2011

U.S. Department of Energy
Joint Genome Institute

PROGRESS REPORT

Sequencing the world
of possibilities for energy
and the environment

Table of Contents

DOE JGI Mission

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

Director's Perspective

Impressive progress has been made in the 15 years since the Department of Energy Joint Genome Institute (DOE JGI) was founded as a virtual entity to invigorate the federally funded effort to sequence the human genome. Two years later, building on this momentum, the Institute opened its doors in Walnut Creek, California, and today, the DOE JGI is the preeminent genomic resource devoted to characterizing plants, microbes, and microbial communities—the living systems that help our planet sustain itself.

The DOE JGI fills a unique scientific niche, providing its user community access to state-of-the-art genomic technologies and a scientific staff committed to helping users exploit these capabilities to solve bioenergy, carbon cycling, and bioremediation problems. By focusing this experimental and analysis expertise on the best peer-reviewed ideas drawn from a broad community of scientists, the DOE JGI continuously advances creative scientific solutions to challenges relevant to the DOE's mission.

Our success is realized and reinforced by high-profile publications in the scientific literature. From 2009-2011, the DOE JGI published more than 400 papers, with 38 in the journals *Science* and *Nature* alone. These papers, and the 10 plant, 69 fungal, four algal, and 476 prokaryotic genomes, as well as the 193 metagenomic data sets that the DOE JGI contributed to public databases during the same time period, provide an essential infrastructure for energy and environmental research.

These projects were all underpinned with genome sequence data. In 2011, the DOE JGI far exceeded its previous record of sequencing output, generating more than 30 terabases (trillion nucleotides). This is the equivalent of sequencing 10,000 human genomes, though the output is actually of plant, fungal, microbial, and metagenomic data—a thirtyfold increase over the past three years.

Last year, the DOE JGI focused its sequencing assets on the Illumina HiSeq and PacBio platforms, achieving efficiencies and additional savings by doing so. However, for the DOE JGI, it's not only about generating sequence—informatics supports all that the Institute does. Its various data-management and analysis portals (the Genome Portal, IMG family, Phytozome, MycoCosm) strive to deliver the highest-quality products to an ever-expanding user community. These in-house resources were featured in a recent special database issue of *Nucleic Acids Research* and have contributed to the completion of an unprecedented 800 projects this year alone.

Translating sequence data into high-impact science is an essential part of the Institute's own "DNA." A selection of these highlights:

- Identified nearly 30,000 genes from the cow rumen that encode possible plant biomass degrading enzymes useful for producing simple sugars, the essential first step in cellulosic biofuel production (*Science*)

- Described how methane-producing microbes isolated from the frigid soil in the Arctic permafrost are responding to their warming environment (*Nature*)

Eddy Rubin

- Identified, using single-cell genomics, the predominant bacterial lineages involved in carbon fixation in the dark ocean (*Science*)

- Elucidated the plant cell-wall decomposing machinery from a boreal forest brown rot that causes millions of dollars of damage to buildings, offering lessons for biofuel pretreatment and a better understanding of the role of fungus in the global carbon cycle (*Science*)

- Characterized the genome of a bacterium found in a marsupial gut that could help explain why its methane emissions are lower than those of livestock (*Science*)

- Afforded, by analysis of the genome of the tiny spike moss, a comparative genomics approach to identify the core genes likely to be present in a common ancestor to land plants, as well as a way to study lignin evolution (*Science*)

- Compared the genomes of two rust fungi to identify the characteristics by which these pathogens can invade their plant hosts and to develop methods to control the damage they can cause (*PNAS*)

- Completed and annotated the genome sequence of a brown-tide species responsible for reducing the light and oxygen available in the ecosystem by discoloring coastal waters (*PNAS*)

- Revealed the most gene-packed animal characterized to date—the water flea—a tiny crustacean that is a keystone species in freshwater ecosystems, and that has been used for decades to develop and monitor environmental regulations (*Science*)

Over the next few years, both predicted and unexpected technological innovations will continue to propel genomic science forward. Existing grand challenges will be surmounted; new grand challenges will emerge.

A complete list of our publications can be found in Appendix F.

There are more exciting research developments in the offing, enabled by the DOE JGI's latest capabilities in DNA synthesis and single-cell genomics. These and other resources have been driven by the Institute's strategic planning. The DOE JGI has been on a deliberate path from moderate-scale user sequencing projects to much larger, complex projects whose targets are driven by systematic and problem-focused approaches. These larger projects, of foundational relevance to energy and environmental research, require both massive-scale sequencing and analysis capabilities along with a focused commitment to cultivating these community resources, vital infrastructure for the bioenergy, and environmental research communities. Examples of this have been the large phylogeny-based genomic surveys of bacteria, archaea, and fungi and the systematic exploration of plant genomes relevant to bioenergy feedstock development.

Through these activities, the DOE JGI has emerged as the world's largest producer of prokaryotic, fungal, and plant genomic data. In addition, several large-scale metagenomic projects focusing on the in-depth characterization of DOE-relevant environments have

been completed (cow rumen, permafrost, etc.). Entire communities of scientists working in a particular field, such as biofuel feedstock improvement or biomass degradation, have come to rely on and grow their research directions around information produced by the DOE JGI.

Necessitated by the extremely dynamic scientific and technical developments in the field of genomics, in 2011 the Institute crafted a "Strategic Vision: Forging a Future for the DOE JGI." This draft outlines how, in parallel with sequence data, an increasing emphasis will be placed on developing a series of complementary presequencing and postsequencing capabilities that focus on the conversion of genomic data into biological insights. These initiatives include:

- Large-scale rapid DNA synthesis to accelerate the linking of sequence to function for testing genomics-derived hypotheses, creation of synthetic pathways, and functional exploration of metagenomic and other sequence data sets

- High-performance computing infrastructure for data processing that, in partnership with Lawrence Berkeley National Laboratory's (Berkeley Lab's) National Energy Research Scientific Computing

Center (NERSC) and other supercomputing centers, will enable the DOE JGI to efficiently process and integrate the rapidly increasing number and size of sequence data sets

• Massive-scale and customizable sample processing to develop capabilities that include the implementation of automated DNA/RNA extractions able to process tens of thousands of samples and implement large-scale single-cell and single-chromosome isolation techniques

• Organization of mission-oriented user communities to ensure continuity of access to state-of-the-art genomics capabilities and strategies, and facilitate data sharing and integration to speed progress toward solving DOE's and the nation's most pressing challenges in alternative fuels, carbon management, and climate and environmental remediation

Our 10-Year Strategic Vision is still a work in progress. It is evolving with the feedback we receive from our user community and guiding the appropriate suite of capabilities required for carrying out future large-scale environmental and systems biological studies.

While the DOE JGI explores the latest in pre- and postsequencing capabilities, it is already offering access to large-scale single-cell genomics to study hard-to-culture environmental microbes. A rapid increase in the requests for this capability has led to a corresponding production increase, from 14 genomes in 2009 to 320 in 2011. Among the postsequencing

capabilities being pursued are: gene expression (RNA) profiles and metatranscriptomics as a function of natural or experimental perturbation, and a broad diversity of customized data-analysis resources.

Technological innovation in genomics continually shifts our perspective on the questions that are feasible to address. The organization of the DOE JGI into four scientific programs—Plants, Microbes, Fungi, and Metagenomes—is an important foundation of its scientific successes. We are, however, aware that this administrative structure may represent in some ways an artificial parsing of many modern biological questions, as the dynamic interaction among each of these domains in the environment is increasingly being appreciated. As the structure of the DOE JGI has evolved, integrative studies are blurring the boundaries between domains and programs. This is reflected in recently approved projects such as the analysis of plant-microbe interactions in the rhizosphere to explore the systems biology of symbiosis.

Over the next few years, both predicted and unexpected technological innovations will continue to propel genomic science forward. Existing grand challenges will be surmounted; new grand challenges will emerge. Throughout these changes, the mission of the DOE JGI—to provide users with access to the latest advancements in nucleic acid-based approaches as well as other synergistic capabilities—remains unchanged and will continue to reinforce the Institute's unique position as the world's leading user facility in energy and environmental genomics.

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

DOE Mission Areas

Bioenergy

The United States is the world's single largest consumer of petroleum-based energy, and imports more than half of the amount used. Two-thirds of the energy consumed goes to transportation and industry, which drives the DOE's focus on developing clean and sustainable alternative fuel sources such as cellulosic biofuels. Sequencing projects at the DOE JGI that contribute to meeting this goal focus on one of three categories: developing plants that can be used as feedstocks for biofuel production, using enzymes from fungi and microbes to break down the lignin and cellulose in plant walls, and identifying organisms that can ferment sugars into biofuels.

Carbon Cycle

The global carbon cycle regulates the levels of atmospheric carbon dioxide and the Earth's climate. It is heavily dependent on the microbes that fix atmospheric carbon, promoting plant growth and degrading organic material. As microbes constitute the largest component of the Earth's biodiversity, understanding how they metabolize carbon, and how environmental changes affect these processes, is crucial. The DOE JGI is sequencing several microbes and microbial communities that influence carbon cycling. With this information, researchers can develop better predictive models that could provide more effective contributions toward reducing the effects of increasing carbon dioxide emissions on the global climate.

Cross section of a stem of switchgrass, a candidate bioenergy feedstock crop at micrometer scale. The green fluorescent areas indicate regions with more chlorophyll, while the blue regions designate xylem tissues involved in transporting water and nutrients.
*(Image by BESC researcher Shi-You Ding, NREL)*
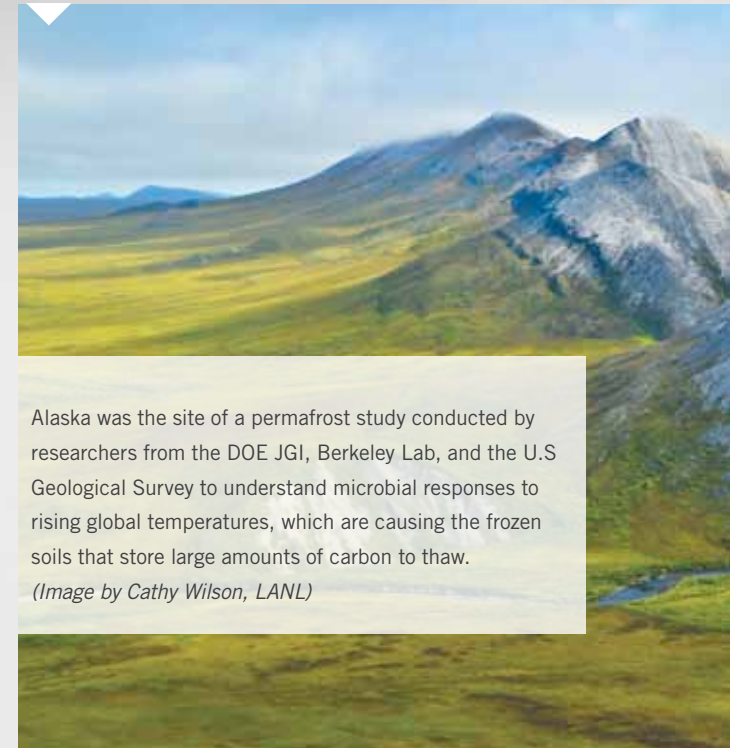
Alaska was the site of a permafrost study conducted by researchers from the DOE JGI, Berkeley Lab, and the U.S Geological Survey to understand microbial responses to rising global temperatures, which are causing the frozen soils that store large amounts of carbon to thaw.
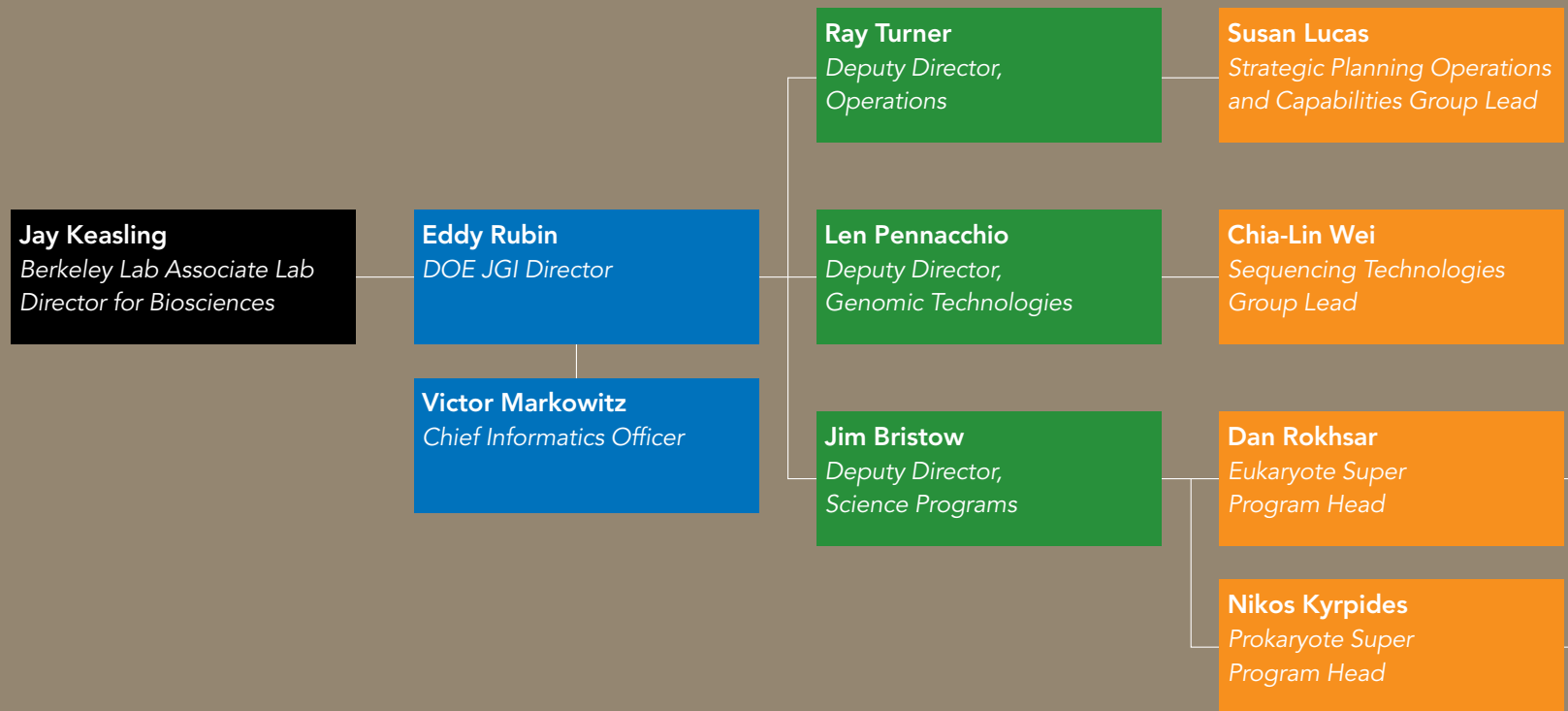*(Image by Cathy Wilson, LANL)*

Biogeochemistry

The carbon cycle is not the only process that regulates the natural environment, and the field of biogeochemistry explores the full spectrum of biological, physical, geological, and chemical processes and reactions involved. Microbes and microbial communities that can degrade or otherwise transform environmental contaminants such as toxic chemicals or heavy metals present another area of focus for the DOE JGI.
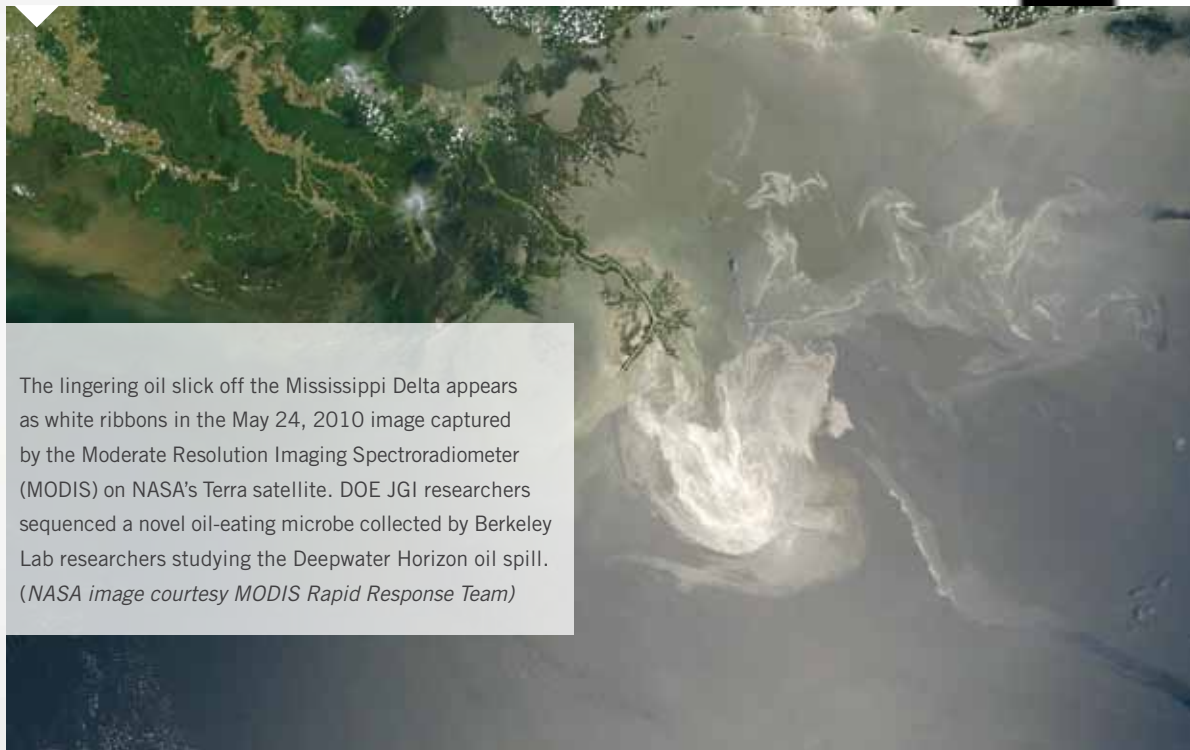
The lingering oil slick off the Mississippi Delta appears as white ribbons in the May 24, 2010 image captured by the Moderate Resolution Imaging Spectroradiometer (MODIS) on NASA's Terra satellite. DOE JGI researchers sequenced a novel oil-eating microbe collected by Berkeley Lab researchers studying the Deepwater Horizon oil spill. (*NASA image courtesy MODIS Rapid Response Team*)

Organizational Structure

Strategic Management

Jay Keasling
Berkeley Lab Associate Lab Director for Biosciences

Eddy Rubin
DOE JGI Director

Victor Markowitz
Chief Informatics Officer

Ray Turner
Deputy Director, Operations

Len Pennacchio
Deputy Director, Genomic Technologies

Jim Bristow
Deputy Director, Science Programs

Susan Lucas
Strategic Planning Operations and Capabilities Group Lead

Chia-Lin Wei
Sequencing Technologies Group Lead

Dan Rokhsar
Eukaryote Super Program Head

Nikos Kyrpides
Prokaryote Super Program Head

The DOE Joint Genome Institute Partnership

The DOE JGI's Walnut Creek, California, headquarters draws its workforce from Lawrence Berkeley National Laboratory (Berkeley Lab) and Lawrence Livermore National Laboratory (LLNL). Additional partners and their respective roles include:

Berkeley Lab's National Energy Research Scientific Computing Center (NERSC):
Computational and storage systems and support; high-performance computing applications

HudsonAlpha Institute for Biotechnology:
Plant genome assembly, eukaryotic genome improvement

Los Alamos National Laboratory (LANL):
Microbial genome improvement, metagenome assembly R&D

Oak Ridge National Laboratory (ORNL):
Plant genome biology

Igor Grigoriev
Fungal Program Lead

Jeremy Schmutz
Plant Program Lead
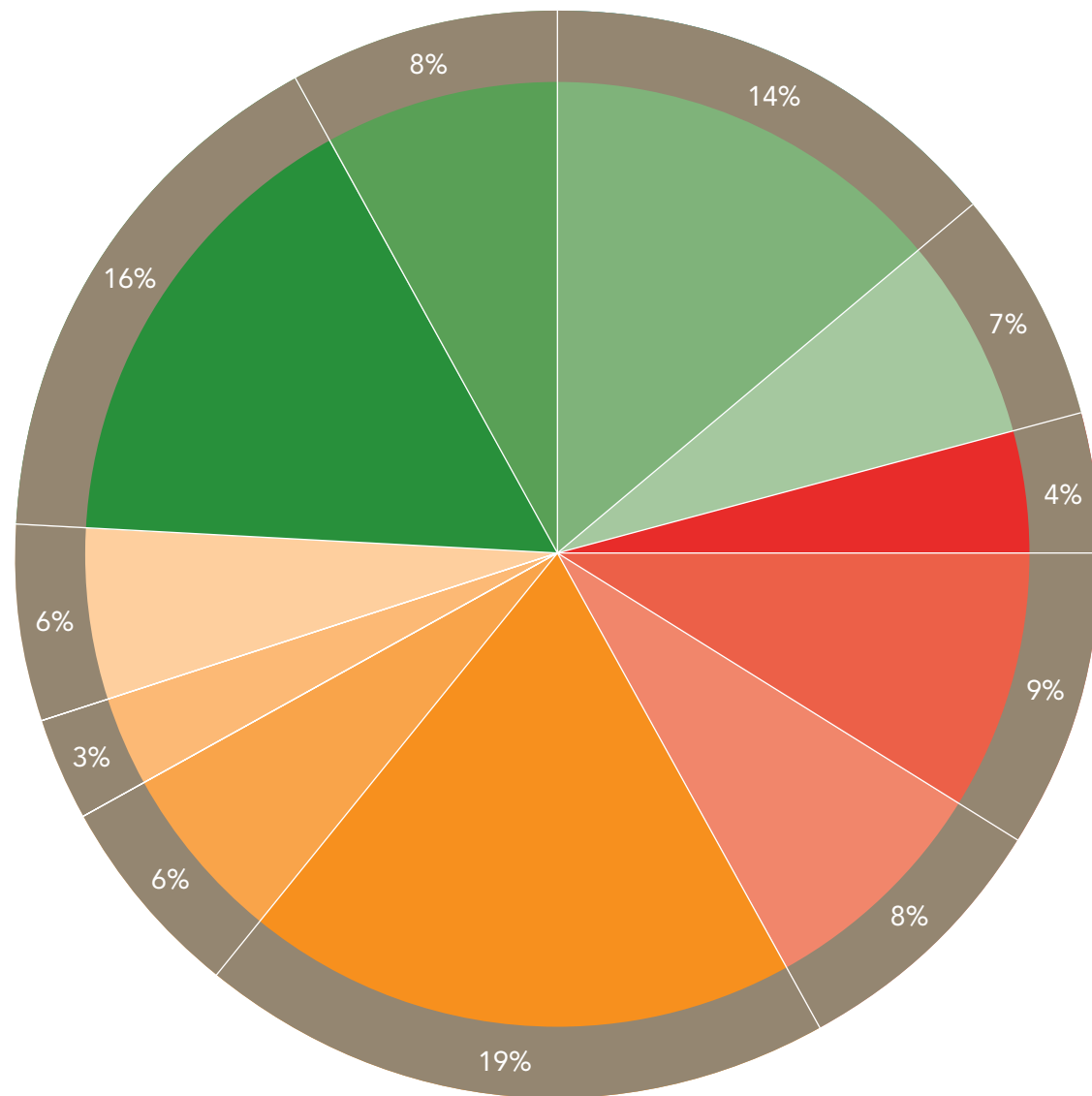
Susannah Tringe
Metagenome Program Lead

Tanja Woyke
Microbial Program Lead

FY2011 Operating Budget ($69.3M)

Plant $11.0

Fungal $5.3

Microbial $9.9

Metagenome $5.0

Management $2.7

Operations $6.4

Facility $5.2

Sequencing $13.2

IT Infrastructure $4.4

Science $2.3

Genomic Technologies $3.9

DOE JGI Sequence Productivity

Total Bases
(in billions of bases or Gb)

Year	Value
2006	32.74
2007	30.31
2008	61.76
2009	1,003.86
2010	6,040.796
2011	29,903

Impact

North America 1,153

United States 1,049
Canada 97
Mexico 7

Number of Users

1,106 DOE-defined Unique Users

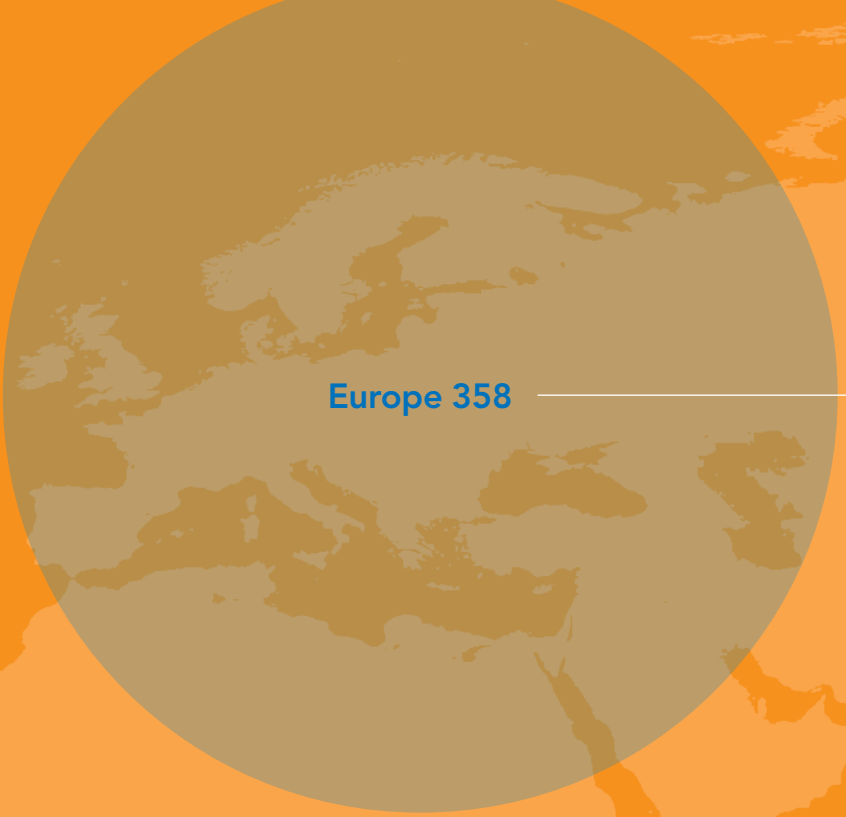(PIs, Co-PIs, collaborators, annotators)

512 Unique IMG-ER Users

156 Educators

1,713 Total Unique Users

(Unique users counted once and sourced from the above groups)

Brazil 23
Chile 6
Colombia 1
Argentina 6
Antigua and Barbuda 1
Peru 1
Uruguay 1

South America 39

Europe 358

United Kingdom 49
Germany 66
France 38
Netherlands 38
Spain 28
Sweden 24
Austria 13
Czech Republic 4
Denmark 16
Italy 15
Finland 10
Switzerland 11
Greece 8
Norway 8
Portugal 8
Russia 5
Ireland 5
Belgium 6
Slovenia 2
Turkey 2
Estonia 1
Ukraine 1

Asia 104

China (incl. HK and Taiwan) 37
India 21
Israel 11
South Korea 10
Japan 16
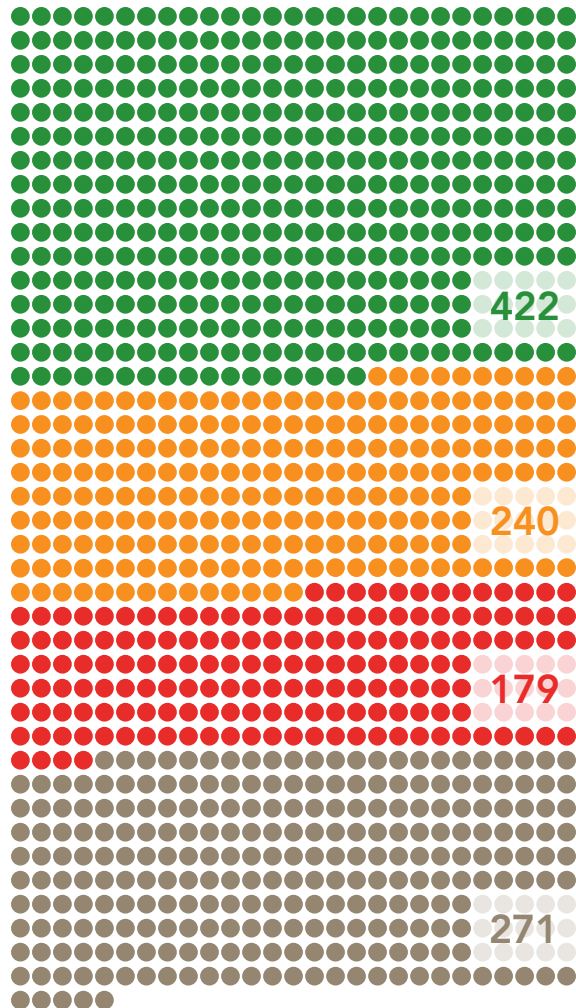Thailand 2
Afghanistan 2
Saudi Arabia 4
Qatar 1

Africa 5

South Africa 4
Nigeria 1

Australia / New Zealand 54

Australia 41
New Zealand 13

Selected Workshop Participation

422 — 6th Annual Genomics of Energy & Environment Meeting
(workshops include Phytozome, MycoCosm, and RNA-Seq)

240 — 6th Annual Sequencing, Finishing & Analysis in the Future Meeting

179 — Microbial Genomics & Metagenomics Workshops
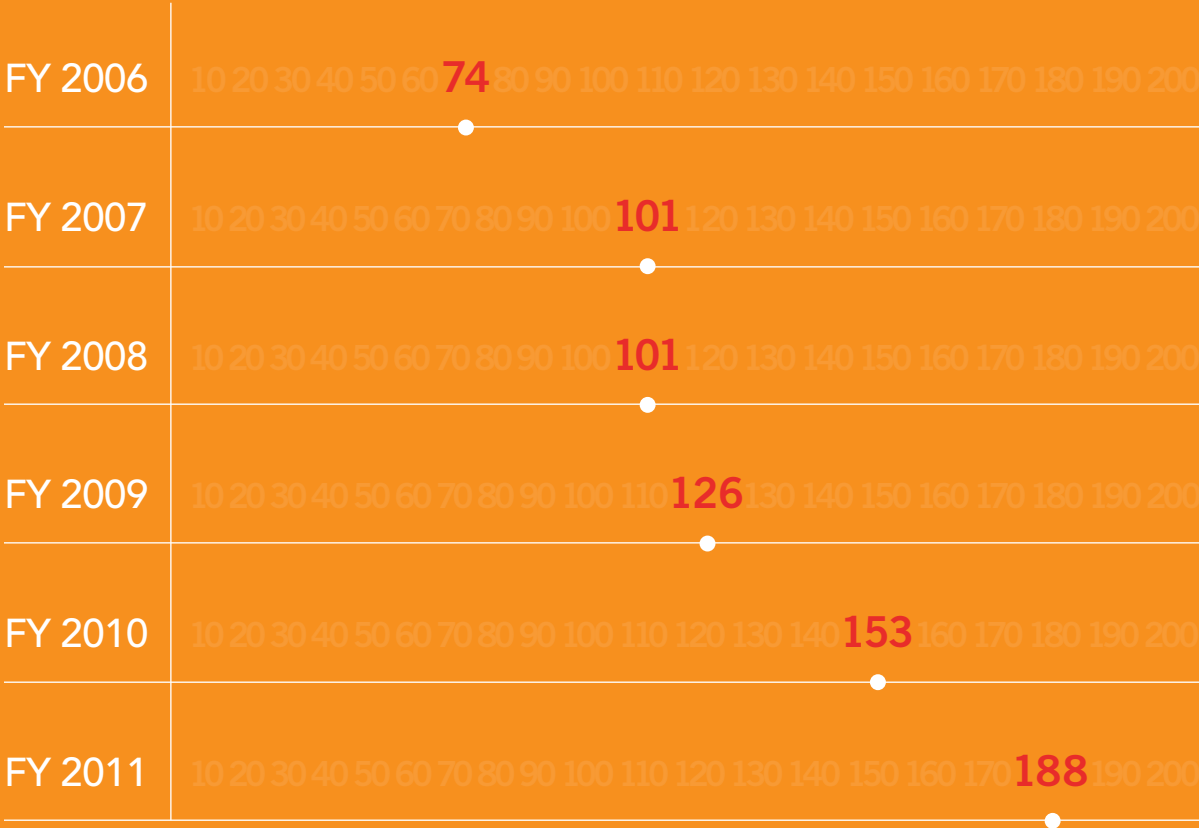*(at four meetings)*

271 — Fungal Program Workshops
(at seven meetings)

Number of participants

Contributions to the Scientific Literature

	10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200
FY 2006	**74**
FY 2007	**101**
FY 2008	**101**
FY 2009	**126**
FY 2010	**153**
FY 2011	**188**

Number of publications

DOE JGI Projects

Projects Completed by Fiscal Year

Project Type	FY 2008	FY 2009	FY 2010	FY 2011	TOTALS
Fungal	4	4	23	40	71
Metagenome	32	57	109	200	398
Microbial	144	64	153	171	532
Plant	73	108	83	67	331
Resequencing	3	29	206	304	542
RNAseq	31	28	78	92	229
TOTALS	287	290	652	874	2,103

Projects Completed by User Program

User Program	FY 2008	FY 2009	FY 2010	FY 2011	TOTALS
American Recovery and Reinvestment Act				38	38
Bioenergy Research Centers	9	34	211	266	520
Director's Discretionary	39	14	66	78	197
Grand Challenge		27	53	74	154
Work For Others	11	14	7	14	46
Community Sequencing Program/DOE	228	201	315	404	1,148
TOTALS	287	290	652	874	2,103

Building Communities: *Daphnia*

An example of how DOE JGI sequencing projects build scientific communities involves the water flea *Daphnia pulex*. This tiny organism is a keystone freshwater species and was the first crustacean to have its genome sequenced. *(Image by Jan Michels, Christian-Albrechts-University, Kiel)*

2003: Genome sequencing project approved	2007: Genome sequence released	2011: Genome sequence published in *Science*
26 Investigators **9** Publications	**150** Investigators **18** Publications	**475** Investigators **50** Publications

In the News

DOE JGI projects and people were featured
by major news agencies in 2011.

Top Left: In July, the *New York Times* featured cellulosic
biofuels work being done at the DOE JGI, touching on
several fungal and metagenome projects including the
cow rumen and tammar wallaby gut. Read the piece at
http://nyti.ms/pGqCGc.

Top Right: DOE JGI Metagenome Program Lead Susannah
Tringe received a $2.5 million grant to study the microbial
communities of restored wetlands in the San Joaquin Delta
region of California and made the *Popular Science* "Brilliant
10" list in May. More on this at http://1.usa.gov/SGTPop.

Bottom Right: The *Washington Post* ran a story in November
on the importance of cow rumen research, and by extension,
the importance of the DOE JGI's programs and projects.
Read the piece at http://bit.ly/rCl8UD.

Bottom Left: The *San Francisco Chronicle* extensively
covered the publication of the water flea genome in February.
Read the story at http://bit.ly/JGIDpS.

The New York Times

July 13, 2011

No Eureka Moments in Long U.S. Campaign to Crack Cellulosic Code

By PAUL VOOSEN of

Fourth in an occasional series on the bioeconomy.

Down on a farm in Illinois, his forearm stuck inside the noisome gut of a living and otherwise unperturbed brown cow, Matthias Hess, a German-born microbiologist and geneticist, felt far removed from the white hum of his biology lab.

Hess had been fishing in the cow's rumen, its largest stomach, for a nylon mesh sack resembling an oversized teabag. The stink of vomit mixed with rotten eggs and fertilizer. Working through a permanent rubberized hole carved into the heifer's side, Hess waited for its half-digested slop to churn, freeing his hand. Then he pulled out the teabag, which three days earlier he had stuffed with pulverized prairie grass.

The rumen is like a huge bathtub, he said, holding about 50 large soda bottles' worth of fluid redolent with bacteria. Relying on these symbiotic microbes, cows eat up to 150 pounds of grass a day, a food inedible to most animals, including humans. Hess was after those microbial secrets, and the placid heifer was happy to oblige.

"You can just do your experiment," he said later. "The cows don't really care."

Unglamorous as it may sound, Hess and his fellow researchers are at the forefront of one of the defining scientific pursuits of our time. It's a hunt that, if successful, could reshape the world's landscape, sending biofuel prices through the floor and allowing a drastic reduction in the country's oil use. It's a long campaign, and the opposition is all around.

It is a war, at its most fundamental, against plants.

For eons, plants have locked the sun's energy into complex strands of sugar, used to build their stems and leaves. These chains are far different from table sugar or grain starch; they cling together, providing the meat of tree trunks and cotton strands. They are the most

POPSCI — GADGETS / CARS / SCIENCE / TECHNOLOGY / DIY

Brilliant 10: Sludge Miner

SUNDAY INSIGHT
Archive | Feedback

Water flea vital to our survival

David Perlman
Sunday, February 13, 2011

SFGate
San Francisco Chronicle

The genes of ants held center stage in the world of science a week ago, and now it's the genes of water fleas, those tiny shrimp-like critters whose living bodies have long served as sentinels to warn of polluted water.

Almost every week these days, another organism reveals the sequence of its genes to scientists who are learning how specific genes and groups of genes in animals and plants can control their evolution, their structure, their behavior, and their ability to help or harm human health and the environment.

Scientists have decoded the genes of more than 125 different organisms - from bacteria, plants and insects, to mammals from mice to elephants.

We humans, it turns out, have more than 23,000 genes in our entire genome - 25,947 is the best guess - while the genome of the tiny water flea, surprisingly, is much larger - it holds 30,907 genes altogether, the gene sequencers say.

The increasing ability of scientists to learn how specific genes are expressed in normal or abnormal body shapes and sizes, and in the body's ability to resist or succumb to germs and other stresses, is one of the most significant advances of modern basic research.

The Washington Post — Make us your start page

POSTOPINIONS

Who needs government-funded research on a cow's stomach? We all do.

By Annalee Newitz, Published: November 4

Programs

In 2011, with a view to optimizing the alignment of scientific activities, the DOE JGI established two "super programs," under which the Institute organized its sequencing targets.

The Eukaryote Super Program

The Eukaryote Super Program comprises the Plant and Fungal Genomics Programs, each responsible for administrating and coordinating the sequencing and analysis of projects approved through the Community Sequencing Program (CSP) and Bioenergy Research Centers, and for developing sustained programmatic efforts coordinated with the appropriate user community, such as the Gene Atlas (for plants) and Genomic Encyclopedia of Fungi. The scientific aims of the Eukaryote Super Program are to produce, annotate, and analyze plant, fungal, and other eukaryotic genomes that are either directly relevant to the DOE mission, or that underlie fundamental biology related to this mission.

In the summer of 2011, the DOE JGI's sequencing efforts were grouped under two super programs: eukaryote and prokaryote. Each program has a scientific advisory board that meets several times a year to review program activities and provide guidance on prioritization of ongoing activities and new projects.

In the February 21, 2011 issue of *Proceedings of the National Academy of Sciences*, the 56 million base-pair genome of *Aureococcus anophagefferens*, better known as the cause of brown tide, was published. Algae play key roles in the global carbon cycle, helping sequester significant amounts of carbon. However, some algal species can bloom or become so numerous that they discolor coastal waters and reduce the amount of light and oxygen available in the ecosystem. In the case of brown tide, the algae outcompete with the other marine phytoplankton in the area, damaging the food chains in marine ecosystems as well as economically impacting the shellfish industry. Economic losses attributed to this and other harmful algal bloom phenomena in the United States over the course of the past decade are estimated at $1 billion.

In another genome project published on May 5, 2011 in *Science*, a team of researchers from more than 60 institutions reported the sequence of *Selaginella moellendorffii* and used a comparative

genomics approach to identify the core genes likely to be present in this common ancestor to land plants. The *Selaginella* genome fills a large gap in plant evolution, from the unicellular green alga *Chlamydomonas*, sequenced at the DOE JGI and published in 2007, to flowering plants with vascular systems and grasses, including *Brachypodium*, whose genome was published by the DOE JGI last year.

Fungi, as symbionts, pathogens, and biocontrol agents, dramatically affect plant health. Under the Fungal Genomics Program, new sequencing technologies and comparative genomics analysis combine to address large and complex sequencing projects such as the survey of the broad phylogenetic and ecological diversity of fungi and capturing genomic variation in natural populations and engineered strains.

For instance, in industrial applications, fungi provide a source of enzymes to catalyze such processes as generating biofuels from plant biomass. However, fungi may also pose challenges to biomass crop productivity. Rust plant pathogens that cannot survive on their own use crops as hosts, leading to reduced yields and potentially hindering efforts to grow biomass for fuel. The 101 million base-pair genome of *Melampsora larici-populina*, the first tree pathogen sequenced, was made publicly available in 2008. Rust outbreaks on poplar leaves weaken the trees, which are another candidate bioenergy feedstock whose genome sequence was published by the DOE JGI in 2007. In the *Proceedings of the National Academy of Sciences* published the week of May 2, 2011 the DOE JGI and the Broad Institute collaborated on their first joint fungal genomics

study, comparing two rust pathogens to reveal the role they play in infecting the host plant and acquiring nutrients.

Another fungal genomics publication focused on *Aspergillus niger*, used by industry for citric acid production. Harnessing this well-known fermentation process could inform the development of a biorefinery where organic compounds replace the chemical building blocks normally derived from petroleum. In the May 4, 2011 issue of *Genome Research*, the genome sequences of two *A. niger* strains, one of which was sequenced at the DOE JGI, were compared. Learning more about the genetic bases of the behaviors and abilities of these two industrially relevant fungal strains will allow researchers to exploit their genomes toward the more efficient production of organic acids and other compounds, including biofuels.

Still another fungal genome published July 14, 2011 in *Science* compared *Serpula lacrymans*, the second brown-rot fungus to have its genome sequenced, against 10 other published fungal genomes. The analysis not only allowed researchers to understand the chemical reactions involved in the mechanism by which *Serpula* breaks down cellulose, it also shed light on the role of brown-rot fungi in the development of the largest terrestrial ecosystem—the subarctic cool-climate boreal forest—and therefore the fungi's role in the global carbon cycle.

DOE supports several projects that focus on identifying enzymes from fungi and microbes—such as cellulases and heat-tolerant industrial-strength host-cell systems—for use in breaking down and

converting plant mass into sugars for biofuel production. In the *Proceedings of the National Academy of Sciences* published the week of July 25, 2011 a Great Lakes Bioenergy Research Center team partnered with the DOE JGI to sequence the genomes of two types of fungi that reside in the habitats of bark beetles. Woody biomass like bark contains a lot of xylose, and these fungi were well adapted to using this type of sugar to both grow and provide nutrients for the beetles. The team identified several new genes that improve yeast's
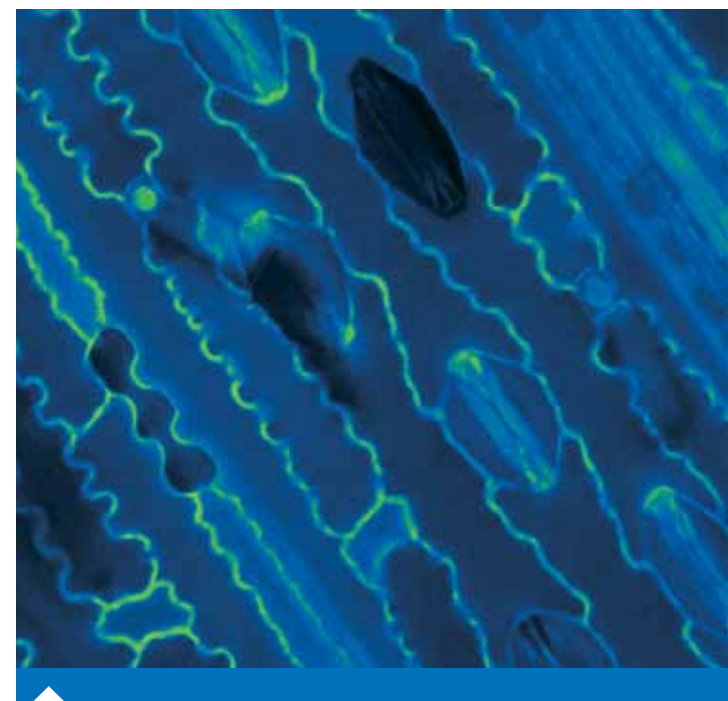
Micrograph of miscanthus leaf. *(Image provided by Mayandi Sivaguru and Steve Long, IGB Microscopy Facility and EBI-UIUC)*

ability to use xylose, a five-carbon sugar that can make up nearly half of available plant sugars. If researchers can coax yeast into using most of these sugars, they can improve the efficiency of producing renewable fuels from biomass crops like corn stover or switchgrass.

Many of the cellulases currently used in biofuel production are derived from species that thrive at temperatures of 20°C to 35°C (68°F to 95°F). The time needed to convert plant mass at these temperatures increases the possibility of contaminants, reducing the final yield. Speeding the conversion
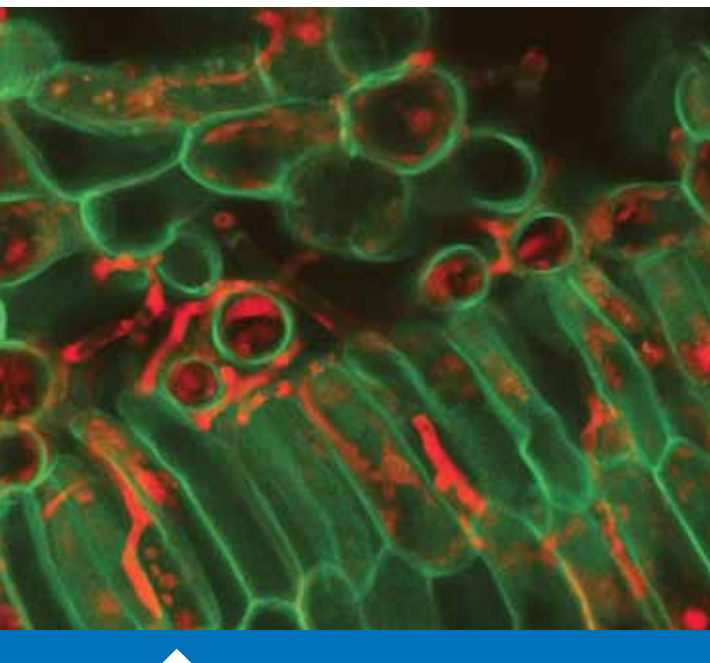
Confocal microscopy image of leaves infected with *M. larici-populina* from poplar. *(Stéphane Hacquard, INRA Nancy)*

process would require enzymes that are stable above current working conditions. Some of these enzymes may be found in thermophilic fungi such as and *Myceliophthora thermophila*, which thrive in high-temperature environments above 45°C. The finished genomes of these two fungi were published online October 2, 2011 in *Nature Biotechnology* by an international team of scientists, including DOE JGI researchers. Cellulases in these fungi are active at temperatures ranging from 40°C to 75°C, and could be useful for accelerating and thus improving the biofuel production process.

Last year saw the culmination of a 10-year collaboration between the DOE JGI and the *Daphnia* Genomics Consortium. The water flea, *Daphnia pulex*, is a keystone species in freshwater ecosystems that is used as a sentinel for environmental monitoring. The water flea has also been designated as a model system for biomedical research by the National Institutes of Health, joining established models like the fruit fly *Drosophila* and the worm *Caenorhabditis elegans*. The analysis of *Daphnia's* 200 million base genome was described in the February 4, 2011 issue of *Science*.

On the heels of the water flea genome release, the genome of another arthropod selected for sequencing by the DOE JGI was published in the November 24, 2011 issue of *Nature*. The two-spotted spider mite can extract needed nutrients from more than 1,000 plant species, including bioenergy feedstocks. Its publicly available, 90 million nucleotide genome is being used by researchers around the world to study novel pest-control strategies that could be used in place of chemicals. The spider

mite is also being considered by researchers for applications in pest-plant interaction studies and biomedicine.

Prokaryote Super Program

The Prokaryote Super Program is founded on four pillars for understanding of microbial life on Earth:

- More than 60 percent of the planet's biomass and biodiversity is microbial.

- Microbes control most of the major biogeochemical cycles on Earth and can therefore have significant impact on weather, climate, and the environment.

- The vast majority of microbes cannot, or have not yet been, cultured and thus cannot be studied in isolation.

- All life forms exist in a continuum of interdependent communities of organisms, in which microbial life is an integral part.

While the sequencing and assembly strategies for the Microbial Genomics and Metagenomics Programs are quite different, their scientific goals are very much aligned, with an underlying common objective to understand the structure and function of microorganisms and microbial communities with mission relevance to DOE. To achieve this, program scientists are working to sequence nucleic acids from both individual microbes and entire communities of organisms. This combined approach offers unique opportunities to the scientific community, but also requires coordinated activities between metagenomic

and individual microbial genome sequence generation and data analysis.

The Microbial Genomics Program is a key provider of high-quality bacterial and archaeal genomes and their analysis, supporting DOE user communities while aligning with DOE missions of clean bioenergy, carbon cycling, and biogeochemistry. The ongoing development of the Genomic Encyclopedia of Bacteria and Archaea (GEBA) initiative aims to systematically fill in the gaps in sequencing along the bacterial and archaeal branches of the tree of life. The success of the GEBA pilot project helped launch several related projects such as GEBA-Cyano (for the cyanobacterial tree of life), GEBA-RNB (for the root-nodulating bacteria), and others under way.

For the pilot GEBA study, the DOE JGI sequenced 53 bacterial and three archaeal novel and highly diverse genomes, representing a first step toward a phylogenetically balanced sequence space in the microbial tree of life. One of the archaeal genomes sequenced was used for testing a new class of solvents known as ionic liquids, which have been reported to be highly efficient in treating biomass and enhancing the yield of sugars from it. Ionic liquids can hinder the ability of the cellulases usually derived from fungi to produce sugars after pretreatment. In a study focused on identifying new enzymes that are tolerant of ionic liquids, first published June 30, 2011 in *Green Chemistry*, researchers from the DOE JGI and the Joint BioEnergy Institute at Berkeley Lab employed a cellulose-degrading enzyme from a salt-tolerant microbe that was isolated from the Great Salt Lake in Utah.

Another microbial publication focused on microorganisms found in the "twilight zone," located between 200 and 1,000 meters below the ocean surface. The microbes at these depths capture carbon dioxide, which they then use to survive and reproduce. Details are emerging, and in the September 2, 2011 issue of *Science*, researchers—including those from the DOE JGI and longtime collaborators at the Bigelow Laboratory Single Cell Genomics Center—described a microbial metabolic pathway that helps solve the mystery of how certain bacteria do this in the dark ocean and what happens to the carbon that is fixed in the oceans every year. In the September 2, 2011 issue of *Science*, the team relied on single-cell genomics to identify bacteria that fix carbon in the dark ocean, a task that until then had only been attributed to archaea.

Understanding the structure and function of microbial communities requires knowing the organisms that compose them, as well as their specific roles. To achieve that, we need to dissect each community and identify its individual components, an approach that directly depends on the availability of sequenced reference microbial genomes. Accordingly, it has become evident that the road to success in metagenomics is through microbial genomics.

The metagenome of the cow rumen was a massive-scale sequencing project conducted at the DOE JGI with support from the Energy Biosciences Institute. As described in the January 28, 2011 issue of *Science*, bioenergy researchers who study the rumen microbial community hope to harness the microbes' ability to efficiently break down plant biomass for use in large-scale biofuels production.

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

To identify microbes of interest in the cow rumen, researchers used samples of the candidate bioenergy feedstock switchgrass. More than a quarter-billion bases of sequence from the microbes that adhered to the switchgrass samples were generated for this project. In the analysis, they were able to identify 27,755 candidate genes that encoded carbohydrate-active enzymes (CAZymes), which can break down plant polysaccharides such as cellulose into small sugars.
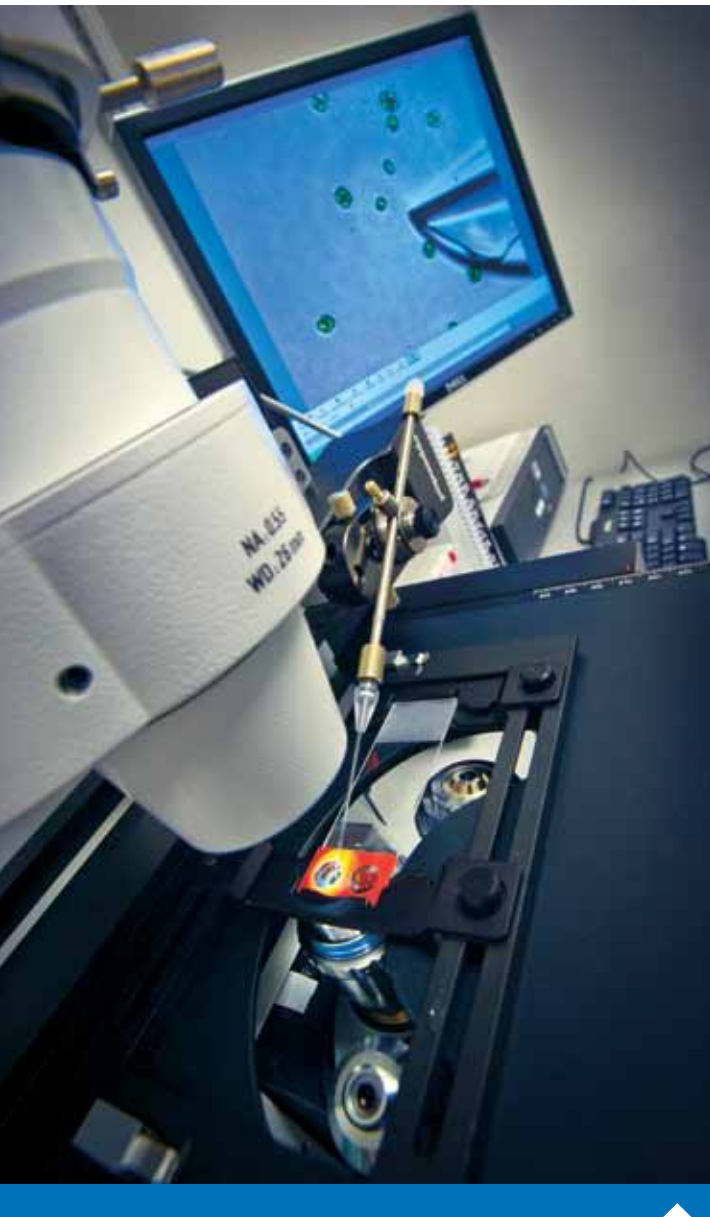
Micromanipulator for single-cell genomics.

All life forms exist in a continuum that forms interdependent communities of organisms. Comprising more than 60 percent of the planet's biodiversity and biomass, microbial life is an integral part of these communities.

This pioneering work illustrates how researchers can now explore a vast array of cellulose-degrading enzymes from the "microbial dark matter" that cannot be grown in culture, and addresses industry's need to develop better ways to break down biomass for a new generation of renewable biofuels.

Another metagenome project from the DOE JGI focused on the microbial communities inside the tammar wallaby, a plant-eating marsupial related to the kangaroo whose digestive system has been compared to that of ruminants such as cows and sheep. In the June 30, 2011 issue of *Science*, an international team of scientists, including researchers at Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO) and the DOE JGI, built on the wallaby rumen metagenome project, describing the genome of a bacterium found in the wallaby's gut that could

help explain why the methane emissions in these Australian marsupials are lower than in those of livestock. The information could be used to reduce the emission of methane (an even more potent greenhouse gas than carbon dioxide) from livestock not just in Australia but worldwide.

The year 2011 began with a large-scale metagenome project focused on the cow rumen and ended with one on the Arctic permafrost, which is believed to keep nearly 1,700 billion metric tons of carbon out of the Earth's atmosphere. Published November 6, 2011 in *Nature*, this project probed the possible impacts on the global carbon cycle when—due to rising global temperatures—the permafrost thaws and releases the carbon that has been trapped for eons. Researchers from the DOE JGI, the Berkeley Lab Earth Sciences Division, and the U.S. Geological Survey collaborated to

understand how microbes in permafrost respond to their warming environment.

The DOE JGI team generated nearly 40 billion bases of raw DNA sequence, necessary due to the high microbial diversity of the soil for this project. The team identified several microbes that produced methane as a byproduct, and successfully assembled, for the first time, a draft genome of a novel microbe that produces methane. The researchers also identified many genes involved in carbon and nitrogen cycling in the metagenomic data, and found that their levels of abundance shifted in response to a thawing habitat. These analyses conducted by the team provide support for the conceptual models of carbon and nitrogen cycling in Arctic soils, and underscore the importance of the microbial communities' response to the thawing permafrost and potential impact on the global carbon cycle.

Selaginella moellendorffii. (Jing-Ke Weng, Salk Institute)

Genomic Technologies

Six groups are focused on developing and effectively applying genomic technologies to accelerate DOE JGI users' science.

One of these groups focuses on production-sequencing capabilities so that the platforms in use provide an efficiently operating, low-cost, high-accuracy sequencing capacity for DOE JGI users. Over the past two years, the DOE JGI has consolidated primary sequencing operations onto the Illumina HiSeq 2000 platform, and is currently using eight of these instruments. Five older-generation Illumina GAIIx sequencers were phased out in September 2011, along with two Roche/454 FLX-Ti sequencers that had run in limited operation. Capillary-based Sanger sequencing was phased out completely in October 2010. Current developmental efforts focus on process optimization for increasing throughput and improving data quality, improving operational efficiencies, eliminating process waste, and ultimately driving down the cost of sequencing.

In 2011, the DOE JGI acquired and installed the second of its two new Pacific Biosciences (PacBio) RS single-molecule DNA sequencers. It is expected that long reads (over 2,000 bases) from the PacBio instrument will make a significant contribution to certain user projects in FY 2012, and PacBio reads primarily will be used to improve coverage of regions that are difficult to cover by Illumina. For a detailed description of the Illumina HiSeq and PacBio RS systems, see Appendix B.

The primary product from the Genomic Technologies Department continues to be the latest advances in high throughput DNA sequencing and analysis, coupled to robust sample management and library construction. Additional efforts are devoted to exploring new opportunities for technological access to DOE JGI user services, including the use of large-scale single-cell genomics to study hard-to-culture environmental microbes. An increasing number of post-sequencing capabilities are also being offered to users, including gene expression (RNA) profiles and metatranscriptomics as a function of natural or experimental perturbation, DNA fragment synthesis for functional characterization of genes identified in metagenomic data sets, and a broad diversity of customized data-analysis resources.

Single-Cell Genomics

At present, only a minute fraction of microbes can be cultured in the laboratory, presenting a substantial obstacle for exploring the biology of the vast majority of microbes. The uncultured majority includes large numbers of microbes relevant to energy and environmental applications. Culture-independent approaches such as metagenomics and, more recently, metatranscriptomics—the expressed subset of genes within a microbial community at a certain point in time—provide a path to understanding the uncultured microbial biosphere and to tackling many questions of DOE relevance.

However, most such techniques have considerable limitations for exploring individual species that are members of complex ecological communities. Emerging single-cell technologies provide a powerful complementary strategy to access the genetic makeup of individual uncultured community members, eliminating key challenges of metagenomic approaches, such as the proper assembly and binning of complex data sets.

Requests have rapidly increased for this capability—which was initially offered by the DOE JGI three years ago—from 14 genomes in 2009 to 320 in 2011. Large-scale single-cell analysis is expected to enable users to leverage the potential of these techniques for energy and environmental studies. Further method development will be aggressively pursued to mitigate the technical challenges that still limit the throughput of single-cell techniques. In particular, methods will be streamlined using micro- and nanofluidics approaches to increase sample

throughput by orders of magnitude. This will also enable cell preparations for complementary single-cell proteomics and metabolomics studies of the same specimens by users, in order to enable systems-level studies at single-cell resolution.

DNA Synthesis

To accelerate the linking of sequence to function, the DOE JGI is exploring the development of rapid and inexpensive approaches to designing and creating DNA fragments encompassing genes

and larger segments of DNA. These capabilities will be available to users for testing genomics-derived hypotheses, creating synthetic pathways, and for the functional exploration of metagenomic and other sequence data sets.

The most expensive and time-consuming step in large-scale synthetic DNA projects is the assembly of small oligonucleotides (short nucleic acid polymers) into larger fragments. The DOE JGI is at the forefront of implementing new technologies into the DNA synthesis pipeline.

The cell sorter separates single cells in suspension based on their expression of fluorescent reporter genes.

The primary product from the Genomic Technologies Department continues to be the latest advances in high-throughput DNA sequencing and analysis, coupled to robust sample management and library construction. Additional efforts are devoted to exploring new opportunities for technological access to DOE JGI user services, including the use of large-scale single-cell genomics to study hard-to-culture environmental microbes.

The Illumina HiSeqs generate the bulk of the sequence data at the DOE JGI.

Defining a set of sequences to be synthesized is a multistep process involving:

• The mining of available sequence repositories for "raw" sequences with desired characteristics, such as encoding proteins with desired catalytic/ structural properties or gene regulatory elements with desired response/activity profiles

• The computational optimization of individual sequences, taking into account the properties of the eventual host system that may require optimization of codons (a series of three adjacent bases that code for a specific amino acid), functionally neutral replacement of "prohibited" sequence motifs, and hypothesis-driven alterations to change the function of protein-coding or regulatory sequences

• Devising a synthesis and assembly strategy compatible with the target sequence, as well as the characteristics of the assembly/host system

The DOE JGI's primary role in synthetic DNA projects will be to support users in the computational design of desired target constructs, in the creation of these large and complex DNA molecules, and in their introduction into suitable host cells. In contrast, in-depth functional characterization of the resulting transformed host organisms will primarily rely on expertise and assays established in the respective users' laboratories. Nevertheless, to support users in their ability to generate synthetic systems required to address energy and environmental challenges, the DOE JGI will also develop experimental strategies in which functional readouts can be closely linked to synthetic sequence.

The PacBio RS sequencing system.

Informatics

Informatics supports three main areas of the DOE JGI's activities: sequencing project management, sequencing, and scientific programs. Specific science program informatics systems provide support for sequencing data processing, analysis, integration, and publication. Informatics systems seek to maintain high throughput (data sets processed weekly, monthly, and quarterly) and quality of service (reliability, robustness, and performance).

Hopper Cray XE6 supercomputer.

Science Programs Support

The DOE JGI is the global leader in generating genome sequences of plants, fungi, microbes, and metagenomes. As such, the genome sequence data processing and integration activities of its Science Programs—Plant, Fungal, Microbe, Metagenome— have a "production" nature, in which tools developed based on computational biology methods are applied on data sets within the context of program-specific informatics systems. The Institute's comparative analysis systems have matured over the past years and are recognized as important resources for conducting genome and metagenome studies, empowering scientists around the world to conduct studies that otherwise would be very expensive or out of reach. They allow users to analyze and improve the functional characterization of a vast number of publicly available genomes and metagenomes.

These data-management and comparative-analysis systems were featured in the database issue of *Nucleic Acids Research*, published online in November 2011.

One of the issue's featured articles, which the editors deemed to be in the top 5 percent in terms of originality, significance, and scientific excellence, covered the Genome Portal through which the DOE JGI's nearly 4,000 publicly available genomes and metagenomes can be accessed via a "Tree of Life" graphical catalog. Users also use it to navigate to the Institute's domain-specific comparative-analysis systems: MycoCosm (fungal genomes), Phytozome (plant genomes), IMG (microbial genomes), and

The DOE JGI's comparative analysis systems have matured over the past years and are recognized as important resources for conducting genome and metagenome studies worldwide, empowering scientists around the world to conduct studies that otherwise would be very expensive or out of reach.

IMG/M (metagenomes). MycoCosm, which was highlighted in the article, was released two years ago in response to requests for a central portal by which to access both fungal genomes and the tools for analyzing them. The DOE JGI Fungal Genomics Program alone aims to double sequencing and analysis throughput every year.

Since 2008, Phytozome has been updated eight times, most recently in January 2012. The Phytozome v.8 release included updates to flagship genomes *Brachypodium*, maize, and *Medicago* and links to the Gene Atlas pilot project. The updates also included newly released DOE JGI data sets for the common bean *Phaseolus* and the *Arabidopsis* comparator *Capsella rubella*, as well as externally generated genomes for the apple and strawberry.

Microbial and metagenomic genomes sequenced at the DOE JGI rely on the Integrated Microbial Genomes (IMG) system and are included into the

IMG Expert Review (IMG/ER) system for annotation review and comparative analysis. Since IMG's release in March 2005, the catalog has grown to nearly 7,000 genomes with more than 11 million genes. In November 2011, the Microbial Genomes with Microbiome Samples (IMG/M) system exceeded the 1 billion genes mark. Since January 2009, more than 930 metagenome sample data sets have been annotated using the IMG/M Expert Review annotation pipeline, with about 60 percent of these samples sequenced at the DOE JGI.

While the DOE JGI's Genome Portal offers users a way to track only the DOE JGI's ongoing projects, the Genomes OnLine Database (GOLD), launched in 1997, allows them to monitor genome and metagenome projects worldwide. With more than 11,000 projects documented, the GOLD database relies heavily on the use of genome standards and has pioneered the development and implementation of several standards for various data types.

Computational Infrastructure

The use and effectiveness of the DOE JGI's computing environment depends on the efficiency of the data-processing workflow and the organization underlying genome and metagenome data interpretation. Through a memorandum of understanding revised in 2011, NERSC provides high-performance and throughput computing support for the DOE JGI. Over the past two years, several large-scale assembly tasks capable of handling terabases of sequence data have been implemented by the Genome Analysis group of the Genomic Technologies Department. As of September 2011, about 7 million core hours on NERSC systems were used to support several genome- and metagenome-analysis large-scale computations and to develop parallel tools for next-generation genome and metagenome analysis. The DOE JGI's current core computing environment has grown steadily from a compute cluster of 120 nodes (eight cores, 32 GB memory per node) and 1.2 PB network-attached storage in December 2008, to a cluster of 610 nodes and 2 PB network-attached storage in September 2011. The DOE JGI relies on NERSC's Hopper—a Cray XE6 supercomputer—and Carver—an IBM iDataPlex system—both served by IBM's General Parallel File System (GPFS). NERSC also provides its High Performance Storage System (HPSS) for archiving the DOE JGI's sequence data sets.
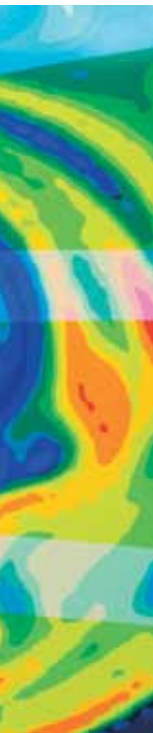
Education

The DOE JGI's faculty-development efforts stimulate the formation of a large network of collaborative faculty educators nationwide that become an important part of the User Community and play a key role in the expansion of the Education Program into functional genomics.

The JGI Science @ the Lesher panel discussion, held May 9, 2011 in downtown Walnut Creek, California, on "The Deal with Carbon: How Earth's Mighty Microbes Respond" focused on the role of microbes in the global carbon cycle. Berkeley Lab microbial ecologist Terry Hazen (far left), holder of the University of Tennessee-Oak Ridge National Laboratory Governor's Chair for Environmental Biotechnology, revisited his keynote talk at the 6th Annual DOE JGI User Meeting regarding microbes and the 2010 Deepwater Horizon oil spill. DOE JGI's Jonathan Eisen (far right) discussed the importance of studying microbes, touching briefly upon the ongoing GEBA project, while former DOE JGI postdoctoral fellow Rachel Mackelprang (center right), now at California State University-Northridge, spoke on the thawing Arctic permafrost as a potent source of greenhouse gases. KTVU Channel 2 Health and Science Editor John Fowler (center left) moderated the panel. Video of the event is available online at http://bit.ly/kgaThH.

Education and workforce training was identified as a long-term strategic goal in the 2010 DOE *Grand Challenges for Biological and Environmental Research: A Long-Term Vision*. Driven by this mandate, the evolution of the DOE JGI from a sequence-production facility to a next-generation genome center with advanced capabilities is accompanied by a focused effort in training future users.

The DOE JGI Education Program trains undergraduate and graduate faculty in incorporating genomics and bioinformatics into their life-science curricula. Faculty development is a key part of helping the United States maintain a competitive edge in science and technology; keeping teachers current on the latest in genomics-based research is an ongoing challenge. The DOE JGI's efforts in this area include workshops both at the Institute and at national scientific organization meetings such as the American Society of Microbiology. The goal is to form a large network of collaborative faculty educators nationwide who will become an important part of the DOE JGI user community and play a key role in the expansion of the DOE JGI Education Program into functional genomics. A systematic approach to functional genomics at the undergraduate level will allow students to take their bioinformatics-generated hypotheses and test them in the wet lab.

The Education Program is overseen by Cheryl Kerfeld, recipient of the 2011 American Society for Biochemistry and Molecular Biology Award for Exemplary Contributions to Education. Kerfeld was recognized for "encouraging effective teaching and learning of biochemistry and molecular biology through her own teaching, leadership in education,

writing, educational research, mentoring and public enlightenment."

Within the Education Program, the Undergraduate Research in Microbial Genome Annotation—or Interpret a Genome Program—provides college and university students access to recently sequenced microbial genomes, such as those of organisms from little-known branches of the tree of life selected as part of the DOE JGI's Genomic Encyclopedia of Bacteria and Archaea (GEBA) project. The students analyze and annotate the genomes in the context of their own coursework, gaining hands-on knowledge of genomics and bioinformatics.

As their annotation platform, students use the Integrated Microbial Genomes Annotation Collaboration Toolkit (IMG-ACT), a wiki/Web portal fusion that lets them work with existing genome data sets and record their discoveries. The platform is the result of a collaboration between the DOE JGI's Education Program and faculty members from several universities around the country.

Since the program launch in 2008, 4,145 students and 156 instructors from 80 institutions have used IMG-ACT. An additional 35 instructors from 18 institutions attended the January 2012 workshop at the DOE JGI headquarters in Walnut Creek, California.

Additional faculty training is supported by a National Science Foundation grant. IMG-ACT is in turn linked to other databases used in microbial genome annotation, including IMG/EDU, the educational version of the Integrated Microbial Genomes database widely used by researchers

in genome biology. A complementary metagenomics annotation tool, IMG-ACTM, developed by the DOE JGI Education Program, is being tested by a first group of instructors.

In parallel with the Education Program's efforts, the DOE JGI has been proactive in putting its computational tools in the hands of the burgeoning genomics community. Several times per year, the DOE JGI hosts five-day workshops on Microbial Genomics and Metagenomics. Built around the IMG system, these workshops are a community resource for comparative analysis and annotation of all publicly available genomes from three domains of life in a uniquely integrated context. Each workshop includes two days of intensive seminars and three days of hands-on tutorials. In addition, at its annual User Meeting and at other external venues, the DOE JGI holds tutorials focused on its other data repositories, including Phytozome plant genomics portal and MycoCosm fungal genomics portal. From 2009 through 2011, the DOE JGI sponsored more than 50 meetings for users to derive scientific insights from data generated by the Institute, drawing more than 2,500 participants.

Safety

The DOE JGI Safety Team has made significant operational efficiency improvements, resulting in a dramatically improved safety record. By building a stronger employee safety culture; increasing management feedback and involvement in safety; and improving existing safety programs, policies and procedures, workplace injuries and associated losses in productivity have dropped well below Bureau of Labor Statistics levels.

For instance, in FY 2008, there were 14 Occupational Safety and Health Administration (OSHA) recordable injuries, resulting in 27 lost workdays and 402 restricted workdays. In the three-year period that has followed, only five OSHA recordable injuries occurred, none resulting in lost or restricted workdays.

The data prove that employee-led safety is the most effective way to develop a mature safety culture and achieve an injury-free workplace. Several notable employee-led safety groups actively promote and support safety at the Walnut Creek headquarters. One of the most active is the Safety Culture Committee, composed of approximately 20 non-management employees who promote safety culture and provide safety-related feedback to management and the Employee Safety Committee. The group generates safety posters, conducts periodic safety-related surveys, sponsors an employee safety recognition program, sponsors safety theme months, and conducts annual safety

fairs. This group routinely submits an entry for the annual Applied Ergonomics Ergo Cup® competition.

In 2011, the DOE JGI entry to the internationally recognized competition was "It Is So Easy Being Green," a presentation of low/no cost solutions to reduce ergonomic risks. The recognition gained from this suite of tools builds on the DOE JGI's Ergo Cup® victories in 2010 for its entry "Empowering Employees in Ergonomics," and in 2007 for the "Shake and Plate" device.

The DOE JGI has shared its success in changing safety culture and creating a proactive program by contributing several best practices and other resources, which can be found at: http://www.jgi. doe.gov/whoweare/ergonomics/

Another important employee led-safety group is the Emergency Response Team (ERT). In the event of a major disaster, this group of 16 volunteer employees provides emergency response and first-aid services until professional responders can arrive The ERT also runs annual emergency evacuation drills, using Community Emergency Response Training (CERT), which is sponsored by the Federal Emergency Management Agency (FEMA), as the training model for its members. The CERT-based training program developed at the DOE JGI has been adopted by Berkeley Lab, with several DOE JGI ERT members serving as instructors.

The sonicator shears DNA fragments for library creation.

Appendix A: Glossary

Acronyms

ARRA
American Recovery and Reinvestment Act

ACT
Annotation Collaboration Toolkit

BESC
Bioenergy Sequencing Center (at ORNL)

BRC
Bioenergy Research Center

CERT
Community Emergency Response Training

CSIRO
Commonwealth Scientific and
Industrial Research Organisation

CSP
Community Sequencing Program

DD
Director's Discretionary [Program]

DNA
Deoxyribonucleic acid

DOE
Department of Energy

ERT
Emergency Response Team

FEMA
Federal Emergency Management Agency

GEBA
Genomic Encyclopedia of Bacteria and Archaea

GLBRC
Great Lakes Bioenergy Research Center

GOLD
Genomes OnLine Database

GPFS
General Parallel File System

HPSS
High-Performance Storage System

IAC
Informatics Advisory Committee

IMG
Integrated Microbial Genomes system

JBEI
Joint BioEnergy Institute

JGI
Joint Genome Institute

LANL
Los Alamos National Laboratory

LBNL
Lawrence Berkeley National Laboratory

LLNL
Lawrence Livermore National Laboratory

NASA
National Aeronautics and Space Administration

NERSC
National Energy Research Scientific
Computing Center

NREL
National Renewable Energy Laboratory

ORNL
Oak Ridge National Laboratory

OSHA
Occupational Safety and Health Administration

PacBio
Pacific Biosciences

PCR
Polymerase chain reaction

PI
Principal investigator

PNAS
*Proceedings of the National Academy
of Sciences*

PNNL
Pacific Northwestern National Laboratory

PSP
Proposal Study Panel

RNA
Ribonucleic acid

RNB
Root-nodulating bacteria

SAC
Scientific Advisory Committee

SMRT
Single-molecule real time

WFO
Work for others

WGS
Whole genome sequencing

ZMW
Zero-mode waveguide

Definitions

Annotation: The process of identifying the locations of genes in a genome and determining what those genes do to improve accuracy of genetic information collected.

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

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

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

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

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

Bioremediation: The use of microorganisms to break down contaminants and other unwanted substances in waste and other substances.

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

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

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

Carbon cycle: The biogeochemical process by which carbon is exchanged between the planet's atmosphere, land, and oceans.

CAZymes: Carbohydrate-Active enzymes that can break down plant polysaccharides such as cellulose into small sugars.

Cloning: The use of 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: A group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome.

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

Curation: Analysis of genome annotations to improve and maintain data presentation.

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

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

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

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

Fluorescence-activated cell sorting: A specialized type of flow cytometry used to study and purify cells. A heterogeneous mixture of cells passes through laser beams and is sorted into two or more containers, one cell at a time, based upon the specific light-scattering and fluorescent characteristics of each cell.

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

GenBank: Open-access, publicly available collection of annotated sequences submitted by individual laboratories and large-scale sequencing centers that is overseen by the National Center for Biotechnology Information.

Halophile: A microbe that thrives in environments with high salinity.

Informatics: The study of the science of information.

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

Mapping: Charting the location of genes on chromosomes.

Metagenomics (also environmental genomics or community genomics): The study of genomes recovered from environmental samples through direct methods that do not require isolating or growing the organisms present. This field of research allows the genomic study of organisms that are not easily cultured in a laboratory.

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

Microbe: Another name for a microorganism.

Microbiome: A defined environment within which a community of microbes exists and interacts.

Multiple displacement amplification (MDA): Method of amplifying tiny amounts of DNA in a cell so that it can be used for sequencing through single-cell genomics.

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

Polymerase chain reaction (PCR): A method of DNA amplification.

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

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

Psychrophile: A cold-loving microbe that optimally grows in environments with temperatures of 15°C (60°F) or less.

Read length: The number of nucleotides tabulated by the DNA analyzer per DNA reaction well.

Rhizosphere: Microecosystem defined by a thin layer of soil where plant roots interact with microorganisms in the soil.

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

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

Single-cell genomics: Method for sequencing a genome using DNA derived from a single cell that is used to study uncultured or nonculturable organisms.

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

Synthetic biology: For the DOE JGI's objectives, this process refers to assembling DNA sequence fragments with the goal of synthesizing sequences to experimentally validate their functions and applications.

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

Appendix B: Sequencing Technologies
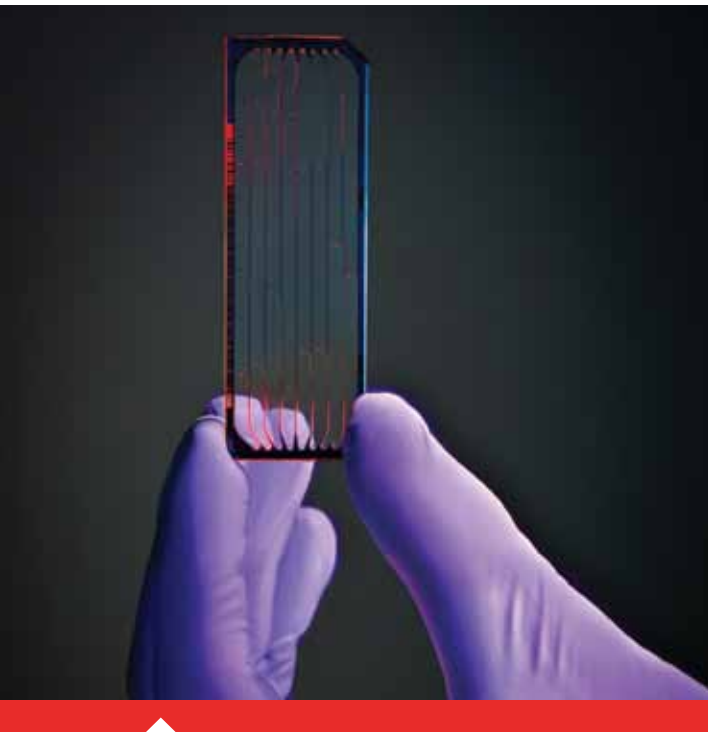

Illumina HiSeq flow cell.

The adoption of the Illumina and Roche/454 next-generation technologies allowed the DOE JGI to significantly increase sequencing throughput over the past three years. Roche/454 technology in 2008 provided a threefold increase in sequence data generation as compared with the Sanger platform in 2007, and adoption of the Illumina technology and subsequent technology improvements on the platform allowed for an eightfold increase in 2009, an additional sixfold increase in 2010, and a further fivefold increase in 2011. The DOE JGI has consolidated primary sequencing operations onto the Illumina HiSeq 2000 platform and has installed a second PacBio RS single-molecule DNA sequencer.

Illumina HiSeq Sequencing Technology

The Illumina approach relies on attaching fragmented genomic DNA prepared in a sample library to a planar, optically transparent surface on a flow cell. These templates are sequenced using a four-color DNA sequencing-by-synthesis technology that employs reversible terminators with removable fluorescence. This highly parallel approach can generate 325 billion bases (gigabases) per 2x150 flow cell run.

Labeled nucleotides are incorporated at each cycle and high-sensitivity fluorescence detection is achieved using laser excitation and total internal reflection optics. Images are compiled and processed to produce base sequences for each DNA template. Applications are de novo sequencing where there is no reference available, and resequencing where short sequence reads are aligned against a reference. The genetic differences on the sequences are called using a specially developed data pipeline.
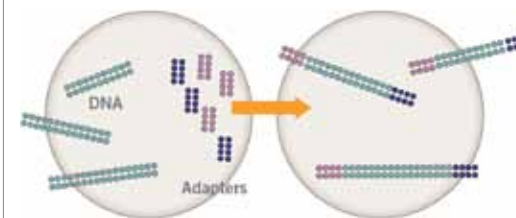
1 Sonication
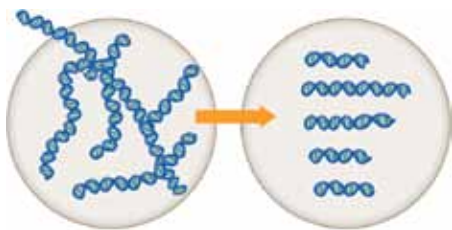
Genomic DNA is fragmented into 100-500 base-pair fragments by sonication to create a library.

2 Fragment End Repair
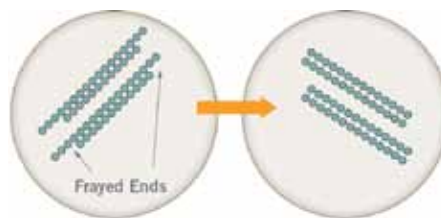
Sonication creates frayed DNA ends that must be blunted or repaired.

3 A-Tailing and Adapter Ligation

Adapters are ligated to each end of the A-tailed DNA fragment.

4 QC Check
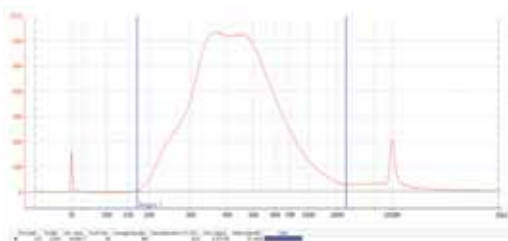
The electropherogram shows the size and concentration of the final library. This library size also confirms the ligation of adapters.

5 cBOT Cluster Generation System
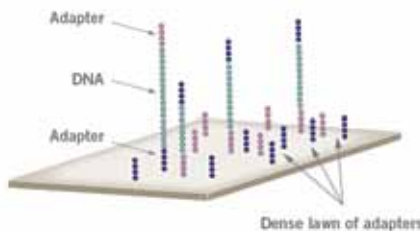
Sodium hydroxide creates single-stranded DNA that is then randomly bound to the top and bottom of each channel in the flow cell.

6 Bridge Formation

Free DNA end binds to complementary primer to form a bridge.

7 Bridge Amplification

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification. Fragments become double-stranded DNA bridges. Thirty-five cycles of amplification create clusters of identical DNA fragments.

8 Finished Flow Cell

By completion of amplification, several million dense clusters of single-stranded DNA have been generated in each channel of the flow cell, with a sequencing primer attached.
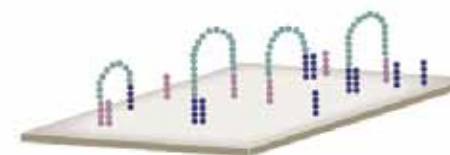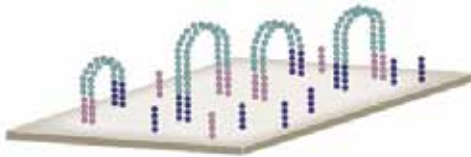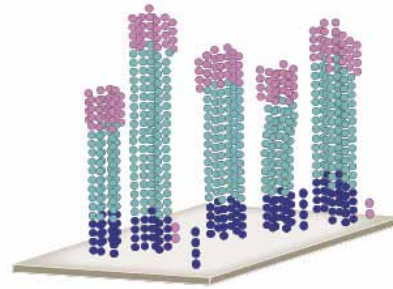
9 DNA Sequencing
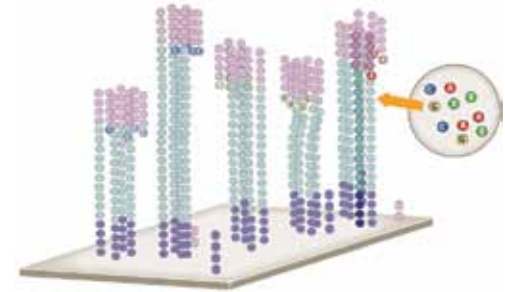
To initiate the first sequencing cycle and determine the first base, all four labeled reversible terminators and DNA polymerase enzyme are first added. Only one base can incorporate at a time.

10 Base Calling
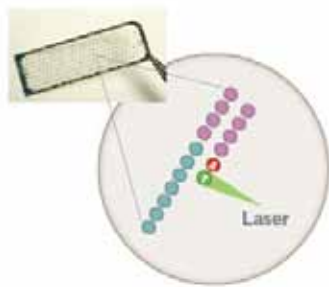
Lasers excite the fluorescent tags and the images are captured via CCD camera. The identity of the first base in each cluster is recorded, then the fluorescent tag is removed.

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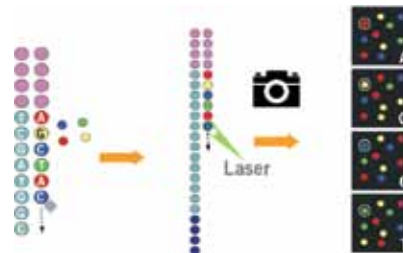
In the first cycle, the first base is incorporated. Its identity is determined by the signal given off and then recorded. In subsequent cycles, the process of adding sequencing reagents, removing unincorporated bases, and capturing the signal of the next base to identify is repeated.

12 Dual Flow Cells

Once the top surface of the flow cell channel has been scanned, the imaging step is repeated on the bottom surface.

In the past year, the DOE JGI has also acquired two MiSeqs, Illumina's latest instrument, which offer a smaller-scale, cost-effective alternative rapid-sequencing platform capable of performing 2x150 in runs in 27 hours as compared with the 16.5 days it takes on the HiSeq. While the MiSeq only generates a fraction of the data of a HiSeq—about 1 Gb on a MiSeq flow cell versus 300 Gb on a HiSeq flow cell—the rapid run times make it beneficial for applications where minimal sequencing reads are needed or when sequencing data are needed quickly.

PacBio Sequencing Technology

The Pacific Biosciences single-molecule real-time (SMRT™) DNA sequencer monitors the enzyme DNA polymerase as it attaches to a strand of DNA, examines the base at the point of attachment, and determines which nucleotide is required to replicate the base. With the aid of proprietary phospholinked nucleotides and a zero-mode waveguide to track the events at the nanoscale level, researchers can study variations at a structural and cellular level. Other applications include transcription, RNA sequencing, and translation. The sequencer allows templates to be made without polymerase chain reaction (PCR)

amplification and can generate reads that are thousands of bases long. This approach currently yields up to 20-35 million bases per SMRT cell and the instrument has an option to load multiple SMRT cells in single run.
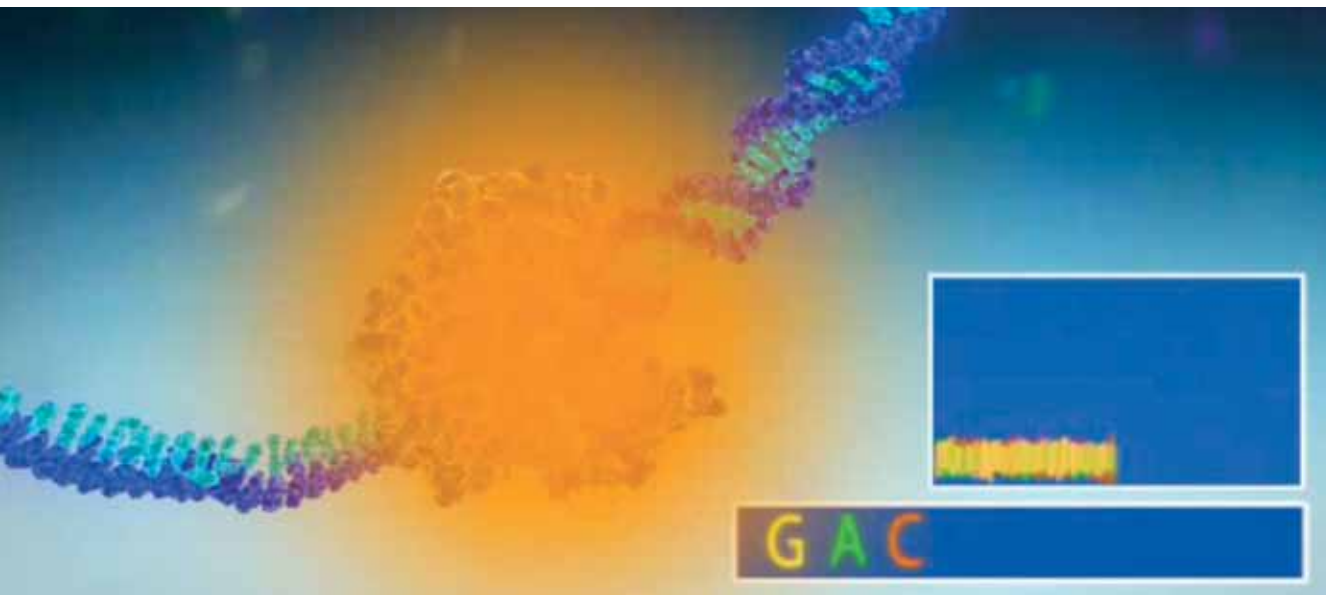
The SMRT sequencer relies on PacBio's RS DNA Template Preparation Kit to convert sample DNA into the proprietary SMRTbell™ library format for single-molecule real-time sequencing. The SMRTbell DNA template preparation method creates a unique, structurally linear, and topologically circular DNA morphology.


Illumina MiSeq

PacBio sequencer

1. DNA sample prep is done away from the instrument and requires 500 nanograms (ng) of starting material. The starting DNA is sheared into double-stranded linear structures with sizes ranging from 200 base pairs to 10 kilobase pairs, and then attached to the SMRT adapters, which produce a topologically closed circle, enabling consensus sequencing of the same template. The front of the machine contains two drawers for sample loading—one for DNA and reagents, the other for up to 96 SMRT cells.

2. SMRT Cell: Each SMRT Cell is patterned with 150,000 zero-mode waveguides (ZMWs) measuring 100 nm across, and each ZMW contains a single DNA polymerase. The ZMW is the window through which DNA sequencing can be monitored in real time. The PacBio RS system continuously monitors ZMWs in sets of 75,000 at a time. Each SMRT cell can be run in minutes.
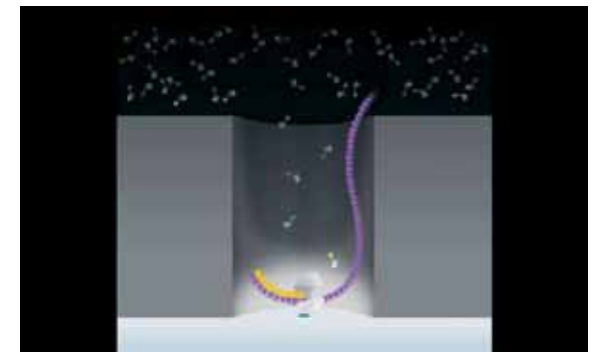
3. SMRT Sequencing: When an active polymerase is immobilized at the bottom of each ZMW, nucleotides diffuse into the chamber. Each of the four nucleotides is tagged with fluorescent markers on the terminal phosphate, not the base. Since only the bottom 30 nm of the ZMW is illuminated, only those nucleotides near the bottom fluoresce. When the correct nucleotide is detected by the polymerase, it is incorporated into the growing DNA strand in a process that takes milliseconds.

4. Base Calling: Four light-sensitive cameras collect the pulses emitted by fluorescent tags, allowing the observation of biological processes. Algorithms then translate the information captured by the optics system and convert the light pulses into either an A, C, G, or T base call. A consensus sequence can then be assembled by aligning the different fragments from each ZMW based on common sequences. A technique known as "strobe sequencing," in which the lasers in a sequencer are turned on and off during a run, can increase the effective generated read length.

Images ©Pacific Biosciences

Appendix C: FY2012 CSP Projects

The 2012 Community Sequencing Program (CSP) call invited researchers to submit proposals for projects that advance capabilities in fields such as plant-microbe interactions, microbes involved in carbon capture and greenhouse gas emission, and metagenomics—the characterization of complex collections of microbes from particular environmental niches. The total allocation for the CSP 2012 portfolio will exceed 30 Tb, a hundredfold increase compared with just two years ago, when just a third of a terabase was allocated to more than 70 projects. This amounts to the equivalent of at least 10,000 human genomes in data.

As an indicator of the increasing use of sequencing to study whole biological systems rather than individual organisms, more than half the approved proposals include sequencing of multi-organism samples either instead of or in addition to individual genomes.

Proposer	Affiliation	Project Description
Acinas, Silvia	ICM-CSIC, Spain	Microbial metagenomics and transcriptomics from a global deep-ocean expedition
Andresson, Olafur	University of Iceland	Sequencing of the three cultured partners of the lichen *Lobaria pulmonaria* and the sequencing of the transcriptomes from the natural tripartite lichen under selected and controlled conditions
Banfield, Jill	University of California, Berkeley	Terabase sequencing for comprehensive genome reconstruction to assess metabolic potential for environmental bioremediation
Brodie, Eoin	Berkeley Lab	Mediterranean Grassland Soil Metagenome (MGSM): Enabling a systems view of soil carbon and nitrogen biogeochemistry under a changing climate
Brutnell, Thomas	Boyce Thompson Institute for Plant Research	Development of sequence-based community tools for *Setaria viridis*—a model genetic system for C4 grasses
Bucking, Heike	South Dakota State University	Exploring the transcriptome of perennial grasses in association with beneficial microorganisms to increase biomass production and environmental sustainability of bioenergy production
Cary, Stephen	University of Delaware	Understanding terrestrial microbial biocomplexity in an Antarctic desert landscape: resolving universal drivers of community structure and function in a trophically simple system
Crouch, Jo Anne	USDA-ARS	Genomic signatures of pathogenicity and endophytism in five species of grass-associated *Colletotrichum* impacting the health and production of bioenergy feedstocks, agriculture, and the environment

Proposer	Affiliation	Project Description
Dangl, Jeff	University of North Carolina at Chapel Hill	Plant-associated metagenomes: microbial community diversity and host control of community assembly across model and emerging plant ecological genomics systems
DeAngelis, Kristen	University of Massachusetts	Microbial ecology and genomics of carbon-storing bacteria in rhizosphere soils
Dubilier, Nicole	Max Planck Institute for Marine Microbiology, Germany	Understanding novel pathways for energy and carbon use in bacterial symbionts of gutless marine worms
Emerson, David	Bigelow Laboratory for Ocean Sciences	Single-cell genome sequencing of biomineralizing bacteria
Fierer, Noah	University of Colorado	Cross-site metagenomic analyses to assess the impacts of experimental nitrogen additions on belowground carbon dynamics
Fredrickson, Jim	Pacific Northwest National Laboratory	Microbial interactions in extremophilic mat communities
Gilbert, Jack	Argonne National Laboratory	Creating a successional model for carbon remediation in the Gulf of Mexico
Gross, Stephen	DOE JGI	The agave microbiome: exploring the role of microbial communities in plant adaptations to desert environments
Hazen, Samuel	University of Massachusetts	Creating a multifunctional library of grass transcription factors for the energy crop model system *Brachypodium distachyon*
Hess, Matthias	Washington State University	Expression profile of biomass-degrading fungi inhabiting the cow rumen
Kelly, William	AgResearch, New Zealand	The Hungate 1000. A catalog of reference genomes from the rumen microbiome

Proposer	Affiliation	Project Description
Kerfeld, Cheryl	DOE JGI	Enhancing bacterial carbon capture and sequestration: synthesis of building blocks for the carboxysome, a metabolic module for CO_2 fixation
Kyrpides, Nikos	DOE JGI	Genomic Encyclopedia of Type Strains, Phase I: the 1,000 Microbial Genomes (KMG) project
Laplaze, Laurent	Institut de Recherche pour le Developpement (IRD), France	Transcriptome analysis of salt tolerance in *Casuarina* trees
Martin, Francis	INRA, France	Metatranscriptomics of soil forest ecosystems
McKay, Robert	Bowling Green State University	Metagenomics and metatranscriptomics of the Lake Erie "dead zone": a seasonal source of greenhouse gases
McMahon, Katherine	University of Wisconsin, Madison	Dynamics of microbial carbon processing pathways across a decade in a freshwater eutrophic lake revealed through metagenomic sequencing
Mock, Thomas	University of East Anglia, UK	Sea of change: eukaryotic phytoplankton communities in the Arctic Ocean
Mohn, William	University of British Columbia, Canada	Metagenomic and metatranscriptomic analysis of forest soil communities across North America
Moran, Mary Ann	University of Georgia	The genetic basis for heterotrophic carbon processing in the sea
Murray, Alison	Desert Research Institute	Lake Vida brine microbial community (LVBMCo) genomics and transcriptomics—a window into diversity, adaptation, and processes in extreme cold
Muyzer, Gerard	Delft University of Technology, Netherlands	Genome sequencing of 100 strains of the haloalkaliphilic chemolithoautotrophic sulfur-oxidizing bacterium *Thioalkalivibrio*

Proposer	Affiliation	Project Description
Nealson, Kenneth	University of Southern California	Life at the edge: community cooperation and success in a very extreme (ultrabasic and ultrareducing) environment
Ohm, Robin	DOE JGI	Toward functional genomics: development of *Schizophyllum commune* as a model system to study lignocellulose degradation
Pester, Michael	University of Vienna, Austria	Targeted metagenomics and metatranscriptomics of a sulfate-reducing rare biosphere member and potentially novel sulfate reducers that impact methane emission from peatlands
Powell, Amy	Sandia National Laboratories	A phylogenomic framework to investigate fungal thermophily
Pukkila, Patricia	University of North Carolina at Chapel Hill	Functional genomics in the model mushroom *Coprinopsis cinerea*
Rodrigues, Jorge	University of Texas at Arlington	Profiling metagenomic consequences of Amazon deforestation at different spatial scales
Schadt, Christopher	Oak Ridge National Laboratory	Defining the *Populus* microbiome: role of genotype by environment interactions in shaping the rhizosphere microbiome of *Populus trichocarpa*
Schrenk, Matthew	East Carolina University	Metagenome-enabled investigations of carbon and hydrogen fluxes within the serpentinite-hosted subsurface biosphere
Spatafora, Joseph	Oregon State University	1,000 fungal genomes
Stepanauskas, Ramunas	Bigelow Laboratory for Ocean Sciences	Dark ocean microbial single-cell genomics
Wing, Rod	University of Arizona	Empowering functional plant genomics with genomes and transcriptomes of the top 20 *Brassicales*
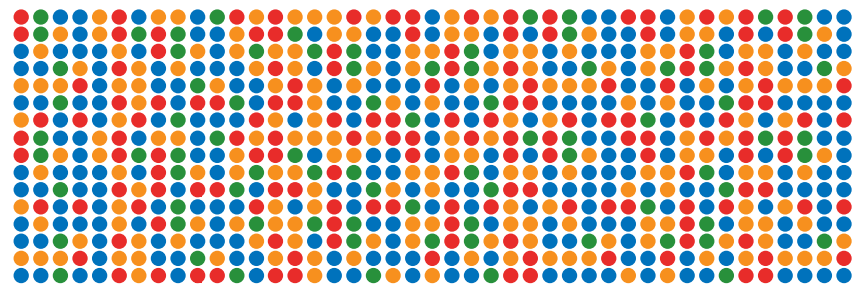
Appendix D: Advisory and Review Committee Members

The Scientific Advisory Committee (SAC)

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

Members

Bruce Birren (Chair), Broad Institute

Toby Bloom, Broad Institute

Jeff Dangl, University of North Carolina

Joe Ecker, The Salk Institute for Biological Studies

Jim Krupnick, Lawrence Berkeley National Laboratory

Eric J. Mathur, SG Biofuels

Nancy Moran, Yale University

Julian Parkhill, The Sanger Institute (UK)

Doug Ray, Pacific Northwest National Laboratory

James Tiedje, Michigan State University

Alexandra Z. Worden, Monterey Bay Aquarium Research Institute

CSP Proposal Study Panel (PSP) Members

Nina Agabian, University of California, San Francisco

Chris Amemiya, Benaroya Research Institute at Virginia Mason

Gary L. Andersen, Lawrence Berkeley National Laboratory

Jo Ann Banks, Purdue University

John Battista, Louisiana State University

Fred Brockman, Pacific Northwest National Laboratory

Zac Cande, University of California, Berkeley

Patrick Chain, Los Alamos National Laboratory

Jonathan C. Cohen, UT Southwestern Medical Center

Nigel Dunn-Coleman, Genencor International

Joe Ecker, The Salk Institute for Biological Studies

Katrina Edwards, Woods Hole Oceanographic Institution

Kelly Frazer, Perlegen Sciences, Inc.

Richard Harland, University of California, Berkeley

Derek Lovley, University of Massachusetts

David Mills, University of California, Davis

Alison Murray, Desert Research Institute

Nipam Patel, University of California, Berkeley

Karin Remington, National Institute of General Medical Sciences

Arend Sidow, Stanford University

John Taylor, University of California, Berkeley

Naomi Ward, The Institute for Genomic Research

Bart Weimer, Utah State University

DOE JGI *Ex-Officio* Members

James Bristow (PSP Chairman), DOE Joint Genome Institute

Daniel Rokhsar, DOE Joint Genome Institute

Eddy Rubin, Director, DOE Joint Genome Institute

DOE Representative

David Thomassen, U.S. Department of Energy

The Informatics Advisory Committee (IAC)

Adam Arkin, Division Director, Physical Biosciences Division, Lawrence Berkeley National Laboratory

David Dooling, Assistant Director, Genome Center, Washington University St. Louis

Saul Kravitz, Principal Systems Engineer, Center for Connected Government MITRE

Stan Letovsky, Vice President and Chief Informatics Officer, SynapDx

Jill Mesirov (IAC Chair), Associate Director and Chief Informatics Officer, Broad Institute

Granger Sutton, Senior Director of Informatics, J. Craig Venter Institute

Kathy Yelick, Associate Laboratory Director, Computing Sciences, Lawrence Berkeley National Laboratory

Plant Program User Advisory Committee Members

Jeff Dangl, University of North Carolina

Joe Ecker, The Salk Institute for Biological Studies

Eva Huala, Carnegie Institute/TAIR

Sabeeha Merchant, University of California, Los Angeles

Thomas Mitchell-Olds, Duke University

Stephen Moose, University of Illinois

Gary Stacey, University of Missouri

Fungal Program User Advisory Committee Members

Scott Baker, Pacific Northwestern National Laboratory

Joan Bennett, Rutgers University

Randy Berka, Novozymes

Daniel Cullen, Forest Products Laboratory

Ronald De Vries, CBS (Netherlands)

Daniel Eastwood, Warwick University (UK)

Audrey Gasch, Great Lakes Bioenergy Research Center

Stephen Goodwin, Purdue University

Joseph Heitman, Duke University

David Hibbett, Clark University

Thomas Jeffries, USDA Forest Service Forest Products Laboratory

Gert Kema, Plant Research International (Netherlands)

Christian Kubicek, Vienna University of Technology (Austria)

Cheryl Kuske, Los Alamos National Laboratory

Francis Martin, INRA (France)

Conrad Schoch, National Center for Biotechnology Information

Joseph Spatafora, Oregon State University

John Taylor, University of California, Berkeley

Adrian Tsang, Concordia University (Canada)

Gillan Turgeon, Cornell University

Rytas Vilgalys, Duke University

Prokaryote Super Program Advisory Committee Meeting Members

Cameron Currie, University of Wisconsin

Ed DeLong, MIT

Jed Fuhrman, University of Southern California

George Garrity, Michigan State University

Steve Hallam, University of British Columbia, (Canada)

Phil Hugenholtz, University of Queensland (Australia)

Bob Landick, Great Lakes Bioenergy Research Center

Folker Meyer, Argonne National Laboratory

Mary Ann Moran, University of Georgia

Nancy Moran, Yale University

Karen Nelson, J. Craig Venter Institute

Rich Roberts, New England BioLabs

Doug Rusch, J. Craig Venter Institute

Ramunas Stepanauskas, Bigelow Laboratory for Ocean Sciences

Niels van der Lelie, RTI

Terry Hazen

Persis Drell

Keynote Speakers:

Persis Drell, Director of the SLAC National Accelerator Laboratory, delivered a keynote speech on how SLAC has retooled to offer new resources and applications for a diversifying community of collaborators. The talk was timely as the DOE JGI transitions from a single service to offering multiple applications due to its genome-sequencing portfolio and expanding data sets.

Terry Hazen, a microbial ecologist at the Berkeley Lab and currently University of Tennessee-Oak Ridge National Laboratory Governor's Chair for Environmental Biotechnology, delivered a keynote speech on the research done by his team after the Deepwater Horizon oil spill on April 20, 2010. The team tracked the thousands of bacterial and archaeal species using a DNA-based array developed by Berkeley Lab, and took advantage of the DOE JGI's single-cell genomics expertise to identify a new "oil-seeking" species related to *Oceanospirillales*.

Learn more about the Meeting talks at http://1.usa.gov/UM6agenda

**Other Featured Speakers
(in order of appearance):**

Gene Robinson, University of Illinois
at Urbana-Champaign

Peer Bork, European Molecular Biology Laboratory,
Heidelberg (Germany)

Jerry Tuskan, Oak Ridge National Laboratory

Pam Silver, Harvard University

Rob Knight, University of Colorado

Len Pennacchio, Deputy Director, DOE JGI

Eddy Rubin, Director, DOE JGI

Jim Bristow, Deputy Director, DOE JGI

Scott Hodges, University of California, Santa Barbara

Mike Thomashow, Michigan State University

Stephan Schuster, Penn State University

Ruth Ley, Cornell University

Ed Buckler, USDA-ARS, Cornell University

Mary Ann Moran, University of Georgia

Dan Distel, Ocean Genome Legacy

Christopher Scholin, Monterey Bay Aquarium
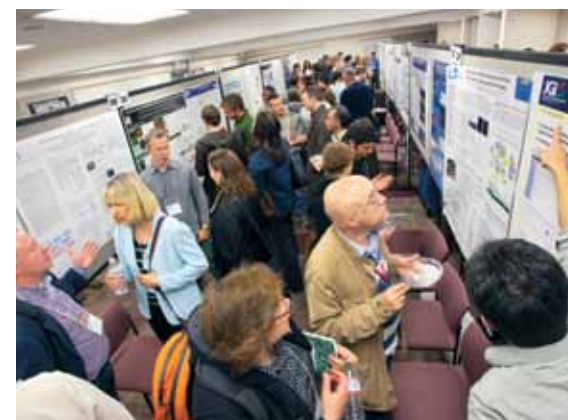Research Institute

Kathy Yelick, Lawrence Berkeley
National Laboratory

Tom Juenger, University of Texas at Austin

Magnus Nordborg, Gregor Mendel Institute,
Vienna (Austria)

Jim Tiedje, Michigan State University

Zhong Wang, DOE JGI

Abraham P et al. Defining the boundaries and characterizing the landscape of functional genome expression in vascular tissues of *Populus* using shotgun *Proteomics. J Proteome Res.* 2012 Jan 1;11(1):449-60. Epub 2011 Nov 9.

Abt B et al. Complete genome sequence of *Cellulophaga algicola* type strain (IC166). *Stand Genomic Sci.* 2011 Feb 22;4(1):72-80.

Abt B et al. Complete genome sequence of *Leadbetterella byssophila* type strain (4M15). *Stand Genomic Sci.* 2011 Mar 4;4(1):2-12.

Alverson AJ et al. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. *Plant Cell.* 2011 Jul;23(7): 2499-513. Epub 2011 Jul 8.

Andersen MR et al. Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Res.* 2011 May 4.

Anderson I et al. Novel insights into the diversity of catabolic metabolism from ten haloarchaeal genomes. *PLoS One.* 2011;6(5):e20237. Epub 2011 May 25.

Anderson I et al. Complete genome sequence of *Nitratifractor salsuginis* type strain (E9I37-1). *Stand Genomic Sci.* 2011 Jul 1;4(3):322-30.

Anderson I et al. Complete genome sequence of the hyperthermophilic chemolithoautotroph *Pyrolobus fumarii* type strain (1A). *Stand Genomic Sci.* 2011 Jul 1;4(3):381-92.

Anderson I et al. Complete genome sequence of *Methanothermus fervidus* type strain (V24S). *Stand Genomic Sci.* 2010 Nov 20;3(3):315-24.

Anderson I et al. Complete genome sequence of *Ferroglobus placidus* AEDII12DO. *Stand Genomic Sci.* 2011 Oct 15;5(1):50-60. Epub 2011 Oct 1.

Baker SE et al. Phylogenomic analysis of polyketide synthase-encoding genes in *Trichoderma. Microbiology.* 2012 Jan;158(Pt 1):147-54. Epub 2011 Nov 17.

Banks JA et al. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science.* 2011 May 5.

Beckham GT et al. Harnessing glycosylation to improve cellulase activity. *Curr Opin Biotechnol.* 2011 Dec 18.

Berg Miller ME et al. Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. *Environ Microbiol.* 2011 Oct 17. doi: 10.1111/j.1462-2920.2011.02593.x.

Berka RM et al. Comparative genomic analysis of the thermophilic biomass-degrading fungi *Myceliophthora thermophila* and *Thielavia terrestris. Nat Biotechnol.* 2011 Oct 2. doi: 10.1038/nbt.1976.

Blumer-Schuette SE et al. Complete genome sequences for the anaerobic, extremely thermophilic plant biomass-degrading bacteria *Caldicellulosiruptor hydrothermalis*, *Caldicellulosiruptor kristjanssonii, Caldicellulosiruptor kronotskyensis, Caldicellulosiruptor owensensis*, and *Caldicellulosiruptor lactoaceticus. J Bacteriol.* 2011 Jan 7.

Boden R et al. Complete genome sequence of the aerobic marine methanotroph *Methylomonas methanica MC09. J Bacteriol.* 2011 Dec; 193(24):7001-2.

Brambilla E et al. Complete genome sequence of *Methanoplanus petrolearius* type strain (SEBR 4847). *Stand Genomic Sci.* 2010 Oct 27;3(2):203-11.

Bresson A et al. Qualitative and quantitative resistances to leaf rust finely mapped within two nucleotide-binding site leucine-rich repeat (NBS-LRR)-rich genomic regions of chromosome 19 in poplar. *New Phytol*. 2011 Jun 10. doi: 10.1111/j.1469-8137.2011.03786.x.

Brown SD et al. Complete genome sequence of the haloalkaliphilic, hydrogen-producing bacterium *Halanaerobium hydrogeniformans*. *J Bacteriol*. 2011 Jul;193(14):3682-3. Epub 2011 May 20.

Brown SD et al. Genome sequence of mercury-methylating *Desulfovibrio desulfuricans* ND132. *J Bacteriol*. 2011 Apr;193(8):2078-9.

Brumm P et al. Functional annotation of *Fibrobacter succinogenes* S85 carbohydrate active enzymes. *Appl Biochem Biotechnol*. 2011 Mar;163(5):649-57. Epub 2010 Aug 28.

Cai B et al. MicroSyn: A user friendly tool for detection of microsynteny in a gene family. *BMC Bioinformatics*. 2011 Mar 18;12(1):79.

Chábera P et al. Excited-state properties of the 16kDa red carotenoid protein from *Arthrospira maxima. Biochim Biophys Acta*. 2011 Jan;1807(1):30-5. Epub 2010 Sep 8.

Chain PS et al. Genomics for key players in the N cycle from Guinea pigs to the next frontier. *Methods Enzymol*. 2011;496:289-318.

Chain PS et al. Genome of *Ochrobactrum anthropi* ATCC 49188 T, a versatile opportunistic pathogen and symbiont of several eukaryotic hosts. *J Bacteriol*. 2011 Aug;193(16):4274-5. Epub 2011 Jun 17.

Chang YJ et al. Noncontiguous finished genome sequence and contextual data of the filamentous soil bacterium *Ktedonobacter racemifer* type strain (SOSP1-21). *Stand Genomic Sci*. 2011 Oct 15;5(1):97-111. Epub 2011 Oct 1.

Chapman JA et al. Meraculous: *de novo* genome assembly with short paired-end reads. *PLoS One*. 2011;6(8):e23501. Epub 2011 Aug 18.

Chertkov O et al. Complete genome sequence of *Thermomonospora curvata* type strain (B9). *Stand Genomic Sci*. 2011 Feb 20;4(1):13-22.

Chiang YM et al. Characterization of a polyketide synthase in *Aspergillus niger* whose product is a precursor for both dihydroxynaphthalene (DHN) melanin and naphtho-γ-pyrone. *Fungal Genet Biol*. 2011 Apr;48(4):430-7. Epub 2010 Dec 19.

Colbourne JK et al. The ecoresponsive genome of *Daphnia pulex. Science*. 2011 Feb 4;331(6017):555-61.

Coleman NV et al. Genome sequence of the ethene- and vinyl chloride-oxidizing actinomycete *Nocardioides* sp. strain JS614. *J Bacteriol*. 2011 May 6.

Daligault H et al. Complete genome sequence of *Haliscomenobacter hydrossis* type strain (O). *Stand Genomic Sci*. 2011 Jul 1;4(3):352-60. Epub 2011 Jun 30.

DeAngelis KM et al. Characterization of trapped lignin-degrading microbes in tropical forest soil. *PLoS One*. 2011 Apr 29;6(4):e19306.

Del Rio TG et al. Complete genome sequence of *Intrasporangium calvum* type strain (7 KIP). *Stand Genomic Sci*. 2010 Nov 16;3(3):294-303.

Djao OD et al. Complete genome sequence of *Syntrophothermus lipocalidus* type strain (TGB-C1). *Stand Genomic Sci*. 2010 Dec 15;3(3):268-75.

Druzhinina IS et al. *Trichoderma*: the genomics of opportunistic success. *Nat Rev Microbiol*. 2011 Sep 16;9(10):749-59. doi: 10.1038/nrmicro2637.

Duplessis S et al. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc Natl Acad Sci U S A*. 2011 May 2.

Earl D et al. Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. *Genome Res*. 2011 Nov 2.

Eastwood DC et al. The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. *Science*. 2011 Aug 5;333(6043):762-5. Epub 2011 Jul 14.

Ellison CE et al. Massive changes in genome architecture accompany the transition to self-fertility in the filamentous fungus *Neurospora tetrasperma*. *Genetics*. 2011 Sep;189(1):55-69. Epub 2011 Jul 12.

Feinberg L et al. Complete genome sequence of the cellulolytic thermophile *Clostridium thermocellum* DSM1313. *J Bacteriol*. 2011 Apr 1.

Field D et al. The genomic standards consortium. *PLoS Biol*. 2011 Jun;9(6):e1001088. Epub 2011 Jun 21.

Fostona M et al. Chemical, ultrastructural, and supramolecular analysis of tension wood in *Populus tremula x alba* as a model substrate for reduced recalcitrance. *Energy Environ. Sci*. 2011. Advance Article. DOI:10.1039/C1EE02073K

Frese SA et al. The evolution of host specialization in the vertebrate gut symbiont *Lactobacillus reuteri*. *PLoS Genet*. 2011 Feb;7(2):e1001314. Epub 2011 Feb 17.

Galardini M. Exploring the symbiotic pangenome of the nitrogen-fixing bacterium *Sinorhizobium meliloti*. *BMC Genomics*. 2011 May 12;12:235.

Gerasimova A. Comparative genomics of the dormancy regulons in mycobacteria. *J Bacteriol*. 2011 Jul;193(14):3446-52. Epub 2011 May 20.

Gladden JM et al. Glycoside hydrolase activities of thermophilic bacterial consortia adapted to switchgrass. *Appl Environ Microbiol*. 2011 Aug;77(16):5804-12. Epub 2011 Jul 1.

Gobler CJ et al. Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics. *Proc Natl Acad Sci U S A*. 2011 Mar 15;108(11):4352-7. Epub 2011 Feb 23.

Göker M et al. Complete genome sequence of *Isosphaera pallida* type strain (IS1B). *Stand Genomic Sci*. 2011 Feb 20;4(1):63-71.

Göker M et al. Complete genome sequence of the acetate-degrading sulfate reducer *Desulfobacca acetoxidans* type strain (ASRB2). *Stand Genomic Sci*. 2011 Jul 1;4(3):393-401. Epub 2011 Jun 30.

Göker M et al. Complete genome sequence of *Odoribacter splanchnicus* type strain (1651/6). *Stand Genomic Sci*. 2011 Apr 29;4(2):200-9.

Goodstein DM et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. 2011 Nov 22.

Goodwin SB et al. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet*. 2011 Jun;7(6):e1002070. Epub 2011 Jun 9.

Gori F et al. The metagenomic basis of anammox metabolism in *Candidatus 'Brocadia fulgida.'* *Biochem Soc Trans*. 2011 Dec 1;39(6):1799-804.

Grbic M et al. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature*. 2011 Nov 23;479(7374):487-92. doi: 10.1038/nature10640.

Grigoriev IV et al. The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res*. 2012 Jan;40(1):D26-32. Epub 2011 Nov 22.

Gronow S et al. Complete genome sequence of *Bacteroides salanitronis* type strain (BL78). *Stand Genomic Sci*. 2011 Apr 29;4(2):191-9.

Gronow S et al. Complete genome sequence of *Paludibacter propionicig enes* type strain (WB4). *Stand Genomic Sci*. 2011 Feb 20;4(1):36-44.

Grubbs KJ et al. The complete genome sequence of *Streptomyces cf. griseus* (XyelbKG-1), an ambrosia beetle-associated Actinomycete. *J Bacteriol*. 2011 Apr 1.

Han C et al. Complete genome sequence of *Syntrophobotulus glycolicus* type strain (FlGlyR). *Stand Genomic Sci*. 2011 Jul 1;4(3):371-80.

Han C et al. Complete genome sequence of *Treponema succinifaciens* type strain (6091). *Stand Genomic Sci*. 2011 Jul 1;4(3):361-70. Epub 2011 Jun 30.

Handoko L et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet*. 2011 Jun 19;43(7):630-8. doi: 10.1038/ng.857.

Harhangi HR et al. Hydrazine synthase, a unique phylomarker to study the presence and biodiversity of anammox bacteria. *Appl Environ Microbiol*. 2011 Dec 2.

Hauser LJ et al. The complete genome sequence and updated annotation of *Desulfovibrio alaskensis* G20. *J Bacteriol*. 2011 Jun 17.

Hellsten U et al. A segmental genomic duplication generates a functional intron. *Nat Commun*. 2011 Aug 30;2:454. doi: 10.1038/ncomms1461.

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

Hsu CY et al. Flowering locus T duplication coordinates reproductive and vegetative growth in perennial poplar. *Proc Natl Acad Sci U S A*. 2011 Jun 28;108(26):10756-61. Epub 2011 Jun 8.

Hu TT et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet*. 2011 May;43(5):476-81. Epub 2011 Apr 10.

Huntemann M et al. Complete genome sequence of the thermophilic sulfur-reducer *Hippea maritima* type strain (MH(2)). *Stand Genomic Sci*. 2011 Jul 1;4(3):303-11. Epub 2011 Jun 30.

Ivanova N et al. Complete genome sequence of the extremely halophilic *Halanaerobium praevalens* type strain (GSL). *Stand Genomic Sci*. 2011 Jul 1;4(3):312-21. Epub 2011 Jun 30.

Ivanova N et al. Complete genome sequence of *Truepera radiovictrix* type strain (RQ-24T). *Stand Genomic Sci*. 2011 Feb 22;4(1):91-9.

Jackson SA et al. Sequencing crop genomes: approaches and applications. *New Phytol*. 2011 Jun 27. doi: 10.1111/j.1469-8137.2011.03804.x.

Jiang Z et al. Development and initial characterization of a HAPPY panel for mapping the *X. tropicalis* genome. *Int J Biol Sci*. 2011;7(7): 1037-44. Epub 2011 Aug 24.

Jones FC et al. A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Curr Biol*. 2011 Dec 20.

Kalamorz F et al. Draft genome sequence of thermoalkaliphilic *Caldalkalibacillus thermarum* strain TA2.A1. *J Bacteriol*. 2011 Jun 17.

Kallimanis A et al. Complete genome sequence of *Arthrobacter phenanthrenivorans* type strain (Sphe3). *Stand Genomic Sci*. 2011 Apr 29;4(2): 123-30.

Kallimanis A et al. Complete genome sequence of Mycobacterium sp. strain (Spyr1) and reclassification to *Mycobacterium gilvum Spyr1*. *Stand Genomic Sci*. 2011 Oct 15;5(1):144-53. Epub 2011 Oct 1.

Kant R et al. Genome sequence of *Pedosphaera parvula* Ellin514, an aerobic verrucomicrobial isolate from pasture soil. *J Bacteriol*. 2011 Apr 1.

Kant R et al. Genome sequence of *Chthoniobacter flavus* Ellin428, an aerobic heterotrophic soil bacterium. *J Bacteriol*. 2011 Apr 1.

Karpowicz SJ et al. The GreenCut2 Resource, a phylogenomically derived inventory of proteins specific to the plant lineage. *J Biol Chem*. 2011 Jun 17;286(24):21427-39. Epub 2011 Apr 22.

Kerfeld CA and Scott KM. Using BLAST to teach "e-value-tionary" concepts. *PLoS Biol*. 2011 Feb 1;9(2):e1001014.

Kim TW et al. High-throughput in vitro glycoside hydrolase (HIGH) screening for enzyme discovery. *Angew Chem Int Ed Engl*. 2011 Sep 16. doi: 10.1002/anie.201106146.

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

Kiss H et al. Complete genome sequence of *Thermobaculum terrenum* type strain (YNP1). *Stand Genomic Sci*. 2010 Oct 27;3(2):153-62.

Klassen JL et al. Draft genome sequence of Streptomyces sp. strain Wigar10, isolated from a surface-sterilized garlic bulb. *J Bacteriol*. 2011 Dec;193(24):6999-7000.

Klippel B et al. Complete genome sequence of the marine cellulose- and xylan-degrading bacterium Glaciecolasp. strain 4H-3-7+YE-5. *J Bacteriol*. 2011 Sep;193(17):4547-8. Epub 2011 Jun 24.

Klippel B et al. Complete genome sequences of Krokinobactersp. strain 4H-3-7-5 and Lacinutrixsp. strain 5H-3-7-4, polysaccharide-degrading members of the family Flavobacteriaceae. *J Bacteriol*. 2011 Sep;193(17):4545-6. Epub 2011 Jul 1.

Kochetov AV et al. Tandem termination signal in plant mRNAs. *Gene*. 2011 Jul 15;481(1):1-6. Epub 2011 Apr 22.

Kouvelis VN et al. Genome sequence of the ethanol-producing *Zymomonas mobilis subsp. pomaceae lectotype* ATCC 29192. *J Bacteriol*. 2011 Jul 8.

Kubicek CP et al. Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of *Trichoderma*. *Genome Biol*. 2011 Apr 18;12(4):R40.

Land M et al. Noncontiguous finished genome sequence of *Bacteroides coprosuis* type strain (PC139). *Stand Genomic Sci*. 2011 Apr 29;4(2):233-43.

Lang E et al. Complete genome sequence of *Weeksella virosa* type strain (9751T). *Stand Genomic Sci*. 2011 Feb 22;4(1):81-90.

Lapidus A et al. Genome sequence of the moderately thermophilic halophile *Flexistipes sinusarabici* strain (MAS10). *Stand Genomic Sci*. 2011 Oct 15;5(1):86-96. Epub 2011 Sep 23.

Lapidus A et al. Genomes of three methylotrophs from a single niche uncover genetic and metabolic divergence of *Methylophilaceae. J Bacteriol*. 2011 May 27.

Liolios K et al. Complete genome sequence of the gliding, heparinolytic *Pedobacter saltans* type strain (113). *Stand Genomic Sci*. 2011 Oct 15;5(1):30-40. Epub 2011 Sep 23.

Liu S et al. Complete genome sequence of *Lactobacillus buchneri* NRRL B-30929, a novel strain from a commercial ethanol plant. *J Bacteriol*. 2011 Aug;193(15):4019-20. Epub 2011 May 27.

Lukashin I et al. VISTA Region Viewer (RViewer)—a computational system for prioritizing genomic intervals for biomedical studies. *Bioinformatics*. 2011 Sep 15;27(18):2595-7. Epub 2011 Jul 26.

Luo C et al. Individual genome assembly from complex community short-read metagenomic datasets. *ISME J*. 2011 Oct 27. doi: 10.1038/ismej.2011.147.

Lykidis A et al. Multiple syntrophic interactions in a terephthalate-degrading methanogenic consortium. *ISME J*. 2011 Jan;5(1):122-30. Epub 2010 Aug 5.

Mackelprang R et al. Deep metagenome sequencing illuminates permafrost response to thaw. *Nature*. 2011 Nov 6. doi:10.1038/*Nature*10576.

Markowitz, V. et al. IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res*. 2012 Jan;40(1):D115-22.

Markowitz VM et al. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res*. 2012 Jan;40(1):D123-9. Epub 2011 Nov 15.

Martin F et al. Sequencing the fungal tree of life. *New Phytol*. 2011 Mar 14. doi:10.1111/j.1469-8137.2011.03688.x.

Martin JA et al. Next-generation transcriptome assembly. *Nat Rev Genet*. 2011 Sep 7. doi: 10.1038/nrg3068.

Mavromatis K et al. Complete genome sequence of *Spirochaeta smaragdinae* type strain (SEBR 4228). *Stand Genomic Sci*. 2010 Oct 18; 3(2):136-44.

Mavromatis K et al. Complete genome sequence of *Riemerella anatipestifer* type strain (ATCC 11845). *Stand Genomic Sci*. 2011 Apr 29;4(2):144-53.

McCluskey K et al. Rediscovery by whole genome sequencing: classical mutations and genome polymorphisms in *Neurospora crassa*. G3: *Genes, Genomes, Genetics*. September 2011 1:303-316; doi:10.1534/g3.111.000307.

Medina M et al. Crawling through time: Transition of snails to slugs dating back to the Paleozoic, based on mitochondrial phylogenomics. *Mar Genomics*. 2011 Mar;4(1):51-9. Epub 2011 Feb 4.

Miller DA et al. Complete genome sequence of the cellulose-degrading bacterium *Cellulosilyticum lentocellum*. *J Bacteriol*. 2011 May; 193(9): 2357-2358.

Monier A et al. Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. *Environ Microbiol*. 2011 Sep 14. doi: 10.1111/j.1462-2920.2011.02576.x.

Morrison N et al. Data shopping in an open marketplace: Introducing the Ontogrator Web application for marking up data using ontologies and browsing using facets. *Stand Genomic Sci*. 2011 Apr 29;4(2):286-92.

Munk AC et al. Complete genome sequence of *Rhodospirillum rubrum* type strain (S1). *Stand Genomic Sci*. 2011 Jul 1;4(3):293-302. Epub 2011 Jun 30.

Munk AC et al. Complete genome sequence of *Tsukamurella paurometabola* type strain (no. 33). *Stand Genomic Sci*. 2011 Jul 1;4(3):342-51. Epub 2011 Jun 30.

Muyzer G et al. Complete genome sequence of *Thioalkalivibrio sulfidophilus* HL-EbGr7. *Stand Genomic Sci*. 2011 Feb 14;4(1):23-35.

Nawabi P et al. Engineering *E. coli* for biodiesel production utilizing a bacterial fatty acid methyl-transferase. *Appl Environ Microbiol*. 2011 Sep 16.

Nolan M et al. Complete genome sequence of *Ferrimonas balearica* type strain (PAT). *Stand Genomic Sci*. 2010 Oct 27;3(2):174-82.

Pagani I et al. Complete genome sequence of *Desulfobulbus propionicus* type strain (1pr3T). *Stand Genomic Sci*. 2011 Feb 20;4(1):100-10.

Pagani I et al. Complete genome sequence of *Marivirga tractuosa* type strain (H-43). *Stand Genomic Sci*. 2011 Apr 29;4(2):154-62.

Pagani I et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*. 2012 Jan;40(1):D571-9. Epub 2011 Dec 1.

Pappas KM et al. Genome sequence of the ethanol-producing *Zymomonas mobilis subsp. mobilis lectotype* ATCC 10988. *J Bacteriol*. 2011 Jul 1.

Pati A et al. Noncontiguous finished genome sequence of the opportunistic oral pathogen *Prevotella multisaccharivorax* type strain (PPPA20). *Stand Genomic Sci*. 2011 Oct 15;5(1):41-9. Epub 2011 Sep 23.

Pati A et al. Complete genome sequence of *Cellulophaga lytica* type strain (LIM-21). *Stand Genomic Sci*. 2011 Apr 29;4(2):221-32.

Pati A et al. Complete genome sequence of *Oceanithermus profundus* type strain (506). *Stand Genomic Sci*. 2011 Apr 29;4(2):210-20.

Pati A et al. Complete genome sequence of *Bacteroides helcogenes* type strain (P 36-108). *Stand Genomic Sci*. 2011 Feb 22;4(1):45-53.

Pati A et al. ClaMS: A classifier for metagenomic sequences. *Stand Genomic Sci*. 2011 Nov 30;5(2): 248-53. Epub 2011 Nov 28.

Persson T et al. Genome sequence of *Candidatus Frankia datiscae* Dg1, the uncultured microsymbiont from nitrogen-fixing root nodules of the dicot *Datisca glomerata*. *J Bacteriol*. 2011 Dec;193(24):7017-8.

Pitluck S et al. Noncontiguous finished genome sequence of *Aminomonas paucivorans* type strain (GLU-3). *Stand Genomic Sci*. 2010 Nov 20;3(3):285-93.

Pitluck S et al. Complete genome sequence of *Calditerrivibrio nitroreducens* type strain (Yu37-1). *Stand Genomic Sci*. 2011 Feb 20;4(1):54-62.

Pope PB et al. Isolation of Succinivibrionaceae implicated in low methane emissions from tammar wallabies. *Science*. 2011 Jun 30.

Pukall R et al. Complete genome sequence of *Deinococcus maricopensis* type strain (LB-34). *Stand Genomic Sci*. 2011 Apr 29;4(2):163-72. Epub 2011 Apr 25.

Reddy AP et al. Bioenergy feedstock-specific enrichment of microbial populations during high-solids thermophilic deconstruction. *Biotechnol Bioeng*. 2011 Sep;108(9):2088-98. doi: 10.1002/bit.23176. Epub 2011 May 2.

Roberts EW et al. Isolation and characterization of the *Prochlorococcus carboxysome* reveals the presence of the novel shell protein CsoS1D. *J Bacteriol*. 2011 Dec 9.

Sales CM et al. Genome sequence of 1,4-dioxane degrading *Pseudonocardia dioxanivorans* strain CB1190. *J Bacteriol*. 2011 Jul 1.

Sandh G et al. Comparative proteomic profiles of the marine cyanobacterium *Trichodesmium erythraeum* IMS101 under different nitrogen regimes. *Proteomics*. 2011 Feb;11(3):406-19. doi: 10.1002/pmic.201000382. Epub 2011 Jan 5.

Scholz MB et al. Next-Generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr Opin Biotechnol*. 2011 Dec 7.

Schuster A et al. A versatile toolkit for high throughput functional genomics with *Trichoderma reesei*. *Biotechnol Biofuels*. 2012 Jan 2;5(1):1

Seidl-Seiboth V et al. Novel hydrophobins from *Trichoderma* define a new hydrophobin subclass: protein properties, evolution, regulation and processing. *J Mol Evol*. 2011 Mar 22.

Sessitsch A et al. Functional characteristics of an endophyte community colonizing rice roots as revealed by metagenomic analysis. *Mol Plant Microbe Interact*. 2011 Oct 4.

Siddaramappa S et al. Horizontal gene transfer in *Histophilus somni* and its role in the evolution of pathogenic strain 2336, as determined by comparative genomic analyses. *BMC Genomics*. 2011 Nov 23;12(1):570.

Siddaramappa S et al. Comparative genomic analyses of *Francisella novicida*-like isolates from North America and Australia reveal common ancestry and novel genetic traits. *Appl Environ Microbiol*. 2011 Jun 10.

Siebers B et al. The complete genome sequence of *Thermoproteus tenax*: a physiologically versatile member of the Crenarchaeota. *PLoS One*. 2011;6(10):e24222. Epub 2011 Oct 7.

Sikorski J et al. Complete genome sequence of *Sulfurimonas autotrophica* type strain (OK10). *Stand Genomic Sci*. 2010 Oct 27;3(2):194-202.

Sikorski J et al. Complete genome sequence of *Ilyobacter polytropus* type strain (CuHbu1). *Stand Genomic Sci*. 2010 Dec 15;3(3):304-14.

Sikorski J et al. Complete genome sequence of *Mahella australiensis* type strain (50-1 BON). *Stand Genomic Sci*. 2011 Jul 1;4(3):331-41. Epub 2011 Jun 30.

Singer E et al. The genomic potential of *Marinobacter aquaeolei*—A biogeochemical opportunitroph. *Appl Environ Microbiol*. 2011 Apr; 77(8): 2763-2771.

Skennerton CT et al. Phage Encoded H-NS: A potential Achilles heel in the bacterial defence system. *PLoS One*. 2011;6(5):e20095. Epub 2011 May 18.

Smith DR et al. The GC-rich mitochondrial and plastid genomes of the green alga *Coccomyxa* give insight into the evolution of organelle DNA nucleotide landscape. *PLoS One*. 2011;6(8):e23624. Epub 2011 Aug 26.

Spring S et al. The genome sequence of *Methanohalophilus mahii* SLPT reveals differences in the energy metabolism among members of the Methanosarcinaceae inhabiting freshwater and saline environments. *Archaea*. 2010 Dec 23;2010:690737.

Srivastava A et al. A zinc finger family transcription factor AbVf19 is required for the induction of a gene subset important for virulence in *Alternaria brassicicola*. *Mol Plant Microbe Interact*. 2011 Dec 20.

Starkenburg SR et al. Genome of the cyanobacterium *Microcoleus vaginatus* FGP-2, a photosynthetic ecosystem engineer of arid land soil biocrusts worldwide. *J Bacteriol*. 2011 Sep;193(17):4569-70. Epub 2011 Jun 24.

Stein LY et al. Genome sequence of the methanotrophic *Alphaproteobacterium, Methylocystis sp. Rockwell* (ATCC 49242). *J Bacteriol*. 2011 May;193(10):2668-9. Epub 2011 Mar 25.

Studer MH et al. Lignin content in natural *Populus* variants affects sugar release. *Proc Natl Acad Sci U S A*. 2011 Apr 12;108(15): 6300-5. Epub 2011 Mar 28.

Sucgang R et al. Comparative genomics of the social amoebae *Dictyostelium discoideum* and *Dictyostelium purpureum*. *Genome Biol*. 2011 Feb 28;12(2):R20.

Suen G et al. The complete genome sequence of *Fibrobacter succinogenes* S85 reveals a cellulolytic and metabolic specialist. *PLoS One*. 2011 Apr 19;6(4):e18814.

Suen G et al. Complete genome of the cellulolytic ruminal bacterium *Ruminococcus albus* 7. *J Bacteriol*. 2011 Oct;193(19):5574-5.

Svenning MM et al. Genome sequence of the Arctic methanotroph *Methylobacter tundripaludum* SV96. *J Bacteriol*. 2011 Nov;193(22):6418-9. Epub 2011 Jul 1.

Swarbreck SM et al. Analysis of leaf and root transcriptomes of soil-grown *Avena barbata* plants. *Plant Cell Physiol*. 2011 Feb;52(2):317-32.

Swithers KS et al. Genome sequence of *Kosmotoga olearia* strain TBF 19.5.1, a thermophilic bacterium with a wide growth temperature range, isolated from the Troll B Oil Platform in the North Sea. *J Bacteriol*. 2011 Oct;193(19):5566-7.

Swithers KS et al. Genome sequence of *Thermotoga* sp. strain RQ2, a hyperthermophilic bacterium isolated from a geothermally heated region of the seafloor near Ribeira Quente, the Azores. *J Bacteriol*. 2011 Oct;193(20):5869-70.

Tang KH et al. Complete genome sequence of the filamentous anoxygenic phototrophic bacterium *Chloroflexus aurantiacus*. *BMC Genomics*. 2011 Jun 29;12(1):334.

Tisserant E et al. The transcriptome of the arbuscular mycorrhizal fungus *Glomus intraradices* (DAOM 197198) reveals functional tradeoffs in an obligate symbiont. *New Phytol*. 2011 Nov 16. doi: 10.1111/j.1469-8137.2011. 03948.x.

Vanden Wymelenberg A et al. Gene expression of wood decay fungi *Postia placenta* and *Phanerochaete chrysosporium* is significantly altered by plant species. *Appl Environ Microbiol*. 2011 May 6.

Vandera E et al. Heterologous expression and characterization of two 1-hydroxy-2-naphthoic acid dioxygenases from *Arthrobacter phenanthrenivorans*. *Appl Environ Microbiol*. 2011 Nov 18.

van Passel MW et al. Genome sequence of the Verrucomicrobium *Opitutus terrae* PB90-1, an abundant inhabitant of rice paddy soil ecosystems. *J Bacteriol*. 2011 May; 193(9): 2367-2368.

van Passel MW et al. The genome of *Akkermansia muciniphila*, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. *PLoS One*. 2011 Mar 3;6(3):e16876.

van Passel MW et al. Genome sequence of *Victivallis vadensis* ATCC BAA-548, an anaerobic bacterium from the phylum Lentisphaerae, isolated from the human gastro-intestinal tract. *J Bacteriol*. 2011 Mar 11.

Vishnivetskaya TA et al. Complete genome sequence of the thermophilic *Exiguobacterium* sp. AT1b. *J Bacteriol*. 2011 Apr 1.

Voolstra, CR et al. Rapid evolution of coral proteins responsible for interaction with the environment. *PLoS One*. 2011;6(5):e20392. Epub 2011 May 25.

Wang L et al. Mapping N-linked glycosylation sites in the secretome and whole cells of *Aspergillus niger* using hydrazide chemistry and mass spectrometry.

J Proteome Res. 2012 Jan 1;11(1):143-56. Epub 2011 Dec 2.

Weilharter A et al. Complete genome sequence of the plant growth-promoting endophyte *Burkholderia phytofirmans* Strain PsJN. *J Bacteriol*. 2011 Jul;193(13):3383-4. Epub 2011 May 6.

Wirth R et al. Complete genome sequence of *Desulfurococcus mucosus* type strain (O7/1). *Stand Genomic Sci*. 2011 Apr 29;4(2):173-82.

Wohlbach DJ et al. Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. *Proc Natl Acad Sci USA*. 2011 Aug 9;108(32):13212-7. Epub 2011 Jul 25.

Wortman J et al. A data analysis and coordination center for the human microbiome project. *Genome Biol*ogy. 2010, 11(Suppl 1):O13.

Woyke T et al. Complete genome sequence of the gliding freshwater bacterium *Fluviicola taffensis* type strain (RW262). *Stand Genomic Sci*. 2011 Oct 15;5(1):21-9. Epub 2011 Sep 23.

Woyke T et al. Decontamination of MDA reagents for single cell whole genome amplification. *PLoS One*. 2011;6(10):e26161. Epub 2011 Oct 20.

Yang X et al. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Res*. 2011 Apr;21(4):634-41.

Yang X et al. Identification of candidate genes in *Arabidopsis* and *Populus* cell wall biosynthesis using text-mining, co-expression network analysis and comparative genomics. *Plant Sci*. 2011 Dec;181(6):675-87.

Ye CY et al. Comparative analysis of GT14/GT14-like gene family in *Arabidopsis, Oryza, Populus, Sorghum,* and *Vitis. Plant Sci*. 2011 Dec;181(6):688-95.

Yilmaz, P et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications *Nat Biotechnol*. 2011 May;29(5):415-20.

Yuichi S et al. Genome sequence of *Nitrosomonas* sp. strain AL212, an ammonia-oxidizing bacterium sensitive to high levels of ammonia. *J Bacteriol*. 2011 Sep 1;193(18)5047-5048.

Zeytun A et al. Complete genome sequence of *Hydrogenobacter thermophilus* type strain (TK-6). *Stand Genomic Sci*. 2011 Apr 29;4(2): 131-43.

Zhao B et al. Complete genome sequence of the anaerobic halophilic alkalithermophile *Natranaerobius thermophilus* JW/NM-WN-LFT. *J Bacteriol*. 2011 Jun 3.
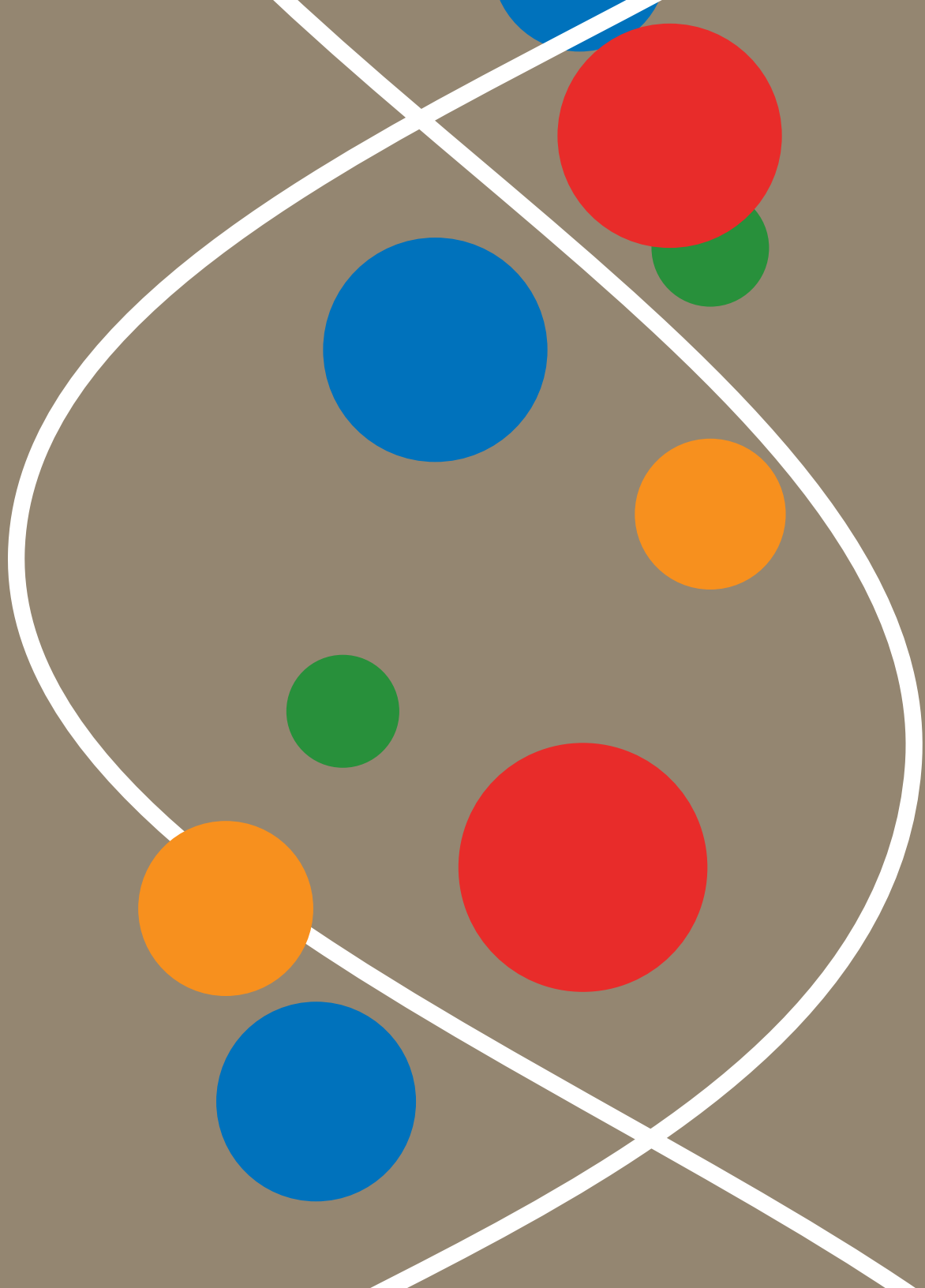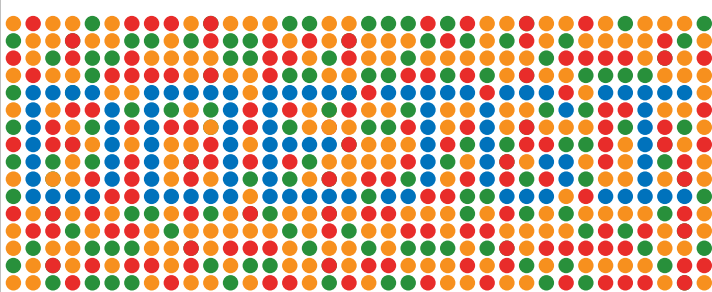
Comments?

David Gilbert
DOE JGI Public Affairs Manager
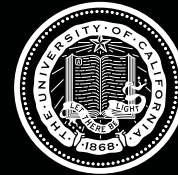degilbert@lbl.gov

Take Our Survey

http://bit.ly/11JGIr

Follow Us

U.S. DEPARTMENT OF
ENERGY
Office of Science

BERKELEY LAB

All photography, unless otherwise noted, by Roy Kaltschmidt, Berkeley Lab.

CSO23169

2011

U.S. Department of Energy Joint Genome Institute

PROGRESS REPORT

PROGRESS REPORT

U.S. Department of Energy Joint Genome Institute

JGI

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE