;11(1):119.

Hyten DL et al. High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics.* 2010 Aug 16;11:475.

International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon. Nature.* 2010 Feb 11;463(7282):763-8.

Ivanova N et al. Complete genome sequence of *Geodermatophilus obscurus* type strain (G-20T). *Standards in Genomic Sciences.* 2010:2(2): 158-167.

Ivanova N et al. Complete genome sequence of *Haliangium ochraceum* type strain (SMP-2T). *Standards in Genomic Sciences.* 2010:2(1): 96-106.

Ivanova N et al. Complete genome sequence of *Gordonia bronchialis* type strain (3410T). *Standards in Genomic Sciences.* 2010:2(1): 19-28.

Ivanova N et al. A call for standardized classification of metagenome projects. *Environ Microbiol.* 2010 Jul;12(7):1803-5.

Jackson DJ et al. Parallel Evolution of Nacre Building Gene Sets in Molluscs. *Mol Biol Evol.* 2010 Mar;27(3):591-608.

Janssen PJ et al. The complete genome sequence of *Cupriavidus metallidurans* strain CH34, a master survivalist in harsh and anthropogenic environ-

ments. *PLoS One.* 2010 May 5;5(5):e10433.

Junier P et al. The genome of the Gram-positive metal- and sulfate-reducing bacterium *Desulfotomaculum reducens* strain MI-1. *Environ Microbiol.* 2010 Oct;12(10):2738-54.

Keim P et al. The genome and variation of *Bacillus anthracis. Mol Aspects Med.* 2009 Dec;30(6):397-405.

Kerfeld CA et al. Bacterial microcompartments. *Annu Rev Microbiol.* 2010 Oct 13;64:391-408.

Kerfeld CA and Gross L. Open education, open minds. *PLoS Biol.* 2010 Oct 5;8(10). pii: e1000508.

Kielak A et al. Phylogenetic and metagenomic analysis of *Verrucomicrobia* in former agricultural grassland soil. *FEMS Microbiol Ecol.* 2010 Jan;71(1):23-33.

Kim MY et al. Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proc Natl Acad Sci* U S A. 2010 Dec 21;107(51):22032-7.

Kinney JN et al. Comparative analysis of carboxysome shell proteins. *Photosynth Res.* 2011 Jan 30.

Kiss H et al. Complete genome sequence of *Denitrovibrio acetiphilus* type strain (N2460T) *Standards in Genomic Sciences.* 2010:2(3): 270-279.

Kouvelis VN et al. Complete genome sequence of ethanol producer *Zymomonas mobilis* NCIMB 11163. *J Bacteriol.* 2009 Nov;191(22):7140-1.

Kyrpides NC et al. Meeting report from the Genomic Standards Consortium (GSC) workshop 8. *Standards in Genomic Sciences.* 2010:3(1): 93-96.

LaButti K et al. Complete genome sequence of *Planctomyces limnophilus* type strain (Mü 290T). *Standards in Genomic Sciences.* 2010:3(1): 47-56.

LaButti K et al. Permanent draft genome sequence of *Dethiosulfovibrio peptidovorans* type strain (SEBR 4207T). *Standards in Genomic Sciences.* 2010:3(1): 85-92.

Lail K et al. Complete genome sequence of *Spirosoma linguale* type strain (1T). *Standards in Genomic Sciences.* 2010:2(2): 176-185.

Land M et al. Complete genome sequence of *Beutenbergia cavernae* type strain (HK1 0122T). *Standards in Genomic Sciences.* 2009:1(1):21-28.

Liolios K et al. The Genomes On Line Database (GOLD) in 2009: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D346-54.

Liolios K et al. Complete genome sequence of *Thermobispora bispora* type strain (R51T). *Standards in Genomic Sciences.* 2010:2(3): 318-326.

Lykidis A et al. The Complete multipartite genome sequence of *Cupriavidus necator* JMP134, a versatile pollutant degrader. *PLoS One.* 2011 Jan;5(1):122-30.

Lykidis A et al. Multiple syntrophic interactions in a terephthalate-degrading methanogenic consortium. *ISME J.* 2010 Aug 5.

Ma LJ et al. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium. Nature.* 2010 Mar 18;464(7287):367-73.

Macagno ER et al. Construction of a medicinal leech transcriptome database and its application to the identification of leech homologs of neural and innate immune genes. *BMC Genomics.* 2010 Jun 25;11:407.

Markowitz VM et al. The integrated microbial genomes system: An expanding comparative analysis resource. *Nucleic Acids Res.* 2010 Jan;38 (Database issue)D382-90.

Martin J et al. *Bacillus anthracis* genome organization in light of whole transcriptome sequencing. *BMC Bioinformatics.* 2010 Apr 29;11 Suppl 3:S10.

Martinez D et al. Annotation of protein-coding genes in fungal genomes. *Appl. Comput. Math.*, V.9, Special Issue, 2010, 56-65.

Mavromatis K et al. Complete genome sequence of *Alicyclobacillus acidocaldarius* type strain (104-IAT). *Standards in Genomic Sciences.* 2010:2(1): 9-18.

Mavromatis K et al. Gene context analysis in the integrated microbial genomes (IMG) data management system. *PLoS One.* 2009 Nov 24:4(11):e7979.

Mavromatis K et al. Complete genome sequence of *Coraliomargarita akajimensis* type strain (04OKA010-24 T). *Standards in Genomic Sciences.* 2010:2(3): 290-299.

Mavrommatis K et al. Complete genome sequence of *Vulcanisaeta distributa* type strain (IC-017T). *Standards in Genomic Sciences.* 2010:3(2): 117-125.

Mavrommatis K et al. Complete genome sequence of *Spirochaeta smaragdinae* type strain (SEBR 4228T). *Standards in Genomic Sciences.* 2010:3(2): 136-144.

McCarren J et al. Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proc Natl Acad Sci* U S A. 2010 Sep 21;107(38):16420-7.

McKinlay JB et al. A genomic perspective on the potential of *Actinobacillus succinogenes* for industrial succinate production. *BMC Genomics.* 2010 Nov 30;11:680.

McMurdie PJ et al. Localized plasticity in the streamlined genomes of vinyl chloride respiring *Dehalococcoides*. *PLoS Genet.* 2009 Nov;5(11):e1000714.

Miller TR et al. The genome sequence of the dioxin mineralizing bacterium *Sphingomonas wittichii* RW1. *J Bacteriol.* 2010 Nov;192(22):6101-2.

Munk C et al. Complete genome sequence of *Stackebrandtia nassauensis* type strain (LLR-40K-21T). *Standards in Genomic Sciences.* 2009:1(3):292-299.

Nalbantoglu U et al. Large direct repeats flank genomic rearrangements between a new clinical isolate of *Francisella tularensis* subsp tularensis A1 and Schu S4. *PLoS One.* 2010 Feb 3;5(2):e9007.

Nelson KE et al. A catalog of reference genomes from the human microbiome. *Science*. 2010 May 21;328(5981):994-9.

Nicolas FE et al. Endogenous short RNAs generated by Dicer 2 and RNA-dependent RNA polymerase 1 regulate mRNAs in the basal fungus *Mucor circinelloides*. *Nucleic Acids Res.* 2010 Sep;38(16):5535-41.

Nolan M et al. Complete genome sequence of *Rhodothermus marinus* type strain (R-10T). *Standards in Genomic Sciences.* 2009:1(3):283-291.

Nolan M et al. Complete genome sequence of *Streptosporangium roseum* type strain (NI 9100T). *Standards in Genomic Sciences.* 2010:2(1): 29-37.

Nolan M et al. Complete genome sequence of *Streptobacillus moniliformis* type strain (9901T). *Standards in Genomic Sciences.* 2009:1(3):300-307.

Novichkov PS et al. RegPrecise: A database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D111-8.

Novichkov PS et al. RegPredict: An integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Res.* 2010 Jul 1;38 Suppl:W299-307.

Ohm RA et al. Genome sequence of the model mushroom *Schizophyllum commune*. *Nat Biotechnol.* 2010 Sep;28(9):957-63. Epub 2010 Jul 11.

Pati A et al. Complete genome sequence of *Sphaerobacter thermophilus* type strain (S 6022T). *Standards in Genomic Sciences.* 2010:2(1): 49-56.

Pati A et al. GenePRIMP: A gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods.* 2010 Jun;7(6):455-7.

Pati A et al. Complete genome sequence of *Brachyspira murdochii* type strain (56-150T). *Standards in Genomic Sciences.* 2010:2(3): 260-269.

Pati A et al. Complete genome sequence of *Arcobacter nitrofigilis* type strain (CIT). *Standards in Genomic Sciences.* 2010:2(3): 300-308.

Pearson T et al. Phylogeographic reconstruction of a bacterial species with high levels of lateral gene transfer. *BMC Biol.* 2009 Nov 18;7:78.

Penn K et al. Genomic islands link secondary metabolism to functional adaptation in marine *Actinobacteria*. *ISME J.* 2009 Oct;3(10):1193-1203.

Pitluck S et al. Complete genome sequence of *Thermosediminibacter oceani* type strain (JW/IW-1228PT). *Standards in Genomic Sciences.* 2010:3(2): 108-116.

Pope PB et al. Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proc Natl Acad Sci U S A*. 2010 Aug 17;107(33):14793-8.

Prochnik SE et al. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science.* 2010 Jul 9;329(5988):223-6.

Pukall R et al. Complete genome sequence of *Slackia heliotrinireducens* type strain (RSH 1T). *Standards in Genomic Sciences.* 2009:1(3):234-241.

Pukall R et al. Complete genome sequence of *Jonesia denitrificans* type strain (Prevot 55134T). *Standards in Genomic Sciences.* 2009:1(3):262-269.

Pukall R et al. Complete genome sequence of *Conexibacter woesei* type strain (ID131577T). *Standards in Genomic Sciences.* 2010:2(2): 212-219.

Pukall R et al. Complete genome sequence of *Kribbella flavida* type strain (IFO 14399T). *Standards in Genomic Sciences.* 2010:2(2): 186-193.

Raha D et al. Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc Natl Acad Sci* U S A. 2010 Feb 23;107(8):3639-44. Epub 2010 Feb 5.

Ran L et al. Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS One.* 2010 Jul 8;5(7):e11486.

Reeve WG et al. Complete genome sequence of *Rhizobium leguminosarum* bv. *trifolii* strain WSM2304, an effective microsymbiont of the South American clover *Trifolium polymorphum*. *Standards in Genomic Sciences.* 2010:2(1): 66-76.

Reeve WG et al. Complete genome sequence of the Medicago microsymbiont *Ensifer (Sinorhizobium) medicae* strain WSM419. *Standards in Genomic Sciences.* 2010:2(1): 77-86.

Reeve WG et al. Complete genome sequence of *Rhizobium leguminosarum* bv. *trifolii* strain WSM1325, an effective microsymbiont of annual Mediterranean clovers. *Standards in Genomic Sciences.* 2010:2(3): 347-356.

Saunders E et al. Complete genome sequence of *Haloterrigena turkmenica* type strain (4kT). *Standards in Genomic Sciences.* 2010:2(1): 107-116.

Schmutz J et al. Genome sequence of the palaeopolyploid soybean. *Nature.* 2010 Jan 14; 463(7278):178-83.

Sieber JR et al. The genome of *Syntrophomonas wolfei:* New insights into syntrophic metabolism and biohydrogen production. *Environ Microbiol.* 2010 August: 12(8): 2289–2301.

Seidl V et al. Transcriptomic response of the mycoparasitic fungus *Trichoderma atroviride* to the presence of a fungal prey. *BMC Genomics*;10:567.

Sikorski J et al. Complete genome sequence of *Sulfurospirillum deleyianum* type strain (5175T). *Standards in Genomic Sciences.* 2010:2(2): 149-157.

Sikorski J et al. Complete genome sequence of *Segniliparus rotundus* type strain (CDC 1076T). *Standards in Genomic Sciences.* 2010:2(2): 203-211.

Sikorski J et al. Complete genome sequence of *Meiothermus silvanus* type strain (VI-R2T). *Standards in Genomic Sciences.* 2010:3(1): 37-46.

Sikorski J et al. Complete genome sequence of *Acetohalobium arabaticum* type strain (Z-7288T). *Standards in Genomic Sciences.* 2010:3(1): 57-65.

Spring S et al. Complete genome sequence of *Thermosphaera aggregans* type strain (M11TLT).

Standards in Genomic Sciences. 2010:2(3): 245-259.

Spring S et al. Complete genome sequence of *Desulfotomaculum acetoxidans* type strain (5575T). *Standards in Genomic Sciences.* 2009:1(3):242-253.

Spring S et al. Complete genome sequence of *Desulfohalobium retbaense* type strain (HR100T). *Standards in Genomic Sciences.* 2010:2(1): 38-48.

Srivastava M et al. Early evolution of the LIM homeobox gene family. *BMC Biol.* 2010 Jan 18;8:4.

Srivastava M et al. the *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature.* 2010 Aug 05;466(7307): 470-726.

Stein LY et al. Genome sequence of the obligate methanotroph, *Methylosinus trichosporium* strain OB3b. *J Bacteriol.* 2010 Dec;192(24):6497-8.

Stone CL et al. Analysis of the complete mitochondrial genome sequences of the soybean rust pathogens *Phakopsora pachyrhizi* and *P. meibomiae. Mycologia.* 2010 Jul-Aug;102(4):887-97.

Strnad H et al. Complete genome sequence of the photosynthetic purple nonsulfur bacterium *Rhodobacter capsulatus* SB 1003. *J Bacteriol.* Jul;192(13):345-6.

Suen G et al. An insect herbivore microbiome with high plant biomass-degrading capacity. *PLoS Genet.* 2010 Sep 23;6(9). pii: e1001129.

Tagmount A et al. The porcelain crab transcriptome and PCAD, the Porcelain Crab Microarray and Sequence Database. *PLoS One.* 2010 Feb 19;5(2):e9327.

Tice H et al. Complete genome sequence of *Nakamurella multipartita* type strain (Y-104T). *Standards in Genomic Sciences.* 2010:2(2): 168-175.

Tindall BJ et al. Complete genome sequence of *Halomicrobium mukohataei* type strain (arg-2T). *Standards in Genomic Sciences.* 2009:1(3):270-277.

Tindall BJ et al. Complete genome sequence of *Meiothermus ruber* type strain (21T). *Standards in Genomic Sciences.* 2010:3(1): 26-36.

Tyler L, Bragg JN, Wu J, Yang X, Tuskan GA, Vogel JP. Annotation and comparative analysis of the glycoside hydrolase genes in *Brachypodium distachyon. BMC Genomics.* 2010 Oct 25;11:600.

Vanden Wymelenberg A et al. Comparative transcriptome and secretome analysis of wood decay fungi *Postia placenta* and *Phanerochaete chrysosporium. Appl Environ Microbiol.* 2010 Jun;76(11)3599-610.

Visel A et al. Targeted deletion of the 9p21 noncoding coronary artery disease risk interval in mice. *Nature.* 2010 Mar 18;464(7287):409-12.

Voelker SL et al. Antisense down-regulation of 4CL expression alters lignification, tree growth and saccharification potential of field-grown poplar. *Plant Physiol.* 2010 Oct;154(2)874-86.

von Jan M et al. Complete genome sequence of *Archaeoglobus profundus* type strain (AV18T). *Standards in Genomic Sciences.* 2010:2(3): 327-346.

Walker CB et al. *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci* U S A. 2010 May 11;107(19):8818-23.

Walsh DA et al. Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science.* 2009 Oct 23;326(5952):578-82.

Wang S et al. Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies. *Genome Biol.* 2010 Jan 22;11(1):R8.

Wang Z et al. SoyDB: A knowledge database of soybean transcription factors. *BMC Plant Biol.* 2010 Apr;76(7):2091-7.

Wei D et al. Laccase and its role in the production of extracellular reactive oxygen species during wood decay by the brown rot basidiomycete *Postia placenta. Appl Environ Microbiol.* 2010 Feb 12. [Epub ahead of print]

Wilson A et al. Structural determinants underlying photoprotection in the photoactive orange carotenoid protein of cyanobacteria. *J Biol Chem.* 2010 Jun 11;285(24):18364-75.

Wirth R et al. Complete genome sequence of *Thermocrinis albus* type strain (HI 11/12T). *Standards in Genomic Sciences.* 2010:2(2): 194-202.

Woyke T et al. One bacterial cell, one complete genome. *PLoS One.* 2010 Apr 23;5(4):e10314.

Wu D. A phylogeny-driven genomic encyclopedia of bacteria and archaea. *Nature.* 2009 Dec 24;462(7276):1056-60.

Wurtzel O et al. A single-base resolution map of an archaeal transcriptome. *Genome Res.* 2010 Jan;20(1):133-41.

Yan H et al. Genome-wide mapping of cytosine methylation revealed dynamic DNA methylation patterns associated with genes and centromeres in rice. *Plant J.* 2010 August.

Yasawong M et al. Complete genome sequence of *Arcanobacterium haemolyticum* type strain (11018T). Standards in *Genomic Sciences.* 2010:3(2): 126-135.

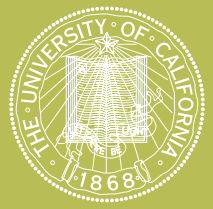
Yilmaz S et al. Fixation-free fluorescence *in situ* hybridization for targeted enrichment of microbial populations. *ISME J.* 2010 Oct;4(10):1352-6.

Yilmaz S et al. Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat Methods.* 2010 Dec;7(12):943-4.

Yin T, Zhang X, Gunter L, Priya R, Sykes R, Davis M, Wullschleger SD, Tuskan GA. Differential detection of genetic Loci underlying stem and root lignin content in *Populus. PLoS One.* 2010 Nov 22;5(11):e14021.

Young M et al. Genome sequence of the Fleming strain of *Micrococcus luteus,* a simple free-living actinobacterium. *J Bacteriol.* 2010 Feb;192(3):841-60.

U.S. DEPARTMENT OF
ENERGY



The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

CSO J020845


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