DOE Joint Genome Institute
Strategic Planning Update

JGI | JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

U.S. DEPARTMENT OF ENERGY | Office of Science

Mission

To advance energy and environmental science by providing access to state-of-the-art genomics capabilities in support of the US Department of Energy's research mission.

Vision

To continually evolve as a leading edge genome science user facility developing and applying genomics capabilities to solve the most pressing worldwide energy and environment challenges.

Table of Contents

The Annual Genomics of Energy and Environment Meeting, as well as related workshops, continue to be primary ways how the JGI engages in face-to-face interactions with its users and leads the development of user communities in areas of strategic interest.

I.   Introduction

A central theme to the 2012 DOE JGI Ten-Year Strategic Vision was the recognition that a fundamental unsolved problem in genomics is the need for high-throughput approaches to bridge the gap between the availability of DNA sequence data and our ability to assign biological function to it. Now in 2016, we are engaged in reviewing the 2012 plan, exploring progress that has been made in implementing the plan and assessing the landscape going forward. One major conclusion from this analysis is that the central theme articulated in the 2012 Ten-Year Strategic Vision remains. The challenge is no longer generating massive amounts of sequence data, but imbuing these data with meaning; thus the DOE JGI's niche is the development of a diversity of large-scale experimental and computational capabilities to link sequence to biological insights relevant to energy and environmental research. This will range from preparing material and applying functional capabilities prior to genomic analysis to post-sequence processing and manipulations to enable the Institute's users to carry out sequence-to-function studies that are beyond the capabilities of individual laboratories.

New sequence-to-function capabilities that have been established at the DOE JGI since the publication of the 2012 plan include epigenomics and methylation analyses, enabling exploration of this "second code" of modifications to the DNA bases; transposon-mediated mutagenesis paired with sequencing, to assess the genes required for survival in various contexts; transcriptomics and metatranscriptomics, to investigate when genes are transcribed into RNA; metabolomics, to characterize the basic biochemistry of cells; and exome and whole-genome resequencing to survey whole populations rather than single individuals. In each case the capability is coupled with appropriate analysis tools. A DNA Synthesis Science Program has also been developed to enable users to combine mining of the DOE JGI databases for hypothesis generation with design and synthesis of genes and gene pathways that can be used to study functional features of the sequence in appropriate hosts. These capabilities, which we expect to evolve and grow with time, align with the DOE JGI's transition from a production sequencing center to a genomic science analysis resource, allowing researchers to both experimentally and computationally convert sequence data into biological insights.

One major initiative associated with implementing the Vision that is now in full swing is the Emerging Technologies Opportunity Program (ETOP). This program was designed to facilitate the availability to DOE JGI users of cutting-edge sequence-to-function capabilities developed by external groups collaborating with the DOE JGI. Through focused ETOP efforts, where the DOE JGI supports relevant activities in leading laboratories around the US and the world, the Institute has developed partnerships with pioneers in microfluidics, Raman spectroscopy and single-cell function-driven genomics, advanced DNA synthesis technologies, metagenomic assembly, and plant and fungal material acquisition.

Another new direction for the DOE JGI is a collaborative program with the Environmental Molecular Sciences Laboratory (EMSL) at the Pacific Northwest National Laboratory (PNNL). EMSL offers a large collection of proteomics and imaging tools and, together, DOE JGI and EMSL are making their respective capabilities available to users. This combined user program supports exciting science that synergistically leverages each User Facility's strengths to benefit user science projects.

In the pages that follow, we go into some detail about how we have been advancing many aspects of the 2012 DOE JGI Strategic Vision. In particular, this document reviews the 2015 Strategic Planning Retreat held April 13-14, 2015, in Pacifica, CA. At this retreat we evaluated our successes and challenges in the implementation of the 2012 Vision, developed time lines and goals to ensure forward progress, explored new scientific and technological areas that may be relevant to the DOE JGI, and discussed areas that may be of reduced priority in the future.

Over the next five years, the DOE JGI will transition fully to a next-generation Genome Science User Facility while maintaining its primary identity as an Office of Science National User Facility focused on solving fundamental DOE mission challenges in bioenergy and the environment. To accomplish this ambitious goal, the DOE JGI will need to develop and apply new advanced and high-throughput technologies, approaches and computational resources, and it will also need to collaborate with the world's best scientists in relevant research areas to assign computationally predicted and experimentally validated functions to genes and genomes critical to DOE mission interests.

The Integrative Genomics Building (IGB is the prospective new home of the DOE JGI at Lawrence Berkeley National Laboratory where KBase staff will be collocated. Groundbreaking is anticipated in 2016 with occupancy in 2018.

II. Strategic Planning at the DOE JGI

Development of the Strategic Vision

From 2010 to 2012, the DOE JGI underwent a major strategic planning effort to develop a new scientific and organizational vision. In developing this vision, the DOE JGI considered and actively sought input from a variety of sources, including the 2010 BER Grand Challenges Long-Term Vision, a strategic planning session at the 2011 DOE JGI User Meeting, feedback from all DOE JGI advisory committees, several dedicated workshops organized by the DOE JGI, and a workshop organized by BER in 2012 (see workshop report "DOE Joint Genome Institute—Strategic Planning for the Genomic Sciences"). The strategic planning process culminated in the publication of "Forging the Future of the DOE JGI – A 10-Year Strategic Vision" in 2012.

The 2012 Strategic Vision document describes the transformation of the DOE JGI from a sequencing center to a "Next-Generation Genome Science User Facility" that provides its users with access to cutting-edge genomic technologies to make biological discoveries relevant to the DOE mission. This transition is organized around a set of science drivers, i.e. high-level scientific directions related to areas such as energy security, climate, and environment that the DOE JGI will address in the future. It also outlines general paradigm changes such as a shift from reductionist studies of individual organisms to systems-level studies, and fundamentally new areas of science and technology that the DOE JGI should expand into, such as DNA synthesis and the use of advanced computing for biological data analysis.

Along with these science drivers, the 2012 Strategic Vision also outlines a set of more concrete goals that need to be met in order to address these high-level challenges. Organized in three pillars, called "Experimental Data Generation", "Biological Data Interpretation", and "User Interactions", an extensive set of specific capabilities are described that the DOE JGI will seek to implement and make available to its users. Examples include state-of-the-art DNA sequencing technologies, single-cell analysis tools, DNA synthesis platforms, and high-throughput sequence-to function assays (Pillar 1); functional interpretation of sequence, utilization of high-performance computing platforms, and multidimensional data integration (Pillar 2); and organization of user communities and development of interactive data platforms (Pillar 3).

Assessing Progress and Developing Implementation Milestones

While the 2012 Strategic Vision was published with a 10-year time horizon in mind, strategic planning at the DOE JGI is a continuous, ongoing process. As programmatic priorities, scientific knowledge, research paradigms, and genomic technologies continue to evolve, it is necessary

to revisit the Strategic Vision at regular intervals. The purpose of this process is to assess progress towards the implementation of the Strategic Vision, refine and refocus the long-term vision as needed, and develop and adjust short- and mid-term implementation plans to guide scientific and organizational decision making processes.

The DOE JGI 2015 Strategic Planning Retreat brought together approximately 20 DOE JGI management and program staff and 12 outstanding external scientists working in fields directly relevant to DOE JGI science (see *Appendix* for a list of participants and the retreat agenda). Summarizing discussions at and conclusions from this retreat, the present document provides a brief review of the implementation accomplishments to date, organized by program area. This is followed by a set of two- and five-year milestones as well as stretch goals to provide clear targets and useful measures of success as the DOE JGI Strategic Vision continues to be implemented. Finally, the report describes several new strategic initiatives that were developed at the retreat.
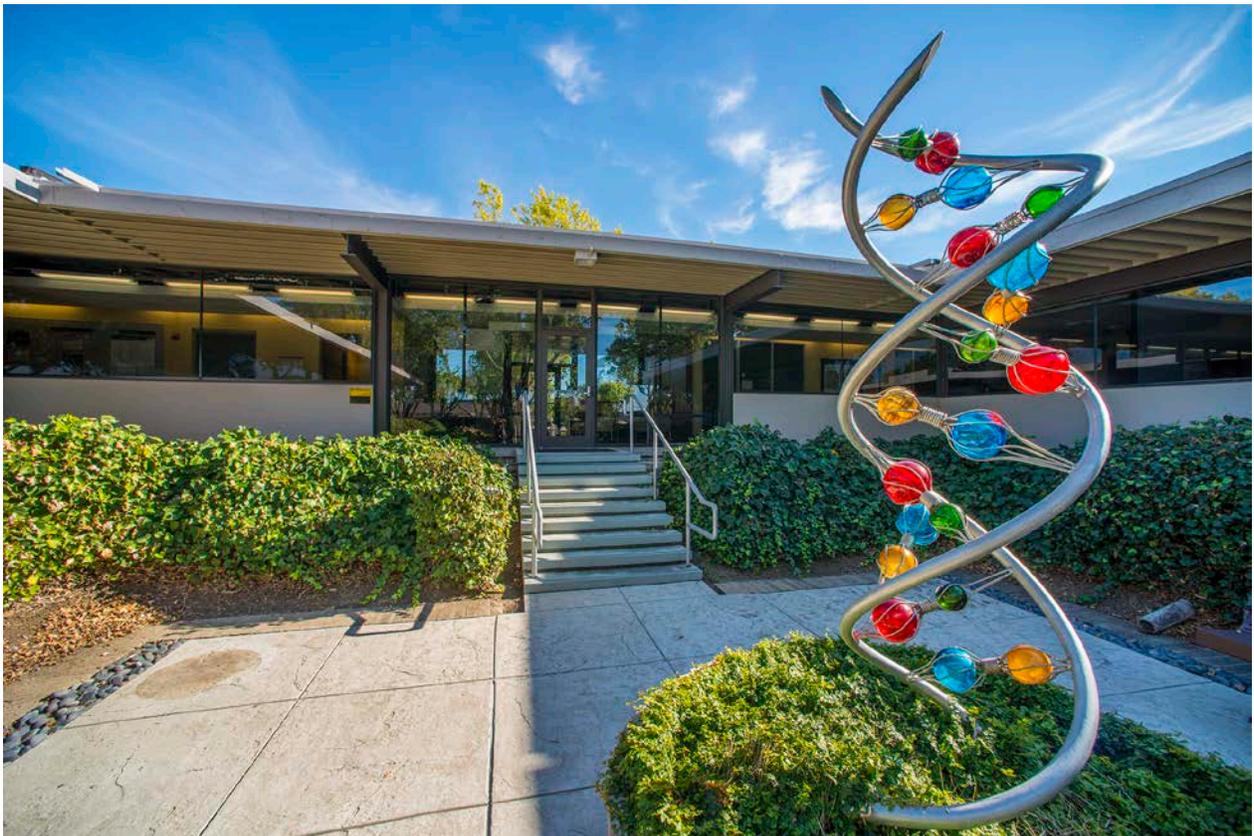
The DOE Joint Genome Institute campus has been located in Walnut Creek, California since 1999.

III. Implementing the 2012 Strategic Vision

This section provides an overview of implementation efforts towards the 2012 Strategic Vision over the past four years. For each program area, we describe specific steps that have already been taken, as well as possible adjustments to the directions outlined in the strategic plan. This section also describes, in the form of two- and five-year implementation milestones and stretch goals, a roadmap toward further implementation of the DOE JGI's Strategic Vision.

Genomic Technologies and Functional Genomics

Implementation Progress

A major thrust of the DOE JGI Strategic Vision has been to continue to provide leadership in genomics and access to state-of-the-art technologies, as well as to grow capabilities to understand genome function. Toward these aims, the DOE JGI has had numerous notable accomplishments as described below. In summary, these technological investments have been critical to realizing the scientific goals of the DOE JGI Strategic Vision.

New Genomic and Sequence-to-Function Technologies

The DOE JGI invested in a single-cell genomics pipeline and built out a significant DNA synthesis pipeline that is now associated with a new Science Program. As part of the single cell pipeline, the DOE JGI now also routinely isolates hundreds of cells from environments of interest and screens them for those most desirable genomes for DNA sequencing and assembly. The DNA synthesis pipeline and its outputs are described separately in the "DNA Synthesis Science Program" section of this document. Finally, the DOE JGI greatly expanded the number of library types to enable functional genomic studies, which now include epigenomic techniques such as chromatin immunoprecipitation and sequencing (ChIP-seq), Assay for Transposase-Accessible Chromatin (ATAC-seq), and bisulfite-seq to analyze methylation patterns, to name a few.

Remaining State of the Art in DNA Sequencing

The DOE JGI has remained state of the art in DNA sequencing with strategic partnerships with Illumina, Pacific Biosciences, and Oxford Nanopore. The DOE JGI now operates a variety of Illumina sequencers including HiSeq 2000s, HiSeq 2500s, a NextSeq 500, and MiSeqs. The DOE JGI is currently exploring the latest HiSeq 4000 platform which should enable a further 30% cost reduction in DNA sequencing. In addition, the DOE JGI has three Pacific Biosciences instruments. We have scaled microbial and fungal *de novo* genome assemblies on this platform for a final product that exceeds even the quality of reference genomes obtained by the Sanger method. Consistent with the DOE JGI's strategic plan to remain state of the art in sequencing,

the Institute also has five Oxford Nanopore MinION units as part of the early access program and is seeing steady progress in data output and quality.

To sustain the increased output of all of these sequencing technologies, the Genomic Technologies Department has also scaled upstream sample processing. This includes robust automated 96-well plate processing in an end-to-end workflow for a substantial number of products including shotgun sequencing, small RNAs, transcript profiling, and exome capture. A new freezer management system has also been introduced for more automated sample storage and tracking.

Developing New Genomic Technologies with External Partners

Complementary to in-house development of new technologies, the DOE JGI has developed strategic partnerships through the Emerging Technologies Opportunity Program (ETOP), which includes securing both fungal (Jon Magnuson, Pacific Northwest National Laboratory [PNNL]) and plant (Rod Wing, University of Arizona) DNA samples for DOE JGI processing. The DOE JGI also supported two ETOP projects focused on improving single-cell omics (Steve Quake, Stanford) and functional screening using Raman microspectroscopy (Roman Stocker, MIT). An additional ETOP partnership has been established in the area of developing microfluidic solutions for customized library construction at the nanoscale (Paul Blainey, MIT).

In addition to these new capabilities, the DOE JGI has also collaborated with the Environmental Molecular Sciences Laboratory (EMSL) to develop a new user program, the JGI-EMSL Collaborative Science Initiative. Through this program, users can propose research projects that combine sequence-based capabilities located at DOE JGI and advanced analytical capabilities located at EMSL through a single joint proposal, review and project management process.

Implementation Milestones

Area	Two years	Five years
High-Throughput DNA Sequencing	Establish nanopore sequencing for user access	Apply nanopore sequencing to many DOE JGI standard products
	Identify next disruptive technology	Pilot-scale implementation of next disruptive technology
	Establish miniaturization of various library construction processes	Reduce reagent cost for all DOE JGI processes through microfluidics
	Assess microfluidic versus emulsion approaches	Routinely use a new microfluidic or emulsion-based approach as part of one or several sample preparation workflows
	Generate 200 Tbp of sequence	Generate 1+ Pbp of sequence

Experimental Functional Annotation	Initiate three new ETOPs aimed at complementary technologies for experimental functional annotation	Routinely apply and offer to users three new complementary technologies for experimental functional annotation
	Assess impact of metabolomics linked to pathway engineering activities in the DNA Synthesis Science Program	Routinely provide metabolomics capabilities to users as part of a sequence-to-function validation pipeline
Single-Cell Genomics	Initiate 16S-independent workflow for single cell genomics to discover cryptic phyla and provide alternative approach to shotgun sequencing	Sequence 10K single cells per year
	Implement approach developed through ETOP to sequence single cells selected based on functional traits revealed by Raman microspectroscopy	Offer Raman-based functional analysis and single cell sequencing through Community Science Program
	Develop high-throughput, single-cell transcriptomics pipeline for unicellular eukaryotes	Sequence 1,000 single fungal cell transcriptomes in support of Community Science Program projects
	Develop flow cytometry based approached for functional and phylogenetic targeting of cells and through ETOP	Sequence 1,000 genomes per year from targeted single cells
Sample Preparation	Assess additional ETOPs for high quality nucleic acid preparation	Routinely use new capabilities from ETOP for high quality nucleic acid preparation

Stretch Goals

- Develop and implement a single efficient preparation technology for DNA regardless of the source
- Establish methods for direct sequencing of samples without nucleic acid purification
- Develop methods for gap-free single cell sequence assemblies of microbial and fungal genomes

DNA Synthesis Science Program

Implementation Progress

The DNA Synthesis Science Program at the DOE JGI started in 2011 and has made significant progress over the past five years.

DNA Synthesis as an Integral Component of a Sequence-to-Function Workflow

To date, much of the most impactful work in synthetic biology and pathway engineering has been carried out by self-identified "synthetic biologists", i.e. experts in the requisite techniques. In contrast, many investigators working on major problems in energy and environmental genomics do not routinely utilize DNA synthesis for their projects because of the significant technical hurdles. In particular, while commercial providers now routinely offer custom synthesis of DNA for a fee, these services are typically disconnected from upstream sequence analysis and design steps, as well as downstream host engineering and functional analysis capabilities. A central goal of the DOE JGI's Synthesis Science Program is to provide users with end-to-end capabilities to a) develop hypotheses based on DOE JGI-generated and/or externally produced sequence data, b) design DNA constructs to test those hypotheses using cutting-edge design and synthesis approaches at very large scale and c) assess the results though sensitive and precise phenotyping (e.g., metabolomics, transcriptomics).

Increase Throughput

The Synthetic Biology group has developed a robust platform technology capable of synthesizing 5-6 million basepairs (Mbp) of DNA annually. In FY15, it offered 4 Mbp of DNA synthesis capacity to the scientific user community, with 1.5 Mbp reserved for the DOE Bioenergy Research Centers, 2.1 Mbp reserved for the Community Science Program, and 0.4 Mbp for the internal Director's Science (DS) program. Since FY13, the available DNA synthesis capacity has increased by 10% per year.

Size of constructs	CT in FY14 (Q1/Q2)*	CT in FY15 (Q1/Q2)*
Small Size Constructs (< 1 kbp)	129 calendar days	49 calendar days
Medium Size Constructs (1-5 kbp)	165 calendar days	95 calendar days
Large Size Constructs (5-10 kbp)	382 calendar days	104 calendar days
Very Large Size Constructs (> 10 kbp)	464 calendar days	173 calendar days

for comparison, cycle time (CT) is reported for the first half (quarters 1 and 2, Q1/2) of each fiscal year (FY)

Decrease Project Cycle Times

The Synthetic Biology group significantly improved the production cycle time of the DNA synthesis platform in the past year as described below, with two major modifications to the synthesis pipeline: (i) starting material (single strand oligo DNA) was replaced with double stranded DNA of up to 3 kbp in size; (ii) yeast TAR-based cloning is now used to assemble constructs larger than 10 kbp.

Growth of a Synthesis Science Program

To support the growth of the DNA Synthesis Science Program, the DOE JGI hired Dr. Yasuo Yoshikuni as a new investigator to lead the program.

The DNA Synthesis Science Program has attracted increasing numbers of users. The number of proposals increased to 17 proposals in the Q2 FY15 CSP semiannual call, which represents a 30% increase compared to the previous call. In alignment with major strategic initiatives of the DOE JGI, several of these projects were aimed at the synthesis of secondary metabolite biosynthetic pathways.

As part of its outreach activities and to support the growth of new user communities around capabilities offered by the DOE JGI, the DNA Synthesis Science Program hosted two workshops in FY15, a DNA Synthesis Science Workshop at the DOE JGI User Meeting, and a workshop "Biological Systems Modulated by Secondary Metabolites" held at the DOE JGI in April 2015.
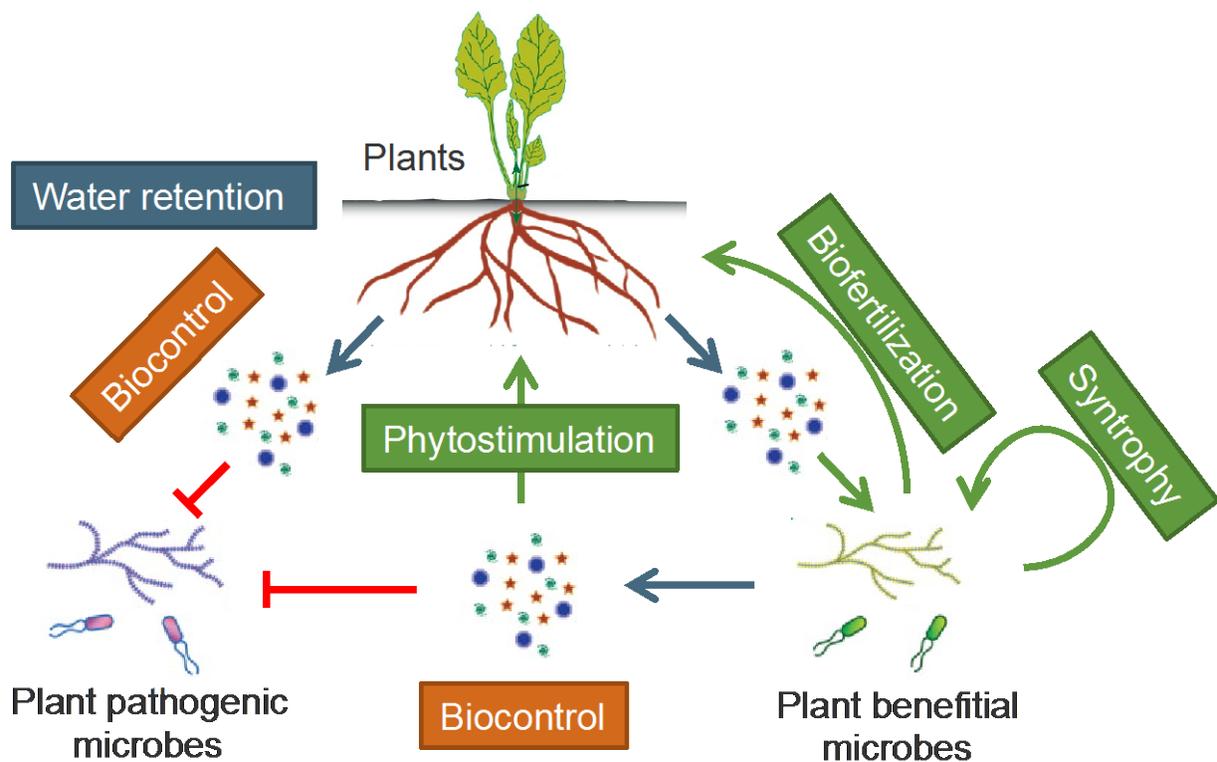
Continue to Develop State-of-the-Art DNA Synthesis Capabilities

To further enhance the DNA synthesis capabilities of the DOE JGI, an ETOP project was initiated with ETOP partner Prof. Jay Shendure (U. Washington) in FY14. The goals of this ETOP included the development of a "dial-out PCR" method to facilitate the synthesis of double-strand DNA in the 1-3 kbp size range. While partially successful, due to general alterations to our synthesis procedure this technology became obsolete during the course of the ETOP. In parallel, Dr. Shendure's group successfully developed methods for transposase-based high-throughput library construction ("tagmentation") for efficient and cost-effective synthesis validation. This technology was subsequently transferred to the DOE JGI. Currently, the Synthetic Biology group is further improving the throughput and the cost for synthesis validation.

Implementation Milestones

Area	Two years	Five years
Pathway Design	1,200 pathways per year designed	6,000 pathways per year designed
Pathway Assembly	120 pathways per year assembled	600 pathways per year assembled
Pathway Synthesis Capacity	10+ Mbp	50+ Mbp
Strain Engineering - Hosts	*E. coli*, *S. cerevisiae*, *P. fluorescens*, *Streptomyces sp.*, *B. subtilis*	Other Pseudomonads, other Streptomycetes, Firmicutes, *Aspergillus nidulans*
User Program Capabilities	Design, refactor, and synthesize pathways of moderate complexity	Design, refactor, and synthesize large pathways
	Pathway insertion into *E. coli, S. cerevisiae, P. fluorescens, Streptomyces sp.*, and *B. subtilis*	Pathway insertion into *E. coli, S. cerevisiae*, Pseudomonads, Streptomycetes and Firmicutes
	Fermentation up to 2 L-scale	Fermentation up to 10 L-scale

Stretch Goals

- Enable the modulation of biomes via secondary metabolites

- Develop a synthesis-enabled sequence-to-function pipeline for secondary metabolite pathways that can complete in one month the selection, design, refactoring, synthesis, introduction into a suitable host strain, pathway expression, and characterization of a secondary metabolite

Plant Genomics Program

Implementation Progress

The DOE JGI Plant Genomics Program, one of its primary activities, has made significant progress in the past four years in the areas of sample preparation, assembling genomes from

diverse data types, collecting information about the function of plant genes, and improving the ability to disseminate data through interactive tools.

Improving Access to Large Numbers of DNA and RNA Samples

A significant bottleneck in the ability of the DOE JGI to accelerate plant-related user science is the difficulty in isolating large numbers of high quality DNA and RNA samples. It is now routine for plant projects to include 400 or 500 samples. To develop isolation capabilities to offer DOE JGI plant users, the Plant Program has established a working model under the Emerging Technical Opportunities Program (ETOP) with the laboratory of Dr. Rod Wing at the University of Arizona, who is an expert in extracting nucleic acid material from plant tissues. With this ETOP, users send frozen tissue directly to the Wing laboratory to provide high molecular weight DNA for DOE JGI sequencing pipelines.

Developing New Methods for Plant Genome Assembly

In order to meet the strategic needs of working in complex and polyploid genome sequences, the DOE JGI has developed new methods for complex plant genome assembly and incorporated new data types into the process. These methods include new ways of localizing sequence using high-density genetic mapping by sequencing and a fast, multithreaded assembler for short read data called Meraculous. We have integrated long read single molecule sequence from Pacific Biosciences into plant genome assemblies and continue to develop methods to take advantage of this new data type.

Plant Functional Data Across the Flagship Genomes

To address the DOE JGI Strategic Vision's focus on understanding gene function in plants, we have begun major efforts to collect gene expression data across the DOE JGI Plant Flagships including the DOE JGI Plant Gene Atlas project. These data enable the identification of genes that work together to perform specific functions within complex plant biochemical pathways. In addition, such data will also enable and accelerate the development of urgently needed improved tools for plant genome annotation. With such tools in hand, DOE JGI users can infer function about important unknown genes by examining their expression throughout the plant kingdom.

Identifying Key Agronomic Genes

With the DOE JGI partners at the Bioenergy Research Centers (BRCs) and key academic collaborators, we have made progress in identifying allelic variants related to important agronomic and DOE-relevant traits in poplar through large-scale common garden and genome-wide association studies. We are in the process of extending this methodology into foxtail millet, switchgrass, and *Brachypodium* in order to capture the functions of genes related to drought tolerance and abiotic stress in DOE JGI plant flagship genomes.

Improving Data Dissemination and Interactive Data Tools

We continue to extend and improve the online analysis, visualization, and data query and access tools at Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html) to deliver the full

scientific value of DOE JGI plant genomic data to our users. We have migrated our software platform to use modern industry-standard web frameworks (e.g., Google Web Toolkit) for both sustainability and scalability. We have also upgraded our genome browser and data warehouse components to JBrowse and InterMine, respectively, providing users a Google Maps-style interface and state-of-the-art data exploration tools. We have enhanced our database back-end and visualization tools to accommodate the increased focus on resequencing and gene expression data from DOE JGI flagship genome projects and the DOE JGI Plant Gene Atlas, with new gene-centric as well as genome-wide interactive views of population diversity and co-expression. These changes enable our users to ask and answer increasingly sophisticated questions about gene function and plant adaptation across conditions, geography, populations and evolutionary time.

Implementation Milestones

Area	Two years	Five years
Sample Preparation	Develop capability to isolate 5,000 high-quality, high-molecular weight RNA / DNA samples per year	Develop capability to perform expression analysis on picogram quantities from small volumes of plant cells
Assembling Sequence Data	Generate 8 alternative reference genomes from bioenergy-relevant subspecies/genotypes of DOE JGI flagship plants including two *Populus* genotypes, *Setaria*, *Sorghum*, three *Brachypodium* genotypes, and *Oryza*	Develop Pan and Core genomes for relevant DOE JGI Plant Flagships (*Populus*, *Brachypodium*, *Sorghum*, *Setaria*, *Panicum*) including annotation and structural variation. Develop computational methods to analyze genomic assays (RNA-seq, resequencing) against Pan and Core genomes
Experimental Functional Annotation	Establish in-house experimental capabilities for a model plant	Offer users experimental capabilities to explore plant processes such as metabolism and plant-microbe interactions relevant to bioenergy
Functional Discovery and Annotation	Increase capacity for structural annotation 10x and develop a system for multiple-genotype annotations that produces complete genes	Apply comparative gene expression data for functional gene prediction
Multidimensional Data Integration	Capture QTL/eQTL data for ongoing flagship projects	Apply these data to improve cross-plant analysis of gene function and provide cross-species co-expression networks for target sets
Tools for	Design additional tools to identify	Use these tools to characterize the

Scientific Discovery	genes and transcription factors underlying plant processes of BER mission interest in a flagship plant	gene network(s) underlying abiotic stress response and water and nutrient use
Community Development	Coordinate workshop to address the function of unknown genes in JGI plants, including computational and experimental methods for functional validation	Develop a computational toolkit to mine data from common gardens of flagship species

Stretch Goals

- High-confidence functional prediction or experimental validation for >50% of genes in flagship genomes

- Enable the identification of molecular mechanisms and environmental responses that differentiate perennials from annual plants

Fungal Genomics Program

Implementation Progress

The DOE JGI Fungal Genomics program is making progress in all areas of the 2012 DOE JGI Strategic Vision. In particular, the Fungal Program significantly increased throughput in genome and transcriptome sequencing and annotated its first Pacific Biosciences-sequenced genomes.

Access to Environmental Fungal Material

In collaboration with the Metagenome program, we developed the ITS (internal transcribed spacer) amplicon sequencing and analysis pipeline to explore the species composition of fungal communities, the first step towards fungal metagenomics. In partnership with the Microbial Program, we piloted fungal single-cell genomics to explore uncultured fungi and released annotation of the first three fungal genomes sequenced using this approach. To address supply chain challenges, one of the ETOP projects is focused on high-throughput culturing of fungi and nucleic acid extraction in partnership with the Fungal Biotechnology group at PNNL led by Jon Magnuson.

Automation to Increase Throughput

Assembly of Illumina minimal draft genomes was completely automated and integrated into the DOE JGI RollingQC system. For high-performance computing-based annotation, we developed workflow management and visualization tools to run and monitor fungal annotation across different National Energy Research Scientific Computing Center (NERSC) platforms and modules for targeted annotation of gene families including secondary metabolite biosynthesis gene clusters. The latter also provides targets for the DOE JGI DNA Synthesis Program.

Multidimensional Data Integration

The latest release of our interactive data platforms, the fungal genome portal MycoCosm (jgi.doe.gov/fungi) offers an interactive fungal tree of life enabling exploration of the taxonomic positions of all sequenced fungi and multidimensional data integration (genomes, transcriptomes, proteomes, and resequencing data). It is equipped with tools for web-based comparative analysis, community-based genome curation, and support for nominating new species for sequencing within the context of the 1000 Fungal Genomes Project.

Increase Outreach and User Interactions of the Program

The Fungal Program also engaged in productive interactions with its users. We organized DOE JGI workshops and tutorials as part of the DOE JGI User Meeting and many large fungal biology meetings such as the bi-annual Fungal Genetics conference at Asilomar and the European Conference on Fungal Genetics, as well as annual meetings of the Mycological Society of America and the American Phytopathological Society. In addition, we hosted 11 scientists and graduate students associated with fungal projects carried out by the DOE JGI for periods of 1-12 months for training and research.

Implementation Milestones

Area	Two years	Five years
High-Throughput DNA Sequencing	Move complex fungal genomes to Pacific Biosciences and Oxford Nanopore sequencing	Complete the 1000 Fungal Genomes Project
Experimental Functional Annotation	Generate transcriptomic and epigenomic data for select model fungi	Produce data and tools for metabolic modeling and regulatory networks in select model fungi
Single Cell Genomics	Develop fungal single-cell genomics and transcriptomics pipeline	Develop function-based single-cell environmental genomics
Sample Preparation	Offer fungal sample preparation capabilities developed through an ETOP with PNNL to individual DOE JGI users	Connect to citizen scientists to collect DNA/RNA materials from environmental samples
Interactive Data Platforms	Build interactive tools for transcriptomic data analysis	Develop tools for monitoring fungal interactions

Stretch Goals

- For fungal isolates and simple systems of fungi interacting with plants, other fungi or microbial communities, produce multi-dimensional omics datasets and convert data into gene networks and metabolic models

- Identify generic genetic traits of fungal parasitism and mutualism and biomass decomposing potential to enable genome-based fungal species characterization across the fungal kingdom

Microbial Genomics Program

Implementation Progress

Since the release of the DOE JGI Strategic Vision, the Microbial Genomics Program has taken several key steps towards implementing the 2012 vision.

Development of Function-Driven Genomic Strategies

In the area of function-driven single-cell genomics, an ETOP project for Raman-based functional sorting of single cells was solicited and accepted. This effort has met all its milestones and is expected to provide a new and powerful methodology in the sequence-to-function toolkit for uncultivated microorganisms. Twenty functionally Raman-sorted microcolonies have already been sequenced and analyzed. Additionally, BONCAT (bioorthogonal noncanonical amino acid tagging)-labeled clusters of anaerobic methanotrophic archaea (ANME) and associated bacteria were sorted at the DOE JGI in collaboration with the laboratory of Victoria Orphan (CalTech), providing a high throughput methodology for the enrichment of translationally active environmental cells. This is the first instance of combining BONCAT with flow cytometry and genomic sequencing and highly complementary to the aforementioned activity-sorting capabilities. Lastly, a Laboratory Directed Research and Development (LDRD) project on fluorescently-labeled cellulose-enabled single-cell sorting was initiated, for which benchmark testing is currently under way.

Exploring the Phylogenetic Diversity of Microbial Life on Earth

In the area of phylogenetic diversity, an automated decontamination tool (ProDeGe; Protocol for Decontamination of Genomes) was developed at the DOE JGI to support the quality control of single-cell genomes and genomes from metagenomes, many of which are of high phylogenetic novelty. Approximately 150 single-cell amplification products from candidate phyla were generated for genome sequencing under the Microbial Dark Matter II proposal umbrella. These are from several DOE-relevant sites and include representatives from uncultivated phyla without sequenced genomes, as well as completely novel phyla. Such previously unknown phyla without any existing 16S rRNA gene sequence information represent current taxonomic blind spots. To further explore and find members of novel uncultivated phyla, >100 single-cell amplification products without 16S rRNA gene amplification products have been prepared for sequencing. Lastly, a pilot study for Pacific Biosciences sequencing of Actinomycetes isolates has been successfully completed.

Capture All Microbial Data

Continued developments in IMG (Integrated Microbial Genomes) are critical to keep pace with novel data types and the continuously evolving scientific directions of the Microbial Program. IMG has continued the integration of novel -omics data types (e.g., methyl-seq, RNA-seq), has implemented biosynthetic cluster identification for isolates and single cells, and has added Average Nucleotide Identity (ANI) analysis, which was published earlier this year, to the available toolkit. As part of the onsite training and user interactions, 2-3 onsite workshops have been hosted per year by the IMG team.

Implementation Milestones

Area	Two years	Five years
Function-Driven (Single-Cell) Genomics	Analyze 100 functionally targeted microcolonies	Routinely generate functionally targeted Single Amplified Genomes (SAGs)
	Benchmark fluorescently-labeled substrate approach for cell isolation	Apply fluorescently-labeled substrate approach to environmental samples
	Pilot exometabolomics to define metabolite exchange capabilities of individual microbial community members	Exometabolomics between co-cultured organisms or enriched clusters
Phylogenetic Diversity	Sequence 500 SAGs from >20 candidate phyla	Discover major novel branches to the tree of life
	Analyze genomes from 100 single cells that could not be taxonomically identified with traditional 16S rRNA-based approaches	Provide automated analysis pipeline for the 16S rRNA gene independent analysis of 1000s of single cells per year
	Establish high-throughput pipeline for Pacific Biosciences sequencing of microbial genomes	Complete Pacific Biosciences sequencing of 1000 Actinomycete genomes and analysis for biosynthetic clusters
GOLD/IMG	Incorporation of experimental data of function-driven genomics projects	New visualization methods to enable analysis of 100,000s of isolates/single cells

Stretch Goals

- For single cells, provide basic cell imaging, transcript, protein and metabolite profiling, and complete genomes

- Model syntrophic networks of microbial assemblages and predict community behavior through the integration of multi-omics data in collaboration with KBase *

- Enable the identification of novel mechanisms of microbe-microbe and plant-microbe interactions *

- Enable the understanding and exploitation of molecular underpinnings of microbial plant growth-promoting activities *

*Note: * joint Microbial/Metagenome Program goals*

Metagenome Program

Implementation Progress

As described in detail below, the Metagenome Program has made significant progress towards the goals laid out in the 2012 DOE JGI Strategic Vision. While sequencing of microbial genomes and communities are currently overseen by two distinct scientific programs, Microbial and Metagenomics, many aspects of their analysis are merged under the umbrella of the Prokaryote Super Program. An increasing number of proposals span the two programs; for example, single cell genome sequencing is now routinely combined with metagenomic approaches to gain insight into the composition of environmental communities, as well as the genomes of individual uncultivated organisms from these communities. Moreover, some project types, such as phylogenetically selected, flow sorted populations, do not clearly fit either program. As a result, the appropriate program structure within the Prokaryotic Superprogram will continue to be evaluated. Beyond our own science programs, the DOE JGI is actively involved in shaping community-wide efforts such as the "Unified Microbiome Initiative" (Alivisatos et. al. [2015], Science, 350: 507-508) to which the DOE JGI microbial and metagenomics efforts are expected to make significant contributions.

Increase the Throughput and Quality of Metagenomic Data

In the pillar of experimental data generation, a key objective was to increase data output, particularly in areas beyond straightforward shotgun sequencing of genomic DNA. In line with this, metatranscriptome sample throughput has risen rapidly, doubling from FY13 to FY14 and on track to more than double again in FY15. New library construction methods enable sequencing of extremely low quantities of DNA (1 nanogram or even less), which has expanded the range of communities able to be interrogated with these methods. In particular, a joint project with the Microbial Program has demonstrated the feasibility of so-called "mini-metagenome" sequencing of flow-sorted cells, paving the way towards functional targeting of metagenome sequencing.

Improve Assembly and Analysis

The metagenome assembly pipeline has undergone frequent benchmarking and updating. An ETOP for assembly and binning of individual genomes from metagenome data was solicited and funded, and is producing tools for leveraging different data types (e.g. relative abundance,

long read data from single molecule sequencing) to better assemble genomes from metagenomes.

A number of new analytical tools enhance biological data interpretation, including a recently published metagenome visualization tool, Elviz, which allows for integration of both assembly and annotation results and metadata. Porting of the metagenome annotation pipeline to a high-performance computing (HPC) environment has allowed the IMG/M (Integrated Microbial Genomes with Microbiome samples) pipeline to maintain the same high quality of annotations despite the ever-increasing data load. Tools for improved prediction of biosynthetic clusters in metagenomes were also developed within IMG, as described in a recent publication, and now enable in-depth studies of microbial communication via secondary metabolites. A workshop "Genomes to Secondary Metabolites" was held at the DOE JGI in May 2015.

User Interactions

User interactions continue to be facilitated by workshops on Microbial Genomics and Metagenomics (MGM) held at the DOE JGI. User relationships have been enhanced by the hire of a scientist to provide user support (Adam Rivers), and by opening our doors to outside scientists via the Visiting Scholar Program, including Fulbright Scholar Carly Rosewarne of CSIRO (Australia) in Summer 2015.

Implementation Milestones

Area	Two years	Five years
Metagenome Data Generation	Use single-molecule long read sequencing to improve assembly and detect base modifications in microbial community data	Use single-molecule sequencing to detect novel base modifications in community DNA
	Offer at least one technology for functionally targeted metagenomes, either based on stable isotope probing (SIP) or flow sorting, with appropriate analytical tools	Enable IMG-GOLD metadata submission via smartphone
Metagenome Data Inter-pretation	Assemble terabase-scale datasets	Enable large-scale exploration of phylogenetic diversity in metagenome datasets
	Annotate gene families found only in uncultivated lineages	Build improved IMG/M resident tools for the annotation of metagenomes
	Reconstruct >50 genomes from a single environment	Provide tools for iterative binning and assembly of specific target genomes including those present at low coverage in multiple samples
	Include automated genome binning and identification of potential novel	Build community-level metabolic models, e.g. to identify metabolic

	lineages for all metagenome sequence datasets	partners, via assembly, binning, annotation and metabolic reconstruction software available through JGI, IMG and KBase
Metagenome Discovery	Reconstruct a genome of an uncultivated microbe with experimentally characterized functional activity	Reconstruct multiple genomes with distinct experimentally characterized activities from a specific environment
	Discover novel DNA and RNA virus families from metagenome and metatranscriptome data	Provide global biogeographic maps of viral distribution
	Detect and categorize base modifications in uncultivated lineages	Discover novel base modifications in environmental DNA

Stretch Goals

- Single-molecule sequencing and detection of labeled community DNA (e.g. Bromodeoxyuridine [BrdU] or stable isotope probing [SIP])

- Model syntrophic networks of microbial assemblages and predict community behavior through the integration of multi-omics data in collaboration with KBase *

- Enable the identification of novel mechanisms of microbe-microbe and plant-microbe interactions *

- Enable the understanding and exploitation of molecular underpinnings of microbial plant growth promoting activities *

*Note: * joint Microbial/Metagenome Program goals*

Computation

Implementation Progress

Scientific Computing at DOE JGI

Computational processing and analysis of large genome-scale datasets is central to the success of the DOE JGI. Since 2012, the DOE JGI has collaborated with its partner, the National Energy Research Scientific Computing Center (NERSC), on several computing initiatives aligned with its Strategic Vision. There have been more than 20 computational training sessions and three bi-weekly study groups to help DOE JGI staff learn software best practices and high-performance computing (HPC) processes. Staff are currently exploring HPC methods for improving existing production pipelines in accordance with those best practices. In 2013, a centralized data archive and metadata organizer system, the JGI Archive and Metadata

Organizer (JAMO), was established across the DOE JGI. Pipelines were updated and some repetitive tasks have been automated.

Increased Efficiency and Algorithm Development

Several HPC initiatives were started in FY14, of which three have been completed to date. The focus areas for the completed initiatives were (1) the development of a scalable algorithm for plant assembly, (2) migration of DOE JGI pipelines from cluster hardware to the Cray supercomputers, and (3) improving computational efficiency of common bioinformatics algorithms. There is a new version of a DOE JGI short read assembler, Meraculous, which now scales to several thousand processors on Edison and thereby enabled a high-quality assembly of a large wheat genome. A set of production pipelines is ready to be moved to the Cray supercomputers and run using the allocation from BER, maximizing use of resources and increasing flexibility. Knowledge gained during this process has also been used by NERSC staff members to inform the configuration and hardware that will be set up for their next supercomputer, Cori, so it will be able to handle the DOE JGI's data-intensive workloads. These initiatives and coordination with NERSC enabled the DOE JGI to make significant progress on all of its strategic computing goals for Biological Data Interpretation.

Toward State-of-the-Art Scientific Computing

Massive improvements have been made to the DOE JGI computing infrastructure and operations. In the following two to five years we will continue to improve computational efficiency and expand our HPC initiatives to include work to support data accessibility and analysis through the DOE JGI web portals. A fully interoperable database system that allows users to query across plant, fungal and microbial databases will enable scientists to ask deeper questions.

In order for NERSC and other DOE computing facilities to move to exascale computing, the hardware and computing architectures must change. Bioinformatics tools and algorithms must be adapted so the DOE JGI can pursue its mission. The new version of Meraculous is a promising start in this algorithm development, however assembly is just one aspect of the DOE JGI workload, which is dominated by similarity search algorithms for quality control, comparative analysis, SNP calling and other applications. As such, we are interested in optimizing the pipelines and codes that use these methods. In 2015, the DOE JGI will explore the possibility of an ETOP proposal for the design of a fast, accurate similarity search tool that can run on the Cori architecture. This algorithmic development will be an important step towards the implementation of all-to-all comparisons across metagenomic data sets and eventually expose this as a capability for the DOE JGI user community.

KBase Initiatives

The push-to-KBase capability developed at DOE JGI can be strengthened to facilitate access to new and historic DOE JGI data. Furthermore, the development of a unified back-end for the DOE JGI databases will facilitate access to a larger number of datasets for analysis on the

KBase platform. As the capabilities of KBase develop, we will identify additional opportunities for collaboration.

Additional interactions between IMG, Phytozome and MycoCosm with KBase will be developed and in-demand older DOE JGI data will be upgraded with necessary metadata for easier compatibility with KBase. One step that DOE JGI is exploring is the modification of DOE JGI tools to the Docker suite of "wrappers" so that they can more easily be ported to other bioinformatics systems such as KBase.

Implementation Milestones

Area	Two years	Five years
Improved Computational Efficiency	All pipelines that use more than 10% of the Genepool cluster resources will be 10% more efficient	All pipelines will improve I/O efficiency by 20% (either in data storage or run time)
Migration to NERSC Supercomputers	Move BLAST workload to NERSC's Cori supercomputer	Port 75% of DOE JGI's analysis codes to the predominant NERSC architectures
	Prototype core bioinformatics algorithms using Many Integrated Core (MIC) architecture	Production use of core bioinformatics algorithms on future NERSC hardware using MIC architecture
Data Integration	Development of a prototype Global Environmental Database	Back-end integration of DOE JGI databases (plants, fungi, microbes) to allow common functionality and interoperability
Big Data Computations	Prototype new methods to enable all vs all data analysis	Support all vs all comparisons in a production environment to support all users
General Data Exchange and KBase Interactions	Prototype of a web Application Program Interface (API) to DOE JGI Data that will support KBase	Enable direct access to DOE JGI data via a web Application Program Interface (API)
	Make DOE JGI tools Docker-compliant	Enable ready capture of DOE JGI tools into other data portals, such as, but not limited to, KBase

Stretch Goals

- Run all DOE JGI compute on the NERSC-9 system in 2019. This stretch goal will drive the DOE JGI to be a leader in the biological computing space and provide access to analytical tools and capabilities that cannot be found at other sequencing centers

- Develop approaches for unstructured and/or novel complex queries of sequence and biological databases to take advantage of networks of genetic, biochemical, physiological, and environmental datasets associated with DOE JGI and other sequence data

Emerging Technologies Opportunity Program (ETOP)

Engaging Partners for Development of New DOE JGI Capabilities

The DOE JGI 2012 Strategic Vision described the transformation of the DOE JGI into a Next-Generation Genome Science User Facility. This will require an expansion of the DOE JGI's capability portfolio. While many of the required new capabilities can be developed in house, taking advantage of the DOE JGI's experience and track record in the development of high-throughput methods, some require specialized expertise not currently present onsite. As a key initiative for the implementation of the Strategic Vision, the DOE JGI launched the Emerging Technologies Opportunity Program (ETOP) in 2013. Through this program, the DOE JGI provides funding to external ETOP partners, who are typically leaders in the development and application of specific highly specialized technologies and approaches. The goal of these projects is to develop and implement cutting-edge new technologies specifically for the purposes of the DOE JGI. Depending on the project and technology, capabilities are either transferred to the DOE JGI or can potentially be made available to DOE JGI users through longer-term ETOP partnerships with external groups. Below, we describe the outcomes of two previous rounds of ETOP projects that started in FY14 and FY15, as well as the focus areas of a new round of ETOP projects that are currently being solicited and scheduled to commence in FY16.

Overview of Active and Completed ETOP Projects

Tools for Recovery of Microbial Genomes from Metagenomic Data

This ETOP project, led by PIs Jill Banfield (UC Berkeley) and Chongle Pan (Oak Ridge National Laboratory), was initiated in FY14 and aimed at the development of new methods for extracting near-complete and complete microbial genomes from complex metagenomic data sets. Through this project several experimental and computational approaches to obtaining higher-quality metagenome data were explored. Synthetic long-read data (Moleculo) were determined to be of limited use for whole genome assembly from metagenomes due to low sequence coverage, but of significant value in accessing rare genomes otherwise too poorly sampled for contig

assembly. Improved methods for error correction and assembly of Pacific Biosciences long-read metagenome data were developed, and approaches and pipelines for genome binning were tested and applied. This project ended as planned at the end of FY15.

High-Throughput Methods for Fungal Culturing and Nucleic Acid Isolation

This ETOP project, led by PI Jon Magnuson (Pacific Northwest National Laboratory), was initiated in FY14 and aimed at the development of efficient methods for culturing fungi under a variety of conditions as well as isolation of DNA and RNA from fungal cultures. Procedures for culturing fungi in microplates and measuring their growth on different substrates in a high throughput manner were developed along with methods for DNA and RNA isolation from very diverse fungi. The first genomes were sequenced and annotated. Robust and reproducible procedures for use of Promega's Maxwell 16 DNA/RNA extraction and purification instrument were developed and used to produce samples for DOE JGI in a high-throughput fashion. This project will be extended into FY16 through a no-cost extension. Furthermore, plans are underway to engage the Magnuson group as a longer-term ETOP partner for obtaining nucleic acids from fungal cultures for user projects.

Through the ETOP, Jon Magnuson's team from PNNL is developing flexible methods for high throughput fungal culturing.

Single-Cell Approaches to Metagenomics

This ETOP project, led by PI Stephen Quake (Stanford) was initiated in FY14 and was aimed at the development of a suite of microfluidic tools and single molecule genomic analyses, for potential implementation into DOE JGI workflows. The Quake group developed and benchmarked protocols that can be used on the Fluidigm C1 microfluidic platform, namely the

microfluidic mini-metagenome method and the single-cell yeast and fungal transcriptomic method. While the Fluidigm C1 microfluidic system provides simple sample handling, it was not fully compatible with DOE JGI's single-cell workflow, provided less overall flexibility and came at a greater cost. DOE JGI thus decided to maintain its existing single-cell and mini-metagenome workflow, which uses fluorescent-activated cell sorting and provides greater flexibility for the diverse set of environmental samples the DOE JGI is handling for its DOE user community. This project ended as planned at the end of FY15.

Dial-Out PCR for Retrieval of Sequence-Verified Synthetic DNA Molecules

This ETOP project, led by PI Jay Shendure (U Washington), was initiated in FY14 and aimed at the development of a "dial-out PCR" strategy for targeted retrieval of sequence-verified molecules from pools of synthetic DNA, with the goal to increase the efficiency of the DOE JGI DNA synthesis pipeline. During the course of the project comparable technologies were developed and productionized by synthesis companies such as Gen9 and Twist, which led us to re-focus the project to address other bottlenecks in our pipeline, namely the sequence verification steps. The Shendure lab developed a Nextera-based sequence verification protocol that works directly from colonies for rapidly assessing the sequence of strains carrying synthetic constructs. The protocol was transferred to the DOE JGI and was miniaturized using the Echo acoustic deposition platform to reduce cost to ~$1/library. It is expected that this new workflow will replace the current PacBio-based workflow in FY16Q3. The ETOP project itself ended as planned at the end of FY15.

Raman Microspectroscopy for Function-Based Isolation of Microbial Cells

This ETOP project, led by PIs Roman Stocker (MIT) and Michael Wagner (U Vienna), was initiated in FY14 and aims at coupling Raman microspectroscopy to microfluidics and flow cytometry in order to enable high-throughput sorting of microbial cells with specific functional traits for single cell genomics. During the first two years of this project, the Wagner group has benchmarked and developed a method for the Raman-based isolation of deuterium-labeled active cells. Using this method, the Raman instrument was fine-tuned to minimize spectra acquisition time and a software component for spectra interpretation was implemented. The MIT team has in parallel developed a microfluidics device for automated Raman-based sorting of samples. This project was continued for a third year into FY16. During this extension the developed microfluidics component will be benchmarked and the whole system will be tested on an environmental sample for full end-to-end integration. The third year will also focus on making the system more robust and user-friendly and transferring it to the DOE JGI, so it can be used for future DOE JGI user projects.

Plant Vouchering and Isolation of Plant Nucleic Acids

This ETOP project, led by PI Rod Wing (Arizona Genomics Institute, AGI), was initiated in FY14 and aimed at vouchering of plants studied by the DOE JGI, isolation of high-molecular weight DNA and RNA from plants, and generation of large-insert sequencing libraries. AGI has delivered high molecular weight DNA for twelve plant user projects and delivered three large-

insert BAC libraries for important reference projects. As part of this ETOP, AGI has developed high throughput 96-well DNA extractions from plant tissues that have been used for over 1,000 DNA extractions. AGI has also completed development on a high throughput RNA isolation protocol that is currently being tested on DOE JGI user projects. This project will be extended into FY16 through a no-cost extension. Furthermore, plans are underway to engage the Wing group as a longer-term ETOP partner for plant vouchering and isolation of nucleic acids for DOE JGI user projects.

Microfluidic Platform for High-Sensitivity DNA Library Construction

This ETOP project, led by PI Paul Blainey (MIT), was initiated in FY15 and aims at the implementation of ultra-low-input library preparation protocols on microfluidic platforms with the goal to reduce sample amount and increase throughput of DNA library preparation workflows. So far, we have established one functional prototype microfluidics system at DOE JGI and successfully performed routine whole genome shotgun (WGS) library construction, which largely completed the first milestone. We are aggressively evaluating the capability of this system in the single cell sequencing workflow to replace MDA for coverage and assembly improvement. This project is in progress and currently scheduled to end in FY16. Milestones for FY16 include the creation of automated workflows for transcriptome and epigenome analyses, as well as high impact scientific projects to demonstrate the microfluidics capabilities.

Future ETOP Projects

Building on the success of previous ETOP projects, the DOE JGI decided to release a new call for ETOP proposals in 2015, with the goal to start 1-3 new ETOP projects in early 2016. The focus areas for the FY16 ETOP call include:

- Methods for targeted sequencing of uncultivated microbes, such as:
    - High throughput single-cell isolation of specific phylogenetic groups
    - Techniques to physically enrich microbial consortia for genomic characterization
    - Automated or semi-automated density gradient centrifugations to enable large-scale SIP experiments for metagenomics
- Plant technologies to access single cell types of complex tissues for RNA-seq and related genomic assays
- Efficient computational approaches to genome-resolved metagenomics and/or reference-independent comparative metagenomics
- Infinitely (horizontally) scalable data stores to replace/supplement the existing monolithic relational database systems that support the metagenomic, microbial, fungal, and plant genomic portals and analytic workflows

- Development or adaptation of software for the Intel XeonPhi architecture to implement JGI-defined algorithms, e.g. fast, accurate and sensitive similarity search tools that can run efficiently on NERSC's Cori supercomputer
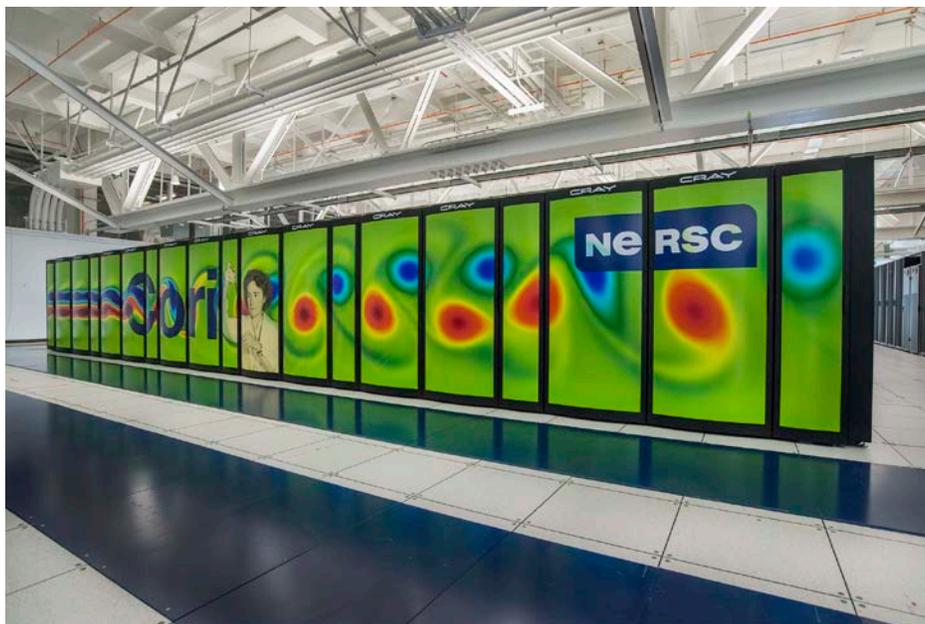
Berkeley Lab's recently opened Shyh Wang Hall houses the National Energy Research Scientific Computing Center (NERSC) and Cori, a new Cray XC40 supercomputer designed for data-intensive science.

JGI-EMSL Collaborative Science Initiative

In today's research climate, sequence data are rarely analyzed in isolation but are interpreted in the light of extensive additional molecular data and metadata. As a result, many DOE JGI users also make use of the type of molecular capabilities offered by another DOE User Facility, the Environmental Molecular Sciences Laboratory (EMSL). To facilitate such combined use projects, the JGI-EMSL Collaborative Science Initiative was launched in 2013. Successful proposals have combined DOE JGI and EMSL capabilities such as genomics, transcriptomics and proteomics of anaerobic gut fungi (PI: Michelle O'Malley); metatranscriptomics, metaproteomics and metabolomics of plant root-associated communities (PI: Mary Firestone); and genomics, transcriptomics, genetic manipulation via DNA synthesis and proteomics of cyanobacterial / moss symbioses (PI: Philip Weyman). We expect to continue supporting such innovative and multifaceted endeavors through this program in the future. Furthermore, the DOE JGI will explore the possibility of establishing similar joint programs in conjunction with other DOE User Facilities with the goal to provide an expanded portfolio of orthogonal capabilities that can be coupled to sequence-based analysis at the JGI through joint calls.

New Strategic Initiatives

In addition to the intensive dive into the DOE JGI's two- and five-year goals and stretch goals, the last third of the 2015 Strategic Planning Retreat also focused on brainstorming about distinct capabilities and products that the DOE JGI should seriously consider for the future. The external retreat participants largely led these discussions, presenting examples of applications from their own work and highlighting how these or related technologies may be useful additions to the portfolio of the DOE JGI to support the vision of a Next-Generation Genome Science User Facility. A particular focus of these sessions was the identification of both scientific questions and technology areas that may have been missed in prior strategic planning discussions or may have newly emerged since the last major strategic planning effort of the DOE JGI. While many of the suggestions and preliminary conclusions from these sessions are the subject of ongoing discussions, we provide here a summary of those that we consider to be the most promising areas that were discussed. We expect these to serve as nucleation points for further DOE JGI activities, particularly those that continue to resonate following further evaluation.

Syntrophy

Many organisms, and in particular microorganisms, critically depend on syntrophic relationships with other species to survive and thrive within their natural environments. These syntrophic relationships are critical drivers of community composition and impact on a variety of environmental processes, yet due to technical hurdles they have been difficult to explore systematically. The DOE JGI is considering developing a combined scientific and technological focus on being able to explore both experimentally and computationally syntrophic microbial complementarities and mutualisms.

As an initial step to explore this potential focus area in more detail, the DOE JGI plans to organize a workshop on this topic that will include scientists that have worked with relevant experimental systems and technologies. This will include both experimental and computational approaches. Providing these capabilities and exploring their applications through a targeted CSP focus area is a possible outcome.

Stable Isotope Probing (SIP) Capabilities

Stable isotope probing (SIP)-based assays offer a plethora of powerful approaches to identify and study organisms that are living and active in environmental samples. Several groups, including some represented at the 2015 Strategic Planning Retreat, have successfully developed and deployed SIP technologies as a means to selectively isolate active microbes metabolizing specific substrates from microbial communities. The overall conclusion of the retreat discussion was that this is an important technology to functionally screen microbes prior

to sequencing. While some of the required techniques may be challenging to establish, it is expected that they would be of major interest to the DOE JGI user community and complement the existing method portfolio for functional studies of environmental communities. Establishment of relevant methods at the DOE JGI may be facilitated via an ETOP project engaging external investigators as technology development partners.

All-Against-All Metagenomic Searches

Currently, metagenomic sequence comparisons are largely limited to reference-based searches, i.e. comparing the sequence of metagenomic data sets to a reference database of individual isolate genomes. In discussions at the 2015 Strategic Planning Retreat, there was extensive support for the notion that the ability to perform all-against-all searches, i.e. comparing metagenomes to metagenomes, will potentially enable transformative progress in the field and would be a capability of major interest to future DOE JGI users.

While new JGI/NERSC computation platforms and technologies may in principle put the development of such a capability within reach, many details remain to be resolved and substantial development effort will be required. Proposed next steps to facilitate the development of all-against-all searches for metagenomic data include focused interactions with NERSC, as well as consideration of an ETOP in this area.

Development of New User Communities

The need for the DOE JGI to continue to evolve its user community was discussed. The DOE JGI capabilities expected to become available over the next five years are anticipated to meet the scientific needs of new sets of users. At the 2015 Strategic Planning Retreat, there were extensive discussions about ways to increase the visibility of the DOE JGI and its evolving suite of capabilities. Activities identified as potentially effective include: 1. Workshops in new strategic focus areas such as secondary metabolites, 2. Targeted advertisements for the DOE JGI, directed specifically at investigators working in DOE-funded areas such as EERE, BER-funded Next-Generation Ecosystem Experiments (NGEE; arctic and tropics), as well as other BER-funded investigators, 3. Presence of DOE JGI at and on the programs of meetings such as Gordon Research Conferences and other DOE JGI-relevant meetings accompanied by presentations/advertisements for the programs.

Searchable All-Inclusive Environmental Database

Currently, the DOE JGI offers several separate sequence databases that are organized along its scientific program structure and contain microbial and metagenome, plant, and fungal data, respectively. While each of these databases is a world leading resource in its own area, there is currently limited integration and interoperability across these resources, impairing the users' ability to analyze data and work on projects spanning more than one program. As an example,

supporting projects studying plant-microbial interactions currently requires customized data integration efforts across different DOE JGI Science Programs.

To better support such integrative systems-level analyses as outlined in the 2012 DOE JGI Strategic Vision, it will be critical to offer DOE JGI data users a unified access portal with clear DOE JGI brand identity, as well as cross-program search capabilities. With its substantial existing data resources and platforms, the DOE JGI is uniquely positioned to become the provider of such a central "one-stop-shop" data portal that caters sequence data and primary analysis tools across the different domains of life to researchers in energy and environmental genomics. Steps required for implementation of this goal are currently being discussed in collaboration with the DOE KBase project in order to avoid duplication of efforts, enable users to move smoothly between systems, and facilitate integration of DOE JGI data and analysis resources with KBase and other external data sources and tools.

Additional Possible Future Directions

In addition to the Strategic Initiatives described above, several other possible future directions, both in the areas of science drivers and technologies were discussed at the retreat or in follow-up of the retreat. While these are not currently actively being pursued as strategic initiatives, the DOE JGI will evaluate them.

Develop a Defined Microbiome

Past attempts to study microbes and their interactions have focused on either experimental manipulation of isolated microbes, which provides limited insights into microbial interactions, or metagenomic approaches, which are typically limited to observational studies as the possibilities to experimentally manipulate complex communities within their environment are limited. As microbial community studies become a larger focus area for the DOE JGI, development of appropriate model systems for study is critical. One possible approach is isolating microbes from relevant sites and systems, characterizing them in isolation, and then building defined communities to explore the critical elements and how they interact. Availability of this approach would benefit a range of DOE JGI activities including studies of microbe-microbe and plant-microbe interactions, and the role of natural products in these interorganismal communications.

Pan-Genomics

Work on *Emiliania huxleyi*, as well as an increasing number of microbes, highlights that the species concept is becoming increasingly insufficient to capture the complexities of phylogenetic relationships and population structure in meaningful ways. Extensive sequence analyses can illuminate the composition of "pan"-genomes of a microbial, archaeal, or eukaryotic "cliques" and potentially help bound their genome space. Such data in turn could help define the interactions and activities of which such cliques are capable.

Microbial – Climate Interactions

While still limited in number, publications are appearing that suggest that microbes and fungi can contribute significantly to ice nucleation in clouds implying that there is a linkage between terrestrial (perhaps marine as well) microbial activities and climate processes. Since clouds move and deposit rainfall, this could even be a manner of microbial dispersal with the possibility of gene flow across significant distances. This is largely unexplored science.

Interactions with Other User Facilities

A very powerful opportunity that DOE JGI is already exploring and will continue to explore is the science that can result when several complimenting DOE User Facilities collaborate to "gang-tackle" a hard scientific challenge. DOE JGI is already working with EMSL but will continue to explore working with DOE funded light sources, among them the Advanced Photon Source at Argonne National Lab and the Linac Coherent Light Source at the SLAC National Accelerator Lab. Linking sequence analyses to proteomic analyses to protein structural and imaging analyses offers the prospect of going much deeper into the 3-D complexities of cellular architecture and functioning.

Further Ideas for Future Consideration

- Methods for improved coverage in single-cell genomes

- Manipulative metagenomics, i.e. controlled experimental changes of conditions

- Develop and apply genome editing techniques (CRISPR/Cas9)

- Citizen Science – Harnessing the power of interested amateur scientist communities, e.g. for large-scale collection of environmental samples

- Expanded outreach through social media platforms as a way to build communities

- Establishing anaerobic host systems, e.g. for archaea

- Develop Brainbow-like methods to visualize spatial aspects of microbial communities

- Screens for antifungal components

- Expand access to NERSC for DOE researchers with large-scale genome analysis needs

- Expand Tn-Seq phenotyping to additional types of conditions

- Network analyses to identify microbial interactions and pathways

- Single-cell targeting of "unusual" microbes

- Develop additional environmentally relevant model systems/engineered hosts

- Integrative omics analysis (genomics, transcriptomics, proteomics, metabolomics) of co-cultured organisms

- Spatiotemporal measurements of microbes and microbial communities

- Measure all data possible (sensor and imaging) on single-cell, cell-cell and community samples

- Use synthesis to generate full-length transcription factor libraries

- Targeted identification of differences between *Arabidopsis* and *Brachypodium*

- Identify and pursue large DOE-relevant biomanufacturing project

- Expand user support in DNA synthesis design and refactoring

- Develop a toolkit of molecular reporters

- Methyl-seq of populations

- Develop functionally rich reference maps for defining biological pathways

IV. Appendices

Retreat Participants

JGI

Adam Deutschbauer

Axel Visel

Chia-Lin Wei

Daniel Rokhsar

David Gilbert

Edward Rubin

Igor Grigoriev

James Bristow

Jan-Fang Cheng

Jeremy Schmutz

Kjiersten Fagnan

Len Pennacchio

Matthew Blow

Nikos Kyrpides

Ray Turner II

Rex Malmstrom

Samuel Deutsch

Shane Canon

Susannah Tringe

Tanja Woyke

Tootie Tatum

Trent Northen

Yasuo Yoshikuni

Zhong Wang

External Participants

Adam Arkin – LBNL

Brenda Andrews - University of Toronto

Deirdre Meldrum - Arizona State

Gary Stacey - University of Missouri

Jeffery Dangl - University of North Carolina

Jennifer Pett-Ridge - Lawrence Livermore Nat'l Lab

Jonathan Eisen - UC Davis

Mark Adams - J. Craig Venter Institute

Samuel Hazen - University of Massachusetts Amherst

Sarah Grant - University of North Carolina at Chapel Hill

Victoria Orphan - CalTech

Evening Speaker

Pascal Lee - NASA/SETI

Retreat Agenda

Monday, April 13th

9:30a-9:40a *Eddy Rubin:* Welcome and Introduction

9:40a-10:00a *Axel Visel:* Strategic Planning at JGI – History and Perspective

10.00a-5.00p **Implementing the 2012 Strategic Vision**

Format for each 40min session:

- 10-15min: Overview Presentation by ***JGI Speaker(s)***

    o Relevant goals in Strategic Vision

    o Progress towards their implementation

    o Proposed specific two-year/five-year implementation goals

    o Possible adjustments/changes in strategic directions

- General Discussion (everybody):

    o Is JGI doing the right things to implement the 2012 vision?

    o How can we track our success towards implementation of the vision?

    o Are the proposed two-/five-year implementation goals relevant, meaningful and realistic measures of success? If not, what else?

    o Generally, what areas described in the 2012 vision should continue to be high priority, what should be de-emphasized?

Session (JGI speakers):

10.00a-10.40a   Microbial **(Tanja Woyke)**

11.00a-11.40a   Metagenome **(Susannah Tringe)**

11.40a-12.20p   Plants **(Jeremy Schmutz)**

1.20p-2.00p   Computation **(Dan Rokhsar/KjierstenFagnan/Shane Canon)**

2.40p-3.20p   Synthesis **(Sam Deutsch/Yasuo Yoshikuni)**

3.40p-4.20p   Fungi **(Igor Grigoriev)**

4.20p-5.00p   Functional Genomics **(Matthew Blow/Adam Deutschbauer/Trent Northen/Len Pennacchio/Chia-Lin Wei)**

7:30-8:30   After-Dinner Talk: ***Pascal Lee*** - Mission to Mars

Tuesday, April 14th

9:30a-12:30   New Science, Technologies and User Communities

Format for each 1 hour session:

- Very brief intro by **Jim Bristow** about 4 major science emphasis areas

- 2-3 "flash talk" presentations (5min each) by **external speakers**

- General Discussion (guided by JGI chairs)

    o   What grand scientific challenges within each of these science themes should JGI aim to tackle over the next 10 years?

    o   How would JGI engage relevant new user communities?

    o   What fundamentally new technologies relevant to this area should JGI invest in?

9.30a-10.30a   ***Phylogenetic Diversity and New Life***

JGI session chair: **Tanja Woyke**

External presenters: **Jonathan Eisen, Victoria Orphan**

10.30a-11.30a   ***Microbe-Microbe and Microbe-Plant Interactions***

JGI session chairs: **John Vogel, Susannah Tringe**

External presenters: **Jeff Dangl, Sarah Grant, Jennifer Pett-Ridge**

11.30a-12.30a   ***Synthetic Biology and Secondary Metabolites***

JGI session chairs: **Sam Deutsch, Yasuo Yoshikuni**

External presenters: **Mark Adams, Sam Hazen**

1.45p-2.45p   ***Sequence-to-Function Capabilities***

JGI session chair: **Len Pennacchio**

External presenters: **Gary Stacey, Brenda Andrews, Deirdre Meldrum**

2:45p-5:00p   OPEN DISCUSSION

Questions to guide the discussion:

- What general science themes (grand scientific challenges) not currently covered by the Vision but within JGI/BER mission and JGI's broad capability portfolio should JGI aim to tackle over the next 10 years?

- What existing or nascent communities doing JGI/BER-relevant science but currently not interacting with JGI would benefit from having access to JGI capabilities or JGI's help in building these communities?

- What fundamentally new technologies should the JGI invest in to enable new science/user communities?

- Potential ETOP projects/partners?

- What would you like to do with the JGI two or five years from now?

- How do we prioritize with a finite budget?

5.00p END OF RETREAT

Retreat Group Photo

Front: Rex Malmstrom, Jeremy Schmutz, Dan Rokhsar, Kjiersten Fagnan, Jonathan Eisen, Sam Deutsch, Jennifer Pett-Ridge, Jeff Dangl

Middle (standing): Tootie Tatum, Sam Hazen, Gary Stacey, Matt Blow, Igor Grigoriev, Brenda Andrews, Shane Canon, Victoria Orphan, Deirdre Meldrum, Susannah Tringe, Adam Deutschbauer, Jan-Fang Cheng, Sarah Grant, Natalia Ivanova, Yasuo Yoshikuni, Nikos Kyrpides, Tanja Woyke, Axel Visel, Zhong Wang, Melissa Trevizo

Back (on structure): Mark Adams, Len Pennacchio, Ray Turner, Trent Northen, Jim Bristow, Adam Arkin, Eddy Rubin

Abbreviations

ANI	Average Nucleotide Identity
ANME	anaerobic methanotrophic archaea
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
BC	Biosynthetic Cluster
BER	Office of Biological and Environmental Research
BONCAT	bioorthogonal noncanonical amino acid tagging
BRC	Bioenergy Research Center
BrdU	Bromodeoxyuridine
ChIP-seq	Chromatin Immunoprecipitation followed by sequencing
CSP	Community Science Program
DOE	Department of Energy
EERE	Office of Energy Efficiency & Renewable Energy
EMSL	Environmental Molecular Sciences Laboratory
ETOP	Emerging Technologies Opportunity Program
FY	fiscal year
Gbp	billion basepairs
HPC	High Performance Computing
IMG	Integrated Microbial Genomes (a DOE JGI data portal)
JAMO	JGI Archive and Metadata Organizer
JGI	Department of Energy Joint Genome Institute
kbp	thousand basepairs
LDRD	Laboratory Directed Research and Development
Mbp	million basepairs
NERSC	National Energy Research Scientific Computing Center
NGEE	Next-Generation Ecosystem Experiments
PNNL	Pacific Northwest National Laboratory
Pbp	quadrillion basepairs
SAG	Single-Cell Amplified Genome
SIP	Stable Isotope Probing
Tbp	trillion basepairs
Tn-Seq	Transposon saturation mutagenesis followed by sequencing
WGS	Whole genome shotgun