U.S. DEPARTMENT OF ENERGY

JOINT GENOME INSTITUTE PROGRESS REPORT 2007

U.S. DEPARTMENT OF ENERGY

Office of Science
U.S. DEPARTMENT OF ENERGY

JGI
DOE JOINT GENOME INSTITUTE
U.S. DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

On the cover: The eucalyptus tree was selected in 2007 for sequencing by the JGI. The microbial community in the termite hindgut of *Nasutitermes corniger* was the subject of a study published in the November 22, 2007 edition of the journal, *Nature.*

JGI Mission

The U.S. Department of Energy Joint Genome Institute, supported by the DOE Office of Science, unites the expertise of five national laboratories—Lawrence Berkeley, Lawrence Livermore, Los Alamos, Oak Ridge, and Pacific Northwest — along with the Stanford Human Genome Center to advance genomics in support of the DOE missions related to clean energy generation and environmental characterization and cleanup. JGI's Walnut Creek, CA, Production Genomics Facility provides integrated high-throughput sequencing and computational analysis that enable systems-based scientific approaches to these challenges.

U.S. DEPARTMENT OF ENERGY JOINT GENOME INSTITUTE PROGRESS REPORT 2007

JGI PROGRESS REPORT 2007

DIRECTOR'S PERSPECTIVE

In this past year, we have seen a continuing emphasis on the importance of developing biologically-derived alternatives to fossil fuels driven by a confluence of political, economic, and environmental factors. In connection with this, the Department of Energy recently funded three new Bioenergy Research Centers to accelerate basic research for the development of cellulose-derived fuels. The JGI will play an important role—through sequencing selected genomes and conducting data analysis—in enabling the groundbreaking science of these Centers. While the JGI will sustain its other user programs, the DOE Bioenergy Research Centers represent a substantial and important new community of users for the JGI.

The Community Sequencing Program (CSP), now in its fourth year, continues to be the main conduit for our users. The largest CSP project that the JGI will be tackling in the next year is the eucalyptus tree genome. As is typical for many JGI projects, the consortium of investigators who proposed this effort draws expertise from dozens of institutions and hundreds of researchers worldwide. Eucalyptus are rapidly growing trees, and likely to gain more attention as a prospective bioenergy crop. Knowledge of this genome will accelerate the work of researchers who are seeking to adapt this tree and other plant species as a biomass feedstock.

The JGI's productivity in the past year can be assessed in the basic

metric of nucleotides sequenced, but a more important measure is the increasing number of researchers whose science was enabled by the new data and its analysis available from the various JGI databases. This contribution is only in part reflected in the peer-reviewed publications authored by JGI researchers. During 2006–2007, JGI and its collaborators published over 170 papers, with more than 20 papers in the journals *Science* and *Nature*. These papers have ranged from an analysis of microbial communities in the termite hindgut as a means to identify cellulose-degrading machinery, to revealing the genes of a green alga genome and how its genes enable efficient carbon capture.

The JGI hosts a variety of forums to encourage interactions with its users. In March, the JGI's annual user meeting drew over 400 participants to Walnut Creek, California. This event provided a forum for investigators to discuss sequence-based science, genomic applications to issues of bioenergy, carbon cycling, and bioremediation, and informatics tools for sequence analysis. In addition to the user meeting, the JGI hosted numerous informatics workshops, annotation jamborees, and focused meetings throughout the year at the JGI's Walnut Creek Production Genomics Facility (PGF).

A crucial change presently underway at the PGF is a technological one. For the past 20-plus years, the Sanger-based sequencing process dominated how sequencing was done. Over the years, as this technology has yielded increases in throughput and reduction in costs, the number of JGI's Sanger-based sequencing machines eventually

DOE Bioenergy Research Centers

grew to more than 110. Recently, a new generation of sequencers has emerged, offering even more spectacular economies of scale and capability. Accordingly, the JGI has been exploring how to add these new platforms and their capabilities to meet JGI's sequencing needs. This last summer, as part of that transition, the JGI retired 36 Sanger-based DNA sequencers and has added several next-generation sequencers to the DNA sequencing production line.

With these new technologies, DNA sequencing centers, such as the JGI, have progressively become fire hoses of DNA sequence data with the accompanying challenge of supplying this information to users in an accessible form. In response to this challenge, the JGI is engaged in developing new ways to improve community access to its data. This has included increases in our informatics capabilities, improvements in our plant and microbial portals, as well as expansion of our educational activities.

Essential to the efficient operation of the JGI is maintaining a safe workplace and a motivated workforce. Understanding that the nature of production sequencing is repetitive, the JGI has taken enormous care to monitor various aspects of the production activities and to implement the necessary changes to ensure a safe workplace. Despite our efforts and successes, ergonomics remains a serious concern at the PGF. For the month of December, 2007, reinforcing our commitment to maintain a safe workplace, we seized an opportunity to step back from the day-to-day production schedule to evaluate our standard operating procedures and implement bottom-to-top process improvements.

On the operations front, a crucial accomplishment in the past year has been securing a five-year lease with an option for an additional five years on our Walnut Creek facility—our home for the entirety of the JGI's 10-year history. It is only with support from the Department of Energy and the University of California that we were able to marshal the resources to secure our future. This new lease will enable us to expand our campus, to consolidate our administrative and informatics staff, and effectively meet the expanding thirst for DNA sequence and analysis.

We look ahead to contributing to advances that will be catalyzed at the interface of biology, information technology, energy, and the environment. I encourage you to read on about the successes of the JGI in the past year—and our prospects for the future.

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


JGI
JOINT GENOME INSTI
UNITED STATES DEPARTMENT OF E
2800 MITCHELL D

JGI HISTORY

The U.S. Department of Energy Joint Genome Institute (JGI) was created in 1997 to unite the expertise and resources in genome mapping, DNA sequencing, technology development, and information sciences pioneered at the DOE genome centers at Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (LANL).

In 1999, the University of California, which managed the three national labs for the DOE, leased 60,000 square feet of laboratory and office space in a light industrial park in Walnut Creek, California, to consolidate activities and accommodate the JGI's growing workforce in what is now known as the Production Genomics Facility (PGF).

In 2004, with the successful completion of the human genome, the JGI reoriented its focus to become a user facility driven by DOE mission needs in the areas of bioenergy, carbon cycling, and bioremediation. This transition has benefited from broader involvement of the DOE national laboratory system, including the formal participation of Oak Ridge National Laboratory (ORNL), Pacific Northwest National Laboratory (PNNL), and the Stanford Human Genome Center, in the activities of the JGI. The importance of the JGI as essential infrastructure for DOE science was codified through signing of a new MOU between the five partner labs in 2006.

In 2007, with the launching of the three DOE Bioenergy Research Centers (http://genomicsgtl.energy.gov/centers/) to accelerate basic research in the development of next-generation biofuels, the JGI was acknowledged as a vital contributor in its capacity to provide DNA sequencing and its analysis in support of these centers.

For more information about the JGI's history, facility, partners, and science, visit our Web site at http://www.jgi.doe.gov/whoweare/index.html.


Berkeley Lab
Lawrence Livermore
National Laboratory
Los Alamos
NATIONAL LABORATORY
Oak
Ridge
National Laboratory
Pacific Northwest
National Laboratory
Operated by Battelle for the
U.S. Department of Energy
shgc

PARTNER LABORATORIES

The JGI primarily draws its employees for the Production Genomics Facility (PGF) in Walnut Creek, California, from Lawrence Berkeley National Laboratory (about 160) and Lawrence Livermore National Laboratory (about 50), but there are a total of five national laboratories participating in the activities of the institute:

LAWRENCE BERKELEY NATIONAL LABORATORY

The JGI draws on the scientific and high-performance computing expertise of LBNL. The JGI partners with the Biological Data Management and Technology Center (BDMTC) on the Integrated Microbial Genomes (IMG) data management system, which is updated quarterly with new public and DOE JGI genomes.

The JGI also taps into VISTA, a comprehensive suite of programs and databases for comparative analysis of genomic sequences, for sequence submission, for the generation of alignment analyses (VISTA servers), and for the examination of pre-computed whole-genome alignments of different species.

As one of the world's largest sequencing facilities, the JGI generates millions of trace data files each month. These trace fragments are assembled into genomic sequences and refined through comparison with other sequences to confirm the assembly. A sequence is compared and contrasted with sequence from the same organism and other organisms to understand its function. All of this requires massive amounts of data storage, data handling, and retrieval capacity, which is provided by the National Energy Research Scientific Computing (NERSC) facility. NERSC is a critical partner providing scalable resources to tackle the massive data management tasks on the next-generation sequencing frontier. Currently, JGI transfers about 600 gigabytes per month in raw sequence data, and will soon be ramping up to 1.2 terabytes per month.

LAWRENCE LIVERMORE NATIONAL LABORATORY

LLNL staffs the PGF with bioscientists, bioinformaticists, senior managers, and support personnel. Nearly all of these employees are from LLNL's Biosciences and Biotechnology Division (BBTD) or Science and Technology Computing (S&TC) Division and are integral to the PGF mission. This year's largest impacts were made in several key areas including production informatics, global project tracking, finishing informatics, and computing infrastructure.

In FY07, LLNL was also specifically tasked with generating finished DNA sequence data for JGI projects. LLNL employee Patrick Chain led this important activity. He is also the JGI point of contact for finishing, organizing meetings to coordinate tasks, and obtaining updates on projects from the various JGI finishing groups. LLNL's finishing team is highly productive and has excellent collaborative relationships with a large number of JGI users. This team has produced about 30% of the finished sequence for the JGI.

At the end of 2007, LLNL launched an LDRD-funded three-year research project in computational metagenomic analysis and pathway identification. This research will leverage the strength of the LLNL world-class bioinformatics expertise and the developing systems biology capabilities targeted to the bioenergy field. The resulting novel analytical methods and tools will be integrated into the JGI's IMG system, extending its capabilities to the broad user base of the JGI.

LOS ALAMOS NATIONAL LABORATORY

LANL serves as the primary finishing arm of the JGI for prokaryotic (microbial) genomes. JGI LANL has developed and implemented a high-throughput finishing pipeline where experimental and computational activities are tightly linked. Nearly everything in the finishing process is automated, creating a platform specifically optimized for finishing microbial genomes. As a result, LANL has finished more than

120 genomes in the past three years—more than any other single institution in the world. Finished genomes are critical for comparative genomics studies, for development of single nucleotide polymorphism (SNP) and sequence-based detection strategies, and as a reference set for metagenomics studies.

The LANL group conducts a number of education and outreach activities for the scientific community. In addition to working with individual JGI users on comparative genomic analysis and manuscript preparation, LANL hosted four annotation jamborees in 2007 on eukaryotic genomes: two *Trichoderma* species; one *Aspergillus*; and one *Postia*. Workshops on soil microbial communities, metagenomics, and IMG annotation were also hosted. In addition, LANL scientists teach classes on genome annotation and comparative genomics at multiple scientific meetings each year.

In summer of 2007, JGI LANL hosted a major international meeting called "Finishing in the Future." Over 140 scientists and industry technologists from the U.S. and Europe participated in this forum for discussion about new sequencing technologies, sequence-finishing strategies, and genomics-based science.

Information on the 2008 "Finishing in the Future" meeting can be found at http://www.lanl.gov/finishingin thefuture/.

OAK RIDGE NATIONAL LABORATORY

The JGI partners with the Genome Analysis and Systems Modeling Group of the Life Sciences Division of ORNL to annotate genomes, i.e., to assign putative roles and functions to genes. The ORNL Group conducts research and systems development in microbial genome analysis and protein structure analysis and provides bioinformatics and analytic services and resources to the JGI with a goal of predicting prospective gene and protein models for analysis. An annotation constitutes a prediction, amenable to experimental testing.

As hundreds of microbial genomes churn through JGI sequencers, efforts are underway to make their annotation—the determination of what biological functions these sequences actually encode—faster, more accurate, and cheaper. One of the challenges in assigning function to sequence is that the generation of raw sequence has far outpaced experimental data-gathering. To help close this gap, the annotation process is guided by the principle that, if genes found in two different organisms look similar, there is a high probability that they will function in a similar way. If the sequenced DNA contains a gene that looks like another gene that has been experimentally proven to be responsible for the production of a specific protein, there is a high probability that the gene of interest has the same function. This correlation is far from perfect, however.

A look at the microbial pipeline illustrates how automation is helping to speed up the annotation process. The microbial annotation pipeline begins with the PGF uploading data from completely assembled microbial genomes to the JGI FTP site, where ORNL can access it. Then, with the help of a series of computational tools ORNL generates a master list of genes.

After all the database searches are conducted and predictions finalized, there is then a hierarchy of information that can support a function or gene product description. The annotated genome is sent back to the PGF, where the information is uploaded into the Integrated Microbial Genomes (IMG) data management system, the JGI portals, and the National Center for Biotechnology Information's (NCBI's) GenBank for use by the global research community.

PACIFIC NORTHWEST NATIONAL LABORATORY

PNNL provides biological expertise in the areas of environmental microbial genomics, metagenomics, and proteomics. Capabilities and expertise in microbial genomics are applied to eukaryotic microbes, such as fungi and algae, to microbial communities and microbes involved in bioremediation. PNNL proteomics capabilities have been

applied to aid in gene model validation for a number of microbes. Proteomic data from eukaryotic microbes has been visualized using the eukaryotic genome portals. Examples of these projects include PNNL's leadership role in the characterization, physiology, and comparative bioinformatics of *Shewanella* species. The JGI has sequenced 18 *Shewanella* strains of the 22 released.

Currently, in an expanding interaction with the JGI, PNNL is supplying single-pass proteomics characterization of all Genomic Encyclopedia of Bacteria and Archaea (GEBA) microbes to provide data for annotation and initial characterization of strains.

PNNL contributes to the JGI through development of computational biology tools such as ScalaBlast. ScalaBLAST is a parallel implementation of the original NCBI sequence-alignment algorithm (BLAST), which is used to rapidly identify sequences that are similar to a set of protein sequences supplied by a user. ScalaBLAST achieves speedup in a multiprocessor environment by a unique combination of breaking the query list and managing memory in an efficient way so as to more easily manage large-scale projects.

PARTNER LABORATORIES

STANFORD HUMAN GENOME CENTER

The Stanford Human Genome Center (SHGC) began collaborating with the JGI in 1999, with an emphasis on finishing the DOE commitment to the Human Genome Project. Since the completion of the sequencing of the human genome, SHGC has kept pace with this rapid expansion of JGI projects and currently plays several key project roles in the JGI organization.

The greatest contribution of SHGC to advancing JGI science has been assessing, assembling, improving, and finishing eukaryotic whole-genome shotgun (WGS) projects for which the large-scale shotgun sequence is generated at the Production Genomics Facility (JGI PGF). (Recall that JGI LANL does this for microbial genomes.) SHGC generates additional genomic resources (including BAC end sequences, genetic maps, and full-length cDNA sequences) and combines them with the draft WGS to produce improved genomic sequences.

Having the ability to assemble and evaluate WGS-sequenced genomes, combined with the infrastructure to generate additional resources for problematic genomes, allows the SHGC to produce JGI-sequenced eukaryotic genomes of the highest quality possible for the greatest benefit to downstream functional and translational research.

JGI DEPARTMENTS AND PROGRAMS

SEQUENCING DEPARTMENT, SUSAN LUCAS

The JGI Sequencing Department resides at the heart of the JGI Production Genomics Facility. The Department generates high-quality sequence in a cost-efficient manner. As genomics is a rapidly changing field, the department constantly adapts to take advantage of new technological developments that have substantially increased throughput while decreasing the cost of DNA sequencing. The JGI Sequencing Department comprises several subgroups, including: Cloning Technologies, Sequencing Preparation and Generation, Finishing, Quality Control, Sequence Assessment and Analysis, Process Optimization, and Instrumentation. In FY07, the department adopted the Roche 454 sequencing technology. Using the Roche 454, MegaBACE 4500, and ABI 3730xl technologies, the group produced 38 billion base pairs of sequence. In FY08, the department anticipates that it will produce about 40 billion base pairs.

INFORMATICS DEPARTMENT, YAKOV GOLDER

The Informatics Department manages the tracking, assembly, analysis, and distribution of an ever-increasing stream of DNA sequence coming through the facility. There has been a steady increase over the past year in both the volume of sequence being generated and number of different projects that the JGI is working on. To meet these challenges, there has been an ongoing investment in laboratory tracking systems, project management tools, better DNA assembly algorithms, annotation pipelines, and portals for distributing results. Recognizing the need to help the community deal with this tremendous volume of data, the Integrated Microbial Genomics (IMG) project, now in its eighth major release, holds the genomes of over 750 organisms, allowing researchers to make quick comparisons between whole genomes. In 2006-2007, the JGI finished 82 prokaryotic genomes representing 311 million bases of finished sequence, bringing the total to over 350 draft and finished microbial genomes.

In 2006-2007, the Informatics Department released assemblies and annotations of over 45 eukaryotic genomes to collaborators and the broader genomic community through JGI portals. In support of growing user communities, we organized a large number of training sessions and participated in 17 jamborees or genome annotation workshops.

Supporting this ever-growing volume of data, a new computing facility was

brought online this year. Combining over 400 CPUs and 150 terabytes of storage, computation involving tens of thousands of CPU days is routinely performed. The Informatics Department comprises several subgroups, including: Production Informatics, Assembly, Comparative Genomics, Software Support, IT, Genome Data Systems, and Genome Annotation, and brings together a diverse set of computational skills to serve the JGI mission.

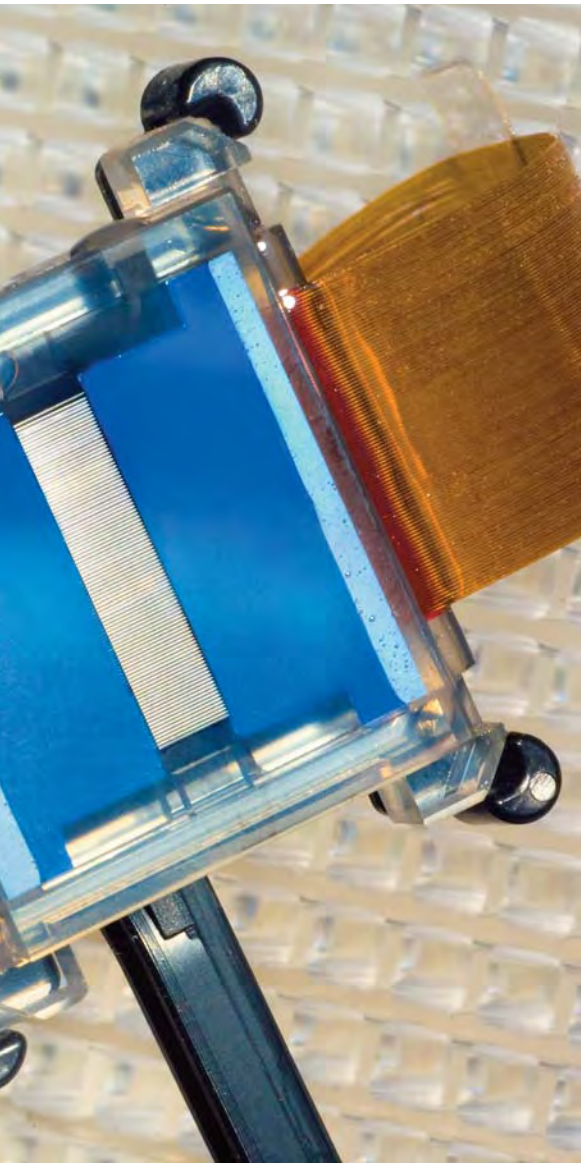
COMPUTATIONAL GENOMICS PROGRAM, DAN ROKHSAR

The Computational Genomics Program develops new analytical tools and data management systems that transform the raw data produced by the JGI into biologically valuable information and insights. These tools are designed to facilitate the use of JGI data by the biological community with a particular focus on eukaryotic genomes. This work is essential for managing and visualizing the expanding body of genome-scale data and linking it to functional and phenotypic information generated at the JGI and elsewhere. A major focus of the program is to work with communities interested in the genomes of the different large eukaryotes sequenced by the JGI to bring these completed genomes to publication.

GENOMIC TECHNOLOGIES PROGRAM, LEN PENNACCHIO

The Genomic Technologies Program develops procedures for implementation of new methods and instruments to increase sequencing capacity and expand genomic capabilities of the JGI. The group tests and develops laboratory protocols for sample preparation and quality control measures to ensure efficient use of the sequencing instruments. The group evaluates new sequencing technologies and makes recommendations for implementing new platforms at the institute. This year, the Genomic Technologies Program developed standard operating procedures for the first next-generation sequencer that is not based on the traditional Sanger sequencing process. The new instrument, the Roche 454, utilizes pyrosequencing technology that performs sequencing onboard the instrument and produces over 100 million bases of data per run. The group trained production personnel and transferred the process to the production group, which now runs the new instruments routinely—primarily to support the microbial genome sequencing efforts.

This year, the group also tested and evaluated another next-generation DNA sequencing platform, the Applied Biosystems SOLiD™ System. This new in-



strument has the potential to produce over one billion bases per run. However, the lengths of the individual reads (about 35 bases) are much shorter than Sanger (about 600 bases) or 454 (200 bases) platforms. A major research effort at the JGI is aimed at developing methods to use this new type of data in existing projects as well as develop new applications that can take advantage of the unique properties of these short-read platforms. The group is currently working on sample preparation methods and developing applications for implementation in production projects. This is a rapidly changing time in the genomics field and the group will continue to explore new technologies that promise greater sequencing throughput and new ways to use this information.

MICROBIAL ECOLOGY PROGRAM, PHILIP HUGENHOLTZ

The Microbial Ecology Program (MEP) uses sequence-based technologies and a combination of computational and experimental methods to obtain a deeper understanding of microbial communities. MEP is heavily involved in the emerging field of metagenomics—cloning, sequencing, and characterizing DNA extracted directly from environmental samples—to obtain an overview of community function and population dynamics. Since environmental shotgun sequencing is in its infancy, MEP is exploring ways to analyze and visualize metagenomic data together with the Microbial Genome Biology Program. With the Genomic Technologies Program, MEP is currently exploring new technological directions: metatranscriptomics for de novo expression profiling of microbial communities, and population sorting using flow cytometry to improve genomic recovery of individual species in communities. Microbial ecosystems that MEP is studying through metagenomics and related techniques include activated sludges, the termite hindgut, Guerrero Negro hypersaline mat in Baja, Mexico, and Lake Vostok in Antarctica.

MICROBIAL GENOME BIOLOGY PROGRAM, NIKOS KYRPIDES

The goal of the Genome Biology Program is to understand the structure and function of various microorganisms and microbial communities and elucidate the evolutionary dynamics that shape their genomes. To accomplish that, members of the program are developing pipelines for processing sequencing data, as well as methods and toolkits that facilitate their own genome analysis efforts and support those of the scientific community at large.

The Genome Biology Program also has a major research focus on bioenergy

with respect to archaeal genomics and methane production, genomics of microorganisms and communities that decompose plant biomass, and bioengineering of lipid metabolism for biodiesel production.

EVOLUTIONARY GENOMICS PROGRAM, PILAR FRANCINO

Research in this program develops computational and experimental approaches to the study of evolutionary genomics. Computationally, the group addresses molecular evolution subjects at different scales, ranging from the impact of mutational biases during DNA sequence evolution to the evolution of new genes and their regulatory regions, to the coevolution of different genomic traits. Experimentally, program members investigate the metagenomics of complex microbial niches to understand the development, population genetics, and coding capabilities of the bacterial communities in these environments. Their newest endeavor aims at establishing an experimental system of bacterial cultures continuously selected for increased efficiency in biofuel production, whose analysis will provide detailed records of actual pathways of short-term genomic adaptation. In this experimental research, members are collaborating with the Genomic Technologies program to develop and apply new approaches that exploit the recent surge of new high-throughput sequencing technologies.

USER SUPPORT PROGRAM, JIM BRISTOW

The User Support Department is charged with facilitating user interactions with the JGI. This group manages the application process and peer review for the Community Sequencing Program. Once proposals are approved for sequencing, the department has responsibility for all aspects of project management, including project planning and initiation, communication of progress or problems with collaborators, coordination of data analysis, and project closeout. The Project Management Office is comprised of managers from LANL and the PGF.

The User Support Department also coordinates the annual user meeting, which in 2007 brought more than 400 of the JGI's users and potential users together with JGI staff to hear state-of-the-art talks on all aspects of se-

quence-based science as well as tutorials on JGI processes, user interfaces, and genomic tools such as the Integrated Microbial Genome tools. Finally, the User Support Department serves as a conduit for feedback from users to the JGI through periodic user surveys, end-of-project questionnaires, and interaction with the JGI Users Executive Committee. An important new cohort of users will be the three DOE Bioenergy Research Centers. It is expected that these Centers will be heavily reliant on all aspects of JGI User Support. For a full list of the CSP 2008 sequencing projects, see http://www.jgi.doe.gov/sequencing/cspseqplans2008.html

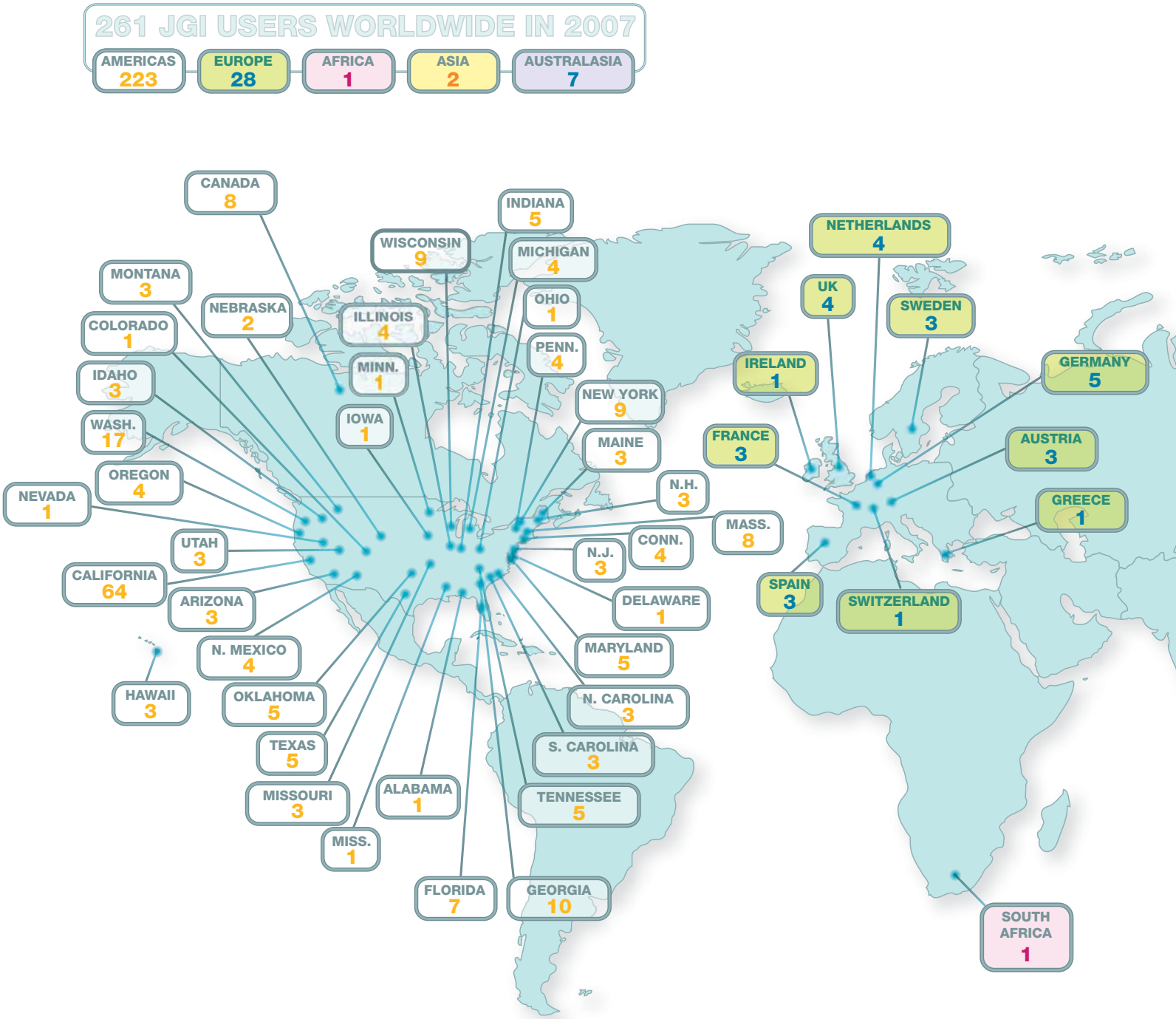
EDUCATION PROGRAM, CHERYL KERFELD

The JGI Education Program develops educational workshops and tools centered on large-scale DNA sequencing and its bioinformatic analysis. These activities target a range of institutions, from research universities to community colleges. The products are also being adapted to high school outreach programs.

The Education Program will host genome-based projects at high schools and undergraduate institutions. The concept is to build a variety of genome-scale research projects specifically for students, and tailored to the existing curriculum and interests of educators.

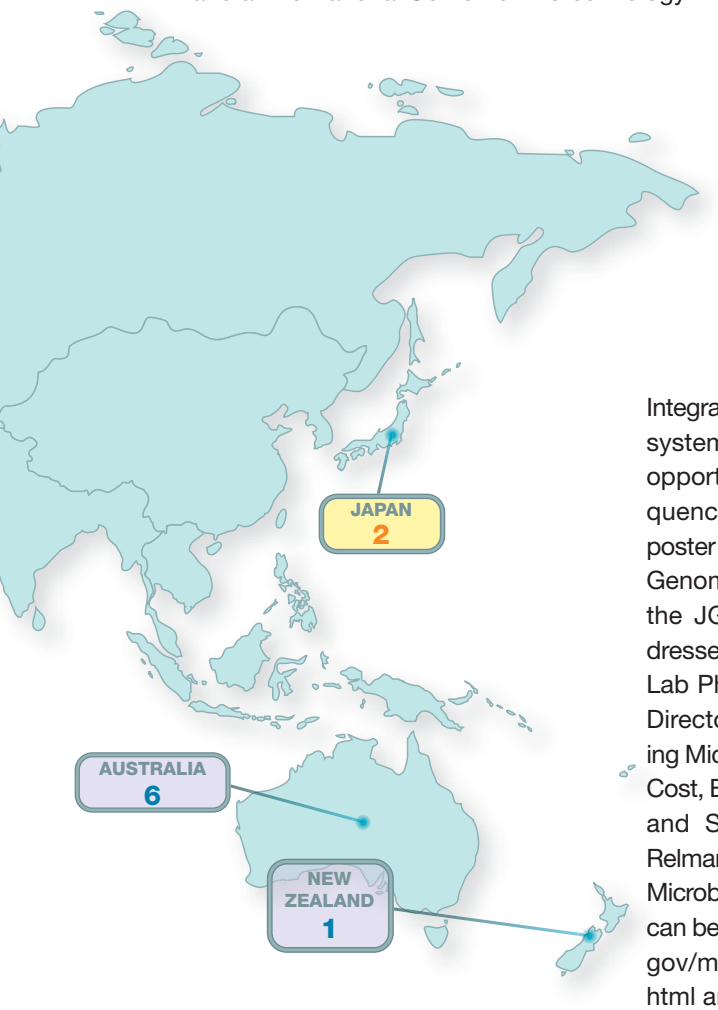
LABORATORY SCIENCE PROGRAM (LSP), GERALD TUSKAN

The Laboratory Science Program (LSP) has provided DOE national laboratory researchers with broad access to high-throughput DNA sequencing in support of DOE mission-relevant projects. First, the LSP has fostered large-scale strategic sequencing projects across the national laboratory system that are aligned with future funding opportunities in DOE's biology programs. Second, it provided small-scale sequencing that met the needs of individual investigators at the national laboratories. In 2008, the LSP, having successfully promoted state-of-the-art DOE mission-based science by investigators at the DOE national laboratories, will be folded into the CSP. For the list of LSP projects, see Appendix D.


261 JGI USERS WORLDWIDE IN 2007
AMERICAS 223
EUROPE 28
AFRICA 1
ASIA 2
AUSTRALASIA 7
CANADA 8
INDIANA 5
WISCONSIN 9
MICHIGAN 4
NETHERLANDS 4
UK 4
SWEDEN 3
MONTANA 3
NEBRASKA 2
OHIO 1
ILLINOIS 4
COLORADO 1
PENN. 4
IRELAND 1
GERMANY 5
IDAHO 3
MINN. 1
NEW YORK 9
IOWA 1
WASH. 17
MAINE 3
FRANCE 3
AUSTRIA 3
OREGON 4
N.H. 3
NEVADA 1
GREECE 1
MASS. 8
UTAH 3
CONN. 4
N.J. 3
CALIFORNIA 64
SPAIN 3
SWITZERLAND 1
ARIZONA 3
DELAWARE 1
N. MEXICO 4
MARYLAND 5
HAWAII 3
OKLAHOMA 5
N. CAROLINA 3
TEXAS 5
S. CAROLINA 3
MISSOURI 3
ALABAMA 1
TENNESSEE 5
MISS. 1
FLORIDA 7
GEORGIA 10
SOUTH AFRICA 1

JGI USER COMMUNITY

The JGI user community draws heavily from academic research institutions in the U.S., along with the national laboratories, federal agencies, and a small number of companies. The broader user community entails collaborations that are global in nature. There are over 300 collaborators on active projects. In addition, the JGI's influence is extensive, considering that hundreds more every year tap sequence data posted by the JGI on its numerous genome portals and at the National Center for Biotechnology Information's (NCBI) GenBank.



The second annual JGI User Meeting, convened from March 28 through March 30, 2007, brought over 400 registrants to Walnut Creek to hear about state-of-the-art, sequence-based science, including some of the most exciting JGI projects. The meeting also featured an Integrated Microbial Genomes (IMG) system workshop and offered the opportunity to learn about new sequencing instrumentation, attend poster sessions, tour the Production Genomics Facility, and mingle with the JGI community. Keynote addresses were delivered by Berkeley Lab Physical Biosciences Division Director Jay Keasling, ("Engineering Microbes for Production of Low-Cost, Effective, Anti-Malarial Drugs") and Stanford University's David Relman ("Explorations of the Human Microbiome"). The complete agenda can be found at: http://www.jgi.doe.gov/meetings/usermtg07/agenda.html and in Appendix F.

An electronic book of abstracts from the meeting can be downloaded at: http://www.jgi.doe. gov/meetings/usermtg07/abstracts.html

The next JGI User Meeting, "Genomics of Energy & Environment," will be held from March 26 through March 28,2008 in Walnut Creek, California. This meeting will specifically emphasize the genomics of renewable energy strategies, biomass conversion to biofuels, environmental gene discovery, and engineering of fuel-producing organisms. A series of presentations by leading scientists advancing these topics will feature a keynote address by Nobel Laureate Steve Chu, Director of Lawrence Berkeley National Laboratory. The meeting will also include informatics workshops and tutorials for the analysis of prokaryotic and eukaryotic genomes, and the evaluation of new sequencing platforms and their applications.

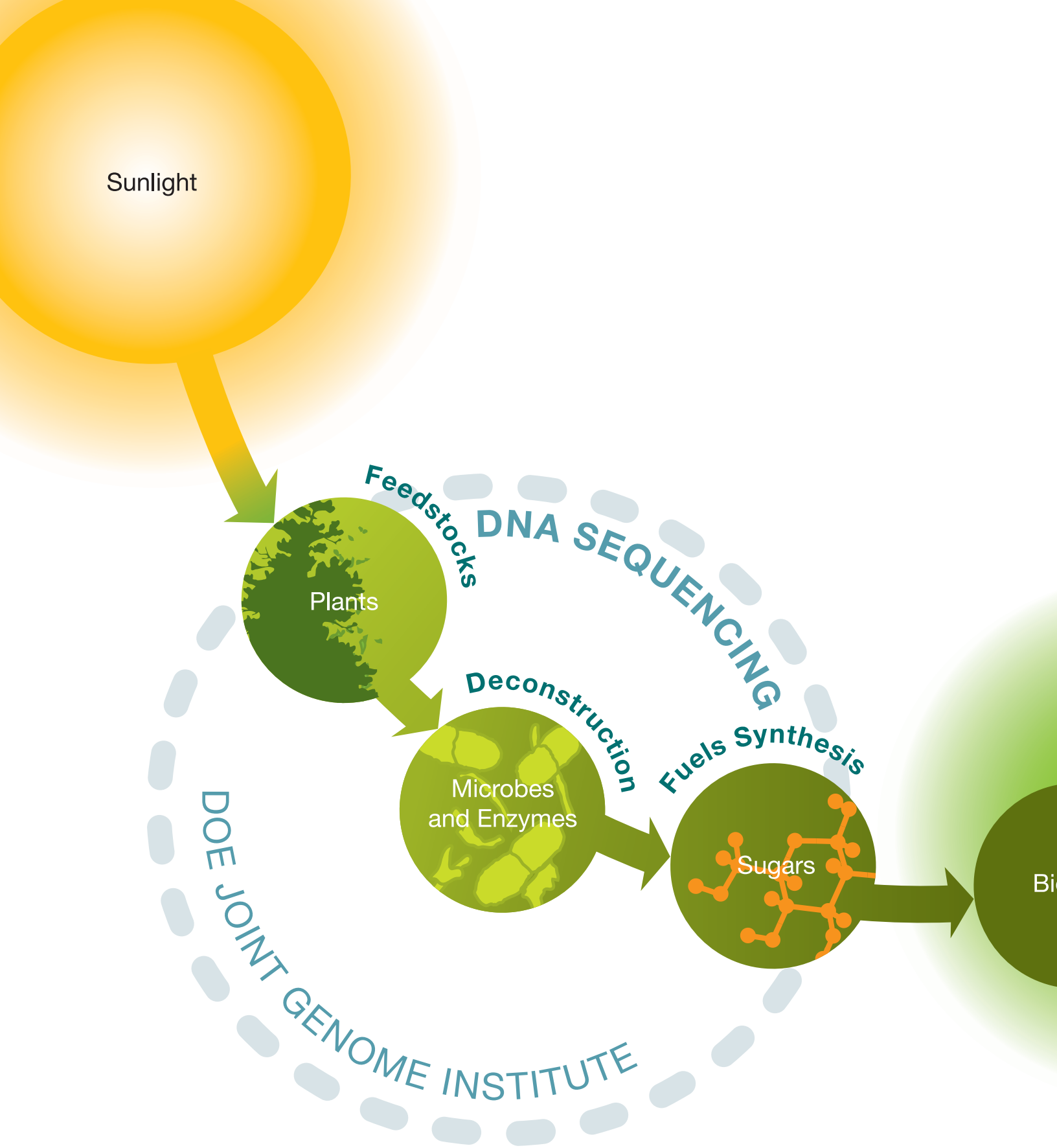
JGI USER MEETING 2007 BY THE NUMBERS:

Registrations:
404

Posters:
71

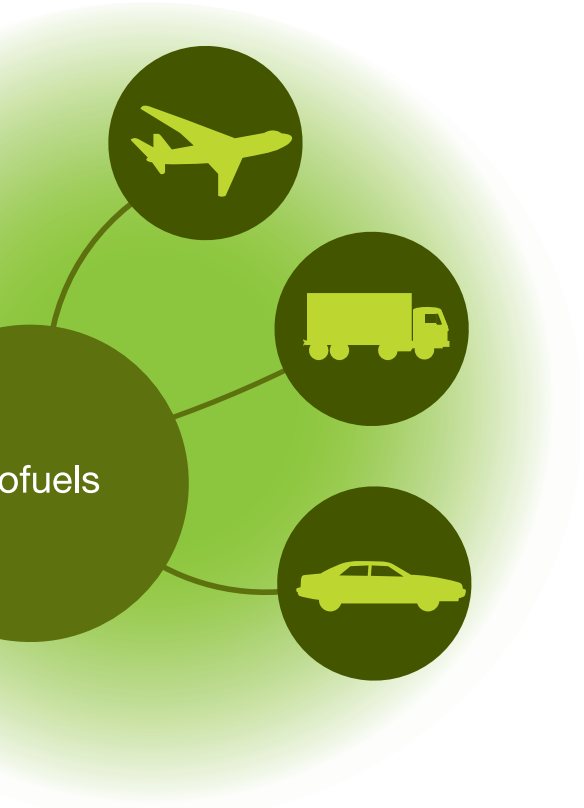
IMG workshop attendees:
137

JGI workshop participants:
248


Sunlight
Feedstocks
Plants
DNA SEQUENCING
Deconstruction
Microbes
and Enzymes
Fuels Synthesis
Sugars
DOE JOINT GENOME INSTITUTE

GENOMICS APPROACHES TO ADVANCING NEXT-GENERATION BIOFUELS

Plants store solar energy through photosynthetic production of sugars that are biochemically transformed into cell wall polymers (long macromolecules, such as cellulose, made of similar or identical subunits linked together). These sugars, mostly glucose, can be fermented into environmentally friendly fuels such as cellulosic ethanol. However, there is a catch. Lignin is a chemical compound that is an integral part of the cell walls of plants and trees, providing strength to cellulose fibers while conferring flexibility to the plant structure. Lignin makes up about 25–30% of the content of most dry wood. In the conversion of cellulose to liquid fuel, lignin stands in the way. Among the keys to overcoming this barrier, while achieving more efficient production of biofuels from woody plant matter, is breaking down the recalcitrant lignin polymer that complicates the efficient extraction of cellulose.



WHAT IS BIOMASS?

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

WHAT IS BIOFUEL?

Biofuel is any fuel derived from biomass. Agricultural products that are currently converted to biofuels include corn for ethanol and soybeans for biodiesel. Efforts are underway to facilitate the generation of cellulosic ethanol. This requires the conversion of non-grain crops—such as switchgrass, a variety of trees, and other woody crops—to biofuels. JGI research supports the DOE commitment to displacing 15% of projected U.S. gasoline consumption with renewable fuels by 2017.

HOW ARE BIOFUELS CURRENTLY PRODUCED?

At present, most ethanol in the U.S. is made from corn. When corn is harvested, the kernels constitute about half of the above-ground biomass, and corn stover (e.g., stalks, leaves, cobs, husks) makes up the other half. Ethanol production from corn grain involves one of two different processes: wet milling or dry milling. In wet milling, the corn is soaked in water or dilute acid to separate the grain into its component parts (e.g., starch, protein, germ, oil, kernel fibers) before converting the starch to sugars that are then fermented to ethanol. In dry milling, the kernels are ground into a fine powder and processed without fractionating the grain into its component parts. Most ethanol comes from dry milling.

GENOMICS APPROACHES TO ADVANCING NEXT-GENERATION BIOFUELS

WHAT ARE THE SHORTCOMINGS OF ETHANOL FROM CORN KERNELS?

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

WHAT IS CELLULOSIC ETHANOL?

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

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

Image Courtesy DOE/NREL

Image Courtesy DOE/NREL

GENOMICS APPROACHES TO ADVANCING NEXT-GENERATION BIOFUELS

Image Courtesy DOE/NREL

WHAT IS BIODIESEL?

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

WHAT IS BUTANOL?

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

WHY DO BIOFUELS PRODUCE FEWER GREENHOUSE GASES?

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

HOW WILL GENOMICS ADVANCE BIOFUELS RESEARCH?

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

GENOMICS APPROACHES TO ADVANCING NEXT-GENERATION BIOFUELS

JGI'S PLANT BIOMASS PORTFOLIO

Raw plant material, or dedicated bioenergy crops, can provide the starting point for making biofuels. The JGI's growing plant portfolio includes many that are, or will soon be, targets of keen interest in the biofuels research community. These include switchgrass, poplar, soybean, eucalyptus, sorghum, foxtail millet, cassava, cotton, and corn.

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

A significant step in realizing the potential of cellulosic feedstocks for ethanol production was achieved with the completion and publication by JGI and its collaborators of the genome sequence of the poplar, *Populus trichocarpa,* the first tree to have its DNA decoded. The publication of the annotated poplar genome has facilitated rapid and effective analysis of the gene network underlying traits related to tree growth, drought tolerance, pest resistance, cell wall composition, and other traits relevant to the DOE mission. Already, modifications have been made in poplar that may accelerate its domestication as a viable bioenergy feedstock.

EUCALYPTUS

Several significant bioenergy-relevant plant genomes are currently being sequenced by the JGI. The largest new plant project approved for the 2008 Community Sequencing Program (CSP) is the eucalyptus tree genome, with a 600-million-nucleotide genome. The biomass production and carbon sequestration capacities of eucalyptus trees match DOE's and the nation's interests in alternative energy production and global carbon cycling. The eucalyptus consortium draws its expertise from dozens of institutions and hundreds of researchers worldwide.

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

The genome of eucalyptus, only the second tree to be sequenced, will advance the community's understanding of eucalyptus's superior properties and enable fruitful comparisons with other species such as poplar.

This international effort, geared to the generation of resources for renewable energy, is led by Alexander Myburg of the University of Pretoria,

South Africa, with Gerald Tuskan of Oak Ridge National Laboratory (and DOE JGI), and Dario Grattapaglia, of EMBRAPA Genetic Resources and Biotechnology (Brazil).

FOXTAIL MILLET

The second largest CSP project selected for sequencing in 2008 is foxtail millet *(Setaria italica).*

This forage crop is a close relative of several prospective biofuel crops, including switchgrass, napiergrass, and pearl millet. In the U.S., pearl millet is grown on some 1.5 million acres. Pearl millet would be useful as a supplement or replacement for corn in regions that suffer from drought and low-fertility soils.

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

SOYBEAN

The soybean, *Glycine max*, is the principal U.S. source of the renewable alternative fuel, biodiesel. Biodiesel has the highest energy content of any current alternative fuel and is much more environmentally friendly than comparable petroleum-based fuels. Biodiesel degrades rapidly in the environment and burns more cleanly than conventional fuels, releasing only half the pollutants and reducing the production of carcinogenic compounds by more than 80%.

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

GENOMICS APPROACHES TO ADVANCING NEXT-GENERATION BIOFUELS

Soybean sequencing will allow for crop improvements and the effective application of this plant for renewable fuel production. Knowing which genes control specific traits, researchers could change the type and quantity of oil produced by the crop, as well as produce soybean plants that are more resistant to drought and disease. Principal investigators on this project are Gary Stacey (National Center for Soybean Biotechnology, University of Missouri), Randy Shoemaker (USDA Agricultural Research Service), Scott Jackson (Purdue University), Bill Beavis (National Center for Genome Resources), Daniel Rokhsar (DOE JGI).

SORGHUM

In 2007, the JGI released an assembly of the 770 million base sorghum genome through Phytozome (http://www.phytozome. net/sorghum). One of the world's leading grain crops, sorghum is also an important model for tropical grasses, and is a logical complement to *Oryza* (rice), the first monocot plant to be sequenced. Sorghum is representative of the tropical grasses in that it uses "C4" photosynthesis, with a complex combination of biochemical and morphological specializations resulting in more efficient carbon assimilation at high temperatures. By contrast, rice is more representative of temperate grasses, using "C3" photosynthesis.

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

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

BRACHYPODIUM— A MODEL GRASS

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

The sequencing of *Brachypodium* will be undertaken via a two-pronged strategy: the first, a whole-genome shotgun sequencing approach, is a collaboration between John Vogel and David Garvin, both of the USDA, and

Michael Bevan at the John Innes Centre in England; and the second, an expressed gene sequencing effort, led by Todd Mockler and Jeff Chang at Oregon State University, with Todd Michael of The Salk Institute for Biological Studies, and Samuel Hazen from The Scripps Research Institute.

CASSAVA

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

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

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

PHYSCOMITRELLA—THE FIRST MOSS GENOME

Messages from nearly a half-billion years ago, conveyed via the inventory of genes sequenced from a present-day moss, provide clues about the earliest colonization of land by plants. The JGI was among the leaders of an international effort uniting more than 40 institutions to complete the first genome sequencing project of a nonvascular land plant, the moss *Physcomitrella patens.* The team's insights into the code that enabled this seminal emergence and dominance of land by plants were published December 13, 2007 online in Science Express.

Physcomitrella is to flowering plants what the fruit fly is to humans; that is, in the same way that the fly and mouse have informed animal biology, the genome of this moss will advance our exploration of plant genes and their functions and utility. Traits such as those that allow plants to survive and thrive on dry land will be useful in the selection and optimization of crops that may be domesticated for biomass-to-biofuels strategies.

Physcomitrella, with a genome of just under 500 million nucleotides and possessing nearly 36,000 genes (about 50% more than are thought to be in the human genome), is the first bryophyte to be sequenced. Bryophytes are nonvascular land plants that lack specialized tissues (phloem or xylem) for circulating fluids. Rather, they possess specialized tissues for internal transport. They neither flower nor produce seeds, but reproduce via spores.

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

much more rapid lifecycle, which also greatly facilitates experimental studies of cell walls. Thus, the completion of the genome is an important step forward in facilitating basic research concerning the development of cellulosic biofuels.

There is also a clear connection with this work and the intensifying interest in the global carbon cycle. The moss system is proving useful for studies of photosynthesis, among many other processes. One of these is the ability of mosses to withstand drought and, in some cases, complete desiccation, which will provide us with a model experimental system to identify genes and gene networks that might be involved in, and related to, seed desiccation in flowering plants. *Physcomitrella* is well placed phylogenetically to fill in the large gap between the unicellular green alga *Chlamydomonas,* also sequenced by the JGI, and the flowering plants.

It is anticipated that having the full *Physcomitrella* genome available to the public greatly advances bioinformatic comparisons and functional genomics in plants—an example of how phylogenetics can integrate with functional and applied studies.

Furthermore, unlike vascular plant systems, specific moss genes can be targeted and deleted to study their function in important crop processes. They can also be replaced with genes from crop plants to enable the study the evolution of gene function.

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

CONIFERS—ABUNDANT BIOMASS

Pines and most conifers have extremely large genomes—many-fold larger than poplar, for instance—having accumulated multiple genome duplications over evolutionary time.

This phenomenon has made sequencing pines prohibitively expensive and time-consuming. A strategy to more rapidly extract valuable information from these genomes is to sequence expressed sequence tags (ESTs). ESTs are fragments of DNA sequence that serve as a tool for the identification of genes and prediction of their protein products and their function. Conifer forests are among the most productive in terms of annual lignocellulosic biomass generation, and coniferous trees are the preferred feedstock for much of the forest products industry. Climate change and exotic forest pests are threatening conifer populations.

Breeding programs to improve conifers will benefit from access to this EST genomic resource. The principal investigators for this project are Jeffrey Dean (University of Georgia), Glenn T. Howe (Oregon State University), Kathleen D. Jermstad (U.S. Forest Service), David B. Neale and Deborah L. Rogers (University of California, Davis).

HARNESSING WASTE BIOMASS

It is estimated that 236 million tons of municipal solid waste is produced annually in the U.S., 50% of which is biomass. Converting organic waste to renewable biofuel represents an appealing option to exploit this potential resource. In California alone, it is estimated that 22 million tons of organic waste is generated annually, which if converted by microbial digestion, could produce biogas—primarily methane and carbon dioxide—equivalent to 1.3 million gallons of gasoline per day. When biogas is cleaned of its particulates and carbon dioxide is removed, it has the same characteristics as natural gas, also known as biomethane. Yet little is known about the microorganisms involved and their biology. In 2008, the JGI will undertake a genomic study of a biogas-degrading community with the aim to optimize the anaerobic digestion process and promote conversion of biomass into biofuel.

Image Courtesy DOE/NREL

JGI'S MICROBIAL PORTFOLIO

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

Symbiotic Organisms

LACCARIA

The DNA sequence of *Laccaria bicolor,* a fungus that forms a beneficial symbiosis with poplar and other trees and inhabits one of the most ecologically and commercially important microbial niches in North American and Eurasian forests, has been determined by the JGI. The complete *Laccaria* genome sequence was announced July 23, 2006, at the Fifth International Conference on Mycorrhiza in Granada, Spain, by an international consortium comprised of the JGI, Oak Ridge National Laboratory (ORNL), France's National Institute for Agricultural Research (INRA), the University of Alabama in Huntsville (UAH), Ghent University in Belgium, and additional groups in Germany, Sweden, and France. In December 2007, the analysis of the *Laccaria* genome, yielding insights into the mechanisms of symbiosis between the fungus and the roots of plants, was accepted for publication in the journal *Nature*.

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

The *Laccaria* genome sequence will provide the global research community with a critical resource to develop faster-growing trees for producing more biomass that can be converted to fuels, and for trees capable of capturing more carbon from the atmosphere.

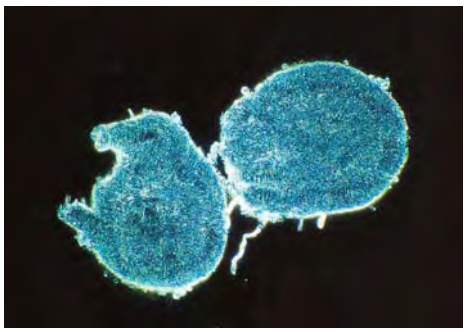
This research will advance the understanding of how functional genomics of this symbiosis enhances biomass production and carbon management, particularly through the interaction with the poplar tree, also sequenced by the JGI. It will now be possible to harness the interaction between these species and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

PLANT PATHOGENS THAT HINDER BIOMASS YIELD

The soybean pathogen *Heterodera glycines* has been selected for sequencing in 2008. Soybean is a major oil, feed, and export crop, with $17 billion annually in unprocessed crop value in the U.S. alone. Soy biodiesel is a leading contender for a renewable, alternative vehicle fuel with a high energy density. Soybean has the environmental and energy advantage of not requiring the use of nitrogen fertilizer. *H. glycines* is the most significant soybean pathogen in the U.S.; thus, sequencing its genome would aid in the development of control strategies and directly contribute to soybean yield enhancement.

Laccaria bicolor

Heterodera glycines



Frankia

GENOMICS APPROACHES TO ADVANCING NEXT-GENERATION BIOFUELS

Microbes That Break Down Biomass

TERMITE HINDGUT MICROBES

Termites—notorious for their voracious appetite for wood and causing billions of dollars in damage per year—are providing insights into the molecular machinery that can enable the efficient breakdown of lignocellulose into fermentable substrates for biofuels. Termite stomachs are a rich source of microbes, producing enzymes that can be employed to improve the conversion of wood or waste biomass to valuable biofuels.

Termites digest massive quantities of wood employing an array of specialized microbes in their hindguts to break down the cell walls of plant material and catalyze the digestion process. Industrial-scale DNA sequencing by the JGI was key to identifying the genes employed by the termite gut microbes in the breakdown of lignocellulosic material. The task now is to discover the metabolic pathways generated by these structures to figure out how nature digests plant materials. The plan is to use this information to synthesize the novel enzymes discovered in this project to explore their capabilities for lignocellulose degradation with the view to eventually facilitate industrial-scale processing of biomass for biofuels.

How the study was conducted Like cows, termites have a series of stomachs, each harboring a distinct community of microbes under precisely defined conditions. These microbes within the insect are tasked with particular steps along the conversion pathway of woody polymers to sugars. The mandibles of the insect chomp the wood into bits, but the real work is conducted in the dark recesses of the gut, where the enzymatic juices exuded by microbes attack and deconstruct the lignocellulose.

The insects were collected in the rainforest of Costa Rica, the world's geographic hotbed of biodiversity for termites. The team discovered a massive, tumor-like nest of termites clinging to an otherwise nondescript tree. With a flick of a machete, the contents of this dense network of tunnels forged from wood waste were revealed, along with a frenzy of higher termites from the genus *Nasutitermes*, which are only about the size of the date imprinted on a penny.

Foregoing the funnel-headed "soldiers," the project focused on the larger "workers," with bulbous heads and inflated bellies. In the laboratory of INBio, researchers armed with fine forceps and needles painstakingly extracted the contents of the workers' hindguts. Each sample was barely visible to the naked eye, and care was taken not to contaminate it with material from neighboring stomachs. Contents from 165 specimens were purified, yielding only a few valuable drops—a veritable microbial mosh pit—that was sent on ice to Verenium for DNA extraction and preparation, then on to the JGI for sequencing.

What was discovered From the sample, about 71 million letters of fragmented genetic code were elaborated and computationally reassembled, teasing out the identities of the microbial players in the mixture and the metabolic profile of the enzymes that they produce. From this reconstructed liquid puzzle emerged the identities of a dozen different phyla—broad groupings of microbial life forms.

In the termite hindgut alone, more than 500 genes related to the enzymatic deconstruction of cellulose and hemicellulose were identified. The termite gut meta-genome dataset is publicly available, along with an annotated view, through the DOE JGI's metagenome data management and analysis system, IMG/M (http://img. jgi.doe.gov/m).

While termites can efficiently convert milligrams of lignocellulose into fermentable sugars in their tiny bioreactor hindguts, scaling up this process so that biomass factories can produce biofuels more efficiently and economically is another story. To get there, the set of genes with key functional attributes for the breakdown of cellulose will need to be

GENOMICS APPROACHES TO ADVANCING NEXT-GENERATION BIOFUELS

refined, and this study represents an essential step along that path.

The termite hindgut project represents the first of several metagenomic projects at the JGI exploring the molecular machinery nature employs to digest lignocellulose.

The genomic sequencing and analysis of termite gut microbes by the JGI, the California Institute of Technology, Verenium Corporation (a biofuels company), INBio (the National Biodiversity Institute of Costa Rica), and the IBM Thomas J. Watson Research Center, was highlighted in the November 22 edition of *Nature.*

Other projects in the sequencing queue at the JGI are metagenomes from cow rumen, the Tammar wallaby forestomach, the Asian longhorned beetle gut, and other exotic species that promise to be treasure troves of enzymes involved in cellulose deconstruction. Additional JGI projects that entail prospecting for promising sources of novel enzymes extend to the hot pools of Yellowstone National Park. Here, extremophiles—microbes that thrive in extreme conditions—represent targets for rapid translation into products and processes for the biofuels and biotechnology industries.

Microbes That Ferment Sugars into Ethanol

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

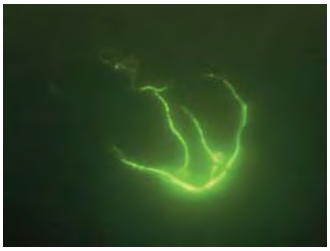
Saccharomyces cerevisiae, brewer's yeast, has long been employed for the fermentation of plant sugars into ethanol—beer and wine being prime examples. However, there are ways to improve the process, yielding higher levels of ethanol from lignocellulose.

One major challenge has been addressed with the characterization of the genetic blueprint of the unicellular fungus, *Pichia stipitis*. The research, entailing the identification of numerous genes in *P. stipitis* responsible for its fermenting and cellulose-bioconverting prowess, and an analysis of these metabolic pathways, was featured last year in *Nature Biotechnology.*

P. stipitis is the most proficient microbial fermenter in nature of the five-carbon "wood sugar" xylose—abundant in hardwoods and agricultural leftovers. Increasing the capacity of *P. stipitis* to ferment xylose, and using this knowledge to improve xylose metabolism in other microbes—such as the yeast *Saccharomyces*—offers a strategy for improved production of cellu-

Asian longhorned beetle

Tammar wallaby



Yellowstone National Park hot pool

losic ethanol. In addition, this strategy could enhance the productivity and sustainability of agriculture and forestry by providing new outlets for agricultural and wood harvest residues.

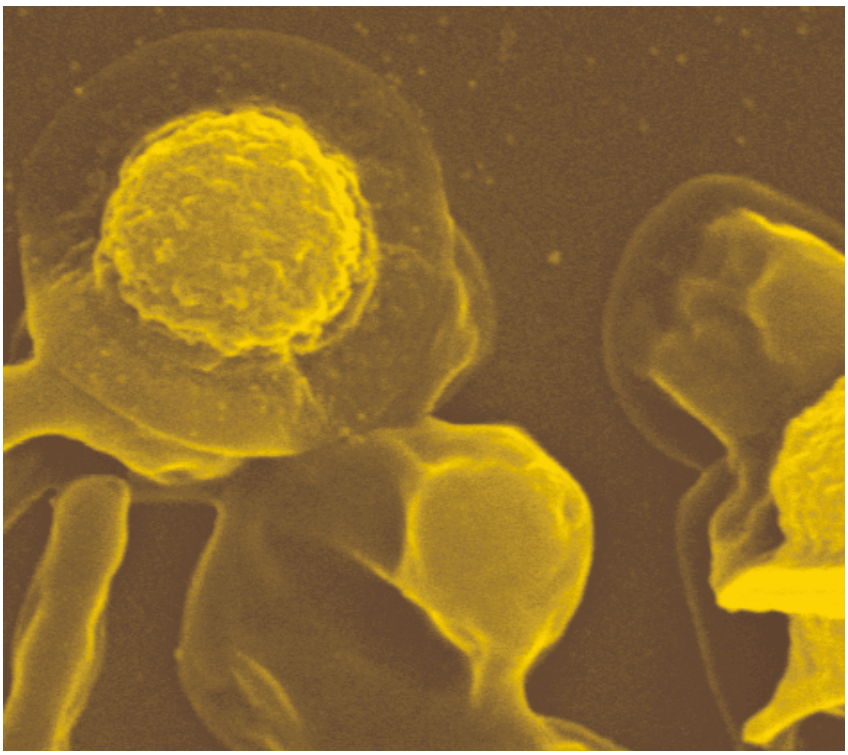
The information embedded in the genome sequence of *Pichia* has helped to identify several gene targets to improve xylose metabolism. Efforts are underway to engineer these genes to drive the fermentation process to higher concentrations of ethanol.

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

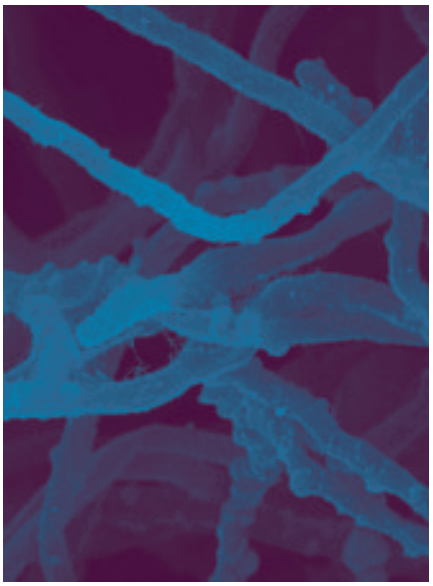
Developing *Pichia* for xylitol and ethanol production has drawn heavily on the JGI sequence data. The sequence data has helped identify previously unknown enzymes that enable this organism to use the diverse sugar components of the plant material. This could lead to improved organisms for simultaneous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

Pichia is but one fungus in an expanding portfolio of targets sequenced by the JGI that can be directed to steps in the biofuel process, such as fermentation. Others include *Clostridium thermocellum,* capable of directly converting cellulosic substrate into ethanol, the white rot fungi *Phanerochaete chrysosporium,* and the oyster mushroom, *Pleurotus ostreatus* that are capable of lignin deconstruction—a necessary step for making cellulose accessible to further enzymatic processes. Another organism slated for sequencing in 2008 is the leaf-degrading fungus *Agaricus bisporus.* Genomic studies of *A. bisporus* target enhanced understanding of the mechanisms employed for efficient conversion of lignocellulose—crucial for the production of fuels and products from renewable biomass. The principal investigator on this project is Thomas W. Jeffries (U.S. Forest Service, Forest Products Laboratory)

Pichia stipitis spores



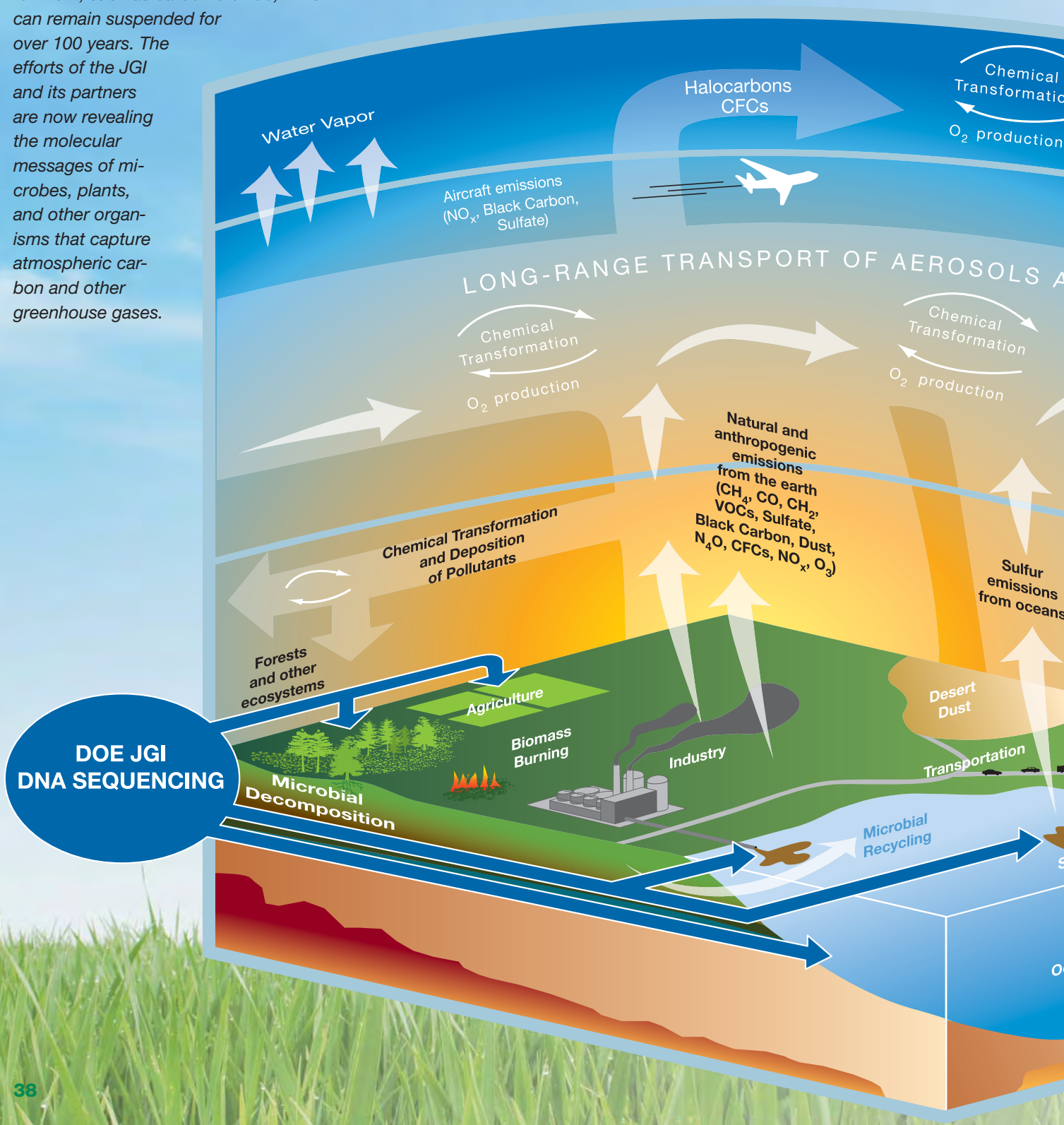
Phanerochaete chrysosporium



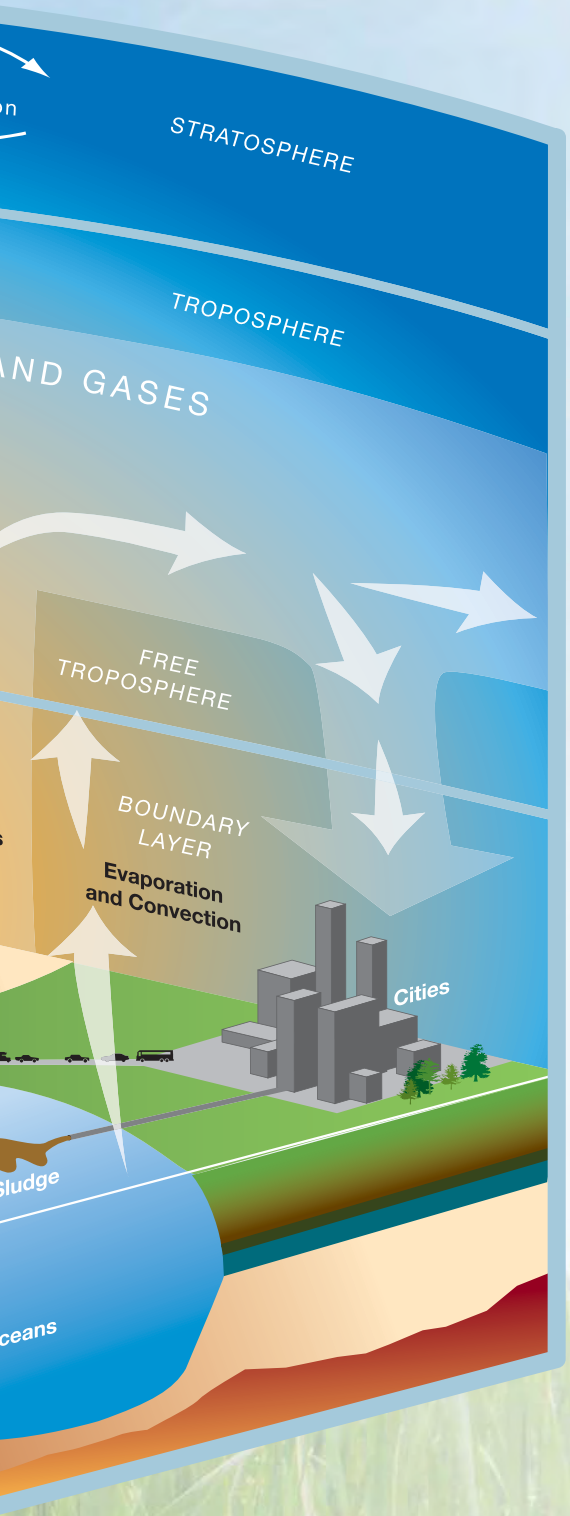
Agaricus bisporus

Pleurotus ostreatus

The atmosphere is a conduit of change, directly affecting human welfare as it circulates around the globe interacting with oceans, land, plants, and animals. Emissions from natural sources and human activities enter the atmosphere and are transported to new geographical locations or to higher altitudes. In this regard, the atmosphere is a long-term reservoir for trace gases that can modify the global environment, such as carbon dioxide, which can remain suspended for over 100 years. The efforts of the JGI and its partners are now revealing the molecular messages of microbes, plants, and other organisms that capture atmospheric carbon and other greenhouse gases.



CARBON CYCLING



Investigations by the JGI and its partners are shedding light on the cellular machinery of organisms and microbes that capture carbon dioxide from the atmosphere. While research continues into new sources of energy that emit little or no carbon, strategies must also be explored for capturing atmospheric carbon dioxide (CO_2)—a chief contributor to global climate change—generated by the use of fossil fuels. Research into the role that microorganisms play in the Earth's natural carbon cycle may lead to new strategies for reducing atmospheric carbon and other greenhouse gases.

Understanding Algae's Role in Photosynthesis and Carbon Capture

The JGI is among the world's leaders in characterizing the genomes of algae. Across a diverse spectrum—from single-celled species to complex multicellular "seaweeds" colored blue-green, red, and brown—algae all possess photosynthetic machinery that capture light and carbon dioxide to produce oxygen and water as byproducts of photosynthesis.

OSTREOCOCCUS

Ocean-dwelling phytoplankton from the genus *Ostreococcus* emerge at the primitive root of the green plant lineage, dating back nearly 1.5 billion years. Today, these microscopic, free-living creatures, among the smallest eukaryotes ever characterized, barely a micron in diameter, contribute to a significant share of the world's total photosynthetic activity. These "picophytoplankton" also exhibit great diversity that contrasts sharply with the dearth of ecological niches available to them in aquatic ecosystems. This observation, known as the "paradox of the plankton," has long puzzled biologists.

The analysis of DNA sequence from these green algae have provided new insights into the mystery of how new species of plankton evolve—and further highlights their critical role in managing the global cycling of carbon. This study, from a group led by the JGI; the Scripps Institution of Oceanography; the University of California, San Diego; and the Pierre & Marie Curie University, was published last May in the Proceedings of the National Academy of Sciences (PNAS).

Plumbing the depths of molecular-level information of related species, genomics offers a novel glimpse into this paradox. The researchers compared the genomes of two *Ostreococcus* species, *O. lucimarinus* and *O. tauri,* and saw dramatic changes in genome structure and metabolic capabilities.

Assimilation of atmospheric CO_2 by marine phytoplankton is a global-scale process that is responsible for about half of the biosphere's net

primary production. This active absorption of hundreds of millions of tons of carbon per day is essential for maintaining control of the planet's climate by counteracting greenhouse effects due to human activities. This storage capacity is affected by changes in the photosynthetic efficiency of the algae, which in turn is linked to the environmental conditions experienced by these organisms in their environment.

The ecology of picoeukaryotes has thus become an intense field of investigation over the last decade as these microalgae, although representing a minor component of the plankton, nevertheless play major roles in oceanic biomass production.

With even more picoplankton genomes in the sequencing queue, the JGI will enable the community to secure a better grasp on the mechanisms of species adaptation and the great diversity of biological pathways operating in the oceans. Also, the comparative analyses that lie ahead will help to predict the roles of these organisms in contributing to primary marine productivity.

CHLAMYDOMONAS

The single-celled alga *Chlamydomonas reinhardtii,* while less than a thousandth of an inch in diameter, or about one-fiftieth the size of a grain of salt, is packed with many ancient and informative surprises. Affectionately known to its large research community as *"Chlamy,"* the alga is a powerful model system for the study of photosynthesis and cell motility. The genes that encode the alga's "flagella," which propel it much like a human sperm tail, were also cataloged in this study. Defects in these genes are associated with a growing list of human diseases.

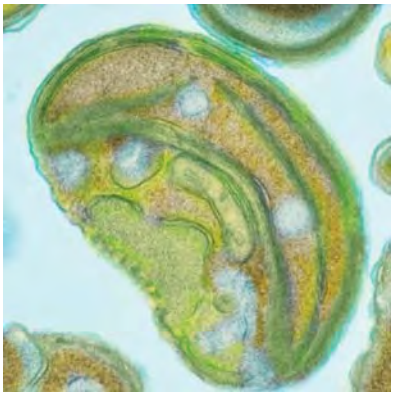
The genome analysis of this tiny green alga has uncovered hundreds of genes that are uniquely associated with carbon dioxide capture and generation of biomass. Among the 15,000-plus genes revealed in the study are those that encode the structure and function of the specialized organelle that houses the photosynthetic apparatus, the chloroplast, which is responsible for converting light to chemical energy. The genome also provides a glimpse back through time to the last common ancestor of plants and animals.

The *Chlamy* genome is like a green time capsule that affords a view into the complex core machinery that gave rise to today's energy-capturing and oxygen-producing chloroplasts. Interest in *Chlamy* centers on its keen ability to efficiently capture and convert sunlight into energy, and its role in managing the global pool of carbon. The sequence analysis presents a comprehensive set of genes—the molecular and biochemical instructions—required for these capabilities. With the data now publicly available, new strategies will begin to surface for biology-based solar energy capture, carbon assimilation, and detoxification of soils by employing algae to remove heavy metal contaminants. The analysis will also shed light on the capabilities of related algae that can produce biodiesel and biocrude as alternatives to fossil fuels.

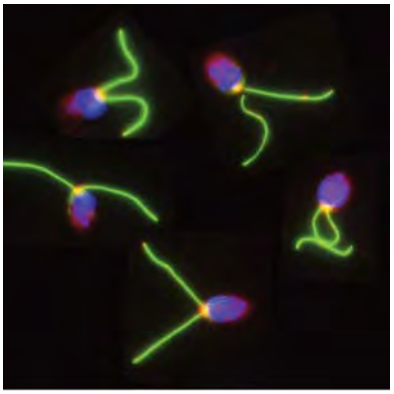
The work has generated a clear roadmap for exploring the roles of numerous genes in photosynthetic function, for defining the structure and dynamic aspects of flagellar function, and for understanding how the soil environment, with its large fluctuations in nutrients, has molded the functionality of organisms through evolutionary time.

The published analysis of approximately 120 million units of DNA sequence generated by the JGI showed that *Chlamy* shares nearly 7,000 genes with other organisms; more than a third of these are shared by both humans and flowering plants, which helps support the argument for their common ancestry. Many of these genes are normally associated with animals, such as those that describe the circuitry for flagella, enabling this alga to swim. Others have affinity with the earth's early photosynthetic organisms, cyano-bacteria, dating back more than three billion years ago into Precambrian times when biodiversity began its explosive proliferation.

*Ostreococcus**



Chlamydomonas



Porphyra

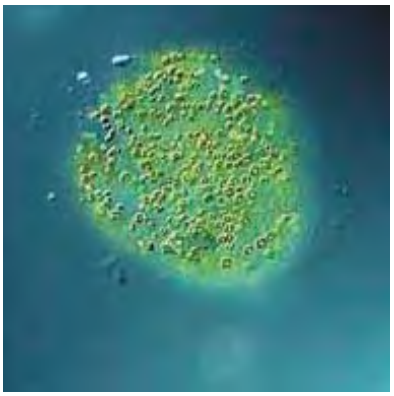
The project, led by the JGI; the University of California, Los Angeles; the Carnegie Institution; and including contributions from over 100 international collaborators, was featured in the October 12 edition of the journal *Science.*

PORPHYRA

Among the largest genome projects selected for 2008 is the marine red alga *Porphyra purpurea.* The ocean plays a key role in removing carbon dioxide from the atmosphere with the help of marine photosynthetic organisms like *Porphyra* consuming the carbon and releasing oxygen. *Porphyra* species are among the most common algae in the intertidal and subtidal zones of temperate rocky shores in both the northern and southern hemispheres. Understanding the effects of elevated climatic stresses on photosynthetic organisms would benefit from genome-enabled studies of carbon fixation in *Porphyra*, because of this organism's great diversity of light-harvesting and photo-protective strategies. The *Porphyra purpurea* project is led by Susan H. Brawley (University of Maine).

PHAEOCYSTIS

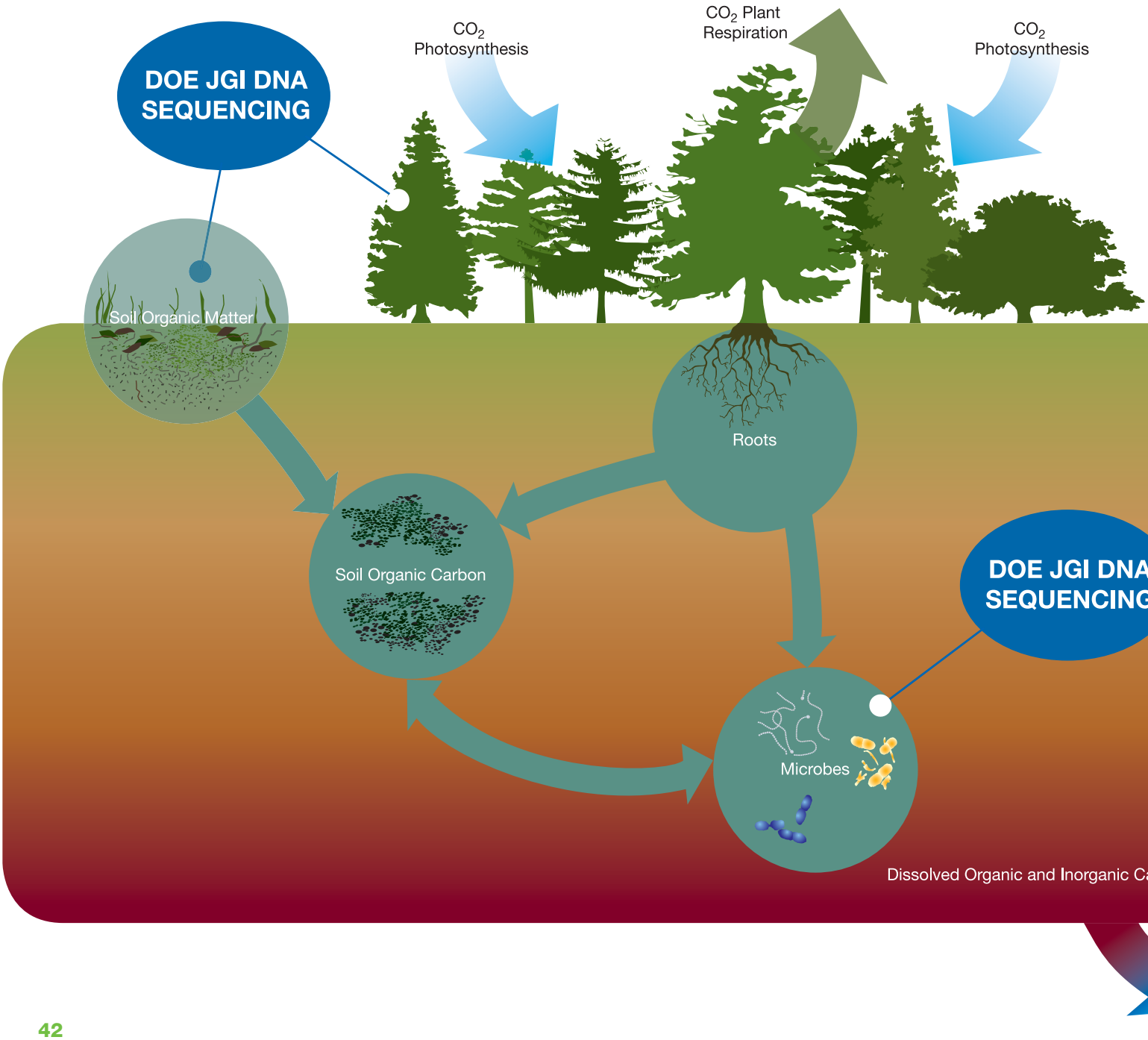
The *Phaeocystis* genus contributes approximately 10% of annual global marine primary photosynthetic production, equivalent to four billion metric tons of carbon dioxide captured or "fixed" annually—reinforcing its importance for the study of the global carbon cycle and carbon sequestration. The *Phaeocystis globosa* project is led by Andy E. Allen, Ian T. Paulsen, and Jonathan H. Badger (The Institute for Genomic Research); and Peter G. Verity and Marc E. Frischer (Skidaway Institute of Oceanography)



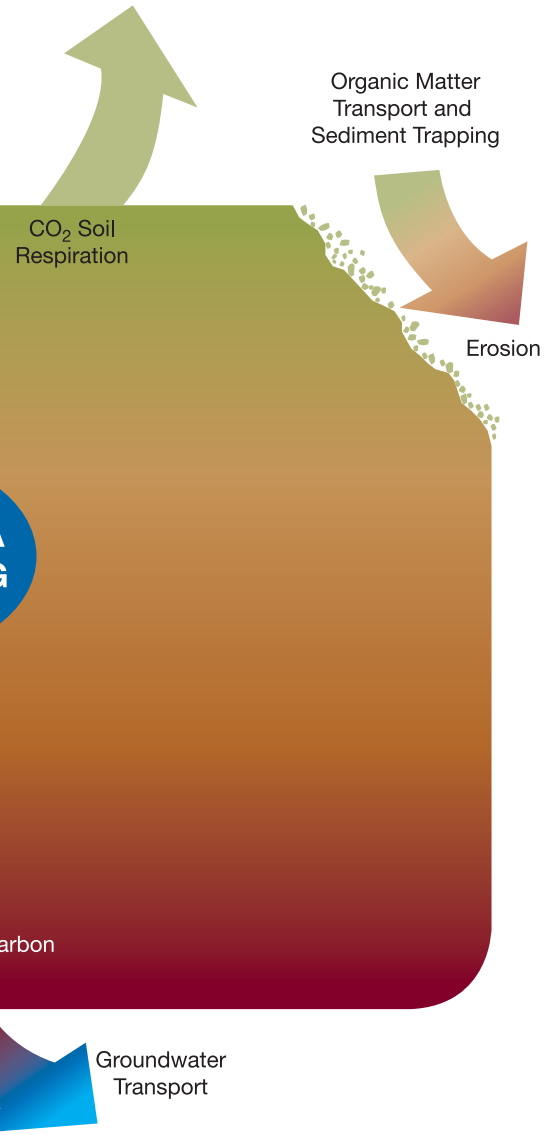
Phaeocystis

*This Wikipedia image is licensed under the Creative Commons Attribution ShareAlike 2.5 License.

Microbial processes influence the absorption and movement of carbon in soil and plants, the conversion of sunlight to energy, and detoxification of soils through the removal of heavy metals. Over 75% of the carbon in terrestrial ecosystems is stored in forests, with over half of this carbon found in soil organic matter (SOM). JGI efforts continue to reveal the genes associated with photosynthetic function, the enhancement of carbon storage in biomass and soils, and how nutrient fluctuations in organic matter shape the functionality of soil organisms.



MICROBIAL BIOREMEDIATION



Bioremediation is a technology that can be used to reduce, contain, or eliminate hazardous waste. Microorganisms can transform and degrade many contaminants in soil and water. Organic contaminants such as hydrocarbon fuels can be degraded into carbon dioxide, while toxic and radioactive metals can be immobilized or removed from the environment. These solutions are made possible by the microorganisms that have adapted to live in contaminated environments. DNA sequencing provides unique signatures of individual microbes and their communities, enabling the JGI to help identify new solutions for cleaning up hazardous waste sites and restoring the environment.

Among the CSP selections to be sequenced in 2008 is an organism that holds promise for bioremediation applications. The first ciliated protozoan genome, *Tetrahymena thermophila,* is a microbial model organism for discovering fundamental principles of eukaryotic biology. It will allow improved construction and stability of cell lines for the over-expression of proteins, including cellulase enzymes to overcome the limiting hurdle of biomass-to-biofuel production, and metal-chelating proteins to enhance the already superior capacity of ciliates for bioremediation of toxic heavy metals in industrial effluents.

Over 75% of the carbon in terrestrial ecosystems is stored in forests. More than half of this carbon is found in soil organic matter (SOM). Recent studies have indicated that ectomycorrhizal fungi like *Paxillus involutus* provide the dominant pathway through which carbon enters the SOM. These fungi are also known to protect plants from toxic metals. Thus, in 2008, the JGI will sequence *Paxillus* to inform the further development of metal-tolerant fungal associations that would provide a strategy for active remediation of metal-contaminated soils.

Microbial Managers of the Nitrogen Cycle

The Earth's atmosphere consists of about 78% nitrogen. The global nitrogen cycle describes a set of biogeochemical processes crucial for maintaining the food chain and the balance of life on earth. Nitrogen is incorporated in all nucleic acids and is essential for photosynthesis and the further generation of biomass. Nitrogen fixation is the process that converts nitrogen from the atmosphere into forms usable by organisms. Some nitrogen-fixing microorganisms form a symbiotic relationship with plants, capturing nitrogen and converting it into compounds, such as ammonia, that enrich the soil. Some forms of nitrogen are highly soluble and thus can penetrate through the soil matrix and accumulate and ultimately contaminate groundwater.

ANAMMOX

Anammox bacteria are able to synthesize the rocket fuel hydrazine from ammonia and hydroxylamine. Insight into the genes and proteins involved in this reaction may be the basis for further optimization of the production of this potent fuel in a suitable biological system. Also, *anammox* bacteria are responsible for about 50% of the processing of ammonia to nitrogen gas in the ocean. In marine ecosystems, the carbon and nitrogen cycles are closely connected. More information about the regulation and mechanism of CO_2 sequestration by anammox bacteria in the ocean will contribute to our understanding of the global biogeochemical cycles and their impact on climate change.

Microbial Management of Wastewater

In addition to employing microbial strategies to cope with accumulation of nitrates in groundwater, there are projects underway to deal with the delicate balance of phosphorus in wastewater.

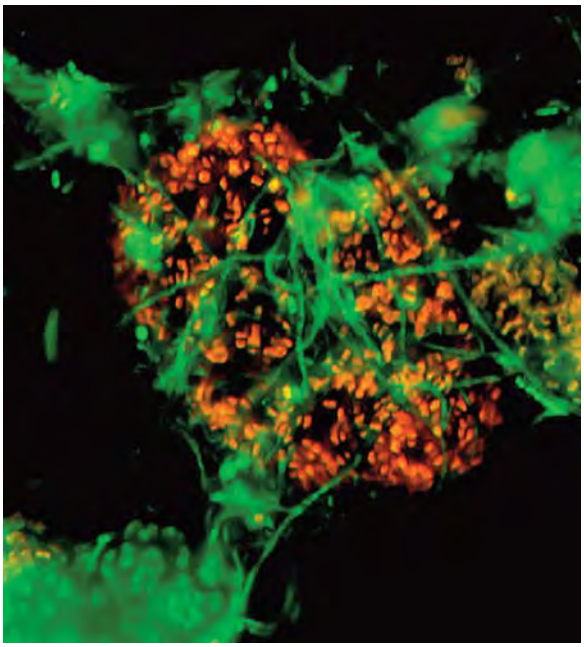
ACCUMULIBACTER

Enhanced Biological Phosphorus Removal (EBPR) is a wastewater treatment process used throughout the world to protect surface waters from accelerated stagnation and depletion of oxygen. EBPR can be unreliable and often requires expensive backup chemical treatments to protect sensitive receiving waters. This project will shed light on the microbial population dynamics leading to better use and management of these important environmental systems.

The metagenomic strategy entails generating DNA sequence information directly from samples of sewage sludge to provide a blueprint of the genes and hence the metabolic potential of microbial communities in the wastewater environment, with a view to understanding how the system works for predicting and averting failures or crashes. The researchers were able to obtain a complete genetic blueprint for a key player, *Candidatus Accumulibacter phosphatis.*

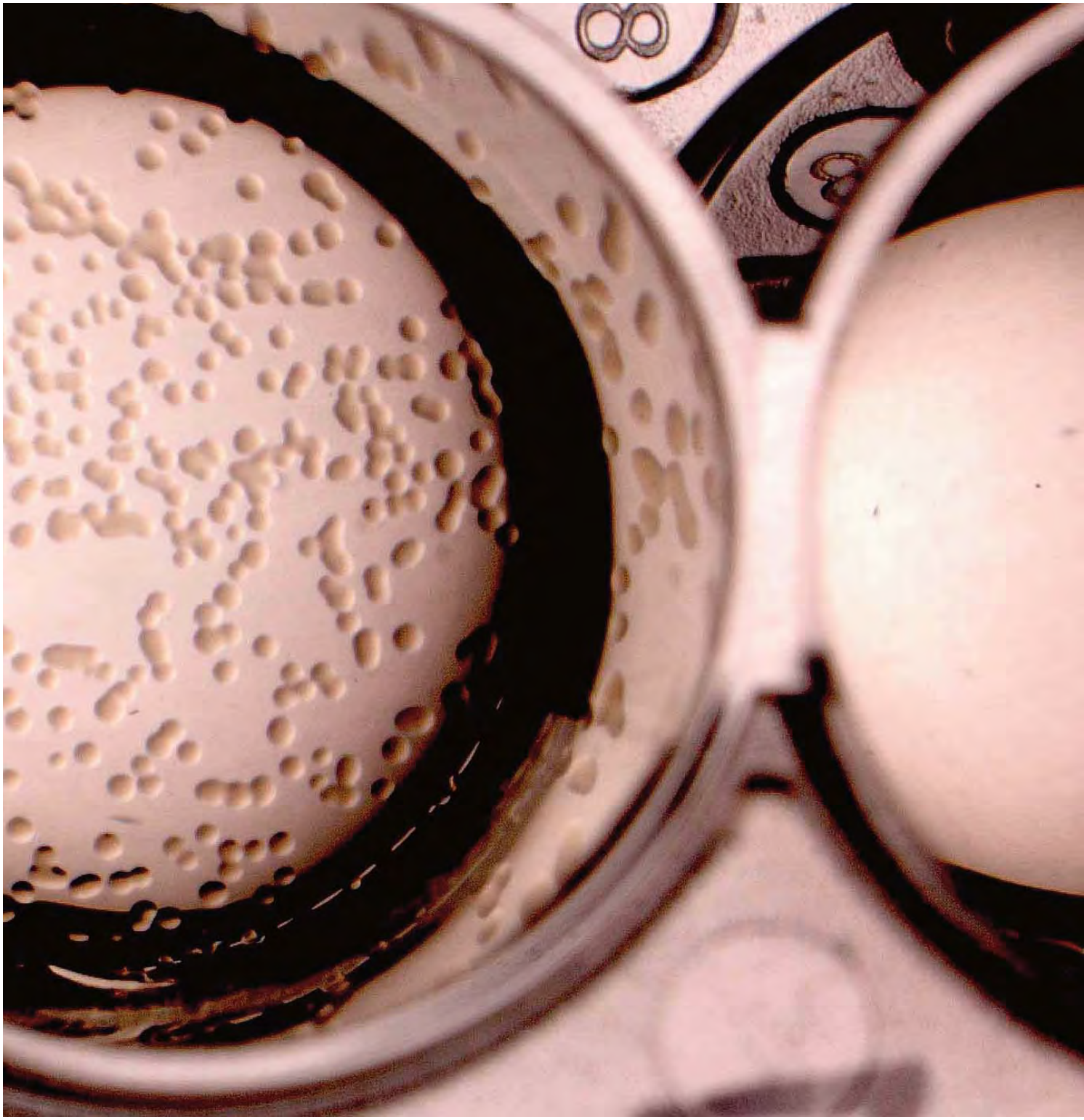
Background image: Anammox

Wastewater sludge sampling



Accumulibacter

Bacteria that grow normally on a plate (left) cannot grow when a toxic gene is transferred into them (right).



EXPLORATORY SEQUENCE-BASED SCIENCE

Photo by Rotem Sorek, JGI

While the vast majority of projects at the JGI relate directly to DOE mission areas, a small number of projects are either deemed to be pioneering next-generation applications of DNA sequencing by the JGI senior management, or are supported by non-DOE funding.

Genomic Encyclopedia for the Bacteria and Archaea (GEBA)

Bacterial and archaeal genome sequences currently available to the scientific community are highly biased with respect to the narrow phylogenetic diversity of the species from which they come.

To address this bias, the JGI launched a pilot project in 2007 under the umbrella of the Genomic Encyclopedia for the Bacteria and Archaea (GEBA), to select and sequence 100 bacterial and archaeal genomes based on the phylogenetic position of organisms in the tree of life. Jonathan Eisen, who has appointments with the JGI and UC Davis, is driving the GEBA campaign.

GEBA offers a systematic approach that addresses areas of DOE emphasis and advances broader scientific community interests. The goals include:

- Improved protein family identification across species
- Addition of metagenomic data
- Targeting novel organisms to facilitate gene discovery
- Better tracking of classification and evolutionary history of microbial species
- To contribute to the discovery of enzymes and pathways of DOE mission relevance
- Anchors metagenomic sequencing projects to reflect a more accurate representation of microbial community composition
- More refined microbial taxonomy enables more accurate characterization of sequenced genomes

The organism selection process for the pilot project is based on a combination of objective analysis of the ribosomal RNA tree of life and consultation with a scientific advisory board (see Appendix E). All genome sequence data is released to the community through the JGI's Web site and GenBank.

Functional Analysis of Horizontal Gene Transfer

In nature, microbes share genetic information so readily that using genes to infer their species position on the evolutionary tree of life was thought to be futile. Thus, the whole concept of a tree of life has been called into question, in favor of depicting relationships between living things by a web or net-

work. A study led by JGI scientists, published in the October 19 edition of the journal *Science,* has characterized barriers to this gene transfer by identifying genes that kill the recipient bacterium upon transfer. These lethal genes provide better reference points for building phylogenic trees—the means to verify evolutionary relationships between organisms. While not resolving the tree-versus-web debate, the survey does provide additional information with which to refine the arguments.

Sequencing a genome is like conducting a massive experiment in gene transfer. The industrial-scale "shotgun" DNA sequencing strategy typically involves shearing the organism's DNA into manageable fragments, and then inserting these fragments into a strain of *E. coli,* which is used as an enrichment culture—to grow vast amounts of the target DNA. This sequencing process mimics the transmission of DNA from one organism to another, the mechanism referred to as horizontal gene transfer. This phenomenon occurs in nature, allowing one organism to acquire and use genes from other organisms. While this is an extremely rare event in animals, it does occur frequently in microorganisms and is one of the main sources for the rapid spread of antibiotic resistance among bacteria.

The research, entailing a systematic analysis of the massive backlog of microbial genome sequences from the public databases, revealed specific genes that kill *E. coli,* the bacteria employed in the sequencing process.

The genes categorized, while providing a lesson in the evolutionary history of the organism, now suggest a new strategy for screening molecules that may represent the next generation of broad-spectrum antibiotics.

Anemone Genome Gives Glimpse of Multicelled Ancestors

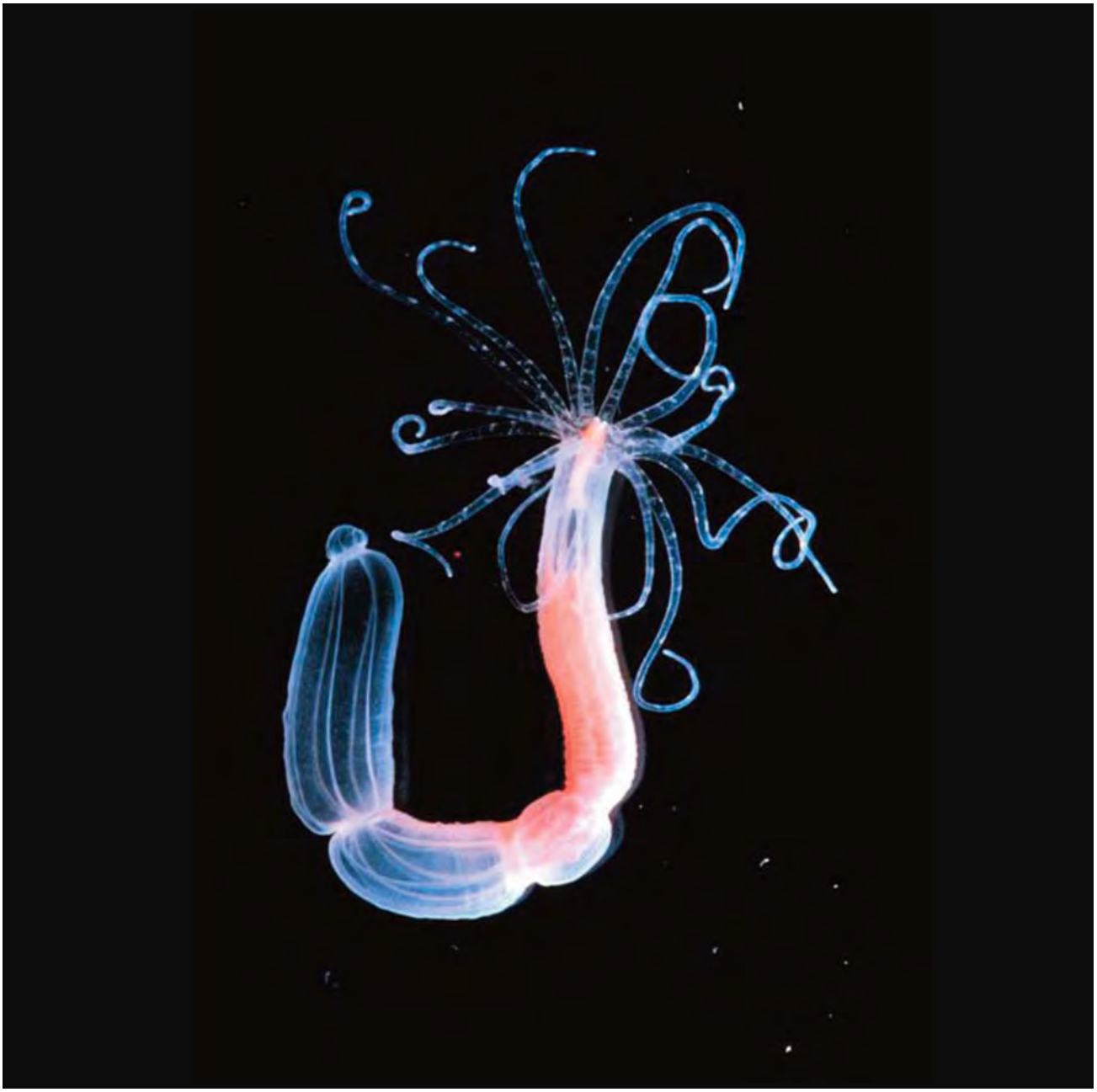
The genome of the starlet sea anemone—*Nematostella vectensis*, a delicate, few-inch-long animal in the form of a transparent, multitentacled tube—was sequenced at the JGI as part of the Community Sequencing Program. The analysis, led by JGI Computational Genomics Program head Daniel Rokhsar, was reported in the July 6, 2007 issue of the journal *Science.*

The anemone burrows in the mud in brackish water along the east and west coasts of the United States and in the British Isles, and is becoming an increasingly important model system for the study of development, evolution, genomics, reproductive biology and ecology.

The JGI-led genome analysis, the first of a sea anemone, has revealed it to be nearly as complex as the human genome, leading researchers farther down the trail of a common ancestor of not only humans and sea anemones, but of nearly all multicelled animals.

Insights from the analysis shed light on how genes and genomes evolved throughout the history of animals, affording scientists a better understanding not only of animal origins, but also how biodiversity is shaped by genomic change.

Nematostella vectensis



DNA SEQUENCE ANALYSIS TOOLS

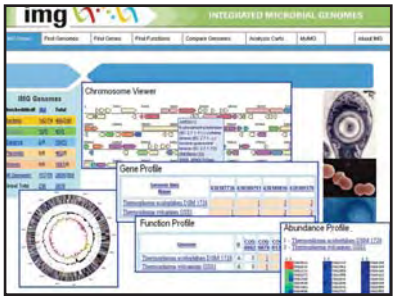
To fully realize the value of DNA sequence information for use by the greater scientific community, the JGI develops resources and tools to enable rapid access and analyses of these data. IMG is one prominent contribution to this effort.

Integrated Microbial Genomes (IMG) Data Management System

The Integrated Microbial Genomes (IMG) data management system is a collaboration between the JGI and the Lawrence Berkeley National Laboratory Biological Data Management and Technology Center (BDMTC). The IMG system is designed as an easy-to-use tool for investigators wanting to extract information from microbial sequence information. IMG responds to the urgent and increasing need for a means to handle the vast and growing spectrum of datasets emerging from genome projects taken on by the JGI and other public DNA sequencing centers. This important computational tool enables scientists to tap the rich diversity of microbial environments and harness the possibilities that they hold for addressing challenges in environmental cleanup, agriculture, industrial processes, and alternative energy production.

In 2007, IMG was upgraded to include fungi, protists (eukaryotic unicellular organisms), and plant genomes to enhance the breadth of comparative analysis. A new addition of particular interest to DOE is *Pichia stipitis* CBS 6054, a fungus with the exceptional capability to ferment xylose, a five-carbon wood sugar, and yield high levels of ethanol.

Version 2.4 included new microbial genomes from the Version 25 release of the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq). IMG 2.4 also contains a total of 3,637 genomes consisting of 818 bacterial, 50 archaeal, and 40 eukaryotic genomes; 2,042 viruses and bacterial phages; and 687 plasmids. Among these genomes, there are 3,334 finished and 303 draft genomes, of which 256 (185 finished and 71 draft) are genomes sequenced by the DOE JGI.



The IMG native controlled vocabulary of functional terms and pathways has been extended to 4,148 terms and 524 pathways, with 546,169 genes characterized using IMG terms. Functional annotation of genomes in IMG has been further enhanced through the computation of fused genes and by filling in various RNA genes missing from the original genome data sets.

IMG's user interface has been reorganized and enhanced to improve ease of use. IMG's user manual, *Using IMG,* has been revised and extended. For more details, see http://img.jgi.doe.gov/.

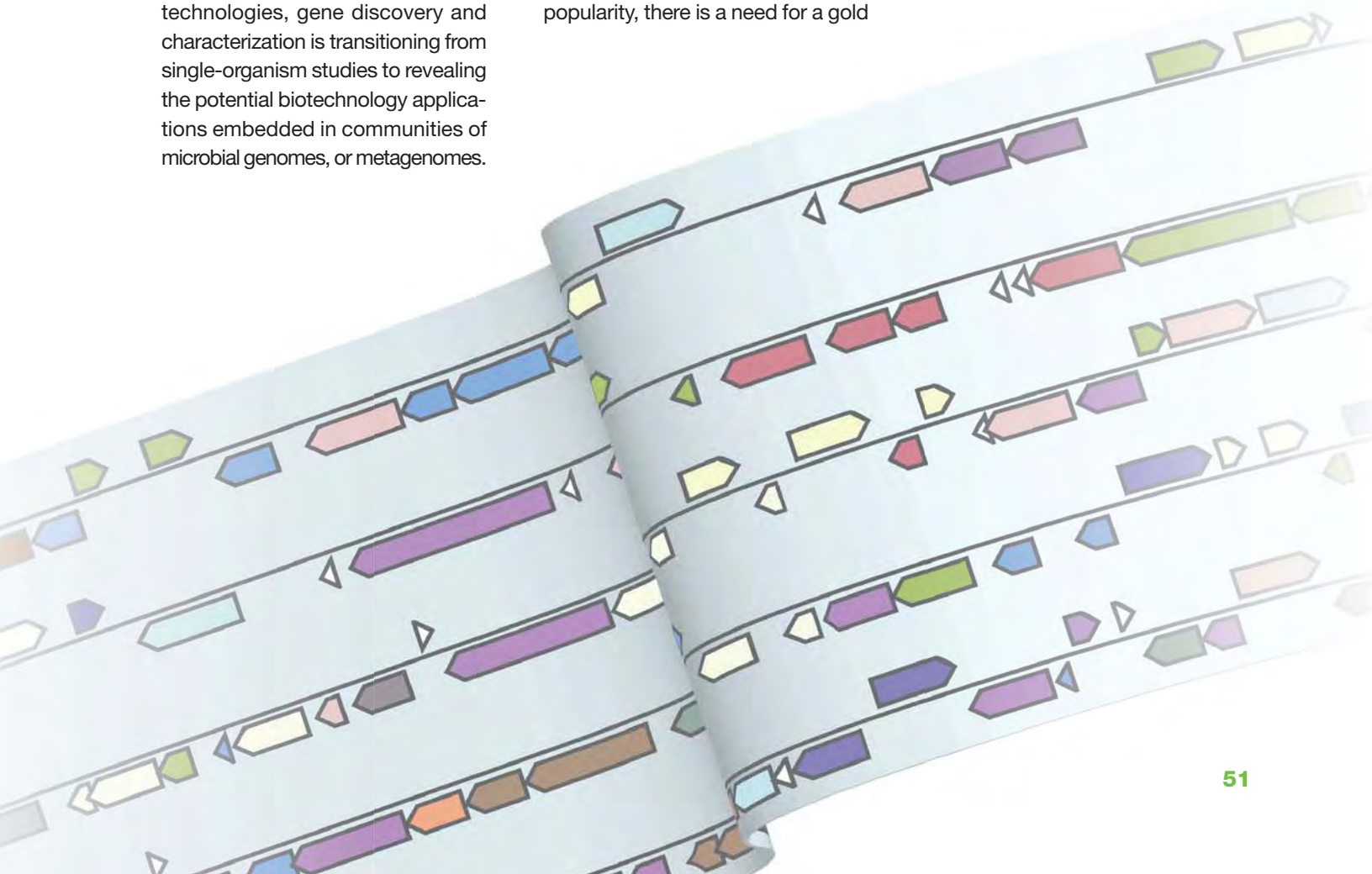
IMG/M "Gold Standard" for Metagenome Analysis

Integrated Microbial Genomes for Microbiome Samples (IMG/M) integrates data from diverse environmental microbial communities with isolated microbial genome data from the JGI's IMG system. This allows the application of IMG's comparative analysis tools to metagenome data for examining the functional capability of microbial communities and isolated organisms of interest, and the analysis of strain-level heterogeneity within a species population in metagenome data.

With the advent of more powerful and economical DNA sequencing technologies, gene discovery and characterization is transitioning from single-organism studies to revealing the potential biotechnology applications embedded in communities of microbial genomes, or metagenomes.

IMG/M contains metagenome data generated from microbial community samples that have been the subject of recently published studies, including two biological phosphorus removing sludge samples, two human distal gut samples, a gutless marine worm sample, and obese and lean mouse gut samples. In addition, IMG/M includes three of the simulated metagenome data sets employed for benchmarking several assembly, gene prediction, and binning methods.

As prospecting for potential utility in microbial communities is gaining popularity, there is a need for a gold standard by which useful data can be evaluated. The JGI has provided such a standard, in work led by Konstantinos Mavrommatis, a postdoctoral fellow in the Genome Biology Program, and published in the May edition of *Nature Methods.* The method described provides researchers the means to compare their metagenome dataset with a simulated metagenome and receive guidance as to which are the optimal tools for analysis—akin to how consumers now rely on certain Web sites to compare new electronic products.



NEW SEQUENCING TECHNOLOGIES

New technologies being explored at the JGI promise an even faster and more efficient future for sequencing. There are two new methods. One method uses the Roche 454 machine and involves a process called emulsion PCR along with pyrosequencing. The other method uses the Illumina machine and involves bridge PCR along with reversible-dye terminators. Both technologies eliminate the time-consuming steps of in vivo cloning, colony picking, and capillary electrophoresis. Both machines are able to produce approximately 100 times more bases per week than the previous generation of sequencers. Both technologies have the added benefit of eliminating bias against in vivo genomic regions. The new sequencing-by-synthesis (SbS) approach builds a picture of a newly synthesized DNA fragment one base at a time. The addition of each base is detected in real time eliminating the need for separating molecules according to size using capillary electrophoresis.

ROCHE 454 SEQUENCING TECHNOLOGY

Roche's 454 sequencing technology uses emulsion PCR DNA amplification combined with pyrosequencing, which enables one person to prepare and sequence an entire genome. The 454 instrument uses a parallel processing approach to produce over 100 megabases (100 million bases) of DNA sequence per 4.5-hour sequencing run (7 hours including preparation steps).

ILLUMINA SEQUENCING TECHNOLOGY

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

The approach begins by attaching randomly fragmented genomic DNA to a glass slide, the flow cell, at a density of about 1 molecule per 5 μm^2. Each molecule is amplified by bridge PCR to create about 1,000 copies of template per cluster.

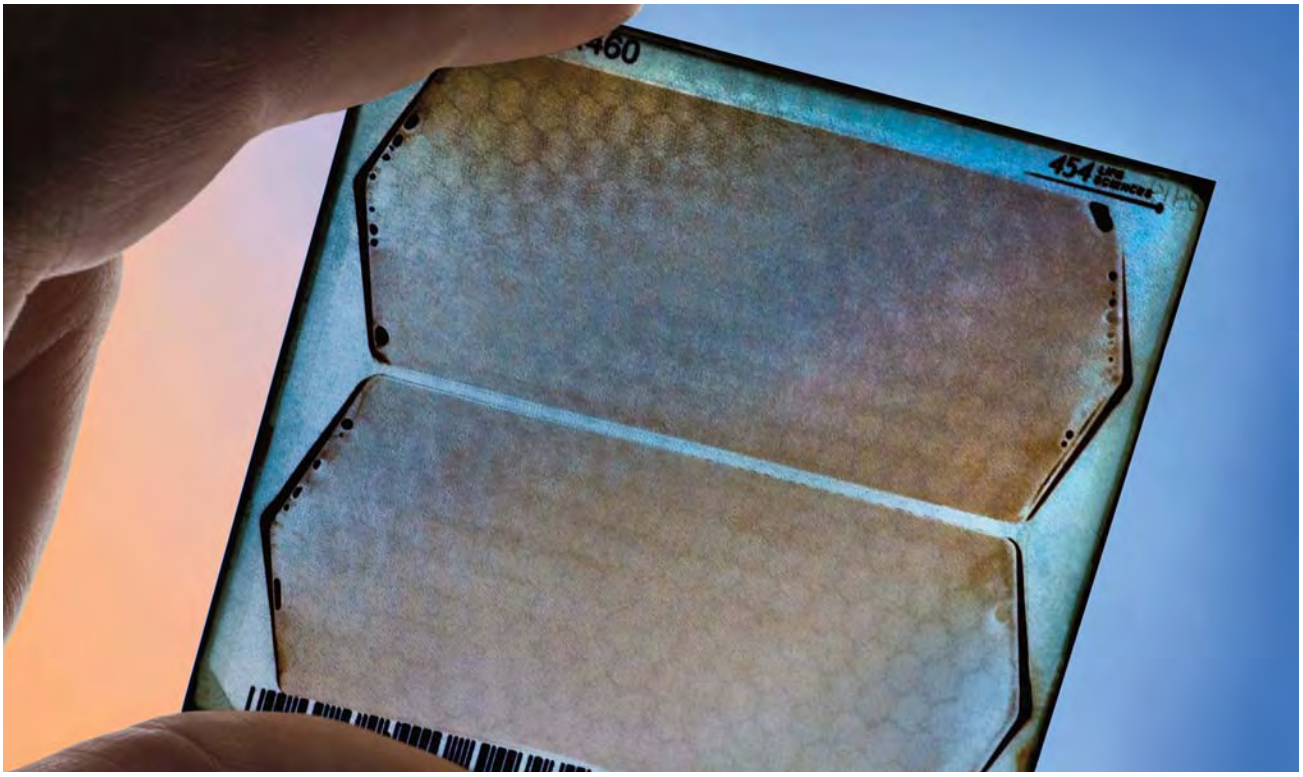
The flow cell then contains up to 40 million photographable clusters, each giving 40 bases per 40 SBS cycles. These templates are sequenced using a four-color DNA sequencing-by-synthesis technology. During each sequencing cycle, reagents are washed over the slide and only one base of the four colored bases is incorporated.

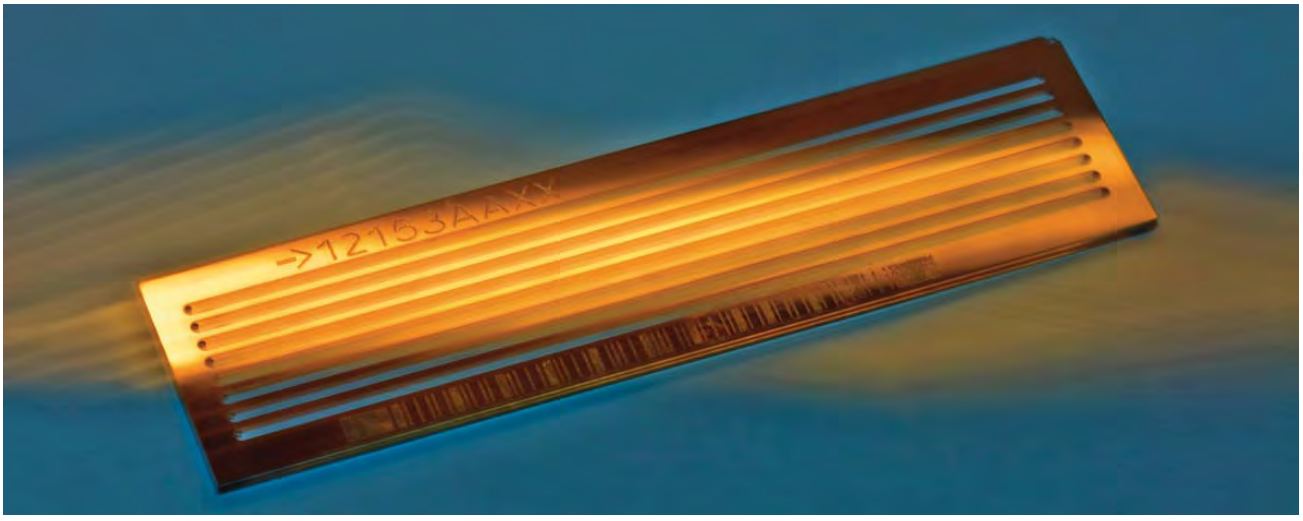
The bases used in this process have the ability to have their terminator and fluorescent capabilities removed at the end of each cycle. Unincorporated bases and the colored tag are then removed. This SBS process differs from the 454 process in that all four labeled nucleotides are used in one reaction. Because only one base is added at a time, the Illumina has no problems with homopolymer repeats and is highly accurate.

The identity of each base of a cluster is read off from sequential images. Short sequence reads (25–40 bases) are aligned against a reference genome and genetic differences are called using a specially developed data pipeline.

APPLICATIONS AND COMPARISONS OF TECHNOLOGIES

As standalone processes, both the 454 and the Illumina technologies have some shortcomings. Many of the problems have to do with the short read lengths of only 35 to 200 bases. This makes it difficult to determine the relative position of each fragment and deal with long repeats during the assembly process. The pyrosequencing process has the added problem of not being able to accurately determine the number of nucleotides from homopolymer repeats (e.g., AAAA, TTTT, etc.). Advantages regarding clarity may be gained by assembling the small reads with longer


454 LIFE SCIENCES


->12153AAXX

Sanger reads. This produces a more comprehensive draft sequence. However, assembly and annotation require intensive informatics work because the base quality scores need to be made comparable to conventional sequence scoring methods. Though the per-base cost of new sequencing technology is much cheaper than the traditional Sanger sequencing, more depth of coverage is required to achieve high-quality assemblies. Most of the cost savings has come from the capture of genomic regions that were unable to be processed by the in vivo methods used in the Sanger method.

An additional "next generation" technology, the Applied Biosystems SOLiD™ System, was brought into the JGI for evaluation in late 2007. The system is a genetic analysis platform that enables massively parallel sequencing of clonally amplified DNA fragments linked to beads. The sequencing methodology is based on sequential ligation with dye-labeled oligonucleotides. The SOLiD System promises more than three billion bases of data per run.

In spring 2008, the JGI will consider the resource allocation for acquiring additional new sequencers.

2006-2007 Jamboree Roundup

Jamborees are working meetings at which members of a scientific community gather to discuss and annotate (assign function to) the genome of an organism (or family of organisms) of common interest. The goal is to identify genes and generate high-quality annotations, ultimately leading to publications in the scientific literature.

ANNOTATED EUKARYOTES FY2007

	ANNOTATED	PUBLIC	JAMBOREE	PUBLISHED
Chlamydomonas reinhardtii	●	●	●	•
Xenopus tropicalis	●	●	•	○
Laccaria bicolor	●	•	•	•
Pichia stipitis	●	•		•
Ostreococcus lucimarinus	●	•	●	•
Nematostella vectensis	•	•		•
Physcomitrella patens	•	•	•	•
Branchiostoma floridae	•	•		○
Monosiga brevicollis	•	•		○
Trichoderma reesei, finished	•	•		○
Naegleria gruberi	•	•		○
Aspergillus niger	•	•	•	○
Thalassiosira pseudonana, finished	•	•	•	○
Phaeodactylum tricornutum, finished	•	•	•	○
Mycosphaerella graminicola	•	•	•	○
Sporobolomyces roseus	•	•	○	
Nectria haematococca MPVI, finished	•	•	•	
Phycomyces blakesleeanus	•	•	•	○
Daphnia pulex	•	•	•	○
Postia placenta	•	•	•	○
Volvox carteri	•	•	•	○
Aureococcus anophagefferens	•	•	•	
Lottia gigantea	•	•		
Mycosphaerella fijiensis	•	•		
Capitella sp. I	•	•		
Helobdella robusta	•	•		
Trichoplax adhaerens	•	•		○
Phytophthora capsici	•		•	
Selaginella moellendorfii	•	•		
Trichoderma virens	•	•		
Micromonas pusilla NOUM17, finished	•	•	•	○
Micromonas pusilla CCMP1545, improved	•	•	•	○
Emiliania huxleyi	•	○		
Batrachochytrium dendrobatidis JAM81	•	○		
Chlorella sp. NC64A	•	○		
Ostreococcus RCC809	•			
Dictyostelium purpureum	•			
Sorghum bicolor	○			○
Trichoderma atroviride	○			

● prior to 2006; • done 2006; • done 2007; ○ in progress

Micromonas *jamboree, April 2007. Front row: Hervé Moreau, Evelyn Derelle, Andy Allen, Sarah McDonald; Middle row: Uwe John, Marie Cuvelier, Micaela Parker, Meredith Everett, Alex Worden; Back row: Thomas Mock, Rory Welsh, Kemin Zhou, Igor Grigoriev, Pierre Rouze*

JGI Sequencing Department Head Susan Lucas instructing the next generation of scientists at the Expanding Your Horizons event, March 2007.

In October 2007, JGI's Undergraduate Research Program convened the first Workshop in Microbial Genome Annotation. Nineteen educators from 14 institutions participated, representing a diverse cross-section of research universities, state and liberal arts colleges. The goal of the workshop was to provide the tools and ideas for advancing genomics and bioinformatics across undergraduate curricula.

Front to back, left to right; Front row: Cheryl Bailey, Kelynne Reed, Sharyn Freyermuth, Ferda Soyer, Zhaohui Xu. Second row: Kathleen Scott, Sabine Heinhorst, Cheryl Kerfeld, Tuajuanda Jordan, Jay Lennon. Back row: Daniela Bartels, Brad Goodner, Erin Sanders-Lorenz, Christopher Kvaal, Mitrick Johns, Folker Meyer, Jayna Ditty, Tobias Paczian. Stuart Gordon not pictured.

EDUCATION AND OUTREACH

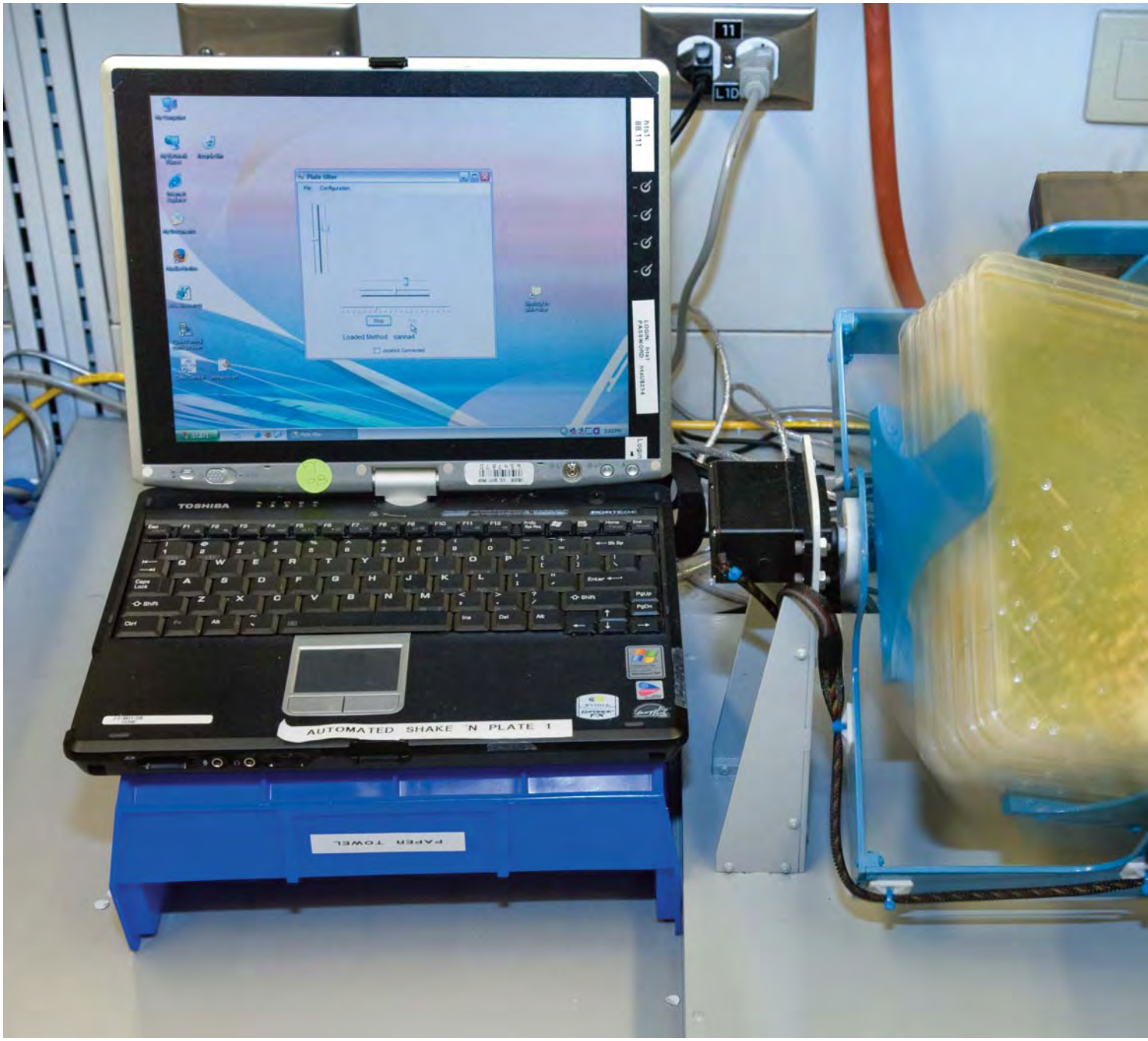
In 2007, the JGI launched a formal Education Program, focusing on undergraduate and graduate education, led by Cheryl A. Kerfeld, Ph.D., who joined the JGI from the University of California, Los Angeles, where she was director of the UCLA Undergraduate Genomics Research Initiative. While at UCLA, she exploited the user facility resources of the JGI to provide her students with an interdepartmental multicourse collaboration with the central theme of sequencing and analyzing the genome of a bacterium.

TEACHING THE TEACHERS

In October 2007, the JGI's Education Program convened the first Workshop in Microbial Genome Annotation for undergraduate educators. Nineteen educators from 14 institutions participated, representing a diverse cross-section of research universities, state, and liberal arts colleges. The goal of the workshop was to provide the tools and ideas for advancing genomics and bioinformatics across undergraduate curricula. Cheryl Kerfeld, JGI's Education Program head, presented a survey of current bioinformatics tools and strategies for integrating annotation across the life sciences curriculum, including algorithms for gene calling, pathway annotation, and characterization of hypothetical protein function.

TEACHING THE NEXT-GENERATION RESEARCHERS

The University of California, Merced, a UC campus with a significant student enrollment of minorities underrepresented in science, which opened in 2005, is the 10th UC campus and located about two hours' drive from the JGI Production Genomics Facility. Former JGI researcher Mónica Medina is now on the faculty at UC Merced in the Natural Sciences Department and has been working with her colleagues at the JGI to develop an innovative theoretical and experimental course in genomics. Starting in January 2007 and running through the end of April, the course covered every aspect of the JGI sequencing and bioinformatics process.


11
L1D
TOSHIBA
AUTOMATED SHAKE 'N PLATE 1
PAPER TOWEL

SAFETY AND ERGONOMICS

Safety is of paramount importance throughout every activity of the JGI Production Genomics Facility. Due to the potential for repetitive strain injuries resulting from some of the tasks associated with the sequencing line, the JGI has been vigilant about conducting proactive safety and ergonomic assessments in all workstations.

In December 2007, the JGI production team participated in a "stand-down" from their regular daily activities in order to focus on resolving ergonomic issues. Eighty people participated in team-building exercises geared to engender improved communication, respect, trust, ergonomics, and safety. As a result, this team has committed to strive for the goal of zero injuries in 2008.

JGI TAKES THE PRIZE

A team of scientists and engineers from the JGI and LBNL won the prestigious 2007 Ergo Cup, competing against the likes of 28 international finalists with their innovative "Shake 'N Plate" instrument. The 10th Annual Applied Ergonomics Conference, held March 12–15 in Dallas, Texas, hosted the awards. The Ergo Cup, presented by the Institute of Industrial Engineers, provides opportunities for institutions to highlight successful ergonomic innovations. "Shake 'N Plate," which won in the "team-driven workplace solutions" category, is a simple device designed to alleviate upper body fatigue associated with bacterial culture plating. Operators were manually processing stacks of 22 cm x 22 cm gel-filled plates weighing up to 7 pounds. The hand-grasp forces and total weight during long processing times made this an unpleasant and fatiguing task. The "Shake 'N Plate is a lightweight sheet metal platform mounted on a ball joint similar to a camera tripod. In fact, one of the operators used the camera tripod idea as her original design suggestion to the JGI Instrumentation Group. The design was quickly fabricated in the LBNL shops and put into production use. The device removes almost all of the weight from the operator's arms and actually improves the process throughput. "Shake 'N Plate" won against such larger competitors as Boeing, GE, Harley Davidson, and Toyota. The winning team of JGI and LBNL Engineering Division staff was comprised of Christine Naca, Martin Pollard, Diane Bauer, Catherine Adam, Simon Roberts, Karl Petermann, Charlie Reiter, Ira Janowitz, Karli Ikeda, Miranda Harmon-Smith, Sanna Anwar, and Damon Tighe.


JGI
DOE JOINT GENOME INSTITUTE
THE ERGO CUP
HONORS
THE MOST OUTSTANDING
TEAM-DRIVEN WORKPLACE SOLUTION
PRESENTED AT THE
ANNUAL
APPLIED ERGONOMICS CONFERENCE

AWARDS AND HONORS

The JGI and its partners are comprised of a legion of talented scientists, engineers, and support staff united in their vision for advancing the frontiers of science with DNA sequence information. In 2007, two of the JGI's own received high honors.

Len Pennacchio, the JGI's Genetic Analysis Program Head, was among the year's recipients of The Presidential Early Career Award for Scientists and Engineers (PECASE). The PECASE award was created to honor and support the extraordinary achievements of young professionals at the outset of their independent research careers in the fields of science and technology. The Presidential Award embodies the high priority placed by the government on maintaining the leadership position of the United States in science by producing outstanding scientists and engineers who will broadly advance science and the missions important to the participating federal agencies.

Pennacchio was recognized for his significant contributions to the generation and interpretation of the human genome sequence. Specifically, the committee cited his systematic assignment of gene regulatory function to the human genome through the coupling of vertebrate comparative genomics and large-scale studies in mice, using a world-class mouse resource that he established.

Gerald Tuskan, the JGI's Laboratory Science Program lead and senior scientist in the Environmental Sciences Division at Oak Ridge National Laboratory, was recognized with the ORNL Director's Award for Outstanding Team Accomplishment. Tuskan led the ORNL team that participated in the international effort to sequence the poplar tree genome, which is regarded as a key step toward advancing biomass as an alternative energy source. Tuskan is credited by his colleagues with combining enthusiasm, organizational skills, and scientific prowess to achieve a complex scientific goal.

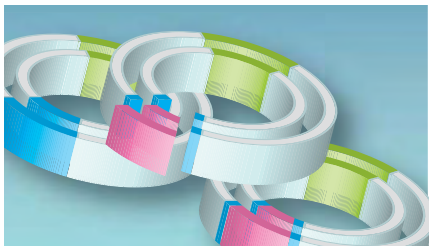
APPENDICES

Appendix A: JGI Sequencing Process

Receive DNA from collaborators 1

Shear DNA 2

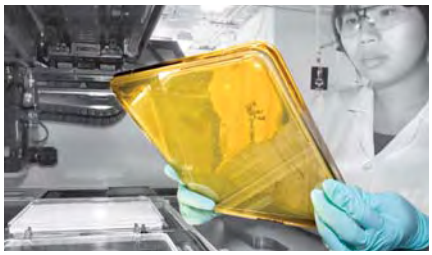
Insert DNA fragments into vectors 3



Introduce vectors in bacteria 4

Plate bacteria 5

Pick bacteria that contain vectors with inserts 6



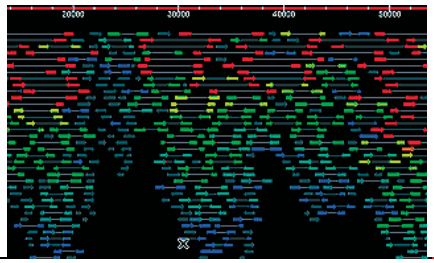
Amplify vectors with inserts 7

Produce dye-labeled DNA fragments 8

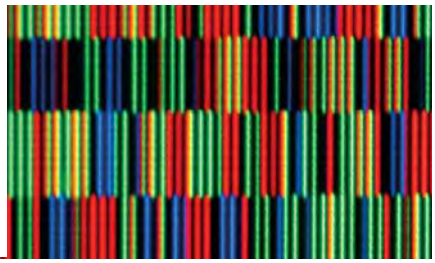
Clean up dye-labeled DNA fragments 9

Sequence by capillary electrophoresis 10

Assemble genome from sequence 11



Release genomic data to the world 12



DNA: LIFE'S CODE

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

WHAT IS DNA SEQUENCING?

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

JGI DNA SEQUENCING PROCESS

Whole-genome shotgun sequencing is a technique for determining the precise order of the letters of DNA code of a genome. First, DNA received from JGI collaborators **(1)**, or users, is sheared into small fragments that are easier for sequencing machines to handle **(2)**. These fragments are biochemically inserted into a plasmid vector **(3)**—a loop of nonessential bacterial DNA—and mixed into a solution of *E. coli* bacteria. An electric shock allows the plasmids to enter the bacterial cells **(4)**. The bacterial cells are moved to an agar plate **(5)** and incubated with a nutrient and antibiotic to suppress the growth of unwanted cells. Over a night of incubation, colonies form, and each contains about a million bacteria. The clones in the colonies that contain the inserted DNA fragments required for sequencing are distinguished by color, picked **(6)**, and incubated with nutrients again to make many more copies, which can then be sequenced. These DNA of interest are then duplicated, or amplified **(7)**, through a process initiated with another enzyme called polymerase and an abundant supply of the chemical building-blocks of DNA, or nucleotides, from which the polymerase can assemble new copies. After many heating and cooling cycles, the ends of the DNA fragments are labeled with fluorescent markers **(8)**. The DNA fragments are cleaned—isolated from the bacteria—using magnetic beads in an ethanol solution **(9)**. In a sequencing machine, an electrical charge is applied to the samples, pulling the DNA fragments through an assembly of fine glass tubes filled with a gel-like matrix, smaller fragments traveling faster than the larger, toward a laser detector system, which excites the fluorescent tags on each fragment by length and counts them to determine the sequence of As, Ts, Cs, and Gs **(10)**.

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

Appendix B: Genomics Glossary

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

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

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

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

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

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

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

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

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

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

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

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

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

Mapping: Charting the location of genes on chromosomes.

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix C: CSP Projects

COMMUNITY SEQUENCING PROGRAM SEQUENCING PLANS FOR 2008

ORGANISM	COLLABORATOR	INSTITUTION
LARGE EUKARYOTES		
Eucalyptus tree	Myburg	Univ. of Pretoria
Foxtail millet *(Setaria italica)*	Bennetzen	Univ. of Georgia
Porphyra purpurea (a marine red alga)	Brawley	Univ. of Maine
SMALL EUKARYOTES		
Agaricus bisporus (a leaf-litter degrading homobasidiomycete)	Challen	Univ. of Warwick
Heterodera glycines (soybean cyst nematode)	Lambert	Univ. Illinois Urbana-Champaign
Marchantia polymorpha	Bowman	Monash Univ./UC Davis
Paxillus involutus (an ectomycorrhizal fungus)	Tunlid	Lund Univ.
Phaeocystis antarctica: A dominant phytoplankter and ice alga in the Southern Ocean	Berg	Stanford Univ.
Phaeocystis globosa	Allen	The Institiute for Genomic Research
ESTs for pines and other conifers	Dean	Univ. of Georgia
Tetrahymena thermophila strain SB210	Collins	Univ. of California Berkeley
METAGENOMES		
Type I *Accumulibacter*	McMahon	Univ. of Wisconsin-Madison
Anammox bacteria *(Scalindua marina, Brocadia fulgida,* and *Anammoxglobus propionicus)*	Jetten	Radboud Univ.
A biogas-producing microbial community	Wu	Univ. of California Davis
Extreme microbial habitats across the Yellowstone geothermal ecosystem	Inskeep	Montana State Univ.
ISOLATES		
Allochromatium vinosum DSM 180(T)	Dahl	Univ. of Bonn
UNCULTIVATED METHANE-OXIDIZING ARCHAEON		
ANME-1		Hallam
Budding and non-budding stalked bacteria from aquatic environments *(Asticcacaulis biprosthecum, Asticcacaulis excentricus, Brevundimonas subvibrioides, Ancalomicrobium adetum, Hyphomicrobium denitrificans,* and *Rhodomicrobium vannielii)*	Brun	Indiana Univ.
Diaphorobacter sp. strain TPSY, *Ferrutens nitratireducens* strain 2002, and *Azospira suillum* strain PS	Coates	Univ. of California, Berkeley

COMMUNITY SEQUENCING PROGRAM SEQUENCING PLANS FOR 2008

ORGANISM	COLLABORATOR	INSTITUTION
Frankia strains (EuI1c, BCU110501, R43, BMG5.12, and AmMr)	Tisa	Univ. of New Hampshire
Haloalkaliphilic sulfate-, thiosulfate- and sulfur-reducing bacteria *Desulfonatronovirga dismutans* ASO3-1, *Desulfovibrio alkaliphilus* AHT2, and *Dethiobacter alkaliphilus* AHT1	Muyzer	Delft University of Technology
Halothiobacillus neapolitanus and *Thiomonas intermedia*	Heinhorst	Univ. of Southern Mississippi
Thermophilic or hyperthermophilic methanoarchaea within the *Methanococcales* (*Methanothermococcus okinawensis* IH1, *Methanotorris igneus* Kol 5, *Methanotorris formicicus* Mc-S-70, *Methanocaldococcus fervens* AG86, *Methanocaldococcus infernus* ME, *Ethanocaldococcus vulcanius* M7, *Methanocaldococcus* strain FS406-22)	Whitman	Univ. of Georgia
TYPE I AND TYPE II METHANOTROPHIC BACTERIA		
Methylomicrobium album BG8 and *ethylosinus trichosporium* OB3b)	Stein	Univ. of California Riverside
Two *Micromonosporas* (*aurantiaca* and L5)	Hirsch	Univ. of California Los Angeles
Natrialba magadii ATCC 43099	Maupin-Furlow	Univ. of Florida
Pseudonocardia dioxanivorans CB1190	Mahendra	Univ. of California Berkeley
Selenospirillum indicus	Bini	Rutgers Univ.
Starkeya novella	Kappler	Univ. of Queensland
Thermovibrio ammonificans DSM 15698	Vetriani	Rutgers Univ.
Variovorax paradoxus strains (S110 and EPS)	Han	Rensselaer Polytechnic Inst.
Zymomonas mobilis strains: subsp. *mobilis* ATCC 10988; subsp. *mobilis* ATCC 29191; subsp. *mobilis* ZM4 (ATCC 31821); subsp. *pomaceae* ATCC 29192; subsp. *mobilis recifensis,* industrial strain	Pappas	Univ. of Athens

Appendix C: CSP Projects (continued)

COMMUNITY SEQUENCING PROGRAM. SEQUENCING PLANS FOR 2007

ORGANISM	COLLABORATOR	INSTITUTION
LARGE EUKARYOTES		
Aquilegia formosa	Hodges	UC Santa Barbara
Brachypodium distachyon (Poaceae)	Vogel	USDA-ARS
Gossypium (cotton)	Paterson	Univ. of Georgia
Manihot esculenta (cassava)	Fauquet	Danforth Plant Science Ctr.
SMALL EUKARYOTES		
Cryphonectria parasitica (chestnut blight fungus)	Nuss	Univ. of Maryland Biotech. Inst.
Reef-building corals and dinoflagellate symbionts (ESTs from *Acropora palmata, Montastraea faveolata,* and *Symbiodinium* clade A and clade B)	Medina	Univ. of California, Merced
Fragilariopsis cylindrus (a diatom)	Mock	Univ. of Washington
Guillardia theta and *Bigelowiella natans*	Archibald	Dalhousie Univ.
Heterobasidion annosum	Stenlid	Swedish Univ. of Agri. Sciences
Three species of *Neurospora* (*N. discreta, N. tetrasperma* FGSC2508, *N. tetrasperma* FGSC2509)	Taylor	Univ. of California, Berkeley
Peronosporomycete mtDNAs (26)	Hudspeth	Northern Illinois Univ.
Pleurotus ostreatus (oyster mushroom)	Pisabarro	Public Univ. of Navarre
Riftia pachyptila (deep-sea tubeworm)	Girguis	Harvard Univ.
Switchgrass	Tobias	USDA-ARS
Tetranychus urticae (two-spotted spider mite)	Grbic	Univ. of Western Ontario
Thellungiella halophila	Schumaker	Univ. of Arizona
Mating loci from *Volvox carteri* and *Chlamydomonas reinhardtii*	Umen	Salk Inst.
BACTERIA AND ARCHAEA		
Candidatus Amoebophilus asiaticus and *Encarsia* symbiont *Cand. Cardinium hertigii*	Horn	Univ. of Vienna
Actinobacteria (*Arthrobacter chlorophenolicus* and *Micrococcus luteus*	Jansson	Swedish Univ. of Agri. Sciences
Beggiatoa alba	Mueller	Morgan State Univ.
Burkholderia	Tiedje	Michigan State Univ.

COMMUNITY SEQUENCING PROGRAM. SEQUENCING PLANS FOR 2007

ORGANISM	COLLABORATOR	INSTITUTION
Anaerobic benzene-degrading methanogenic consortium *(Chloroflexi, Desulfobacterium* sp., *Desulfosporosinus* sp., *Methanomicrobiales*-like sp., *Methanosaeta* sp., *Methanosarcinales*-like spp)	Edwards	Univ. of Toronto
Crenothrix polyspora enrichment	Wagner	Univ. of Vienna
Cyanothece strains	Pakrasi	Washington Univ.
Dechlorinating community (KB-1) *(Dehalococcoides, Geobacter, Methanosarcina, Spirochete, Sporomusa)*	Edwards	Univ. of Toronto
Lithifying mat communities of marine stromatolites (*Desulfovibrio* sp. H0407_12.1Lac, *Schizothrix gebeleni* sp. A and sp. B, *Solentia* sp., Sulfate-reducing Bacterium sp. B, and sulfur-oxidizing bacteria)	Decho	Univ. of South Carolina
Candidatus endomicrobium trichonymphae elusimicrobium minutum Pei191	Brune	Max Planck Institute for Terrestrial Microbiology
Symbiont from the basal clade of the *Frankiaceae*	Benson	Univ. of Connecticut
Six freshwater iron-oxidizing bacteria *(Gallionella ferruginea* - 2 samples, *Leptothrix cholodnii, Rhodobacter* sp. str. SW2, *Rhodopseudomonas palustris,* and *Sideroxydans lithotrophicus)*	Emerson	Amer. Type Culture Collection
Microbiome resident in the foregut of the tammar wallaby *(Macropus eugenii)*	McSweeney	CSIRO
Methanomicrococcus blatticola	Hackstein	Radboud University Nijmegen
Methylocella silvestris BL2, *Methylocapsa acidiphila* B2, and *Beijerinckia indica* subsp.	Indica	Dunfield
Pedomicrobium manganicum	Mackenzie	Univ. of Texas, Houston
Microbial community (plasmid mobilome) in wastewater treatment plant	van der Meer	Univ. of Lausanne
Rhizobium leguminosarum bv *trifolii* (strains WSM1325 and WSM2304)	Reeve	Murdoch Univ.
Near-shore anoxic basin: Saanich Inlet	Hallam	Univ. of British Columbia
Thauera sp. MZ1T	Sayler	Univ. of Tennessee
Thermolithobacter	Wiegel	Univ. of Georgia
Haloalkaliphilic sulfur-oxidizing bacteria (*Thioalkalivibrio* sp. HL-EbGR7, and T. sp. K90 mix)	Muyzer	Delft Univ. of Technology

Appendix D: 2007-2008 LSP Projects

TITLE	PRINCIPAL INVESTIGATOR	LAB
Genome sequencing of *Burkholderia cepacia* Bu72	van der Lelie	BNL
Genome sequencing of *Sulfobacillus acidophilus*	Garcia	LLNL
Gene silencing using microRNAs to modulate biomass growth and cell wall synthesis in bioenergy crop	Hong-Geller	LANL
Comparative metagenomics of grass-feeding termite hindgut communities	Hugenholtz	LBNL
Metagenomics-enabled analysis from the sediment of an anoxic meromictic lagoon in Greece.	Kyrpides	LBNL
Genome-wide sequencing and analysis of nuclear matrix attachment regions (MARs)	Doggett	LANL
454-Based Eco-transcriptomics to identify key microbial consortia essential for *Dehalococcoides*-mediated bioremediation of chlorinated pollutants	Brodie	LBNL
In-depth microbial diversity and population dynamics during a biomass degrading composting process	D'haeseleer	LLNL
Characterization of the glycoside hydrolase content of the fiber-associated microbiome from the bovine rumen: Studying the co-evolution of microorganism and enzyme content.	White	ANL
Archaeal virus community genomics of Yellowstone's high-temperature environments	Roberto	INL
Genome sequencing of *Leonotis nepetifolia*	Shanklin	BNL
Genome sequencing of *Phycomyces blakesleeanus*	Baker	PNNL
Poplar in response to dehydration	Yang	ORNL
Hypersaline microbial mat	Raymond	LLNL
Medaka fish	Glenn	SRNL
454 sequencing of viruses	Han	LANL
Genome sequencing of *Peromyscus*	Glenn	SRNL

Appendix E: Review Committees & Board Members

JOINT GENOME INSTITUTE POLICY BOARD

The JGI Policy Board serves two primary functions:

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

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

JGI POLICY BOARD

Gerry Rubin, (Chair)
Howard Hughes Medical Institute

Richard Gibbs
Baylor College of Medicine

Stephen Quake
Stanford University

David Galas
Institute for Systems Biology, Seattle, WA
Battelle Memorial Institute, Columbus, OH

Edward DeLong
Massachusetts Institute of Technology

Penny Chisholm
Massachusetts Institute of Technology

Melvin Simon
California Institute of Technology

Chris Somerville
Stanford University

James Tiedje
Michigan State University

Susan Wessler
University of Georgia

Appendix E: Review Committees & Board Members (continued)

JGI SCIENTIFIC ADVISORY COMMITTEE

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

2007 JGI SCIENTIFIC ADVISORY COMMITTEE

Mark Adams
Case Western Reserve University

Ginger Armbrust
University of Washington

Bruce Birren
Broad Institute

Edward DeLong
Massachusetts Institute of Technology

Joseph Ecker
Salk Institute for Biological Studies

Richard Harland
University of California, Berkeley

Jim Krupnick
Lawrence Berkeley National Laboratory

Marco Marra
Michael Smith Genome Sciences Centre

Eric Mathur
Synthetic Genomics

Doug Ray
Pacific Northwest National Laboratory

George Weinstock
Baylor College of Medicine

EUKARYOTIC PROPOSAL REVIEWERS

Nina Agabian
University of California, San Francisco

Jody Banks
Purdue University

Zac Cande
University of California, Berkeley

Jeff Dean
The University of Georgia

Nigel Dunn-Coleman
Genencor International

Daniel Ebbole
Texas A&M University

Mike Freeling
University of California, Berkeley

Mónica Medina
University of California, Merced

Andrew Paterson
University of Georgia

Arend Sidow
Stanford University

Shauna Somerville
Stanford University

Rob Steele
University of California, Irvine

John Taylor
University of California, Berkeley

Gary Andersen
Lawrence Berkeley National Laboratory

Donald Bryant
Pennsylvania State University

Patrick Chain
Lawrence Livermore National Laboratory

Ludmila Chistoserdova
University of Washington

Emilio Garcia
Lawrence Livermore National Laboratory

Steven Hallam
The University of British Columbia

Frank E. Loeffler
Georgia Institute of Technology

David Mead
Lucigen Corporation

David Mills
University of California, Davis

Frank Robb
Center of Marine Biotechnology

Kathleen Scott
University of South Florida

Craig Stephens
Santa Clara University

Tamás Török
Lawrence Berkeley National Laboratory

Bart Weimer
Utah State University

Daniel (Niels) van der Lelie
Brookhaven National Laboratory

Appendix F: 2007 User Meeting Agenda Speakers List

Nikos Kyrpides
Joint Genome Institute

Igor Grigoriev
Joint Genome Institute

Jim Bristow
Joint Genome Institute

Susan Lucas
Joint Genome Institute

Darren Platt
Joint Genome Institute

David Bruce
Joint Genome Institute

Paul Richardson
Joint Genome Institute

Jay Keasling
Lawrence Berkeley National Laboratory

David Relman
Stanford University

Eddy Rubin
Joint Genome Institute

Jim Frederickson
Pacific Northwest National Laboratory

Caroline Harwood
University of Washington

Jim Tiedje
Michigan State University

Larry Wackett
University of Minnesota

Jonathan Eisen
University of California, Davis

David Stahl
University of Washington

Nancy Moran
University of Arizona

Bryan White
University of Illinois

Craig Stephens
Santa Clara University

Scott Geib
Pennsylvania State University

Kirk Harris
University of Colorado

Scott Baker
Pacific Northwest National Laboratory

Lillian Fritz-Laylin
University of California, Berkeley

Steve Goodwin
Purdue University

John Vogel
U.S. Department of Agriculture, Agricultural Research Service

John Archibald
Dalhousie University

Brian Palenik
University of California, San Diego

Stephen Kingsmore
The National Center for Genome Resources

David Mead
Lucigen

Robert Nutter
Applied Biosystems

Steven Quake
Stanford University

Rotem Sorek
Joint Genome Institute

Michael Thelen
Lawrence Livermore National Laboratory

Appendix G: JGI Publications 2006–2007

Adamska, M, et al.
The evolutionary origin of hedgehog proteins. *Current Biology,* 17 (19): R836-R837, Oct. 9, 2007

Ahituv, N., et al.
Deletion of ultraconserved elements yields viable mice. *PLoS Biology*, 5 (9): 1906-1911, September 2007

Ahituv, N, et al.
Medical sequencing at the extremes of human body mass. *American Journal of Human Genetics*, 80 (4): 779-791 April 2007

Ahituv, N., et al.
A PYY Q62P variant linked to human obesity. *Human Molecular Genetics*, 15 (3): 387-391 Feb. 1, 2006

Allen, E.E., et al.
Genome dynamics in a natural archaeal population. *PNAS*, 104 (6): 1883-1888, Feb. 6, 2007

Alonso, C., et al
High local and global diversity of *Flavobacteria* in marine plankton. *Environmental Microbiology*, 9 (5): 1253-1266, May 2007

Anderson, S.J., et al.
Phenotypic and genetic variation among soybean rust isolates. *Phytopathology*, 97 (7): S4-S4 Suppl. S, July 2007

Auerbach, R.K., et al.
Yersinia pestis evolution on a small time scale: Comparison of whole genome sequences from North America. *PLoS ONE* 2:e770.DOI: 10.1371/journal.pone.oooo770.

Baker, S.E., et al.
Aspergillus genomics and DHN-melanin conidial pigmentation. In Aspergillus *Systematics in the Genomic Era*, eds. J. Varga and R. Samson (in press).

Baker, S.E., et al.
Aspergillus niger genomics: Past, present and into the future. *Medical Mycology*. 44:S17-21.

Baker, S.E., et al.
Genome and proteome analysis of industrial fungi. In *Exploitation of Fungi*. eds. G.D. Robson, P. van West, G.M. Gadd. Cambridge Press, pp 3–9.

Barns, S.M., et al.
Acidobacteria phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within the phylum. *Applied Environmental Microbiology*, 73:3113-6.

Bazykin, G.A., et al.
Extensive parallelism in protein evolution. *Biology Direct,* 2: Art. No. 20, Aug. 16, 2007

Beckstrom-Sternberg, S.M., et al.
Complete genomic characterization of a pathogenic A.II strain of *Francisella tularensis subspecies tularensis*. Submitted to *PLoS*, July 2007.

Bejerano, G., et al.
A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, 441 (7089): 87-90, May 4, 2006

Beller, H.R., et al.
The genome sequence of the obligately chemolithoautotrophic, facultatively anaerobic bacterium *Thiobacillus denitrificans. Journal of Bacteriology*, 188 (4): 1473-1488 February 2006

Bland, C., et al.
CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, 8: Art. No. 209, June 18, 2007

Bleyl, S.B., et al.
Candidate genes for congenital diaphragmatic hernia from animal models: Sequencing of FOG2 and PDGFR alpha reveals rare variants in diaphragmatic hernia patients. *European Journal of Human Genetics*, 15 (9): 950-958, September 2007

Boore, J.L., et al.
The complete sequence of the mitochondrial genome of *Nautilus macromphalus* (*Mollusca*: *Cephalopoda*). *BMC Genomics*, 7: Art. No. 182, July 19, 2006

Boore, J.L.
The use of genome-level characters for phylogenetic reconstruction. *Trends in Ecology and Evolution*, 21 (8): 439-446, August 2006

Brodie, E.L., et al.
Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. *Applied and Environmental Microbiology*, 72 (9): 6288-6298, September 2006

Budowle, B., et al.
Quality sample collection, handling, and preservation for an effective microbial forensics program. *Applied Environmental Microbiology* 72:6431-6438.

Cai, Z.Q., et al.
Complete plastid genome sequences of *Drimys*, *Liriodendron*, and *Piper*: Implications for the phylogenetic relationships of magnoliids. *BMC Evolutionary Biology*, 6: Art. No. 77, Oct. 4, 2006

Cameron, R.A., et al.
Unusual gene order and organization of the sea urchin Hox cluster. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution*, 306B (1): 45-58, Jan. 15, 2006

Carapelli, A., et al.
The mitochondrial genome of the entomophagous endoparasite *Xenos vesparum* (*Insecta*: *Strepsiptera*). *GENE*, 376 (2): 248-259, July 19, 2006

Appendix G: JGI Publications 2006–2007 (continued)

Chain, P.S.G., et al.
Burkholderia xenovorans LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. *PNAS*, 103 (42): 15280-15287, Oct. 17, 2006

Chain, P.S.G., et al.
Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: Evidence of gene reduction in an emerging pathogen. *Journal of Bacteriology*, 188 (12): 4453-4463, June 2006

Challacombe, J.F., et al.
The complete genome sequence of *Bacillus thuringiensis* Al Hakam. *Journal of Bacteriology*, 189 (9): 3680-3681, May 2007

Challacombe, J.F., et al.
Complete genome sequence of *Haemophilus somnus* (*Histophilus somni*) strain 129Pt and comparison to *Haemophilus ducreyi* 35000HP and *Haemophilus influenzae* Rd. *Journal of Bacteriology*, 189 (5): 1890-1898, March 2007

Chistoserdova, L., et al.
Genome of *Methylobacillus flagellatus*, molecular basis for obligate methylotrophy, and polyphyletic origin of methylotrophy. *Journal of Bacteriology*, 189 (11): 4020-4027, June 2007

Coleman, M.L., et al.
Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science*, 311 (5768): 1768-1770, March 24, 2006

Collins, A.G., et al.
Medusozoan phylogeny and character evolution clarified by new large and small subunit rDNA data and an assessment of the utility of phylogenetic mixture models *Systematic Biology*, 55 (1): 97-115 February 2006

Cooper, S.J.B., et al.
The mammalian alpha(D)-globin gene lineage and a new model for the molecular evolution of alpha-globin gene clusters at the stem of the mammalian radiation. *Molecular Phylogenetics and Evolution*, 38 (2): 439-448 February 2006

Daly D.S., et al.
A mixed effects statistical model for comparative LC-MS proteomics studies. *Journal of Proteome Research* (in press).

Danilov, I.G., et al.
The type series of "*Sinemys*" *wuerhoensis*, a problematic turtle from the Lower Cretaceous of China, includes at least three taxa. *Palaeontology*, 50: 431-444 Part 2, March 2007

Danilov, I.G.,
A redescription of "*Polesiochelys*" *tatsuensis* from the Late Jurassic of China, with comments on the antiquity of the crown clade *Cryptodira*. *Journal of Vertebrate Paleontology*, 26 (3): 573-580, Sept. 11 2006

de Broin, F.D., et al.
Eurotestudo, a new genus for the species *Testudo hermanni* Gmelin, 1789 (*Chelonii, Testudinidae*). *Comptes Rendus Palevol* 5 (6): 803-811, September 2006

DeGusta, D., et al.
Primate phylogeny and divergence dates based on complete mitochondrial genomes. *American Journal of Physical Anthropology*: 78-78 Suppl. 44 2007

DeSantis, T.Z., et al.
Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72 (7): 5069-5072, July 2006

DeSantis, T.Z., et al.
NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Research*, 34: W394-W399 Sp. Iss. SI, July 1, 2006

Dehal, P.S., et al.
A phylogenomic gene cluster resource: The Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics*, 7: Art. No. 201, April 11, 2006

Di Gregorio, A., et al.
Gene regulation in the ancestral notochord: Insights from a collection of cis-regulatory elements from the ascidian *Ciona intestinalis*. *Developmental Biology*, 306 (1): 350-350 155 June 1, 2007

Doggett, N.A., et al.
A 360-kb interchromosomal duplication of the human HYDIN locus. *Genomics*, 88 (6): 762-771, December 2006

Field, D., et al.
The positive role of the ecological community in the genomic revolution. *Microbial Ecology*, 53 (3): 507-511, April 2007

Flanagan, J.L., et al.
Loss of bacterial diversity during antibiotic treatment of intubated patients colonized with *Pseudomonas aeruginosa*. *Journal of Clinical Microbiology*, 45 (6): 1954-1962, June 2007

Giraud, E., et al.
Legumes symbioses: Absence of Nod genes in photosynthetic bradyrhizobia. *Science*, 316 (5829): 1307-1312, June 1, 2007

Gowda, M., et al. Interrogation of the RNA species in *Magnaporthe grisea*. *Phytopathology*, 97 (7): S42-S42 Suppl. S, July 2007

Gowda, M., et al.
Robust analysis of 5 '-transcript ends (5 '-RATE): a novel technique for transcriptome analysis and genome annotation. *Nucleic Acids Research*, 34 (19): DOI:10.1093/nar/gkl522, November 2006

Appendix G: JGI Publications 2006–2007 (continued)

Grigoriev I.V., et al.
Fungal genomic annotation. *Applied Mycology and Biotechnology, Volume 6: Bioinformatics*. 2006 Elsevier, Amsterdam p 350. (Martinez corresponding author)

Grossman, A.R., et al. Novel metabolism in *Chlamydomonas* through the lens of genomics. *Current Opinion in Plant Biology*, 10 (2): 190-198, April 2007

Gu, S.Y., et al.
TreeQ-VISTA: an interactive tree visualization tool with functional annotation query capabilities. *Bioinformatics*, 23 (6): 764-766 March 15, 2007

Hackstein, J.H.P, et al.
Hydrogenosomes of anaerobic chytrids: An alternative way to adapt to anaerobic environments. In *Hydrogenosomes and Mitosomes: Mitochondria of Anaerobic Eukaryotes*. Ed. J. Tachezy. *Microbiology Monographs* (in press).

Hallam, S.J., et al.
Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *PNAS*, 103 (48): 18296-18301, Nov. 28, 2006

Hallam, S.J., et al.
Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine *Crenarchaeota*. *PLoS Biology*, 4 (4): 520-536, April 2006

Han, C.S., et al.
Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis*. *Journal of Bacteriology*, 188 (9): 3382-3390, May 2006

Han, C.S, et al.
Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis*. (Vol. 188, p 3382, 2006) *Journal of Bacteriology*, 188 (21): 7711, November 2006

Hellsten, U., et al.
Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis*. *BMC Biology*, 5: Art. No. 31, July 25, 2007

Hixson, K.K., et al.
Biomarker candidate identification in *Yersinia pestis* using organism wide semi-quantitative proteomics. *Journal of Proteome Research*. 5:3008-17.

Horner-Devine, C., et al.
A comparison of taxon co-occurrence patterns for macro and microorganisms. *Ecology* (in press).

Housman, D.C., et al.
Heterogeneity of soil nutrients and subsurface biota in a dryland ecosystem. *Soil Biology & Biochemistry*, 39:2138-2149.

Huerta, A.M., et al.
Selection for unequal densities of sigma(70) promoter-like signals in different regions of large bacterial genomes. *PLoS Genetics*, 2 (11): 1740-1750, November 2006

Huerta, A.M., et al.
Positional conservation of clusters of overlapping promoter-like sequences in enterobacterial genomes. *Molecular Biology and Evolution*, 23 (5): 997-1010, May 2006

Hugenholtz, P., et al.
Increasing focus on environmental biotechnology in the face of pressing environmental challenges. *Current Opinion in Biotechnology*, 18 (3): 235-236, June 2007

Hugenholtz, P.
Riding giants. *Environmental Microbiology*, 9 (1): 5-5, January 2007

Imachi, H., et al.
Non-sulfate-reducing, syntrophic bacteria affiliated with Desulfotomaculum cluster I are widely distributed in methanogenic environments. *Applied and Environmental Microbiology*, 72 (3): 2080-2091 March 2006

Janga, S.C. et al.
The distinctive signatures of promoter regions and operon junctions across prokaryotes. *Nucleic Acids Research*, 34 (14): 3980-3987 2006

Jeffries, T.W., et al.
Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nature Biotechnology*, 25 (3): 319-326 March 2007

Johnson, S.L., et al.
Export of nitrogenous compounds due to incomplete cycling within biological soil crusts of arid lands. *Environmental Microbiology*, 2007. 9 (3): 680-689.

Kane, S.R., et al.
Whole-genome analysis of the methyl tert-butyl ether-degrading beta-proteobacterium *Methylibium petroleiphilum* p.m.1. *Journal of Bacteriology*, 189 (5): 1931-1945, March 2007

Kane, S.R., et al.
Whole-genome analysis of the methyl tert-butyl ether-degrading beta-proteobacterium *Methylibium petroleiphilum* p.m.1. (Vol. 189, p 1931, 2007). *Journal of Bacteriology*, 189 (13): 4973-4973, July 2007

Kazakov, A.E., et al.
RegTransBase—a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Research*, 35: D407-D412 Sp. Iss. SI, January 2007

Kelleher, C.T., et al.
A physical map of the highly heterozygous *Populus* genome: Integration with the genome sequence and genetic map and analysis of haplotype variation. *Plant Journal*, 50 (6): 1063-1078, June 2007

Appendix G: JGI Publications 2006–2007 (continued)

Klotz, M.G., et al.
Complete genome sequence of the marine, chemolithoautotrophic, ammonia-oxidizing bacterium *Nitrosococcus oceani* ATCC 19707. *Applied and Environmental Microbiology*, 72 (9): 6299-6315, September 2006

Krampis, K., et al.
Extensive variation in nuclear mitochondrial DNA content between the genomes of *Phytophthora sojae* and *Phytophthora ramorum*. *Molecular Plant-Microbe Interactions*, 19 (12): 1329-1336, December 2006

Kryukov, G.V., et al.
Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *American Journal of Human Genetics*, 80 (4): 727-739 April 2007

Kubicek, C.P., et al.
Purifying selection and birth-and-death evolution in the class II hydrophobin gene families of the ascomycete *Trichoderma*/*Hypocrea*. *BMC Evolutionary Bioogy*. (in press).

Kunin, V., et al. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biology*, 8 (4): Art. No. R61 2007

Larroux, C., et al.
The NK homeobox gene cluster predates the origin of Hox genes. *Current Biology*, 17 (8): 706-710, April 17, 2007

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

Li, H., et al.
TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research*, 34: D572-D580 Sp. Iss. SI, Jan. 1, 2006

Li, M., et al.
Confidence-based classifier design. *Pattern Recognition*, 39 (7): 1230-1240, July 2006

Li, M.K., et al.
Confidence-based active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (8): 1251-1261, August 2006

Liolios, K., et al.
The genomes on line database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Research*, 34: D332-D334 Sp. Iss. SI, Jan. 1, 2006

Lo, I., et al. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature*, 446 (7135): 537-541, March 29 2007

Loots, G.G., et al.
Array2BIO: from microarray expression data to functional annotation of co-regulated genes. *BMC Bioinformatics*, 7: Art. No. 307, June 16, 2006

Lykidis, A.
Comparative genomics and evolution of eukaryotic phospholipid biosynthesis. *Progress in Lipid Research*, 46 (3-4): 171-199, May-July 2007

Lykidis, A.
Genomics and evolution of eukaryotic phospholipid biosynthesis. *FASEB Journal*, 21 (6): A976-A976, April 2007

Lykidis, A., et al.
Genome sequence and analysis of the soil cellulolytic actinomycete *Thermobifida fusca* YX. *Journal of Bacteriology*, 189 (6): 2477-2486, March 2007

Ma, A., et al.
Multimedia Content Representation, Classification and Security, 4105: 753-760 2006 Book series *Lecture Notes in Computer Science*, 2006

Maeder, D.L., et al.
The *Methanosarcina barkeri* genome: Comparative analysis with *Methanosarcina acetivorans* and *Methanosarcina mazei* reveals extensive rearrangement within methanosarcinal genomes. *Journal of Bacteriology*, 188 (22): 7922-7931, November 2006

Maeder, D.L., et al.
The *Methanosarcina barkeri* genome: Comparative analysis with *Methanosarcina acetivorans* and *Methanosarcina mazei* reveals extensive rearrangement within methanosarcinal genomes. (Vol. 188, p 7922, 2006) *Journal of Bacteriology*, 189 (4): 1488-1488, February 2007

Makarova, K,S,, et al.
Deinococcus geothermalis: The pool of extreme radiation resistance genes shrinks. *PLoS ONE*, 2007 Sep 26;2(9):e955.

Marco, E.J., et al.
ARHGEF9: Identification of a novel X-linked mental retardation and behavior disorder gene. *Annals of Neurology*, 60 (5): 635-636, November 2006

Marcy, Y., et al.
Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *PNAS*, 104 (29): 11889-11894, July 17, 2007

Markowitz, V.M., et al.
The Integrated Microbial Genomes (IMG) system. *Nucleic Acids Research*, 34: D344-D348 Sp. Iss. SI, Jan. 1, 2006

Martin, F.N., et al. Mitochondrial genome sequences and comparative genomics of *Phytophthora ramorum* and *P-sojae*. *Current Genetics*, 51 (5): 285-296, May 2007

Martin, H.G., et al.
Metagenomic analysis of two enhanced biological phosphorus re-

Appendix G: JGI Publications 2006–2007 (continued)

moval (EBPR) sludge communities. *Nature Biotechnology*, 24 (10): 1263-1269, October 2006

Mavromatis, K., et al. The genome of the obligately intracellular bacterium *Ehrlichia canis* reveals themes of complex membrane structure and immune evasion strategies. *Journal of Bacteriology*, 188 (11): 4015-4023, June 2006

McHardy, A.C., et al. Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, 4 (1): 63-72, January 2007

McMahon, K.D., et al. Integrating ecology into biotechnology. *Current Opinion in Biotechnology*, 18 (3): 287-292, June 2007

McPherson, R., et al. A common allele on chromosome 9 associated with coronary heart disease. *Science*, 316 (5830): 1488-1491, June 8, 2007

Medina, M. Collins, et al. Naked corals: Skeleton loss in *Scleractinia*. *PNAS*, 103 (24): 9096-9100, June 13, 2006

Meier, S., et al. Continuous molecular evolution of protein-domain structures by single amino acid changes. *Current Biology*, 17 (2): 173-178, Jan. 23, 2007

Merchant, S.S., et al. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science,* 318 (5848): 245-251, Oct. 12, 2007

Merchant, S., et al. Establishing potential chloroplast function through phylogenomics. *Photosynthesis Research*, 91 (2-3): 271-272 PS1813, February-March 2007

Montsant, A., et al. Identification and comparative genomic analysis of signaling and regulatory components in the diatom *Thalassiosira pseudonana*. *Journal of Phycology*, 43 (3): 585-604, June 2007

Moran, M.A., et al. Ecological genomics of marine roseobacters. *Applied and Environmental Microbiology*, 73 (14): 4559-4569, July 2007

Newton, I.L.G., et al. The *Calyptogena magnifica* chemoautotrophic symbiont genome. *Science*, 315 (5814): 998-1000, Feb. 16, 2007

Noonan, J.P., et al. Sequencing and analysis of Neanderthal genomic DNA. *Science*, 314 (5802): 1113-1118, Nov. 17, 2006

Normand, P., et al. Genome characteristics of facultatively symbiotic *Frankia* sp strains reflect host range and host plant biogeography. *Genome Research*, 17 (1): 7-15, January 2007

Nosenko, T., et al. Chimeric plastid proteome in the Florida "red tide" dinoflagellate *Karenia brevis*. *Molecular Biology and Evolution*, 23 (11): 2026-2038, November 2006

Oudot-Le Secq, M.P., et al. Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage. *Molecular Genetics and Genomics*, 277 (4): 427-439, April 2007

Palenik, B., et al. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *PNAS*, 104 (18): 7705-7710, May 1, 2007

Parham, J.F., et al. The phylogeny of Mediterranean tortoises and their close relatives based on complete mitochondrial genome sequences from museum specimens. *Molecular Phylogenetics and Evolution*, 38 (1): 50-64 January 2006

Parham, J.F., et al. The complete mitochondrial genome of the enigmatic bigheaded turtle (Platysternon): description of unusual genomic features and the reconciliation of phylogenetic hypotheses based on mitochondrial and nuclear DNA. *BMC Evolutionary Biology*, 6: Art. No. 11 Feb. 7, 2006

Parmley, D., et al. Diverse turtle fauna from the Late Eocene of Georgia including the oldest records of aquatic testudinoids in southeastern North America. *Journal of Herpetology*, 40 (3): 343-350, September 2006

Pennacchio, L.A., et al. *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature*, 444 (7118): 499-502, Nov. 23, 2006

Pennacchio, L.A., et al. Predicting tissue-specific enhancers in the human genome. *Genome Research*, 17 (2): 201-211, February 2007

Pandit, B., et al. Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. *Nature Genetics,* 39 (8): 1007-1012, August 2007

Podsiadlowski, L., et al. The mitochondrial genomes of *Campodea fragilis* and *Campodea lubbocki* (*Hexapoda*: *Diplura*): High genetic divergence in a morphologically uniform taxon. *Gene*, 381: 49-61, Oct. 15, 2006

Prabhakar, S. et al. Accelerated evolution of conserved noncoding sequences in humans. *Science*, 314 (5800): 786-786, Nov. 3, 2006

Appendix G: JGI Publications 2006–2007 (continued)

Prabhakar, S., et al.
Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Research*, 16 (7): 855-863, July 2006.

Prochnik, S.E., et al.
Evidence for a microRNA expansion in the bilaterian ancestor. *Development Genes and Evolution*, 217 (1): 73-77, January 2007

Putnam, N.H., et al.
Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, 317 (5834): 86-94, July 6, 2007

Raubeson, L.A., et al.
Comparative chloroplast genomics: Analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics*, 8: Art. No. 174, June 15, 2007

Rensing, S., et al.
The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science,* DOI: 10.1126/science.1150646

Romeo, S., et al.
Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature Genetics*, 39 (4): 513-516, April 2007

Rubin, E.M., et al.
Comparing Neanderthal and human genomes – Response. *Science*, 315 (5819): 1664-1664, March 23 2007

Schulte, J.A., et al.
A genetic perspective on the geographic association of taxa among and North American lizards of the *Sceloporus magister* complex (*Squamata: Iguanidae: Phrynosomatinae*). *Molecular Phylogenetics and Evolution*, 39 (3): 873-880, June 2006

Scott, K.M., et al.
The genome of deep-sea vent chemolithoautotroph *Thiomicrospira crunogena* XCL-2. *PLoS Biology*, 4 (12): 2196-2212, December 2006

Shoguchi, E., et al.
Chromosomal mapping of 170 BAC clones in the ascidian *Gona intestinalis*. *Genome Research*, 16 (2): 297-303 February 2006

Simison, W.B., et al.
Rolling circle amplification of metazoan mitochondrial genomes. *Molecular Phylogenetics and Evolution*, 39 (2): 562-567, May 2006

Slavotinek, A.M., et al.
Array comparative genomic hybridization in patients with congenital diaphragmatic hernia: mapping of four CDH-critical regions and sequencing of candidate genes at 15q26.1-15q26.2. *European Journal of Human Genetics*, 14 (9): 999-1008, September 2006

Smith T.J., et al.
Analysis of the neurotoxin complex genes in *Clostridium botulinum* subtype A1-A4 and B1 strains – BoNT/A3, BoNT/A4 and BoNT/B1 clusters are located within plasmids. *PLoS ONE* 2007:2:e1271, DOI:10.1371/journal.pone.0001271.

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

Sorek, R., et al.
Genome-wide experimental determination of barriers to horizontal gene transfer. Sciencexpress/www.sciencexpress.org/18 October 2007/Page 4/10.1126/ science. 1147112

Starkenburg, S.R., et al.
Genome sequence of the chemolithoautotrophic nitrite-oxidizing bacterium *Nitrobacter winogradskyi* Nb-255. *Applied and Environmental Microbiology*, 72 (3): 2050-2063 March 2006

Stillman, J.H., et al.
Construction and characterization of EST libraries from the porcelain crab, *Petrolisthes cinctipes*. *Integrative and Comparative Biology*, 46 (6): 919-930, December 2006

Stuart, B.L., et al.
Recent hybrid origin of three rare Chinese turtles. *Conservation Genetics*, 8 (1): 169-175, February 2007

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

Tartaglia, M., et al. Gain-of-function SOS1 mutations cause a distinctive form of Noonan syndrome. *Nature Genetics*, 39 (1): 75-79, January 2007

Tang, Y.J., et al. Flux analysis of central metabolic pathways in *Geobacter metallireducens* during reduction of soluble Fe(III)-nitrilotriacetic acid. *Applied and Environmental Microbiology*, 73 (12): 3859-3864, June 2007

Timme, R.E., et al.
A comparative analysis of the *Lactuca* and *Helianthus* (*Asteraceae*) plastid genomes: Identification of divergent regions and categorization of shared repeats. *American Journal of Botany*, 94 (3): 302-312, March 2007

Tuanyok, A., et al.
A horizontal gene transfer event defines two distinct groups within the *Burkholderia pseudomallei* that have dissimilar geographic distributions. *Journal of Bacteriology*, 24:9044-9049.

Tuskan, G.A., et al.
The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).

Appendix G: JGI Publications 2006–2007 (continued)

Science, 313 (5793): 1596-1604, Sept. 15, 2006

Tyler, B.M., et al.
Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*, 313 (5791): 1261-1266, Sept. 1, 2006

Udwary, D.W., et al.
Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *PNAS*, 104 (25): 10376-10381, June 19, 2007

Valles, Y., et al.
Lophotrochozoan mitochondrial genomes. *Integrative and Comparative Biology*, 46 (4): 544-557, August 2006

Vanden Wymelenberg, A., et al.
Computational analysis of the *Phanerochaete chrysosporium* v2.0 genome database and mass spectrometry identification of peptides in ligninolytic cultures reveal complex mixtures of secreted proteins. *Fungal Genetics and Biology*, 43 (5): 343-356, May 2006

Vannini, C., et al.
Endosymbiosis *in statu nascendi*: close phylogenetic relationship between obligately endosymbiotic and obligately free-living *Polynucleobacter* strains (*Betaproteobacteria*). *Environmental Microbiology*, 9 (2): 347-359, February 2007

Visel, A., et al. Enhancer identification through comparative genomics. *Seminars in Cell and Developmental Biology*, 18 (1): 140-152, February 2007

Visel, A., et al.
VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Research*, 35: D88-D92 Sp. Iss. SI, January 2007

Volik, S., et al.
Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Research*, 16 (3): 394-404 March 2006

von Mering, C., et al.
Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315 (5815): 1126-1130, Feb. 23, 2007

Wang, Q.F., et al.
Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. *Genome Biology*, 8 (1): Art. No. R1 2007

Wang, Q.F., et al.
Primate-specific evolution of an LDLR enhancer. *Genome Biology*, 7 (8): Art. No. 68, 2006

Ward, R.D., et al.
Comparative genomics reveals functional transcriptional control sequences in the Prop1 gene. *Mammalian Genome*, 18 (6-7): 521-537, July 2007

Warnecke, F., et al.
Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature,* Vol. 450|22 November 2007| DOI:10.1038/nature06269

Warren, R.L.; et al.
Physical map-assisted whole-genome shotgun sequence assemblies. *Genome Research*, 16 (6): 768-775, June 2006

Weng, L., et al.
Application of sequence-based methods in human microbial ecology. *Genome Research*, 16 (3): 316-322 March 2006

Williams, L.E., et al.
Facile recovery of individual high-molecular-weight, low-copy-number natural plasmids for genomic sequencing. *Applied and Environmental Microbiology*, 72 (7): 4899-4906, July 2006

Wong, R., et al.
The story told by 100 shotgun-sequenced microbial genomes—Why the clonability for genomes are different. *Plasmid*, 57 (2): 239-239, March 2007

Woyke, T., et al.
Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, 443 (7114): 950-955, Oct. 26, 2006

Yamada, T., et al.
Characterization of filamentous bacteria, belonging to candidate phylum KSB3, that are associated with bulking in methanogenic granular sludges. *ISME Journal*, 1 (3): 246-255, July 2007

Xie, G., et al.
Genome sequence of the cellulolytic gliding bacterium *Cytophaga hutchinsonii*. *Applied and Environmental Microbiology*, 73 (11): 3536-3546, June 2007

Zhang, K., et al.
Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology*, 24 (6): 680-686, June 2006

Zhou K., et al.
Proteomics for validation of automated gene model predictions. In *Mass Spectrometry of Proteins and Peptides*, eds: M Lipton, L Paša-Tolic. *Humana Press* (in press).

DISCLAIMER

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231, and Los Alamos National Laboratory under Contract No. W-7405-ENG-36.

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

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

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

LBNL-63471
CSO 14128