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

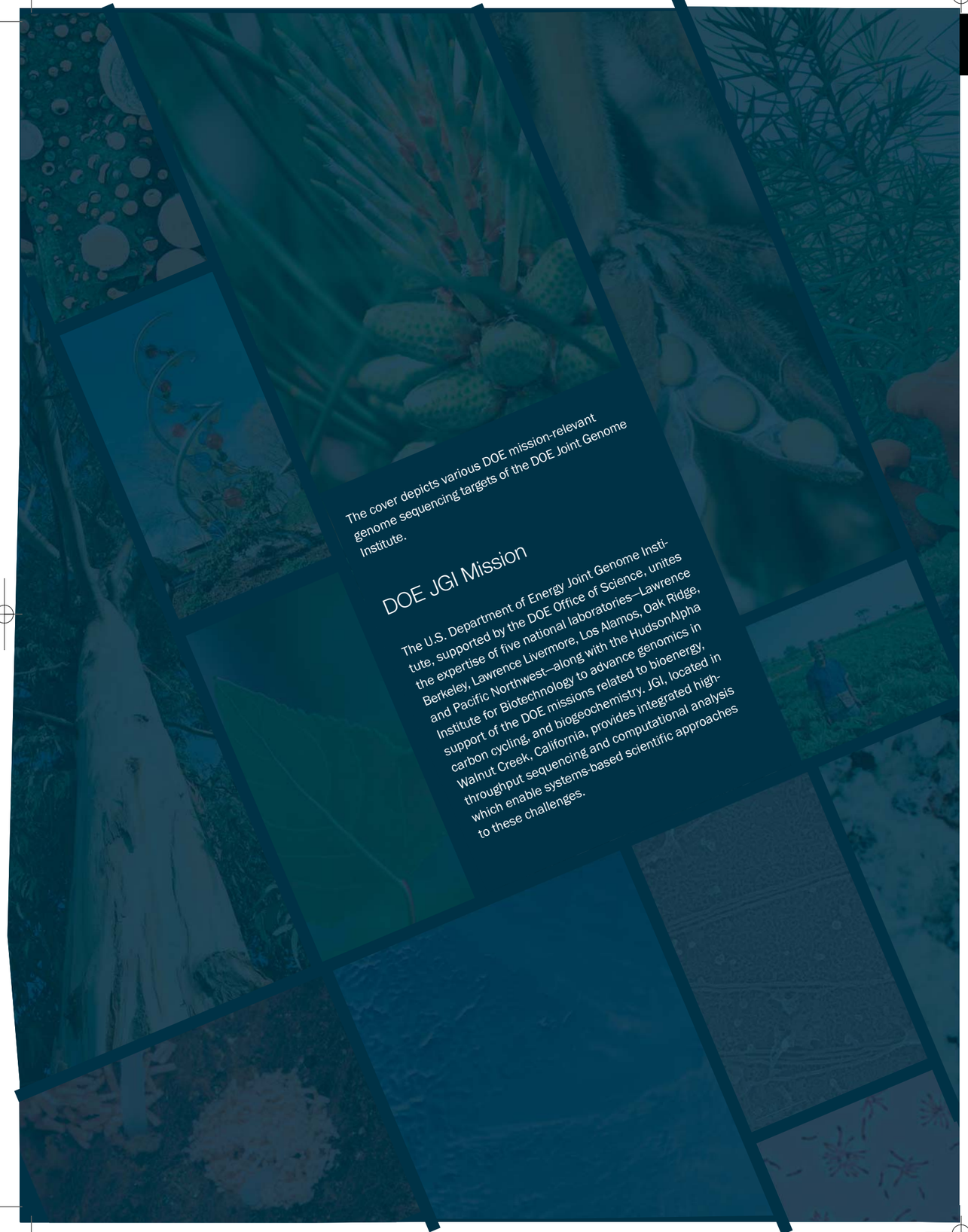
U.S. DEPARTMENT OF ENERGY

Joint Genome Institute

U.S. DEPARTMENT OF ENERGY
Office of Science

2008
Progress
Report

DOE JGI—powering a sustainable future with the science we need for biofuels, environmental cleanup, and carbon capture.



This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

For more information about JGI, contact:
David Gilbert, Public Affairs Manager
DOE Joint Genome Institute
2800 Mitchell Drive
Walnut Creek, CA 94598
e-mail: degilbert@lbl.gov
phone: (925) 296-5643

JGI Web site:
http://www.jgi.doe.gov/

Published by the Berkeley Lab Creative Services Office in collaboration with DOE JGI researchers and staff for the U.S. Department of Energy.

LBNL-1597E
CSO 15866


JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE
U.S. DEPARTMENT OF ENERGY
Joint Genome Institute
U.S. DEPARTMENT OF
ENERGY
Office of Science
2008
Progress
Report

table of contents

While initially a virtual institute, the driving force behind the creation of the DOE Joint Genome Institute in Walnut Creek, California in the Fall of 1999 was the Department of Energy's commitment to sequencing the human genome. With the publication in 2004 of a trio of manuscripts describing the finished "DOE Human Chromosomes," the Institute successfully completed its human genome mission. In the time between the creation of the DOE JGI and completion of the Human Genome Project, sequencing and its role in biology spread to fields extending far beyond what could be imagined when the Human Genome Project first began. Accordingly, the targets of the DOE JGI's sequencing activities changed, moving from a single human genome to the genomes of large numbers of microbes, plants, and other organisms, and the community of users of DOE JGI data similarly expanded and diversified. Transitioning into operating as a user facility, the DOE JGI modeled itself after other DOE user facilities, such as synchrotron light sources and supercomputer facilities, empowering the science of large numbers of investigators working in areas of relevance to energy and the environment.

The JGI's approach to being a user facility is based on the concept that **by focusing state-of-the-art sequencing and analysis capabilities on the best peer-reviewed ideas drawn from a broad community of scientists, the DOE JGI will effectively encourage creative approaches to DOE mission areas and produce important science.** This clearly has occurred, only partially reflected in the fact that the DOE JGI has played a major role in more than 45 papers published in just the past three years alone in *Nature* and *Science*. The involvement of a large and engaged community of users working on important problems has helped maximize the impact of JGI science.

A seismic technological change is presently underway at the JGI. The Sanger capillary-based sequencing process that dominated how sequencing was done in the last decade is being replaced by a variety of new processes and sequencing instruments. The JGI, with an increasing number of next-generation sequencers, whose throughput is 100- to 1,000-fold greater than the Sanger capillary-based sequencers, is increasingly focused in new directions on projects of scale and complexity not previously attempted.

These new directions for the JGI come, in part, from the 2008 National Research Council report on the goals of the National Plant Genome Initiative as well as the 2007 National Research Council report on the New Science of Metagenomics. Both reports outline a crucial need for systematic large-scale surveys of the plant and microbial components of the biosphere as well as an increasing need for large-scale analysis capabilities to meet the challenge of converting sequence data into knowledge. The JGI is extensively discussed in both reports as vital to progress in these fields of major national interest. JGI's future plan for plants and microbes includes a systematic approach for investigation of these organisms at a scale requiring the special capabilities of the JGI to generate, manage, and analyze the datasets. JGI will generate and provide not only community access to these plant and microbial datasets, but also the tools for analyzing them. These activities will produce essential knowledge that will be needed if we are to be able to respond to the world's energy and environmental challenges.

As the JGI Plant and Microbial programs advance, the JGI as a user facility is also evolving. The Institute has been highly successful in bending its technical and analytical skills to help users solve large complex problems of major importance, and that effort will continue unabated. The JGI will increasingly move from a central focus on "one-off" user projects coming from small user communities to much larger scale projects driven by systematic and problem-focused approaches to selection of sequencing targets. Entire communities of scientists working in a particular field, such as feedstock improvement or biomass degradation, will be users of this information. Despite this new emphasis, an investigator-initiated user program will remain. This program in the future will replace small projects that increasingly can be accomplished without the involvement of JGI, with imaginative large-scale "Grand Challenge" projects of foundational relevance to energy and the

DIRECTOR'S PERSPECTIVE:

DOE JGI at a decade and looking forward

environment that require a new scale of sequencing and analysis capabilities.

Close interactions with the DOE Bioenergy Research Centers, and with other DOE institutions that may follow, will also play a major role in shaping aspects of how the JGI operates as a user facility.

Based on increased availability of high-throughput sequencing, the JGI will increasingly provide to users, in addition to DNA sequencing, an array of both pre- and post-sequencing value-added capabilities to accelerate their science. The obstacles for doing genomic-based investigations in the future will be less the generation of sequence and more the isolation of the specific material that investigators want to sequence and the informative analysis of the sequence data after it is generated. Among the pre-sequencing capabilities that the DOE JGI is beginning to and will increasingly offer to users in the future are a variety of robust ways of isolating DNA from hard-to-culture environmental microbes (single-cell sorting, amplification, high-throughput culturing) as well as scaleable capabilities for capturing specific sequences of biological interest from huge plant genomes and environmental metagenomes.

A crucial post-sequencing bottleneck that the JGI is increasingly focusing on is the analysis of genomic data. It is already apparent that the application of genomics to problems in energy and the environment is being limited by the scientific community's inability to capture and in-

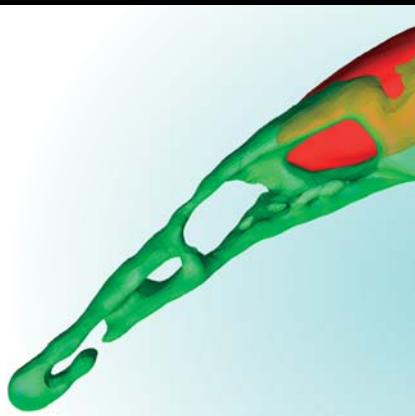
terpret the dramatically increasing volume of data being generated. The JGI's development of tools and infrastructure for assisting users in the analysis of sequence data are helping advance the JGI as a user facility providing powerful and diverse services for the derivation of useful knowledge from genomic data. With its capabilities, including experienced individuals knowledgeable in state-of-the-art sequence analysis as well as computing resources far beyond that found in small centers, the JGI has and will continue to offer a unique breadth of genomic resources to enable users to accomplish the science necessary for meeting present and future energy and environmental challenges.

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

The helical sculpture was created by Jeff Brees of Markleeville, California who specializes in topiary and garden sculpture in wire and hand-wrought metal. He says of the sculpture: "Here's a DNA molecule you can see without a microscope!"

A new BenchCel Microplate Handling System robot was borrowed from the DNA sequencing line to cut the ribbon and usher in JGI's second decade and to dedicate a newly renovated building at its Walnut Creek facility. From left, JGI's Operations Manager Ray Turner, JGI Director Eddy Rubin, Congresswoman Ellen Tauscher's Field Representative Erik Ridley, and Walnut Creek City Council member Susan McNulty Rainey.

In July 2008, JGI opened a laboratory for two new sequencing technologies. Both the Illumina *(left top)* and 454 *(left bottom)* technologies have since been integrated into JGI's pipeline, and the throughput is currently being scaled for both platforms. These rooms were designed with the technician in mind and feature advanced ergonomic workstations, equipment, and tools that allow the technicians to safely and efficiently prepare and sequence samples.



2008 news highlights

The DOE JGI reported a number of accomplishments in 2008, from the completion of the soybean genome to the identification of a fungus that could help improve biofuel production.

Soybean Genome Completed

The DOE JGI released a complete draft assembly of the soybean *(Glycine max)* genetic code in December, making it widely available to the research community to advance new breeding strategies for one of the world's most valuable plant commodities. Soybean accounts for 70% of the world's edible protein and is an emerging feedstock for biodiesel production. Soybean is second only to corn as an agricultural commodity and is the leading U.S. agricultural export. DOE JGI's interest in sequencing the soybean centers on its use in biodiesel, a renewable alternative fuel, with the highest energy content of any alternative fuel. According to 2007 U.S. Census data, soybean is estimated to be responsible for more than 80% of biodiesel production.

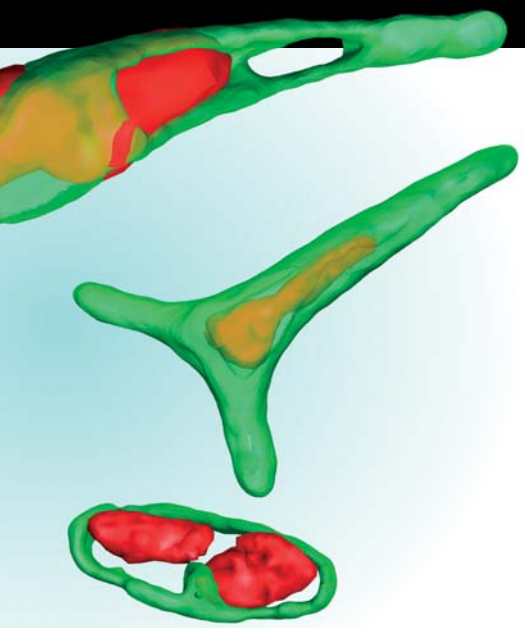
The soybean genome project is already making its mark out in the field. "Now every breeder can go into this valuable library for the information that will help speed up the breeding process," said Rick Stern, a New Jersey soybean farmer and chair of the Production Research program for the United Soybean Board (USB). "It should cut traditional breeding time by half from the typical 15 years."

This soybean genome effort was led by Dan Rokhsar and Jeremy Schmutz of the DOE JGI, Gary Stacey of the University of Missouri-Columbia, Randy Shoemaker of the U.S. Department of Agriculture-Agricultural Research Service (USDA-ARS), and Scott Jackson of Purdue University, with support from the DOE, the USDA, and the National Science Foundation (NSF), the United Soybean Board, the North Central Soybean Research Program, and the Gordon and Betty Moore Foundation.

Diatom Genome Helps Explain Their Success in Capturing Carbon

Diatoms, mighty microscopic algae, have profound influence on climate, producing 20% of the oxygen we breathe by capturing atmospheric carbon and, in so doing, countering the greenhouse effect. Since their evolutionary origins, these photosynthetic wonders have come to acquire advantageous genes from bacterial, animal, and plant ancestors, enabling them to thrive in today's oceans. These findings, based on the analysis of the latest sequenced diatom genome, *Phaeodactylum tricornutum,* were published in October in the journal *Nature* by an international team of researchers led by the DOE JGI and the Ecole Normale Supérieure of Paris.

"These organisms represent a veritable melting pot of traits—a hybrid of genetic mechanisms contributed by ancestral lineages of plants, animals, and bacteria, and optimized over the relatively short evolutionary timeframe of 180 million years



since they first appeared," said first author Chris Bowler of the Ecole Normale Supérieure. Bowler speculates that the diatom uses urea to store nitrogen, not to eliminate it like animals do, because nitrogen is a precious nutrient in the ocean. What's more, the alga draws the best of both worlds—it can convert fat into sugar, as well as sugar into fat—extremely useful in times of nutrient shortage.

Diatoms, encapsulated by elaborate lacework shells made of glass, are only about one-third of a strand of hair in diameter. "The diatom genomes will help us to understand how they can make these structures at ambient temperatures and pressures, something that humans are not able to do. If we can learn how they do it, we could open up all kinds of new nanotechnologies, like for building miniature silicon chips or for biomedical applications," said Bowler.

hypersaline mat

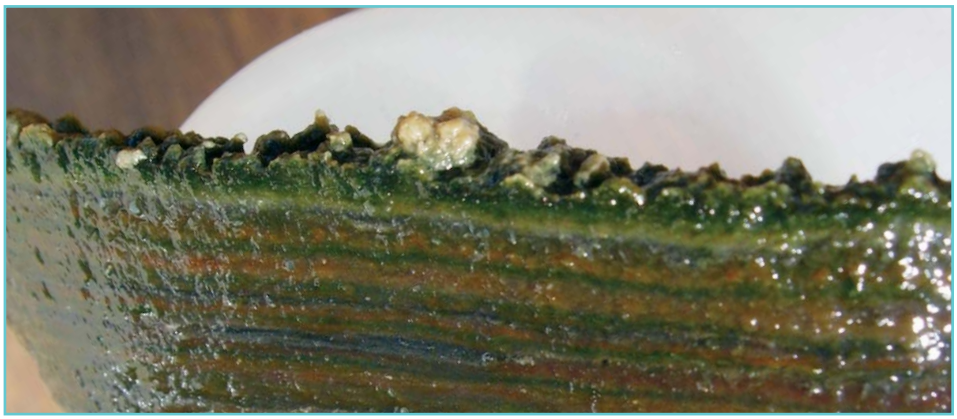


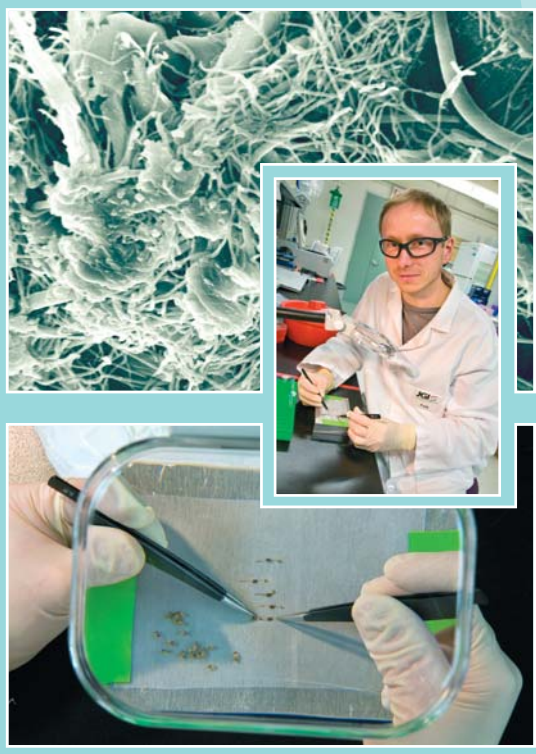
Image courtesy of John R. Spear, Colorado School of Mines

Metagenomics Comes of Age

Mostly hidden from the scrutiny of the naked eye, microbes have been said to run the world. The challenge is how best to characterize them, given that fewer than 1% of the estimated hundreds of millions of microbial species can be cultured in the laboratory. The answer is metagenomics—an increasingly popular approach for extracting the genomes of uncultured microorganisms and discerning their specific metabolic capabilities directly from environmental samples. Now, 10 years after the term was coined, metagenomics is already paying dividends, according to a Q&A by the head of DOE JGI's microbial ecology program, Philip Hugenholtz, and Massachusetts Institute of Technology researcher Gene Tyson, which was published in September in the News and Views section of the journal *Nature*.

"Metagenomic tools are becoming more widely available and improving at a steady pace, but there are still computational and other bottlenecks, such as the high percentage of uncharacterized genes emerging from metagenomic studies," Hugenholtz said. Hugenholtz and Tyson go on in the *Nature* article to cite the emergence of the next-generation of sequencing technologies, already creating a deluge of data that has outstripped the computational power available to cope with it. Hugenholtz cautions that it is not necessary to compare all the data to glean biological insights. "What we can capture will help steer the direction toward a relevant data subset to investigate," he said. "We are still far from capturing and characterizing the dazzling diversity of the microbial life on earth—but at least we have hit upon the gold standard for scratching the surface."

JGI IN



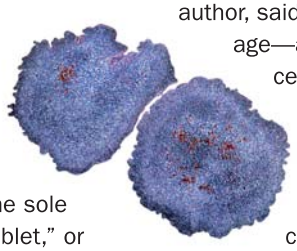
Data Management System Extended to Education Community

In September, a special version of the Integrated Microbial Genomes (IMG) data management system, IMG/EDU, was established to support DOE JGI's Education Program in Microbial Genome Annotation for teaching microbial genome analysis and annotation, using specific microbial genomes in the comparative context of all the genomes available in IMG.

IMG/EDU will serve as the core of a Web-based portal that enables undergraduates to participate in microbial genome annotation, said Cheryl Kerfeld, head of DOE JGI's Education Program. Currently, students at 12 schools nationally are using the portal in molecular biology, genetics, microbiology, and biochemistry courses in which they examine gene predictions and annotate genes and biochemical pathways. By helping to build curated genomes with researchers across the globe, undergraduates will discover the concepts and applications of bioinformatics using IMG/EDU. An additional resource, IMG/ACT, provides support for managing student classes and assignments, as well as for sharing teaching materials and guiding students in their study of gene predictions and functional annotations.

Genome of Simplest Animal Reveals Ancient Lineage

As Aesop said, appearances are deceiving—even in life's tiniest creatures. *Trichoplax adhaerens,* a simple and primitive animal, was first detected in the 1880s clinging to the sides of an aquarium, but its recent characterization by the DOE JGI reveals that it harbors a more complex suite of capabilities than meets the eye. The findings, reported in August in the journal *Nature*, establish a group of organisms as a branching point of animal evolution and identify sets of genes, or a "parts list," employed by organisms that have evolved along particular branches.



The analysis of the 98 million base-pair genome of *Trichoplax* (literally "hairy-plate") illuminates its ancestral relationship to other animals. *Trichoplax* is the sole member of the placozoan ("tablet," or "flat" animal) phylum, whose relationship to other animals, such as bilaterians (which include humans, flies, worms, and snails) and cnidarians (jellyfish, sea anemones, and corals), and sponges is much debated. Originally collected from the Red Sea, and cultured over the last 40 years in the laboratory, *Trichoplax* is a two-millimeter flat disk containing fluid sandwiched between two cell layers. It lacks organs and has only four or five cell types.

Mansi Srivastava, the study's first author, said, "*Trichoplax* is an ancient lineage—a good representation of the ancestral genome that is shedding light on the kinds of genes, the structures of genes, and even how these genes were arranged on the genome in the common ancestor 600 million years ago. It has retained a lot of primitive features relative to other living animals." Srivastava, a graduate student at the Center for Integrative Genomics at the University of California, Berkeley, works under the direction of Daniel Rokhsar, the publication's senior author, who is DOE JGI's head of Computational Genomics Program and a professor of Genetics, Genomics, and Development at UC Berkeley.

THE NEWS

Smithsonian

Termite Bellies and Biofuels

(Published in Smithsonian, *August 1, 2008)*

Scientist Falk Warnecke's research into termite digestion may hold solutions to our energy crisis. The insects are remarkably efficient at turning cellulose into sugar—the first step in making fuel from plants like switchgrass or poplar trees. Scientists can't compete with termites. They can break apart cellulose's tough bonds in the lab, but the enzymes they use are wildly, prohibitively expensive. That's where Warnecke comes in. His research has some people salivating at the prospect of dipping into the termites' microbial stew and pulling out a few enzymes that would finally make it possible to produce ethanol from cellulose on an industrial scale. *(www.smithsonianmag.com/science-nature/termites-bellies-biofuels.html)*

the Atlantic

Gut Reactions

(Published in The Atlantic, *August 20, 2008)*

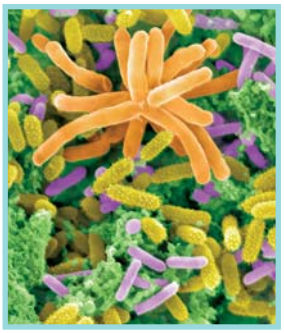
The termite's stomach, of all things, has become the focus of large-scale scientific investigations. Where humans have failed, the termite succeeds—spectacularly. A worker termite tears off a piece of wood with its mandibles and lets its guts work on it like a molecular wrecking yard, stripping away sugars, CO_2, hydrogen, and methane with 90% efficiency. Offer a termite this page, and its microbial helpers will break it down into two liters of hydrogen, enough to drive more than six miles in a fuel-cell car. *(www.theatlantic.com/doc/200809/termites)*

Biofuel's Holy Grail: Shipworms?

(Published in The Perch, *blog of* Audubon Magazine, *October 9, 2008)*

A group of Philippine and American scientists were awarded a $4-million grant from the National Institutes of Health, with support from the U.S. Department of Energy and the National Science Foundation, to survey mollusks in the Philippine archipelago for chemicals that could yield nervous system disorder and cancer drugs and enzymes capable of breaking down cellulose into ethanol, considered by some energy experts as the holy grail in the quest for viable biofuels. An estimated 10,000 marine mollusks inhabit the Philippines, including the giant clam and blue-ring octopuses, but it's the shipworms, which rely on bacteria to metabolize wood fibers, that pose the most exciting prospects for alternative energy. *(magblog.audubon.org/node/165)*

Lake Washington Microbes Yield Clues to Methane Generation



Today's powerful sequencing machines can rapidly read the genomes of entire communities of microbes, but the challenge is to extract meaningful information from the jumbled reams of data. In *Nature Biotechnology* in August, a paper by researchers at the University of Washington and the DOE JGI described a novel approach for extracting single genomes and discerning specific microbial capabilities from mixed community ("metagenomic") sequence data. The research team explored the sediments in Lake Washington, near Seattle, and characterized biochemical pathways associated with nitrogen cycling and methane utilization in an effort to understand methane generation and consumption by microbes.

Most of the microbes that oxidize single-carbon compounds are unculturable and therefore unknown, as are the vast majority of microbes on Earth. To find species of interest, the researchers sequenced microbial communities from Lake Washington sediment samples because lake sediment is known to be a site of high methane consumption. However, these sediment samples contained over 5,000 species of microbes performing a complex array of biochemical tasks.

Using an enrichment technique applied for the first time to microbial community samples, the researchers adapted a technique called stable isotope probing to enrich the samples for the microbes of interest. According to the paper's senior author, Ludmila Chistoserdova, a microbiologist at the University of Washington, five different single-carbon compounds were labeled with a heavy isotope of carbon and fed to a separate sediment sample. The microbes that could consume the compound incorporated the labeled carbon into their DNA, Chistoserdova said, while organisms that could not use the compound did not incorporate the label. The labeled DNA was then separated out and sequenced, producing microbial "subsamples" highly enriched for organisms that could metabolize methane, methanol, methylated amines, formaldehyde, and formate.

DOE JGI Director Sows New Ideas for Biofuels

Genomics is accelerating improvements for converting plant biomass into biofuel, Eddy Rubin, director of the DOE JGI, reported in August in the journal *Nature*. In an article entitled "The Genomics of Cellulosic Biofuels," Rubin argued that despite the barriers to improving biofuels, genetics and genomics can catalyze progress towards delivering economically viable and more socially acceptable biofuels based on lignocellulose.

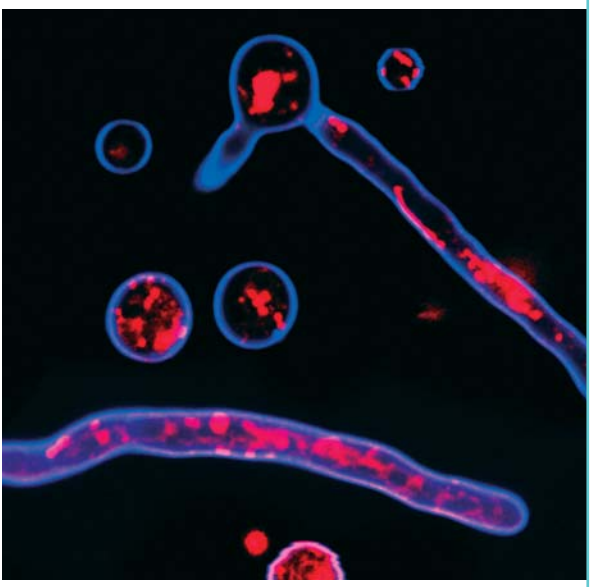
The production processes involved include the harvesting of biomass, pretreatment, and saccharification, which results in the deconstruction of cell wall polymers into component sugars and the conversion of those sugars into biofuels through fermentation. "With the data that we are generating from plant genomes," Rubin said, "we can home in on relevant agronomic traits such as rapid growth, drought resistance, and pest tolerance, as well as those that define the basic building blocks of the plants' cell wall—cellulose, hemicellulose and lignin." Biofuels researchers are able to use this information to optimize the plants' use as biofuels feedstocks—for example, altering branching habit, stem thickness, and cell wall chemistry to produce plants that are less rigid and more easily broken down.

Lancelet Genome Shows How Genes Quadrupled During Vertebrate Evolution

The newly sequenced genome of a dainty, quill-like sea creature called a lancelet provides the best evidence yet that vertebrates evolved over the past 550 million years through a four-fold duplication of the genes of more primitive ancestors.

The late geneticist Susumu Ohno argued in 1970 that gene duplication was the most important force in the evolution of higher organisms, and Ohno's theory was the basis for original estimates that the human genome must contain up to 100,000 distinct genes. Instead, the Human Genome Project found that humans today have only 20,000 to 25,000 genes, which means that, if our ancestors' primitive genome doubled and redoubled, most of the duplicate copies of genes must have been lost. An analysis of the lancelet, or amphioxus, genome, published in the June 19 issue of *Nature*, shows this to be the case.

"Amphioxus and humans had a common ancestor 550 million years ago, which allows us to use amphioxus as a surrogate for that ancestor in terms of understanding how vertebrate genomes evolved," said Daniel S. Rokhsar, program head for computational genomics at the DOE JGI. Rokhsar and JGI post-doctoral fellow Nicholas H. Putnam performed the sequencing, assembly, and genome-wide analyses of the amphioxus genome and are lead authors of the *Nature* paper.



Micrograph of T. reesei *hyphae with vesicle membranes stained red and cell wall chitin in blue, courtesy of Mari Valkonen, VTT Finland.*

Tent Fungus Could Help Produce Biofuels

The bane of military quartermasters may soon be a boon to biofuels producers. In World War II, *Trichoderma reesei* was identified as the culprit responsible for the deterioration of fatigues and tents in the South Pacific. This progenitor strain has since yielded variants for broad industrial applications and is known today as an abundant source of enzymes, particularly cellulases and hemicellulases, which are being explored to catalyze the deconstruction of plant cell walls as a first step towards the production of biofuels from lignocellulose.

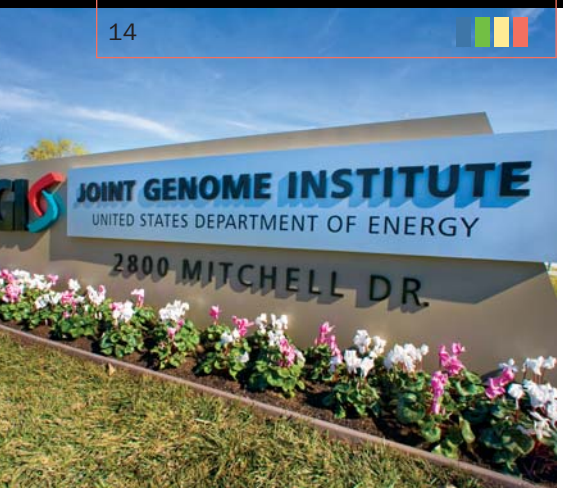
A genome analysis published in May in *Nature Biotechnology* by researchers from DOE JGI and Los Alamos National Laboratory revealed the surprisingly slim repertoire of genes that T. reseei uses to break down plant cell walls. This provides a roadmap for accelerating research to optimize fungal strains for reducing the steep cost of converting lignocellulose to fermentable sugars. Improved industrial enzyme "cocktails" from *T. reseei* and other fungi, according to Eddy Rubin, director of DOE JGI and one of the paper's senior authors, will enable more economical conversion into next-generation biofuels of biomass from such feedstocks as the perennial grasses *Miscanthus* and switchgrass, wood from fast-growing trees like poplar, agricultural crop residues, and from municipal waste.

Fruiting body of the ectomycorrhizal basidiomycete Laccaria bicolor *S238N interacting with Douglas fir seedlings (© Francis Martin, INRA)*

DOE JGI Collaborates on Study of Plant-Fungi Symbiosis

Plants gained their ancestral toehold on dry land with considerable help from symbiotic fungi. Now, that partnership is being exploited as a way of bolstering biomass production for next-generation biofuels. The genetic mechanism of this symbiosis, which contributes to the delicate ecological balance in healthy forests, provides many insights into plant health. This may show the way to more efficient carbon sequestration and phytoremediation, using plants to clean up environmental contaminants.

DOE JGI's genome analysis of the symbiotic fungus *Laccaria bicolor*—with collaborators from INRA, the National Institute for Agricultural Research in Nancy, France, and contributions from 16 institutions—was published in March in the journal *Nature.* Trees' ability to generate large amounts of biomass or store carbon is underpinned by their interactions with soil microbes known as mycorrhizal fungi, which process scarce nutrients, such as phosphate and nitrogen, which are transferred to the growing tree. Forests around the world rely on the partnership between plant roots and soil fungi and the environment they create, the rhizosphere. The *Laccaria* genome represents a valuable resource, the first of a series of tree community genomics projects to have passed through the DOE JGI production sequencing line.



The DOE JGI in Walnut Creek, California carries out production-scale sequencing and analysis augmented by the specialized tasks drawn from the specific capabilities provided by its partner laboratories, Lawrence Berkeley National Laboratory (LBNL), 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 (formerly associated with the Stanford Human Genome Center).

The DOE JGI Director, working with the DOE JGI Joint Management Board (JMB), which is comprised of two 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 JMB, communicates the JGI's technological and strategic vision to the DOE's Office of Biological and Environmental Research (OBER), and funding decisions are based upon these discussions.

The JGI headquarters in Walnut Creek is staffed by LLNL and LBNL employees, while the partner laboratories represent components of large national laboratories and one research institute.

JGI Directors

Eddy Rubin, JGI Director, provides scientific and managerial guidance for all aspects of the JGI. This includes responsibilities toward the JGI internal scientific programs, the partner DOE National Laboratories, and external collaborations. In addition to serving as the DOE JGI Director, Rubin is Genomics Division Director at the Lawrence Berkeley National Laboratory. He received a B.S. in Physics from the University of California, San Diego and a Ph.D. in Biophysics and an M.D. from the University of Rochester. Following clinical training in Medical Genetics at the University of California, San Francisco, he moved across the bay to join LBNL to develop computational and biological approaches to the analysis of DNA sequence data.

In 2002, he became the Director of the JGI. Under his leadership, the JGI was one of the five leading institutions involved in the Human Genome Project. Following the completion of the Human Genome Project, he led the transition of the JGI into a genomic user facility focused on the sequencing and analysis of organisms of relevance to bioenergy, carbon cycling, and bioremediation. Rubin is internationally recognized for his research on the conversion of genomic data to biological information in a variety of fields. He has authored more than 200 research papers, and the current work in his laboratory includes the development of computational tools for cross-species sequence analysis, genome-wide studies of gene regulation, Neanderthal genomics, and microbial community genomics.

Jim Bristow, Deputy Director of Scientific Programs, provides direction, leadership, and oversight for all JGI scientific programs, including the Evolutionary Genomics, Genomic Technologies, Genetic Analysis, Computational Genomics, Microbial Ecology, Microbial Genome Analysis, and Computational Genomics Programs. In addition, the Deputy is responsible for oversight and administration of the Community Sequencing Program (CSP). He is also responsible for formulation and implementation of scientific policy, engaging in interactions with the scientific community at large, managing JGI's long-term capacity plan, and overseeing JGI user programs and the project management office.

After receiving his medical degree from Harvard University, Bristow was a member of the UCSF faculty for over 15 years, directing a molecular genetics research lab and working as a pediatric cardiologist. He has held leadership positions on several grant review boards, including the American Heart Association and the National Institutes of Health. He joined the JGI in January 2004 and has developed and implemented the CSP, which provides large-scale DNA sequencing support for investigator-initiated projects spanning the range of modern biology, from microbial communities to human variation and disease. As Deputy Director of Programs, he oversees the various research programs at the JGI as well as its activities as a user facility. His research interests include extracellular matrix biology and genetics and recently, microbial diversity.

Vito Mangiardi, Deputy Director for Business Operations and Production, is responsible for leading and managing the Production Sequencing, Informatics, and Operations Departments at the JGI. He ensures the highest quality operational effectiveness for the JGI and coordinates with the Deputy Director of Programs to facilitate efficient production and operational processes. He provides leadership for the activities of multiple departments and oversight and direction of the safety group and Environmental Health and Safety efforts. He is expected to elevate the JGI to dominance in high-throughput genomic data generation through strategic planning for production sequencing, informatics and operational efficiencies.

Mangiardi joined the JGI in October 2008, bringing over 24 years of national and international management experience in the pharmaceutical, biotechnology, health services, and diagnostic industries. He has held positions including President/CEO, Executive Vice President, Senior Vice President/COO, and Operations Manager and has been responsible for Management, Operations, Sales/Marketing, and Sciences. Most recently, he has served as Chief Executive Officer at Bilcare, Global Clinical Supplies.

JGI Department Heads

Susan Lucas, Sequencing Department Head, provides direction and leadership for the Production Sequencing efforts at the JGI. She received her B.A. in biology from the University of Oregon and joined Lawrence Livermore National Lab in 1997 to work on the Human Genome Project. She started work as a technician running ABI 377s and worked her way through the sequencing process. During her time in production, Lucas has overseen a tremendous increase in throughput and reduction in sequencing cost per lane. This growth has enabled the sequencing of three human chromosomes and many model organisms.

Eveline Dube, Informatics Department Head (acting), assumed an interim leadership role of the Informatics Department in January 2009. The Department Head provides leadership for computational science, data management, software and web development, and systems administration, and is responsible for defining, assessing, communicating, and implementing JGI's informatics infrastructure plan. The Department Head has broad latitude in planning and scheduling work to meet JGI informatics goals and objectives, including the formulation and direction of informatics projects to enhance the JGI's informatics infrastructure. Dube is the former Division Lead of the Science & Technology Computing division at LLNL.

Ray Turner, Operations Department Head, provides oversight and leadership of JGI operations infrastructure, including space and facilities, finance and procurement, human resources, and administrative activities. Turner received his B.S. in Finance from the University of Utah and was commissioned into the U.S. Navy in 1980. After completion of flight training, he served in various Strategic Management and training positions within the Navy. In 1995, he received a Master's Degree in financial management from the Naval Postgraduate School and served as the comptroller at Naval Air Station Alameda during the base closure process. After completion of a 21-year Navy career, he joined a San Francisco Bay Area healthcare information technology company where he served as the Vice President of Finance and Vice President/General Manager of the Health Information Management Division. Turner joined the JGI in October 2005 and oversees the Finance, Human Resources, Facilities, and Administration departments at the facility.

DOE JGI departments and programs

JGI Department Heads *continued*

Daniel Rokhsar, Computational Genomics Department Head, oversees the development of new analytical tools and data management systems that transform the raw data produced by the JGI into biologically valuable information and insights. These tools are designed to facilitate the use of JGI data by the biological community. This work is essential for managing and visualizing the expanding body of genome-scale data and linking it to functional and phenotypic information generated at the JGI and elsewhere. Rokhsar is also Professor of Molecular Cell Biology and Physics at the University of California, Berkeley. After receiving a Ph.D. at Cornell University in theoretical physics and conducting postdoctoral research at IBM, Rokhsar joined the physics faculty at UC Berkeley in 1989. In the mid-1990s his research interests shifted from materials physics to biophysics, computational biology, and genomics. In 2000 he joined the JGI to lead its computational biology program. He is a former NSF Presidential Young Investigator, Sloan Foundation Fellow, Miller Research Professor, and Guggenheim Foundation Fellow. His research interests are in computational, comparative, and functional analysis of eukaryotic genomes. Rokhsar also serves as the DOE JGI Plant Program Lead.

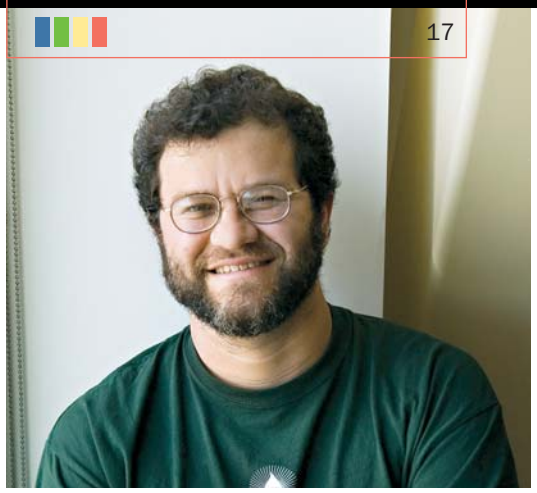
Len Pennacchio, Genomic Technologies Department Head, serves as the JGI's primary point of contact for new genomic technologies. This group is also involved in all projects seeking to identify genetic variation within a given species. In addition, scientific research is being conducted to derive biological insights into sequence data being generated by the JGI. One current area of focus is the assessment and utilization of new DNA sequencing technologies to capture genetic diversity as well as to aid in genome assemblies.

Pennacchio earned his Bachelor's degree in biology from Sonoma State University. He received his Ph.D. in 1998 from the Department of Genetics at Stanford University. During his graduate studies, he worked with Richard Myers to uncover the genetic cause of a rare form of human epilepsy and subsequently generated one of the first mouse models for epilepsy. In 1999, he was an Alexander Hollaender Distinguished Fellow at LBNL, where he identified a novel apolipoprotein involved in human and mouse triglyceride metabolism. In 2007, he took charge of the Genomic Analysis Department at the JGI. Pennacchio is a recipient of the Presidential Early Career Award for Scientists and Engineers. His research is focused on understanding how DNA sequence variation contributes to phenotypic (trait) differences within a species. In 2008, Pennacchio was named by *Genome Technology* magazine as one of its 30 rising young stars in its third annual "Tomorrow's PIs" special edition.

JGI Program Heads

Nikos Kyrpides, Genome Biology Program Head, has led the development of the suite of resources entailed under the Integrated Microbial Genome (IMG) system—for interpreting the newly sequenced genomes and for the analysis of existing genomic sequence data on a comparative level. Additionally, Kyrpides' team is engaged in the development of new tools and strategies that can be used for the prediction of gene functions.

Kyrpides received his Ph.D. in Molecular Biology from the Institute of Molecular Biology and Biotechnology, Crete, Greece, in 1996. He then pursued postdoctoral work in the laboratory of Carl Woese at the University of Illinois at Urbana-Champaign's Department of Microbiology. In 1998, he moved to the department of Mathematics and Computer Science at Argonne National Laboratory, where he did research in genome-wide analysis of microbial organisms. In 1999, he turned to industry by joining the newly founded Integrated Genomics Inc, leading the development of its genome analysis pipeline. Kyrpides joined JGI in 2004 to lead its Genome Biology Program. His group is primarily responsible for the analysis of microbial genomes as well as the development of novel methods and approaches for comparative analysis, including large-scale function prediction and the design of cellular and metabolic pathway collections.

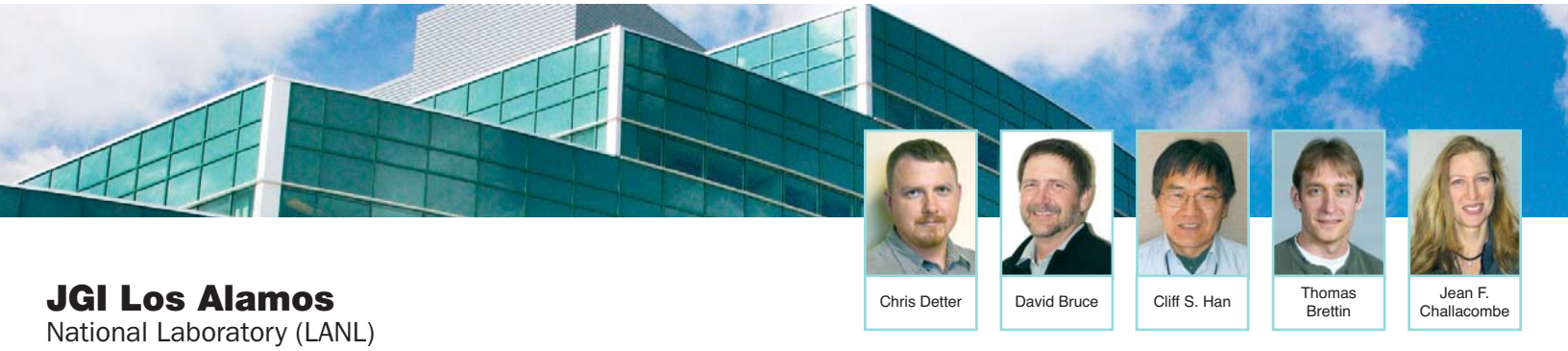


Philip Hugenholtz, Microbial Ecology Program (MEP) Head, leads the effort to use sequencing-based technologies to understand microbial communities via a combination of computational and experimental methods. Hugenholtz received his Ph.D. in Microbiology from the University of Queensland, Brisbane, Australia, in 1994. He then pursued postdoctoral work with Norman Pace in the Department of Biology at Indiana University and later in the Department of Plant and Microbial Biology at the University of California, Berkeley. In 1998 he returned to the University of Queensland to study the microbial ecology of activated sludges with Linda Blackall. In 2001, he joined the Computational Biology and Bioinformatics group in the Mathematics Department at UQ and learned to interact with mathematicians and programmers. In 2002 he ventured back to the U.S., lured by the promise of applying genomic methods to microbial ecosystems in the group of Jill Banfield at UC Berkeley. And indeed, the promise was fulfilled by the publication of the first metagenomic study of a simple acid mine drainage biofilm community. Buoyed by this success, Hugenholtz joined the DOE JGI in May 2004 to lead the Microbial Ecology Program. His group is developing methods for analyzing metagenomic datasets and applying them to a number of interesting communities, including enhanced biological phosphorus-removing and other sludges, termite hindguts, and Lake Vostok accretion ice.

Cheryl Kerfeld, Education Program Head, is developing programs and tools to train the next generation of genomic scientists. The focus is centered on large-scale DNA sequencing and its bioinformatic analysis. This effort relates to the need to fill a void in life sciences training as recommended by the National Research Council in 2005. It is expected that these programs and tools will be useful to a range of institution types, from research universities to community colleges, recognizing that faculty at smaller schools are often unfamiliar with bioinformatics and genome-scale research. The products also have the potential for adaptation to high-school outreach programs.

Prior to joining the DOE JGI, Kerfeld developed and directed the Undergraduate Genomics Research Initiative at UCLA. Kerfeld is also a structural biologist; her current research involves the structural and functional characterization of bacterial microcompartments and of proteins involved in photoprotection in photosynthetic organisms. She has degrees in Biology (B.A. and Ph.D.) and English (B.A. and M.A.) and has longstanding interests in bioinformatics and research-based and interdisciplinary education. Kerfeld is also an Adjunct Professor of Plant and Microbial Biology at the University of California, Berkeley.

Jonathan Eisen, Phylogenomics Group Head, focuses on the development and use of methods in the emerging field of phylogenomics. Phylogenomics involves the integration of evolutionary reconstructions (e.g., phylogenetics) and genome sequence analysis. This approach allows both for a better understanding of the mechanisms of evolution as well as for improved interpretation of genome sequence data. In addition to his DOE JGI role, Eisen is an evolutionary biologist and a Professor at the University of California, Davis, holding appointments in the Section on Evolution and Ecology and the Department of Medical Microbiology and Immunology. His research focuses on the mechanisms underlying the origin of novelty (how new processes and functions originate). Most of his work involves the sequencing of the genomes of microorganisms and the development and use of phylogenomic methods to analyze the genome data. Currently, he is using phylogenomic methods to study microbes in their natural habitats, including symbionts living inside host cells and planktonic species in the open ocean. Prior to moving to UC Davis, he worked at The Institute for Genomic Research (TIGR) in Rockville, Maryland. He earned his Ph.D. from Stanford University and his undergraduate degree from Harvard College.



Chris Detter

David Bruce

Cliff S. Han

Thomas Brettin

Jean F. Challacombe

JGI Los Alamos
National Laboratory (LANL)

Los Alamos National Laboratory (LANL) was one of the founding members of the original DOE Joint Genome Institute in 1997. The Genome Science Group at LANL is composed of approximately 35 full-time employees. With more than 20 years of genomics expertise, the Group focuses on several synergistic genomics activities revolving around sequencing, finishing, analysis, applications, and bioinformatics. The Group's central focus is as a partner in the DOE JGI. The resources of the Genome Sciences Group are augmented by the Bioscience Division, in which it is housed, as well as the over 400 scientists throughout LANL directly engaged in biological research and development.

JGI LANL's main directive is to "finish" microbial organisms drafted at the JGI in a high-throughput fashion. This involves both wet-lab and computational activities to fill in the gaps, resolve repeats, and assemble genomes into a single, high-quality contiguous structure. The JGI LANL Group has finished more than 200 microbial organisms over the past three years, both as part of the DOE JGI and for other funded collaborations. On a case-by-case basis, these activities are often followed by working closely with the project's collaborator to curate automated annotations, analyze, and carry out comparison studies on the genome that leads to publication of the resultant work.

Key JGI LANL Personnel

Chris Detter leads the JGI LANL organization and has a Ph.D. in Molecular Genetics and Microbiology, and 13 years of relevant hands-on experience in the field of high-throughput genomics related to genomic and cDNA library creation, DNA and RNA isolation and purification, and DNA sequencing. He has authored and co-authored more than 40 peer-reviewed publications in the field of high-throughput genomics and DNA sequencing. Detter is the current Genome Science Group Leader within the Bioscience Division at Los Alamos National Laboratory and the JGI-LANL Center Director for the DOE-JGI organization.

David Bruce has a B.S. in Biochemistry, a separate B.S. in Electrical Engineering, and has been a certified Project Management Professional for seven years. Bruce currently leads the Joint Genome Institute Project Management Office with employees at the DOE JGI and LANL. He has over 25 years experience in molecular biology, and 16 years experience in genomics with specializations in project management and process optimization.

Cliff S. Han has a Ph.D. in Genetics and over 16 years of experience in genomics. He received his Ph.D. for his work on mapping a 3 Mb region in the human genome with yeast artificial chromosomes in 1996 and came to LANL to continue working on the Human Genome Project. Since 2003, Han has led the Computational Finishing Team he established.

Thomas Brettin has an M.S. in Genetics, and over 15 years of experience in the field of genomics. He came to LANL and the DOE JGI from the Whitehead Institute/MIT Center for Genome Sciences (now the Broad Institute). He became a Software Architecture Professional from the Software Engineering Institutes at Carnegie Mellon University in the summer of 2005 and has over 10 years of software engineering experience. Brettin has taught Computer Science at the University of New Mexico Los Alamos branch since 2000. He has lead the Informatics Team at LANL since 2002 and is currently on a Change of Station (COS) to the DOE JGI to help provide subject matter expertise in the areas of software architecture and engineering as well as computational biology.

Jean F. Challacombe has a Ph.D. in neuroscience and over eight years of experience in software development, as well as six years of experience in bioinformatics. Challacombe is currently leading the efforts in genome analysis for the Genome Science Group at LANL, including JGI-LANL's activities in this area.

HUDSONALPHA
INSTITUTE FOR BIOTECHNOLOGY

Richard Myers

JGI HudsonAlpha
(formerly Stanford Human Genome Center)

The Stanford Human Genome Center (SHGC) began collaborating with the DOE JGI in the fall of 1999 to finish the human chromosomes that the JGI was shotgun sequencing for the DOE's allocated portion of the human genome. After the completion of the sequencing of the human genome, the DOE JGI rapidly expanded its scientific goals, and the JGI SHGC kept pace, focusing on assessing, assembling, improving, and finishing eukaryotic whole-genome shotgun (WGS) projects for which the shotgun sequence is generated by JGI.

In late 2008, the research program formerly situated at the Stanford Human Genome Center moved from California into the newly built nonprofit HudsonAlpha Institute for Biotechnology in Huntsville, Alabama. HudsonAlpha will continue to work as a partner of the JGI with a focus on plant and eukaryotic genomics. Its new research group, the HudsonAlpha Genome Sequencing Center, will expand its operations into next-generation genome sequencing and add capabilities to collect data to augment the use of JGI genomes in scientific discovery, bioenergy, and other directed plant genomics applications. Its new research facility and laboratory space is a vast improvement over the Stanford infrastructure, and it expects, once fully staffed, to be able to make a larger impact on DOE science.

JGI HudsonAlpha has four main groups consisting of 27 group members: Production Sequencing (with eight members), Library Construction (with 10 members), Computational Finishing (with three members), and Informatics (with six members). Jane Grimwood coordinates the activities of the Library and Finishing Groups. Jeremy Schmutz coordinates the activities of the Sequencing and Informatics Groups. As Principal Investigator, Dr. Richard Myers has overall responsibility for the program.

Key JGI HudsonAlpha Personnel

Richard Myers is Director and a Faculty Investigator of the HudsonAlpha Institute for Biotechnology, a new nonprofit research institute in Huntsville Alabama devoted to research in genetics and genomics. He received his Ph.D. in Biochemistry at the University of California, Berkeley and then did postdoctoral work at Harvard University. His first faculty position was at UCSF, where he developed his research program in human genetics and genomics. There he was director of one of the first U.S. genome centers, and moved the center and his laboratory to the Department of Genetics at

partner laboratories

The DOE JGI partnership produces high-throughput DNA sequencing and analysis for its user community by harnessing expertise in its partner laboratories to carry out specific projects in its sequencing and analysis pipelines.

Jeremy Schmutz

Jane Grimwood

JGI Oak Ridge
National Laboratory (ORNL)

Stanford University School of Medicine in 1993, where his group began their collaboration with the DOE JGI and contributed to sequencing the human genome. The group has continued to contribute to DOE sequencing efforts since the human genome was finished. With Jeremy Schmutz and Jane Grimwood, the sequencing center moved with Meyers and his laboratory to HudsonAlpha Institute in July 2008.

Jeremy Schmutz is a Faculty Investigator and, with Jane Grimwood, leads the HudsonAlpha Genome Sequencing Center. In 1994, he moved directly from his undergraduate studies in Computer Science and Biology into private industry working on algorithm development in support of massively parallel DNA sequencing by hybridization. In 1998, he joined the Stanford Human Genome Center to build the data analysis and data tracking systems for the new pilot NIH genome sequence center at Stanford. His work included creating complex data collection systems that made possible the finishing of the human genome sequence. He also led the NIH sequence quality assessment project that validated the accuracy of finished human genome sequence. After the completion of the human genome, he concentrated on whole-genome shotgun assembly and directed sequence improvement of plant, fungal, and algal genomes. Schmutz currently leads the Informatics and Production Sequencing Groups at the HudsonAlpha Genome Sequencing.

Annotation is currently the principal bottleneck for the exploitation of genome sequence. The Computational Biology and Bioinformatics Group (formerly Genome Annotation and Systems Modeling) at Oak Ridge National Laboratory (ORNL), in partnership with DOE JGI, has developed a large-scale distributed computational framework consisting of analysis tools and databases used to provide comprehensive microbial genome annotation to further the understanding and analysis of microbial biology.

This system, referred to as the Analysis and Annotation Pipeline (AAP), is used to identify sequence-level features such as gene predictions, and via large-scale comparisons, to annotate predictions such as protein function, regulatory signals, and metabolic pathways. The AAP processes both Draft and Finished microbial genomes from the DOE JGI. The resulting annotation files are returned to the DOE JGI ready for submission to GenBank, and provided to the user(s) prepublication, and then to the larger research community after a three-month embargo. Assistance and support is provided to ensure that DOE JGI and its research user community are able to gain maximum scientific value from the genome sequence and its analysis.

Key JGI ORNL Personnel

Bob Cottingham leads the Computational Biology and Bioinformatics group. He recieved a B.S. degree in Mechanical Engineering from the University of Washington in 1973. In 1989 he became the Directeur Informatique at CEPH in Paris, then joined the U.S. Human Genome Project, first as the Co-Director of the Informatics Core in the Baylor College of Medicine Human Genome Center, then as Operations Director of the Genome Database (GDB) at the Johns Hopkins University School of Medicine. Subsequently, he became Vice President of Computing at Celltech Chiroscience, a UK biopharmaceutical company developing drugs based on gene targets. In 2000 he co-founded VizX Labs, a bioinformatics company that developed GeneSifter, the first web-based gene expression microarray analysis service now used by hundreds of labs. In 2008 Cottingham moved to ORNL. His interests include the development of computational systems as embodiments of scientific models that not only capture data and support analysis but can guide and direct research to new discovery.

Gerald Tuskan, Ph.D., Distinguished Scientist, Plant Genetics Group, Environmental Sciences Division, ORNL, has over 15 years experience leading and working with DOE on the development of biomass feedstocks. Tuskan is currently the co-lead for the Joint Genome Institute Plant Genomic effort, the project leader of the International *Populus* Genome Consortium and the Activity Lead for the DOE BioEnergy Science Center *Populus*

Pacific Northwest
NATIONAL LABORATORY

Bob Cottingham

Gerald Tuskan

Miriam Land

Harvey Bolton

Scott Baker

Doug Ray

JGI Pacific Northwest
National Laboratory (PNNL)

Biosynthesis team. In addition, he is currently the lead PI on DOE and NSF projects related to *Populus* genomics and biomass energy. His research focuses on the accelerated domestication of *Populus* through direct genetic manipulation of targeted genes and gene families. Tuskan has published over 76 peer-reviewed publications since 1990 in the areas of genetic and genomics of perennial plants, including 20 publications with nearly 300 citations exclusively related to biomass and bioenergy.

Miriam Land leads the Analysis and Annotation Pipeline project and coordinates the activities, including data exchange between ORNL and JGI, tracks the status of local tasks, writes much of the code for the pipeline, and provides support to researchers. Land received an M.S. in Statistics in 1986 and has been at ORNL since 1991 when she worked on the Oak Ridge Environmental Information System project. In 1996, she joined the Computational Biology group and developed the Web interface for the Genome Channel and has been on the microbial genome annotation team since 2001.

Since 2004, PNNL has collaborated with DOE JGI in the area of fungal genomics. PNNL proposed and is the lead collaborator for the *Aspergillus niger* genome project. When the DOE JGI instituted its Laboratory Science Program (LSP), Scott Baker from PNNL led the development of a fungal genome sequencing "track." Six fungal genome projects were approved in the first year of the LSP. After the LSP was folded into the DOE JGI Community Sequencing Program, DOE JGI management decided that, due to the strong relevance of fungi to DOE Mission Areas, a formal fungal genomics program would be an appropriate addition to existing microbial, metagenomic, and plant genomics programs.

Also, the proteomics facility at PNNL has engaged in a pilot project that is designed to provide proteomic data for 15 microbes currently being sequenced at the JGI. Proteomics is the only reliable method for experimentally validating the accuracy of gene models derived from genome sequencing. The pilot project provided a mechanism for the use of proteomic data for use in genome annotation of microorganisms and is currently being extended for eukaryotic fungi. Initial work focused on the annotation of fungal genomes will provide the framework for expanding the use of proteomic data for the proteome-assisted creation of gene models for all eukaryotic genomes from fungi, plants, and animals.

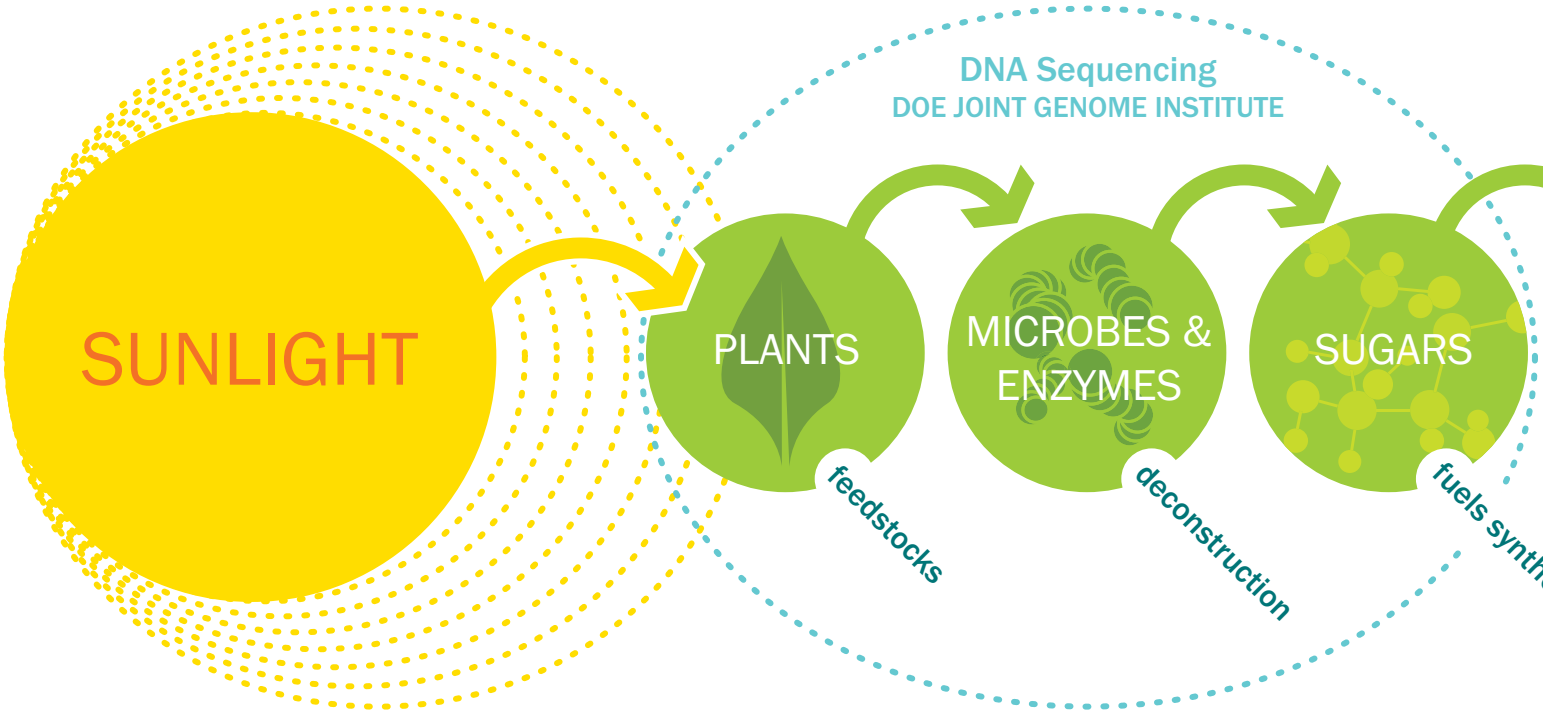
Key JGI PNNL Personnel

In addition to the contribution of Scott Baker, Doug Ray, PNNL Associate Laboratory Director of Fundamental and Computational Sciences Directorate, is a member of the Scientific Advisory Committee for the DOE JGI. Harvey Bolton is the Director of Biological Sciences at PNNL and a PNNL representative on the JGI Management Board. Bolton is responsible for the DOE Office of Biological and Environmental Research (DOE OBER)-funded Biological Systems Science Programs at PNNL, which includes interactions with the DOE JGI. Ray and Bolton coordinate all collaborations between PNNL and the DOE JGI.

PNNL has utilized both independent DOE BER funding and internal Laboratory Directed Research and Development (LDRD) funding for science that has found synergy with the DOE JGI mission. DOE-BER is one of the primary funding agencies of the proteomics facility at PNNL and is currently supporting the fungal research as well. Closer and formal collaboration with the JGI will mutually benefit both the ongoing science at PNNL and the mission of the JGI. Baker, along with Gordon Anderson and Bolton sit on the Joint Management Board, and Mary Lipton, Baker, and Bolton participate in the bi-weekly partner management meetings. Baker leads target selection for the JGI's Fungal Genome Program.

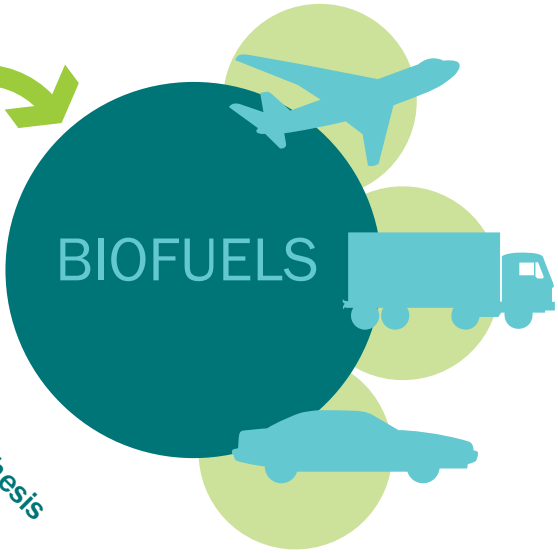
DOE JGI mission areas

In 2004, after completing its commitment to the Human Genome Project, the JGI established itself as a national user facility. The vast majority of JGI sequencing is conducted under the auspices of the Community Sequencing Program (CSP), surveying the biosphere to characterize organisms relevant to the DOE science mission areas of bioenergy, global carbon cycling, and biogeochemistry. The DOE JGI's largest customers are the DOE Bioenergy Research Centers (BRCs), which were launched in 2007 to accelerate basic research in the development of next-generation cellulosic biofuels.



Bioenergy

A major aim of DOE research is to develop technologies to enable the use of biomass for production of alternative fuels. The transportation sector, in particular, is considered a key target. Current biofuel production methods, such as corn-to-ethanol, are a valuable start but are considered too inefficient to replace a substantial fraction of fossil fuels; moreover, it competes with the food supply. Cellulosic materials, such as wood and grasses, are seen as a promising alternative because they do not require high levels of agricultural inputs and can yield more fuel per acre. However, they are difficult to break down into their component sugars, a necessary step in the conversion to biofuels. By sequencing plant genomes, DOE JGI is seeking to identify genes and pathways involved in plant architecture, drought and pest resistance, and biomass yield, which can help inform efforts to optimize feedstocks for biofuels production.



Many systems in nature have evolved to use cellulosic biomass for energy, and all of these depend on microbial metabolism. Termites are a prime example of animals that efficiently break down biomass; the dense microbial community in their hindguts appears to shoulder the burden in biomass breakdown. A metagenomic sequencing project targeting this community found more genes for cellulolytic enzymes, per base sequenced, than any prior sequencing project. These genes, many of which were only distantly related to known enzymes or contained novel domain organizations, enhance our knowledge of biomass breakdown and expand the repertoire of enzymes available for testing in industrial biofuel production processes.

Several ongoing sequencing projects target both symbiotic and free-living communities with biomass degrading capability. The tammar wallaby, a marsupial, has a unique foregut fermentation whose microbial players have recently been subjected to metagenomic sequencing at the DOE JGI. Additional projects target the symbiotic communities of the shipworm, a wood-eating bivalve; the hoatzin, a leaf-eating bird; and the Asian longhorned beetle, a pest that consumes healthy hardwood trees. DOE JGI has also sequenced a free-living microbial community actively degrading poplar biomass in a woodpile.

At present, we produce ethanol from biomass-derived sugars by centuries-old fermentation techniques. However, some sugars released from biomass are not fermented by this process, and it is clear that other alcohols (e.g., butanol) are better biofuels. DOE JGI's microbial sequencing efforts stand to contribute to the development of ethanol-tolerant microbial strains, new fermentation pathways, and engineering of better biofuel products. By combining, comparing, and contrasting sequence data from these widely varied and independently evolved cellulolytic communities, we expect to identify key elements in successful biomass degradation.

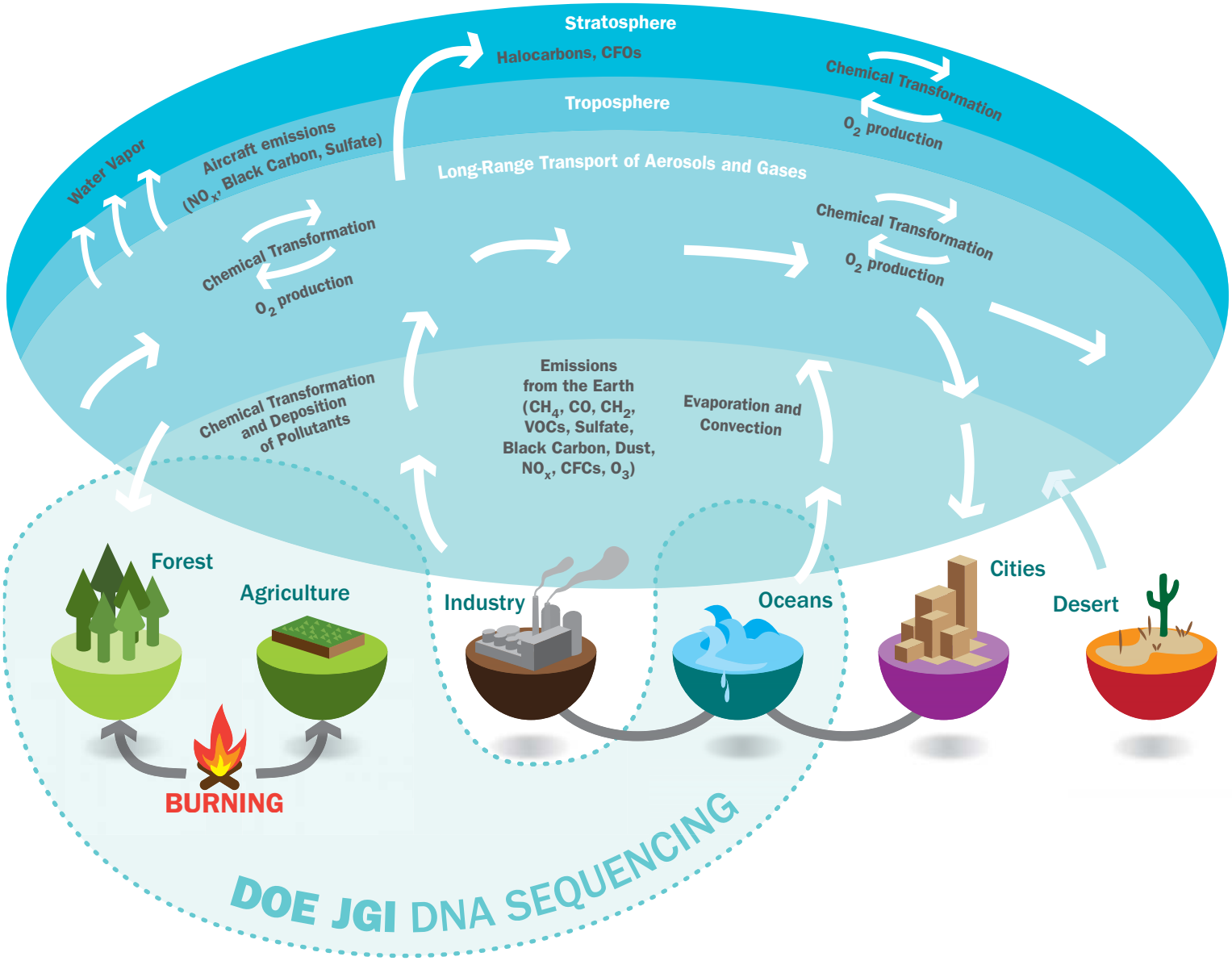
Carbon Cycling

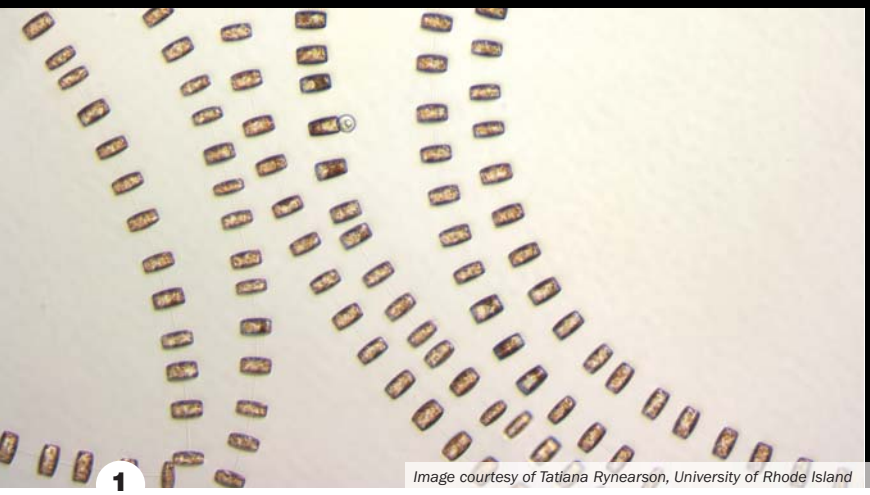
The global carbon cycle depends heavily on microorganisms at all stages, from fixing atmospheric carbon to facilitating plant growth to breaking down dead organic material. Understanding the carbon metabolism of environmental microbes will therefore improve our ability to predict, and perhaps mitigate, the effects of rising carbon dioxide levels on our global climate.

To this end, the DOE JGI is pursuing multiple projects to characterize natural microbial communities with direct or indirect influence on carbon cycling. Research into the role that microorganisms play in the Earth's natural carbon cycle may lead to new strategies for reducing atmospheric carbon and other greenhouse gases. A prime example is a recent project to characterize microbes actively metabolizing single-carbon compounds, such as methane, in freshwater sediments. The "Lake Washington Methylotroph" metagenome used a specialized metabolic labeling technique to specifically target these organisms and revealed a unique community involved in processing each of several different single carbon compounds. Additional projects targeting marine, soil, and sediment communities are expanding our understanding of the roles these microbial communities play in global nutrient cycling.

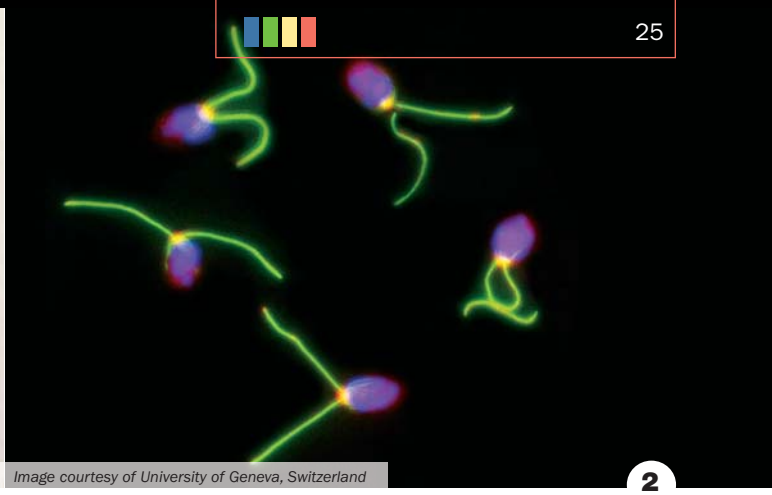
1. Diatoms

Diatoms are responsible for significant amounts of marine primary production. In response to favorable light and nutrient conditions, diatoms rapidly divide and form





1 *Image courtesy of Tatiana Rynearson, University of Rhode Island*



2 *Image courtesy of University of Geneva, Switzerland*

large blooms. As blooms propagate, nutrients are depleted, growth ceases, and cells sink to the deep ocean. The sinking diatom blooms fuel the biological carbon pump and export carbon from the atmosphere to the deep ocean. Diatoms generate about 40% of the 45-50 billion tons of organic carbon fixed annually in the sea and as much as 90% of the photosynthetically produced organic carbon that fuels coastal ecosystems. According to some predictions, their role in global carbon cycling is comparable to that of all terrestrial rain forests combined.

Because of their significance in global carbon cycling, the DOE JGI has pioneered the sequencing of diatom genomes and has completed sequencing of two distantly related species, *Thalassiosira pseudonana* and *Phaeodactylum tricornutum.* Although these genomes are invaluable for understanding the molecular underpinnings of diatom biology, they are primarily laboratory models and do not form large blooms in the ocean. As a CSP 2009 project, JGI plans to use transcriptional profiling approaches to understand how new isolates of a bloom-forming, abundant, and globally distributed diatom in the genus *Thalassiosira*, *T. rotula,* responds to nutrient and light limitation. This research will fundamentally advance our understanding of how diatoms acquire nutrients, how these strategies are regulated, how different isolates of the same species respond to stress, and how they respond to multiple stressors. These data are expected to open the door for molecular investigations of unsequenced diatom species in the field and provide a framework for future genome-enabled studies of diatoms.

The principal investigators are Bethany D. Jenkins (University of Rhode Island), Sonya Dyhrman (Woods Hole Oceanographic Institute), Tatiana Rynearson (University of Rhode Island), and Mak Saito (Woods Hole Oceanographic Institute).

2. *Chlamydomonas*

The single-celled alga *Chlamydomonas reinhardtii,* while less than a thousandth of an inch in diameter, or about one-fiftieth the size of a grain of salt, is packed with many ancient and informative surprises. Affectionately known to its large research community as "*Chlamy*," the alga is a powerful model system for the study of photosynthesis and cell motility. While the *Chlamy* genome has been completed and analyzed by JGI and other contributors, a CSP 2009 project will sequence *Chlamydomonas* genes using newer methods.

Because our knowledge of its gene expression is limited, sequencing *Chlamy* genes via new methods, specifically 454 and Illumina sequencing, will allow researchers to compare the available approaches for analyzing gene expression. They will then be able to use the most appropriate one to more efficiently evaluate the conditions that favor a particular metabolism in *Chlamydomonas*, with specific emphasis on those that generate energy-rich products. As a model organism for photosynthetic eukaryotes, it is expected that lessons learned from *Chlamydomonas* will be applicable to the development of similar pathways in other, more industrially useful algae.

The performance of the 454 and Illumina platforms will be compared to that of high-density microarrays for the study of global gene expression in nutrient- and oxygen-deprived *Chlamydomonas*. Besides confirming the expression characteristics obtained by means of traditional microarrays, this project should also provide more quantitative information regarding high- and moderate-abundance messages and identify novel lower-abundance transcripts that are differentially expressed. It should also identify short gene sequences that have not been detected by other methods.

These alternative methods of transcript profiling should be applicable to any organism for which a genome sequence is available. Indeed, microarray platforms are not available for many ecologically important organisms, including many abundant phytoplankton species in the nutrient-poor oceans. The proposed development of this methodology opens the door to similar analyses for these organisms and to comparative protein expression analyses among the different algae. Such comparisons might be particularly interesting for organisms that have adapted to different environments and have developed different strategies for coping with both prevalent and fluctuating environmental conditions.

The principal investigators are Maria L. Ghirardi and Michael Seibert (National Renewable Energy Laboratory), Arthur R. Grossman (Carnegie Institute of Washington), Sabeeha Merchant and Matteo Pellegrini (UCLA), and Matthew C. Posewitz (Colorado School of Mines, NREL).

Biogeochemistry

The area of biogeochemistry (formerly called bioremediation) is one in which metagenomics can play a critical role, as successful degradation or sequestration of pollutants often depends on consortia of microbes working in concert. Biogeochemistry entails study of the biological, physical, geological, and chemical processes that regulate the natural environment ranging from air, water, and soil to the cycles of matter and energy that transport chemical components in time and space.

Bioreactors, such as sewage sludge (or system that supports a biologically active environment), are particularly well-suited to metagenomics as they often harbor communities whose members cannot be grown in isolation. In one project, scientists are studying samples of sludge from Enhanced Biological Phosphate Removal (EBPR), a wastewater treatment process, to better understand the microbial population dynamics and thus improve these important environmental systems. From 72 Mb of EBPR bioreactor DNA sequence, the DOE JGI team was able to assemble a largely complete genome of the dominant player, *Candidatus Accumulibacter phosphatis*, and use metabolic reconstruction to elucidate the mechanism by which it accumulates phosphate. This, in turn, will aid in understanding and preventing "crashes" that occasionally compromise treatment and result in phosphate being released to the environment.

Additional sequencing projects well underway target bioreactors capable of degrading terephthalate, a byproduct of plastics and polyester production discharged in wastewater; detoxifying chloro-organic pollutants that commonly contaminate groundwater near dry-cleaning sites; and breaking down benzene, a toxic component of gasoline. These projects are yielding important insight into the metabolisms of communities studied, making them more readily exploitable for biogeochemistry.

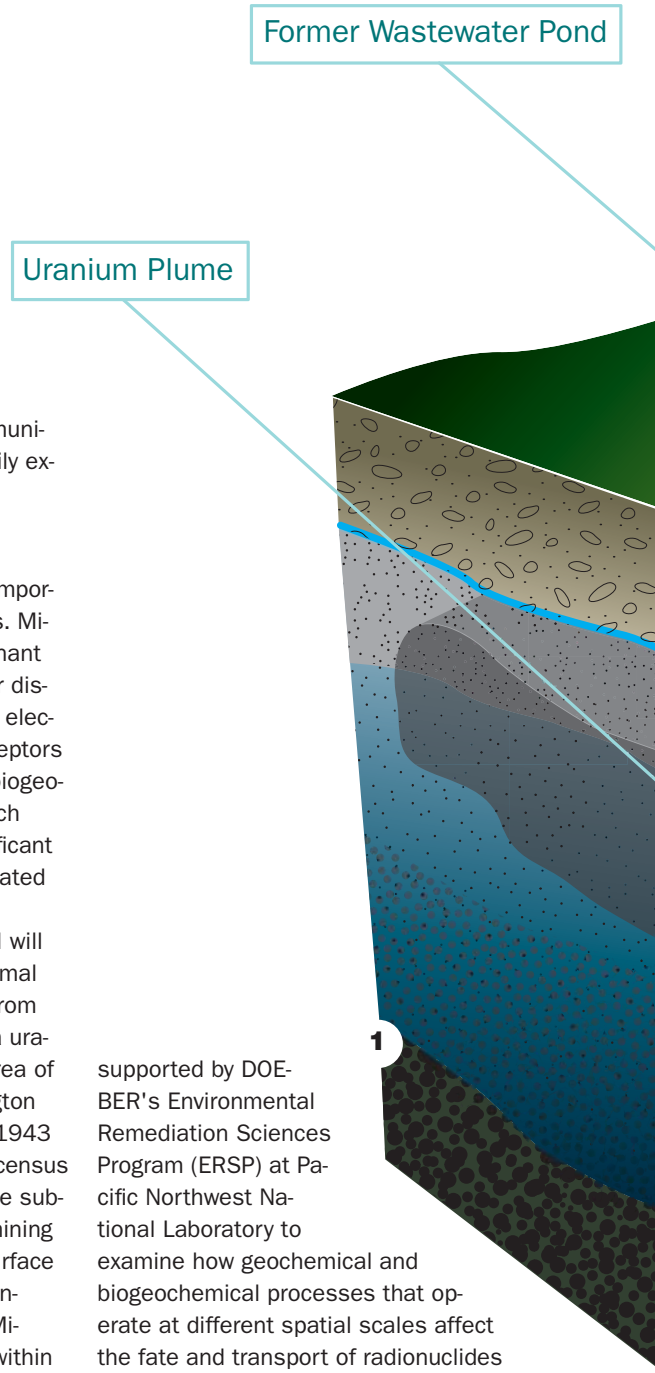
1. Hanford Site

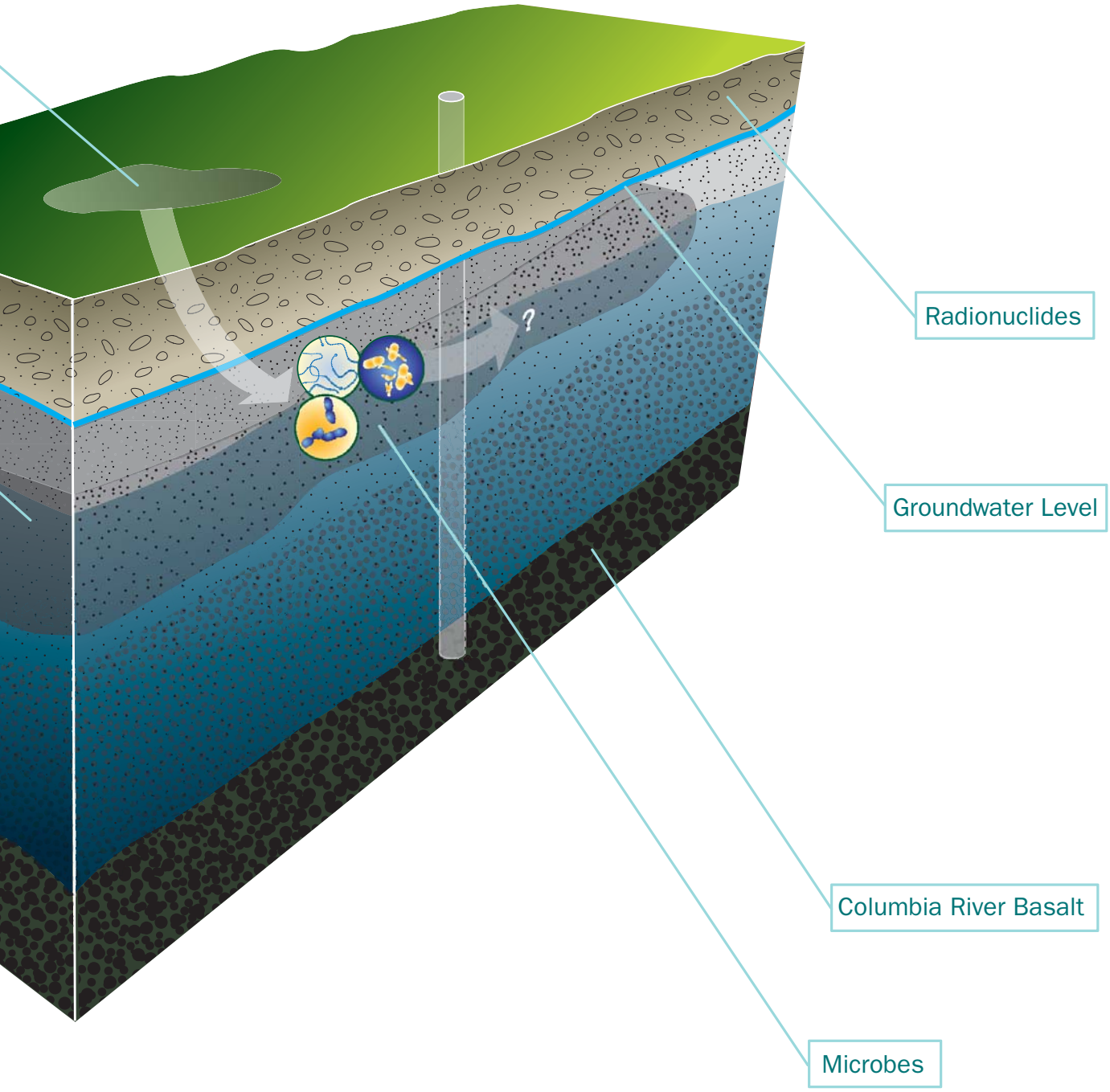
Subsurface microorganisms play an important role in transforming contaminants. Microbial reactions can modify contaminant solubility, result in the precipitation or dissolution of mineral phases, consume electron donors, and reduce electron acceptors (and thereby alter the chemical and biogeochemical reactivity of microsites). Such transformations could be highly significant to long-term stewardship of contaminated subsurface sediments.

For this CSP 2009 project, JGI will sequence 16S rRNA (a type of ribosomal RNA found in bacteria and archaea) from subsurface sediment sampled from a uranium-contaminated site in the 300 Area of the Hanford Site in southern Washington (used for plutonium production from 1943 to 1989). The sequences will yield a census of bacteria and archaea present in the subsurface, aiding researchers in determining the composition and activity of subsurface microbial communities in microenvironments and across transition zones. Microenvironments are small domains within larger ones that exert a disproportionate influence on subsurface contaminant migration. Transition zones are field-scale features where chemical, physical, or microbiologic properties change dramatically over distances less than one meter. As a result, important chemical species such as molecular oxygen (or other potential electron acceptors) and organic carbon have steep transport-controlled gradients in these regions that dramatically impact subsurface contaminant reactivity.

This work is one component of integrated, multidisciplinary research efforts supported by DOE-BER's Environmental Remediation Sciences Program (ERSP) at Pacific Northwest National Laboratory to examine how geochemical and biogeochemical processes that operate at different spatial scales affect the fate and transport of radionuclides and other contaminants in the subsurface at DOE's Hanford Site. The 300 Area subsurface environment, where uranium is the primary contaminant of concern, represents a valuable natural laboratory where hydrologic, mass transfer, and biogeochemical processes controlling contaminant fate and transport can be investigated across a span of biogeochemical environments—the vadose zone (earth above groundwater), aquifer, and hyporheic zone (interface between aquifer and river).

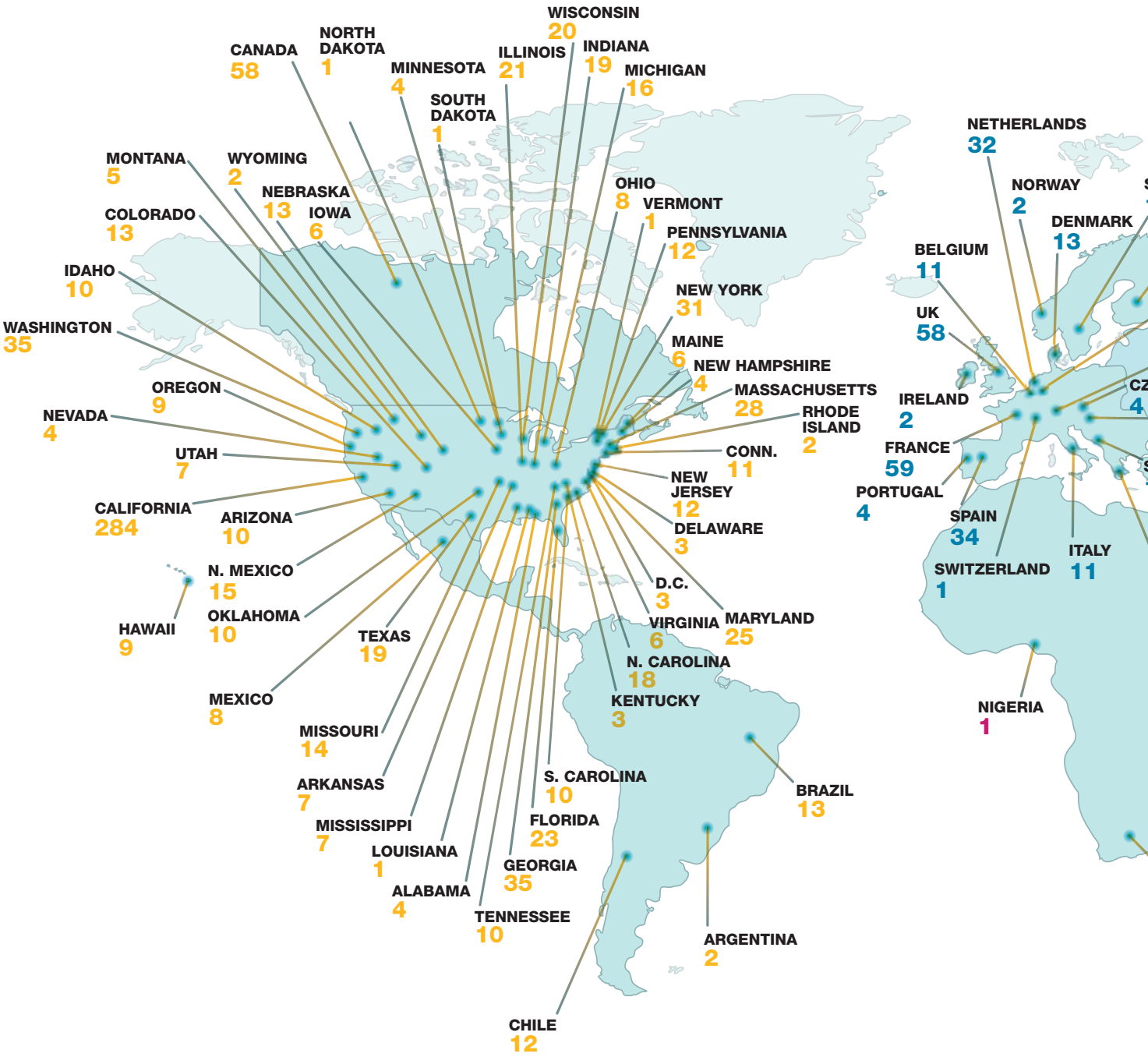
The principal investigators are Allan Konopka and Jim Fredrickson (PNNL).




?
Radionuclides
Groundwater Level
Columbia River Basalt
Microbes

DOE JGI user community

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



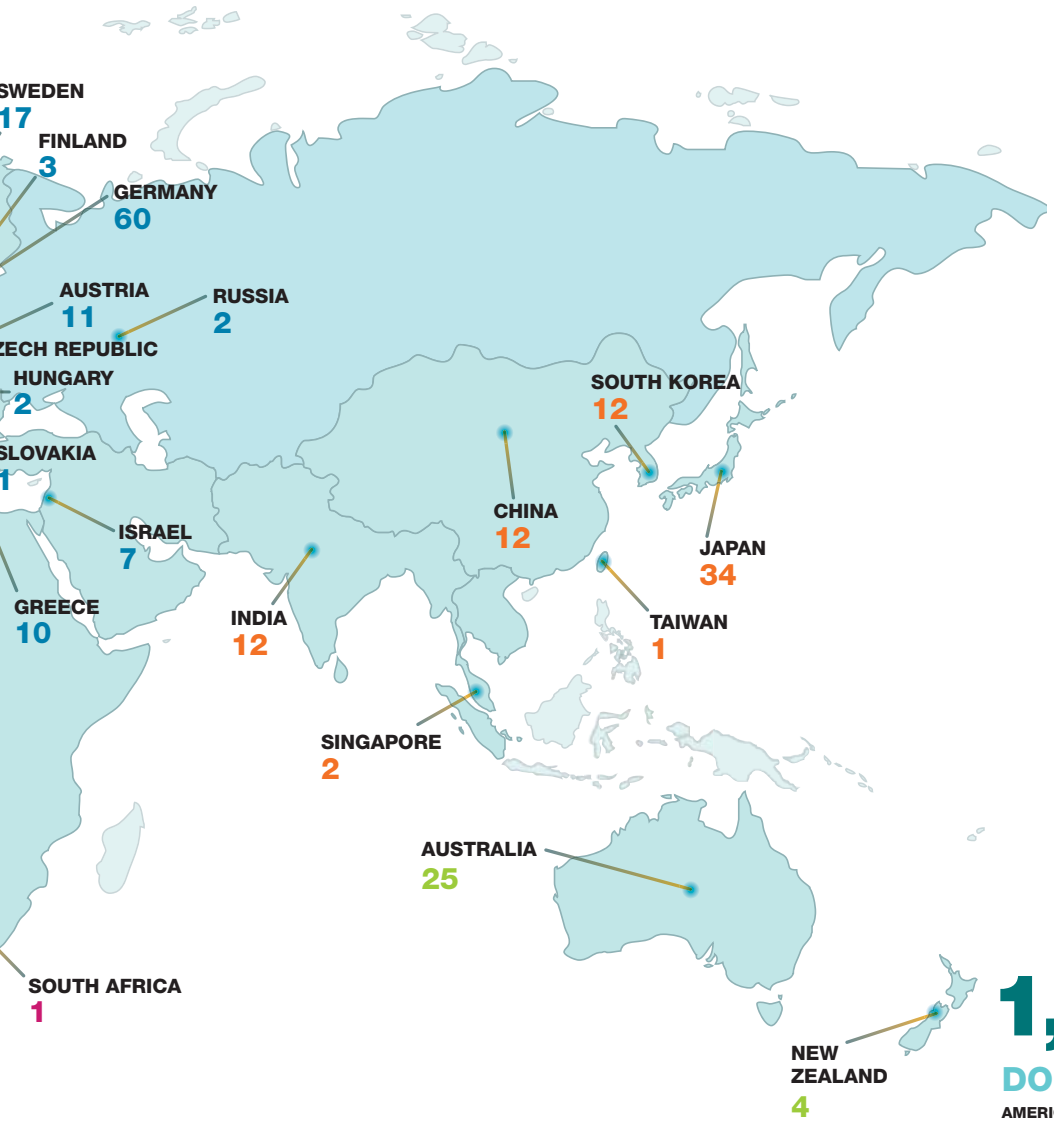
1,300 unique collaborators on active projects. The broader user community is engaged in collaborations that are global in nature. In addition, the JGI's influence is extensive, considering that hundreds more researchers every year tap sequence data posted by the JGI on its numerous genome portals and at the National Center for Biotechnology Information's (NCBI) GenBank.

Building the Community

Since 2006, DOE JGI has convened an annual meeting of its users and collaborators, as well as prospective new users. The DOE JGI User Meeting offers researchers a forum for sequence-based science related to the DOE mission.

The 2008 meeting emphasized the genomics of renewable energy strategies, biomass conversion to biofuels, environmental gene discovery, and engineering of fuel-producing organisms. The highlight of the event was the keynote address by Nobel Laureate Steven Chu, then-Director of the Lawrence Berkeley National Laboratory (now U.S. Secretary of Energy). The meeting also featured informatics workshops and tutorials for the analysis of prokaryotic and eukaryotic genomes and the evaluation of new sequencing platforms and their applications.

The Fourth Annual DOE JGI User Meeting is scheduled for March 25-27, 2009 and will feature several high-profile speakers. Slated keynote speakers include genome sequencing pioneers J. Craig Venter of the J. Craig Venter Institute, George Church, Director of the Center for Computational Genetics at Harvard Medical School, and Chris Somerville, Director of the Energy Biosciences Institute (EBI), a partnership between BP; the University of California, Berkeley; the Lawrence Berkeley National Laboratory; and the University of Illinois.



1,351

DOE JGI Users Worldwide in 2008

AMERICAS 856	EUROPE 336
AFRICA 2	ASIA 64
AUSTRALIA/NEW ZEALAND 29	

DOE bioenergy research center sequencing program

Beginning in FY 2008, DOE JGI began sequencing and analysis on behalf of the DOE's recently funded Bioenergy Research Centers (BRC). The Centers are intended to accelerate basic research in the development of cellulosic ethanol 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 Lawrence Berkeley National Laboratory (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 the DOE, the BRC projects are afforded top priority for sequencing and analysis at the DOE JGI.

To facilitate understanding of BRC sequencing needs and DOE JGI's capabilities and capacity, the JGI convened a two-day meeting in January 2008 of BRC investigators, OBER (DOE's Office of Biological and Environmental Research) program managers, and DOE JGI scientific and technical staff. Critical outcomes of this meeting were:

- understanding of DOE JGI's capabilities by the BRCs
- preliminary understanding of BRC's sequencing and analysis needs by DOE JGI
- agreement that BRC projects would receive top priority for sequencing and analysis
- an agreement on protection of intellectual property for the BRCs.

DOE JGI's extensive experience in working with users to define an effective plan to pursue important questions, often using new technical or analytical approaches, is a significant strength of our organization and is of special value to the BRCs.


BESC
BioEnergy Science Center


GLBRC
Great Lakes Bioenergy Research Center

genomic approaches to solving global challenges

Plants store solar energy through photosynthetic production of sugars that are biochemically transformed into cell wall polymers (long macromolecules, such as cellulose, made of similar or identical subunits linked together). These sugars, mostly glucose, can be fermented into environmentally friendly fuels such as cellulosic ethanol.

An obstacle to the efficient fermentation and conversion of cellulose to liquid is lignin. Lignin is a chemical compound that is an integral part of the cell walls of plant and trees, providing strength to cellulose fibers while conferring flexibility to the plant structure. Lignin makes up about 25–30% of the content of most dry wood. Breaking down the recalcitrant lignin polymer is key to overcoming the inefficient production of biofuels from woody plant matter.

What is biomass? Biomass is organic material from plants (agricultural and forestry residues, terrestrial and aquatic crops) and can include municipal solid and industrial wastes that can be harvested to produce energy. Biomass is an attractive feedstock for fuel because it is a renewable resource. Lignocellulose is biomass composed of cellulose, hemicellulose, and lignin.

What is biofuel? Biofuel is any fuel derived from biomass. Agricultural products that are currently converted to biofuels include corn for ethanol and soybeans for biodiesel. Efforts are underway to facilitate the generation of cellulosic biofuels. This requires the conversion of non-grain crops—such as switchgrass, a variety of trees, and other woody crops—to biofuels. JGI research supports the DOE commitment to displacing a significant amount of projected U.S. gasoline consumption with renewable fuels in the next decade.

How are biofuels currently produced? At present, most ethanol in the U.S. is made from corn. The biomass consists of the kernels and the corn stover, e.g., stalks, leaves, cobs, husks.

What are the shortcomings of ethanol from corn kernels? The energy value of corn is not as high as other plants (such as sugarcane). Moreover, growing corn on a massive scale requires inefficiently large inputs of water, fertilizers, and pesticides, on land that could otherwise be used for growing food. When corn is used as a feedstock to make ethanol, the costs of the inputs required to grow the corn are passed on to the consumer. Using today's technology, conversion of cellulosic biomass to ethanol is less productive and more expensive than the conversion of corn grain to starch ethanol. Genomics will help to identify the genes and enzymatic pathways leading to improved feedstock plant characteristics for easier conversion to biofuels.




Images courtesy of DOE/NREL

What is cellulosic ethanol? Cellulose is the largest component of biomass and is the most abundant organic polymer in nature. Each cellulose molecule is a linear polymer of glucose residues. For the production of cellulosic ethanol, carbohydrate polymers (cellulose and hemicellulose) in plant cell walls are broken down into sugars (glucose) that can be fermented into alcohol and then distilled to achieve fuel-grade ethanol. Cellulosic biomass is a less expensive and more abundant feedstock than corn grain. Sources of such promising renewable biomass include perennial crops—such as switchgrass, which can grow with minimal inputs and tolerate harsh growing conditions—and abundant fibrous and generally inedible plant matter such as wood pulp or corn stover (leftover leaves and stalks).

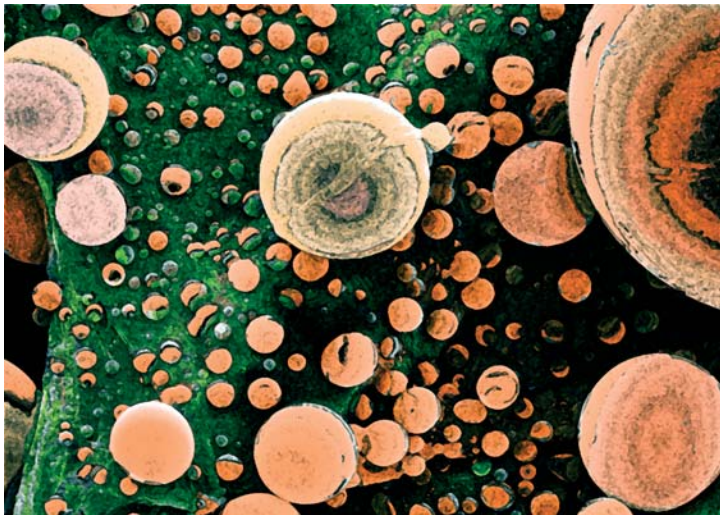
The structural complexity of cellulosic biomass is what makes this feedstock such a challenge to break down into simple sugars that can be converted to ethanol. While near-term research (over the next five years) is focused on more efficient means to convert starchy plants into liquid fuels, longer-term research (over the next 10–15 years) is focused on cellulosic ethanol.

What is biodiesel? Biodiesel can be produced from plant-based oils, including soybean, canola, and waste vegetable oil. It has a higher energy density than ethanol, and some forms can be used in unmodified diesel engines, but current oil crops require much more land than cellulosic ethanol feedstocks.

What is butanol? Butanol, for biofuel, can be produced from biomass (as biobutanol) by using microbes employed in natural fermentation processes. For example, a specific enzyme from the bacterium *Clostridium* can be used to drive the fermentation pathway. The same plants used to make starch ethanol can be used to make biobutanol, which is chemically more similar to gasoline and tolerates water contamination better than ethanol. Therefore, it may more easily be adapted to the existing gasoline transportation pipe-line infrastructure. Current research goals include genomic analyses of potential microbes to optimize biological processes aimed at making butanol at a price that's competitive with gasoline.

Why do biofuels produce fewer greenhouse gases? Fossil fuels, like coal or oil, add carbon to the atmosphere when burned, but biofuels only release carbon recently captured from the atmosphere by the plant during photosynthesis, rendering the latter "carbon neutral." Sustainable growth of trees and perennial plants can remove carbon dioxide from the atmosphere during photosynthesis and store the carbon in plants.

How will genomics advance biofuels research? The information emerging from the JGI's plant genome projects represents a critical foundation for advancing the development of a new generation of biofuels, such as cellulosic ethanol, butanol, and bio-diesel. By sequencing plants, the JGI and its collaborators have identified sets of genes involved in plant cell wall biosynthesis that are now providing clues for improving biofuel crop (or feedstock) domestication. The sequencing of microbial genomes is revealing the genes that encode cellulose and lignin deconstruction pathways, which in turn are informing strategies to increase the efficiency of the conversion of biomass to fuels.



DOE JGI plant genome program

1

2

Raw plant material, or dedicated bioenergy crops, can provide the starting point for making biofuels. With the sequencing and analysis of the genomes of the first tree, the poplar (published in *Science* in September 2006), and the green alga *Chlamydomonas* (published *Science* in October 2007), DOE JGI emerged as an important contributor in the plant genomics arena. Sequencing of diverse plant genomes is driven by the importance of plants as feedstocks for biofuel production, as natural agents of biogeochemistry and carbon cycling, and more generally as key contributors to the global carbon cycle. DOE JGI pioneered the application of the whole-genome shotgun method to complex plant genomes, such as the 730-million nucleotide *Sorghum bicolor* sequence, and the ~1-billion base-pair *Glycine max* (soybean) genome (see *www.phytozome.net)*.

The JGI's growing plant portfolio includes many that are, or will soon be, targets of keen interest in the biofuels research community. These include switchgrass, poplar, soybean, eucalyptus, sorghum, foxtail millet, cassava, cotton, and corn. In addition to bioenergy crops, DOE JGI has sequenced a diverse set of plant genomes, including bryophytes (the model moss *Physcomitrella*) and lycophtyes (the spike moss *Selaginella*).

Forest trees contain more than 90% of the Earth's terrestrial biomass, providing such environmental benefits as carbon capture, renewable energy supplies, improved air quality, and biodiversity. However, little is known about the biology of forest trees in comparison to the detailed information available for food crop plants.

The Plant Genome Program at DOE JGI integrates projects selected through the Laboratory Science Program (LSP), the Community Sequencing Program (CSP), and requests from DOE Bioenergy Research Centers. In recognition of the programmatic importance of plant science to the DOE mission, the DOE JGI Plant Genome Program has assembled an advisory panel of plant researchers to prioritize genomes for sequencing, consider nominations from the community, and make strategic selections of plant sequences guided by both DOE mission relevance and the needs of the plant science community.

A significant strength of DOE JGI's plant genome program can be found in the contributions of the several partner laboratories. DOE JGI's plant program is co-managed by Dan Rokhsar at the JGI and Jerry Tuskan at ORNL. Library construction and production sequencing occur at the JGI, while genome assembly and genome improvement and finishing are carried out at HudsonAlpha. The DOE JGI Computational Genomics group caries out automated annotation of genomes. Genomes are then published on DOE JGI's website, submitted to GenBank, and placed into DOE JGI's Phytozome. A focal point of the DOE JGI Plant Genome Program is Phytozome *(www.phytozome.net),* a Web hub for comparative plant genomics. The aim of Phytozome is to provide uniform access to all sequenced plant genomes, facilitating analysis of sequence evolution and function through gene families and whole-genome comparisons. The Phytozome project integrates not only DOE JGI's plant genome sequences but also other genome-scale datasets.

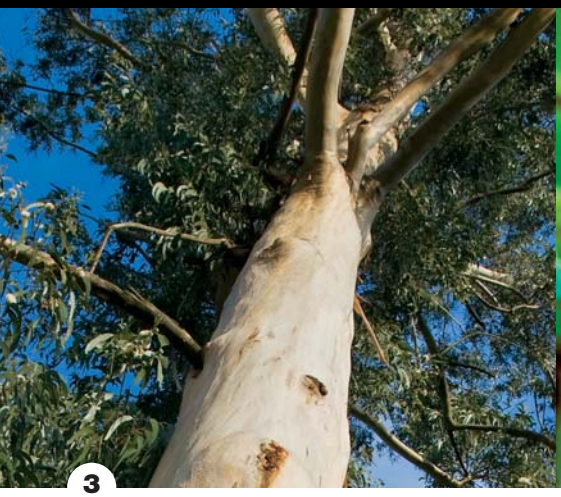
1. Loblolly Pine

Several significant bioenergy-relevant plant genomes are currently being sequenced by the JGI. Loblolly pine *(Pinus taeda)*, approved for the 2009 Community Sequencing Program (CSP), is an organism of tremendous economic and ecological importance and a key representative of the conifers, an ancient lineage of plants that dominates many of the world's temperate and boreal ecosystems. Loblolly pine's fast growth, amenability to intensive tree farming, and high-quality lumber/pulp have made it the cornerstone of the U.S. forest products industry and the most commonly planted tree species in America; approximately 75% of all seedlings planted each year are Loblolly pines.

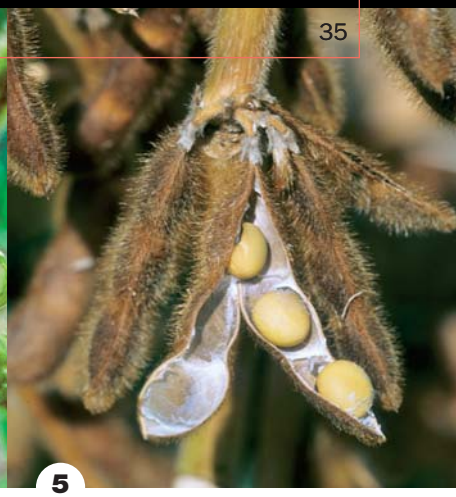
Its ability to efficiently convert CO_2 into biomass and its widespread use as a plantation tree have also made Loblolly pine a cost-effective feedstock for lignocellulosic ethanol production and a promising tool in efforts to curb greenhouse gas levels via carbon sequestration. Despite the importance of Loblolly pine and other conifers, genomic sequence information for this taxon is extremely limited. EST (expressed sequence tag) resources are available as a result of sequencing efforts at JGI and elsewhere, but actual genomic DNA sequence data is needed to enable efficient tree improvement.

JGI plans to sequence 100 clones from an existing BAC (Bacterial Artificial Chromosome) library. Of these clones, 50 will be selected based upon the presence of genes involved in carbon metabolism and/or wood formation, 25 will be chosen based on low-repeat content, and 25 will be picked at random. BAC sequencing will provide unprecedented insight into genome organization in Loblolly pine and conifers in general.

The principal investigators are Daniel G. Peterson (Mississippi State University), Jeffrey F.D. Dean (University of Georgia), C. Dana Nelson (USDA Forest Service), Ronald R. Sederoff (North Carolina State University), and Daniel S. Rokhsar (DOE JGI).


3


4


5

2. Cotton

Cotton *(Gossypium)* is one of the world's most important crops, and it sustains one of the world's largest industries (textiles). The value of cotton fiber and byproducts grown in the United States is typically about $6-7 billion per year. Increased durability and strength of cotton fiber offers the opportunity to replace the synthetic fibers that require more than 200 million barrels of petroleum per year to produce in the U.S. alone, also increasing use of bio-based products and increasing farm income. Its seed oil, and also the byproducts of cotton processing, are potential raw materials for production of biofuels. The unique structure of the cotton fiber makes it useful in bioremediation, and accelerated cotton improvement also promises to reduce pesticide and water use. A CSP 2009 project will sequence the cotton genome.

Cotton fibers, which can be produced in vitro, represent an excellent single-celled genomics platform, building on a distinguished history of seminal contributions to molecular biology by accelerating the study of gene function, particularly regarding cellulose biosynthesis that is fundamental to bioenergy production. A genome sequence for *Gossypium* will facilitate "phylogenetic shadowing" of the Brassicales (an order of flowering plants), including *Arabidopsis* and *Capsella*, enabling discovery of otherwise cryptic genomic features such as conserved non-coding sequences. Sequencing under this proposal will augment the JGI's 2006 CSP pilot sequencing of *Gossypium raimondii,* selected by the worldwide cotton community as the first cotton genotype to be fully sequenced. The United Nations declaration of 2009 as the International Year of Natural Fibers makes this project particularly timely.

The principal investigator is Andrew H. Paterson (University of Georgia).

3. Eucalyptus

The largest new plant project approved for the 2008 Community Sequencing Program was the eucalyptus tree genome, with a 600-million-nucleotide genome. The biomass production and carbon sequestration capacities of eucalyptus trees match DOE's and the nation's interests in alternative energy production and global carbon cycling.

A major challenge for the achievement of a sustainable energy future is our understanding of the molecular basis of superior growth and adaptation in woody plants suitable for biomass production. Eucalyptus species are among the fastest growing woody plants in the world and, at approximately 18 million hectares in 90 countries, the most widely planted genus of plantation forest trees in the world. Eucalyptus, only the second tree to be sequenced, is also listed as one of the DOE's candidate biomass energy crops.

This international effort, geared to the generation of resources for renewable energy, is led by Alexander Myburg of the University of Pretoria, South Africa, with Gerald Tuskan of Oak Ridge National Laboratory (and DOE JGI), and Dario Grattapaglia, of EMBRAPA Genetic Resources and Biotechnology (Brazil).

4. Foxtail Millet

The second-largest CSP project selected for sequencing in 2008 is foxtail millet *(Setaria italica)*. This forage crop is a close relative of several prospective biofuel crops, including switchgrass, napiergrass, and pearl millet. In the U.S., pearl millet is grown on some 1.5 million acres. Pearl millet would be useful as a supplement or replacement for corn in regions that suffer from drought and low-fertility soils.

This project is led by researchers at the University of Georgia, the University of Florida, the University of Missouri, the U.S. Department of Agriculture Agricultural Research Service–Cold Spring Harbor Laboratory, and the University of Tennessee.

5. Soybean

In December 2008 the DOE JGI released a complete draft assembly of the soybean *(Glycine max)* genetic code. Soybean, the principal U.S. source of biodiesel, has the highest energy content of any current alternative fuel and is much more environmentally friendly than comparable petroleum-based fuels. Biodiesel degrades rapidly in the environment and burns more cleanly than conventional fuels, releasing only half the pollutants and reducing the production of carcinogenic compounds by more than 80%.

Over 3.1 billion bushels of soybeans were grown in the U.S. on 75.5 million acres in 2006, with an estimated annual value exceeding $20 billion—second only to corn and approximately twice that of wheat. The soybean genome is about 1.1 billion nucleotides in size.

With the soybean genome sequence, researchers will be able to further enhance crop traits and ensure the effective application of this plant for renewable fuel production. Knowing which genes control specific traits, researchers can change the type and quantity of oil produced by the crop, as well as produce soybean plants that are more resistant to drought and disease.

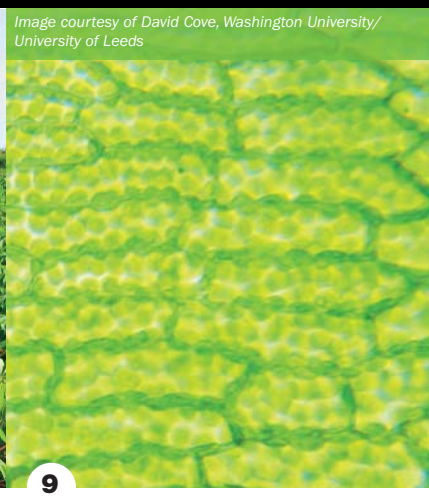
The soybean genome—all one billion nucleotides, roughly one-third the size of the human genome—was one of the largest and most complex plant genomes sequenced by the


6


7


8

Image courtesy of David Cove, Washington University/ University of Leeds


9

whole-genome shotgun strategy. The process entails shearing the DNA into small fragments, enabling the order of the nucleotides to be read and interpreted. Preliminary studies suggest as many as 66,000 genes—more than twice the number identified in the human genome sequence, and nearly half again as many as the poplar genome, sequenced by DOE JGI and published in the journal *Science* in September 2006.

Principal investigators on this project are Gary Stacey (National Center for Soybean Biotechnology, University of Missouri), Randy Shoemaker (USDA Agricultural Research Service), Scott Jackson (Purdue University), Bill Beavis (National Center for Genome Resources), Daniel Rokhsar (DOE JGI).

6. Sorghum

In 2007, the JGI released an assembly of the 730-million nucleotide sorghum genome through Phytozome *(www.phytozome.net/sorghum)*. Sorghum is representative of the tropical grasses in that it uses "C4" photosynthesis, with a complex combination of biochemical and morphological specializations resulting in more efficient carbon assimilation at high temperatures. By contrast, rice is more representative of temperate grasses, using "C3" photosynthesis.

In addition to its intrinsic value, the sorghum sequence will be a valuable reference for assembling and analyzing the four-fold larger genome of maize (corn), a tropical grass that is the leading U.S. fuel ethanol crop (sorghum is second). Sorghum is an even closer relative of sugarcane, arguably the most important biomass/biofuel crop worldwide with annual production of about 140 million metric tons and a net value of about $30 billion. Sorghum and sugarcane are thought to have shared a common ancestor about 5 million years ago, but sorghum's genome is about 25% the size of human, maize, or sugarcane genomes.

The sorghum project is a collaboration between Andrew H. Paterson (lead), John E. Bowers, and Alan R. Gingle (all three from the University of Georgia); C. Thomas Hash (International Crops Research Institute for the Semi-Arid Tropics); Stephen E. Kresovich (Cornell University); Joa-chim Messing (Rutgers); Daniel G. Peterson (Mississippi State University); and Daniel S. Rokhsar (JGI and University of California, Berkeley).

7. *Brachypodium*—A Model Grass

While herbaceous energy crops (especially grasses) are poised to become a major source of renewable energy in the United States, we know very little about the genetic traits that affect their utility for energy production. The temperate wild grass species, *Brachypodium distachyon,* is a new model plant being studied by the JGI for developing grasses into superior energy crops. *Brachypodium* is small in size, can be grown rapidly, is self-fertilizing, and has simple growth requirements. It can be used as a functional model to gain the knowledge about basic grass biology necessary to develop superior energy crops, like switchgrass and *Miscanthus*.

The sequencing of *Brachypodium* is being undertaken via a two-pronged strategy: the first, a whole-genome shotgun sequencing approach, is a collaboration between John Vogel and David Garvin, both of the USDA, and Michael Bevan at the John Innes Centre in England; and the second, an expressed gene sequencing effort, led by Todd Mockler and Jeff Chang at Oregon State University, with Todd Michael of The Salk Institute for Biological Studies, and Samuel Hazen from The Scripps Research Institute.

8. Cassava

The DOE JGI is also sequencing cassava *(Manihot esculenta)*, an excellent energy source and food for approximately one billion people around the planet. Its roots contain 20-40% starch, from which ethanol can be derived, making it an attractive and strategic source of renewable energy. Cassava grows in diverse environments, from extremely dry to humid climates, acidic to alkaline soils, sea level to high altitudes, and in nutrient-poor soil.

Sequencing the cassava genome will help bring this important crop to the forefront of modern science and generate new possibilities for agronomic and nutritional improvement. The cassava project will extend broad benefits to its vast research community, including a better understanding of starch and protein biosynthesis, root storage, and stress controls, enabling crop improvements while shedding light on similar mechanisms in related plants, including the rubber tree and castor bean.

The cassava project is led by Claude M. Fauquet, Director of the International Laboratory for Tropical Agricultural Biotechnology and colleagues at the Danforth Plant Science Center in St. Louis, and includes contributions from the USDA laboratory in Fargo, ND; Washington University in St. Louis; University of Chicago; The Institute for Genomic Research (TIGR); Missouri Botanical Garden; the Broad Institute; Ohio State University; the International Center for Tropical Agriculture (CIAT) in Cali, Colombia; and the Smithsonian Institution.

9. *Physcomitrella*—The First Moss Genome

Messages from nearly a half-billion years ago, conveyed via the inventory of genes sequenced

from a present-day moss, provide clues about the earliest colonization of land by plants. The JGI was among the leaders of an international effort uniting more than 40 institutions to complete the first genome sequencing project of a nonvascular land plant, the moss *Physcomitrella patens.* The team's insights into the code that enabled this seminal emergence and dominance of land by plants were published online in *Science Express* in December 2007.

The availability of the *Physcomitrella* genome is expected to create important new opportunities for understanding the molecular mechanisms involved in plant cell wall synthesis and assembly. The ease with which genes can be experimentally modified in *Physcomitrella* will facilitate a wide range of studies of the cell wall, the principal component of terrestrial biomass.

There is also a clear connection with this work and the intensifying interest in the global carbon cycle. The moss system is proving useful for studies of photosynthesis, among many other processes. One of these is the ability of mosses to withstand drought and, in some cases, complete desiccation, which will provide us with a model experimental system to identify genes and gene networks that might be involved in, and related to, seed desiccation in flowering plants.

The moss genome project, originally proposed by Brent Mishler of UC Berkeley, and Ralph Quatrano of Washington University in St. Louis, was enabled by the CSP.

10. Greater Duckweed

The smallest, fastest growing, and simplest of flowering plants, Greater Duckweed *(Spirodela polyrhiza)* is less than 10 millimeters tall. Nevertheless, its utility is manifold: as a biotech protein factory, toxicity testing organism, wastewater remediator, high-protein animal feed, carbon cycling player, as well as basic research and evolutionary model system. Selected as a CSP 2009 project, duckweed relates to all three of DOE JGI's mission areas: bioenergy, bioremediation, and global carbon cycling.

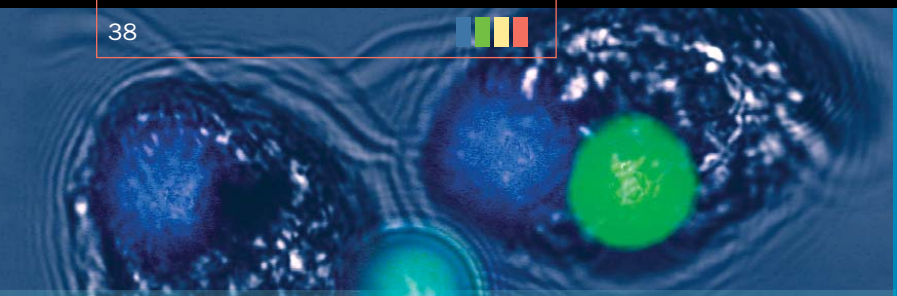
Duckweed plants produce biomass faster than any other flowering plant, and their carbohydrate content is readily converted to fermentable sugars by using commercially available enzymes developed for corn-based ethanol production. Propagated on agricultural and municipal wastewater, *Spirodela* species efficiently extract excess nitrogen and phosphate pollutants. Duckweed growth on ponds effectively reduces algal growth (by shading), coliform bacteria counts, suspended solids, evaporation, biological oxygen demand, and mosquito larvae while maintaining pH, concentrating heavy metals, sequestering or degrading halogenated organic and phenolic compounds, and encouraging the growth of aquatic animals such as frogs and fowl.

This project, submitted by Todd Michael of the Waksman Institute of Microbiology at Rutgers, The State University of New Jersey, unites the efforts of six institutions.

> Sequencing of diverse plant genomes is driven by the importance of plants as feedstocks for biofuel production, and more generally as key contributors to the global carbon cycle.

11. Harnessing Waste Biomass

It is estimated that 236 million tons of municipal solid waste is produced annually in the U.S., 50% of which is biomass. Converting organic waste to renewable biofuel represents an appealing option to exploit this potential resource. In California alone, it is estimated that 22 million tons of organic waste is generated annually, which, if converted by microbial digestion, could produce biogas—primarily methane and carbon dioxide—equivalent to 1.3 million gallons of gasoline per day. When biogas is cleaned of its particulates and carbon dioxide is removed, it has the same characteristics as natural gas, also known as biomethane. Yet little is known about the microorganisms involved and their biology.



DOE JGI microbial genome program

Image courtesy of Emily Roberts, Swansea University, UK

What are Eukaryotes and Prokaryotes?

Eukaryotes are animals, plants, fungi, and protists (microorganisms such as protozoa and algae). Prokaryotes include bacteria and archaea (a recently discovered branch of life intermediate between eukaryotes and prokaryotes). The cells of eukaryotic organisms have a nucleus containing their DNA, as opposed to those of prokaryotic cells, which lack a nucleus. Eukaryotic cells are usually much larger than prokaryotic cells and have a more complex structure. Eukaryotes, bacteria, and archaea comprise the three branches of life.

Eukaryotic microbes play essential roles in biomass degradation, global carbon cycling, and fermentation processes which are currently being investigated for biofuels potential. Given their important role in processes that are central to the DOE Office of Biological and Environmental Research (OBER) mission, it is not surprising that DOE JGI has, through its user programs, sequenced a number of fungal, algal, and protist genomes, with many more on the way. Given the growing number and importance of these genomes, DOE JGI has recently launched the Eukaryotic Microbe Genome Program to facilitate their sequencing, manage user interactions, and most importantly, plan for the growing genomic infrastructure needs of this community.

The motivation for the JGI to target microbial genomes is that, embedded in their sequence information, is the complete gene inventory of those organisms. With the "part list" in hand for microbes, researchers can explore how microbes use these parts to build and sustain key functions of critical importance to DOE. These include the symbiotic organisms, those that support the growth of plant biomass, and the microorganisms that possess enzymatic activity that can break down plant material to produce renewable sources of energy.

Prokaryotic genomes are particularly amenable to shotgun sequencing and assembly because of their compact size and limited repetitive content. The DOE JGI sequenced its first prokaryotic genome in 2000 and explored the limits of high-throughput sequencing later that year by producing draft sequences of 15 organisms in a single month. DOE JGI's first user program, the Microbial Genome Program (MGP), was initiated the following year to provide sequences of organisms from the environment to DOE users. Since then, DOE JGI has produced more than 350 bacterial and archaeal genomes, of which 70% are currently at finished quality (no gaps and <1 error/50kb). According to the Genomes Online Database *(www.genomesonline.org)*, DOE JGI is the world's leading producer of prokaryotic genomes.

A major strength of DOE JGI's microbial genome program can be found in the contribution of the several partner laboratories. Library construction and production sequencing occur at the JGI, while finishing is carried out at LANL. ORNL provides automated annotation of draft genomes and curated annotation of finished genomes. Genomes are then published on DOE JGI's website, submitted to GenBank, and placed into DOE JGI's comparative genomics platform, the Integrated Microbial Genomes (IMG) data management system, which is described in detail on page 50. The bulk of DOE JGI prokaryotic genomes have been proposed through the Community Sequencing Program, the DOE Microbial Genome Program, and the Genomic Encyclopedia of Bacteria and Archaea (GEBA), described in detail on page 42.

1. Predatory Nanoflagellates

Heterotrophic nanoflagellates are a group of marine microbes that prey on other microbes, such as bacteria and phytoplankton. Bacteria and phytoplankton constitute a dominant fraction of the living biomass in marine ecosystems. Their fates are dictated largely by two major forces: predation by protists like the nanoflagellates, and cell death induced by viruses. These two forces have very different consequences on the fate of phytoplankton and bacterial carbon, affecting whether it can be sequestered into larger sinking particles/cells (via predation) or simply released back into the environment (viruses). Thus predatory protists play critical roles in marine carbon cycling.

However, little is known about the basic biology of predatory nanoflagellates. While rates of consumption, growth, and even assimilation and egestion (discharge of undigested food), have been quantified, the underlying mechanisms remain largely unexplored. Thus the molecular underpinnings of heterotrophic nanoflagellate behavior and interactions with prey are still unknown. Currently no sequenced genomes are available from these organisms.

This project, approved for the 2009 Community Sequencing Program (CSP), is designed to allow researchers to "listen to the message" generated by these organisms and to deduce real-time behavior based on genomic composition and mRNA expression. An international group of investigators led by Alexandra Worden (Monterey Bay Aquarium Research Institute) will collaborate in sequencing the genomes of three target species. These are *Paraphysomonas imperforata,* a widespread marine heterotroph; *Ochromonas,* a marine mixotroph (mixotrophs can synthesize nutrients from both inorganic and organic compounds, whereas heterotrophs must rely on organic material alone); and a third organism, *Spumella,* which is a heterotrophic nanoflagellate common in freshwater systems. This suite of organisms provides a unique opportunity for comparative studies (e.g. heterotrophy vs. phototrophy and mixotrophy, marine vs. freshwater habitats) that will greatly facilitate carbon cycling research, addressing evolutionary developments, novel biochemical pathways, and the integration of potentially competing metabolic processes.

The principal investigators are Andrew Allen (J.Craig Venter Institute), Jens Boenigk (Austrian Academy of Sciences), David A. Caron and Karla B. Heidelberg

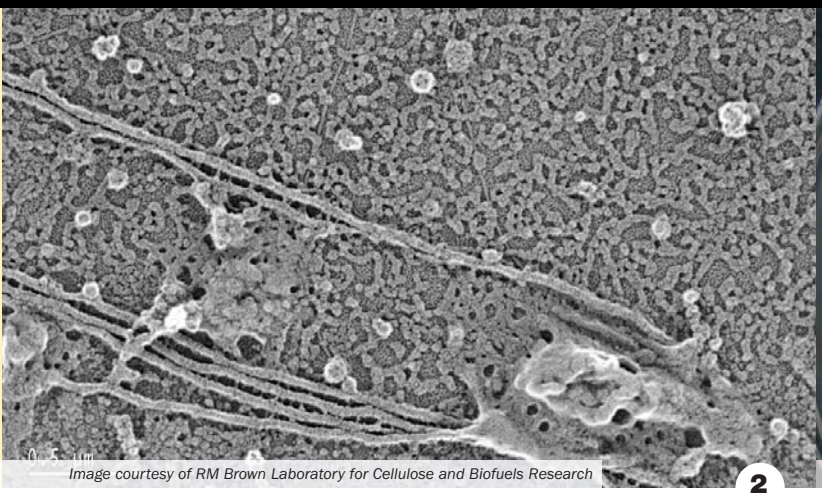


Image courtesy of RM Brown Laboratory for Cellulose and Biofuels Research

2

3

Image courtesy of Andriy Sibirny, NAS of Ukraine

4

(University of Southern California), Emily Roberts (Swansea University), and Alexandra Z. Worden.

2. Cellulose-degrading Bacteria

One of the major DOE missions is the production of renewable fuels to both reduce our dependence on foreign oil and also to take the place of petroleum-based fuels as these resources dwindle. Biologically produced ethanol is one possible replacement for fossil fuels. Currently, ethanol is produced from corn starch, but there is much research into using lignocellulosic materials (those containing cellulose, hemicellulose, and lignin) as the raw material for ethanol production.

Ethanol production from cellulose requires several steps: pretreatment with steam, acid, or ammonia; digestion of cellulose to sugars; and fermentation of sugars to ethanol. The slowest and most expensive step is the breakdown of cellulose, chemically accomplished by cellulases. The second and third steps can be carried out simultaneously in a process called Simultaneous Saccharification and Fermentation (SSF). This prevents inhibition of cellulases by glucose and cellobiose. These two steps can also be combined in one organism in the Consolidated Bioprocess (CBP). CBP can be carried out by adding cellulase genes to fermenting organisms, or introducing a fermentation pathway to a cellulose-degrading organism.

As a CSP 2009 project, JGI will be sequencing three phenotypically and phylogenetically diverse cellulose-degrading microbes to expand the repertoire of cellulases available for biotechnological production of ethanol. *Byssovorax cruenta* is a cellulolytic myxobacterium. *Eubacterium cellulosolvens* is a rumen bacterium from the family Eubacteriaceae of the class Clostridia. *Cellvibrio mixtus* is a multiple polysaccharide degrader from the family Pseudomonadaceae. Cellulolytic and other saccharolytic enzymes have already been identified from some of these organisms, and having the complete genome sequences available will speed up the process of characterizing new enzymes.

The principal investigator is Iain Anderson (DOE JGI).

3. *Hansenula polymorpha*

Another 2009 CSP project is *Hansenula polymorpha* strain NCYC 495 leu1.1, a yeast capable of fermenting xylose, cellobiose, and glucose to ethanol at high temperatures (45–50°C). This particular strain ferments xylose much more efficiently than other strains of *H. polymorpha.* Hence, it is a promising strain for the simultaneous saccharification and fermentation (SSF) process, which combines enzymatic hydrolysis of pretreated lignocellulose by cellulases and hemicellulases (exhibiting optimal activities around 50°C) with simultaneous fermentation of produced hexoses and pentoses to ethanol in the same vessel. Commercially feasible SSF technology has not been developed yet because of the absence of a robust organism capable of efficient ethanolic fermentation of xylose and other lignocellulosic sugars at high temperatures. Still, ethanol yield from xylose fermented by this strain is rather low and must be improved via metabolic engineering approaches. This work can be efficient and successful only when a complete genomic sequence is available.

Sequencing will enable the identification of the limiting steps (bottlenecks) in the fermentation pathway from xylose to ethanol. In addition, *H. polymorpha* NCYC 495 is also a popular system for basic research, such as studies of growth and degradation of peroxisomes (organelles that metabolize fatty acids and toxins), nitrate assimilation, resistance to heavy metals and oxidative stress, as well as enabling organisms to produce proteins of biotechnological significance. Studies of this organism's modes of protein and organelle degradation are important for elucidating the mechanisms of severe human diseases (Huntington's chorea, Alzheimer's disease, some forms of cancer). Scientists from diverse fields will greatly benefit from a complete genomic sequence database of this organism.

The principal investigator is Andriy Sibirny (National Academy Science of Ukraine and Rzeszów University, Poland).

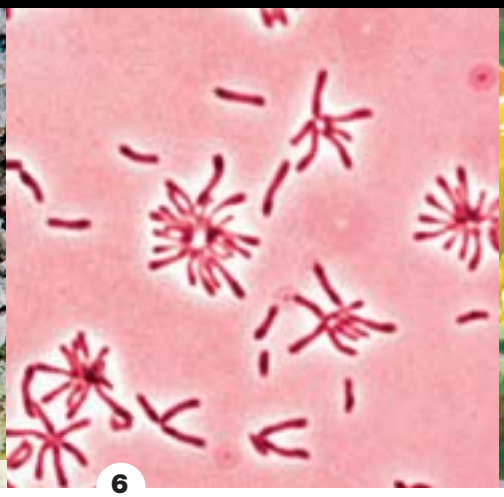
4. Alfalfa bacterium

Sinorhizobium meliloti, is the most well-studied and abundant species of rhizobia, a group of bacteria that infects the roots of legumes and form nodules in which they fix atmospheric nitrogen into plant matter. Rhizobia enable legumes such as peas, beans, clover, and alfalfa to grow in nitrogen-poor soils with no need for costly fertilizers. One strain of *S. meliloti* has been sequenced, but because different strains have adapted to diverse environments worldwide, *S. meliloti* strains exhibit tremendous genetic variability.

A CSP 2009 project will sequence two new strains of *S. meliloti* that each contain desirable traits. Strain AK83 allows alfalfa to thrive in the desert-like North Aral Sea region of Kazakhstan, where extremely salty soils prevent many plants from growing. Strain BL225C comes from an agricultural region in the north of Italy and is highly efficient, producing plants with three to four times as much dry weight as strain AK83.

S. meliloti forms a symbiotic association with alfalfa, a crop traditionally used as animal

5 Image courtesy of Tamás Török, LBNL



6

7

feed but now under study by the USDA and the University of Minnesota as a promising feedstock for biofuel production. Alfalfa's helper bacteria give it a competitive advantage over current bioenergy crops such as corn, which require fertilizers. Because commercial fertilizer production is expensive and produces greenhouse gases, alfalfa would be a more environmentally sustainable biofuel crop than corn and other cereals. In addition, alfalfa can grow on marginal lands, so it would not use up prime cropland.

By comparing the genomes of these two strains with the previously sequenced reference strain, researchers may discover the genetic basis for traits such as salt resistance and nodulation efficiency. This would open the possibility of engineering a new strain by combining desirable traits already present in the species but not normally found together. Such a strain could produce high-yield alfalfa plants that can grow on land unsuitable for agriculture—an ideal bioenergy crop.

The principal investigator is Emanuele Biondi (University of Florence, Italy).

5. *Desulfurococcus*

The genus *Desulfurococcus* represents a unique clade (a group of organisms all descended from a common ancestor) in the domain of archaea that is not currently represented in the publicly available genomic databases. This scenario will change to some extent as *Desulfurococcus* mucosus genome sequence is determined under the GEBA (Genomic Encyclopedia for Bacteria and Archaea) Pilot Project of the JGI; it is also a CSP 2009 project.

The motivation for this project comes from the observation that *Desulfurococcus fermentans* is the only known archaeon that breaks down cellulose, and that, unlike most known microorganisms that carry out fermentation, it produces hydrogen in the presence of hydrogen while fermenting cellulose and starch without experiencing an inhibition of growth. On the other hand, the starch-utilizing *Desulfurococcus amylolyticus* and starch non-utilizing and solely peptide-fermenting *Desulfurococcus mobilis* and *D. mucosus* are inhibited by hydrogen and stimulated by sulfur; *D. fermentans* is not stimulated by sulfur. Since these organisms are closely related, differences in their genomes will highlight differences in their metabolisms.

A comparative genomics investigation involving *D. fermentans, D. amylolyticus, D. mobilis,* and *D. mucosus* will allow a rapid development of hypotheses about the special molecular tools and regulatory processes that *D. fermentans* and *D. amylolyticus* utilize for degrading starch and those that *D. fermentans* employs for fermenting cellulose and producing hydrogen. It will help to reveal the finer details that distinguish proton reduction (producing hydrogen) from sulfur reduction in fermentative archaea and provide hypotheses concerning the evolution of these two fermentation pathways. In addition, since a strict dependence on sulfur is found in several hyperthermophilic fermentative archaea, the genome data will help to define the evolutionary and metabolic relationships of the *Desulfurococcus* species with their archaeal relatives.

The principal investigator is Biswarup Mukhopadhyay (Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University).

6. *Rhodopseudomonas palustris* strain DX-1

To date, six strains of *Rhodopseudomonas palustris* have been sequenced, all by the JGI, but the strain to be sequenced as part of the 2009 CSP has a most shocking ability: it is exoelectrogenic. In other words, it can directly generate electricity from the biodegradation of organic and inorganic matter. In fact, it produces very high power densities in low-internal-resistance microbial fuel cells (MFCs), a technology that shows great promise as a method of bioelectricity production from waste biomass.

In an MFC, exoelectrogenic bacteria oxidize organic matter and transfer electrons to the anode, where they flow to the counter electrode (cathode) and react with protons and oxygen to form water. Our ability to understand factors that affect electricity generation by exoelectrogenic bacteria has been limited by a lack of high-power-producing strains of microorganisms. While some iron-reducing bacteria are known to be able to generate electricity, few microorganisms have been directly isolated from MFCs. Although not many direct comparisons have been made, all isolates so far examined show power densities equal to or less than those produced by acclimated mixed cultures under otherwise identical conditions.

The photosynthetic bacterium *R. palustris* strain DX-1, however, can produce higher power densities than mixed cultures in the same MFC. We anticipate that genome sequence comparisons between DX-1 and the other sequenced strains of *R. palustris* will reveal key biochemical characteristics of strain DX-1 that are critical for its ability to generate power. The genome sequence will be used to develop and test hypotheses about biological mechanisms that drive electricity generation by strain DX-1 and by bacteria in general. The sequence would also make *R. palustris* strain DX-1 readily accessible as a model organism (along with *Shewanella* and *Geobacter*) for MFC investigations.

The principal investigators are Caroline S. Harwood (University of Washington) and Bruce E. Logan (Pennsylvania State University).

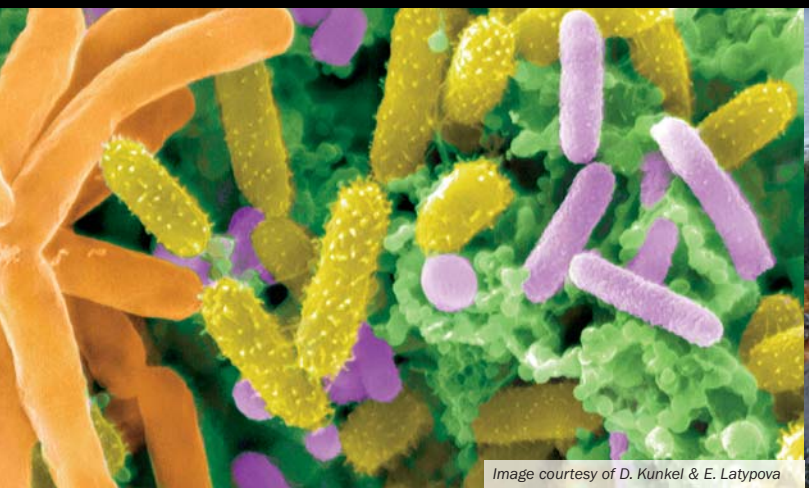


Image courtesy of D. Kunkel & E. Latypova

Image courtesy of William Hickey, University of Wisconsin.

8

7. *Methylotenera*

Metabolism of organic C1 compounds (compounds containing no carbon-carbon bonds) is an important part of the global carbon cycle. Methane has been recognized as one of the major C1 compounds in the environment and a major contributor to the greenhouse effect. While global emissions of other C1 compounds (methanol, methylated amines) have historically attracted less attention, recent models put their emissions on a scale similar to the scale of methane emissions. JGI plans to sequence three methylotrophs (degraders of C1 compounds) of the genus *Methylotenera* as part of its CSP 2009 plans.

Methylotrophic bacteria play a major role in maintaining the balance of C1 compounds in aerobic (oxygenated) environments. They are ubiquitous and are found across a range of oxygen tension, salinity, pH, and temperature. Methylotrophy is also widespread across bacterial groups and is found in alpha, beta, and gamma Proteobacteria; in Gram-positive bacteria; and recently in Verrucomicrobia. In addition to their role in natural environmental processes, methylotrophs have potential in bioremediation of environmental pollutants such as chlorinated solvents and methyl tert-butyl ether (MTBE).

This work follows previous work by JGI in sequencing methylotrophic communities from Lake Washington. Sequencing members of the *Methylotenera* will enable genome-wide comparisons within a single family (*Methylophilaceae*) and across phyla. The datasets generated in this project will also be invaluable for JGI, as these will allow comparisons between metagenomic *Methylotenera* scaffolds and the finished genomes, as a way to test and validate the tools currently used for assembly, quality control, and analysis of metagenomic data.

The principal investigators are Ludmila Chistoserdova and Marina G. Kalyuzhnaya (University of Washington) and Alla Lapidus (DOE JGI).

In 2009 JGI will sequence a microbe with a shocking ability: it's exoelectrogenic, meaning it can directly generate electricity from the biodegradation of organic and inorganic matter.

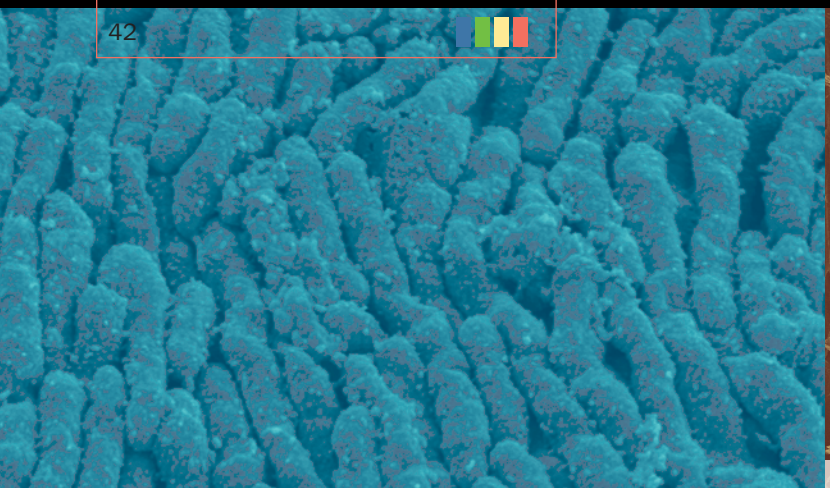
8. PAH-Degrading Soil Bacteria

Polycyclic aromatic hydrocarbons (PAHs) are widespread pollutants of soil and sediment. Many are carcinogenic and are contaminants of concern at DOE sites. The health threats of PAHs are compounded by the fact that they are fat-soluble and have a strong potential to accumulate in tissues and to increase in concentration over time. Thus, PAH-contaminated sediments and soils are high priorities for remediation. A CSP 2009 project will sequence two strains of the PAH-degrading soil bacteria *Burkholderia*.

Bioremediation via the microbial biodegradation of PAH could be an economical means to eliminate the human health threats posed by these compounds. Unfortunately, because key processes that determine its effectiveness are either poorly understood or are simply "black boxes," the success of PAH bioremediation has been limited at best. Making bioremediation a viable remedial technology for PAH cleanup will require improving our understanding of these ill-defined areas, one of the most important being the characteristics of the environment and of the microbes that affect the PAHs' availability for degradation—their bioavailability.

In soil, many factors can affect PAH bioavailability, but two of the most important are diffusion into soil nanopores and uptake by organic matter. Each of the PAH-degrading soil bacteria, *Burkholderia* sp. strains Ch1-1 and Cs1-4, possesses unique features that may enable it to overcome these key bioavailability constraints. Sequencing the genomes of such organisms should enable researchers to determine how the degradation process works. Genome sequencing of strains Ch1-1 and Cs1-4 would also add much needed information about metabolic diversity among *Burkholderia* outside of the well studied *B. cepacia* complex. Collectively, genome sequencing of strains Ch1-1 and Cs1-4 would be valuable in providing insights into the molecular basis by which these organisms affect an environmentally important process, provide further insights into niche adaptation in *Burkholderia*, and expand our knowledge of speciation in less thoroughly studied strains.

The principal investigator is William Hickey (University of Wisconsin-Madison).



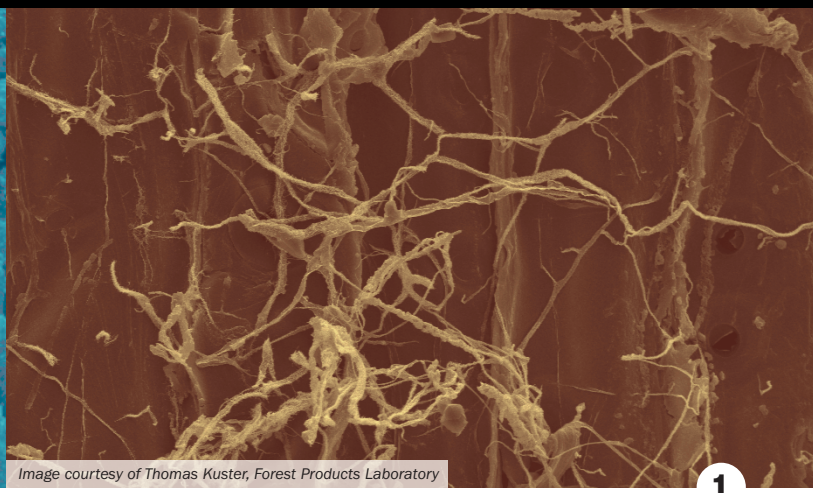


Image courtesy of Thomas Kuster, Forest Products Laboratory

1

A Genomic Encyclopedia of Bacteria and Archaea (GEBA)

The introduction of highly parallel new sequencing technologies allowed DOE JGI to augment user-defined prokaryotic sequencing with a programmatic effort to expand the phylogenetic representation of sequenced genomes through the Genomic Encyclopedia of Bacteria and Archaea (GEBA) program. The GEBA pilot project, approved as part of DOE JGI's Laboratory Science Program, is aimed at systematically filling in the gaps in sequencing along the bacterial and archaeal branches of the tree of life.

Though the wide variety of microbial sequencing projects undertaken throughout the world has created a valuable collection of microbial genomes, strong biases in what has been sequenced thus far are evident. This bias has resulted in a major gap in our knowledge of microbial genome complexity and our understanding of the evolution, physiology, and metabolic capacity of microbes. This is surprising given that there are systematic efforts to sequence genomes from diverse groups in animals, plants, and fungi. Although there have been small efforts in this arena for microbes (e.g., eight genomes from novel branches are being sequenced as part of an NSF Tree of Life project), there has been no sustained, systematic effort to expand phylogenetic diversity among sequenced genomes.

GEBA is a response to the resounding need for a large-scale systematic effort to sequence genomes to fill in gaps in the tree of life. It represents the first systematic attempt to use the tree of life itself as a guide to sequencing target selection. DOE JGI is beginning by collaborating on a pilot project with the German Resource Center for Biological Materials (DSMZ) providing the microbial DNA.

This phylogenomic approach will be of great value in multiple areas of public and general scientific interest. The potential benefits include:

- improved identification of protein families and orthology groups across species, which will improve annotation of other microbial genomes
- improved phylogenetic anchoring of metagenomic data
- gene discovery (which tends to be maximized by selecting phylogenetically novel organisms)
- a better understanding of the processes underlying the evolutionary diversification of microbes (e.g., lateral gene transfer and gene duplication)
- a better understanding of the classification and evolutionary history of microbial species
- improved correlations of phenotype and genotype in microbes.

Fungal Genomics Program

Encoded in the genomes of the organisms of the kingdom Fungi are biological processes with major relevance to DOE missions in bioenergy production, biogeochemistry, and carbon cycling. We know this from both experimental and genome sequence-based evidence. For example, in the area of enzymatic biomass deconstruction, an important process in the production of cellulosic-based renewable fuels, years of experimental biochemical-based evidence suggests that fungi are able to easily deconstruct plant biomass into its component sugar building blocks. Genomic-based evidence suggests that the experimentally described enzymatic activities are merely a shadow of the potential enzyme activities encoded within a fungal genome.

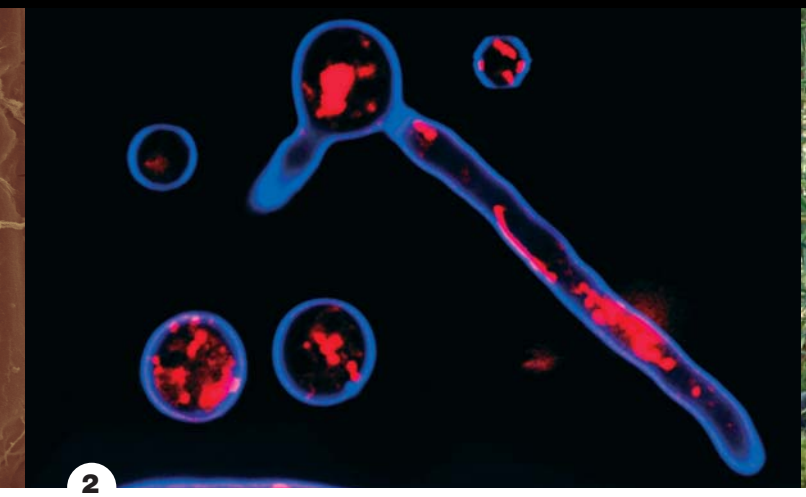
DOE JGI is engaged in sequencing a significant breadth and depth of fungi to establish a rich catalog of tools in the form of pathways and enzymes that can be brought to bear on the important problems related to DOE missions. With the formation of the DOE JGI Fungal Genomics Program comes the opportunity for a unified message to articulate the needs of the fungal biology research community with respect to genomics.

1. Brown Rot Fungus

Among the challenges to more cost-effective production of biofuels from cellulosic biomass—the fibrous material of whole plants—is to find effective means to work around the polymer lignin, the scaffolding that endows the plant's architecture with rigidity and protection from pests. By doing so, the organic compound cellulose—the long chain of glucose (sugar) units—can be unbound, broken down, fermented, and distilled into liquid transportation fuel. This is where the destructive capabilities of rot come in.

An international team led by scientists from the DOE JGI and the U.S. Department of Agriculture Forest Service, Forest Products Laboratory (FPL) have translated the genetic code that explains the complex biochemical machinery making brown-rot fungi *(Postia placenta)* uniquely destructive to wood. The same processes that provide easier access to the energy-rich sugar molecules bound up in the plant's tenacious architecture are leading to innovations for the biofuels industry. The research, conducted by more than 50 authors, is reported in the February 4, 2009 online edition of the *Proceedings of the National Academy of Sciences (PNAS).*

The microbial world represents a little-explored yet bountiful resource for enzymes that can play a central role in the deconstruction of plant biomass—an early step in biofuel production.



2

3

The genome of *Postia placenta* offers a detailed inventory of the biomass-degrading enzymes that this and other fungi possess. *Postia* has, over its evolution, shed the conventional enzymatic machinery for attacking plant material. Instead, the evidence suggests that it utilizes an arsenal of small oxidizing agents that blast through plant cell walls to depolymerize the cellulose. This biological process opens a door to more effective, less energy-intensive, and more environmentally sound strategies for more lignocellulose deconstruction.

Few organisms in nature can efficiently breakdown lignin into smaller, more manageable chemical units amenable to biofuels production. The exceptions are the basidiomycete fungi, which include white-rot and brown-rot, wood-decayers and essential caretakers of carbon in forest systems. In addition, brown-rot fungi have significant economic impact because their ability to wreak havoc with wooden structures. A significant portion of the U.S. timber harvest is diverted toward replacing such decayed materials.

Unlike white-rot fungi, previously characterized by DOE JGI and FPL, which simultaneously degrades lignin and cellulose, brown-rot rapidly depolymerizes the cellulose in wood without removing the lignin. Up until this study, the underlying genetics and biochemical mechanisms were poorly understood.

2. *T. Reesei,* the Tent Fungus

Trichoderma reesei was discovered during World War II, when it was identified as the culprit responsible for the deterioration of fatigues and tents in the South Pacific. The DOE JGI has completed the genome analysis of this champion biomass-degrading fungus and found a surprisingly minimal repertoire of genes that it employs to break down plant cell walls, highlighting opportunities for further improvements in enzymes customized for biofuels production. The results were published in *Nature Biotechnology* in May 2008 by a team of government, academic, and industry researchers led by the DOE JGI and LANL.

The genome analysis of *T. reesei* can facilitate accelerated research into optimizing fungal strains for reducing the prohibitively high cost of converting lignocellulose to fermentable sugars. Improved industrial enzyme "cocktails" from *T. reesei* and other fungi will enable more economical conversion of biomass from such feedstocks as the perennial grasses *Miscanthus* and switchgrass, wood from fast-growing trees like poplar, agricultural crop residues, and municipal waste, into next-generation biofuels.

The research team compared the 34-million-nucleotide genome of *T. reesei* with 13 previously characterized fungi and discovered something counterintuitive. Despite its reputation as an avid plant polysaccharide degrader, *T. reesei,* was found to have the smallest inventory of genes powering its robust degradation machinery. The team also observed clustering of carbohydrate-active enzyme genes, which suggested a specific biological role: polysaccharide degradation.

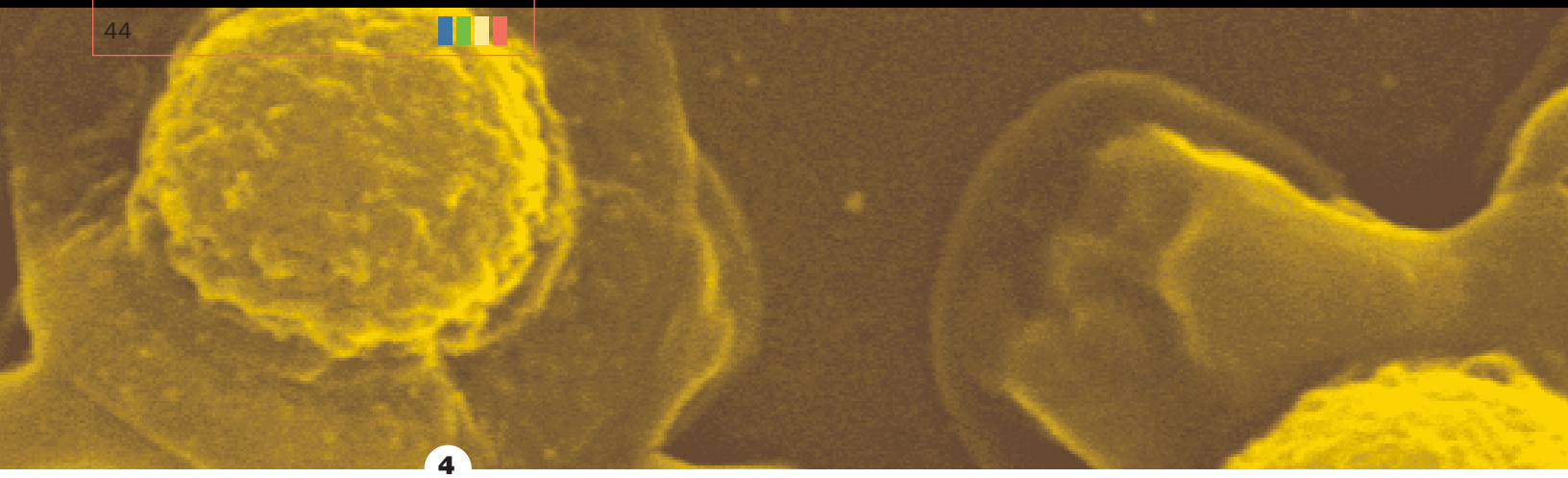
A 2009 CSP project is to resequence *T. reesei,* focusing on mutant strains. Academic and industrial research programs have over several years, through random mutagenesis, produced strains of *T. reesei* whose production of cellulases is several times higher than that of the "original" *T. reesei* cellulose-producing strain, QM6a, isolated from U.S. Army tent canvas in 1944 in the Solomon Islands. Understanding the molecular mechanisms that underlie the improvements made through random mutagenesis of *T. reesei* QM6a could lead to better and more efficient cellulose-producing strains created through targeted molecular genetic manipulation rather than through a random mutagenesis process that leads to collateral and deleterious genome damage. In addition, sequencing one of its natural teleomorphs, which forms fruiting bodies in its habitat, will enable the identification of the genomic basis for meiosis in *Trichoderma*, which can subsequently be used to render the organism susceptible to classic genetic manipulation.

T. reesei is the workhorse organism for a number of industrial enzyme companies for the production of cellulases. The costs associated with enzymes that degrade biomass are considered a bottleneck to economic lignocellulosic fuel ethanol. The DOE has made large investments in bioenergy research with the goal of economically viable cellulosic ethanol. One of the key barriers to a viable cellulosic fuel ethanol process emphasized by program offices within both DOE Offices of Science and Energy Efficiency and Renewable Energy is the cost of enzymes for the degradation of cellulosic biomass. Currently, the cellulases used in test and pilot cellulosic ethanol plants are produced by fungi, in many cases *T. reesei.* The widespread use of *T. reesei* in cellulase production underscores the importance of this organism and understanding the mechanisms behind enzyme secretion.

The principal investigators are Scott E. Baker and Jon Magnuson (PNNL), Christian P. Kubicek (Technical University of Vienna), Randy M. Berka (Novozymes), N. Jamie Ryding (Verenium), and Jan-Fang Cheng (DOE JGI).

3. *Laccaria bicolor*

The complete DNA sequence of *Laccaria bicolor,* a fungus that forms a beneficial symbiosis with poplar and other trees and inhabits one of the most ecologically and commercially important microbial niches in North American



4

and Eurasian forests, was determined by the JGI and was announced in 2006 and published in the journal *Nature* in March 2008. The analysis of the *Laccaria* genome has yielded insights into the mechanisms of symbiosis between the fungus and the roots of plants and also provides insights into plant health that may enable more efficient carbon sequestration and enhanced phytoremediation—using plants to clean up environmental contaminants.

Key factors behind the ability of trees to generate large amounts of biomass, or store carbon, reside in the way that they interact with soil microbes known as mycorrhizal fungi, which excel at procuring necessary, but scarce, nutrients such as phosphate and nitrogen. When *Laccaria bicolor* partners with plant roots, a mycorrhizal root is created, resulting in a mutual relationship and making these nutrients available to their host, and significantly benefiting both organisms. The fungus within the root is protected from competition with other soil microbes, and gains preferential access to carbohydrates within the plant.

The *Laccaria* genome sequence will provide the global research community with a critical resource to develop faster-growing trees for producing more biomass that can be converted to fuels, and for trees capable of capturing more carbon from the atmosphere. The DOE JGI and its collaborators have now embarked on characterizing several other poplar community symbionts that will provide a more comprehensive understanding of the biological community of the poplar forest. These include *Glomus*—a second plant symbiotic fungus, *Melampsora*—a leaf pathogen, and several plant endophytes—bacteria and fungi that live inside the poplar tree.

This research will advance the understanding of how functional genomics of this symbiosis enhances biomass production and carbon management, particularly through the interaction with the poplar tree, also sequenced by the JGI. It will now be possible to harness the interaction between these species and identify the factors involved in biomass production by characterizing the changes in the two genomes as the tree and fungus collaborate to generate biomass. It will also help scientists understand the interaction between these two symbionts within the context of the changing global climate.

4. *Pichia*

Lignocellulosic biomass—the complex of cellulose, hemicellulose, and lignin—is derived from such plant-based feedstocks as agricultural waste, paper and pulp, wood chips, grasses, and trees. Under current strategies for generating lignocellulosic ethanol, these forms of biomass require expensive and energy-intensive pretreatment with chemicals and/or heat to loosen up this complex. Enzymes are then employed to break down complex carbohydrates into sugars, such as glucose and xylose, which can then be fermented to produce ethanol. Additional energy is required for the distillation process to achieve a fuel-grade product. Now, the power of genomics is being directed to optimize this age-old process so that biofuels ultimately become more economically competitive with fossil fuels.

Saccharomyces cerevisiae, brewer's yeast, has long been employed for the fermentation of plant sugars into ethanol—beer and wine being prime examples. However, there are ways to improve the process, yielding higher levels of ethanol from lignocellulose. The information embedded in the genome sequence of *Pichia* has helped to identify several gene targets to improve xylose metabolism. Efforts are underway to engineer these genes to drive the fermentation process to higher concentrations of ethanol.

Commercial development of biofuels from lignocellulose will first require efficient infrastructure for feedstock production, harvesting, and transport. Companies are addressing these issues by developing facilities and partnerships that will provide reliable, economical supplies. In addition, one of the key technical aspects is an efficient system for fractionating and converting the materials into ethanol and other fuels.

Developing *Pichia* for xylitol and ethanol production has drawn heavily on the JGI sequence data. The sequence data has helped identify previously unknown enzymes that enable this organism to use the diverse sugar components of the plant material. This could lead to improved organisms for simultaneous breakdown of the polymers (saccharification) and fermentation of the sugars. In the longer run, the sequence will contribute to the development of strains with higher tolerance to ethanol and potential inhibitors.

Pichia is but one fungus in an expanding portfolio of targets sequenced by the JGI that can be directed to steps in the biofuel process, such as fermentation. Others include *Clostridium thermocellum,* capable of directly converting cellulosic substrate into ethanol, the white rot fungi *Phanerochaete chrysosporium,* and the oyster mushroom, *Pleurotus ostreatus*, which are capable of lignin deconstruction—a necessary step for making cellulose accessible to further enzymatic processes.

5

Image courtesy of University of Jaén

Algal Genomics Program

Algae are unicellular or simple multicellular photosynthetic eukaryotes living mostly in aquatic environments. Although phylogenetically diverse (green, brown, and red algae, diatoms, coccolithophores, dinoflagellates, euglenoids, etc.) they all fix CO_2 during photosynthesis, and in fact, ocean phytoplankton is responsible for nearly half of total global photosynthesis. The diversity and wide geographic range of algae—they thrive from tropical to polar climates and from coastal to open-ocean systems—have global effects on primary production and regulation of atmospheric CO_2. Carbon sequestration is directly linked to photosynthesis and achieves its highest level during phytoplankton blooms. At the same time, toxic algal blooms (often caused by human alterations of coastal systems) cause significant public health, economic, and ecological hazard worldwide.

Algae also have significant potential as a feedstock source for next-generation biofuels. These biofuels have advantages compared with land plant biofuel feedstocks, including higher rates of energy per acre than terrestrial crops, ability to use ocean or waste water instead of fresh water, and ability to use wastelands rather than croplands. The challenges in efficient biodiesel production lie in finding or developing stable and disease-resistant strains with fast growth and high lipid content.

The goals of the algal genome program include the characterization of their:

- ecological and evolutionary diversity
- role in carbon sequestration, biogeochemistry, climate change, and environmental health
- metabolism and metabolic regulation as potential targets for biofuel production.

The key instrument is to develop genomics, transcriptomics, proteomics, and analytical resources for user communities and integrate them with community generated data and tools for wide access and analysis.

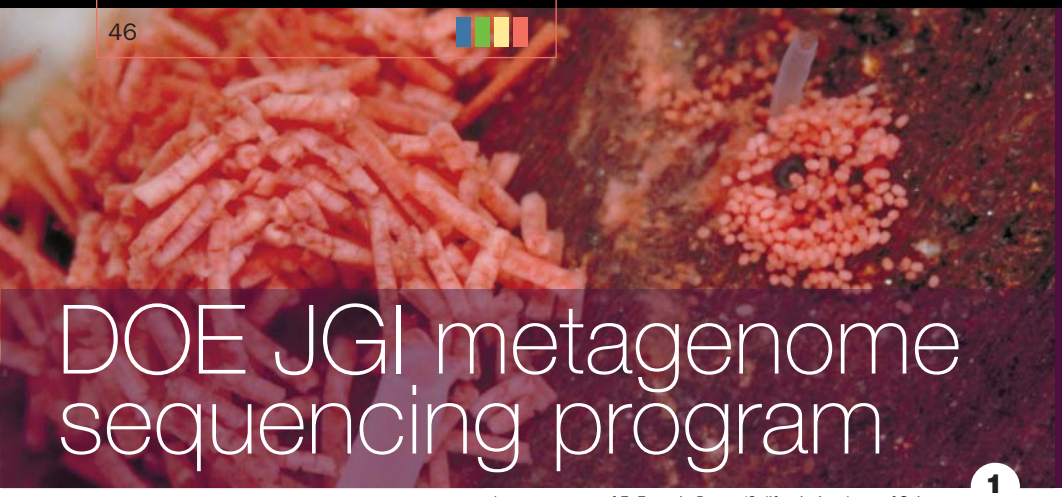
5. Oil-producing Green Microalga

Botryococcus braunii, selected for sequencing as a CSP 2009 Expressed Sequence Tag (EST) project, is a colony-forming green microalga of the species *Chlorophyceae*. It is found in many environments across the globe and has been noted to be capable of growing in both freshwater and brackish environments. During the growth cycle of this organism, the algae synthesize long-chain liquid hydrocarbon compounds and sequester them in the extracellular matrix of the colony to afford buoyancy. (Typical hydrocarbon content of the organism is approximately 30-40% of the dry weight of the cells.)

Three phenotypically distinct isolates, or "races," of *B. braunii* have been reported (races A, B, and L). These races are identified by the type of oil produced and accumulated by the organism. Of the three, the oils produced by race B, a family of isoprenoid compounds termed botryococcenes, hold the most promise as an alternative energy source. Botryococcenes have been converted to fuel suitable for internal combustion engines through caustic hydrolysis, and geochemical analysis has shown that botryococcenes, presumably from ancient *B. braunii* communities, also compose a portion of the hydrocarbon masses in several modern-day petroleum and coal deposits.

Algae are already recognized as a promising carbon sequestration agent. Moreover, the hydrocarbons produced by *B. braunii* hold significant promise as a renewable biofuel. Despite this, little information, either genetic or metabolic, has been reported for this organism. For instance, the identities of the exact genes in the metabolic pathway responsible for hydrocarbon synthesis have been elusive. This knowledge is sorely needed for any efforts to identify and alleviate bottlenecks in metabolic flux through this pathway and thus enable its use as a source of biofuels.

The principal investigators are Andrew Koppisch (LANL and Northern Arizona University), Hakim Boukhalfa, Christy Ruggiero, David T. Fox, and Nathan M. Beck (LANL); Suraj Dhungana (National Institute of Environmental Health Sciences), Joseph Chappell (University of Kentucky), Timothy P. Devarenne (Texas A&M University), Shigeru Okada (Tokyo University), C. Dale Poulter (University of Utah), and Rolf J. Mehlhorn, Jill O. Fuss, and Anastasios Melis (LBNL).



DOE JGI metagenome sequencing program

Image courtesy of E. Eugenia Patten/California Academy of Sciences. 1

What is a Metagenome?

Metagenomes are complex microbial communities that are isolated directly from the environment or reside inside of a larger organism. The study of metagenomes is an increasingly popular approach for discerning specific metabolic capabilities of entire microbial communities.

Metagenome sequencing is a relative newcomer to the DOE JGI science portfolio. This diverse array of projects targets multiple scientific goals, but what the studies share is the sequencing of DNA from a community of organisms rather than from a single isolate. This type of project offers a set of unique opportunities and challenges for the DOE JGI scientific effort.

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 from the community as a whole, and studies of such communities have revealed a diversity of microbes far beyond those found in culture collections. It is suspected that these uncultivated organisms must harbor considerable as-yet undiscovered genomic, functional, and metabolic features and capabilities. Thus to fully explore microbial genomics, it is imperative that we access the genomes of these elusive players.

Several early successes for the DOE JGI metagenome research program sparked a surge in interest in metagenomics research both at the DOE JGI and in the research community as a whole. In 2007, a National Research Council report entitled "The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet" proposed a global metagenomics initiative on the scale of the Human Genome Project. The first metagenomics proposals to the DOE JGI user programs came in 2005, and have continuously increased since then; in the CSP 2008 review metagenomics projects were formally separated from the microbial genome projects and evaluated independently.

Several of the accepted proposals have culminated in high-profile publications, most notably, Enhanced Biological Phosphate Removal (EBPR) sludge, published in *Nature Biotechnology* in October 2006; the termite hindgut community, published in *Nature* in November 2007; the Lake Washington methylotroph community, published in *Nature Biotechnology* in August 2008; and the South African deep gold mine metagenomic analysis revealing a single organism ecosystem deep within the Earth, published in *Science* in October 2008. These published projects showcase the application of metagenomics to the DOE mission areas: bioenergy, carbon cycling, and biogeochemistry. With subsequent calls for proposals we have greatly expanded the metagenome portfolio in these mission areas.

1. Pacific Shipworm

Shipworms, also known as "termites of the sea," are worm-like saltwater clams that feed primarily on submerged wood. Like termites, shipworms depend on symbiotic bacteria in their digestive tract for enzymes, which allow them to digest wood and are of potential interest for the commercial production of ethanol from plant biomass. As part of the DOE's mission to replace fossil fuels with renewable sources for cleaner energy (such as ethanol) sequencing the symbiont metagenome of the giant Pacific shipworm *(Bankia setacea)*, approved as a CSP 2009 project, will provide a valuable source of information for converting biomass into ethanol.

While the mechanisms shipworms use to digest wood remain largely unknown, anatomical considerations indicate that they differ from those observed in terrestrial cellulose consumers such as termites. The digestive systems of most terrestrial cellulose consumers contain dense and diverse populations of symbiotic microbes thought to be involved in cellulose metabolism and nitrogen fixation. Shipworms lack such highly developed and conspicuous microbial populations in their digestive systems. Instead they harbor dense populations of intracellular endosymbiotic bacteria in specialized cells (bacteriocytes) within a specialized organ (the Gland of Deshayes) in their gills.

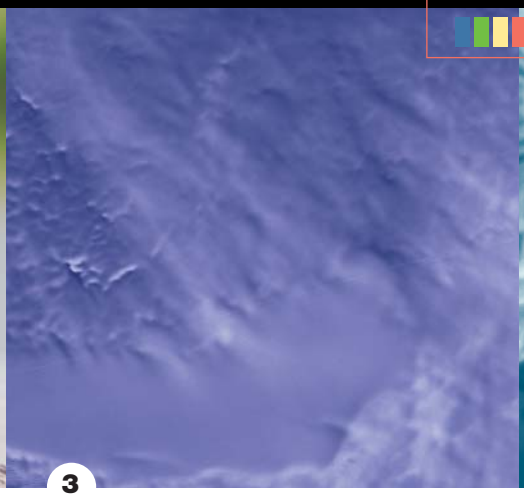
The shipworm symbiont community is far simpler and is phylogenetically distinct from those found in termites and ruminants. The mechanisms of lignocellulose degradation that shipworms employ are unique, placing shipworms and their symbionts among the most promising potential sources of novel enzymes for lignocellulose degradation and ethanol production. Analysis of the shipworm symbiont community metagenome will provide important insights into the composition and function of this unique lignocellulose-degrading bacterial community and will allow valuable comparisons to the recently sequenced termite symbiont metagenome. Unlike termites, shipworms accomplish the complete degradation of lignocellulose with a simple intracellular consortium of just a few related types of microbes.

The principal investigator is Daniel L. Distel (Ocean Genome Legacy).

2. Amazonian Stinkbird

Another CSP 2009 project will sample the foregut of *Opisthocomus hoazin*—a leaf-eating Amazonian pheasant-like stinkbird, or hoatzin. A prehistoric relic, its unique fermentative organ harbors an impressive array of novel microbes, like that of cows and other ruminants. Instead of a rumen, stinkbirds possess a crop, an enlargement of the esophagus where the fermentation takes place—and is the source of the stink. The characterization of its contents will likely

2



3

4

Image courtesy of Keith Ryan, Marine Biological Association & Willie Wilson, Plymouth Marine Laboratory

lead to the identification of novel microbial enzymes that degrade plant cell walls.

The principal investigator is Maria Dominguez-Bello of the University of Puerto Rico.

3. Lake Vostok

Microbes that eke out their existence in a lightless, frigid, buried lake in Antarctica may be the most energy-efficient organisms on earth. To try to unlock the secrets of their remarkable abilities, the DOE JGI will partner with researchers at the University of Alberta and the University of Delaware in a 2009 CSP project to sequence the genomes of this microbial community.

The microbes survive in the hostile environment of Lake Vostok, a freshwater lake about the size of Lake Ontario which has lain entombed under more than three kilometers of ice in the middle of Antarctica for perhaps 15 million years. The lake's near-freezing waters contain no light, no organic matter for food, and no geothermal activity or other energy sources. In addition, the lake's crushing pressure of nearly 400 atmospheres causes oxygen to become trapped in the waters. Lake Vostok is saturated with 50 times normal oxygen concentrations, which would be toxic to most organisms.

Although Lake Vostok itself has never been penetrated, the Vostok microbes were discovered in ice cores taken from just above the lake in 1998. After the microbes are extracted from the ice cores, JGI will sequence and analyze them. Since excess carbon dioxide in the atmosphere contributes to global warming, efficient carbon sequestration methods are of great interest to the DOE. Analysis of the bugs' genomes is expected to reveal novel enzymes and pathways for carbon sequestration, as well as other unusual adaptations to extreme cold, high pressure, and high oxygen. Because these conditions resemble those found on some outer planets and moons—for example, Jupiter's moon Europa—the microbes' genomes could even provide clues to the nature of extraterrestrial life.

The principal investigators are Phil Hugenholtz and Victor Kunin (DOE JGI), Brian Lanoil (University of Alberta), and Craig Cary (University of Delaware).

4. Uncultivated Marine Viruses

Approved as a CSP 2009 project, JGI will sequence three uncultivated viruses (obtained by physical fractionation) from one of the largest biomes on the planet: the oligotrophic (low-nutrient) open ocean. Viruses are an integral part of the marine ecosystem. Every living thing in the ocean appears to be susceptible to disease and death caused by viral infections. Although viruses cannot replicate on their own, they outnumber all forms of cellular life in the oceans by roughly an order of magnitude. Marine viruses are not only numerous, but also extraordinarily diverse, both morphologically and genetically.

Viruses have a particularly significant influence on the cycling of carbon and nutrients in marine planktonic communities. They cause the termination of phytoplankton blooms and lyse a significant fraction of the daily marine bacterial production. Food web models predict that viruses increase respiration and production by bacteria at the expense of larger organisms. Because of the intimate interaction and exchange that occurs among viruses and their hosts, an understanding of any marine organism is incomplete without knowledge of the viruses with which it interacts. The viruses in a sense represent the "extended genotype" of cellular organisms.

The mining of the greatest source of genetic novelty on our

Shipworms and their symbionts are among the most promising potential sources of novel enzymes for lignocellulose degradation and ethanol production.

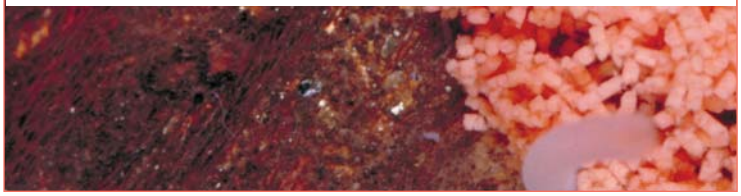
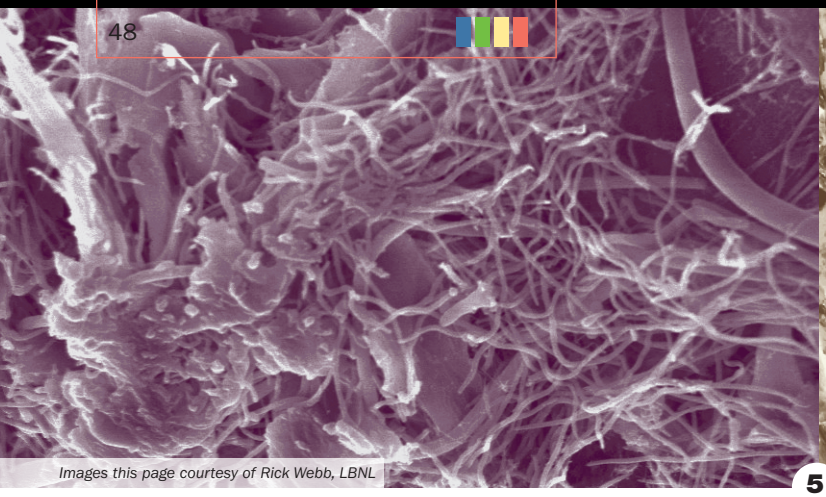
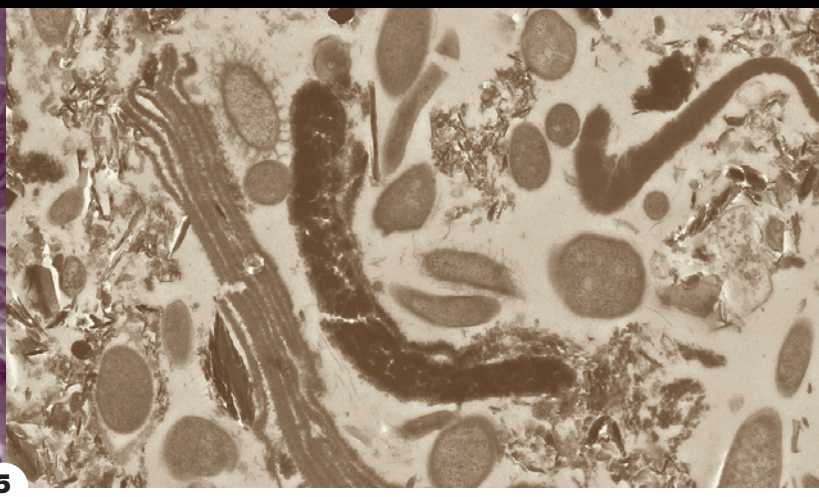


Image courtesy of E. Eugenia Patten at California Academy of Sciences.





Images this page courtesy of Rick Webb, LBNL

5

planet, the viruses, may well result in the discovery of new genes relevant to all three of DOE's missions. One potential outcome is identification of viral sequences useful as vectors for bioengineering of marine bacteria or archaea. Besides the genetic information, the data will lead to a better understanding of a class of organisms that directly affect plankton in the ocean in a number of different ways, including causing mass mortality, mediating genetic exchange, and effecting lysogenic conversion (survival of viruses in an infected bacterium without destroying it). These processes have direct relevance to efforts to exploit marine prokaryotic and eukaryotic plankton for carbon sequestration, alternative energy, or bioremediation.

The principal investigator is Grieg F. Steward (University of Hawaii).

5. Termite Hindgut Microbes

Termites are providing insights into the molecular machinery that can enable the efficient breakdown of lignocellulose into fermentable substrates for biofuels. Termites digest massive quantities of wood employing an array of specialized microbes in their hindguts to break down the cell walls of plant material and catalyze the digestion process. Their stomachs are a rich source of microbes, producing enzymes that can be employed to improve the conversion of wood or waste biomass to valuable biofuels.

Using industrial-scale DNA sequencing, the DOE JGI identified the genes employed by the termite gut microbes in the breakdown of lignocellulosic material. More than 500 genes related to the enzymatic deconstruction of cellulose and hemicellulose were identified. The genes encoding the novel enzymes discovered in this project are potentially useful for improving the conversion of wood or waste biomass to biofuels.

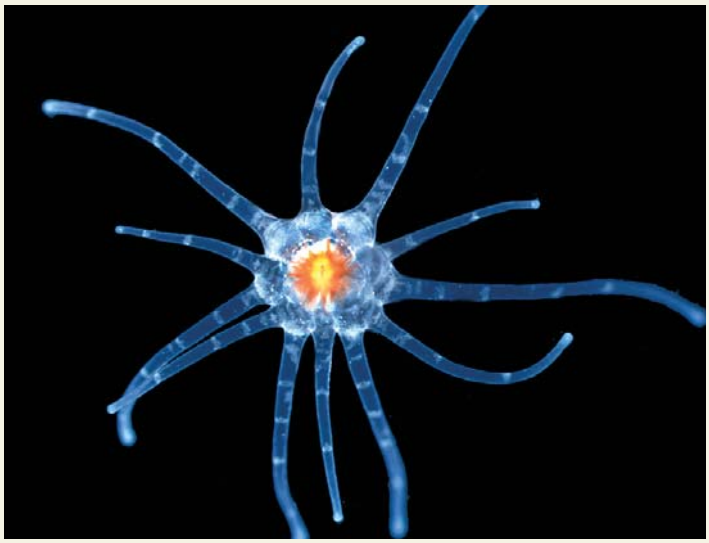
In addition to wood-feeding termites collected from Costa Rica, DOE JGI researchers have also collected grass-feeding termites from Arizona and Texas as part of the Energy Biosciences Institute's enzyme discovery initiative. The grass-feeding species were targeted because of EBI's interest in ultimately using switchgrass and *Miscanthus* as biofuel feedstock.

JGI scientists will catalog, characterize, and compare the enzyme inventories of grass-feeding termites to wood-feeders and to other systems such as the cow rumen to figure out which enzymes are specific to grassy diets. Then, with the help of EBI colleagues, they will characterize the activities of a selection of the most promising enzymes identified in the grass-feeder's hindguts.

Only a handful of model organisms involved in cellulose degradation have been investigated in detail for their enzymes, so the potential to discover novel cellulolytic enzymes is very high. While termites can efficiently convert milligrams of lignocellulose into fermentable sugars in their tiny bioreactor hindguts, scaling up this process so that biomass factories can produce biofuels more efficiently and economically is another story. To get there, the set of genes with key functional attributes for the breakdown of cellulose will need to be refined, and this study represents an essential step along that path.

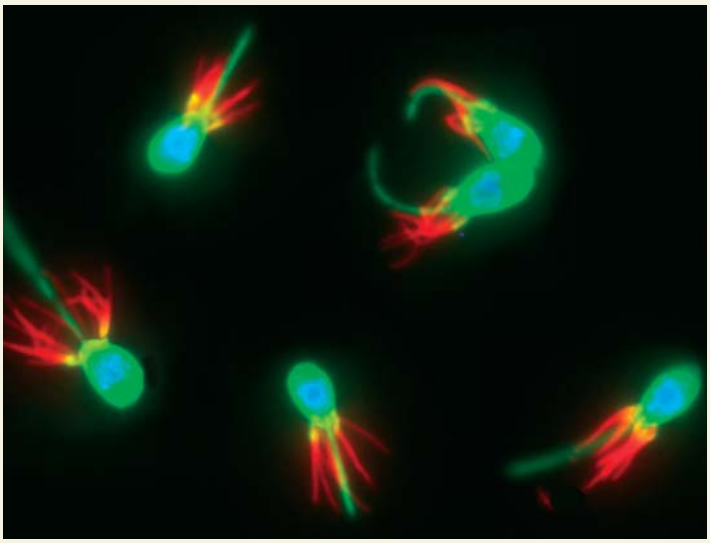
The termite hindgut project represents the first of several metagenomic projects at the JGI exploring the molecular machinery nature employs to digest lignocellulose. The genomic sequencing and analysis of the Costa Rican termite gut microbes by the JGI, the California Institute of Technology, Verenium Corporation (a biofuels company), INBio (the National Biodiversity Institute of Costa Rica), and the IBM Thomas J. Watson Research Center, was highlighted in *Nature* in November 2007. The termite gut metagenome dataset is publicly available, along with an annotated view, through the DOE JGI's metagenome data management and analysis system, IMG/M *(img. jgi.doe.gov/m).*

Photo courtesy of Nicholas Putnam, LBNL



A view into the mouth of the starlet sea anemone, *Nematostella vectensis.* The anemone, only a few inches long and endowed with between 16 and 20 tentacles, lives in the mud of brackish estuaries and marshes. It is becoming a popular laboratory subject for studies of development, evolution, genomics, reproductive biology, and ecology.

Photo courtesy of Monika Abedin, UC Berkeley



Choanoflagellates are aquatic microbes distinguished by a flagellum (green) used for swimming and feeding, surrounded by a collar of tentacles (red) against which bacterial prey are trapped. The nucleus of the one-celled organism is highlighted in blue.

finale for "legacy" genomes

In 2008 DOE JGI made considerable strides toward completing its legacy of animal sequencing projects. These projects were all begun prior to the decision in 2007 that the sequencing activity of the DOE JGI should focus exclusively on DOE mission science. While these activities did not directly contribute to DOE mission science, they provided a hugely important infrastructure for understanding animal evolution and diversity, as well as for plant and microbial eukaryotes.

Early DOE JGI efforts at sequencing individual mouse chromosomes and the Fugu genome (representing the first public whole-genome shotgun sequencing project) grew into a series of genome sequencing and analysis projects spanning diverse animal phyla. There was a special focus on deep chordate branches of the animal tree—tunicates and lancelet, the so-called "basal metazoan"—as well as other species of ecological and evolutionary interest (western clawed frog, water flea, owl limpet, a polychaete worm, leech, and spider mite). These projects entered the sequencing queue initially through direct DOE approval or special panels convened to define projects to fill gaps in phylogenetic coverage and the Community Sequencing Program (CSP).

Comparative analysis of these genomes has provided important insights into animal genome evolution. Chief among them is the recognition that most animal genomes share a relatively consistent complement of gene families, introns (regions between genes), and syntenic relationships (which refers to similar positions on the chromosome across species). This result was surprising given the remarkable divergence of previously sequenced model organisms, such as *Drosophila* and *C. elegans,* which now appear to have severely reduced genomes that masked underlying conservation, meaning these regions have been conserved among different species over time.

Strikingly, even the genome of the apparently simple placozoan *Trichoplax adhaerens,* a primitive animal with only four or five cell types, encodes a panoply of signaling genes and transcription factors usually associated with more complex animals. In a paper in the August 21, 2008 online edition of *Nature*, a team of scientists establishes this organism as a branching point of animal evolution and identifies a set of its genes, or a "parts list," that has evolved along particular branches of the tree of life.

The sequencing of another simple organism, the one-celled aquatic microbe called *Monosiga brevicollis,* a choanoflagellate, is helping scientists to discover the evolutionary changes that accompanied the jump from one-celled life forms to multicellular animals such as humans. In a paper in *Nature* on February 14, 2008 and a complementary paper in *Science* published the same week, the researchers found that the genome of choanoflagellates is surprisingly complex; the findings are helping to assemble a picture of what the original common ancestor of humans and choanoflagellates looked like at least 600 million years ago.

With the impending closure of animal genomic activities at the DOE JGI, efforts directed at complex genomes are now focused on plant and microbial sequences of more direct DOE relevance.

sequence analysis tools

Translating DNA Sequence Data into Useful Information

The sequencing and analysis of genomes is an inherently computational process and requires coordinated efforts of the informatics and computational biology groups at both the JGI and its partner laboratories.

- **Informatics** activities are directed towards three broad areas: computational infrastructure and system administration; the JGI web portal and systems in support of user project submission, review, and collaborations; and data processing and management in support of data submission, project tracking, production sequencing (including the Laboratory Information Management Systems), and quality control. These demands are handled by groups in the Production and Informatics Departments.

- **Computational Biology** activities include genome assembly from raw shotgun data; structural and functional annotation of genomes; scientific data analyses; and the development and support of genomic databases and web portals to enable widespread access to annotations and analyses of JGI data. JGI's computational biology efforts are distributed across scientific programs at both the JGI and its partner laboratories, with focused groups forming the nucleus of the prokaryotic, microbial eukaryotic, and plant programs.

The JGI has sequenced more than 400 prokaryotic isolates, 48 eukaryotic isolates, and a dozen metagenomes. The qualitative difference in size, structure, and organization—and thus complexity—of genomes and genes between prokaryotes and eukaryotes, and the distinct nature of metagenomic data relative to complete genomes, have led to the development of different approaches to assembly and annotation tailored to these three groups, which are described below. Groups responsible for these efforts are typically embedded in their associated scientific program, enabling each program to develop appropriate tools based on scientific needs. The JGI computational biology groups typically work closely with collaborators on analyses and manuscripts, and often are the principal drivers of a publication.

Integrated Microbial Genomes (IMG) portals

The Integrated Microbial Genomes (IMG) data management system *(img.jgi.doe.gov/)* aims to effectively integrate genomic data in order to facilitate and support a comparative analysis of the sequenced genomes and to generate their metabolic and cellular reconstructions.

The IMG system is developed through collaboration between the Genome Biology Program of the DOE-JGI and the Biological Data Management and Technology Center (BDMTC) of the Lawrence Berkeley National Laboratory (LBNL). The system is designed as a user-friendly tool for investigators wanting to extract information from microbial sequences. IMG responds to the increasing need for a means to handle the vast and growing spectrum of datasets emerging from genome projects taken on by the JGI and other public DNA sequencing centers. This tool enables scientists to tap the rich diversity of microbial environments and harness the possibilities that they hold for addressing challenges in environmental cleanup, agriculture, industrial processes, and alternative energy production.

The system is updated every three months with all the publicly available genome

sequence data (either complete or draft) from multiple biological data sources. IMG currently provides data integration of 4,207 genomes (comprised of 1,078 bacteria, 56 archaea, 40 eukaryotes, 803 plasmids, and 2,230 viruses), consisting of 4.7 million genes, with a variety of publicly available metabolic pathway collections, and protein family information. IMG has been cited in more than 80 publications and has been used in the analysis and publication of dozens of genomes. There are approximately 5,000 unique users of the system every month. IMG exists in several flavors:

IMG (*img.jgi.doe.gov/*) Having more than 20% of the worldwide microbial genome production capacity, the DOE JGI is the leading microbial genome sequencing center and therefore has a vested interest in the efficient analysis and interpretation of the produced data. IMG enables the efficient comparative analysis of all complete public genomes, draft or finished,

IMG/M (*img.jgi.doe.gov/m*) The Integrated Microbial Genomes with Microbiome Samples (IMG/M) data management system provides support for comparative analysis of metagenomic sequences generated with various sequencing technology platforms and data processing methods. IMG/M integrates data from diverse environmental microbial communities with microbial isolate genome data from the JGI's IMG system. This system allows the application of IMG's comparative analysis tools to metagenome data for examining the functional capabilities of microbial communities and isolated organisms of interest, and the analysis of strain-level heterogeneity within a species population in metagenome data.

IMG/M contains metagenome data generated from microbial community samples that have been the subject of recently published studies, including two biological phosphorus-removing sludge samples, two human distal gut samples, a gutless marine worm sample, obese and lean mouse gut samples, and a wood-feeding higher termite gut sample (*Nature* 2007). In addition, IMG/M includes three simulated metagenome data sets employed for benchmarking several assembly, gene prediction, and binning methods. The IMG/M system is updated twice a year.

IMG/ER and IMG/M-ER (*img.jgi.doe.gov/er* and *img.jgi.doe.gov/mer*) The IMG "Expert Review" (IMG/ER and IMG/M-ER) systems support individual scientists or groups of scientists for functional annotation and curation of their microbial genomes (and metagenomes) of interest. Often such genomes have not been deposited into the public genome sequence archives, and IMG annotation pipeline provides restricted access and automated analysis. Genomes undergoing curation in IMG/ER are integrated with all publicly available genomes in IMG and are accessible in the same framework for comparative analysis.

The ER systems were launched in November 2007, and there is a rapidly increasing demand for this type of service. The current version of IMG/ER contains 158 private microbial genome projects, and there are over 150 private-access user accounts issued so far. IMG/M-ER is enabling the analysis of 136 metagenomic samples.

IMG/EDU (*img.jgi.doe.gov/edu*) The Integrated Microbial Genomes–Education Site (IMG/EDU) system provides support for training and teaching microbial genome analysis and annotation using specific microbial genomes in the comparative context of all the genomes available in IMG. The current version of the IMG/EDU genome baseline consists of all isolate genomes in IMG 2.6. In addition to the IMG genomes, IMG/EDU contains the draft *Ammonifex degensii* genome and two of the recently sequenced GEBA genomes, *Halorhabdus utahensis* AX-2 and *Planctomyces limnophilus.*

User Training

In parallel with its research efforts, the microbial program has launched a user training program, the Microbial Genomes and Metagenomes (MGM) workshop (*www.jgi.doe.gov/meetings/mgm/index.html*). This program offers three five-day workshops on Microbial Genomics and Metagenomics per year. The goal is to provide strong and solid training in microbial genomic and metagenomic analysis and demonstrate how the cutting-edge science and technology of DOE JGI can enhance the participants' research. The program was launched in January 2008, and more than 150 people from over 20 countries participated in the first three workshops.

Eukaryotic Genome Portal

The Eukaryotic Genome Portal provides access to genome assembly and annotations and equips users with analysis tools for every eukaryote sequenced at the DOE JGI. In addition to the power of comparative genomics, input of biological experts is critical for genome annotation and analysis. The Portal serves as a platform for the JGI Community Annotation model, a scalable approach to integrate automated annotation at JGI with distributed community-wide efforts worldwide, which also includes tools for training, coordinating, and supporting user communities built around different genomes. This enables user communities to actively participate in genome annotation and analysis, to improve genome annotations and build stronger user communities. The Portal also acts as an interactive data repository and publishing tool for user communities with multiple tools to explore genome structure, gene families, and pathways in comparative fashion and to record user comments, references, new gene models, and annotations.

The Eukaryotic Genome Portal includes more than 50 eukaryotic genomes annotated at JGI and further improved by Community Annotation as well as genomes sequenced elsewhere and brought by communities to the DOE JGI for comparative analysis.

User Training

In parallel to the research work, a series of tutorials and jamborees have been hosted by the DOE JGI. Over 800 users have been trained to use the Eukaryotic Genome Portal, annotating more than 70,000 genes. On-line tutorials introduce these tools to an even broader scientific community. The Eukaryotic Genome Portal is responsible for the largest number of external users of the JGI Web site, attracting over 12,000 unique visitors a month.

Phytozome, Metazome, and VISTA

Comparative analysis of plant genomes is facilitated by Phytozome (*www.phytozome.net*), a joint project of the JGI and LBNL's Center for Integrative Genomics to facilitate comparative genomic studies amongst green plants. Phytozome, and its sister portal Metazome for metazoans, is built on open-source components (GMOD, Biomart, Galaxy, etc.) and an internal database that is made available through the Phytozome Web interface and through Biomart. The Phytozome genome browsers integrate JGI-LBNL's VISTA comparative toolkit to provide users access to conserved DNA sequences between groups of plant genomes.

Research interests in the Phytozome group include improved gene structure annotation via comparative methods; accurate identification of groups of orthologous genes through sequence similarity, evolutionary distance metrics, and synteny; analysis of the dynamics of plant genomes including the rich history of genome duplications and rearrangements; molecular evolution of gene families; and the characterization of sequence variation in and across populations.

The VISTA Suite of comparative genomics tools is a widely used resource, originally developed for mammalian genome comparisons. The VISTA Web site (*genome.lbl.gov/vista*) provides several tools for investigating and visualizing cross-species and multi-species conservation. Users can submit sequences of interest to several VISTA servers for various types of comparative analysis. VISTA is based on sequence alignment algorithms both for long genomic sequences (AVID, LAGAN, and Shuffle-LAGAN) and whole-genome assemblies (combined local/global alignment approach based on the LAGAN algorithms). The algorithmic base and the analysis and visualization modules have been adapted for comparative analysis of plant, fungi, and other JGI genomes. VISTA plug-ins are used in Phytozome, and relevant conservation tracks are shown in the Eukaryotic Genome Portal when phylogenetic divergences are appropriate. The value of the VISTA suite is reflected in the several hundreds of papers where VISTA is utilized to obtain and display comparative sequence data (see the selected list at *genome.lbl.gov/vista/opublications.shtml*), including several papers describing the analysis of individual genomic intervals in plants.

new sequencing platforms

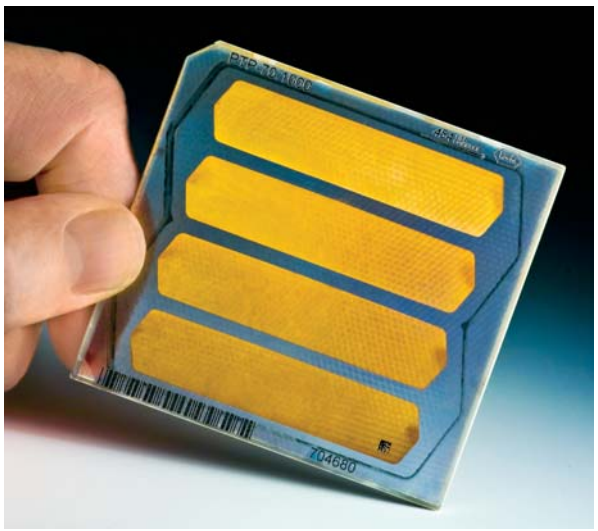
Transitioning to New Sequencing Technologies

Producing high-quality sequence as inexpensively as possible is vital to the continued success of the DOE JGI. JGI has implemented two new sequencing technologies within the last three years that have led to a dramatic increase in DNA sequencing capabilities in a very short period. In 2006, the JGI was running a mix of 35 MegaBACE 4500s and 70 ABI 3730xls sequencers, but in 2007, the JGI scaled back the Sanger line by shutting down the MegaBACE line. This enabled the JGI to free up resources to bring the Roche 454 Genome Sequencer from Technology Development into production. In 2008, the JGI scaled back to 57 ABI 3730xls on the Sanger line, focused on optimization and expansion of the 454 line, and introduced the Illumina Genome Analyzer into production.

ABI 3730xls Sequence Analyzer The Sanger sequencing production line performs standard Sanger sequencing reactions to produce high-quality draft sequence. The platform focuses on de novo eukaryotic sequencing, EST (expressed sequence tag) projects for gene discovery where no reference is present, and metagenomes. Clone libraries of 3kb, 8kb, and 40kb are generated by the Cloning Technology group and handed off as transformation stocks to the Sanger team. Transformation stocks are plated and picked using automated colony pickers, and a template is generated using Templiphi. The template is labeled using Big Dye chemistry and prepared for loading on ABI 3730 sequencers. During 2008, 20.3 gigabases of high-quality sequence were generated on the ABI 3730xl sequencers.

Roche 454 Genome Sequencer The DOE JGI operates eight Roche 454 instruments, which include a combination of FLX and Titanium. The FLX version of this platform is capable of generating 100 Mb of sequence in ~250-base-pair reads in a period of seven to eight hours. The Titanium version of this platform was beta-tested by the JGI beginning in April 2008 and is capable of delivering up to 400 Mb of sequence in 400-bp reads in the same period. The Titanium platform was released commercially in October 2008. The Roche 454 platform has become the primary de novo sequencer for microbial and EST sequencing projects, and will be extended to larger eukaryotic genomes this year. During 2008, 41.2 gigabases of high-quality sequence were generated on the 454 FLX.

ABI 3730, Sanger sequencing



Roche 454

Illumina GAII sequencers The JGI currently has seven Illumina GAII sequencers in operation. The Illumina was transitioned into production in May 2008. The Illumina platform produces ~4 gigabases of sequence per run, but does so with very short reads. As a result, the Illumina platform has become the mainstay sequencing platform for resequencing of organisms for which a reference genome already exists, as well as for polymorphism detection and as a polishing tool used for finishing genomes. During 2008, 64 gigabases of high-quality sequence were generated on the GA I sequencer. The JGI opened a new technology sequencing lab in its production building in July 2008. It has since fully integrated both new sequencing technologies into its pipeline and is currently scaling the throughput on both platforms. The new technology lab rooms were designed with the technicians in mind and feature advanced ergonomic workstations, equipment, and tools that allow the technicians to safely and efficiently prepare and sequence samples.

FY 2009 Production Goals

The FY 2009 sequencing target is 253 gigabases, which is a six-fold increase from FY 2008. There are a number of ongoing process improvement projects (e.g., automation, quality improvement) and staffing plans in place that will allow for scaling the 454 and Illumina sequencing platforms.

To ensure a smooth transition toward incorporating the new sequencing technologies, the Process Optimization team has been working with other production support groups (QC/QA, Instrumentation, Ergonomics, Informatics) in the past year to test out and implement new cloning and sequencing processes into the production pipelines. Their responsibilities include identifying areas of high ergonomic risks, putting in place designs to automate those steps before the production implementation, optimizing process flow for a scale-up operation, and improving steps to increase consistency of data quality.

Illumina

education and outreach

Training the Next Generation of Scientists

Faculty at undergraduate institutions struggle to provide meaningful research experiences to all students. Genomics and bioinformatics provide a scalable approach to undergraduate research. However, the ability of many faculty to mentor research in genomics is limited; those at primarily undergraduate teaching institutions without strong research emphases find it difficult to keep up with the pace of advances in genomics and bioinformatics.

Undergraduate Microbial Genome Annotation Program

The JGI Education Program, in collaboration with the Integrated Microbial Genomes (IMG) system team and the Informatics Department, has developed a set of tools and faculty training modules to integrate microbial genome annotation across the undergraduate life sciences curriculum, from introductory molecular biology to capstone microbiology and biochemistry courses. The annotation platform, IMG/ACT, links to IMG/EDU, a new version of IMG with special features, and to other databases used in microbial genome annotation.

The expansions of IMG/EDU include a six-frame visualization tool, to introduce students to fundamental principles of genome biology, and an upgraded IMG Chromosome Viewer with a GC% Heat Map to help students visualize potential examples of horizontal gene transfer in an organism.

IMG/ACT is a wiki/web portal fusion that guides students through the annotation process and provides a place to document their findings. This provides students an opportunity to work with "real-life" genome datasets. This will enable colleges and universities to "adopt a genome" for manual curation and student research.

Students at 13 colleges and universities are now part of the pilot phase of the program, taking place in the 2008-2009 academic year; about 340 students nationwide, from freshmen to seniors, are currently involved in annotating two of the GEBA (Genomic Encyclopedia of Bacteria and Archaea) genomes. In 2009-2010, the program will add 20 additional schools. The long-term goal is to expand the program nationally, with students participating in the annotation of GEBA genomes while they learn the fundamentals of bioinformatics and genomics. The Undergraduate Microbial Genome Annotation Program is currently being evaluated as a learning tool by the Oak Ridge Institute for Science Education (ORISE).

The JGI Education Program is also currently developing the metagenomics counterpart to IMG/ACT that will be linked to IMG/M.

DOE JGI Pilot Structural Genomics Program

The DOE JGI has established a collaboration with the Midwest Center for Structural Genomics at Argonne National Laboratory, a component of the NIH-funded Protein Structure Initiative, to test the potential of a wider application of high-throughput structural biology to produce proteins and deduce function and evolutionary relationships from structure. Along with members of the JGI User Community, JGI scientists are being invited to nominate targets from genome and metagenome sequences for protein expression, purification, crystallization, and structure determination. Moreover, the expression clones for each of the candidate genes are made available to the investigators for further studies. In addition to targets nominated by the JGI User Community, other targets include members of 382 novel protein families discovered in the first 56 GEBA genomes and the termite hindgut metagenome.

The structural descriptions will not only provide opportunities for researchers to discover the biochemical function of proteins and their distant evolutionary relationships but also, in the case of enzymes, be used to predict and model substrate interactions. More generally, knowledge of an increasingly complete repertoire of protein structures will improve the understanding of the structural basis of protein function structure, inform structure prediction methods, lead to new design strategies for synthetic biology, and ultimately lend insight into molecular interactions and pathways.

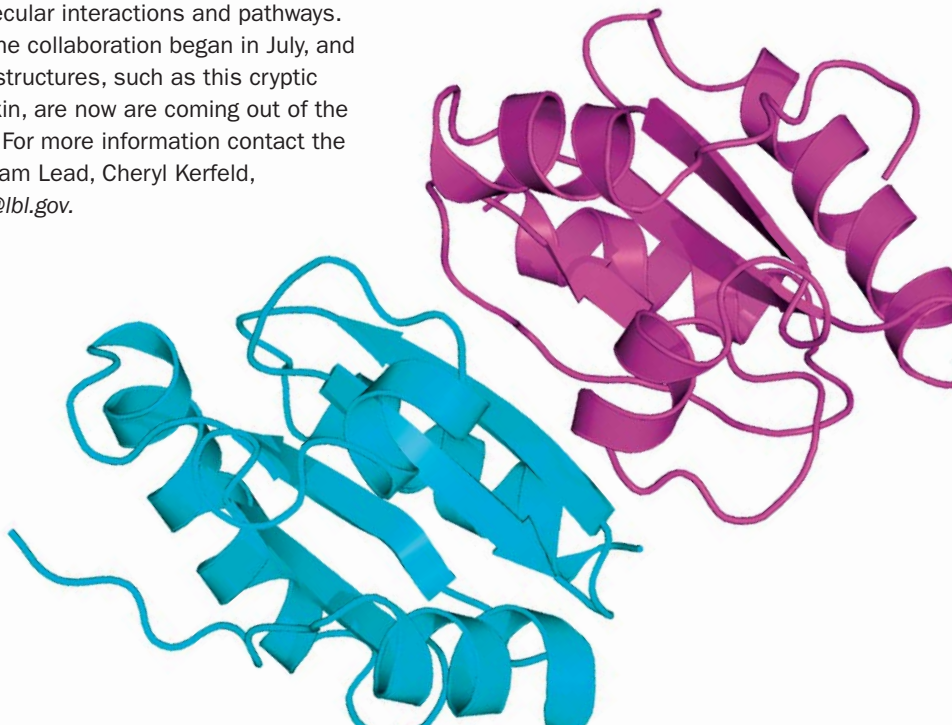
The collaboration began in July, and the first structures, such as this cryptic thioredoxin, are now are coming out of the pipeline. For more information contact the JGI Program Lead, Cheryl Kerfeld, *ckerfeld@lbl.gov.*

Sequencing Training Program

Among the goals of the DOE JGI is to implement programs to inspire the next generation of genomics researchers. By availing high-throughput DNA sequencing capacity to educators through the JGI Sequencing Training Program (STP), we expect that the tools of genomics will become relevant to classroom coursework and laboratory training.

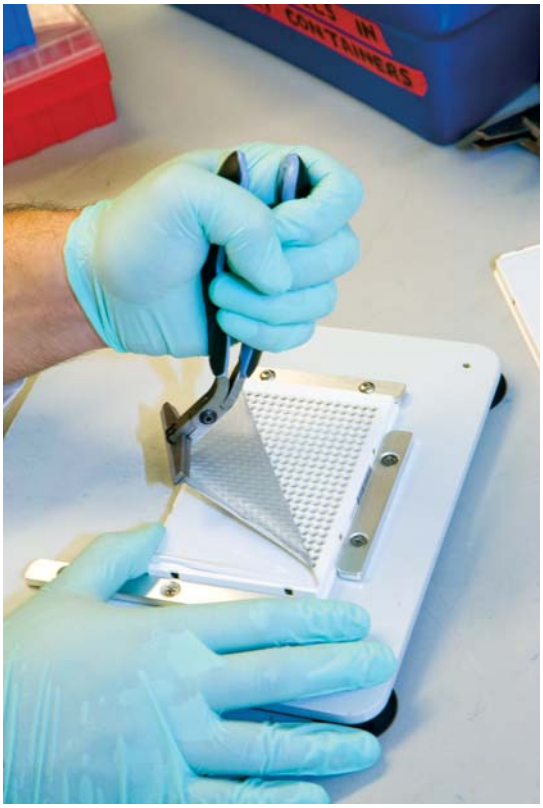
Pending available resources, and at the discretion of the DOE JGI Director, educational institutions can apply for JGI DNA sequencing capacity. A simple registration form is required, stating how the proposed project will advance classroom learning and how the data is to be used. Once approved, educational sponsors send samples to the JGI in a prescribed format designed to easily incorporate into JGI's Sanger production DNA sequencing line.

The projects include standard or kit-based samples, such as GAPDH (Glyceraldehyde 3-phosphate dehydrogenases), a family of enzymes essential for glycolysis—one of the most fundamental metabolic processes of life. They are found in all organisms, and because their enzyme function is so vital to life, their protein sequences are highly conserved both within the family and between organisms. Students can isolate DNA and amplify a highly conserved gene from a plant that can be sequenced.



This cryptic thioredoxin is one of the first structures resulting from the partnership with the Midwest Center for Structural Genomics at Argonne National Laboratory.

DOE JGI Emergency Response Team (ERT) with the Contra Costa County Fire Department



It's A-peeling: before *(left)* and after *(right)*

safety and ergonomics

Safety is a core value of the DOE JGI. With a highly repetitive task-based production environment and informatics teams comprising most of the work force at the JGI, developing a comprehensive and effective safety program for laboratory and office ergonomics has become a central focus for the organization.

At the DOE JGI, Integrated Safety Management (ISM) is the method used to ensure that all operations are planned and executed to protect the health and safety of each employee, the public, and the environment. The JGI Integrated Safety Management Plan is the document that describes in detail how the safety program is implemented at the JGI. The LBNL Job Hazards Analysis (JHA) process serves as the primary method of implementing ISM at the JGI. The JHA serves as an agreement between an employee and his or her supervisor on how they will conduct work safely at the JGI. Alternate and equivalent forms of a JHA are used for authorizing the work of subcontractors at the JGI.

Over the past three years there has been a significant focus on improving the ergonomic work environment throughout the JGI. The JGI staff ergonomist works hand-in-hand with the newly formed Ergonomics Working Group, comprised of volunteers from each of the JGI departments, and the entire JGI staff to develop and implement ergonomic solutions throughout the organization. The JGI Team that comprises the Ergonomics Working Group and JGI Safety Staff has implemented approximately 140 ergonomic solutions at the JGI with another 40 currently in some stage of implementation.

In March 2007, the team presented their "Shake-N-Plate" solution at the 10th Annual Applied Ergonomics Conference in Dallas, Texas, and won the prestigious 2007 Ergo Cup. Shake-N-Plate serves as a solution to the ergonomically challenging process of plating bacteria. It was developed through a collaborative effort between JGI engineers and the production staff and is now incorporated into the production process. In 2008, continuing to leverage the teamwork approach to solving ergonomic challenges, the JGI entered two projects in the 11th annual Ergo Cup competition: "It's A-peeling," which is an automated instrument used to remove seals from PCR (polymerase chain reaction) plates, and "Base Off," a tool designed to remove the lids from ABI 3730xl trays. Both projects were accepted for competition and received honorable mentions.

In 2008 two new Ergonomic Program solutions were implemented throughout the JGI. Remedy Interactive, an online interactive ergonomic risk assessment program, was introduced between May and August of 2008. The Remedy Interactive software automatically encourages employees to lower their ergonomic risk by listing self-led changes to ergonomic work station and work habits. In addition to these employee-led changes, those employees identified as high risk are approached by the staff ergonomist, who will work with the employee to ensure that the work station is optimized and that all possible factors that increase risk are addressed or improved.

In addition, in August 2008 RSI Guard or RAVE software was centrally installed on all Windows, Macs, and Unix computers at the JGI. This software tracks computer usage and reminds and instructs employees to take appropriate computer keyboard and stretch breaks.

Base Off: before *(left)* and after *(right)*

Appendix A:
DOE JGI Sequencing Processes

DNA: LIFE'S CODE

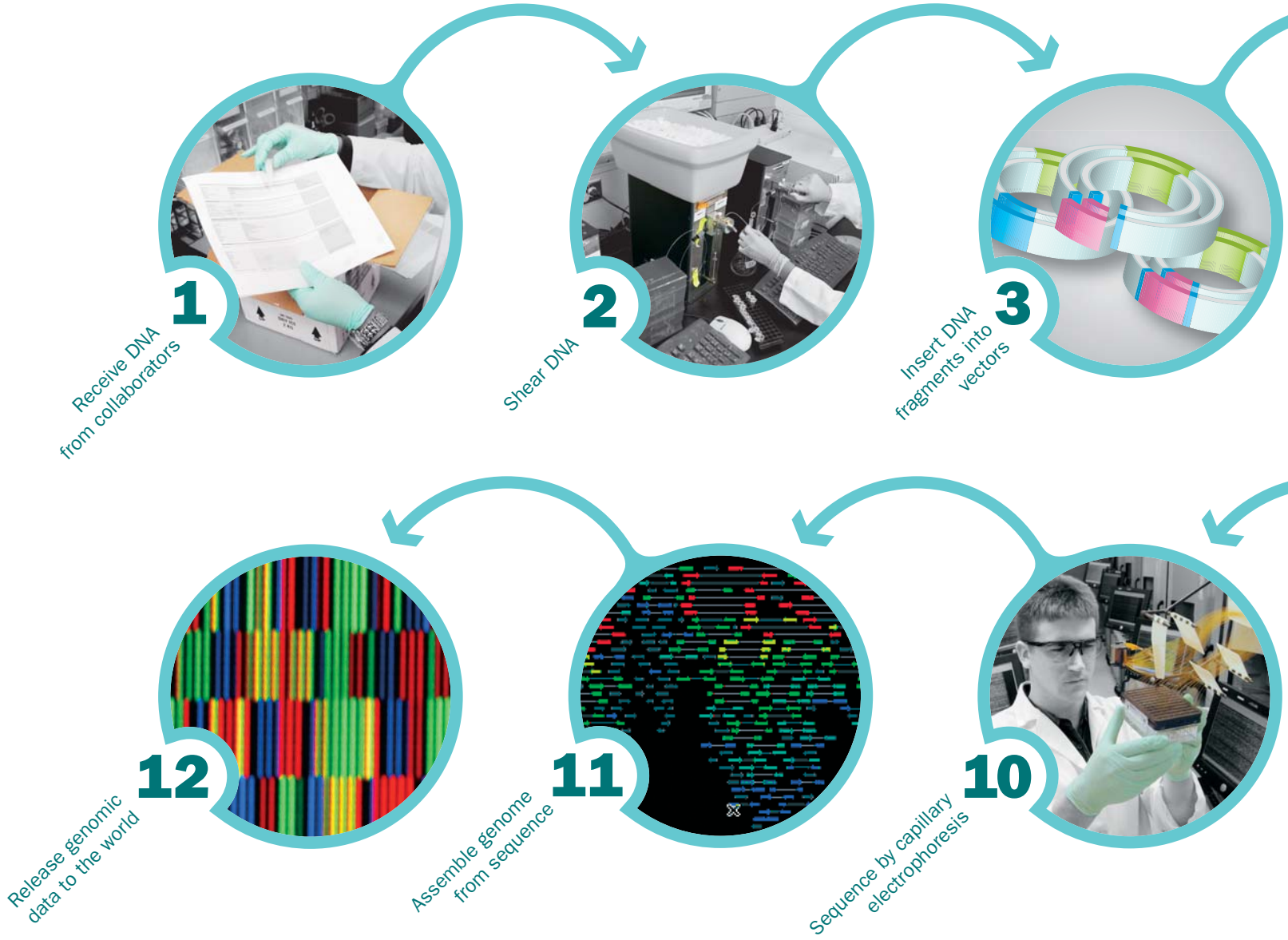
DNA (deoxyribonucleic acid), the information embedded in all living organisms, is a molecule made up of four chemical components—the nucleotides Adenine, Thymine, Cytosine, and Guanine—abbreviated A, T, C, G. These letters constitute the "rungs" of the double-helical ladder/backbone of the DNA molecule, with the As always binding with Ts, and Cs with Gs.

WHAT IS DNA SEQUENCING?

Just as computer software is rendered in long strings of 0s and 1s, the "software" of life is represented by a string of the four chemicals, A, T, C, and G. To understand the software of either a computer or a living organism, we must know the order, or sequence, of these informative bits.

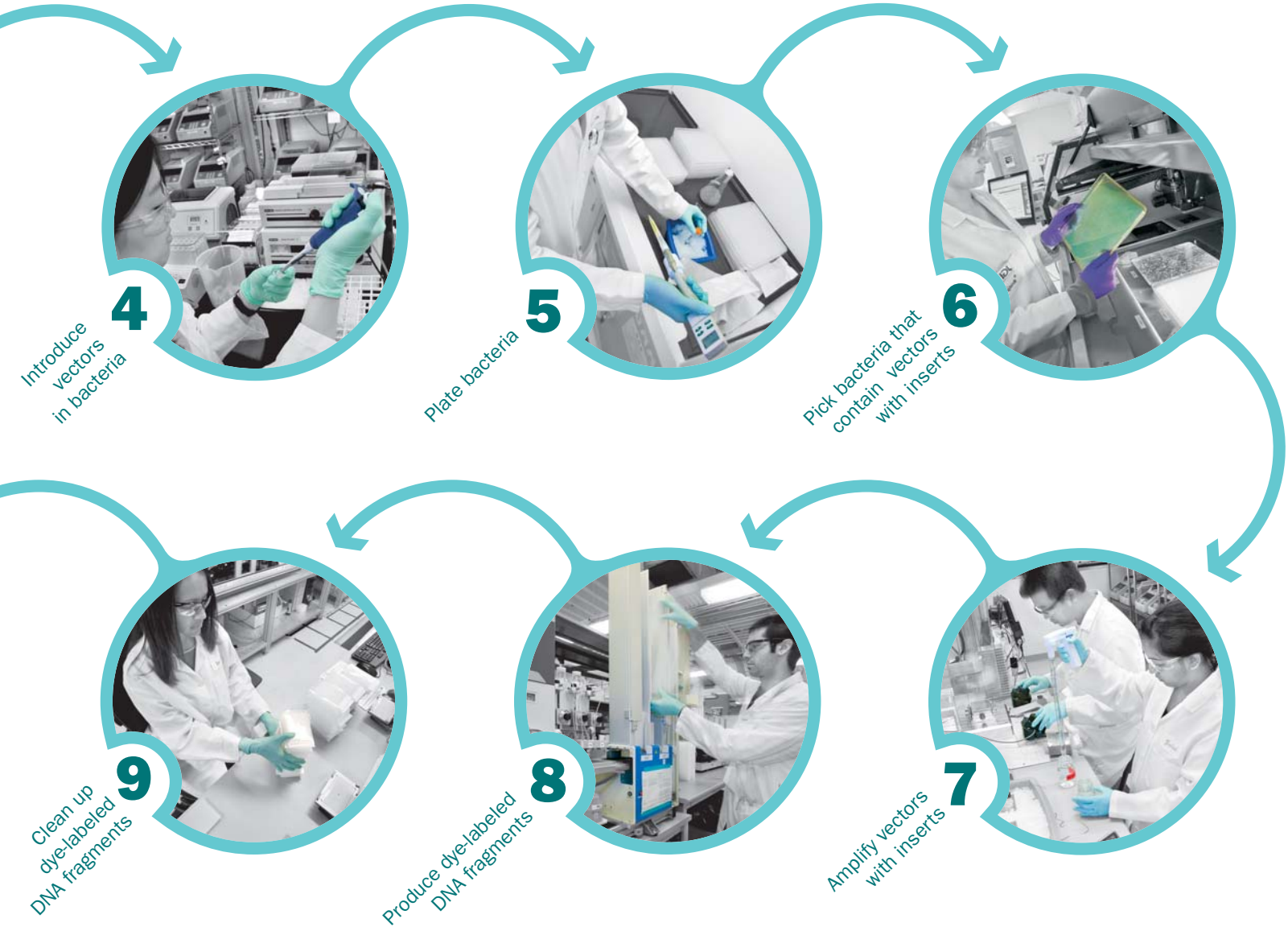
DOE JGI SANGER SEQUENCING PROCESS

Whole-genome shotgun sequencing is a technique for determining the precise order of the letters of DNA code of a genome. First, DNA received from JGI collaborators **(1)**, or users, is sheared into small fragments that are easier for sequencing machines to handle **(2)**. These fragments are biochemically inserted into a plasmid vector **(3)**—a loop of nonessential bacterial DNA—and mixed into a so-



lution of *E. coli* bacteria. An electric shock allows the plasmids to enter the bacterial cells **(4)**. The bacterial cells are moved to an agar plate **(5)** and incubated with a nutrient and antibiotic to suppress the growth of unwanted cells. Over a night of incubation, colonies form, and each contains about a million bacteria. The clones in the colonies that contain the inserted DNA fragments required for sequencing are distinguished by color, picked **(6)**, and incubated with nutrients again to make many more copies, which can then be sequenced. These DNA of interest are then duplicated, or amplified **(7)**, through a process initiated with another enzyme called polymerase and an abundant supply of the chemical building-blocks of DNA, or nucleotides, from which the polymerase can assemble new copies. After many heating and cooling cycles, the ends of the DNA fragments are labeled with fluorescent markers **(8)**. The DNA fragments are cleaned—isolated from the bacteria—using magnetic beads in an ethanol solution **(9)**. In a sequencing machine, an electrical charge is applied to the samples, pulling the DNA fragments through an assembly of fine glass tubes filled with a gel-like matrix, smaller fragments traveling faster than the larger, toward a laser detector system, which excites the fluorescent tags on each fragment by length and counts them to determine the sequence of As, Ts, Cs, and Gs **(10)**.

During the assembly process, the DNA fragments are realigned based on overlaps in their sequences **(11)**. Computer software uses the overlapping ends of different reads to assemble them into a reconstruction of the original contiguous sequence, then the annotated genome is made available to the scientific community **(12)**.



New Sequencing Technologies

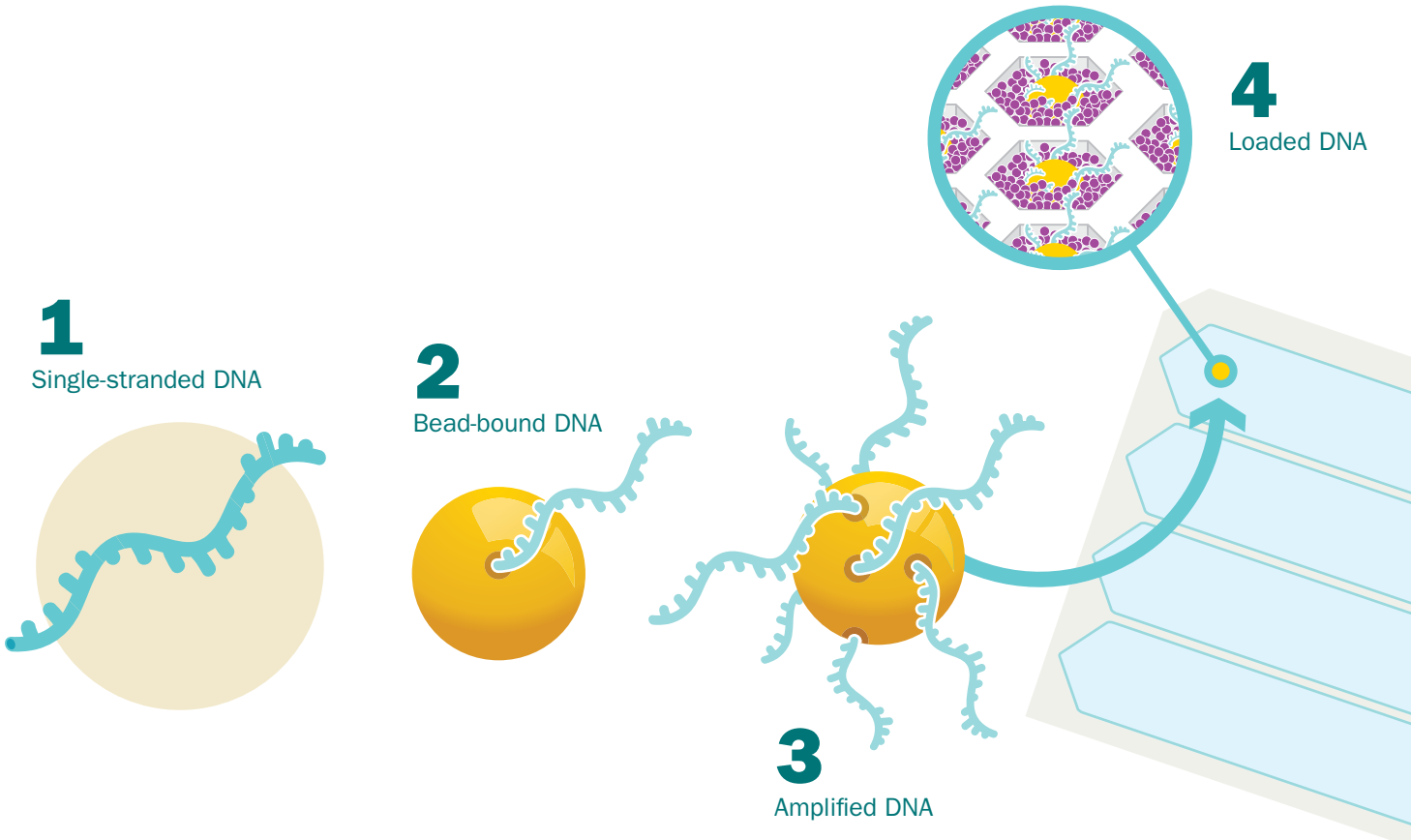
Two new technologies being incorporated into the production line at the DOE JGI promise an even faster and more efficient future for sequencing. One method uses the Roche 454 machine and involves a process called emulsion PCR (polymerase chain reaction) along with pyrosequencing. The other method uses the Illumina machine and involves bridge PCR along with reversible dye terminators. Both technologies eliminate the time-consuming steps of in vivo cloning, colony picking, and capillary electrophoresis. Both machines are able to produce approximately 1,000 times more bases per week than the previous generation of sequencers. Both technologies have the added benefit of eliminating bias against in vivo genomic regions.

The new sequencing-by-synthesis (SbS) approach builds a picture of a newly synthesized DNA fragment one base at a time. The addition of each base is detected in real time, eliminating the need for separating molecules according to size using capillary electrophoresis.

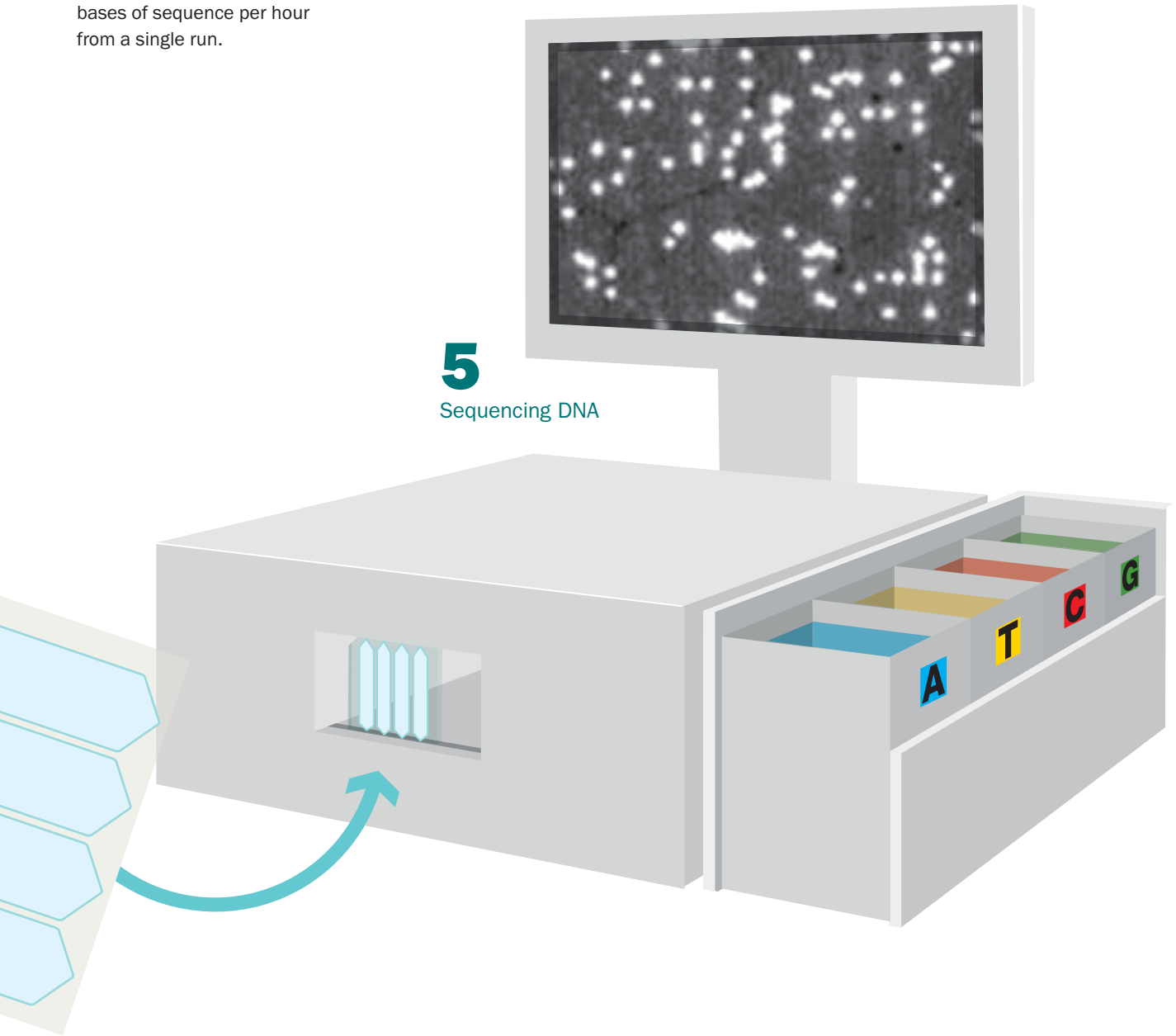
454 Sequencing Technology

Roche's 454 sequencing technology uses emulsion PCR DNA amplification combined with pyrosequencing, which enables one person to prepare and sequence an entire genome. The 454 instrument uses a parallel-processing approach to produce an average of 300-400 megabases (300,000,000 bases) of DNA sequence per 9-hour sequencing run. This means that a single 454 machine has the potential to sequence the equivalent of one human genome (3.1 billion bases) in one month.

1. A single-stranded DNA is attached to one capture bead.
2. Each bead carries millions of copies of a unique single-stranded DNA.
3. PCR amplification is performed on each bead in separated water-in-oil droplet (emulsion) so that there are ten million copies of a fragment on each bead, eliminating the need for cloning and robots.
4. The DNA-coated beads are then loaded into the 3.2 million hexagonal wells of a fiber-optic slide.



5. Solutions containing a single nucleotide type (A, T, C, or G) are consecutively applied over the wells in cycles. As each base (A, T, C, or G) is incorporated into a new DNA strand, a CCD camera records the light flashes generated by the reaction. The sequence of hundreds of thousands of individual reactions is determined simultaneously, producing millions of bases of sequence per hour from a single run.

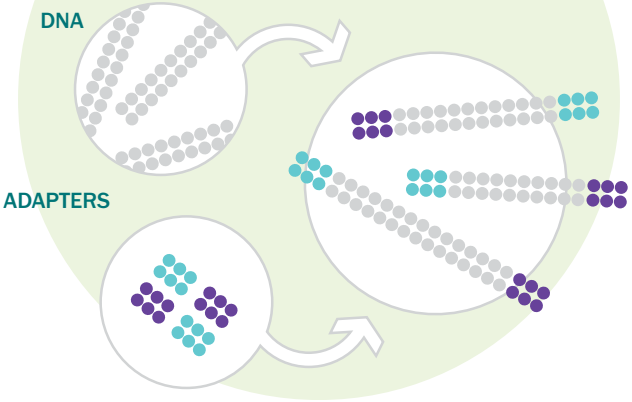


Illumina Sequencing Technology

This platform is based on parallel sequencing of millions of fragments using a proprietary Clonal Single Molecule Array technology—which amplifies the template by bridge PCR onto a glass slide—and a novel reversible terminator-based sequencing chemistry which allows detection of the sequence in real time during the sequencing-by-synthesis (SbS) process.

1

Prepare genomic DNA sample: Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.



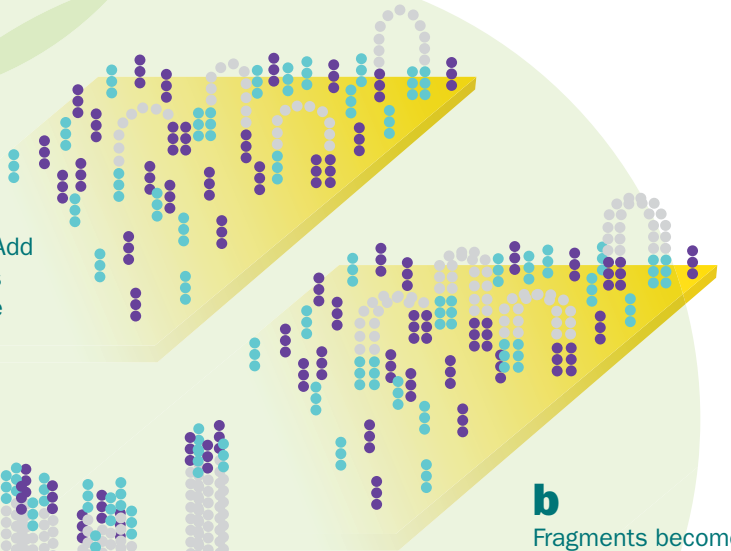
2 Attach DNA to surface: Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3

DNA Amplification

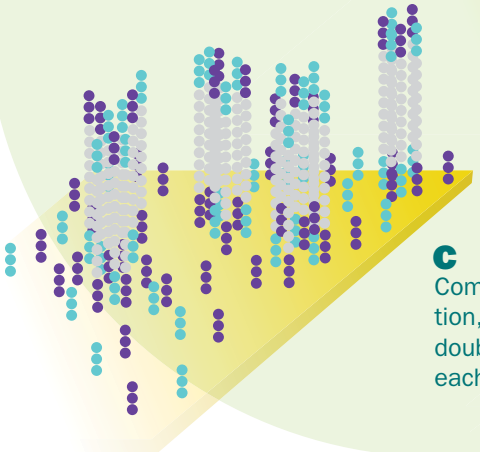
a

Bridge amplification: Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification



b

Fragments become double-stranded

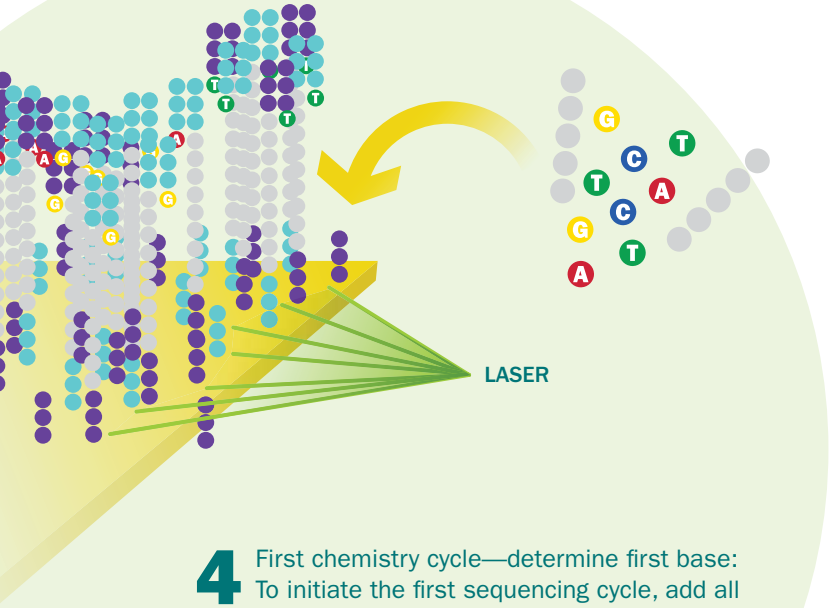


c

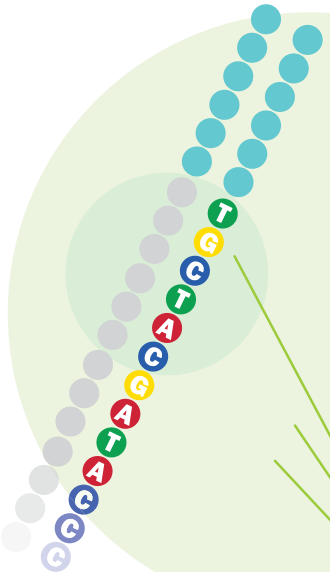
Completion of amplification: On completion, several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Comparison of Sequencing Processes (as of March 2009)

Process steps	Sanger Method	454 – Titanium	Illumina – GAII
Break up genome and isolate separate fragments	• Shear and ligate into vector • Transform bacteria • One fragment per cell	• Shear and ligate adaptors • One fragment per bead	• Shear and ligate adaptors • Hybridize to adaptor complimentary oligos on slide
Isolate DNA fragments	Spread onto bioassay plate to separate bacteria	One bead per emulsion droplet	Fragments are now at one fragment per μm^2
Amplify fragments	Grow colonies on bioassay then RCA	Emulsion PCR	Bridge amplification
Sequencing Chemistry	Cycle Sanger sequencing reaction using dye terminators	Pyrosequencing One type of base per cycle	SBS reversible terminator reaction Four colored bases per cycle
Detection of sequence	Capillary electrophoresis with fluorescence detection	Real-time detection of luminescence each cycle	Real-time detection of fluorescence each cycle
Read Length (max)	900 bases	500 - 600 bases	35 - 100 bases (depending on run type)
Assembly	Paired reads (3kb, 8kb, and 40kb) are assembled using Phrap program (easiest)	Single direction and paired reads available; Assembled using Newbler program	Single direction and paired reads available; various assembly programs
Bases per week per machine	8 million bases	2 billion bases (five runs)	8 billion bases (two runs at 35 cycles)
Advantages	• Accurate • Long read lengths • Large paired reads of 3 sizes	• Fast • Less expensive • Good for capturing uncloned regions	• Fast • Less expensive • Can distinguish bases in a homopolymer
Disadvantages	• Expensive • Slow	• Cannot accurately decipher repetitive regions • Cannot distinguish the number of nucleotides in a long homopolymer	• Cannot accurately decipher repetitive regions • Shorter reads



4 First chemistry cycle—determine first base: To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.



5 In the first cycle, add sequencing reagents, the first base is incorporated, remove unincorporated bases, detect signal. In subsequent cycles, add sequencing reagents and repeat.

Appendix B: Glossary

Annotation: The process of identifying the locations of genes in a genome and determining what those genes do.

Archaea: One of the three domains of life (eukaryotes and bacteria being the others) that subsume primitive microorganisms that can tolerate extreme (temperature, acid, etc.) environmental conditions.

Assembly: Compilation of overlapping DNA sequences obtained from an organism that have been clustered together based on their degree of sequence identity or similarity.

BAC (Bacterial Artificial Chromosome): An artificially created chromosome in which large segments of foreign DNA (up to 150,000 bp) are cloned into bacteria. Once the foreign DNA has been cloned into the bacteria's chromosome, many copies of it can be made and sequenced.

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 one another (A and T or G and C) that join the complementary strands of DNA to form the double helix characteristic of 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.

Electrophoresis: A process by which molecules (such as proteins, DNA, or RNA fragments) can be separated according to size and electrical charge by applying an electric current to them. Each kind of molecule travels through a matrix at a different rate, depending on its electrical charge and molecular size.

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.

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

Library: An unordered collection of clones containing DNA fragments from a particular organism or environment that together represents 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 relatively new field of genetic research allows the genomic study of organisms that are not easily cultured in a laboratory.

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.

Plasmid: Autonomously replicating, extrachromosomal, circular DNA molecules, distinct from the normal bacterial genome and nonessential for cell survival under nonselective conditions. Some plasmids are capable of integrating into the host genome. A number of artificially constructed plasmids are used as cloning vectors.

Polymerase: Enzyme that copies RNA or DNA. RNA polymerase uses preexisting nucleic acid templates and assembles the RNA from ribonucleotides. DNA polymerase uses preexisting nucleic acid templates and assembles the DNA from deoxyribonucleotides.

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.

RCA: Acronym for Rolling Circle Amplification, a randomly primed method of making multiple copies of DNA fragments, which employs a proprietary polymerase enzyme and does not require the DNA to be purified before being added to the sequencing reaction.

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.

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

Transformation: A process by which the genetic material carried by an individual cell is altered by the introduction of foreign DNA into the cell.

Vector: DNA molecule originating from a virus, a plasmid, or the cell of a higher organism into which another DNA fragment of appropriate size can be integrated without loss of the vector's capacity for self-replication; vectors introduce foreign DNA into host cells, where it can be reproduced in large quantities. Examples are plasmids, cosmids, Bacterial Artificial Chromosomes (BACs), or Yeast Artificial Chromosomes (YACs).

Whole-genome shotgun: Semi-automated technique for sequencing long DNA strands in which DNA is randomly fragmented and sequenced in pieces that are later reconstructed by a computer.

Appendix C:
CSP Sequencing Plans for 2009

Organism	Proposer	Affiliation
Eukaryotes		
Resequencing *Trichoderma reesei*	Scott Baker	Pacific Northwest National Laboratory
Rhizopogon salebrosus (ectomycorrhizal fungus)	Thomas Bruns	University of California, Berkeley
Ceriporiopsis subvermispora (lignin-degrading fungus)	Daniel Cullen	USDA Forest Products Laboratory
Gene expression in *Chlamydomonas reinhardtii*	Maria Ghirardi	National Renewable Energy Laboratory
Paralvinella sulfincola (polychaete worm)	Peter Girguis	Harvard University
Thalassiosira rotula (diatom)	Bethany Jenkins	University of Rhode Island
Dendroctonus frontalis (southern pine beetle) ESTs	Scott Kelley	San Diego State University
Botryococcus braunii (Oil-Producing Green Microalga) cDNA	Andrew Koppisch	Los Alamos National Laboratory
Chlamydomonas and *Volvox* transcriptomes	Sabeeha Merchant	University of California, Los Angeles
Spirodela polyrhiza (duckweed)	Todd Michael	Rutgers
Zostera marina (seagrass)	Jeanine Olsen	University of Groningen
Gossypium (cotton)	Andrew Paterson	University of Georgia
Pine BAC Sequencing	Daniel Peterson	Mississippi State University
Hansenula polymorpha strain NCYC 495 leu1.1 (ATCC MYA-335)	Andriy Sibirny	Institute of Cell Biology, Ukraine
Resequencing of *Brachypodium distachyon*	John Vogel	USDA-ARS Western Regional Research Center
Nanoflagellates: *Paraphysomonas, Ochromonas,* and *Spumella*	Alexandra Worden	Monterey Bay Aquarium Research Institute
Bacteria and Archaea		
Four diverse cellulose degrading microbes	Iain Anderson	DOE JGI
Escherichia coli MG1655	John Battista	Lousiana State University
Sinorhizobium meliloti strains AK83 and BL225C	Emanuele Biondi	Universita' di Firenze
Methylotenera species	Ludmila Chistoserdova	University of Washington
SAR11 Genome Evolution	Stephen Giovannoni	Oregon State University
Rhodopseudomonas palustris strain DX-1	Caroline Harwood	University of Washington
Cycloclasticus pugetii (a PAH-Degrading Bacterium)	Russell Herwig	University of Washington
Burkholderia sp. Ch1-1 and *Burkholderia* sp. Cs1-4	William Hickey	University of Wisconsin
Sphaerochaeta pleomorpha and *Sphaerochaeta globus*	Frank Loeffler	Georgia Institute of Technology
Archaeal transcriptomes	Todd Lowe	University of California, Santa Cruz
Dehalogenimonas lykanthroporepellens	William Moe	Louisiana State Univ.
Desulfurococcus (hyperthermophilic archaeon)	Biswarup Mukhopadhyay	Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University
Mesorhizobium ciceri bv biserrulae (strains WSM1271, WSM2073 and WSM2075) (legume symbionts)	Kemanthi Nandasena	Murdoch University
Thermoacidophiles of deep-sea hydrothermal vents	Anna-Louise Reysenbach	Portland State University
Alicycliphilus denitrificans strain BC	Alfons Stams	Wageningen University
Desulfotomaculum species	Alfons Stams	Wageningen University
Freshwater Actinobacteria belonging to the acI lineage	Falk Warnecke	DOE JGI

Organism	Proposer	Affiliation
Metagenomes		
Novel subsurface microbial phylotypes	Jennifer Biddle	University of North Carolina, Chapel Hill
Highly efficient, highly stable, reductive dechlorinating bioreactor	Eoin Brodie	Lawrence Berkeley National Laboratory
Bankia setacea (shipworm) metagenome.	Daniel Distel	Ocean Genome Legacy
Hoatzin crop microbiome	Maria Dominguez-Bello	University of Puerto Rico
Ammonia-oxidizing archaeal enrichment culture	Christopher Francis	Stanford University
Subarctic Pacific Ocean	Steven Hallam	University of British Columbia
Lake Vostok accretion ice	Philip Hugenholtz	DOE JGI
Microbial communities at the Hanford 300A IFC Site.	Allan Konopka	Pacific Northwest National Laboratory
PCE-dechlorinating mixed communities	Ruth Richardson	Cornell University
Uncultivated marine viruses	Grieg Steward	University of Hawaii
Great Salt Lake	Bart Weimer	Utah State University

Appendix D:
GEBA Sequencing Plans

ORGANISM	DOMAIN	PHYLUM
Acidimicrobium ferrooxidans DSM 10331	Bacteria	Actinobacteria
Actinosynnema mirum 101, DSM 43827	Bacteria	Actinobacteria
Alicyclobacillus acidocaldarius 104-IA, DSM 446	Bacteria	Firmicutes
Anaerococcus prevotii PC 1, DSM 20548	Bacteria	Firmicutes
Atopobium parvulum IPP 1246, DSM 20469	Bacteria	Actinobacteria
Beutenbergia cavernosa HKI 0122, DSM 12333	Bacteria	Actinobacteria
Brachybacterium faecium DSM 4810	Bacteria	Actinobacteria
Brachyspira murdochii DSM 12563	Bacteria	Spirochaetes
Capnocytophaga ochracea DSM 7271	Bacteria	Bacteroidetes
Catenulispora acidiphila ID139908, DSM 44928	Bacteria	Actinobacteria
Cellulomonas flavigena 134, DSM 20109	Bacteria	Actinobacteria
Chitinophaga pinensis UQM 2034, DSM 2588	Bacteria	Bacteroidetes
Conexibacter woesei ID131577, DSM 14684	Bacteria	Actinobacteria
Cryptobacterium curtum DSM 15641	Bacteria	Actinobacteria
Denitrovibrio acetiphilus N2460, DSM 12809	Bacteria	Deferribacteres
Desulfohalobium retbaense DSM 5692	Bacteria	Delta proteobacteria
Desulfomicrobium baculatum DSM 04028	Bacteria	Delta proteobacteria
Desulfotomaculum acetoxidans 5575, DSM 771	Bacteria	Firmicutes
Dethiosulfovibrio peptidovorans SEBR 4207, DSM 11002	Bacteria	Aminanaerobia
Dyadobacter fermentans NS 114, DSM 18053	Bacteria	Bacteroidetes
Eggerthella lenta VPI 0255, DSM 2243	Bacteria	Actinobacteria
Geodermatophilus obscurus DSM 43160	Bacteria	Actinobacteria
Gordonia bronchialis DSM 43247	Bacteria	Actinobacteria
Haliangium ochraceum SMP-2, DSM 14365	Bacteria	Delta proteobacteria
Halogeometricum borinquense DSM 11551	Archaea	Halobacteria
Halomicrobium mukohataei arg-2, DSM 12286	Archaea	Halobacteria
Halorhabdus utahensis AX-2, DSM 12940	Archaea	Halobacteria
Jonesia denitrificans DSM 20603	Bacteria	Actinobacteria
Kangiella koreensis SW-125, DSM 16069	Bacteria	Gamma proteobacteria
Kribbella flavida DSM 17836	Bacteria	Actinobacteria
Kytococcus sedentarius DSM 20547	Bacteria	Actinobacteria
Leptotrichia buccalis C-1013-b, DSM 1135	Bacteria	Fusobacteria
Meiothermus ruber DSM 1279	Bacteria	Deinococci
Meiothermus silvanus DSM 9946	Bacteria	Deinococci
Nakamurella multipartita DSM 44233	Bacteria	Actinobacteria
Nocardiopsis dassonvillei DSM 43111	Bacteria	Actinobacteria

ORGANISM	DOMAIN	PHYLUM
Pedobacter heparinus HIM 762-3, DSM 2366	Bacteria	Bacteroidetes
Planctomyces limnophilus DSM 3776	Bacteria	Bacteroidetes
Rhodothermus marinus DSM 4252	Bacteria	Bacteroidetes
Saccharomonospora viridis P101, DSM 43017	Bacteria	Actinobacteria
Sanguibacter keddieii DSM 10542	Bacteria	Actinobacteria
Sebaldella termitidis ATCC 33386	Bacteria	Fusobacteria
Slackia heliotrinireducens DSM 20476	Bacteria	Actinobacteria
Sphaerobacter thermophilus 4ac11, DSM 20745	Bacteria	Chloroflexi
Spirosoma linguale DSM 74	Bacteria	Bacteroidetes
Stackebrandtia nassauensis LLR-40K-21, DSM 44728	Bacteria	Actinobacteria
Streptobacillus moniliformis DSM 12112	Bacteria	Fusobacteria
Streptosporangium roseum NI 9100, DSM 43021	Bacteria	Actinobacteria
Sulfurospirillum deleyianum DSM 6946	Bacteria	Epsilon proteobacteria
Thermanaerovibrio acidaminovorans Su883 DSM 6589	Bacteria	Aminanaerobia
Thermobaculum terrenum YNP1, ATCC BAA-798	Bacteria	Chloroflexi
Thermobispora bispora DSM 43833	Bacteria	Actinobacteria
Thermomonospora curvata DSM 43183	Bacteria	Actinobacteria
Tsukamurella paurometabola DSM 20162	Bacteria	Actinobacteria
Veillonella parvula Te3, DSM 2008	Bacteria	Firmicutes
Xylanimonas cellulosilytica DSM 15894	Bacteria	Actinobacteria

Appendix E:
Review Committees and Board Members

CSP 2009 Reviewers

Eukaryotic Proposal Reviewers

Jody Banks Purdue University

Zac Cande University of California, Berkeley

Mike Freeling University of California, Berkeley

Steven Mayfield The Scripps Research Institute

Kevin McCluskey University of Missouri-Kansas City

Monica Medina University of California, Merced

Andrew Paterson University of Georgia

Arend Sidow Stanford University

Shauna Somerville Stanford University

Rob Steele University of California, Irvine

John Taylor University of California, Berkeley

Prokaryotic Proposal Reviewers

Patrick Chain Lawrence Livermore National Laboratory

Ludmila Chistoserdova University of Washington

Emilio Garcia Lawrence Livermore National Laboratory

Steven Hallam University of British Columbia

Martin Klotz University of Louisville

Todd Lowe University of California, Santa Cruz

David Mead Lucigen Corporation

Sebastien Monchy Brookhaven National Laboratory

Mircea Podar Oak Ridge National Laboratory

Frank Robb University of Maryland Biotechnology Institute

Craig Stephens Santa Clara University

Matthew Sullivan Massachusetts Institute of Technology

Tamas Torok Lawrence Berkeley National Laboratory

Bart Weimer Utah State University

DOE JGI Advisory Committees

DOE JGI Policy Board

The DOE JGI Policy Board serves two primary functions:

1. To serve as a visiting committee to provide advice on policy aspects of JGI operations and long-range plans for the program, including the research and development necessary to ensure the future capabilities that will meet DOE mission needs.

2. To ensure that JGI resources are utilized in such a way as to maximize the technical productivity and scientific impact of the JGI now and in the future. The JGI Policy Board meets annually to review and evaluate the performance of the entire JGI, including its component tasks and leadership. It reports its findings and recommendations to the participating Laboratory Directors and to the DOE BER.

Scientific Advisory Committee (SAC)

The Scientific Advisory Committee (SAC) is a board that the JGI Director convenes to provide a scientific and technical overview of the JGI. Responsibilities of this board include providing technical input on large-scale production sequencing, new sequencing technologies, and related informatics; overview of the scientific programs at the JGI; and overview of the Community Sequencing Program (CSP). A crucial job of the committee is to take the input from the CSP Proposal Study Panel on prioritization of CSP projects and, with DOE Office of Biological and Environmental Research's concurrence, set the final sequence allocation for this program.

MEMBERS

Gerry Rubin Howard Hughes Medical Institute (Chair)

Sallie W. (Penny) Chisholm Massachusetts Institute of Technology

Ed DeLong Massachusetts Institute of Technology

David Galas Institute for Systems Biology, Seattle, Washington, and Battelle Memorial Institute, Columbus, Ohio

Richard Gibbs Baylor College of Medicine

Stephen Quake Stanford University

Melvin Simon California Institute of Technology

Chris Somerville University of California, Berkeley

James Tiedje Michigan State University

Susan Wessler University of Georgia

MEMBERS

Mark Adams Case Western Reserve University

Ginger Armbrust University of Washington

Bruce Birren Broad Institute

Edward F. DeLong Massachusetts Institute of Technology

Joe Ecker Salk Institute

Richard Harland University of California, Berkeley

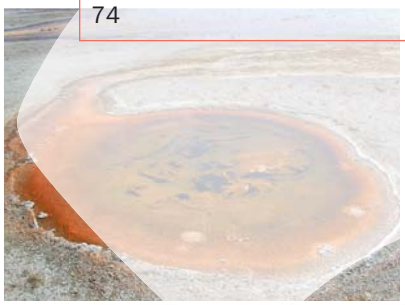
Marco Marra Michael Smith Genome Sciences Centre

Jim Krupnick Lawrence Berkeley National Laboratory

Eric J. Mathur Synthetic Genomics

Doug Ray Pacific Northwest National Laboratory

George Weinstock Baylor College of Medicine

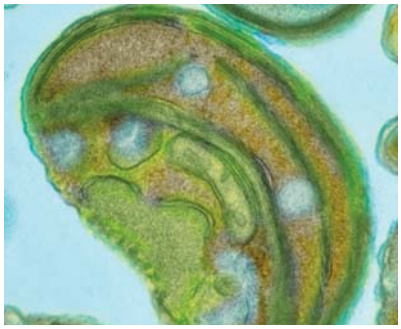



Appendix F:
2008 DOE JGI User Meeting Agenda

The DOE JGI Third Annual User Meeting took place March 26-28, 2008, in Walnut Creek. The keynote address was given by Nobel Laureate **Steven Chu,** then-Director of Lawrence Berkeley National Laboratory and currently the Secretary of Energy.

Other featured speakers were:

Steven Chu, delivering the keynote address at the 2008 DOE JGI User Meeting.


Eddy Rubin DOE Joint Genome Institute

Jerry Tuskan JGI and Oak Ridge National Laboratory

Andrew Paterson University of Georgia

Steve Long University of Illinois

Timothy Tschaplinski Oak Ridge National Laboratory

Dario Grattapaglia Brazilian Agricultural Research Corporation

Dan Rokhsar DOE Joint Genome Institute

Debra Mohnen University of Georgia

Virginia Walbot Stanford University

Mike Himmel National Renewable Energy Laboratory

Martin Keller Oak Ridge National Laboratory

James Liao University of California, Los Angeles

Nikos Kyrpides DOE Joint Genome Institute

Bernhard Palsson University of California, San Diego

Terry Hazen Lawrence Berkeley National Laboratory

John Taylor University of California, Berkeley

Mitch Sogin The Josephine Bay Paul Center

Audrey Gasch University of Wisconsin, Madison

Jim Bristow DOE Joint Genome Institute

Steve Mayfield Scripps Research Institute

John Glass J. Craig Venter Institute

Jay Shendure University of Washington

Jill Banfield University of California, Berkeley

Len Pennacchio DOE Joint Genome Institute

Appendix G:
Publications 2007-2008

Adamska, M., et al.
The evolutionary origin of hedgehog proteins. *Current Biology*, 17 (19):R836-7, Oct 2007

Anderson, I., et al.
Genome sequence of *Thermofilum pendens* reveals an exceptional loss of biosynthetic pathways without genome reduction. *Journal of Bacteriology*, 190 (8):2957-2965, Apr 2008

Angiuoli, S.V., et al.
Toward an online repository of Standard Operating Procedures (SOPs) for (Meta) genomic annotation. *OMICS: A Journal of Integrative Biology*, 12 (2):137-141, Jun 2008

Arp, D.J., et al.
The impact of genome analyses on our understanding of ammonia-oxidizing bacteria. *Annual Review of Microbiology*, 61:503-528, 2007

Auger, S., et al.
The genetically remote pathogenic strain NVH391-98 of the *Bacillus cereus* group is representative of a cluster of thermophilic strains. *Applied and Environmental Microbiology*, 74 (4):1276-1280, Feb 2008

Blow, M.J., et al.
Identification of ancient remains through genomic sequencing. *Genome Research*, 18 (8):1347-53, Aug 2008

Boore, J.L., et al.
Beyond linear sequence comparisons: the use of genome-level characters for phylogenetic reconstruction. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 363 (1496):1445-1451, Apr 27, 2008

Bowler, C., et al.
The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature*, 456:239-244, Oct 15, 2008

Challacombe, J.F., et al.
Comparative Genomics of the *Pasteurellaceae. Pasteurellaceae: Biology, Genomics and Molecular Aspects*, *Horizon Press*, 2008.

Chivian, D., et al.
Environmental Genomics Reveals a Single-Species Ecosystem Deep Within Earth. *Science,* 322:275-278, Oct 10, 2008

Dalevi, D., et al.
Automated group assignment in large phylogenetic trees using GRUNT: Grouping, ungrouping, naming tool. *BMC Bioinformatics*, 8:402, Oct 18, 2007

Dalevi, D., et al.
Annotation of metagenome short reads using proxygenes. *Bioinformatics*, 24 (16):I7-I13, Aug 15, 2008

Elkins, J.G., et al.
A korarchaeal genome reveals insights into the evolution of the Archaea. *Proceedings of the National Academy of Sciences of the United States of America,* 105(23):8102-7, Jun 10, 2008

Field, D., et al.
The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, 26(5):541-7, May 2008

Field, D., et al.
Foreword to the Special Issue on the Fifth Genomic Standards Consortium. *OMICS: A Journal of Integrative Biology*, 12(2):109-13, Jun 2008

Fong, J.J., et al.
A genetic survey of heavily exploited, endangered turtles: caveats on the conservation value of trade animals, *Animal Conservation*, 10 (4):452-460, Nov 2007

Fujita, M.K., et al.
Multiple origins and rapid evolution of duplicated mitochondrial genes in parthenogenetic geckos (*Heteronotia binoei; squamata, gekkonidae*). *Molecular Biology and Evolution*, 24 (12):2775-2786, Dec 2007

Garcia, E., et al.
Molecular characterization of L-413C, a P2-related plague diagnostic bacteriophage. *Virology*, 372 (1): 85-96, Mar 1, 2008

Garrity, G.M., et al.
Toward a standards-compliant genomic and metagenomic publication record. *OMICS: A Journal of Integrative Biology*, 12(2):157-60, Jun 2008

Gordon, L., et al.
Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions. *Genome Research*, 17 (11):1603-1613, Nov 2007

Grimson, A., et al.
Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455 (7217):1193-1197, Oct 30, 2008

Guisinger, N.P., et al.
Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. *PNAS*, 105:18424-18429, 2008

Haberle, R.C., et al.
Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *Journal of Molecular Evolution*, 66 (4):350-361, Apr 2008

Hendrix, D.A., et al.
Promoter elements associated with RNA Pol 11 stalling in the *Drosophila* embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 105 (22):7762-7767, Jun 3, 2008

Hess, M.
Thermoacidophilic proteins for biofuel production. *Trends in Microbiology*, 16(9):414-9, Sep 2008

Hess, M., et al.
Extremely thermostable esterases from the thermoacidophilic euryarchaeon *Picrophilus torridus. Extremophiles*, 12 (3):351-364, May 2008

Hirschman, L., et al.
Habitat-Lite: A GSC case study based on free text terms for environmental metadata. *OMICS: A Journal of Integrative Biology*, 12 (2):129-136, Jun 2008

Holland, L.Z., et al.
The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Research,* 18(7):1100-11, Jul 2008. Erratum in: *Genome Research* 18(8):1380.

Hooper, S.D., et al.
A molecular study of microbe transfer between distant environments. *PLoS One*, 9;3(7):e2607, Jul 2008

Horton, A.C., et al.
Conservation of linkage and evolution of developmental function within the Tbx2/3/4/5 subfamily of T-box genes: implications for the origin of vertebrate limbs. *Development Genes and Evolution*, 218(11):613-628, Sep 25, 2008

Hristova, K.R., et al.
Comparative transcriptome analysis of *Methylibium petroleiphilum* PM1 exposed to the fuel oxygenates methyl tert-butyl ether and ethanol. *Applied and Environmental Microbiology*, 73 (22):7347-7357, Nov 2007

Hugenholtz, P., et al.
Microbiology: Metagenomics, *Nature*, 455:481-483, Sep 25, 2008

Hwang, J.S., et al.
Cilium evolution: Identification of a novel protein, nematocilin, in the mechanosensory cilium of *Hydra* nematocytes. *Molecular Biology and Evolution*, 25 (9):2009-2017, Sep 2008

Isabel, S., et al.
Divergence Among Genes Encoding the Elongation Factor Tu of *Yersinia* Species. *Journal of Bacteriology*, 190(22):7548-7558, Nov 2008

Ishoey, T., et al.
Genomic sequencing of single microbial cells from environmental samples. *Current Opinion in Microbiology*, 11 (3):198-204, Jun 2008

Jansen, R.K., et al.
Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences of the United States of America,* 104 (49):19369-19374, Dec 4, 2007

Kallimanis, A., et al.
Identification of two 1-hydroxy-2-naphthoate dioxygenase genes in *Arthrobacter* sp strain Sphe3. *FEBS Journal*, 275:409-409 Suppl. 1, Jun 2008

Kalluri, U.C., et al.
Genome-wide analysis of Aux/IAA and ARF gene families in *Populus trichocarpa. BMC Plant Biology*, 7:59, Nov 6, 2007

Kalyuzhnaya, M.G., et al.
High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature Biotechnology*, 26 (9):1029-1034, Sep 2008

Kamburov, A., et al.
Denoising inferred functional association networks obtained by gene fusion analysis. *BMC Genomics*, 8:460, Dec 14, 2007

Katsaveli, K., et al.
Microbial community shifts during the annual operation of Messolonghi solar saltern, Greece. *FEBS Journal*, 275:282-282 Suppl. 1, Jun 2008

King, N., et al.
The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature*, 451 (7180):783-788, Feb 14, 2008

Kraaijeveld, K., et al.
The evolution of sperm and non-sperm producing organs in male Drosophila. *Biological Journal of the Linnean Society*, 94 (3):505-512, Jul 2008

Kunin, V., et al.
A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Research*, 18(2):293-7, Feb 2008

Kunin, V., et al.
Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Molecular Systems Biology*, 4:198, 2008.

Kyrpides, N.
Genomics, evolution and evolution of genomics. *FEBS Journal*, 275:9-9 Suppl. 1, Jun 2008

Labbé, J., et al.
A genetic linkage map for the ectomycorrhizal fungus *Laccaria bicolor* and its alignment to the whole-genome sequence assemblies. *New Phytologist*, 180 (2):316-328, Sep 2008

Lapidus, A., et al.
Extending the *Bacillus cereus* group genomics to putative foodborne pathogens of different toxicity. *Chemico-Biological Interactions,* 171 (2):236-249, Jan 30, 2008

Larroux, C., et al.
Genesis and expansion of metazoan transcription factor gene classes. *Molecular Biology and Evolution*, 25(5):980-96, May 2008

Lee, J.H., et al.
Comparative genomic analysis of the gut bacterium *Bifidobacterium longum* reveals loci susceptible to deletion during pure culture growth. *BMC Genomics*, 9:247, May 2008

Leonardi, R., et al.
Localization and regulation of mouse pantothenate kinase 2. *FEBS Letters*, 581 (24):4639-4644, Oct 2, 2007

Liolios, K., et al.
The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 36:D475-D479 Special Issue SI, Jan 2008

Loh, Y.H., et al.
Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids. *Genome Biology,* 9(7):R113, 2008

Mackelprang, et al.
Paleontology—New tricks with old bones. *Science*, 321 (5886):211-212, Jul 11, 2008

Marco, E.J., et al.
ARHGEF9 disruption in a female patient is associated with X linked mental retardation and sensory hyperarousal. *Journal of Medical Genetics*, 45 (2):100-105, Feb 2008

Markowitz, V.M., et al.
IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Research,* 36:D534-D538 Special Issue SI, Jan 2008

Markowitz, V.M., et al.
The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Research*, 36:D528-D533 Special Issue SI, Jan 2008

Martin, F., et al.
The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature*, 452 (7183):88-U7, Mar 6, 2008

Martinez, D., et al.
Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn.

Hypocrea jecorina). *Nature Biotechnology*, 26 (5):553-560, May 2008

Masta, SE, et al.
Parallel evolution of truncated transfer RNA genes in arachnid mitochondrial genomes. *Molecular Biology and Evolution*, 25 (5):949-959, May 2008

Mattes, T.E., et al.
The genome of *Polaromonas* sp. strain JS666: insights into the evolution of a hydrocarbon- and xenobiotic-degrading bacterium, and features of relevance to biotechnology. *Applied and Environmental Microbiology*, 74(20):6405-6416, Oct 2008

Mccoy, S.R., et al.
The complete plastid genome sequence of *Welwitschia mirabilis:* an unusually compact plastome with accelerated divergence rates. *BMC Evolutionary Biology*, 8:130, May 1, 2008

McNeal, J.R., et al.
Systematics and plastid genome evolution of the cryptically photosynthetic parasitic plant genus *Cuscuta* (Convolvulaceae). *BMC Biology*, 5:55, Dec 13, 2007

McNeal, J.R., et al.
Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta. BMC Plant Biology*, 7:57, Oct 24, 2007

Merchant, S.S., et al.
The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science*, 318(5848):245–250, Oct 12, 2007

Minovitsky, S., et al.
Short sequence motifs, overrepresented in mammalian conserved non-coding sequences. *BMC Genomics*, 8:378, Oct 18, 2007

Mueller, R.L., et al.
Genome size, cell size, and the evolution of enucleated erythrocytes in attenuate salamanders. *Zoology*, 111 (3):218-230, 2008

Norton, J.M., et al.
Complete genome sequence of *Nitrosospira multiformis,* an ammonia-oxidizing bacterium from the soil environment. *Applied and Environmental Microbiology,* 74 (11):3559-3572, Jun 2008

Opperman, C.H., et al.
Sequence and genetic map of *Meloidogyne hapla:* A compact nematode genome for plant parasitism. *Proceedings of the National Academy of Sciences of the United States of America,* 105(39):14802-7, Sep 30, 2008

Paterson, A., et al.
The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457, 551-556, Jan 2009

Peterson, S.B., et al.
Environmental distribution and population biology of *Candidatus Accumulibacter,* a primary agent of biological phosphorus removal. *Environmental Microbiology*, 10 (10):2692-2703, Oct 2008

Pierce, E, et al.
The complete genome sequence of *Moorella thermoacetica* (*f. Clostridium thermoaceticum*) *Environmental Microbiology*, 10 (10):2550-2573, Oct 2008

Prabhakar, S., et al.
Human-specific gain of function in a developmental enhancer. *Science*, 321 (5894):1346-1350, Sep 5, 2008

Putnam, N.H., et al.
The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453 (7198):1064-U3, Jun 19, 2008

Quesada, T., et al.
Comparative analysis of the transcriptomes of *Populus trichocarpa* and *Arabidopsis thaliana* suggests extensive evolution of gene expression regulation in angiosperms. *New Phytologist*, 180 (2):408-420, 2008

Ralph, S.G., et al.
Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding. *BMC Genomics*, 9:57, Jan 29, 2008

Rensing, S.A., et al.
The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, 4;319(5859):64-69, Jan 2008

Riley, M., et al.
Genomics of an extreme psychrophile, *Psychromonas ingrahamii*. *BMC Genomics*, 9:210, May 6, 2008

Rubin, E.M.
Genomics of cellulosic biofuels. *Nature*, 454 (7206):841-845, Aug 14, 2008

Savidor, A., et al.
Cross-species global proteomics reveals conserved and unique processes in *Phytophthora sojae* and *P. ramorum. Molecular & Cellular Proteomics*, 7 (8):1501-1516, Aug 2008

Schoenfeld, T., et al.
Assembly of viral metagenomes from Yellowstone hot springs. *Applied and Environmental Microbiology*, 74 (13):4164-4174. Jul 2008

Schwarz, J.A., et al.
Coral life history and symbiosis: Functional genomic resources for two reef building Caribbean corals, *Acropora palmata* and *Montastraea faveolata. BMC Genomics*, 9:97, Feb 25, 2008

Scott, K.M., et al.
Genome of the epsilon proteo bacterial chemolithoautotroph *Sulfurimonas denitrificans. Applied and Environmental Microbiology*, 74 (4):1145-1156, Feb 2008

Shoemaker, R.C., et al.
Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. *Genome*, 51:294-302, Mar 18, 2008.

Sievert, S.M., et al.
The genome of the epsilonproteobacterial chemolithoautotroph *Sulfurimonas denitrificans. Applied and Environmental Microbiology*, 74(4):1145-56, Feb 2008

Smith, D.R., et al.
Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Research*, 18 (910):1638-1642, Oct 2008

Smith, T.J., et al.
Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3, /Ba4 and /B1 clusters are located within plasmids. *PLoS One*, 5;2(12):e1271, Dec 2007

Sorek, R.
The birth of new exons: Mechanisms and evolutionary consequences. RNA-A *Publication of the RNA Society*, 13 (10):1603-1608, Oct 2007

Sorek, R., et al.
CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews Microbiology*, 6 (3):181-186, Mar 2008

Sorek, R., et al.
Genome-wide experimental determination of barriers to horizontal gene transfer. *Science*, 318 (5855):1449-1452, Nov 30, 2007

Srivastava, M., et al.
The *Trichoplax* genome and the nature of placozoans. *Nature*, 454 (7207):955-U19, Aug 21, 2008

Starkenburg, S.R., et al.
Complete genome sequence of *Nitrobacter hamburgensis* X14 and comparative genomic analysis of species within the genus *Nitrobactei. Applied and Environmental Microbiology*, 74 (9):2852-2863, May 2008

Stein, L.Y., et al.
Whole-genome analysis of the ammonia-oxidizing bacterium, *Nitrosomonas eutropha* C91: implications for niche adaptation. *Environmental Microbiology*, 9 (12):2993-3007, Dec 2007

Tanaka, S., et al.
Atomic-level models of the bacterial carboxysome shell. *Science*, 319 (5866):1083-1086, Feb 22, 2008

Tanaka, S., et al.
Structure of the RuBisCO chaperone RbcX from Synechocystis sp PCC6803. *Acta Crystallographica Section D-Biological Crystallography*, 63:1109-1112 Part 10, Oct 2007

Torriani, S.F.F., et al.
Intraspecific comparison and annotation of two complete mitochondrial genome sequences from the plant pathogenic fungus *Mycosphaerella graminicola. Fungal Genetics and Biology*, 45 (5):628-637, May 2008

Tringe, S.G., et al.
A renaissance for the pioneering 16S rRNA gene. *Current Opinion in Microbiology*, 11 (5):442-446, Oct 2008

Tringe, S.G., et al.
The airborne metagenome in an indoor urban environment. *PLoS One*, 2;3(4):e1862, Apr 2008

Tuanyok, A., et al.
A horizontal gene transfer event defines two distinct groups within *Burkholderia pseudomallei* that have dissimilar geographic distributions. *Journal of Bacteriology*, 189(24):9044-9, Dec 2007

Vallès, Y., et al.
Group II introns break new boundaries: presence in a bilaterian's genome. *PLoS One*, 3(1):e1488, Jan 23, 2008

Van Brabant, B., et al.
Laying the foundation for a genomic rosetta stone: Creating information hubs through the use of consensus identifiers. *OMICS: A Journal of Integrative Biology*, 12 (2):123-127, Jun 2008

Visel, A., et al.
Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature Genetics*, 40 (2):158-160, Feb 2008

Warnecke, F., et al.
Building on basic metagenomics with complementary technologies. *Genome Biology*, 8 (12):231, 2007

Warnecke, F., et al.
Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450 (7169):560-U17, Nov 22, 2007

Weiner, R.M., et al.
Complete genome sequence of the complex carbohydrate-degrading marine bacterium, *Saccharophagus degradans* strain 2-40(T). *PLoS Genetics*, 4(5):e1000087, May 2008

Wickett, N.J., et al.
Functional gene losses occur with minimal size reduction in the plastid genome of the parasitic liverwort *Aneura mirabilis. Molecular Biology and Evolution*, 25 (2):393-401, Feb 2008

Wilson, A., et al.
A photoactive carotenoid protein acting as light intensity sensor. *Proceedings of the National Academy of Sciences of the United States of America,* 105 (33):12075-12080, Aug 19, 2008

Yang, X., et al.
F-box Gene Family is Expanded in Herbaceous Annual Plants Relative to Woody Perennial Plants. *Plant Physiology*, 148 (3):1189-1200, Nov 2008

Yin, T., et al.
Genome structure and emerging evidence of an incipient sex chromosome in *Populus. Genome Research*, 18 (3):422-30, Mar 2008