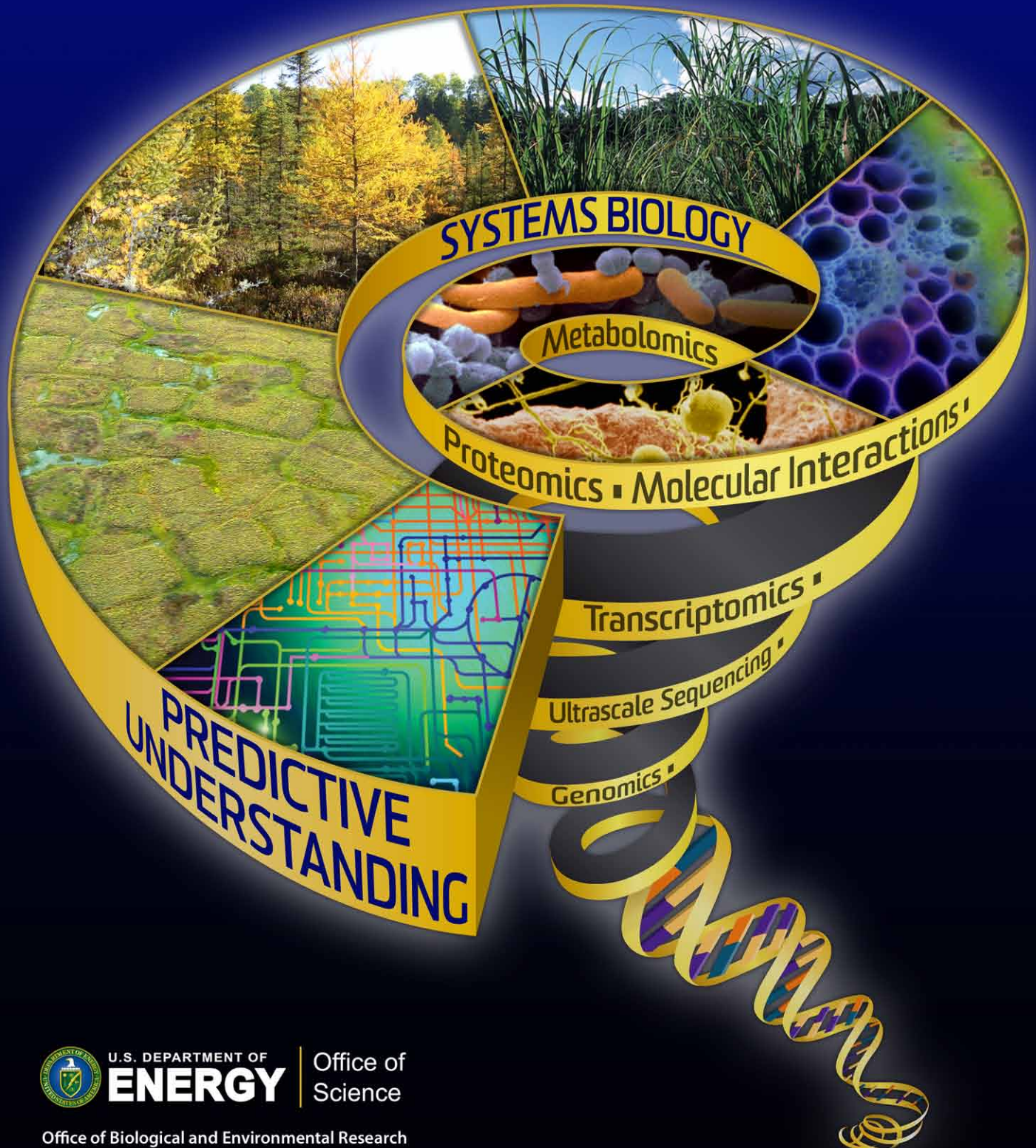


# DOE Joint Genome Institute Strategic Planning for the Genomic Sciences

Workshop Report

DOE/SC-0152



U. S. DEPARTMENT OF  
**ENERGY**

Office of  
Science

Office of Biological and Environmental Research

# DOE Joint Genome Institute Strategic Planning for the Genomic Sciences Report from the May 30–31, 2012, Workshop

Convened by

**U.S. Department of Energy**  
**Office of Science**  
**Office of Biological and Environmental Research**

## Co-Chairs

**Jim Fredrickson** Pacific Northwest National Laboratory  
**Michael Laub** Massachusetts Institute of Technology  
**Jan Leach** Colorado State University

## Breakout Group Chairs

**Richard Michelmore** University of California, Davis  
**Kimmen Sjölander** University of California, Berkeley  
**Tom Schmidt** Michigan State University

## Organizer

**Daniel Drell**

daniel.drell@science.doe.gov, 301.903.4742

---

Web address for this document:

[genomicscience.energy.gov/userfacilities/jgi/futuredirections/](http://genomicscience.energy.gov/userfacilities/jgi/futuredirections/)



## About the Cover

Images on the cover represent a broad range of the complex scales encompassed by the science supported by the Office of Biological and Environmental Research (BER) within the U.S. Department of Energy's (DOE) Office of Science. These scales range from genes at the subcellular level to genomes of microbes and their communities, to the genomics of plant-microbe interactions and plants that could be feedstocks for bioenergy, to the scale of ecosystem and landscape function. The underlying DNA strand represents the DNA sequence data that provide the foundation for further systems-level experimentation. The images culminate in a wired cell that represents the predictive understanding of biological systems sought by BER programs. DNA sequence data generated by the DOE Joint Genome Institute user facility are having a major impact toward the achievement of this goal.

**Image credits:** Microscopic images of organic matter decomposers (copyright Corbis), fungal hyphae on a root surface (copyright Corbis), and cross-section of a switchgrass stem (DOE BioEnergy Science Center and National Renewable Energy Laboratory). Grassland habitat (U.S. Department of Agriculture Natural Resources Conservation Service). Spruce-peatland ecosystem (Oak Ridge National Laboratory). Aerial view of Arctic landscape (Oak Ridge National Laboratory). Cover developed at Oak Ridge National Laboratory.

## Suggested Citation

U.S. DOE. 2012. *DOE Joint Genome Institute Strategic Planning for the Genomic Sciences: Report from the May 30–31, 2012, Workshop*, DOE/SC-0152, U.S. Department of Energy Office of Science.

# DOE Joint Genome Institute Strategic Planning for the Genomic Sciences

## Report from the May 30–31, 2012, Workshop

Published September 2012

Convened by

U.S. Department of Energy  
Office of Science  
Office of Biological and Environmental Research



U.S. DEPARTMENT OF  
**ENERGY**

Office of  
Science



# Contents

Director’s Letter.....	v
Executive Summary.....	vii
Sidebar 1: Major Themes and Needs Emerging from the Workshop.....	vii
Introduction and Background.....	1
Sidebar 2: Decreasing the Lag Time Between Sequencing and Annotation.....	2
Sidebar 3: DOE Biological and Environmental Research Program Perspective.....	3
Sidebar 4: DOE Genomic Science Systems Biology Program.....	5
Grand Challenges.....	9
Next-Generation Enabling Capabilities.....	13
Sidebar 5: Functional Annotation: A Prerequisite for Predictive Biology.....	15
Sidebar 6: Automating Science in a Robotic Laboratory.....	19
Summary.....	21
Appendices.....	23
Appendix 1: Grand Challenges.....	23
Appendix 2: Department of Energy Assets.....	29
DOE Joint Genome Institute.....	29
DOE Environmental Molecular Sciences Laboratory.....	30
DOE Systems Biology Knowledgebase.....	31
DOE Synchrotron and Neutron Beam Facilities for Biology.....	32
Appendix 3: Workshop Agenda, Charge Questions, Participants.....	33
Appendix 4: Bibliography.....	37
Appendix 5: Glossary.....	39
Acronyms and Abbreviations.....	Inside back cover





## Department of Energy

Washington, DC 20585

September 20, 2012

In October 2011, the U. S. Department of Energy (DOE) Joint Genome Institute (JGI) issued a draft “10-Year Strategic Vision: Forging the Future of the DOE JGI.” This document provided a high-level overview of DOE JGI and its plans to evolve as a next-generation genomic science user facility. The intent was to draft a vision for DOE JGI that goes beyond just sequence generation and seeks new technologies and/or capabilities to enhance the interpretation and use of genomic data. The draft document took advantage of a recent assessment (*Grand Challenges for Biological and Environmental Research: A Long-Term Vision* DOE/SC-0135) of the major long-term scientific challenges in energy and the environment that are the core mission areas of the DOE Office of Biological and Environmental Research (BER) and outlines how DOE JGI must evolve to help meet these research challenges.

In May 2012, BER hosted a separate workshop on “DOE JGI Strategic Planning for the Genomic Sciences” to solicit additional community input towards articulating a high-level DOE Office of Science vision for DOE JGI’s role in advancing BER mission science. The intention was not to explore *how* DOE JGI could evolve to be a next-generation genome center but rather to explore *why* DOE JGI should become a next-generation genome center. The workshop attendees focused on the future of genomic science in the context of the scientific challenges central to BER’s mission and the central role that a DOE JGI with enhanced capabilities could play in advancing BER science. The report from this workshop builds on the DOE JGI 10-Year Strategic Vision document ([www.jgi.doe.gov/whoware/10-Year-JGI-Strategic-Vision.pdf](http://www.jgi.doe.gov/whoware/10-Year-JGI-Strategic-Vision.pdf)) but focuses more on the challenges and capabilities envisioned for a next-generation genome center.

The two documents complement each other and recognize that genome sequencing, once a separate goal itself, is now just an initial step towards gaining a functional understanding of biological processes. To capitalize on the benefits of genome sequencing now taking place at ever greater rates, additional capabilities must be developed to bring added value to the sequences produced and to associate those sequences with biological meaning. Both documents inform future efforts at DOE JGI and within BER to accelerate the understanding of biological processes in support of DOE’s energy and environmental missions.

Sincerely,

A handwritten signature in blue ink that reads "R. Todd Anderson".

R. Todd Anderson  
 Director  
 Biological Systems Science Division, SC-23.2  
 Office of Biological and Environmental Research  
 Office of Science





## Executive Summary

The U.S. Department of Energy (DOE) Joint Genome Institute (JGI) Strategic Planning for the Genomic Sciences workshop was convened by the DOE Office of Science’s Biological and Environmental Research Program (BER) on May 30–31, 2012. The goal was to explore DOE JGI’s role in addressing DOE mission-critical scientific questions and in contributing data and knowledge to enable a new generation of systems biology research (see Sidebar 1, this page).

DOE JGI has played a leadership role in genome sequencing, providing a foundation for complex biological studies. As DOE JGI moves forward into the next decade(s), it will have a continued and expanded role in genome sequencing. A major opportunity lies before DOE JGI as it seeks to build on its sequencing strength by providing high-throughput, “value-added” science that can be integrated with massive sequence datasets to accelerate the science underpinning DOE missions in bioenergy and the environment.

In particular, as a DOE scientific user facility that accelerates users’ research with capabilities not available in their own laboratories, DOE JGI could lower access barriers to cutting-edge capabilities and provide expertise to advance mission-relevant science by reducing the gap between genotype data (i.e., the gene sequence) and

phenotype data (what genes and their products do). Addressing this gap is of fundamental importance to the DOE mission (see Appendix 1: Grand Challenges, p. 23).

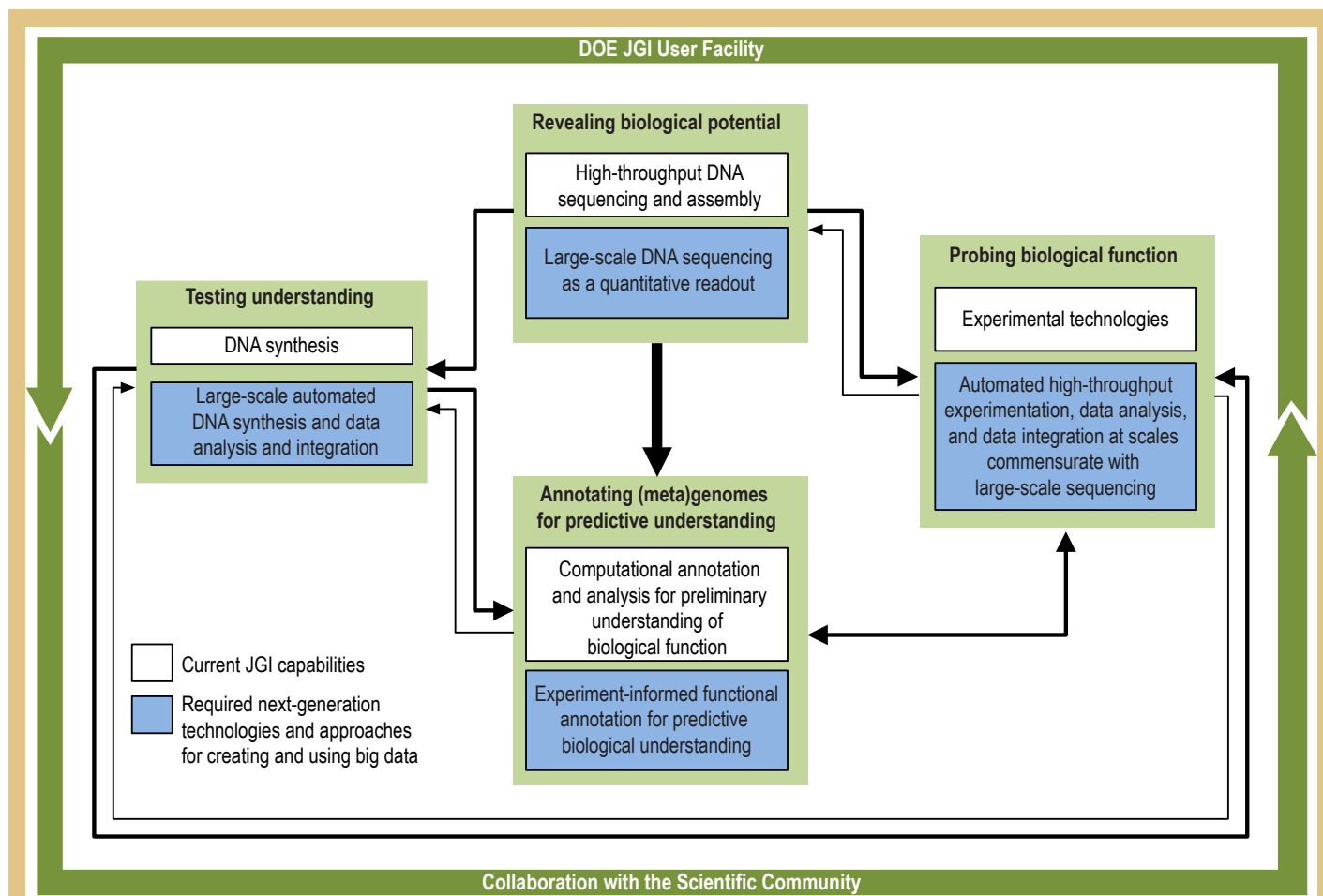
To elicit DOE JGI strategies for achieving these advances, BER invited participants from universities, DOE national laboratories, and other federal agencies with broad expertise in the biological sciences and bioinformatics. The workshop’s conclusions highlight the following capabilities: sequencing

### Sidebar 1

#### Major Themes and Needs Emerging from the Workshop

1. Continued and increasing large-scale sequencing of biologically important organisms and communities. Extant sequencing targets range from viruses to complex, multicellular communities associated with a wide range of environments directly relevant to Department of Energy (DOE) and Biological and Environmental Research Program (BER) missions. Sequencing is becoming increasingly important as an experimental measurement, adding to or even replacing current measurement technologies—a quantitative “readout” rather than just an end in itself.
2. Large-scale functional genomics technologies for high-throughput functional annotation, informed by global measurements of actual cellular activities rooted in genome sequencing. Functional genomics technologies include those provided by whole-expression, proteomic, metabolomic, and genome-wide association studies, at several levels, including individual cells, multicellular organisms, and communities.
3. Extended and improved bioinformatics methods to enable integration and analysis of unprecedented quantities of data and to generate testable hypotheses critical to advancing DOE science.
4. Aggressively expanded capacity not only to sequence (“read”) DNA, but also to synthesize (“write”) DNA, enabling scientists to manipulate genomes. This capability will be essential for directed exploration of gene and genomic manipulations of biological functions relevant to DOE missions.
5. Improved automation of biological experiments to match the throughput now prevalent in sequencing. The costs of these technologies must be reduced so they can keep pace with improvements in sequencing technologies.
6. Communities of scientists led by the DOE Joint Genome Institute (JGI) and organized around key mission-relevant scientific questions. Addressing these questions and challenges will be feasible through the development of new biological approaches made possible by next-generation sequencing and follow-on technologies (that DOE JGI could establish and/or adopt). Novel biological approaches are critical to build an understanding of, and an ability to predict, biological behaviors required for complex applications such as biofuel production and understanding of biological feedbacks to the climate system.

## Executive Summary



**Fig. 1. Pathway to Predictive Biological Understanding.** This figure depicts an integrated view of the current and future capabilities to be explored by the DOE JGI user facility in collaboration with the research community. These capabilities are required for improving the accuracy, efficiency, and effectiveness of annotations resulting from computational analyses. DNA sequence data provide the foundation for predictive understanding, revealing the biological potential in a genome or genomes. Experimentation probes the behavior of biological systems under different environmental conditions and informs hypotheses generated from sequencing and analyses. Because of the output of current and imminent sequencing technologies, however, high-throughput experimental data analysis and data integration capabilities are needed at lower cost and greater level of automation. New understanding gained through these integrated efforts can then be tested by additional experiments, including the use of synthetic techniques to build test systems. The thickness of the arrows represents the level of current challenges to information flow. Importantly, the setting of this conceptual pathway inside of the larger scientific community implies a strong supportive and collaborative relationship. The resulting experiment-informed functional annotations will enable research into more complex biological systems that could not otherwise be studied, leading to the predictive understanding required for DOE missions in bioenergy and the environment.

DNA, annotating DNA, addressing “Big Data” opportunities and challenges, writing DNA and developing associated technologies, implementing high-throughput experimentation, and building research communities (see Fig. 1. Pathway to Predictive Biological Understanding, this page).

### Continuing to Sequence DNA

The impressive accumulated sequencing accomplished to date by DOE JGI and other genome centers is insignificant compared with the diversity and sheer number of microbes, fungi, plants, other eukaryotes, and particularly complex

communities that remain unexplored. This diversity and variation is critical to biological functions relevant to DOE missions and is only beginning to be sampled by current sequencing efforts. Consequently, there was strong consensus among workshop participants that DOE JGI should continue high-throughput and high-quality sequencing. DOE JGI thus will play an important role in refining and developing sequencing-related technologies, including single-cell sequencing and metatranscriptome analysis of diverse species and communities.

Although sequencing capacity continues to grow rapidly, future efforts need to be guided by scientific questions relevant to DOE missions in bioenergy and the environment. To obtain the most value from continued sequencing, carefully selected model species and “model” environments should be identified to provide basic sequence knowledge and to nucleate better functional and structural genome annotations. Continued sequencing will supply the foundational “raw material” for DOE JGI contributions to:

- Building mechanistic models for biological processes.
- Generating a deep understanding of key (“flagship”) organisms.
- Reducing the fraction of genes whose functions remain unknown through improved function prediction protocols.
- Supporting investigations into the individual genetic variation within cells in a species, within species in a population, and between populations.
- Monitoring changes in expression profiles over time and other high-volume, sequence-critical biological measurements.

## Annotating DNA

A consensus of workshop participants agreed that functional annotation following genomic sequencing and structural annotation is one of the biggest challenges confronting the entire biology community (not just DOE JGI). Functional annotations for newly identified genes will require a combination of computational and experimental methodologies, enabling DOE JGI to generate testable hypotheses of function and to capture experimental data where available.

A number of avenues are available for DOE JGI to become a more involved participant in functional annotation and to narrow the gap between generating and understanding genome sequence. Ideas emerging from the workshop included DOE

JGI (1) having a leadership role in integrating different data types from sources worldwide; (2) engaging the scientific community to reveal genomic “dark matter” (e.g., conserved hypothetical genes), perhaps by evolving a competition process or exploiting crowd-sourcing and social networking technologies; (3) exploring the utility of new experimental and computational technologies to integrate data and functional inference protocols for genes, genomes, and metagenomes; (4) generating or acquiring validated high-accuracy genome and gene reference datasets; (5) seeding the development of active end-user communities involved in functional annotation; and (6) encouraging and assisting the development of improved experimental design approaches.

Functional annotations could benefit from novel bioinformatics technologies to accelerate experimental validation of putative functional assignments, as well as from better ways of linking plant or microbial genetic diversity with ecosystem function.

## Addressing “Big Data” Opportunities and Challenges

As sequencing technologies have increased in speed and throughput, data generation has exceeded both storage and analysis capabilities. New computational tools to acquire, curate, analyze, and distribute information from these datasets are needed for all DOE JGI user communities. Addressing scientific questions relevant to DOE missions will require effective data integration. This task is dependent on the generation of complementary data types (e.g., transcriptomics and proteomics) in appropriate volumes; intelligent merging and association of disparate data types from a wide variety of national and international entities; and development of appropriate conventions for controlled vocabularies, genome descriptions, and metadata. To achieve these capabilities, DOE JGI must promote resource integration involving other facilities and utilization of other DOE Office of Science assets (see Appendix 2: Department of Energy Assets, p. 29).

## Writing DNA and Developing Associated Technologies

A clear message from the workshop was that DOE JGI should establish complementary technological capabilities, including “on demand” DNA synthesis and miniaturization of appropriate analysis technologies. “Writing” DNA is viewed

## Executive Summary

as a necessary complement to sequencing in ways that would accelerate DOE JGI science. Several applications are centered on generating desired sequences, introducing them into an appropriate cell, and exploring the effects on physiology, metabolism, cellular architecture, and responses to stimuli. By altering genetic information in defined ways, the process of linking the behaviors of gene products to sequence variants will be accelerated dramatically. Similarly, the relation of sequence to three-dimensional (3D) structures, of both proteins and chromosomes, can be explored much more effectively. DNA synthesis is a powerful tool for investigating biological processes, ranging from the individual gene or gene product (protein) to protein complexes, metabolic pathways, regulatory networks, and even an entire cell and complex communities. This scale remains difficult and challenging but now can be addressed with new high(er)-throughput approaches.

### Implementing High-Throughput Experimentation

An important opportunity for DOE, BER, and DOE JGI is to combine the ideas and algorithms underlying the automation of hypothesis testing with the economies of scale possible with microfluidics devices. These combined

capabilities could dramatically reduce the cost of doing science, by sifting the credible hypotheses (the wheat) from the less credible (the chaff), improving the pace of discovery in functional genomics, directed evolution, and biological design. DOE JGI can accelerate this hypothesis-generation process with advanced automated experimental technologies that contribute to the understanding of genome sequences.

### Building Research Communities

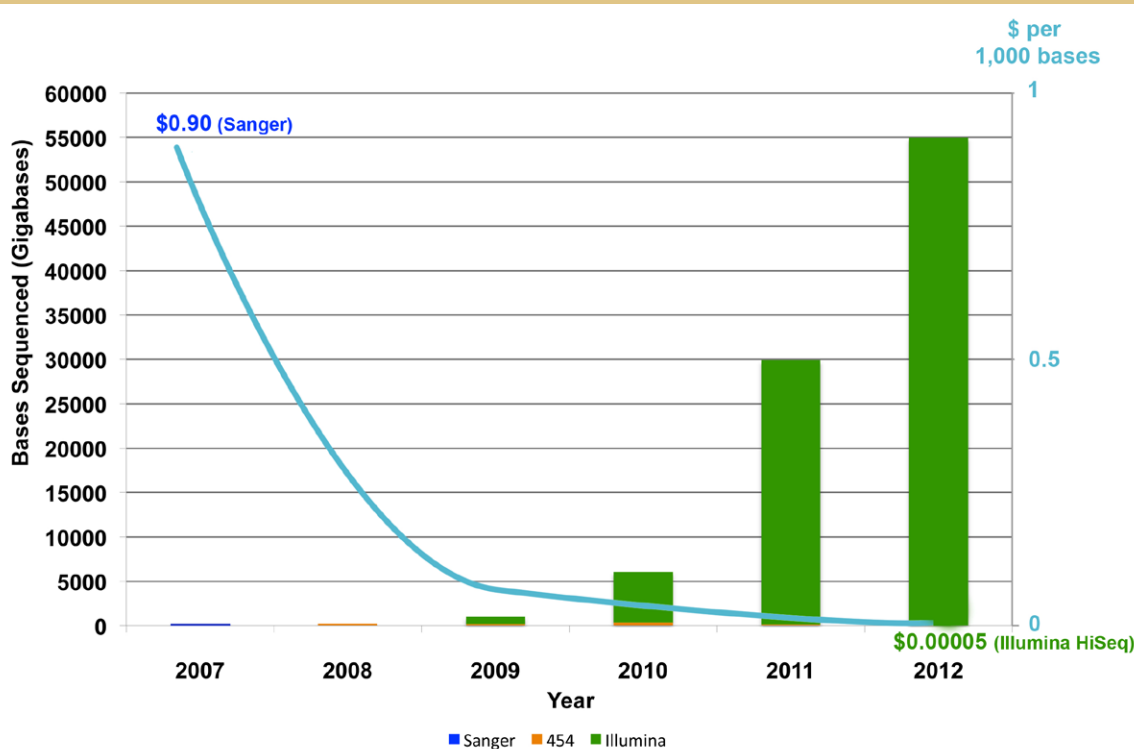
DOE JGI should continue to serve as a user facility, providing a valuable service to the biological community and participating in active collaborations to achieve science that biologists could not readily carry out in their individual laboratories. To this end, DOE JGI should stimulate and support community efforts for large science by providing high-throughput sequencing and other value-added services focused on problems of scale and complexity that exceed the ordinary. This endeavor will require a sustained commitment of time and resources and take advantage of DOE JGI's unique abilities and expertise. DOE JGI should actively seed and promote interdisciplinary teams and initiate stable collaborations among individual researchers, laboratories, and institutions that focus on mission-relevant science.

## Introduction and Background

In 1997, the U.S. Department of Energy (DOE) Office of Biological and Environmental Research (BER) established the DOE Joint Genome Institute (DOE JGI) with the goal of sequencing three human chromosomes totaling about 330 million bases, or 11% of the human genome. Bringing together expertise from multiple national laboratories in DNA sequencing, informatics, and technology development, DOE JGI, in its first year, sequenced 20 million bases of human DNA. In 2012, DOE JGI will sequence 55 trillion bases, an increase of more than six orders of magnitude. This rate of increase in throughput, a consequence of remarkable new sequencing technologies, exceeds Moore's Law for the growth in processing capacity on computer microprocessors. Put another way, the 1997 cost of determining the 3 billion bases of the human genome was about \$1 per "finished" (i.e., high quality) base. Today, DOE JGI sequences DNA at a cost of roughly \$1 per 20 million "raw" bases (see Fig. 2. DOE JGI Sequencing Economies of Scale,

this page), which corresponds roughly to \$3,000 per human genome when raw bases are resolved into finished bases.

The principles that have given rise to the striking improvement of sequencing technologies are now available for the development of other new cutting-edge genomic technologies. However, while sequencing capacity worldwide has increased exponentially, the rest of the biological research pipeline (genome assembly and structural/functional annotation) has not kept pace, and the gap is growing (see Sidebar 2, Decreasing the Lag Time Between Sequencing and Annotation, p. 2). The bioinformatics challenges are technical and scientific. For both prokaryotic and eukaryotic genomes, a single gene can encode multiple proteins, an individual protein can carry out more than one function, and protein function can be influenced by the physical or cellular environment. Gene function must also be interpreted in a systems biology context, as many cell functions are the consequence of multiple gene products



**Fig. 2. DOE JGI Sequencing Economies of Scale.** New sequencing platforms have led to a dramatic increase in throughput and an equally dramatic decrease in cost.

## Sidebar 2

## Decreasing the Lag Time Between Sequencing and Annotation

Although difficult to quantify, but widely held, is the increasing gap between the pace of sequencing technologies and that of sequence annotation. The rate of gene sequencing is readily charted (see Fig. 3. Rate of Sequencing Outpaces Moore's Law, p. 3), but determining the rate of annotation is far from straightforward. Genome *structural* annotation—gene identification and model refinement and the identification of promoter sites and other genome features—has improved significantly. It scales well with the pace of whole-genome sequencing (although challenges remain with microbial community datasets). However, genome *functional* annotation presents a major challenge, in which heuristic and rough annotation methods scale, but precise and informative annotation methods do not. The problem is that not all annotations are equal. Some annotations are clearly more information rich than others, and some are just plain wrong—upwards of 25% of genes are estimated to have errors in their functional annotations. Another 30% or more of genes in a typical genome are labeled as hypothetical or unknown, and this fraction has remained fairly constant over the years. The propagation of dubious annotations to newly sequenced genes and genomes further undermines annotation quality. Though manual curation and annotation are of great value, they do not scale. Experimental data are extremely sparse, with <1% of sequences having any experimental support.

Many difficult questions remain. How can the value of annotations associated with sequences be quantified, and are these quantification methods improving? Is effective understanding growing as fast as sequence databases? Are novel gene families being detected and their functions

determined? Is the fraction of sequences labeled as hypothetical or unknown being reduced? Are annotations increasing in detail and specificity and, as a result, in their utility? If a gene experiment is published, does the published information contribute to that gene's annotation? Are existing annotation errors being detected and corrected, or do they persist and propagate? Can biases be identified that may have skewed and limited understanding? For instance, much is known about certain bacterial species that can be cultured in the laboratory, and almost nothing is known about those that cannot be cultured. Meanwhile, environmental and metagenomic studies are revealing a vast, uncharacterized microbial universe, much of which is phylogenetically distant from any species that have been studied or sequenced. Which tools can be used to understand these data and annotate these sequences?

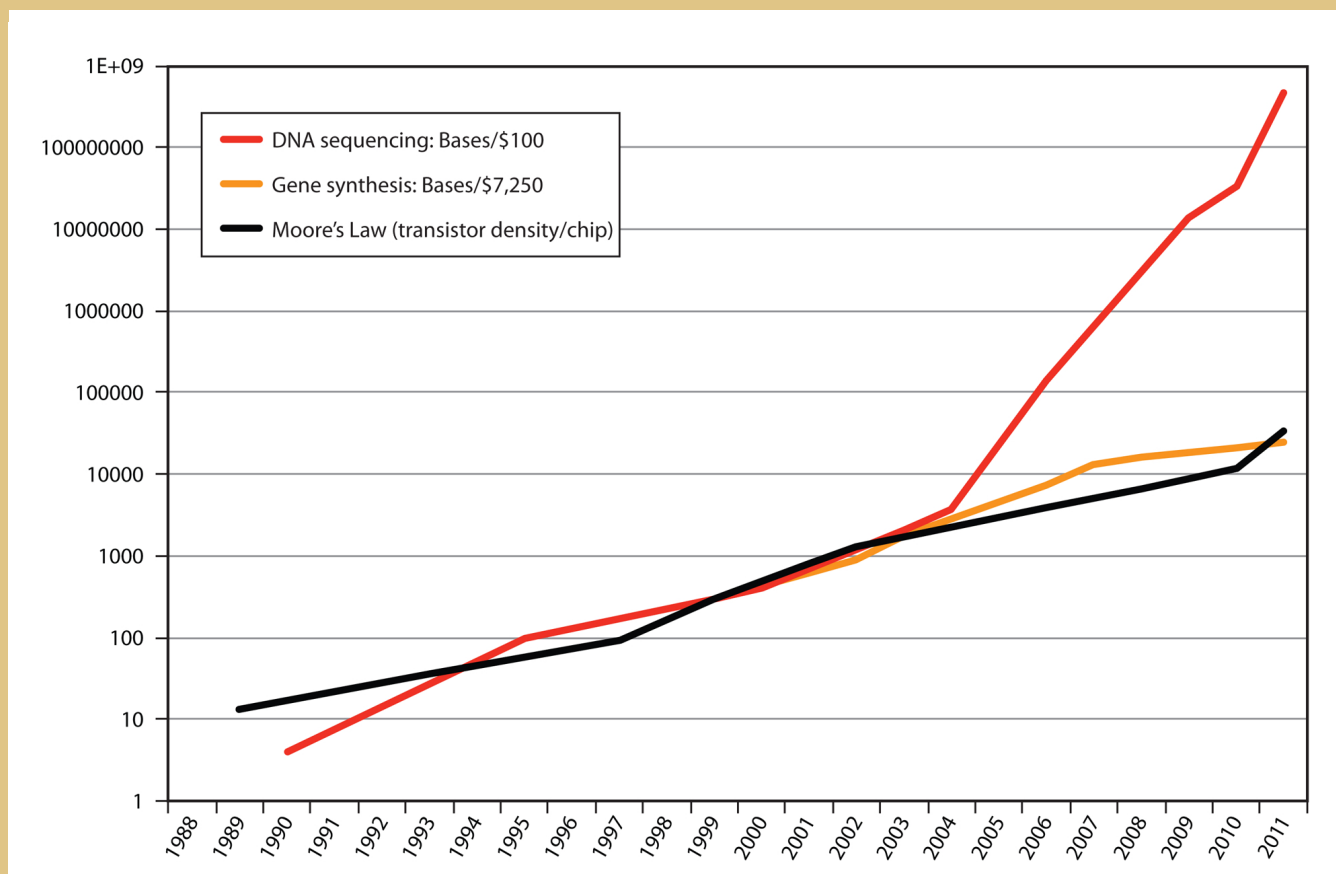
One way to evaluate the actual contribution of a computationally derived sequence annotation (and, by extension, the information gained from genome annotation) is to measure its contribution to the design of an experiment that determines the function empirically. In other words, a computationally produced annotation should be seen not as an answer but as a hypothesis to be tested. An ideal bioinformatics annotation system would not stop at computational prediction; instead it would suggest the experiments needed to confirm or refute the predicted function and then feed these results to a robotized system for high-throughput experiments. The experimental results (prediction was correct or incorrect) could be fed back to the computational method, allowing the system to learn from its mistakes in a powerful feedback loop (see Big Data section, p. 16, and Fig. 5. Toward a More Systematic Annotation Pathway, p. 17).

acting together in tightly regulated “molecular machines” (e.g., the ribosome) embedded in complex networks. Determining the function of a gene's product, therefore, may first require establishing which “machine” is at work and identifying the pathways and networks in which it participates.

Microbial genomes are dynamic and amazingly adaptive, exchanging DNA segments readily through a process known as horizontal gene transfer (HGT). This genomic plasticity confers enormous selective advantages in constantly and rapidly changing environments, permitting microbes to colonize an astonishing variety of environmental niches, including

some at extremes in temperature, acidity, pressure, humidity, and even radiation levels.

Microbes dominate by their sheer numbers. A 1998 perspective by William Whitman and colleagues derived an estimate of 4 to  $6 \times 10^{30}$  for the number of microbes in, on, and under the earth. As a recent National Research Council (NRC) report (*The New Science of Metagenomics*) correctly noted, “Microbes run the world. It's that simple.” The gene space and biochemical abilities of most of this vast biome are simply unknown (see Appendix 4: Bibliography, p. 37).



**Fig. 3. Rate of Sequencing Outpaces Moore's Law.** The chart compares productivity increases in DNA sequencing, gene synthesis, and transistor density (Moore's Law or the observation that the number of transistors on a computer chip doubles about every 2 years). Sequencing productivity is exceeding the other increases by a widening margin. The challenges of representing the gene annotation rate on this curve (much slower than the sequencing rate) are described in Sidebar 2. The data are presented so that productivity increases are seen as upward trends, with all curves normalized at the same point in 1999, the approximate midpoint of the time range. DNA sequencing and gene synthesis are expressed in terms of bases (sequenced or synthesized) per unit money, while transistor improvement is expressed in terms of density per chip (currently about 76,000 per chip). The DNA sequencing and gene synthesis cost curves are adapted from the Carlson cost curves ([www.synthesis.cc/cgi-bin/mt/mt-search.cgi?blog\\_id=1&tag=Carlson%20Curves&limit=20](http://www.synthesis.cc/cgi-bin/mt/mt-search.cgi?blog_id=1&tag=Carlson%20Curves&limit=20)). The transistor curve was adapted from [www.sciencephoto.com/media/348724/enlarge#](http://www.sciencephoto.com/media/348724/enlarge#). (Note that the y axis is a  $\log_{10}$  scale.)

### Sidebar 3

## DOE Biological and Environmental Research Program Perspective

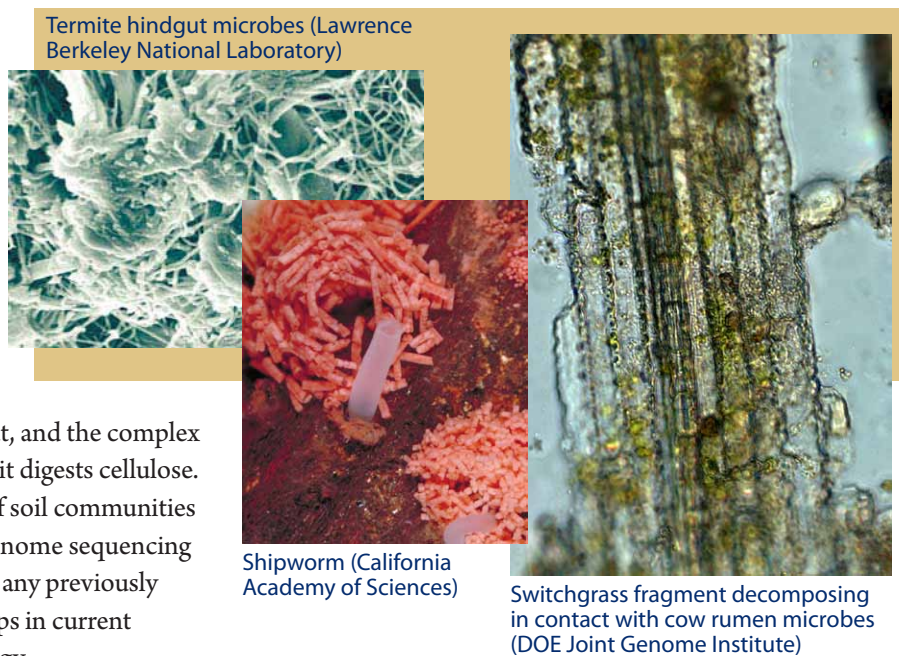
Biological systems science is essential to the U.S. science enterprise and the development of a new bioeconomy. Systems biology is the multidisciplinary study of complex interactions specifying the function of entire biological systems—from pathways and organelles to single cells, populations of cells, multicellular organisms, microbial communities, and interorganismal interactions. Key questions that drive these studies include:

- What information, beyond just the parts list of proteins, is encoded in the genome sequence?

- How is this information translated to functional phenotypes and coordinated among different subcellular constituents?
- What molecular interactions regulate the response of living systems, and how can those interactions be understood dynamically and predictively over time and space?

BER supports DOE JGI to help address these questions and advance DOE missions in sustainable bioenergy production and an in-depth understanding of the roles of biological systems in climate and environmental processes.

DOE JGI, a designated BER user facility (see Appendix 2: DOE Assets, p. 29), is unique not only for its project management and sequencing prowess, but for its leadership in sequencing complex microbial communities found in harsh and extreme environments in which microbes live and carry out reactions relevant to DOE missions. Microbial communities sequenced by DOE JGI have included those within acid mine drainage, hot springs, the wood-digesting termite hindgut, and the complex biochemical cauldron of the cow rumen as it digests cellulose. The wood-boring shipworm and an array of soil communities have been sequenced as well. These metagenome sequencing projects often reveal novel microbes unlike any previously sequenced or studied, highlighting large gaps in current knowledge of microbial diversity and ecology.



Termite hindgut microbes (Lawrence Berkeley National Laboratory)

Shipworm (California Academy of Sciences)

Switchgrass fragment decomposing in contact with cow rumen microbes (DOE Joint Genome Institute)

In summary, sequencing has only skimmed the surface of the microbial world. The number and diversity of microbial species are extraordinary and, as a direct result, so is the repertoire of potential biochemistries and capabilities. However, many microbes are notoriously difficult to culture so they cannot be readily studied. Consequently, most of what is known about microbial biology has been derived from the study of a small number of cultivated microbes. There remains a vast “microbial dark matter” of unexplored organisms and genes that promises many new discoveries pertinent to DOE missions.

This situation is evolving rapidly, bolstered by improved technologies for sequencing the genomes of single cells (circumventing the culturing barrier) and by high-throughput sequencing of environmental DNA samples (metagenomics). Despite such advances, finding ways to help the DOE Genomic Science program achieve its goals and aid the rest of biology in keeping pace is imperative (see Sidebar 3, DOE Biological and Environmental Research Program Perspective, p. 3, and Sidebar 4, DOE Genomic Science Systems Biology Program, p. 5).

Sequencing efforts targeting plants and other important eukaryotes also are enabling new and unexpected insights. These very challenging projects, much larger and often more demanding than microbial sequencing, are identifying unexpected capacities and opportunities for BER mission-relevant

science. For instance, plant genome sequences and analyses provide a basis for understanding plant cell wall composition and properties, growth characteristics, disease resistance, and other important traits. These technologies also are revealing the genomic contributions to plant-microbe interactions, which are vital to plant survival and fitness. Such sequencing projects were unimaginable before the advent of high-throughput sequencing, and DOE JGI has become the prime source for new plant genomes.

In addition to insights into genomic organization and evolutionary relationships (via comparisons with genomes from other related organisms), a genome sequence provides a “parts list” for an organism’s biology. A genome sequence is also a “hypothesis-generation engine” (see Fig. 4. Hypothesis-Generating and Validation Engine for Fundamental Systems Biology, p. 6). The parts list reveals the cell’s genetic potential—both what the cell *can* do in different environmental conditions and what it presumably *cannot* do (i.e., if specific genes are not present). Such a parts list is merely the first step in an analysis, though, because the biological “whole” (the organization and functioning of a cell) is far more than just the “sum of its parts.” As with any hypothesis, experimental validation is required.



## Sidebar 4

## Introduction and Background

## DOE Genomic Science Systems Biology Program

The U.S. Department of Energy's (DOE) Genomic Science program uses microbial and plant genomic data, high-throughput analytical technologies, and modeling and simulation to develop a predictive understanding of biological systems behavior relevant to solving energy and environmental challenges including bioenergy production, environmental remediation, and climate stabilization. As elaborated in the Biological and Environmental Research (BER) program Mission Statement, BER supports fundamental research and scientific user facilities to address diverse and critical global challenges. BER's Genomic Science program seeks to understand how genomic information is translated into functional capabilities, enabling more confident redesign of microbes and plants for sustainable biofuel production, improved carbon storage, or contaminant bioremediation. BER research advances understanding of the roles of Earth's biogeochemical systems (i.e., the atmosphere, land, oceans, sea ice, and subsurface) in determining climate to enable predictions of climates decades or centuries into the future—information needed to plan for future energy and resource needs. Solutions to these challenges are driven by a foundation of scientific knowledge and inquiry in atmospheric chemistry and physics, ecology, biology, and biogeochemistry.

In contrast to a reductionist study of individual components in isolation, systems biology uses comprehensive, multidisciplinary

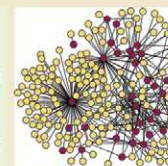
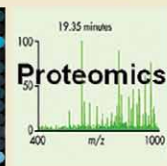
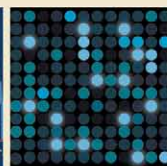
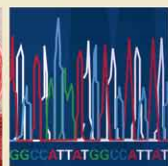
studies of complex interactions to specify the function(s) of an entire biological system, from single cells to multicellular organisms and communities. The Genomic Science program builds on a foundation of sequenced genomes to identify the common fundamental principles that drive living systems. These principles guide the translation of genomic code into functional molecules underlying biological system behavior. For example, understanding these principles could determine the biological mechanisms controlling changes in the production of cellulose-degrading enzymes in an industrial bioreactor or the amount of carbon stored in a plant's roots under different environmental conditions. Knowledge of these common principles revealed by studying organisms for one DOE mission inevitably will lead to breakthroughs in basic biology important to other DOE and national needs.

Addressing extremely complex science questions that span all scales of biology, research supported by DOE's Genomic Science program requires the collective expertise of scientists from many disciplines and the coordinated application of a wide range of technologies and experimental approaches—genome sequencing, gene expression profiling, proteomics, metabolomics, imaging, research technology development, and computational biology. Research is conducted at national laboratories, national user facilities, and universities and spans single-investigator

projects, multi-institutional collaborations, and fundamental research centers. The Genomic Science program is managed by BER within DOE's Office of Science.

## Genomic Science Program Goal and Objectives

## Genome Sequence



## System-Wide Biological Investigations

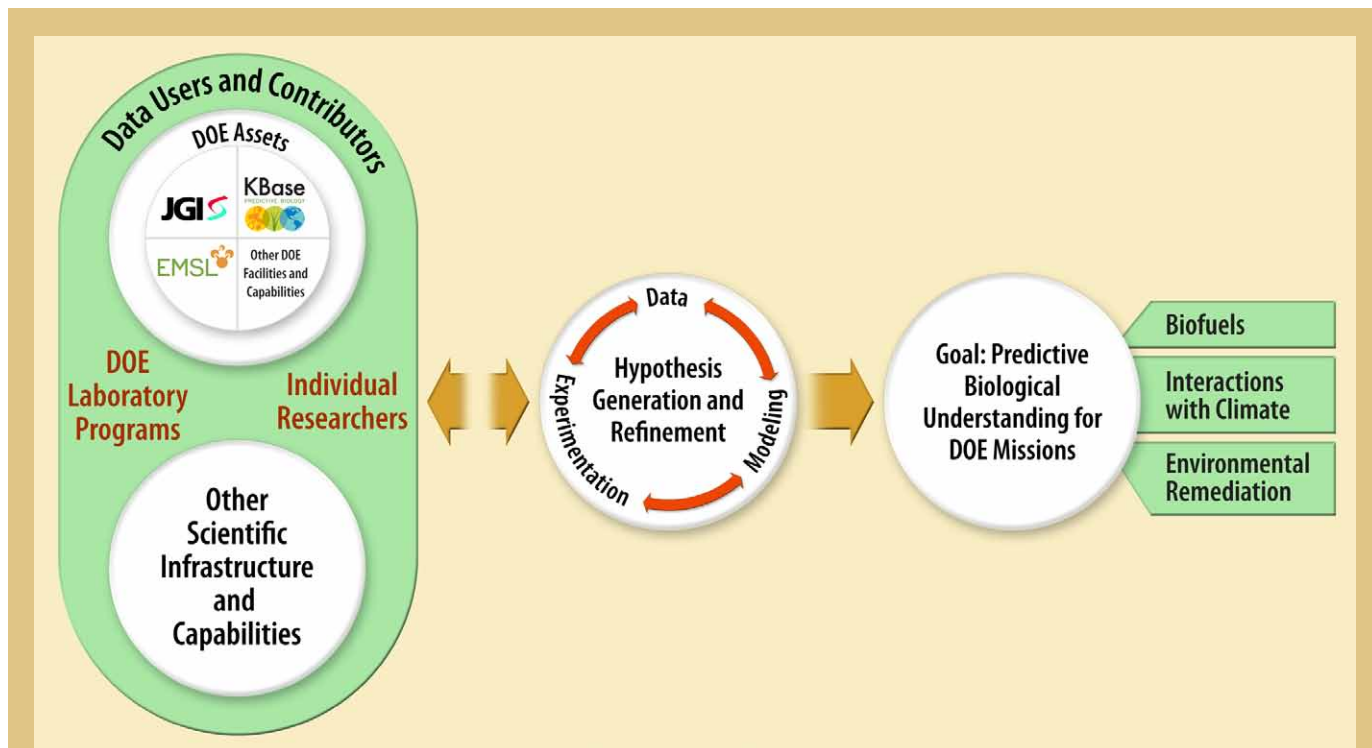
## Predictive Understanding

**Goal:** Achieve a predictive, system-level understanding of plants, microbes, and biological communities, via integration of fundamental science and technology development, to enable biological solutions to DOE mission challenges in energy, environment, and climate.

- **Objective 1:** Determine the genomic properties, molecular and regulatory mechanisms, and resulting functional potential of microbes, plants, and biological communities central to DOE missions.

- **Objective 2:** Develop the experimental capabilities and enabling technologies needed to achieve a genome-based, dynamic system-level understanding of organism and community function.
- **Objective 3:** Develop the knowledgebase, computational infrastructure, and modeling capabilities to advance the understanding, prediction, and manipulation of complex biological systems.

## Introduction and Background



**Fig. 4. Hypothesis-Generating and Validation Engine for Fundamental Systems Biology.** Intense, specialized capabilities accelerate the hypothesis-generating engine, enabling targeted, more effective experiments for answering scientific questions posed by the research community. The DOE JGI, Systems Biology Knowledgebase, Environmental Molecular Sciences Laboratory, and other DOE assets (see Appendix 2: DOE Assets, p. 29) are providing integrative genome sequencing, experimentation, modeling, and database portals that are open access and community driven. The end result advances predictive understanding for DOE missions in bioenergy and the environment.

### DOE JGI: Current and Potential Future Capabilities

DOE JGI will remain a national user facility providing massive-scale DNA sequencing and associated analysis capabilities dedicated to advancing genomics for bioenergy and environmental applications. In addition to its major accomplishments and capabilities in high-throughput sequencing of microbes, microbial communities, and plants, DOE JGI has been part of nascent developments in transcriptomics to understand the expression of proteins under different conditions. The facility also has developed a pipeline for studying metagenomic sequences from a variety of communities and technologies to assemble whole sequences of uncultured microbial organisms using a combination of single-cell sequencing techniques and sophisticated informatics analysis. With respect to plant genomics capabilities, DOE JGI is the premier facility for large-scale sequencing of complex

plants including poplar and switchgrass, both important as potential biofuel sources. As DOE JGI moves forward in plant sequencing science, active goals include genomic analyses of large populations, diverse natural populations, and mutants, as well as transcriptomics of model plant species and their associated microbial communities.

DOE JGI will continue to support bold and innovative BER research programs aimed at addressing “Grand Challenge” scale projects in biology (see Grand Challenges chapter, p. 9, and Appendix 1: Grand Challenges, p. 23). These challenges will require expanding both the scope of genome sequencing and its utility as a tool for the “readout” of systems-level responses to external perturbations. In other words, sequencing is more than an endpoint, it is a key experimental tool.

As DOE JGI sequencing throughput continues to increase, future major challenges to DOE mission-relevant science will require improved analyses of sequence data to enable

### Potential Bioenergy Crops

Switchgrass (Lawrence Berkeley National Laboratory)



Poplar leaf (DOE Joint Genome Institute)



Sorghum (Lawrence Berkeley National Laboratory)

next-generation systems biology research. The application of sequencing technologies to increasingly complex biological questions will present novel challenges and opportunities to generate critically needed knowledge of organisms relevant to bioenergy and the environment. BER's rich history of supporting innovative science and technology has significantly advanced biology. For example, BER-supported DNA sequencing and genomics technologies contributed to the sequencing and assembly of the human genome, numerous microbes, microbial communities, and plants, enabling significant advances to DOE science missions in bioenergy and the environment.

One of the most powerful approaches to inferring gene functions, as well as evolutionary relationships of genes and genomes, is comparative sequence analysis. Robust comparisons to existing, carefully curated sequence can reduce the search space and accelerate scientific discovery. For example, adaptation to new environments and conditions generally proceeds by selection acting on modifications of pre-existing

capabilities favorable to survival. Comparative analyses of related organisms can reveal this genomic variation on which the evolutionary process acts and link function to this variation. Thus, continued sequencing of biologically and/or environmentally related organisms that span the phylogenetic diversity of life will reveal biological functions and relationships that will contribute to DOE mission-relevant science.

Biology is increasingly data rich. Sequencing technologies and throughput are continuing to improve, and data are being generated at rates exceeding the imagination less than a decade ago. But data is not the same as information, particularly with respect to systems-level understanding and prediction. Much of the data being generated are difficult or impossible to interpret with existing bioinformatics technologies. Connecting genotype to phenotype across all of biology remains a major challenge that must be met to convert data to knowledge. There remains a pressing need for a paradigm shift in biology, providing the resources and capabilities in high-throughput reading and synthesis of DNA and phenotypic assays in combination with robust computing infrastructure. These capabilities, connecting to and integrating BER science mission user communities, will provide the scientific underpinnings essential for solving society's most pressing problems.

To explore ways that DOE JGI could better help the scientific community exploit genome sequences, DOE BER organized the May 2012 DOE JGI Strategic Planning for the Genomic Sciences workshop, which resulted in this report.

The following sections of this workshop report provide the collective vision of workshop participants for new and enhanced capabilities in biology over the timeframe of 10 years and longer. These capabilities are beyond those currently available to the scientific community and thus present opportunities for DOE JGI.

At this workshop, 35 scientists from a broad array of backgrounds met to discuss a set of charge questions (see Appendix 3: Workshop Agenda, Charge Questions, Participants, p. 34) focused on next-generation genomics-enabled biology relevant to DOE missions. Among participants were nine DOE national laboratory scientists, 24 academic scientists, and two private-sector scientists, as well as federal agency representatives. They took part in plenary sessions

## Introduction and Background

featuring speakers that provided overviews and perspectives of workshop topics and in breakout discussions organized around the charge questions. Rapporteurs presented their respective breakout group discussions to the entire group following each breakout. A final summary perspective was provided by the co-chairs.

The goal of this workshop was to explore the role of DOE JGI in addressing DOE mission-critical scientific questions and how the data and knowledge generated by DOE JGI can best be utilized to enable a new generation of systems biology research. Given that sequence data continue to rapidly increase in quantity and be highly useful (as repeatedly emphasized by workshop participants), an urgent need identified is for high-throughput, “value-added” science that can be integrated with massive sequence datasets to best accelerate the science underpinning DOE missions in bioenergy and the environment.

At the outset, participants were asked to think boldly, which they did. In his keynote presentation, Greg Petsko (Brandeis University) listed some key criteria for the participants to keep in mind, including:

- It is important to have (or support) platforms that cover the gamut from sequence to functional (and biological) characterizations.
- There are many sequences and annotations, but most of them are probably wrong, at least in part.
- If value can be added to sequences, sequencing can continue.

The ultimate goal is to understand the organization and functions of biological systems relevant to DOE missions, not just the acquisition of data for the sake of acquiring data. Thus, the primary imperative needs to be on adding value to these data, a mission that the next-generation DOE JGI must undertake.

## Grand Challenges

In 2008, at the request of the U.S. Department of Energy (DOE), the National Science Foundation, and the National Institutes of Health, the National Research Council's (NRC) Board on Life Sciences convened a committee to recommend how best to capitalize on recent advances in biology, many enabled by genome sequencing, to predict the behavior of complex biological systems. The “new biology” was defined by the integration of biology’s many subdisciplines and the inclusion of physicists, chemists, computer scientists, engineers, and mathematicians into a community with the capacity to tackle a broad range of scientific and societal problems. The NRC committee also identified several societal challenges for the new biology, including two pertaining directly to DOE missions: to understand and sustain ecosystem function and biodiversity in the face of rapid change and to expand sustainable alternatives to fossil fuels (NRC 2009).

Underlying these societal challenges are scientific grand challenges for DOE’s Office of Biological and Environmental Research (BER) that were defined in a 2010 Biological and Environmental Research Advisory Committee (BERAC) workshop report (BERAC 2010). BER has been a leader in the new biology by emphasizing integrated, multidisciplinary science in its research programs and specifically of the systems biology research it supports (DOE Genomics:GTL Roadmap 2005; DOE Genomics:GTL Strategic Plan 2008; see Sidebar 4, DOE Genomic Science Systems Biology Program, p. 5).

Grand challenges in biological systems identified in the BERAC report were broadly categorized into three groups:

- **Enabling predictive biology.** Use robust biochemical, functional, and experimental evidence to enhance genome and metagenome annotation.
- **Measuring and analyzing biological systems.** Apply advanced computational and analytical capabilities to characterize the molecules and network interactions used by biological systems.
- **Exploring ecosystem function and elemental cycling.** Develop designs for optimizing carbon flow for biomass production, carbon allocation, and biosequestration to reduce rates of atmospheric carbon dioxide (CO<sub>2</sub>) accumulation to increase terrestrial carbon storage by 50% in 20 years.

DOE Joint Genome Institute (JGI) capabilities are important to all three of these broad themes, but current DOE JGI sequencing focuses largely on the third. Below are several exemplar grand challenges for DOE JGI (discussed in greater detail in Appendix 1, p. 23). *These grand challenges are intended to serve as examples of the type of science that would benefit from and be enabled by the advanced capabilities in genomics that a user facility such as DOE JGI could provide.* All require high-throughput generation of multiple types of data (not just sequence). These challenges also require integrated contributions from multiple scientific disciplines. It is not envisioned that DOE JGI could, or should, attempt to accomplish these grand challenges independently; rather, DOE JGI should actively collaborate with the larger scientific community and other resources to develop and apply new capabilities and approaches (see chapter, Next-Generation Enabling Capabilities, p. 13).

Each of these grand challenges represents science of critical importance to DOE missions. This science requires the integration of high-throughput capabilities and multidisciplinary expertise provided by facilities such as DOE JGI, which already has initiated, with its collaborators, pilot-scale efforts in some of these topic areas. These are among the most challenging problems in current science and, as such, represent critical targets for major discoveries. Significant progress on each will necessitate a suite of capabilities beyond that currently provided by DOE JGI. This does not mean, however, that sequencing is passé, that there is no continued need for it; in fact, just the opposite is true, as the grand challenges listed below exemplify. By itself, sequencing is not an endpoint but rather a primary technology that enables diverse biological explorations.

### Grand Challenges

**Designer Phototrophs: Engineering Cyanobacteria to Produce Biofuels.** The diverse realm of cyanobacteria offers considerable promise for the sustainable production of biofuels from light and



Oil-producing green microalga, *Botryococcus braunii* (University of Jaén, Spain)

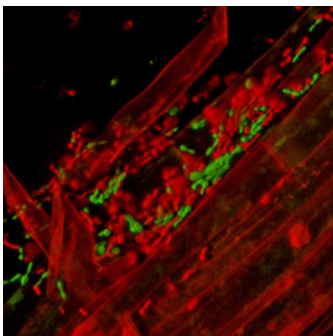
## Grand Challenges

CO<sub>2</sub>. The more than 170 cyanobacterial genomes sequenced to date have produced an extensive potential parts list for bioengineering, but more genomes are needed. In particular, a deeper understanding of the functions of the “parts” is required to advance bioenergy research. One result of greater DOE JGI throughput will be the acceleration of knowledge directly applicable to a key national and DOE priority.

### Understanding Interactions Between Microbes and Climate.

The microbes, fungi, soils, and plants that affect climate processes have not been the subject of sequencing efforts on the same scale as those involved in other activities. Much greater DOE JGI focus will generate the basic data necessary to understand the interactions and contributions of biology to climate cycles. Additionally, plants, microbes, and their interactions are critically important biological drivers of ecosystem function and monitors of changes in ecosystem status. The rhizosphere is the dynamic interface between plant roots and both abiotic and biotic environments; it is central to how plants sense and respond to the soil microbiome and to how plants benefit from the biological composition of the soils in which they reside and grow. DOE JGI has just begun to characterize these interactions at the genomic level. Additional sequencing and functional characterization will enable a transition from “snapshots” of the rhizosphere to “moving pictures,” permitting a deeper understanding of biotic and abiotic soil factors that contribute to useful plant performance.

**Advanced Genomic Capabilities for Biofuel Sustainability.** Alternative fuels from renewable biomass—plant stalks, trunks, stems, and leaves—are expected to significantly



Fluorescent micrograph of microbes isolated from a poplar rhizosphere to determine colonization patterns (Oak Ridge National Laboratory)



Brown-rot fungus, *Postia placenta*, breaks down hemicellulose and cellulose (Forest Products Laboratory)

reduce U.S. dependence on imported oil and decrease the environmental impacts of energy use. Biomass is largely composed of plant cell walls, and the major cell wall polymers are cellulose (sugars) and lignin (phenolic compounds). Controlling the ratio of cellulose to lignin in plant cell walls would enable feedstock optimization for different energy-conversion processes. DOE JGI’s high-throughput sequencing, supplemented by complementary data and improved computational tools, is needed to determine how the interacting networks of genes, proteins, and metabolites control cell wall composition and can be altered to optimize it—all keys to a primary DOE mission.

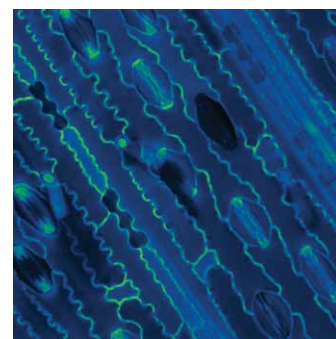
### Mining Natural Variation to Improve Energy Capture (Photosynthesis) in Plants.

Plants capture energy using photosynthesis to fix CO<sub>2</sub> and generate oxygen (O<sub>2</sub>) and biomass. Photosynthesis is famously inefficient. Improving the process has proven to be difficult, but one solution is to exploit natural variation.

The improvements and reduced costs of DOE JGI high-throughput sequencing, combined with new experimental and computational approaches, could lead to identification and characterization of natural variants for enhancing plant photosynthetic capacity and other complex traits relevant to biomass productivity.

### Dynamics of Genomic Participation in CO<sub>2</sub> Cycling in Marine Environments.

The marine biosphere is responsible for half the global primary production of organic compounds and half the annual CO<sub>2</sub> uptake from Earth’s atmosphere. A deeper understanding



Micrograph of a *Miscanthus* leaf, a potential bioenergy crop (IGB Microscopy Facility/EBI-UIUC)



Picophytoplankton, *Micromonas*, influences several biogeochemical cycles, including carbon (National Center for Microscopy and Imaging Research/Monterey Bay Aquarium Research Institute)

of the marine biosphere has tremendous implications for DOE missions in climate science and bioenergy. Marine CO<sub>2</sub> uptake is regulated by a complex array of microbes (comprising ~98% of ocean biomass), whose dynamics, interactions, and overall controls are not well understood. Challenges in this area stem from the largely unmapped taxonomic diversity of marine microbes and the complexity of biogeochemical transformations integrally linked via different modes or “lifestyles.”

#### **Characterizing Horizontal Gene Transfer (HGT).**

HGT—the nonvertical acquisition of genes—is relatively rare across eukaryotic species but surprisingly common between microbial genomes. The increased characterization of HGT from high-throughput sequencing at DOE JGI would accelerate (1) realizing the practical implications that include predicting bacterial metabolic capabilities from sequencing, (2) monitoring community processes in

response to environmental perturbations, and (3) determining the origins of new variants. Potential future HGT applications include altering microbes or plants in defined ways with predictable outcomes.



Genome of *Thermotoga maritima*, a hyperthermophilic bacterium, was sequenced by The Institute for Genomic Research and presented the first concrete evidence of horizontal gene transfer (University of Regensburg, Germany)





# Next-Generation Enabling Capabilities

## Continuing to Sequence DNA

The decreased cost and higher throughput of sequencing technologies already have made possible the sequencing of genomes at a scale that would have been unthinkable even a few years ago. In the years to come, improvements to scalability and cost will continue, enabling the Department of Energy (DOE) Joint Genome Institute (JGI) to contribute even more significantly to this arena. Although the reduced cost of sequencing technologies will enable some individual investigators to take on independent sequencing and assembly, data utilization will become an increasing challenge. DOE JGI's role as a user facility for genome sequencing and analysis thus will have continuing value. Serving as a user facility will enable DOE JGI to ensure standardization of methodologies and data integration and management. Decreasing cost in conjunction with higher fidelity means that sequencing can be applied to a growing number of experimental scenarios and biological investigations.

Sequencing can be subdivided into projects in which the sequence is used to inform downstream functional objectives and those in which the DNA sequence itself is used to draw conclusions (such as in population genetics). These two objectives result in different demands from sequencing and data handling. Downstream functional analyses often need high-quality sequence data, while surveys of large numbers of genotypes within and between species favor higher throughput at the potential expense of accuracy.

### Sequencing for functional objectives

The genome sequence provides a scaffold for functional data (e.g., results of functional assays for individual genes). Having high-quality reference genomes for species relevant to DOE's mission is thus important. However, generating high-quality reference genomes remains a time-consuming, challenging, and continuous process. Assembly algorithms need to be improved and adapted to emerging sequencing technologies. Heterozygosity and structural variability across species, polyploidy, and other issues represent technical hurdles to genome assembly. Metagenomes, particularly of microbial communities, also present unique challenges, especially for fragment assembly into candidate

genomes. New sequencing technologies will help meet these challenges, and DOE JGI should continue to be at the forefront of adopting these technologies and developing both bench-level laboratory and computational approaches to exploit them.

Genome sequencing should be targeted strategically, toward both broad sampling of taxonomic groups of interest to DOE and deep sequencing of selected taxa and specific species. Each sequencing strategy provides specific advantages. Broad genome sequencing strategies that aim to provide representative annotated genomes for all main taxonomic groups are required for DOE objectives in two key areas: functional annotation of genomes and functional and taxonomic annotation of metagenome datasets. Deep sequencing within a lineage—for example, sequencing multiple strains of the same species or closely related species of a genus—is essential for understanding the impact of genetic variation on phenotype.

Novel genome data are required not only to increase understanding of biological diversity (which has large gaps), but to understand the genomes of species *already sequenced*. Broad representation of taxonomic groups is necessary for improving inferences of function based on distant homologies, allowing bioinformatics methods to “connect the dots” and help reduce the number of sequences annotated as hypothetical or unknown. Sufficient genome sequences also are needed for evolutionary studies, which are dependent on dense taxon sampling. Although evolutionary studies are not a specific DOE objective, accuracy in evolutionary reconstruction is needed for interpreting and predicting the origin of species in microbial community datasets. Expanded genome sequencing also will improve the accuracy of multi-gene family trees (i.e., including paralogous genes), allowing phylogenomic methods of function prediction to be used with greater precision.

A small set of key genomes should be selected for in-depth functional annotation, using both manual curation and computational (automated) methods. These genomes can serve as the equivalents of template proteins in the National Institutes of Health (NIH) Protein Structure Initiative. Related genomes can then be annotated using annotation

## Next-Generation Enabling Capabilities

transfer protocols, provided the functional annotations of a small set of representative genomes have been curated to a high accuracy.

New sequencing technologies and new protocols for current technologies increasingly will allow the analysis of epigenetic modifications. As opportunities arise, DOE JGI needs to take a lead role in incorporating epigenomics analyses at the population scale.

### Survey sequencing

The ever-decreasing sequencing costs along with increasing output provide the opportunity to sample biological diversity over time and space. DOE JGI should aim to generate a library of allelic variation in phage, microbes, and plants for phenotypes relevant to DOE missions, such as nitrogen fixation, photosynthetic efficiency, and carbon sequestration. This library will be a critical component for genome-wide association studies (GWAS) and quantitative trait locus (QTL) analyses of numerous large populations, studies that associate genotype to phenotype. Critical issues will be the scope and scale of such projects. The power of GWAS is greatly enhanced by scale. Thus, DOE JGI should focus on enabling genotyping projects that provide unprecedented temporal and spatial resolution not achievable by individual laboratories or larger research efforts. This endeavor will require automated processing and sequencing of thousands of samples to capture genetic diversity.

### Sequencing as a readout

The high information content of DNA sequence data lends itself to serving as a highly informative, quantitative readout for a broad variety of phenotyping at multiple levels, including expression analysis [RNA-Seq and expression quantitative trait locus (eQTL) analysis], high-throughput yeast 2-hybrid (protein-protein and protein–nucleic acid interactions), and selection and phenotyping screens (functional analysis of gene modifications). Sequencing also can be used as a readout for population genetics and for understanding the dynamics of population changes in response to environmental perturbations in natural and agricultural ecosystems as well as in experimental situations.

### *Inferring selective pressures within microbial communities through detection of de novo mutations*

Up to now, metagenomic data have been mined primarily for species composition and secondarily for gene content information. Current technology is enabling the determination of how environmental perturbations affect microbial communities at the level of species composition and physiological adaptation as read through gene expression. However, another important level of adaptation to environmental change is through newly appearing mutations that spread via natural selection. In the near future, advances in the depth of sequencing will allow scientists to resolve new mutations, at the single-base level, arising within individual species embedded in a complex community. Such fine-scale information on species evolution in metagenomic data will enable the identification of key selective pressures acting on each individual species within a community in response to new environmental challenges (e.g., changing climate). The capacity to infer selective pressures acting on each species within a community will be critical to understanding the contribution of each species to the emergent properties of its community. In particular, this methodology will help both to inform predictions about how climate change will affect carbon cycling in natural microbial communities and to facilitate the synthetic construction and directed evolution of microbial communities for environmental process manipulation or biofuel production.

### Annotating DNA

DNA annotation is a multistep process, starting with genome structural annotation (gene identification and the annotation of promoter sites and other genomic features) and a preliminary functional annotation. Ideally, these two aspects of genome annotation would be iterated, making use of new data to correct and refine initial models with potentially limited accuracy. Gene model accuracy remains a challenge for most eukaryotic genomes and, to a somewhat lesser degree, for microbial genomes. Because errors in gene models will impact downstream analyses (e.g., functional annotation), DOE JGI should incorporate advances in gene model prediction tools in its pipeline.

### **Functional annotation: Computational**

Functional annotations—including molecular or biochemical function, metabolic pathway association, protein-protein interaction, subcellular localization, and protein three-dimensional (3D) structure—are most commonly produced using sequence similarity to other genes (e.g., transferring the annotation of the top BLAST hit) (see Sidebar 5, this page). Because annotation-transfer protocols are associated with high functional annotation error rates (as much as 25%, and possibly more, of genes are estimated to have errors in their functional annotations), DOE JGI should be at the forefront of adopting increasingly sophisticated and accurate functional annotation protocols as they are developed (see Schnoes, Dodevski, and Babbitt 2009). Quality scores and provenance of functional annotations should be provided as fundamental attributes of the annotation. Because no automated functional annotation pipeline will ever be completely accurate, mechanisms are needed for revising and updating functional annotations as new data become available. Mechanisms for including manual curation of genes and genomes from biologists external to DOE JGI also should be implemented. Interagency cooperation between NIH, DOE, and other federal agencies may be required to ensure that GenBank annotations reflect the most accurate and current data on gene function.

Gene function encompasses a broad spectrum of meanings, including (but not limited to) biochemical reaction, protein-protein interaction, metabolic or signaling pathway association, cellular localization, phenotype, and changes in protein function that are mediated by shifts in protein structure. Therefore, functional annotation will require integrating heterogeneous and often noisy data from disparate sources. The bioinformatics methods and statistical modeling techniques applied to these data for predicting functional and structural features are equally varied, including hardware features and advanced software such as neural networks, hidden Markov models, and logistic regression models. A functional annotation system ideally would use the most accurate and advanced of these methods, insofar as they can be made scalable for DOE JGI genomes. These challenges likely will require increasing fractions of DOE JGI resources (both human expertise and computational infrastructure) over the next decades.

#### *Sidebar 5*

### **Functional Annotation: A Prerequisite for Predictive Biology**

The term “functional annotation” is meant to be interpreted broadly, encompassing predictions of biochemical function, metabolic or signaling pathway, biological process, cellular location, phenotype, interacting partners, protein 3D structure, orthology identification, and more. Functional annotation in this context is intended to be a hypothesis-generating engine, with the expectation that biologists will be able to design experiments to test predictions.

### **Metagenome and microbial community annotation**

As sequencing technologies continue to evolve, novel experimental and computational methods will be required to explore the implications of these data. These methods will be needed particularly for obtaining robust annotation results from multiple levels of information that are dynamic in response to variations from different sources and of different types. Challenges for DOE JGI will be in automating experimental protocols, data management, and functional and taxonomic annotation. Improved data interpretation methods are necessary to extract maximal information from these data, (e.g., to understand plant-microbe interactions and interpret changes in microbial communities in response to changes in the environment). The vast quantities of data—today in the tens of millions of reads in a typical microbial community dataset—in combination with the noisy and fragmentary nature of individual reads will exhaust the capacities of existing bioinformatics methods for functional and taxonomic annotation. DOE JGI should be active in helping to drive the development of improved methods for these tasks and serve as an early adopter of improved technologies.

### **Genome annotation: Experimental**

The value of DNA sequence data is greatly enhanced by functional annotation. In addition to RNA-Seq and its variants (e.g., digital gene expression and global polyadenylation mapping), a growing number of sequence-based technologies are being developed and used to decorate genomes with various kinds of functional information. Chromatin

## Next-Generation Enabling Capabilities

immunoprecipitation coupled with high-throughput sequencing (ChIP-Seq) provides profiles of eukaryotic epigenetic modifications, as well as protein-DNA binding sites that can help elucidate regulatory networks at the genome scale. Other high-throughput, sequence-based technologies such as interactome sequencing (i.e., genome-wide protein-protein interactions) and chromosome conformation add additional layers of information to the reference genome sequence. A new level of genome understanding will be possible by integrating all these types of sequence-based data with experimental data. These data types include quantitative protein abundance, metabolite abundance, and pathway characterizations, as well as heterogeneous metadata (e.g., experimental design, soil type, climatic conditions, and geographic information system parameters) and image data ranging from the whole organism to the subcellular level. Functional annotation of genes, metabolic profiles, and pathways relating to secondary metabolites and their modifications, such as polysaccharides and lignin, are particularly challenging but of great importance to DOE, and DOE JGI should address this need. DOE JGI should serve as a coordinating entity for studying energy-relevant genes, genomes, and pathways. Leadership particularly is needed in establishing common vocabularies and data structures to enable queries across species and communities and in developing quality metrics and standards for annotation and functional data.

### Addressing “Big Data” Opportunities and Challenges

Several of the opportunities and challenges confronting DOE and JGI will be in the area of Big Data, which is receiving increasing attention from multiple federal agencies (e.g., the joint National Science Foundation–NIH BIGDATA initiative). Technological advances in genomics and metagenomics, including novel sequencing technologies and high-throughput functional genomics and proteomics, will require the development of innovative strategies and tools. Computational methods that can handle vast quantities of heterogeneous and poorly structured data will be needed for data integration and mining.

Although DOE JGI should remain an important generator of sequence data pertinent to its mission, an increasing amount of relevant data will be generated elsewhere, both within and outside DOE laboratories and funding. DOE JGI will require

access to and analysis of vast datasets generated by a variety of entities worldwide. An increasing proportion of DOE JGI resources will need to be devoted to such computational activities in the future.

Amounts of new DNA sequence data alone soon will exceed an exabyte per year. The ability to handle such large amounts of data is beyond the capabilities of most institutions. DOE JGI has a unique opportunity to use the computational expertise and infrastructure of other DOE facilities, building on relationships already initiated [e.g., National Energy Research Scientific Computing Center (NERSC)]. Even if the configurations of current hardware are not designed for genomics, they should be adapted to exploit it. The DOE Systems Biology Knowledgebase (KBBase; see Appendix 2, p. 31) and DOE JGI bioinformatics efforts should be coordinated with efforts in other agencies such as the National Science Foundation’s iPlant. Big Data efforts also should build on expertise in handling large datasets that exist in other agencies.

Storing and curating the vast quantities of sequencing data pose rapidly increasing challenges. Akin to what the high-energy physics, astrophysics, and astronomy communities have had to face for years, hard decisions will have to be made regarding which (raw) data should be stored and which should be distilled into derived secondary data. DOE JGI should assume responsibility for a small number of reference genomes for mission-related organisms that are curated to the highest possible level; this should include storage of raw reads and associated metadata.

Major challenges in biology lie in the area of information integration. Data integration can be achieved programmatically (e.g., through the use of statistical modeling techniques) and via the development of intuitive graphical user interfaces enabling biologists to visualize and navigate complex relationships between data. Algorithmic approaches can and should be integrated with data visualization tools. DOE JGI should identify and include the best tools for these tasks in its genome and metagenome analysis pipelines. As new experimental technologies are developed, data produced by these technologies should be included in statistical models and visualization tools. The informatics infrastructure of DOE JGI needs to expand to model these data (see Fig. 5. Toward a More Systematic Annotation Pathway, p. 17).

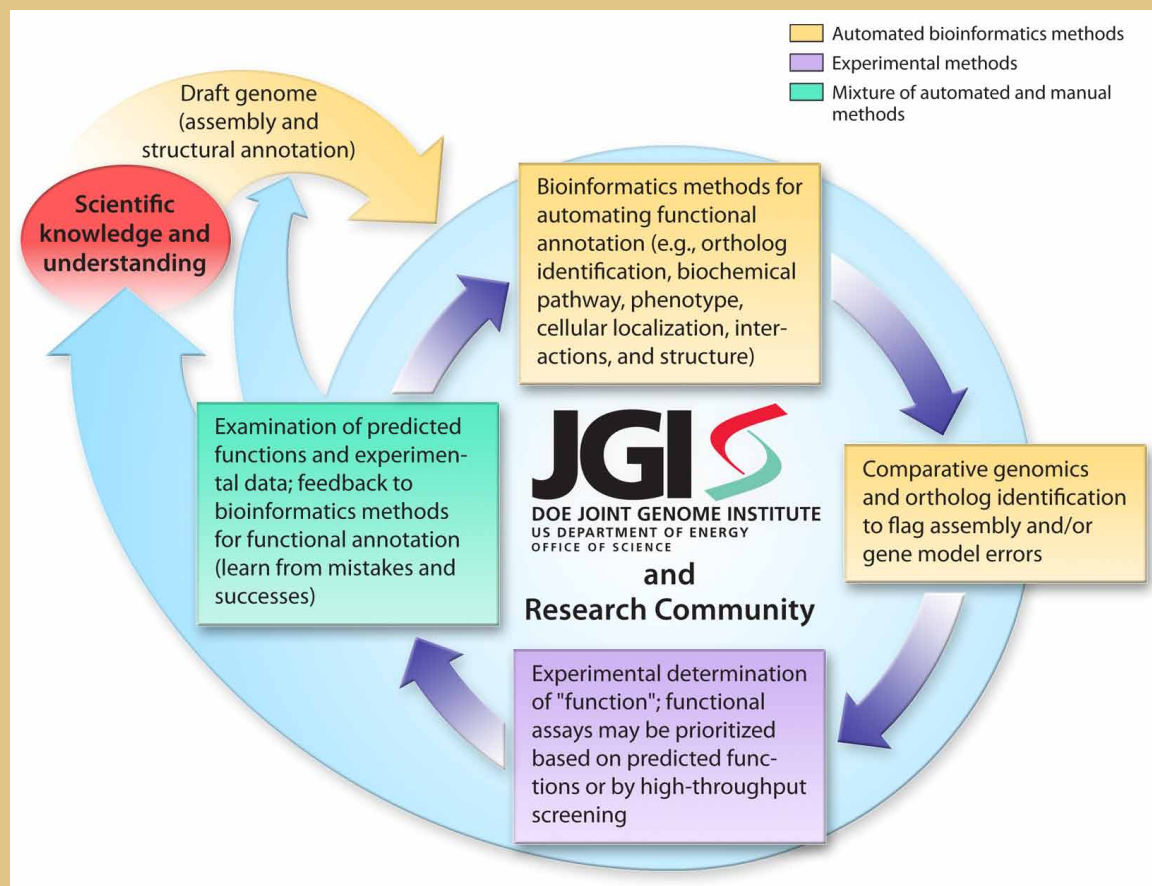
## Next-Generation Enabling Capabilities

DOE JGI should not attempt to solve these technological challenges in data storage, management, and integration independently but should be at the forefront in adopting effective strategies as they are developed by the scientific community. Some level of redundancy in efforts is desirable because optimal solutions are not yet clear. DOE JGI and KBase should coordinate their efforts with other members

of the greater scientific community, particularly iPlant, to meet these challenges. DOE JGI also should play a major role in bringing together the different stakeholders into collaborative working groups to address the specific opportunities in this area. With respect to Big Data in particular, working groups might include investigators from computer science, statistics, biology, and the users and developers of

**Fig. 5. Toward a More Systematic Annotation Pathway.** DOE JGI production of sequence data is the first step in an interactive and iterative process by which computational and experimental methods add value to DOE JGI data. Genome *structural* annotation—identifying genes, promoter sites, and other genome features—is the first annotation process. Bioinformatics methods for gene *functional* annotation come next. They can be used to predict enzymatic reactions, metabolic pathways, gene ontology biological processes, cellular localizations, protein-protein interactions, and protein structures. Many of these analyses can be fed back into earlier steps to refine and improve models. For instance, ortholog identification plays a large role in functional annotation but also can be used to improve gene model accuracy. Experimental investigation, using predicted functions as starting points, is required to prove or disprove any annotations assigned computationally. Computational tool developers need feedback from the results of experimental investigation to improve method accuracy (results showing inaccurate predictions are particularly important). Functional annotations derived computationally and experimentally must be accessible to the greater scientific community, and errors in existing annotations should be revised, along with the provenance and support for annotations. In summary, sequence data feed into a larger (and dynamic and evolving) process, performed by DOE JGI and the research community, to provide feedback for more informed sequencing, associated DNA

annotations, and more informed experiments, ultimately resulting in greater biological understanding. By implication, if the sequence data are not supplied, or are supplied in small amounts, the kinetics of this annotation cycle are slower and thus also the progress of biological understanding pertinent to DOE missions.



## Next-Generation Enabling Capabilities

technologies generating these massive datasets. These working groups should be charged with identifying current limitations of existing data integration tools and methodologies and recommending areas where improvements are needed.

### Writing DNA and Developing Associated Technologies

A major foundational challenge over the next decade will be the ability to synthesize, or “write,” DNA in a rapid, highly automated fashion. This challenge includes the capacity to synthesize and assemble custom-designed DNA on a range of scales, from kilobase-sized fragments to entire chromosomes. Further, it will demand new technologies for introducing synthesized DNA into the genomes of DOE-relevant microbes and plants and for the direct, efficient modification of genomes *in vivo* (i.e., genome engineering). These efforts in synthetic biology will be vital to the DOE missions of understanding gene functions (annotation) and how microbes and plants produce and consume energy and of engineering organisms for improved biofuel production, carbon storage, or remediation. Synthetic biology’s role in driving science and enabling biological engineering is further discussed below.

#### *DNA synthesis as a science driver*

The ability to create molecules and engineer organisms will ultimately produce a revolution in a wide range of DOE-relevant scientific studies. Many current endeavors to understand microbes and plants are fundamentally constrained by the time and effort needed to generate microbial strains or plants with a desired genotype, relying on time-consuming and laborious genetic approaches. For instance, efforts to understand how natural variation influences a trait of interest in plants will often identify multiple candidate alleles of interest. Determining which alleles contribute most to the trait and probing whether individual alleles interact (exhibit epistasis) are major bottlenecks that would be significantly alleviated by an ability to rapidly engineer or edit plant genomes. As another example, structure-function studies of individual proteins are often a crucial element of, or follow up to, functional annotation efforts, but these studies remain labor intensive, in part because of current low-throughput mutagenesis techniques. New technologies for DNA synthesis and the high-throughput generation of mutants promise to

facilitate such studies and enable the investigation of orders of magnitude more mutants. These are just two examples, but, as synthesis technologies improve, more studies will incorporate them—just as the use of DNA sequencing has expanded along with developments in sequencing technology. Finally, recent history clearly shows that the scope and scale of DNA synthesis needed for many future projects of DOE relevance will often demand the economies of scale found only in a large user facility, highlighting DNA synthesis as a key goal for DOE JGI.

#### *Designing molecules and organisms*

Synthetic biology also will be central to future efforts to design molecules with desired functions and organisms with desired capabilities or phenotypes. Modifications that improve plant feedstocks for biofuel production, the design of plants that fix nitrogen, or the efficiency of photosynthesis are long-term goals (see Appendix 1: Grand Challenges, p. 23). Each will require the construction of novel organisms harboring synthetic genes, gene clusters, or even whole chromosomes. Similarly, the microbial research community aims to engineer bacteria and fungi that can produce novel compounds and biofuels, establishing an urgent need for fast, cheap, and reliable DNA synthesis. All these efforts will require many rounds of design, synthesis, and testing; the scale of DNA synthesis and sequence verification needed likely will exceed the capacity of individual laboratories, creating a significant opportunity for DOE JGI. In addition to synthesizing long (kilobase to megabase) *de novo* DNA sequences, a need for, and rise in, genome editing capabilities is anticipated. As these technologies are improved and as alternative technologies take hold, DOE JGI can play a vital role in supporting crop improvement efforts and the modification of other organisms relevant to DOE missions.

#### *Practical considerations*

Although the ability to write DNA should be an important element of DOE JGI’s future, technologies for large-scale, automated DNA synthesis and genome engineering are still in their early stages. Despite some success in synthesizing megabase-sized DNA constructs and in engineering the genomes of bacteria and plants, synthetic biology still is a long way from being a mature discipline. However, similar to the rapid evolution of DNA sequencing technologies over the last decade, DNA synthesis and genome engineering technologies

are anticipated to grow quickly over the next decade. DOE JGI must play a central role in beta-testing these emerging technologies and identifying those that are cheap, scalable, and easily implemented in the context of a user facility. Initial technologies include high-throughput oligonucleotide synthesis, kilobase- and megabase-scale assembly, multiplexed automated genome engineering of *Escherichia coli*, and nuclease-based engineering of plant genomes. Many other approaches are under development as part of work funded by DOE and other agencies, and DOE JGI will need to opportunistically integrate the best and most promising technologies. For DNA synthesis, the priority should be on rapid *de novo* DNA synthesis rather than assembly from pre-existing parts, which often are organism specific, likely to become quickly obsolete, and expensive to archive. As with sequencing projects, DOE JGI partnerships with academic laboratories will be critical for genome engineering to leverage the organism-specific expertise necessary for success.

### Enabling High-Throughput Experimentation via Miniaturization and Automation

In the last decade, two important technologies have emerged with dramatic potential for advancing biological and genomics research. The first is the ability to create inexpensive, small, and reusable liquid-handling devices that can be used to implement miniaturized versions of many traditional laboratory procedures. These microfluidic devices (MFD)

can drastically lower the cost of high-throughput experiments and increase the degree of consistency that can be achieved. Microfluidics has emerged as a set of technologies that can decrease costs of massively parallel assays and screens by significantly reducing the working volumes of reagents and samples. With the addition of discrete droplet technologies, they enable precise manipulation of single cells and engineered microcosms, opening up a frontier of new experimental possibilities. MFDs routinely are used in large-scale screens in the pharmaceutical industry and increasingly are used in instruments such as DNA sequencers. This is a tremendous opportunity for DOE and BER to leverage MFDs to achieve breakthrough levels of parallelism in developing next-generation high-throughput function and phenotype assays for systems biology and to enable the novel ultrahigh-throughput screening needed for directed evolution and synthetic biology research.

The workshop attendees (see Appendix 3, p. 35) suggested that DOE JGI and DOE's Office of Biological and Environmental Research (BER) place a high priority on exploring the use of MFD technology in the near future. Multiple laboratories in the DOE system have capabilities in this area. This technology emphasis also provides a potential opportunity for linkage to DOE nanoscience centers, where advanced lithography and detectors are being developed that extend the capabilities of existing commercial MFD systems.

The second breakthrough in the last decade that could have a transformational impact on BER genomic science is the

#### Sidebar 6

### Automating Science in a Robotic Laboratory

A dramatic demonstration of scientific automation is the work of U.K. researcher Ross King, whose team built a robotic laboratory that automatically devised and carried out experiments to determine the function of unknown genes in *Saccharomyces cerevisiae*. In this case, the robot worked nonstop for days carrying out thousands of experiments to ultimately confirm 14 functional predictions. This robot used conventional laboratory robotics, methods, and instruments (modified only for automated manipulation).

Only high-level input was given to the system (similar to the kind of information that one would give a postdoctoral

researcher or that would come from an advanced bioinformatics tool). The planning and tracking of experiments; physical manipulation of samples, cultures, and instruments; and information management were entirely automated. Although this result was a landmark in scientific automation, it went much further than simply automating laboratory procedures. While currently limited by the use of conventional form factors (e.g., 96-well plate) and scaleup costs, the promise for more automated approaches in conducting biological experiments is evident (see King 2009; Sparkes and Clare 2012).

## Next-Generation Enabling Capabilities

development of robotics that can carry out tens or hundreds of thousands of scientific experiments in the pursuit of validating (or invalidating) high-level biological hypotheses (see Sidebar 6, Automating Science in a Robotic Laboratory, p. 19).

A significant opportunity exists for DOE, BER, and DOE JGI to combine the ideas and algorithms underlying the automation of hypothesis testing with the economies of scale possible with MFDs. The combined capability could dramatically change the face of fields such as functional genomics, directed evolution, and biological design. The expertise needed to pursue this combined goal largely exists in DOE laboratories and would require a long-term and close set of partnerships among computer scientists, engineers, biologists, biochemists, and genomics researchers. The acceleration of DOE-critical science would be unprecedented.

### Building Research Communities

Workshop attendees strongly endorse the strategic vision for user interactions that DOE JGI presents in its current strategic plan, but they provide additional recommendations that will enable DOE JGI to boldly “define the future” of biology, as opposed to simply improving on existing work (see Sidebar 1, p. vii).

There is a fundamental need for DOE JGI to continue expanding its user interactions to include strategic partnering with the community to meet DOE BER scientific grand challenges in genomics. The new model is for DOE JGI to become a research partner in knowledge development, an approach distinguishing DOE JGI from other sequencing facilities and genome centers in the future. The real challenge is how DOE and DOE JGI can work together with the scientific community to integrate such diverse information into knowledge. The development or adaptation of novel approaches to address experimental questions and the “hand-off” of data and results to users represent an outdated mode of operation that needs to transition to longer-term strategic partnerships. In addition to providing state-of-the-science capabilities in genomics and computing, DOE JGI also can serve an important role in coordinating multidisciplinary user groups (e.g., statisticians; experimental biologists; computer scientists; and experts in bioinformatics, databases,

and specific experimental sequencing technologies). The facility also can help integrate activities across various partner institutions (e.g., other national laboratories and sister user facilities such as the DOE Environmental Molecular Sciences Laboratory (EMSL), KBase, and synchrotron and neutron beam facilities for biology (see Appendix 2, p. 29). BER should consider developing a “constellation strategy” that would partner JGI, EMSL, and KBase to enable true grand challenges that span the breadth and depth of genomics; proteomics; and high-throughput phenomics for single microbes, plants, or communities of these organisms. In this role, DOE JGI and partner institutions would develop detailed plans for data collection and analyses prior to these activities to ensure the capture of essential metadata and statistically robust experimental designs. In this manner, DOE JGI can aid the scientific community in capturing and utilizing metadata, bringing awareness of best practices, generating data in a manner that ensures availability and usefulness to others, and employing commonly accepted data standards. One workshop recommendation was to focus efforts on a set of DOE-relevant model organisms that would be subject to deep genomic functional characterization, an area in which DOE JGI could serve a critical community coordinating role. Additional scientific or technical targets included (1) identifying certain gene families or sets of functions for targeted analysis versus pursuing specific organisms or communities of organisms; (2) down-selecting to particular DOE mission-relevant functions; (3) applying comparative analysis across multiple genomes to fill in gaps; and (4) manipulating experimental capability at the level of entire microbial communities to answer questions such as: “What are the mechanisms by which ecotypes and community diversity are generated?”

Community education and training are other critical services DOE JGI can provide. These would include expanding workshops, tutorials, and undergraduate and graduate education, with an emphasis on educating users about data analysis protocols. Workshop topics should include key emerging technologies and major scientific challenges. A primary need is enhancing current efforts in bioinformatics training for the research community and providing better, easier-to-use interfaces for bioinformatics tools. With its broad expertise, DOE JGI could serve as a center to train integrators of diverse datasets to enable construction of robust networks.



## Summary

The main themes emerging from this workshop centered on the needs for more sequencing, more effective annotations, improving data analyses, adding DNA synthesis capability, adapting automation technologies, and assisting community formation around critical biological problems (see Sidebar 1, p. vii). The ever-growing throughput of DNA sequencing technologies permits rapid sequencing of virtually any type and number of organisms. This capability means that continued investment in genomic sequencing and follow-on sequencing studies can be targeted at more difficult challenges such as those central to Department of Energy (DOE) missions in bioenergy and the environment. The enormous diversity of genome-determined capabilities in plants, microbes, fungi, algae, and the repertoire of enzymes and pathways that microbial communities and plant-microbe associations contain make such explorations important to continue. Every individual in a population has its own genome, and any phenotypic differences are ultimately determined by those genetic changes. Therefore, there will be a continuous need for strategically designed genomic sequencing experiments that can capture all those genetic differences and, coupled with other sequence-based technologies as well as other functional assays, can achieve a comprehensive picture of the relationship between genotype and phenotype.

A key imperative recognized at the workshop is to improve the capacity to extract rich and accurate biological information from genomic sequences, at scales ranging from

sequence assembly and functional annotation to insights derived from comparative genomics. Workshop participants emphasized the value and importance of additional tools to accelerate sequencing and sequence interpretation for DOE mission-relevant science. In particular, the scale of this research spotlights the value that can be realized from centralizing appropriate experimental and computational technologies in a high-throughput setting that takes advantage of associated efficiencies and economies of scale. In this way, DOE Joint Genome Institute (JGI) users can do what they do best: explore critical biological problems rather than spend time and resources manually performing tasks that are either amenable to automation or more efficiently performed at a facility such as DOE JGI.

Thus, a broad outcome of this workshop is the recognition of the critical need to accelerate the mission-relevant research best conducted in a user facility with high-throughput, automated, highly precise, and standardized methods of genome annotation and analyses. The elaboration of knowledge from sequencing efforts will increase as a result. Importantly, close interaction with computational modeling resources such as the DOE Systems Biology Knowledgebase (KBase) will allow further hypothesis generation and testing by the DOE JGI user community. KBase, in turn, will depend on high-quality genome sequence and annotation data generated by DOE JGI. Workshop participants saw these as appropriate and critically important directions toward which DOE JGI should move.



## Appendix 1: Grand Challenges

### Designer Phototrophs: Engineering Cyanobacteria to Produce Biofuels

Cyanobacteria offer considerable promise for the sustainable production of biofuels from light and carbon dioxide (CO<sub>2</sub>). Cyanobacteria thrive in a diverse range of challenging environments including oceans, hot springs, hypersaline waters, and desert soil crusts, making these organisms well adapted for growth on nonarable lands and in brackish waste waters. Cyanobacteria have been used to produce a range of biofuels such as hydrogen (H<sub>2</sub>) and biodiesel, as well as value-added coproducts including foods and food additives, fertilizers, dyes, and pigments. Currently, the major obstacle to industrial-scale exploitation of cyanobacteria for economically sustainable biofuel production is their low reaction rate and yield. Because cyanobacteria dynamically regulate their metabolism in a highly complex manner as they adapt to ambient environmental conditions such as light, salinity, and nutrient supply, the traditional metabolic engineering approaches for strain improvement have had limited success. Hence, a systems-level understanding of cyanobacterial metabolic subsystems and their regulation is a critical starting point for enabling high productivity of specific metabolic products at industrial scales.

The more than 170 cyanobacterial genomes sequenced to date have resulted in an extensive potential parts list for bioengineering. To exploit this metabolic diversity in the construction of new, efficient metabolic pathways in cyanobacteria, a much more comprehensive understanding of cyanobacterial gene function is needed. For example, the annotation of all genes to a reasonable level of accuracy in a genetically tractable cyanobacterium is a high priority. Another priority is the annotation of all cyanobacterial metabolic pathways, regulators, and gene networks responsible for important traits such as growth rates, CO<sub>2</sub> fixation efficiency, and resistance to oxidative stress. Further, this information must be integrated with various types of expression and metabolomics data necessary to generate predictive models that incorporate cellular metabolism and regulation. Such models can form the basis for redesigning cyanobacteria to maximize the allocation of carbon and reductant to biomass or products that can be converted to or

used directly as biofuels. The goal is to provide the scientific underpinnings for engineering designer photoautotrophs, enabling the mixing and matching of synthetic pathways and parts for the rational design of cyanobacteria that produce high yields of specific metabolic products at an industrial scale. A related, but longer-term goal would be to build a minimal photosynthetic microbial cell using synthetic biology tools with the ability to engineer for specific metabolic endproducts in a “plug-and-play” mode.

As difficult as predictive understanding may be for a single cyanobacterium, in nature cyanobacteria commonly coexist with a range of other microorganisms in associations that include highly evolved symbiotic relationships such as lichens and structured microbial communities referred to as mats. In these and less complex associations, the partners exchange metabolites and signaling molecules and collectively modify the local environment to provide favorable physical and chemical conditions and protection against stressors. Such associations have the potential to offer significant improvements over axenic (single-organism) cultures for industrial applications because they are far less susceptible to contamination and predation that can lead to significant loss in stability and yield. A goal pertaining to photosynthetic microbial consortia is to understand the interactions to a degree sufficient for engineering them to produce target fuel molecules at high yields in open-air environments. Tools and technologies similar to those described will be required but at a much larger scale.

### Understanding Interactions Between Microbes and Climate

In a report from the American Academy for Microbiology (2011), the urgent need to incorporate microbial processes into global climate models was articulated as follows: “*The most powerful impact of life on the Earth’s climate is made by its smallest inhabitant[s]—the microbes; Bacteria, Archaea, algae, fungi, and other microbes may be too small to see, but are far too important to ignore.*” Human activities increasingly influence the structure and function of microbial communities, whether through the application of fertilizers that has doubled the flux through the microbially driven global

**Appendix 1: Grand Challenges**

nitrogen cycle or the inadvertent discharge of oil or other industrial byproducts that fuel blooms of (or otherwise alter) microbial populations in soil or aquatic environments.

The expanding worldwide demand for food, fuel, fiber, and minerals will continue to impact microbial communities that drive the global cycling of elements, including the production and consumption of greenhouse gases. Fluxes of CO<sub>2</sub>, methane, and nitrous oxide—the three gases that constitute a majority of the warming potential of Earth’s atmosphere—are driven primarily by the metabolism of microbial communities. Although sequencing microbial communities offers the potential to link the taxonomic composition and metabolic potential of microbial communities with global cycles, serious challenges remain in extracting meaningful signals from massive datasets and then moving beyond data correlations to causal relationships.

The data combined from “omics” technologies and flux measurements from the environment have the potential to provide a framework for mathematical models of microbially mediated elemental cycles that predict how and how quickly climate may change. Ultimately, understanding the relationship between microbial communities and the flux of greenhouse gases offers an intriguing opportunity for potentially managing microbial communities to affect the composition of Earth’s atmosphere. This broad mission is central to the Genomic Science program within the Department of Energy’s (DOE) Office of Biological and Environmental Research (BER).

### **Large-Scale Studies of Rhizosphere and Soil Communities Important for Climate, Biofuel Sustainability**

Plants, microbes, and their interactions are significant biological drivers of ecosystem function and sentinels of changes in ecosystem equilibrium. The rhizosphere, as the dynamic interface between plant roots and its abiotic and biotic environments, is central to how plants sense and respond to the soil microbiome. Beneficial interactions with microbes can aid in such functions as plant nutrient and water uptake, growth and development, and pathogen resistance. However, foremost in the list of challenges facing plant microbiome studies is the culturability and identity of the entire gamut of microbes. Beneficial plant-microbe interactions are complex and wide ranging, and the outcomes reflect a vast number of variables along a large

spatiotemporal scale. Understanding the constitution and function of different rhizosphere systems thus presents a grand challenge in plant, microbial, and ecosystem research.

Plant-microbial studies are integral to BER’s core mission within DOE’s Office of Science. Combining nucleic acid sequencing with advanced computational tools is a powerful approach to shedding light on the composition and diversity of the plant microbiome, genetic and molecular bases of interspecies and interkingdom communication, and the outcomes of biological interactions between plants and microbes and their environment. The capabilities of the DOE Joint Genome Institute (JGI) offer an unparalleled opportunity to address this grand challenge.

Efforts in metagenomics, genome sequencing, comparative phylogenomics and phylogenetics, and single-cell genomics will facilitate understanding of rhizosphere composition and dynamics under controlled and natural settings. These approaches have been proposed for *Arabidopsis*, *Populus*, maize, and agave. The ambitious sequencing and other omics needs within these and other model and non-model plant research communities to fully understand the nature of genotype-phenotype interactions and their outcomes in the context of energy production, biogeochemistry, and climate change represent a scientific grand challenge that is as complex and daunting, but of critical importance, as any that DOE may face.

### **Advanced Genomic Capabilities for Biofuel Sustainability**

Alternative fuels from renewable cellulosic biomass—plant stalks, trunks, stems, and leaves—are expected to significantly reduce U.S. dependence on imported oil while enhancing national energy security and decreasing the environmental impacts of energy use. Ethanol and other advanced biofuels from cellulosic biomass are renewable alternatives that could increase domestic production of transportation fuels, revitalize rural economies, and reduce CO<sub>2</sub> and pollutant emissions. According to U.S. Secretary of Energy Steven Chu in the U.S. DOE 2009 overview of the Bioenergy Research Centers, “*Developing the next generation of biofuels is key to our effort to end our dependence on foreign oil and address the climate crisis while creating millions of new jobs that can’t be outsourced.*”

Although biofuels may help address the looming energy crisis, their current contribution to the U.S. energy portfolio is small. A key to successful biofuels production is developing sustainable cropping systems that include the communities of microbes central to soil health and function. Soils are among the most diverse habitats of life on Earth—there are as many as a million species of bacteria in a single gram of soil. Understanding this incredible diversity and how microbial communities shape the environment, interact with plants, and provide a collection of ecosystem services is essential for developing a sustainable agricultural ecosystem. Among the challenges are the relative inefficiency and high cost of currently available enzymatic, chemical, and physical treatments for the breakdown of cellulosic biomass and conversion of the resulting sugars to biofuel compounds. Others are the significant gaps in the fundamental understanding of enzymes and metabolic pathways of microorganisms mediating deconstruction of complex plant biomass and synthesis of ethanol or other potential biofuel compounds. High-throughput sequencing and other omics technologies are needed to develop and characterize the genes and proteins that determine these key capabilities required for sustainable and effective biofuels production.

### **Designer Walls: Improving Plant Cell Walls for Energy Conversion Processes**

“From the perspective of transportation fuels, plants can be viewed as solar energy collectors and thermochemical energy storage systems” (Rubin 2008). Plants efficiently capture light energy and store it in the form of chemical bonds in the major cell wall polymers: cellulose (sugars) and lignin (phenolic compounds). Cellulose and lignin polymers harbor different amounts of energy, and energy capture from these sources depends on the conversion process employed. For example, conversion efficiency for enzymatic processing would be improved by maximizing digestible sugars and minimizing lignin, while higher levels of lignin are desirable for thermochemical processing (i.e., pyrolysis, gasification, and bioelectric) due to its higher energy content. Ultimately, targeted modification of cellulose-lignin ratios will enable process-specific optimization of cell wall composition, thereby increasing the energy captured during conversion.

The plant cell wall determines the architecture and structural properties of the plant, from the strength of a tree trunk to the

flexibility of a blade of grass. The cell wall controls the entry and export of nutrients and products and can protect against pathogens and environmental stresses. It stores carbohydrates that can serve as potential energy sources in the form of biofuels. Lastly, in the form of biomass, the plant cell wall stores carbon acquired from the atmosphere through photosynthesis.

With advances in high-throughput sequencing, omics, and computational biology, it is now feasible to explore how the interacting networks of genes, proteins, and metabolites control cell wall composition and contribute to these attributes. Such a concerted network analysis would identify the gene pathways for plant cell wall polymer biosynthesis and how the pathways are regulated as well as the impacts of altered regulation on cell wall composition. Integration of this information would identify the “knobs and switches” needed to design and tune the flux via these pathways to create optimized energy feedstocks—currently a prime DOE mission.

### **Borrowing from Natural Variation to Improve Energy Capture (Photosynthesis) in Plants**

To be a sustainable biofuel source, plant biomass production must achieve high yields with minimal inputs (e.g., fertilizers). A plant’s photosynthetic capacity results from several factors, one of which is the maximum rate at which its leaves are able to fix carbon into energy-rich sugars. Because photosynthesis is a primary determinant of crop yield but has low overall efficiency, the several known forms of photosynthesis are attractive targets for improvements to plant biomass productivity. Improving photosynthetic capacity might seem eminently tractable, given that the underlying biochemical processes are well understood and essentially invariant across photosynthetic organisms, but numerous challenges remain. First, although the processes are similar, plants vary widely in their photosynthetic capacity. The prevailing wisdom is that natural variation in photosynthetic capacity results from a complex set of components, including species differences in the kinetics, regulation, and expression of key enzymes as well as plant responses to different and fluctuating environments (e.g., changes in light throughout the day). The complexity of causes makes improving photosynthesis a much more difficult task. Instead of identifying individual changes that might affect photosynthesis, one viable option is to exploit natural variation.

## Appendix 1: Grand Challenges

The improvements and reduced costs of high-throughput sequencing combined with new experimental and computational approaches offer novel solutions for using natural variation to improve plants for photosynthetic capacity and other complex traits relevant to biomass productivity. Scores of individual plant genomes that show wide variation in photosynthetic capacity can be sequenced and compared, using genome-wide association studies (GWAS) to enable discovery of the DNA sequence variations associated with photosynthetic variation. Once identified, the targeted genome regions can be introduced rapidly into desired genetic backgrounds using advanced genomics-based breeding techniques such as genotyping by sequencing (GBS), in which entire genetic mapping populations of hundreds to thousands of individual lines are sequenced and compared. GWAS and GBS are feasible for high-diversity, large-genome species, such as the targeted biomass crops *Miscanthus*, switchgrass, and poplar. In this way, comparative genomic analyses exploiting large collections of selected sequenced genomes can provide insights into the variations most relevant to critical characteristics of photosynthetic efficiency and can point the way to improvements of a variety of desirable plant properties.

### Dynamics of Genomic Participation in Marine CO<sub>2</sub> Cycling

The marine biosphere is responsible for half the global primary production of organic compounds and half the annual uptake of CO<sub>2</sub> from Earth's atmosphere. Marine environments have tremendous implications for DOE missions in climate science and bioenergy. Marine CO<sub>2</sub> uptake is regulated by a complex array of microbes (comprising ~98% of ocean biomass) whose dynamics, interactions, and overall controls are not well understood.

Challenges in this area stem from the taxonomic diversity of marine microbes (composed of bacteria, archaea, and eukaryotes) and the complexity of biogeochemical transformations integrally linked via different modes or “lifestyles.” In much of the world's oceans, turnover rates and processing within biogeochemical cycles are so rapid that fluxes cannot be measured and nutrients are below detection. For this reason, oceanographers, biologists, and climate modelers increasingly are moving to sequence-based and other omics approaches to understand marine microbes and their roles in the carbon cycle.

Many insights have been gained from sequence-based studies; however, the level to which these insights can be generalized, and thereby incorporated into global models, is limited. Principally, this results from the absence of sufficient sequencing data and reference genome information for interpreting microbial roles in environmental processes. However, the limited genome sequencing currently available in many lineages presents an important opportunity for DOE JGI. Few genomes are available for predatory marine microbes, and none are available for the most common microbial predators. There also are no genomes available for marine eukaryotic decomposers (fungi) or for the abundant dinoflagellates. While dinoflagellates can cause toxic blooms, both dinoflagellates and decomposers participate in central ecosystem services in CO<sub>2</sub> uptake and provide food resources to marine fisheries and other ecosystems. In sum, the genomic content of marine biota is largely uncharacterized.

Clearly, genomes representing a broader array of marine niches and lifestyles are badly needed, and sequencing efforts also should target uncultured taxa investigated using single-cell and population metagenomic approaches. In addition, seasonal cycles dictate carbon uptake, export, and decomposition and represent a critical element for understanding biological contributions to climate processes. Hence, to understand the mechanisms behind carbon cycling and organism interactions, sampling and sequencing of the oceans must be performed in the context of well-designed studies. To produce statistically informative data, studies must include appropriate biological sample replication and sequencing depth. The focus (in most studies to date) on abundant taxa leads to gaps in understanding how communities respond to perturbation—presumably many of the less abundant taxa are episodic “bloomers” and, in the case of photosynthetic taxa, these episodic blooms are known to sometimes result in major CO<sub>2</sub> drawdown and sequestration. As ocean conditions change, taxon profiles can change, often significantly, and this effect can impact ecosystem functioning in important ways.

Reinvention of how sequencing technology is applied to the oceans is a tremendous challenge but is critical for developing a true understanding of marine systems and carbon cycling. Such a reinvention would be the springboard to an era of “ecosystems biology,” in which investigators can develop network and interaction pathways at levels from the molecular to the ecosystem. Reinventing ocean genomics also will provide

dramatic advances in understanding viral ecology and how viruses differentially regulate distinct microbial populations, which can affect carbon export and sinking processes. Finally, virtually nothing is known about microbial processes in the “twilight zone” or deep ocean, where sequestered carbon can be remineralized, buried, or enter the food chain. Ocean warming is presumed to change ocean circulation and mixing, making knowledge about deep-ocean microbes essential to understanding the global carbon cycle.

Finally, advances in understanding marine CO<sub>2</sub> cycling and the technologies developed to address this grand challenge may shed light on terrestrial CO<sub>2</sub> cycling. High-throughput sequencing for terrestrial CO<sub>2</sub> cycling will be equally applicable and important.

## Characterizing Horizontal Gene Transfer

Horizontal gene transfer (HGT)—the nonvertical acquisition of genes—is relatively rare across eukaryotic species but is surprisingly common in microbial genomes. The frequency of genetic exchange between microbes became apparent from the sequencing of complete microbial genomes. For instance, in the seminal publication of the *Thermotoga maritima* genome in 1999 (Nelson 1999), evidence of the transfer of archaeal genes across domains was presented. Continued whole-genome

sequencing has documented that HGT is widespread in some taxonomic lineages, challenging previously accepted dogma of microbial diversity and ecology and fostering new debate on what constitutes a microbial species. In fact, the basic concept of a species is being re-examined due to these data. HGT complicates phylogenetic reconstruction, because, by definition, HGT introduces nontree-like relationships. Although seminal in first revealing the basic structure of microbial taxonomy, reliance on 16S RNA for phylogenetic inference has been replaced largely by the use of phylogenomic approaches (e.g., using gene matrix methods involving multiple genes), exploiting the data from whole-genome sequencing.

HGT frequency between microbes has practical implications that extend beyond the recasting of perspectives on microbial diversity and the species concept. Functional inferences based on homology are clearly complicated when genes are acquired by HGT. Other HGT implications include predicting bacterial metabolic capabilities from sequencing, monitoring community processes in response to environmental perturbations, and tracking the origins of a new isolate. Future HGT applications may include altering microbes or plants in defined ways with predictable outcomes. The use of sequencing as a monitor or readout of an experiment in microbiology will require a much more thorough understanding of HGT.

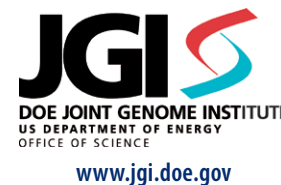




## Appendix 2: Department of Energy Assets

### DOE Joint Genome Institute

*Sequencing the world of possibilities for energy and the environment*



The U.S. Department of Energy (DOE) Joint Genome Institute (JGI) is the only federally funded high-throughput genome sequencing and analysis facility dedicated to genomes of nonmedical microbes, microbial communities, plants, fungi, and other targets relevant to DOE missions in energy, climate, and environment. DOE JGI provides collaborators around the world with access to massive-scale DNA sequencing to underpin modern systems biology research and provide fundamental data on key genes that may link to biological functions, including microbial metabolic pathways and enzymes that are used to generate fuel molecules, affect plant biomass formation, degrade contaminants, or capture carbon dioxide (CO<sub>2</sub>). The information can then be used to optimize organisms for biofuels production and other DOE missions.

Located in Walnut Creek, California, and supported by the Office of Biological and Environmental Research (BER) within the DOE Office of Science, DOE JGI is managed by Lawrence Berkeley National Laboratory, drawing additional complementary capabilities from its partner laboratories: Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, Pacific Northwest National Laboratory, and the HudsonAlpha Institute for Biotechnology.

Among DOE JGI's largest customers are the DOE Bioenergy Research Centers, which were established to accelerate basic research in the development of next-generation cellulosic and other biofuels through focused efforts on biomass improvement, biomass degradation, and strategies for fuels production.

#### Bioenergy

These sequencing projects focus on developing plants that can be used as feedstocks for biofuel production, identifying organisms (e.g., fungi and microbes) with enzymes and

#### JGI Facts

- Sequence production
  - Fiscal year (FY) 2011: 29.9 terabases
  - FY 2012: 55 terabases
- 1,106 users worldwide—individual principal investigators, collaborators, and annotators (who conduct genome analysis) on active projects; 682 in the United States
- >180 JGI-authored publications per year (>20 in top-tiered journals)
  - Science highlights ([jginews.blogspot.com/search/label/Science%20Highlights](http://jginews.blogspot.com/search/label/Science%20Highlights))
  - Notable scientific publications ([www.jgi.doe.gov/News/pubs.html](http://www.jgi.doe.gov/News/pubs.html))
- Annual progress reports ([www.jgi.doe.gov/whoware/progress.html](http://www.jgi.doe.gov/whoware/progress.html))

pathways that can break down the lignin and cellulose in plant cell walls, and characterizing enzymes and pathways that can ferment sugars into biofuels.

#### Carbon Cycle

Because microbes make up the largest component of Earth's biodiversity, understanding how they metabolize carbon and how environmental changes affect these processes is crucial for the development of better predictive models for reducing the effects of increasing CO<sub>2</sub> emissions on the global climate.

#### Biogeochemistry

The field of biogeochemistry explores the full spectrum of biological, physical, geological, and chemical processes and reactions involved in sustaining life on Earth. One area of emphasis targets microbes and microbial communities (or metagenomes) that can degrade or otherwise transform environmental contaminants such as toxic chemicals or heavy metals.

## DOE Environmental Molecular Sciences Laboratory



The Environmental Molecular Sciences Laboratory (EMSL), a U.S. Department of Energy (DOE) scientific user facility located at Pacific Northwest National Laboratory, offers integrated experimental and computational resources for discovery and technological innovation in the environmental molecular sciences to support the needs of DOE and the nation. DOE EMSL provides a suite of next-generation tools for studying cells from complex populations or environmental samples, using multiple and combined approaches.

### Microscopy Capabilities

- Fluorescence microscopy: Super-resolution, single-molecule, and multiphoton
- Electron microscopy: Scanning, transmission, cryogenic, and high-resolution
- Specialized microscopy: Biomolecular imaging, mass, and helium ion
- Influx flow cytometer cell sorting and laser-capture microdissection

These tools support fluorescence imaging of cellular structures and protein complexes, real-time studies of individual protein dynamics in live cells, isolation of organelles and sub-cellular structures as well as cells from tissues or mixed-cell populations, three-dimensional (3D) imaging of living tissues and cells, quantitative investigation of molecular interaction dynamics in living cells, high spatial resolution (0.35 nm beam size) of cell surfaces, high depth-of-field imaging of cell populations or communities, surface and serial section analysis of biofilms or individual cells, and 3D reconstruction of whole cells or large macromolecular complexes.

### Proteomics Capabilities

- High-throughput mass spectrometry (MS): Multiple Orbitrap advanced MS systems
- Specialized MS: 15 Tesla high-field Fourier transform ion cyclotron resonance MS, metallomics MS, and ion-mobility spectrometry-MS proteomics

These tools allow users to apply very high mass resolving power and mass accuracy to intact proteins, identify

### EMSL Facts

- 750 users worldwide
- >280 EMSL-authored publications per year (>130 in top-tiered journals)
  - Science highlights ([www.emsl.pnl.gov/news/archive/](http://www.emsl.pnl.gov/news/archive/))
  - Notable scientific publications ([www.emsl.pnl.gov/root/publications/journals/pubs\\_by\\_year.jsp?year=2011&alpha=A](http://www.emsl.pnl.gov/root/publications/journals/pubs_by_year.jsp?year=2011&alpha=A))
- *The Molecular Bond* quarterly ([www.emsl.pnl.gov/news/newsletter/](http://www.emsl.pnl.gov/news/newsletter/))

metabolites and peptides from complex mixtures, analyze post-translationally modified peptides and metabolites, conduct quantitative proteomics measurements, and study metal interactions and transformations in biological and environmental systems.

### Transcriptomics and Metabolomics Capabilities

- RNA-Seq for transcriptional profiling
- Combined MS/gas chromatography (GC)/nuclear magnetic resonance (NMR) metabolomics system

The transcriptomics capability supports quantitative profiling of gene-expression patterns in prokaryotic and eukaryotic cells. The MS/GC/NMR capability supports global and targeted profiling of important metabolic pathways, fatty acids, and volatile molecules.

### NMR and Electron Paramagnetic Resonance (EPR) Capabilities

- Catalysis/solids 850-MHz wide-bore NMR system
- High-field EPR (95 GHz) system

Solid-state NMR applicable to bio-solids/surface interactions and high field and power EPR to study integer spin metal centers.

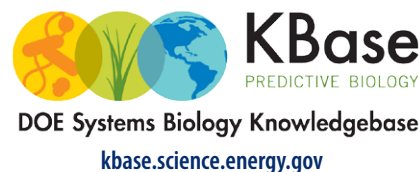
### Bioreactor and Analyzer Capabilities

- Bioscreen-C and Micro-24 bioreactors

These tools support automated, real-time analysis of growth rates of bacteria (including phototrophs) and growth under different conditions (e.g., oxygen levels, pH, and temperature).

## DOE Systems Biology Knowledgebase

### *A community resource*



**T**he DOE Systems Biology Knowledgebase (KBase) is a collaborative effort designed to accelerate scientific inquiries into microbes, plants, and microbial communities toward the ultimate goal of predictive biological understanding. Integrating commonly used tools and their associated data, KBase builds new capabilities and is composed of core biological analysis and modeling functions, including tools to connect different software programs within the community.

The primary KBase missions are to: (1) Create a flexible, extensible, and high-quality framework for experimental, evidence-based functional annotation of genome sequences. A central goal for high-quality genome assessments—and one means of validating genome annotations—is to develop the capability for using sequence to predict overall organismal function and community structure-function relationships in diverse environments (i.e., phenotypes). (2) Enable the scalable and mostly automated creation of high-quality metabolic and regulatory models for validation against experimental data and generation of scientific predictions and testable hypotheses, relying on the KBase backbone of software that facilitates the query, comparison, and visualization of gene function, genome and metagenome organization, and metabolic and regulatory models in a phylogenetic context. (3) Make all algorithms, software, and data free and openly accessible to the community.

These three missions underlie the overall goal of predictive biology with a primary emphasis on improving scientific estimates of genome function. Within these missions lie KBase's scientific drivers in microbes, plants, and communities. KBase will leverage existing integrated bioinformatics systems as core services, enabling research teams to focus on building new capabilities rather than duplicating existing ones. By providing and supporting a diverse set of advanced algorithms for comparative functional modeling, KBase seeks to change the way biologists routinely work with their data and thus allow researchers to use modeling in a far more sophisticated experimental design mode than is currently possible.



### **Microbes**

KBase will maximize understanding of microbial system function, promote sharing of data and findings, and vastly improve

the planning of effective experiments. Early efforts will focus on reconciling metabolic models with experimental data. The ultimate objective is to manipulate microbial function for applications in energy production and remediation by enabling users to expand on a strong foundation of quality genome annotations, reconstruct metabolism and regulation, integrate and standardize “omics” data, and construct genome models.



### **Plants**

A high priority is to link plant genetic variation, phenotypes, molecular profiles, and molecular networks, enabling model-driven phenotype predictions. A second goal is to map plant variability onto metabolic models to create model-driven predictions of phenotypic traits. Initial work will focus on creating a workflow for rapidly converting sequencing reads into genotypes. Tools also will be developed for data exploration and the linking of gene targets from phenotype studies (e.g., genome-wide association studies) with co-expression, protein-protein interaction, and regulatory network models. Users can narrow candidate gene lists by refining targets, visualize subnetworks of regulatory and physical interactions among genes responsible for a phenotype in question, and highlight networks or pathways impacted by genetic variation.



### **Communities**

Comparative analysis of metagenomes acquired over different spatial, temporal, or experimental scales now enables defining how communities respond to and change their environment. KBase will provide the computational infrastructure to study community behavior and build predictive models of community roles in biogeochemical cycles, bioremediation, energy production, and the discovery of useful enzymes. A next-generation metagenomic platform is being developed to provide scalable, flexible analyses; data vectors for models; tools for model creation; data quality control; application programming interfaces; and data and data collection standards compliant with the Genomic Standards Consortium. Initial efforts will target the development of bioprospecting and experimental design tools.

## DOE Synchrotron and Neutron Beam Facilities

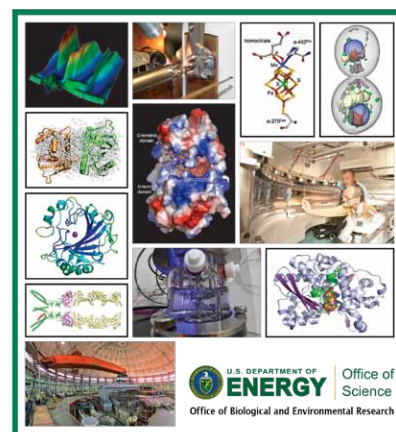
### *Accelerating biological research*

[genomicscience.energy.gov/userfacilities/structuralbio.shtml](http://genomicscience.energy.gov/userfacilities/structuralbio.shtml)

**S**ynchrotron light sources and neutron facilities at the Department of Energy's (DOE) national laboratories enable understanding of the structure of matter down to the atomic or molecular level using approaches not possible with laboratory instrumentation. Synchrotron facilities produce intense beams of photons, from X-rays to infrared to terahertz radiation, while neutron facilities produce beams using particle accelerators or reactors. The beams are directed into experimental stations housing instruments configured for specific biological investigations.

This infrastructure provides user access to beamlines and instrumentation for high-resolution studies of biological organisms and molecules for all areas of research in the life sciences. Users are chosen through a peer-reviewed proposal process managed by each facility.

This activity is supported by DOE's Office of Biological and Environmental Research within the Office of Science and is closely coordinated with other federal agencies and private organizations.



### Structural Biology Experimental Stations

Structural Biology Center at the Advanced Photon Source

*Argonne National Laboratory*

Macromolecular Crystallography Research Resource at the National Synchrotron Light Source

*Brookhaven National Laboratory*

Protein Crystallography Station at the Los Alamos Neutron Science Center

*Los Alamos National Laboratory*

Structural Molecular Biology Center at the Stanford Synchrotron Radiation Lightsource

*SLAC National Accelerator Laboratory*

Structurally Integrated Biology for the Life Sciences Beamline at the Advanced Light Source

*Lawrence Berkeley National Laboratory*

Center for Structural Molecular Biology at the High-Flux Isotope Reactor and Spallation Neutron Source

*Oak Ridge National Laboratory*

Advanced Biological and Environmental X-Ray Spectroscopy at the Advanced Light Source

*Lawrence Berkeley National Laboratory*

Berkeley Synchrotron Infrared Structural Biology Program at the Advanced Light Source

*Lawrence Berkeley National Laboratory*

National Center for X-Ray Tomography at the Advanced Light Source

*Lawrence Berkeley National Laboratory*

## Appendix 3: Workshop Agenda, Charge Questions, Participants

Department of Energy Joint Genome Institute  
Strategic Planning for the Genomic Sciences

### Workshop Agenda

Renaissance D.C. Marriott, 999 Ninth Street, NW, Washington, D.C.

#### Wednesday, May 30, 2012

7:30 a.m. – 8:15 a.m.	<b>Breakfast</b>
8:15 a.m. – 9:00 a.m.	Welcome from BSSD division director: Todd Anderson Welcome, workshop purpose: Daniel Drell and co-chairs
9:00 a.m. – 9:15 a.m.	Around the table introductions
<b>Plenary Session</b>	Two short introductory talks will focus on visionary perspectives on next-generation sequencing and exploitation of sequence data for DOE mission-relevant biology
9:15 a.m. – 9:45 a.m.	<b>Plenary talk:</b> Greg Petsko (Brandeis University)
9:45 a.m. – 10:15 a.m.	<b>Plenary talk:</b> Maureen McCann (Purdue University)
10:15 a.m. – 10:30 a.m.	<b>Break and refreshments</b>
10:30 a.m. – 10:35 a.m.	<b>Charge to Breakout Sessions</b>
10:35 a.m. – 12:30 p.m.	<b>Breakout Sessions I</b> (address first charge question)
12:30 p.m. – 1:30 p.m.	<b>Working lunch</b> (brief reports from Breakout Sessions I)
1:30 p.m. – 3:45 p.m.	<b>Breakout Sessions II</b> (address second charge question)
3:45 p.m. – 4:15 p.m.	<b>Break and coffee</b>
4:15 p.m. – 5:00 p.m.	Reports from Breakout Sessions II
5:00 p.m. – 5:30 p.m.	General discussion and wrapup of day 1: (Are we asking the right questions? Are we getting good ideas?)
5:30 p.m.	Adjourn for the day
6:00 p.m.	<b>Dinner</b> (chairs, breakout leaders, DOE staff; prepare for Thursday)

#### Thursday, May 31, 2012

7:00 a.m. – 8:00 a.m.	<b>Breakfast</b>
8:00 a.m. – 8:15 a.m.	Day 2 logistics, charge to attendees: Drell
8:15 a.m. – 8:45 a.m.	<b>Plenary talk:</b> Claire Fraser-Liggett (University of Maryland)
8:45 a.m. – 9:15 a.m.	<b>Plenary talk:</b> John Gerlt (University of Illinois)
9:15 a.m. – 9:30 a.m.	<b>Break</b>
9:30 a.m. – 11:45 a.m.	<b>Breakout Sessions III</b> (address third charge question)
11:45 a.m. – 12:45 p.m.	<b>Working Lunch</b>
12:45 p.m. – 1:15 p.m.	Reports from Breakout Sessions III
1:15 p.m. – 2:00 p.m.	General discussion
2:00 p.m. – 2:25 p.m.	Summary and wrapup of discussions
2:25 p.m.	Eddy Rubin (DOE JGI)
2:30 p.m.	Closing remarks: Drell
2:35 p.m.	Participants adjourn
3:00 p.m. – 3:30 p.m.	Workshop co-chairs, breakout session leaders, DOE BER staff meet to discuss writing assignments
3:30 p.m. – 5:30 p.m.	Writing session begins (co-chairs, DOE BER staff)
	<b>Dinner on your own</b>

#### Friday, June 1, 2012

7:30 a.m. – 8:30 a.m.	<b>Breakfast</b>
8:30 a.m. – 12:00 p.m.	Writing session (co-chairs, breakout chairs, DOE BER staff)

## Charge Questions

1. What new scientific insights could be enabled by next-generation sequencing? What vision for DOE mission-driven biology, rooted in and building on high-throughput genomic sequencing and analysis, can be identified for the next 5 to 10 years?
  - A. What can be done with (what is the utility of) sequences, from whole genomes to metagenomic fragments?
  - B. How can sequencing and subsequent “omics” advance science?
  - C. What are the current limitations to extracting biological information from sequence data?
  - D. What challenges need to be addressed in the analysis of increasingly complex community metagenomes?
2. What large-scale questions/grand challenges in systems biology, grounded in very high-throughput genomics and post-genomic analyses, will require a user facility to achieve necessary efficiencies and effectiveness and would have the highest impact and value for DOE biology?
  - A. What sequence-based high-throughput approaches would be relevant for DOE JGI (e.g., epigenomics, interactomes, chromatin immunoprecipitation, small RNAs, functional screens, and genome-wide association mapping)?
  - B. How can genomics technologies advance understanding of community interactions between organisms and their environment(s)?
  - C. How can genomics inform predictions/computational models of system and ecosystem response to perturbation?
3. What capabilities and technologies that do not presently exist, presently lack high-throughput capability, or are not generally available to the biological research community, will be required to address the most important questions in biology and to meet the needs of DOE biological science?
  - A. What is needed to more effectively facilitate other post-sequencing forms of omics analysis and subsequent cycles of modeling and experimentation?
  - B. What unique challenges are associated with analysis of eukaryotic (i.e., plants, algae, and fungi) genomes, such as repeat content, polymorphism, polyploidy, and guanine-cytosine bias?
  - C. What is needed to better integrate metadata into the analysis of complex genomic data?
  - D. How can the biological insights obtained from genomic analysis be used to more effectively inform rational design for systems biology experiments or bioengineering applications?
  - E. What are the challenges to annotate and interpret large eukaryotic genomes and metagenomes?

## Workshop Participants and Observers

### Workshop co-chairs

**Jim Fredrickson**  
Pacific Northwest National Laboratory

**Michael Laub**  
Massachusetts Institute of Technology

**Jan Leach**  
Colorado State University

### Participants

**Jill Banfield**  
University of California, Berkeley

**Andrew Bradbury**  
Los Alamos National Laboratory

**Donald Bryant**  
Pennsylvania State University

**Jeffrey Chen**  
University of Texas, Austin

**Paramvir Dehal**  
Lawrence Berkeley National Laboratory

**Elizabeth Edwards**  
University of Toronto

**Claire Fraser-Ligett**  
University of Maryland

**Audrey Gasch**  
University of Wisconsin, Madison

**John Gerlt**  
University of Illinois

**Ryan Gill**  
University of Colorado, Boulder

**Arthur Grossman**  
Stanford University

**Paula Imbro**  
Sandia National Laboratories

**Shawn Kaeppler**  
University of Wisconsin

**Udaya Kalluri**  
Oak Ridge National Laboratory

**Elizabeth Kellogg**  
University of Missouri, St. Louis

**Roy Kishony**  
Harvard University

**Felice Lightstone**  
Lawrence Livermore National Laboratory

**Reinhold Mann**  
Brookhaven National Laboratory

**Maureen McCann**  
Purdue University

**Richard Michelmore**  
University of California, Davis

**James Minor**  
DuPont (retired)

**Debra Mohnen**  
University of Georgia

**Mary Ann Moran**  
University of Georgia

**Thomas Schmidt**  
Michigan State University

**Zach Serber**  
Amyris, Inc.

**Blake Simmons**  
Sandia National Laboratories

**Kimmen Sjölander**  
University of California, Berkeley

**Gary Stacey**  
University of Missouri, Columbia

**Rick Stevens**  
Argonne National Laboratory

**Kathleen Treseder**  
University of California, Irvine

**Doreen Ware**  
Cold Spring Harbor Laboratory

### Observers

**Kevin Anderson**  
U.S. Department of Homeland Security

**Todd Anderson**  
U.S. Department of Energy

**Paul Bayer**  
U.S. Department of Energy

**Dean Cole**  
U.S. Department of Energy

**Daniel Drell**  
U.S. Department of Energy

**Adam Felsenfeld**  
National Institutes of Health

**Joseph Graber**  
U.S. Department of Energy

**Susan Gregurick**  
U.S. Department of Energy

**Roland Hirsch**  
U.S. Department of Energy

**John Houghton**  
U.S. Department of Energy

**Arthur Katz**  
U.S. Department of Energy

**Noelle Metting**  
U.S. Department of Energy

**Pablo Rabinowicz**  
U.S. Department of Energy

**Cathy Ronning**  
U.S. Department of Energy

**Prem Srivastava**  
U.S. Department of Energy

**David Thomassen**  
U.S. Department of Energy

**Sharlene Weatherwax**  
U.S. Department of Energy

Karen Nelson (J. Craig Venter Institute) and Alexandra Worden (Monterey Bay Aquarium Research Institute) also contributed to this report.

Report preparation: Biological and Environmental Research Information System group at Oak Ridge National Laboratory (Kris Christen, Holly Haun, Brett Hopwood, Betty Mansfield, Sheryl Martin, Marissa Mills, and Judy Wyrick)





## Appendix 4: Bibliography

- A New Biology for the 21st Century*. 2009. Committee on a New Biology for the 21st Century: Ensuring the United States Leads the Coming Biology Revolution; National Research Council, Washington, D.C. [www.nap.edu/catalog.php?record\\_id=12764](http://www.nap.edu/catalog.php?record_id=12764).
- A 10-Year Strategic Vision: Forging the Future of the U.S. Department of Energy Joint Genome Institute*. 2012. U.S. Department of Energy Office of Science Joint Genome Institute. [www.jgi.doe.gov/whoweare/10-Year-JGI-Strategic-Vision.pdf](http://www.jgi.doe.gov/whoweare/10-Year-JGI-Strategic-Vision.pdf).
- Eastwood, D. C., et al. 2011. “The Plant Cell Wall–Decomposing Machinery Underlies the Functional Diversity of Forest Fungi,” *Science* **333**, 762–65.
- GenBank. [www.ncbi.nlm.nih.gov/genbank](http://www.ncbi.nlm.nih.gov/genbank).
- Genomics:GTL Program 2008 Strategic Plan*. 2008. U.S. Department of Energy Office of Science Biological and Environmental Research Program. [genomicscience.energy.gov/strategicplan/](http://genomicscience.energy.gov/strategicplan/).
- Genomics:GTL Roadmap: Systems Biology for Energy and Environment*. 2005. U.S. Department of Energy Office of Science Biological and Environmental Research Program. [genomicscience.energy.gov/roadmap/](http://genomicscience.energy.gov/roadmap/).
- Grand Challenges for Biological and Environmental Research: A Long-Term Vision*. 2010. BER Advisory Committee Report. [genomicscience.energy.gov/program/beractv.shtml](http://genomicscience.energy.gov/program/beractv.shtml).
- Incorporating Microbial Processes into Climate Models*. 2011. *American Academy of Microbiology*. [academy.asm.org/index.php/colloquium-program/browse-all-reports/396-incorporating-microbial-processes-into-climate-models](http://academy.asm.org/index.php/colloquium-program/browse-all-reports/396-incorporating-microbial-processes-into-climate-models).
- King, Ross D., et al. 2009. “The Automation of Science,” *Science* **324**(5923), 85–89.
- Mackelprang, R., et al. 2011. “Metagenomic Analysis of a Permafrost Microbial Community Reveals a Rapid Response to Thaw,” *Nature* **480**, 368–71.
- NIH Protein Structure Initiative. [www.nigms.nih.gov/Research/FeaturedPrograms/PSI/](http://www.nigms.nih.gov/Research/FeaturedPrograms/PSI/).
- NSF-NIH BIGDATA Initiative. [www.nsf.gov/news/news\\_summ.jsp?cntn\\_id=123607](http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607).
- Nekrutenko, A., and J. Taylor. 2012. “Next-Generation Sequencing Data Interpretation: Enhancing Reproducibility and Accessibility,” *Nature Reviews Genetics* **13**, 667–72.
- Nelson, K. E., et al. 1999. “Evidence for Lateral Gene Transfer Between Archaea and Bacteria from Genome Sequence of *Thermotoga maritima*,” *Nature* **399**, 323–29. DOI: 10.1038/20601.
- Rubin, E. M. 2008. “Genomics of Cellulosic Biofuels,” *Nature* **454**(84), 20.
- Schnoes, A. M., et al. 2009. “Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies,” *PLoS Computational Biology* **5**(12), e1000605. DOI:10.1371/journal.pcbi.1000605.
- Sparkes, A., and A. Clare. 2012. “AutoLabDB: A Substantial Open Source Database Schema to Support a High-Throughput Automated Laboratory,” *Bioinformatics* **28**(10), 1390–97.
- The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. 2007. Committee on Metagenomics: Challenges and Functional Applications, National Research Council, Washington, D.C. [www.nap.edu/catalog.php?record\\_id=11902](http://www.nap.edu/catalog.php?record_id=11902).
- Tuskan, G. A., et al. 2006. “The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray),” *Science* **313**, 1596–1604.
- U.S. Department of Energy’s Bioenergy Research Centers: An Overview of the Science*. 2009. U.S. Department of Energy Office of Science Biological and Environmental Research Program. [genomicscience.energy.gov/biofuels/](http://genomicscience.energy.gov/biofuels/).
- Warnecke, F., et al. 2007. “Metagenomic and Functional Analysis of Hindgut Microbiota of a Wood-Feeding Higher Termite,” *Nature* **450**, 560–65.
- Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. “Prokaryotes: The Unseen Majority,” *Proceedings of the National Academy of Sciences U.S.A.* **95**, 6578–83.



## Appendix 5: Glossary

**abiotic:** Non-living chemical and physical factors (e.g., soil, water, air, temperature, and sunlight) in the environment that affect ecosystems.

**algae:** Photosynthetic, aquatic, eukaryotic organisms that contain chlorophyll but lack terrestrial plant structures (e.g., roots, stems, and leaves). Algae can exist in many sizes ranging from single cells to giant kelps several feet long.

**allele:** One of two or more forms of a gene or a genetic region (generally containing a group of genes). A population or species of organisms typically includes multiple alleles at each locus distributed among various individuals; except very rarely, each individual can have only two alleles at a given locus. Allelic variation at a locus is measurable as the number of alleles (polymorphism) present, or the proportion of heterozygotes in the population.

**archaea:** Single-celled prokaryotic microbes that are structurally and metabolically similar to bacteria but share some features of their molecular biology with eukaryotes. The archaea are a distinct branch of life from the Bacteria and Eukarya.

**axenic:** Culture of an organism that is entirely free of all other “contaminating” organisms. Axenic culture is an important tool for the study of symbiotic and parasitic organisms in a controlled manner.

**bacteria:** Single-celled prokaryote, typically without a discrete, membrane-bound nucleus.

**bioinformatics:** Science of managing and analyzing biological data using advanced computing techniques.

**biomass (cellulosic):** Plant stalks, trunks, stems, and leaves.

**biotic:** Any living component that affects another organism. Biotic components include plants, animals, fungi, and bacteria.

**carbon dioxide (CO<sub>2</sub>):** Gas that is an important part of the global carbon cycle. CO<sub>2</sub> is emitted from a variety of processes (e.g., cellular respiration, biomass decomposition, and fossil fuel use) and taken up primarily by the photosynthesis of plants and microorganisms where it can become part of the plant’s or microbe’s biomass. CO<sub>2</sub> is a greenhouse gas that absorbs infrared radiation and traps heat in Earth’s atmosphere.

**cellulose:** Linear polysaccharide polymer with many glucose monosaccharide units. Cellulose is the major component of plant cell walls and the most abundant biological material on earth.

**chromatin immunoprecipitation (ChIP):** Method used to determine the location in a genome of DNA binding sites recognized by a particular protein of interest.

**complementary RNA (cRNA):** Synthetic transcripts of a specific DNA molecule or fragment made by an *in vitro* transcription system.

**computational biology:** Development and application of data-analytical and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological systems.

**conserved hypothetical proteins:** The (often large) fraction of genes in sequenced genomes encoding proteins that are found in organisms from several phylogenetic lineages but have not been functionally characterized and described at the protein chemical level. These structures may represent up to half of the potential protein coding regions of a genome.

**cyanobacteria:** Division of photosynthetic bacteria found in many environments, including oceans, fresh water, and soils. Cyanobacteria contain chlorophyll a and other photosynthetic pigments in an intracellular system of membranes called thylakoids. Many cyanobacterial species also are capable of nitrogen fixation.

**DNA (deoxyribonucleic acid):** Molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases adenine (A), guanine (G), cytosine (C), and thymine (T). A pairs with T and C pairs with G.

**DNA annotation:** See *genome annotation*.

**DNA assembly:** See *genome assembly*.

**DNA sequence:** See *genome sequence*.

**epigenome:** Set of chemical compounds that modify, or mark, the genome in a way that tells it what to do, where to do it, and when to do it. The marks, which are not part of the DNA itself, can be passed on from cell to cell as cells divide, and from one generation to the next.

**epigenomics:** Study of the complete set of epigenetic modifications on the genetic material of a cell, known as the epigenome. Epigenetic modifications are reversible modifications on a cell’s DNA or histones that affect gene expression without altering the DNA sequence.

**epistasis:** Phenomenon where the effects of one gene are modified by one or several other genes. The gene whose phenotype is expressed is called epistatic, while the phenotype altered or suppressed is called hypostatic.

**eukaryote:** Single-celled or multicellular organism (e.g., plant, animal, or fungi) with a cellular structure that includes a membrane-bound, structurally discrete nucleus and other well-developed subcellular compartments. See also *prokaryote*.

**expression quantitative trait locus (eQTL):** Genomic locus that regulates expression levels of mRNAs or proteins. See also *quantitative trait locus*.

**functional annotation:** Process of attaching biological information (e.g., biochemical function, biological function, involved regulation and interactions, and expression) to genomic elements. See also *genome annotation*.

**functional genomics:** Study of sequencing data to describe gene (and protein) functions and interactions. Unlike genomics, functional genomics focuses on dynamic aspects such as gene transcription, translation, and protein-protein interactions, as opposed to the static aspects of genomic information such as DNA sequence or structures.

**gene:** Fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides, located in a particular position on a particular chromosome, that encodes a specific functional product (i.e., a protein or RNA molecule). Multiple variants (see *allele*) can exist in a population.

**gene expression:** Process by which a gene's coded information is converted into structures present and operating in the cell. Expressed genes include those transcribed into messenger RNA (mRNA) and then translated into proteins, as well as those transcribed into RNA but not translated into proteins [e.g., transfer (tRNA) and ribosomal RNA (rRNA)].

**gene function:** Biochemical reaction, protein-protein interaction, metabolic or signaling pathway association, cellular localization, phenotype, and changes in protein function that are mediated by shifts in protein structure.

**gene product:** Biochemical material, either RNA or protein, resulting from expression of a gene. The amount of gene product is used to measure a gene's level of expression (transcription).

**gene regulatory network:** Intracellular network of regulatory proteins that control the expression of gene subsets involved in particular cellular functions. A simple network would consist of one or more input signaling pathways, regulatory proteins that integrate the input signals, several target genes (in bacteria a target operon), and the RNA and proteins produced from those target genes.

**genome:** All the genetic material in the chromosomes of a particular organism. Most prokaryotes package their entire genome into a single chromosome, while eukaryotes have different numbers of chromosomes. Genome size generally is given as total number of base pairs.

**genome annotation:** Process of identifying elements in the genome and attaching biological information to these elements. Automatic annotation tools perform this process by computer analysis, as opposed to manual annotation (i.e., curation), which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline. See also *functional annotation and structural annotation*.

**genome assembly:** Process of taking a large number of short DNA sequences and putting them back together to create a representation of the original chromosomes from which the DNA originated. In a shotgun sequencing project, all the DNA from a source is first fractured into millions of small pieces. These pieces are then "read" by automated sequencing machines, which can read up to 1,000 nucleotides or bases at a time. A genome assembly algorithm works by taking all the pieces and aligning them to one another, and detecting all places where two of the short sequences, or reads, overlap. These overlapping reads can be merged, and the process continues.

**genome engineering:** Techniques for the targeted, specific modification of the genetic information (or genome) of living organisms.

**genome sequence:** Order of nucleotides or bases within DNA molecules that make up an organism's entire genome. The four bases are adenine, guanine, cytosine, and thymine, represented as A, G, C, and T.

**genome-wide association study (GWAS):** Examination of many common genetic variants in different organisms to see if any variant is statistically associated with a trait. GWAS are used to identify candidate genes or sequence variants that may link to a condition or property of interest.

**genomics:** The study of genes and their function.

**genotype:** An organism's genetic constitution, as distinguished from its physical characteristics (phenotype).

**hemicellulose:** Any of several polysaccharides (e.g., xylans, mannans, and galactans) that cross link and surround cellulose fibers in plant cell walls. Where cellulose is regular in organization and consequently strong, hemicelluloses are more commonly random in structure and more easily hydrolyzed.

**heterozygous:** Having two different alleles for a single trait, or having two different alleles at a single gene or genetic locus. An allele can be dominant, co-dominant, or recessive.

**high throughput:** Done on a massive, automated scale.

**histone:** Protein that provides structural support to a chromosome. For very long DNA molecules to fit into the cell nucleus, they wrap around complexes of histone proteins, giving the chromosome a more compact shape. Some histones variants are associated with the regulation of gene expression.

**horizontal gene transfer:** Exchange of genetic material between two different organisms (typically different species of prokaryotes). This process gives prokaryotes the ability to obtain novel functionalities or cause dramatic changes in community structure over relatively short periods of time.

**interaction network:** Diagram that shows numerous molecular interactions of a cell. Each point or node on the diagram represents a molecule (typically a protein), and each line connecting two nodes indicates that two molecules are capable of interacting.

**interactome:** Molecular interactions of a cell, typically used to describe all protein-protein interactions or those between proteins and other molecules.

**in silico:** Research performed on a computer or via computer simulation.

**in vitro:** Outside of a living organism.

**in vivo:** Within a living organism.

**lignin:** Complex, insoluble polymer whose structure, while not well understood, gives strength and rigidity to cellulose fibers in the cell walls of woody plants. Lignin makes up a significant portion of the mass of dry wood and, after cellulose, is the second most abundant form of organic carbon in the biosphere.

**loci:** Chromosomal locations of genes or genetic markers. (singular: locus)

**messenger RNA (mRNA):** RNA that serves as a template for protein synthesis. See also *transcription* and *translation*.

**metabolic engineering:** Optimizing genetic and regulatory processes within cells to increase the cells' production of a certain substance.

**metabolism:** Collection of all biochemical reactions that an organism uses to obtain the energy and materials it needs to sustain life. An organism uses energy and common biochemical intermediates released from the breakdown of nutrients to drive the synthesis of biological molecules.

**metabolomics:** Type of global molecular analysis that involves identifying and quantifying the metabolome—all metabolites present in a cell at a given time.

**metadata:** Data that describe specific characteristics and usage aspects (e.g., what data are about, when and how data were created, who can access the data, and available formats) of raw data generated from different analyses.

**metagenome:** Genetic material recovered directly from environmental samples.

**metagenomics:** Study of the collective DNA isolated directly from a community of organisms living in a particular environment.

**metaomics:** High-throughput, global analysis of DNA, RNA, proteins, or metabolites isolated directly from a community of organisms living in a particular environment.

**metatranscriptome:** Transcriptome of a group of interacting organisms or species.

**microfluidics:** Technology platforms that deal with the behavior, precise control, and manipulation of fluids that are geometrically constrained to a small, typically sub-millimeter, scale.

**microorganism:** Any unicellular prokaryotic or eukaryotic organisms, sometimes called a microbe.

**model:** Mathematical (or other, e.g. engineering) representation used in computer simulations to calculate the evolving state of dynamic systems.

**model organism:** Organism studied widely by a community of researchers. Biological understanding obtained from model-organism research is used to provide insights into the biological mechanisms of other organisms. Microbial model microorganisms include the bacteria *Escherichia coli*, the yeast *Saccharomyces cerevisiae*, and the mustard weed *Arabidopsis thaliana*.

**modeling:** Use of statistical and computational techniques to create working computer-based models of biological phenomena that can help to formulate hypotheses for experimentation and predict research outcomes.

**molecular machine:** Highly organized assembly of proteins and other molecules that work together as a functional unit to carry out operational, structural, and regulatory activities in the cells.

**oligonucleotide:** Short nucleic acid polymer, typically with fifty or fewer bases.

**omics:** Collective term for a range of new high-throughput biological research methods (e.g., transcriptomics, proteomics, and metabolomics) that systematically investigate entire networks of genes, proteins, and metabolites within cells.

**organelle:** Specialized subunit within a cell that has a specific function and is usually separately enclosed within its own lipid bilayer.

**ortholog:** Similar gene or gene segments appearing in the genomes of different species but resulting from speciation and mutation.

**pathway:** Series of molecular interactions that occur in a specific sequence to carry out a particular cellular process (e.g., sense a signal from the environment, convert sunlight to chemical energy, break down or harvest energy from a carbohydrate, synthesize ATP, or construct a molecular machine).

**phenology:** Study of recurring biological phenomena.

**phenomics:** Collective study of multiple phenotypes (e.g., all phenotypes associated with a particular biological function).

**phenotype:** Physical characteristics of an organism.

**photosynthesis:** Process by which plants, algae, and certain types of prokaryotic organisms capture light energy and use it to drive the transfer of electrons from inorganic donors (e.g., water) to carbon dioxide to produce energy-rich carbohydrates.

**phototroph:** Organism capable of photosynthesis.

**phylogeny:** Evolutionary history that traces the development of a species or taxonomic group over time.

**phylogenetics:** Study of evolutionary relation among groups of organisms (e.g., species, populations), based on their DNA sequences.

**phylogenomics:** Comparison and analysis of entire genomes, or large portions of genomes, to determine the relationship of the function of genes to their evolution.

**phytoplankton:** Free-floating, microscopic photosynthetic organisms (e.g., algae, cyanobacteria, dinoflagellates found in the surface layers of marine and freshwater environments).

**polymorphism:** Occurs when two or more clearly different phenotypes exist in the same population of a species (i.e., the occurrence of more than one form or morph).

**polyploid:** Cells and organisms that contain more than two paired (homologous) sets of chromosomes. Most eukaryotic species are diploid, meaning they have two sets of chromosomes—one set inherited from each parent. However, polyploidy is found in some organisms and is especially common in plants.

**population genetics:** Study of allele frequency distribution and change under the influence of the four main evolutionary processes: natural selection, genetic drift, mutation, and gene flow. Population genetics also encompasses the factors of recombination, population subdivision, and population structure and attempts to explain such phenomena as adaptation and speciation.

**primary production:** the synthesis and storage of organic molecules (biomass) starting with fixation of CO<sub>2</sub> by photosynthesis, in plants and microorganisms

**prokaryote:** Single-celled organism lacking a membrane-bound, structurally discrete nucleus and other subcellular compartments. Bacteria and archaea are prokaryotes. See also *eukaryote*.

**promoter:** DNA site to which RNA polymerase will bind and initiate transcription.

**protein:** Large molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene that codes for the protein. Proteins maintain distinct cell structure, function, and regulation.

**protein complex:** Aggregate structure consisting of multiple protein molecules.

**protein expression:** Subcomponent of gene expression. It consists of the stages after DNA has been transcribed to mRNA. The mRNA is then translated into polypeptide chains, which are ultimately folded into proteins.

**proteome:** Collection of proteins expressed by a cell at a particular time and under specific conditions.

**proteomics:** Large-scale analysis of the proteome to identify which proteins are expressed by an organism under certain conditions. Proteomics provides insights into protein function, modification, regulation, and interaction.

**quantitative trait locus:** Stretch of DNA containing or linked to the genes that underlie a quantitative trait. See also *expression quantitative trait locus*.

**regulatory elements:** Segments of the genome (e.g., regulatory regions, genes that encode regulatory proteins, or small RNAs) involved in controlling gene expression.

**regulatory region or sequence:** Segment of DNA sequence to which a regulatory protein binds to control expression of a gene or group of genes that are expressed together.

**rhizosphere:** Narrow zone of soil surrounding a plant root, typically inhabited by microbial community(ies) that interact with the root.

**RNA (ribonucleic acid):** Molecule that plays an important role in protein synthesis and other chemical activities of the cell. RNA's structure is similar to that of DNA. Classes of RNA molecules include messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and other small RNAs, each serving a different purpose.

**RNA-Seq:** Use of high-throughput sequencing technologies to sequence complementary DNA (cDNA) and obtain information about a sample's RNA content.

**ribosomal RNA (rRNA):** Specialized RNA found in the catalytic core of the ribosome, a molecular machine that synthesizes proteins in all living organisms.

**simulation:** Combination of multiple models into a meaningful representation of a whole system that can be used to predict how the system will behave under various conditions. Simulations can be used to run *in silico* experiments to gain first insights, form hypotheses, and predict outcomes before conducting more expensive physical experiments.

**species:** Taxonomic group of closely related organisms sharing structural and physiological features that distinguish them from individuals belonging to other species. In organisms capable of sexual reproduction, individuals of the same species can interbreed and generate fertile offspring. For microorganisms, a species is a collection of closely related strains.

**structural annotation:** Process of identifying gene elements such as coding regions, gene structure, regulatory motifs, and open reading frames (ORFs). See also *genome annotation*.

**symbiosis:** Ecological relationship between two organisms in which both parties benefit.

**synthetic biology:** Field of biological research and technology that combines science and engineering with the goal of designing and constructing new biological functions and systems not found in nature. Essential synthetic biology tools include DNA sequencing, fabrication of genes, modeling how synthetic genes behave, and precisely measuring gene behavior.

**systems biology:** Use of molecular analyses (e.g., measurements of all genes and proteins expressed in a cell at a particular time) and advanced computational methods to study how networks of interacting biological components determine the properties and activities of living systems.

**taxa:** Categories (e.g., phylum, order, family, genus, or species) used to classify animals and plants. (singular: taxon)

**taxonomy:** Hierarchical classification system for naming and grouping organisms based on evolutionary relationships.

**transcript:** RNA molecule (mRNA) generated from a gene's DNA sequence during transcription.

**transcription:** Synthesis of an RNA copy of a gene's DNA sequence; the first step in gene expression. See also *translation*.

**transcription factor:** Protein that binds to regulatory regions in the genome and helps control gene expression.

**transcriptome:** Set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA produced in one or a population of cells.

**transcriptomics:** Global analysis of expression levels of all RNA transcripts present in a cell at a given time.

**transfer RNA (tRNA):** RNA that transports amino acids to ribosomes for incorporation into a polypeptide undergoing synthesis.

**translation:** Process in which the genetic code carried by mRNA directs the synthesis of proteins from amino acids. See also *transcription*.

## Acronyms and Abbreviations

<b>3D</b>	three dimensional
<b>BER</b>	Office of Biological and Environmental Research
<b>BERAC</b>	Biological and Environmental Research Advisory Committee
<b>ChIP-Seq</b>	chromatin immunoprecipitation sequencing
<b>CO<sub>2</sub></b>	carbon dioxide
<b>DOE</b>	U.S. Department of Energy
<b>EMSL</b>	DOE Environmental Molecular Sciences Laboratory
<b>EPR</b>	electron paramagnetic resonance
<b>eQTL</b>	expression quantitative trait locus
<b>GBS</b>	genotyping by sequencing
<b>GC</b>	gas chromatography
<b>GWAS</b>	genome-wide association study
<b>H<sub>2</sub></b>	hydrogen
<b>HGT</b>	horizontal gene transfer
<b>JGI</b>	DOE Joint Genome Institute
<b>KBase</b>	DOE Systems Biology Knowledgebase
<b>MFD</b>	microfluidic device
<b>MS</b>	mass spectrometry
<b>NERSC</b>	National Energy Research Scientific Computing Center
<b>NIH</b>	National Institutes of Health
<b>NMR</b>	nuclear magnetic resonance
<b>NRC</b>	National Research Council
<b>O<sub>2</sub></b>	oxygen
<b>QTL</b>	quantitative trait locus

