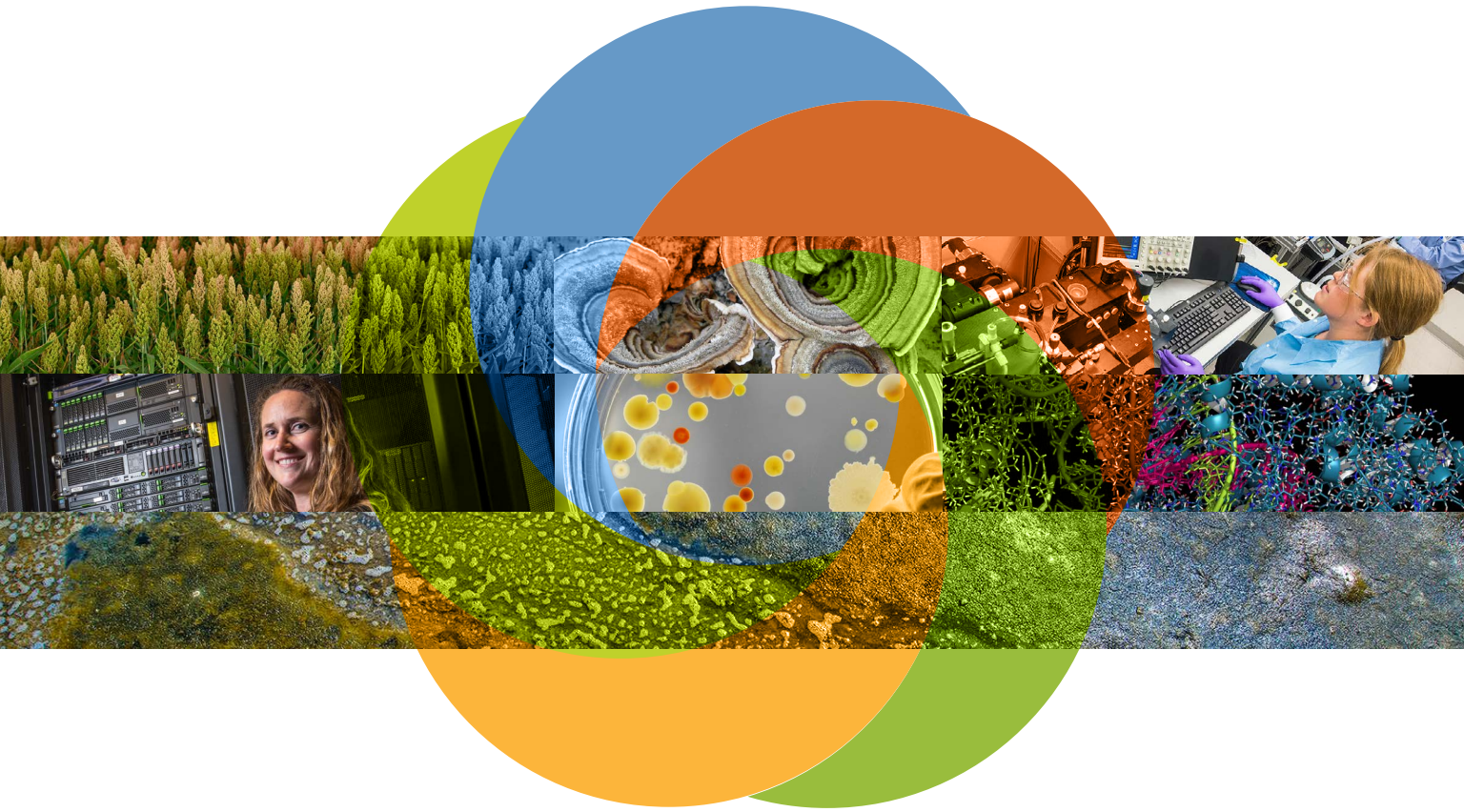
5-Year Strategic Plan

U.S. Department of Energy
Joint Genome Institute

December 2018

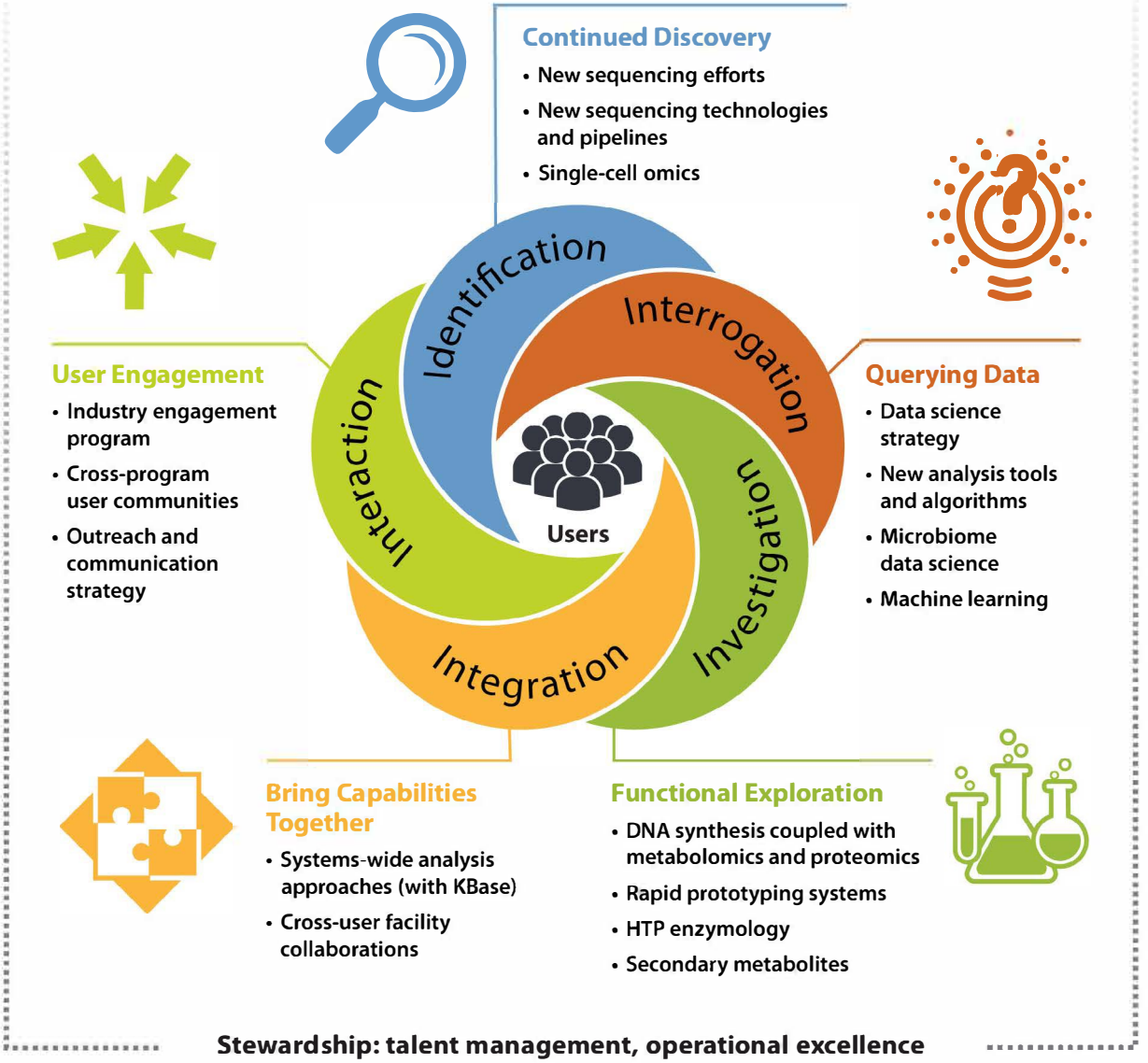
Beyond Basepairs

A Vision for Integrative and Collaborative Genome Science



JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

An Integrative Genome Science User Facility



Stewardship: talent management, operational excellence

Table of Contents

Vision

The leading integrative genome science user facility enabling researchers to solve the world's evolving energy and environmental challenges

Mission

To provide the global research community with access to the most advanced integrative genome science capabilities in support of the DOE's research mission

Executive Summary

The U.S. Department of Energy (DOE) Joint Genome Institute (JGI) provides cutting-edge genomic capabilities to a worldwide community of researchers addressing the most pressing scientific challenges related to energy production and environmental systems within the mission of the DOE Office of Science. Since its inception in 1997, the JGI has pioneered the development of new fields of genomic science, generated almost 1 quadrillion base pairs of sequence data, and served thousands of users.

Today's pace of sequence data generation outweighs our abilities to ascribe function and derive biological insights. Major breakthroughs in advancing our understanding of biology are thus only possible through the integrative use of complementary technologies. To ensure the JGI remains relevant and enables its users to make transformative scientific discoveries, it is imperative that the JGI evolves towards more integration. Therefore, we present a new vision for the JGI as an Integrative Genome Science User Facility that brings together the JGI's core capabilities in DNA sequencing and synthesis, high-performance computation, and metabolomics, and draws upon multidisciplinary partnerships with other user facilities to enable insights into the biological functions of genomic sequence. The JGI is well positioned to address specific challenges in the areas of biological systems science, such as plant and microbial metabolism and interaction, engineering of diverse organisms, integrative analyses of complementary data sets, and establishing experimentally validated links between genotype and phenotype.

Our **Vision** is to be the prominent genomic sciences facility that attracts users to access our multidisciplinary state-of-the-art technologies and to provide expertise that allows our users to harness complementary capabilities to answer their most pressing scientific questions related to energy and environmental challenges.

Our **Mission** is to make these technologies and expertise readily available to our users through peer-reviewed user programs, collaborative and sponsored projects, and user-friendly data portals. We work with our users to ensure we have the most advanced capabilities to enable their science in the best possible ways.

The implementation of this vision will be facilitated by the **I5 Strategic Framework**, which encapsulates a set of overarching guiding principles, each of which is applicable to a wide range of scientific activities, platforms, and user groups.

Identification: The JGI will harness cutting-edge sequencing-based methods to enable discovery of new organisms and biological functions from the analysis of bacterial, archaeal, viral, fungal, algal, and plant genomes as well as environmental metagenomes and single cells.

Interrogation: The JGI will develop and employ advanced and accessible computational approaches in partnership with the National Energy Research Scientific Computing Center (NERSC), the DOE Systems Knowledgebase (KBase), and the DOE Energy Sciences Network (ESNet) to enable the integrative analysis and interpretation of multiple orthogonal data types including transcriptomic, epigenomic, and metabolomic results in conjunction with genomic information.

Investigation: The JGI will expand its experimental capabilities for investigating and validating the biological function of sequence through further growth of DNA synthesis, metabolomics, and high-throughput functional genomic capabilities and introduce new approaches, such as rapid cell-free and cell-based prototyping systems.

Integration: The JGI will broaden the integrative programs with other user facilities that offer complementary research capabilities to further enhance the opportunities for its users to make groundbreaking discoveries in the DOE mission space.

Interaction: The JGI will lead scientific communities comprised of researchers from both the academic and private sectors in foundational and use-inspired research and facilitate communication and education within and between them.

In addition to our major scientific programs, we have identified two scientific areas for which we believe our integrative approaches will be especially powerful. These scientific strategic thrusts, microbiome data science and secondary metabolites, will be new focus areas for the JGI and complement our existing scientific programs.

This vision will be stewarded through a set of specific two- and five-year milestones which will facilitate scientific, organizational, operational, and budgetary decision-making processes, and enable tracking and reporting of the overall completion status of this vision. We will place increased emphasis on the diversity, development, and recognition of our existing talent and attract new talent. The JGI will also continue to focus efforts on operational excellence and financial stewardship, especially with its move to the Integrative Genomics Building (IGB) in 2019. We will continue to ensure that safety and diversity, equity and inclusion remain core institutional values of the JGI.

1. An Integrative Genome Science User Facility

A New Vision for the Joint Genome Institute

The Joint Genome Institute (JGI) continues to be the U.S. Department of Energy (DOE)'s flagship user facility for providing advanced genomic capabilities to a worldwide community of researchers addressing the most pressing questions within the mission of the DOE Office of Science.

Over the past two decades, the JGI has continually evolved, from a major sequencing center supporting the DOE's mission in the Human Genome Project, to a production sequencing facility supporting users in the generation of sequence data for energy and environmental research, to a Next-Generation Genome Science User Facility focused on the adoption of advanced DNA sequence and synthesis-based methods aimed at translating sequence to function (Fig. 1.1).

Building on this trajectory of successful transformations, this strategic plan describes our vision for the next phase in the history of the JGI, which is to become an Integrative Genome Science User Facility. This new vision is founded on a continuation of core principles and directions described in our previous strategic plan, such as a shift in emphasis away from merely producing large amounts of sequence data, towards the adoption of new experimental and analytical approaches that elucidate the function and thus the biological meaning of data. However, these guiding principles are now contained within an expanded set of strategic principles, captured in the **I5 Strategic Framework**. The overarching goal of

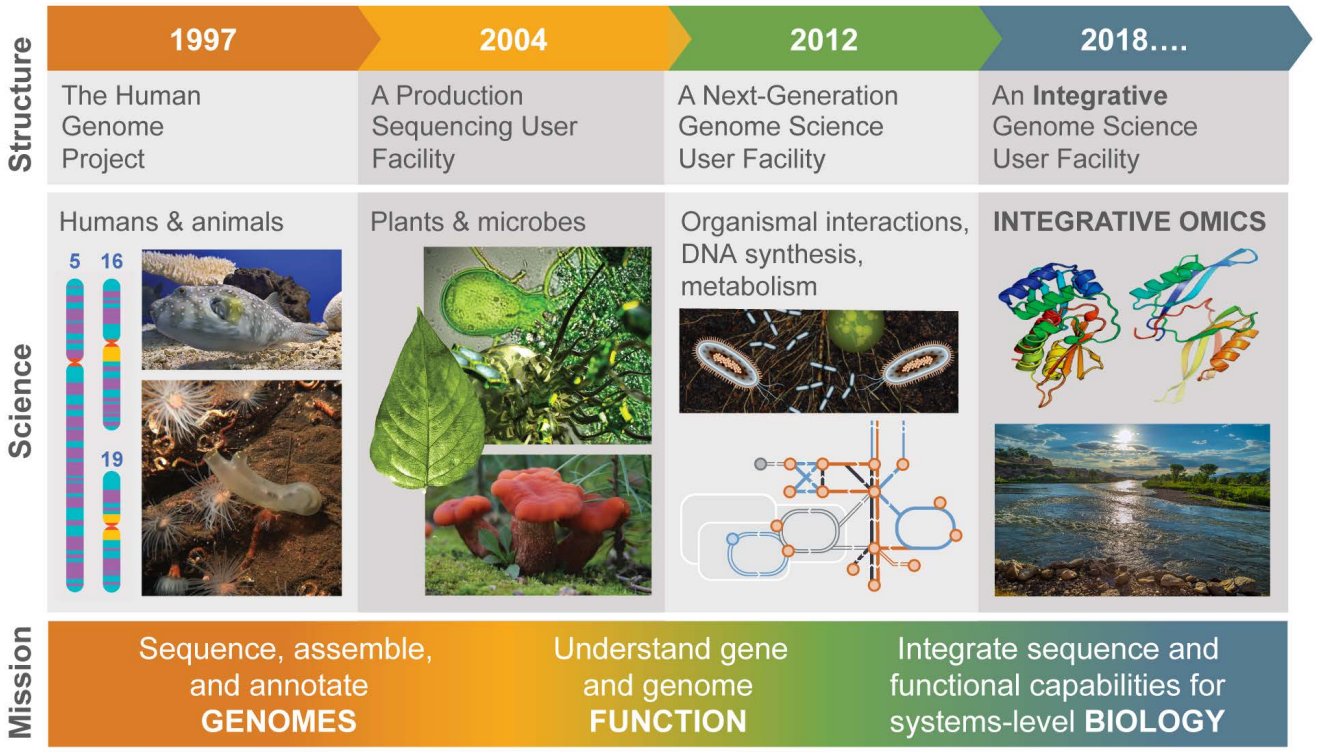


Figure 1.1. Phases of evolution of the JGI.

this framework is to enable JGI users to access a suite of "integrative omics" capabilities that is not limited by the scientific and organizational boundaries of the JGI, and facilitates access to the entire realm of capabilities available at the JGI and other DOE user facilities, with the goal to maximize the impact of DOE mission-relevant user science. Furthermore, the JGI's move into the new Integrative Genomics Building (IGB) on the Lawrence Berkeley National Lab (Berkeley Lab) campus in 2019 will foster stronger partnerships with the DOE Systems Knowledgebase (KBase), the National Energy Research Scientific Computing Center (NERSC), and the other areas/divisions at Berkeley Lab.

This vision will be stewarded through a set of specific two- and five-year milestones, which will facilitate scientific, organizational, operational, and budgetary decision-making processes, and enable tracking and reporting of the overall completion status of this vision. We believe that this time frame is appropriate given the rapid technology advancements in our field and the evolving landscape of energy and environmental research. This time frame also aligns with the DOE review cycles of the JGI, which will facilitate external assessment of implementation progress.

A Strong Team in Support of User Science

User facilities are a primary avenue through which the DOE enables scientific and technological progress towards solving the most pressing questions related to energy and the environment. They offer large-scale capabilities that would be impractical for any single institution to invest in or to house. Designated user facilities, such as the JGI, are typically purpose-built and feature an open access operating mode to accelerate advancement of science and technology to meet DOE mission needs. To encourage innovation and the exploration of new scientific knowledge, the DOE fully supports the design, construction, and operational costs of these facilities and does not charge access fees as long as researchers plan to openly publish their results and make available their data.

The JGI was established in 1997 by the DOE Office of Biological and Environmental Research (BER) with the goal of sequencing three human chromosomes totalling about 330 million bases, or 11% of the human genome. Since the successful completion of the Human Genome Project, the JGI has evolved first into a production sequencing facility focusing on microbes and plants and later into a genomic science user facility combining sequencing, functional genomics, synthetic biology, metabolomics, and advanced computational analyses to serve a global user community in creating biological insights. In fiscal year 2018, the JGI sequenced more than 200 **trillion** bases, an increase of more than seven orders of magnitude compared with its output during the Human Genome Project. This was possible due to the rapid advances in sequencing technologies that have dramatically reduced cost and increased throughput. Today, whilst remaining a large-scale sequencing facility, the JGI has integrated a suite of technologies and approaches all aimed at addressing the gap between genomic sequence and functional biological understanding. The JGI is part of Berkeley Lab and its Biosciences Area, facilitating alignment of activities and close interactions and collaborations with Berkeley Lab scientists.

While the scientific and technological directions of the JGI continue to evolve, as outlined throughout this document, a key approach underlying the success of the U.S. national laboratories remains central to the JGI's operation: that harnessing transdisciplinary and diverse teams of researchers and engineers is necessary to solve the most difficult scientific and technological questions. This approach, enabled through the unique organizational structure of the national laboratories and the national user facilities they house, was key to massive-scale endeavors such as the Human Genome Project. The collaborative spirit established through these successes continues today at the JGI, where a large team of scientists, technologists, and operational support staff work hand in hand, guided by scientific proposals submitted by the user community, to complete scientific projects that would be difficult or impossible for individual scientific investigators to tackle. In recognition of this well-established principle, talent management has been elevated to become a strategic priority at the JGI, as described in the section "**Stewarding the Strategic Plan — Talent Management Strategy**," see page **71**.

An Inclusive Visioning Process

The JGI approaches strategic planning as an inclusive process, tapping into the vast scientific and technological expertise of its staff, its advisory committees, its user communities, and the broader scientific community to develop strategic directions that ensure the JGI focuses on the most pressing questions of relevance to the DOE mission that can be addressed using genomics-based capabilities.

As described in the "Strategic Drivers" section, a main reference point for identifying grand challenges that the JGI will address is the 2017 report from the DOE Biological and Environmental Research Advisory Committee, entitled *Grand Challenges for Biological and Environmental Research: Progress and Future Vision*. This report outlines an exciting vision for how BER may align its scientific portfolio and its user facilities to support research related to energy production and environmental systems. In particular, this report identifies specific challenges in the areas of biological systems science that the JGI may be well positioned to address, such as studies of the biological complexity of plant and microbial metabolism; the development of engineering possibilities in bacteria, fungi, archaea, viruses, plants, and mixed communities; the integrative analysis of "omics" and complementary biochemical and biophysical data sets; the establishment of links between genotypes and phenotypes in single organisms and communities; and the adoption of new and emerging technologies, such as miniaturization, to accelerate biological discoveries. Additional challenges in scaling from microbial to Earth systems also link to JGI capabilities, such as characterizing microbial communities across space and time and building trait-based models starting with genome data. The report also calls out relevant opportunities in analytics and data science, and highlights specific emerging technologies that BER seeks to expand, such as single-cell methods and metabolomics, to enable progress towards these grand challenges.

To explore how these challenges might be best addressed by the JGI, a large number of external experts in relevant areas of science and technology were consulted. Input collected from the diverse JGI advisory committees over the past two years was taken into account, including the Scientific Advisory Committee; the Informatics Advisory Committee; the Prokaryote, Fungal, and Plant Program User Advisory Committees; and the DNA Synthesis Science User Advisory Committee. As a focused effort to support the development of the present plan, in April 2018 the JGI convened a strategic planning retreat that brought together 25 members of the JGI leadership, management, and scientific staff and 15 external experts with relevant scientific, technological, and organizational expertise. Over the course of two days, this group extensively discussed all aspects of the JGI's strategic directions described in this plan, and external participants in this retreat provided input for the present document. Finally, input from staff all across the JGI was solicited and considered, relying on the substantial expertise and experience available internally.

The present plan builds on the *JGI Strategic Planning Update*, published in 2016. The 2016 update contained more than 100 specific strategic milestones with a two- or five-year completion horizon. Continuous tracking of the progress towards implementation of these milestones provided a critical tool for realizing the vision articulated in the 2016 planning document. At this point, the majority of the two-year milestones, as well as a substantial number of five-year milestones, have been completed. Examples of successful implementation of goals articulated in the 2016 *Strategic Planning Update* include a major expansion of the JGI's single-cell sequencing capabilities for studies of phylogenetic diversity; the establishment of in-house model plant growth capabilities; the development and application of tools for terabase-scale metagenome datasets; and the move of major compute workloads to NERSC's Cori supercomputer. Following this successful implementation model, the present plan provides a new set of milestones (see **Appendix I: Implementation Milestones**, see page **77**). During the 2018 planning process, all milestones whose implementation is still in progress were revisited and their continued relevance evaluated. In most cases, we decided to continue to pursue our original directions, and updated the milestones quantitatively and/or qualitatively to align them with our revised strategic directions. In addition, entire new sets of milestones were added to provide a robust framework for implementation of new directions. We will continue to use these milestones to facilitate scientific, organizational, operational, and budgetary decision-making processes, and to track and report overall completion status of this vision through regular assessment of completion status of individual milestones.

The I5 Strategic Framework

To facilitate progress towards the vision of the JGI as an Integrative Genome Science User Facility, the JGI has implemented the **I5 Strategic Framework**. The five cornerstones of this framework are **Identification**, **Interrogation**, **Investigation**, **Integration**, and **Interaction** (Fig. 1.2). Each of these terms encapsulates a set of overarching guiding principles, each of which is applicable to a wide range of scientific activities, platforms, and user groups. These principles transcend any single area of science the JGI pursues, any single user community the JGI works with, any single technological platform the JGI offers to its users, or any single organizational unit within the JGI. Taken together, the five guiding principles of the **I5 Strategic Framework** are expected to enable, catalyze, and accelerate the successful transformation of the JGI as it enters a new chapter as a user facility.

Identification

Identification captures a set of goals that are historically at the core of the JGI's mission. In particular, the JGI will continue to harness cutting-edge sequencing-based methods in order to enable discoveries of relevance to the DOE mission. Sequencing enables the discovery of new biological functions from the vast sequence space that becomes accessible through the analysis of bacterial, archaeal, viral, fungal, and plant genomes as well as environmental metagenomes and single cells. Studies of pangenomes of natural microbial and wild plant populations, as well as sequencing of diverse plant

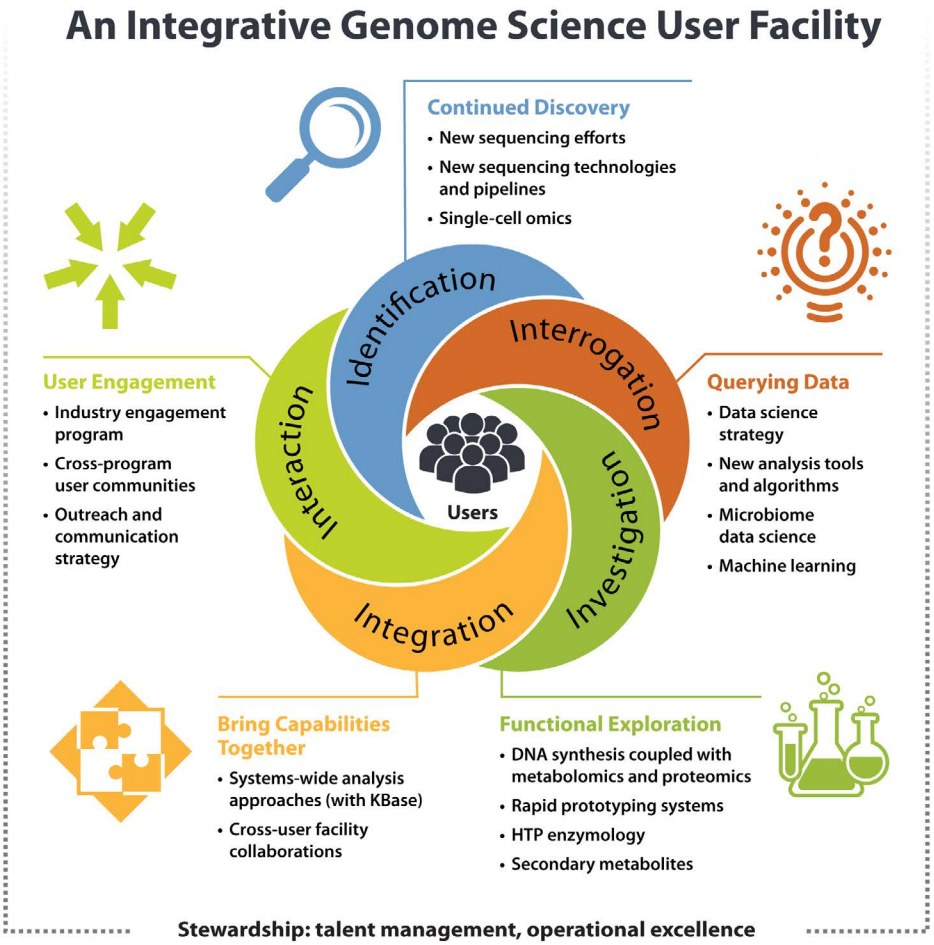


Figure 1.2. The I5 Strategic Framework. This framework will guide the JGI's transformation into an Integrative Genome Science User Facility.

populations in conjunction with careful phenotyping under controlled conditions, have begun to provide powerful tools for linking genetic and genomic variation to phenotypes relevant for the improvement of bioenergy feedstocks. Sequencing also makes it possible to identify "biological dark matter," such as the vast majority of environmental microbes and viruses that cannot be cultured and therefore have evaded discovery by conventional microbiological and virological methods. For this purpose, it is essential to further develop approaches for the sequencing and analysis of single prokaryotic cells. These methods are increasingly complemented by "function-driven" pre-sequencing steps that can enrich for organisms of interest prior to sequencing. Recognizing these important developments over the past years, this strategic vision provides a set of bold milestones for each of the branches of life, with the goal to continue to enable sequencing-based identification of new organisms and biological functions.

To support this vision, the JGI will continue to aggressively push the boundaries of sequencing technology. In particular, we will continue to tap into commercially developed high-throughput short-read platforms that provide the largest sequence output at the lowest cost and maximize throughput and cost efficiency for many types of sequencing projects. However, recognizing the unique set of challenges that the JGI faces in the generation and assembly of phylogenetically diverse genomes from DOE-relevant organisms, whose genomes often have highly unfavorable sequence and structural properties, the JGI will also continue to expand its effort in bringing complementary technologies on board. This includes long-read sequencing by synthesis technologies, which produce reads of high value for genome assembly, but also nascent technologies, such as nanopore-based approaches. Complementary to sequencing platforms, **Identification** goals also include the development of new technology platforms such as DNA synthesis, single-cell analysis, and strain engineering pipelines with the goal to provide scaled, user-accessible platforms to enable biological **Investigation** (see the "**Investigation**" section, this page).

Interrogation

The output of sequencing methods, and along with it, the amount of sequence that the JGI and its users produce in any given year, continue to grow dramatically. This sheer volume of data now necessitates the use of advanced computational approaches to enable meaningful interpretation. Even relatively simple types of analyses, such as sequence comparisons, can become challenges when they need to be performed across billions of individual sequences and trillions of base pairs in parallel. To add further to these complications, the JGI and its partners now provide capabilities that go significantly beyond the sequencing, assembly, and annotation of reference genomes. It will be necessary to integrate genome sequence in meaningful ways with transcriptomic, epigenomic, and metabolomic information and other orthogonal data types.

Fortunately, recent years have seen an unprecedented pace of innovation in the area of data science, including the development of exciting new methods for the application of artificial intelligence (AI) and machine learning approaches to "big data." Many of these approaches are applicable to the genomic and complementary omics data sets that the JGI and its users produce. Through its partnership with NERSC and the DOE Energy Sciences Network (ESnet), the JGI has unique access to both expertise and infrastructure to facilitate these efforts. Particular areas of emphasis will be the development of tools and infrastructure for microbiome data science, and the development of new informatics tools specifically optimized for high-performance computing environments. We will also make these tools portable so the software can be executed on any system, and will adhere to available data standards to facilitate data access and integration.

Investigation

While sequencing provides a rapid and largely unbiased route to the discovery of new proteins and pathways in organisms and environments of interest, in isolation it is often insufficient to reveal their biological function. This is particularly the case for proteins and pathways from phylogenetically distant organisms, and for proteins whose structure does not resemble any previously identified proteins. However, in many cases these novel proteins and pathways are those that hold the largest innovation potential, as they may potentially unlock access to novel biochemistry and metabolites, or provide insight into higher-order organismal functions.

The JGI has already begun to pursue several avenues to complement sequencing by experimental approaches for investigating the biological function of sequence.

A DNA Synthesis Science Program and an associated DNA Synthesis production platform are now firmly established, and a growing number of high-impact publications by users taking advantage of these capabilities illustrates the value of this effort. The JGI has also established a high-throughput metabolomics platform for functional genomics that is now routinely used, often in conjunction with DNA synthesis, to explore novel metabolism and, in particular, secondary metabolites important to environmental processes. The JGI has also utilized high-throughput functional genomic approaches, such as transcriptomics, epigenomics, resequencing of mutant populations, and barcoded transposon saturation mutagenesis of bacterial strains, to assign biological function to genomic sequence. Moving forward, the JGI will continue this successful thrust and further enhance its **Investigation** capabilities through the introduction of new approaches, such as rapid prototyping systems, including cell-free transcription-translation systems for high-throughput enzymology and pathway prototyping, as well as the development of analysis strategies and tools to correlate data obtained through these diverse approaches.

Integration

While the JGI offers a vast array of capabilities and massive throughput, in particular in the area of sequencing, it does not exist in isolation. It is only one part of the BER-funded portfolio of research activities and user facilities, many of which are highly complementary to the JGI's activities. Thus, better integration with these efforts can further enhance the value of the JGI to researchers aiming to make discoveries in the DOE mission space. This is best illustrated by the highly successful collaboration of the JGI with its BER-funded sister user facility, the Environmental Molecular Sciences Laboratory (EMSL). A joint user program between the JGI and EMSL (FICUS), established in 2013, has facilitated coordinated access to sequence-based capabilities offered by the JGI, and advanced molecular characterization methods offered by EMSL, through a single proposal. Building on this successful model, more recently a FICUS program jointly offered by the JGI and NERSC has been established to support microbiome data science projects. As outlined in this vision, the JGI aims to further broaden these integrative programs. In particular, this will include efforts to offer programs together with other experimental user facilities, such as the X-ray light sources supported by DOE Basic Energy Sciences. It will also increase integration with the DOE Systems Biology Knowledgebase (KBase) to enable systems-wide analysis approaches of data produced by the JGI and others. This will be performed in conjunction with efforts to further improve JGI-internal integration across currently disparate computational resources.

Interaction

The JGI continues to be first and foremost a user facility, catering to the needs of its users by providing capabilities and expertise as required by its existing user communities. However, the JGI will continue to take an active approach in developing and organizing these user communities. This philosophy is based on the historic recognition that entire user communities have formed around transformative resources and capabilities initially provided by the JGI, such as flagship genomes or single-cell sequencing methods. Thus, the JGI's Science Programs will continue to be present at the forefront of science in their respective domains, and provide leadership for communities. Besides technological innovation, the JGI is in a particularly favorable role to guide the development of data standards by communities, which are critical to collective progress.

While historically the majority of the JGI's users have been from the academic or government sector, there is an increasing recognition of the opportunities arising from closer interactions with partners and users from industry. In light of the JGI's leadership in many methods that are of topical interest not just for foundational science by academic users but also to use-inspired research in the commercial sector, the JGI has established an Industry Engagement Program (IEP) through which it will increasingly interact with industry partners and through a portfolio of different modes of engagement, will make its capabilities and expertise available.

Finally, the JGI recognizes the importance of outreach and communications to inform the public and the JGI's stakeholders about the impact of tax-funded user science at the JGI on the advancement of science and the development of solutions of national relevance. In order to support this effort, the JGI has now put forward a specific set of strategic goals related to communication and outreach.

2. Strategic Drivers

To remain relevant as a user facility, the JGI must provide cutting-edge technologies with applications to DOE mission-relevant science. Thus, the JGI's strategic directions are heavily influenced by technological advances as well as shifts in DOE priorities. These changes have driven the JGI's evolution from a sequence production facility to a Next-Generation Genome Science User Facility. However, the ever-growing pace of sequence data generation continues to widen the gap between our abilities to read and ascribe function to sequence and derive biological insights. Major breakthroughs in advancing our understanding of biology increasingly require the integrative use of complementary technologies. To ensure the JGI remains relevant and enables its users to make transformative scientific discoveries, it is imperative that the JGI evolves towards more integration. To address this critical need, this new strategic plan describes the next phase of the JGI's transition towards an **Integrative** Genome Science User Facility.

Technological Drivers

Since the JGI's inauguration as a user facility, DNA sequencing has dramatically increased in throughput and decreased in cost. This has enabled not only the sequencing of more genomes, but explorations of evolution, population structure, and organismal associations not possible without broad and deep sequencing. The increase in sequencing capacity has been parallelled by advances and innovations in library construction and functional genomics technologies, both at the JGI and elsewhere. These have expanded the JGI toolkit to include diverse techniques in genome, epigenome, transcriptome, and metagenome sequencing. Next-generation DNA sequencing has also become a capability within easy access of most academic labs, meaning investigators come to the JGI not simply for sequence data but for the many complementary experimental and analytical capabilities offered. This toolkit should continue to expand to meet users' needs and pursue cutting-edge science.

Single molecule sequencing methods have also advanced considerably in recent years, including well-established sequencing-by-synthesis methods as well as rapidly expanding nanopore-based, synthesis-independent sequencing, which aspires to enable sequencing in every laboratory and even at field sites. The JGI has leveraged long-read technologies to produce high-quality de novo genomes for bacteria, archaea, fungi, and plants. As these capabilities become more portable and ubiquitous, the role of a genome center needs to be rethought. The JGI will continue to focus on producing high-throughput, high-quality data demanding the resources, expertise, and capabilities of a dedicated user facility.

New bioinformatics methods and computing techniques have dramatically improved our ability to analyze and interpret sequence data, and the growing field of data science has presented new and creative applications for sequence that in turn drive data generation needs. Increasing volumes of data and diversity of data types have demanded a rethinking of the ways data are processed, analyzed, presented, and stored. The JGI's partnerships with NERSC and KBase position it for leadership in biological data science.

Beyond sequencing, advances in DNA synthesis have driven the JGI to build capabilities in gene and pathway synthesis and novel host organism engineering that allow for direct functional testing of sequence-based hypotheses. Similarly, mass spectrometry-based metabolomics has opened the door for directly measuring metabolic activities, investigating secondary metabolites and microbial food webs, and profiling enzyme activities via targeted assays. By offering all these technologies through user programs, the JGI can advance systems-level interrogation of microbes, fungi, and plants in their natural environments. The portfolio of non-sequencing capabilities will continue to expand and evolve in the coming years.

Science and Policy Drivers

The DOE BER, which supports the JGI, also supports a portfolio of research projects in energy and environment, and priorities for both are periodically updated through visioning and strategy workshops. The most recent of

these strategic planning sessions took place in 2017 and resulted in a report entitled *Grand Challenges for Biological and Environmental Research: Progress and Future Vision* (Table 2.1). It outlines key challenges in the BER research domain and recommends directions to overcome them. Some overarching themes include obtaining spatially and functionally resolved data, linking genotype to phenotype and function, and integrating large datasets.

Table 2.1. JGI strategies for addressing the BER grand challenges

BER Grand Challenge	JGI Strategy
Biological Systems	
2.1 Understand the biological complexity of plant and microbial metabolism and interfaces across scales spanning molecules to ecosystems.	The JGI will develop and improve tools for studying plant-microbe interactions, including endophyte sequencing approaches, exometabolomics techniques, and functional characterization of secondary metabolites.
2.2 Develop technologies to identify DOE mission–relevant metabolic capabilities and engineering possibilities in bacteria, fungi, archaea, viruses, plants, and mixed communities.	Multiple methods in metabolomics, stable isotope probing, synthetic biology, and metabolic modeling will be deployed to understand cellular- and community-level metabolism across a broad range of organisms.
2.3 Optimize the use of large datasets that integrate omics surveys with biochemical and biophysical measurements to generate knowledge and identify biological paradigms.	Biological data science, including data mining and data integration, will be a major focus of the JGI's efforts and its partnership with KBase in coming years.
2.4 Understand the links between genotype and phenotype in single but very diverse organisms and in communities of organisms that interact in terrestrial ecosystems.	Plant and microbial genomics projects will be increasingly coupled with phenotypic characterization to improve genotype-phenotype linkages.
2.5 Effectively exploit new and emerging technologies in systems biology and physical measurements (e.g., miniaturization) to accelerate biological discoveries.	A complete design-build-test-learn (DBTL) cycle for synthetic biology will be implemented at the JGI for characterizing key pathways such as those for secondary metabolite biosynthesis.
BER Grand Challenge	**JGI Strategy**
Earth and Environmental Systems	
4.1 Characterize the biogeochemical exchanges driven by food web and plant-microbe interactions and evaluate their process-level impacts, sensitivity to disturbances, and shifting resource availability under changing environmental regimes.	A major focus on interorganismal interactions will exploit methods such as deep metagenome sequencing and genome extraction, metabolomics, stable isotope probing, and synthetic community experiments to investigate biogeochemical exchanges.
4.2 Define the sphere of influence and key elements of microbial communities in space and time relevant for predicting larger-scale ecosystem phenomena for Earth system understanding.	Single-cell and in situ methods will improve spatial resolution while time-series and spatially resolved metagenomics will characterize community changes with space and time.
4.3 Integrate molecular and process data to improve the ability to define ecologically significant traits of individual taxa and communities and use trait-based models to develop predictive links between community dynamics and ecosystem processes.	Metagenome sequence data, including that derived from stable isotope probing (SIP) experiments, will be coupled to biogeochemical and phenotype data to build trait-based models and develop predictive methods.

BER Grand Challenge	JGI Strategy
Earth and Environmental Systems	
4.4 Align and deepen connections among conceptual understanding, measurements, and models related to the roles of microbes in determining the rate of transformation, uptake, and loss of chemical elements from ecosystems.	The JGI will functionally characterize microbial communities through function-driven single-cell and metagenomic approaches such as bioorthogonal noncanonical amino acid tagging (BONCAT) and SIP as well as engage with KBase to improve community metabolic modeling.

BER Grand Challenge	JGI Strategy
Data Analytics and Computing	
6.1 Develop robust approaches for large-scale data collection, curation, annotation, and maintenance.	The JGI's data systems will adhere to findable, accessible, interoperable, and reusable (FAIR) principles and incorporate robust metadata linkages.
6.2 Develop computing and software infrastructure to enable large-scale data (i.e., Big Data) storage and analysis.	The JGI, with its partners NERSC and KBase, will be a leader in software infrastructure for large-scale biological data storage and analysis.
6.3 Conduct research to develop suitable algorithms and programming models that can harness current and future computer architectures to effectively model complex coupled systems and analyze extreme-scale data.	The JGI will continue to work with NERSC and the Computational Research Division (CRD) at Berkeley Lab to provide benchmarks and use cases for testing next-generation compute architectures. The JGI will also continue work on the ExaBiome project, a collaboration with the CRD to extend critical bioinformatics algorithms to exascale.
6.4 Engineer advanced computational modeling combined with data integration across temporal and spatial scales.	The JGI, NERSC, and KBase will work with each other, other user facilities, and users to improve data integration. The JGI will collaborate with applied mathematicians and statisticians to develop the appropriate multiscale models to support these analyses.
6.5 Conduct research and develop activities that support human understanding of large-scale, multimodal data streams, including the ability to steer experiments in real time.	The JGI's expanding efforts in biological data science include methods to better interpret biological "big data." The JGI will work with partners to develop explainable-AI methods for terabyte-sized multi-omics data sets.

BER Grand Challenge	JGI Strategy
User Facilities and Research Infrastructure	
7.1 Foster a spirit of collaboration to enable integrative capabilities among BER and Office of Science user facilities, as well as other federal research facilities and infrastructure, thereby promoting a fully interdisciplinary approach to BER-relevant science.	The JGI is partnering closely with EMSL, KBase, and NERSC and aims to expand these partnerships to additional facilities. The JGI's active membership with the Society for Science at User Research Facilities is providing guidance on how to streamline processes for user access across facilities through mechanisms such as FICUS.
7.2 Solicit input from the BER research community regarding researchers' needs and train them in new experimental, observational, and modeling approaches, thus propagating the knowledge and skills for generating high-impact scientific results.	Extensive outreach efforts are planned to better engage the BER research community, including co-promotion campaigns with EMSL, NERSC, KBase, and the Atmospheric Radiation Measurement (ARM) Climate Research Facility, and exploring reciprocal resource promotion opportunities with the Nanoscale Science Research Centers supported by DOE Basic Energy Sciences.

7.3 Develop innovative enabling technologies and construct and acquire state-of-the-art instruments that exploit the world-leading characteristics of each user facility. This will boost capabilities for basic research in biological systems and Earth and environmental systems science, thereby providing DOE and the nation with leading-edge capabilities for biological and environmental science.	The JGI will continue to provide cutting-edge capabilities to the user community, including early access partnerships with relevant commercial vendors, engagement in the scientific community through meeting attendance and literature mining, recruitment of key staff, leveraging the JGI's Scientific Advisory Committee for critical feedback, and the development of new Emerging Technologies Opportunity program (ETOP) focus areas as opportunities arise.
7.5 Build upon existing investments and capabilities at the DOE Office of Science light and neutron science user facilities, continuing to align them with BER missions.	The JGI will partner with light sources to offer imaging and structural biology capabilities coupled with omics through a joint FICUS call.
7.6 Further develop the necessary infrastructure at user facilities to study organisms in their natural habitats.	The JGI is actively developing methods to better characterize organisms in their natural habitats, including BONCAT coupled with flow sorting, SIP, and Drop-seq.
7.7 Develop and adopt technologies to convert genome sequence data into functional understanding at appropriate BER user facilities.	The JGI's strategic efforts are heavily aimed at improving functional annotation of genome data through new or expanded experimental and computational methods.
BER Grand Challenge	**JGI Strategy**
Emerging Technologies	
8.1 Characterize the genotype and phenotype of individual cells, including genomics, transcriptomics, proteomics, and metabolomics, to enable high-resolution predictive biology.	A suite of single-cell methods exists or is in development at the JGI, including untargeted and targeted single-cell genomics, Drop-seq, and in situ sequencing.
8.2 Increase throughput and integration of genomics, transcriptomics, proteomics, and metabolomics to enable improved translation from the molecular to cellular realm.	The JGI is continually increasing the throughput of genomics, transcriptomics, and metabolomics capabilities, and the Data Science and Informatics team is developing strategies to present these data in an integrated framework.
8.4 Integrate data covering broad time and length scales—from seconds to years and from Ångströms to the Earth scale—to enable multiscale comprehension and simulation.	The JGI will expand its efforts in macro-scale microbial ecology by explicitly incorporating spatial, temporal, and phylogenetic data through the application of new computational methods and statistical approaches.
8.5 Develop integrative and interpretive computational approaches that can handle large, disparate data types from multiple and heterogeneous sources using advanced and exascale computing.	Biological data science will be a major thrust for the JGI, in cooperation with partners NERSC and KBase. The JGI will also continue work on the ExaBiome project, a collaboration with the CRD to extend critical bioinformatics algorithms to exascale.
8.6 Explore the development and application of nanoscience and nanotechnology to further understand biological systems at nanoscales or larger scales (e.g., nanosensors).	The JGI will engage with scientists and facilities at Berkeley Lab and elsewhere to productively couple genomics with nanotechnology.

Numbers refer to the 2017 report *Grand Challenges for Biological and Environmental Research: Progress and Future Vision* by the DOE Biological and Environmental Research Advisory Committee.

Partner Relationships

All JGI efforts take place in the context of our many productive partnerships with other government, academic, and industry institutions. Our relationships with partner labs and user facilities, and their complementary missions and capabilities, provide synergistic opportunities to enhance our internal capacity and expertise.

Berkeley Lab (Biosciences, NERSC, and ESnet)

The JGI is part of the Biosciences Area at Berkeley Lab. The Area's mission, to "use integrated research teams to solve national challenges in energy, environment, health, and biomanufacturing," complements and expands upon the JGI's mission. The Biosciences Strategic Plan lays out a series of goals for advancing energy and environmental science, many of which are shared with the JGI. Additionally, several JGI staff have affiliate appointments with Biosciences Divisions at Berkeley Lab under which they conduct their own scientific research programs.

NERSC and ESnet also reside at Berkeley Lab and are critical partners in the JGI's informatics activities. All of the JGI's computing takes place at NERSC, and the rapid exchange of information between these physically separate locations is made possible by ESnet's high-speed network. The JGI's Data Science and Informatics and Genome Analysis staff work closely with NERSC to enable the large-scale data analyses critical to JGI science. The JGI's computing strategy is built on and aligned with NERSC's capabilities and future plans.

HudsonAlpha Institute for Biotechnology

The JGI plant program is executed by the HudsonAlpha Institute for Biotechnology, a genomics research facility whose mission encompasses medical, agricultural, and environmental science. Through this JGI-funded partnership, the JGI is able to leverage experimental, computational, and personnel resources at HudsonAlpha, and the active exchange of technical and scientific knowledge between institutes is an important contributor to the JGI's strategic directions.

Environmental Molecular Sciences Laboratory

The JGI has partnered with EMSL for the past several years to support users through the Facilities Integrating Collaborations for User Science (FICUS) initiative. EMSL possesses a suite of spectroscopic, imaging, and computational capabilities that complement and enhance those at the JGI, with a focus on understanding molecular processes and ecosystem functions. The scientific focus and goals of the FICUS efforts are jointly established by the two user facilities, and the JGI's strategic planning aims to enhance support for our joint user base.

DOE Knowledgebase (KBase)

KBase provides an analytical forum for systems biology of plants, microbes, and their communities, including a variety of tools for omics data analysis. KBase is a critical partner in enabling systems biology analyses, such as metabolic modeling, that go well beyond genome annotation and analysis. The JGI and KBase are developing complementary and integrated high-performance tools that provide users with the ability and infrastructure to explore complex and diverse datasets to extract deeper biological insights. Through joint strategic leadership, joint tactical teams, and frequent communication, both entities are partnering to accomplish the goals laid out on a joint roadmap while leveraging capabilities across Berkeley Lab. Successful partnership between the JGI and KBase will enable users to seamlessly access and analyze data on scales ranging from single genes and individual genomes to metagenomes to systems-level modeling and understanding.

Bioenergy Research Centers

The four Bioenergy Research Centers (BRCs) are funded by BER to lay the scientific groundwork for a new bio-based economy. Since the creation of the first three BRCs in 2007, up to 30% of the JGI's capacity has been reserved for their genomics needs. Beyond making JGI capabilities available to BRC researchers, this close relationship has also resulted in numerous successfully completed and ongoing scientific collaborations that take advantage of the complementary and often synergistic expertise and capabilities available at the BRCs and the JGI. Many capabilities at the JGI have been built or adapted to meet BRC requests. Continued support of BRC work in feedstock development, biomass deconstruction, biofuel and bioproduct conversion, and sustainability, as well as resulting scientific collaborations, will remain key drivers for the JGI.

Knowledge Gaps

All these scientific and technological inputs have aided in identifying a set of knowledge gaps and learning opportunities of critical importance for energy and environmental science, which motivate the JGI's strategic vision and consequently the efforts proposed in this plan. They include the following (Fig. 2.1):

- **Phylogenetic and functional diversity:** Although sequence data have been generated from an extensive array of organisms and environments, many branches in the tree of life remain undiscovered or uncharacterized. We aim to continue to expand the known world of genes, genomes, and function.
- **Pangenomics:** Sequence data from diverse organisms is illustrating the surprising and important diversity in gene content across the genomes from individuals of the same "stable" species. We aim to continue to explore this in microbes, fungi, and plants.
- **Sequence to function:** Even for relatively well-characterized organisms, many genes remain unannotated, minimally annotated, or are assigned incorrect annotations which continue to be propagated. We aim to pursue experimental and computational techniques to improve annotations and functional predictions across all genomes.
- **Genotype to phenotype:** The phenotypic effects of variation in sequence or gene content remain difficult to predict. We will leverage tools for high-throughput phenotyping and genotyping to contribute to crop breeding efforts, generation of biofuels, bioproducts and biomaterials, metabolic modeling, and synthetic biology.
- **Inter-organismal interactions:** Most genomic studies have focused on individual organisms in isolation, yet many genes likely function in a competitive or cooperative context with other organisms. We aim to characterize the physical, chemical, and genetic factors mediating interactions of these organisms using controlled laboratory ecosystems combined with omics techniques.
- **Spatially resolved / single-cell omics:** Most genomics studies rely on bulk tissue or environmental samples to yield sufficient nucleic acid for sequencing, but single-cell methods reveal linkages and inter-cell variations that bulk methods do not, and spatially resolved methods reveal physical features masked by other approaches. We will pursue methodologies to characterize organisms and communities at the single-cell level to better understand organism and community properties.
- **Metabolism:** Primary and secondary metabolites are the means by which cells exchange nutrients and information and survive under suboptimal conditions. We aim to characterize the diverse metabolites produced by bacteria, archaea, plants, fungi, and algae and discern their roles in nutrient cycling and organismal interactions in the environment.

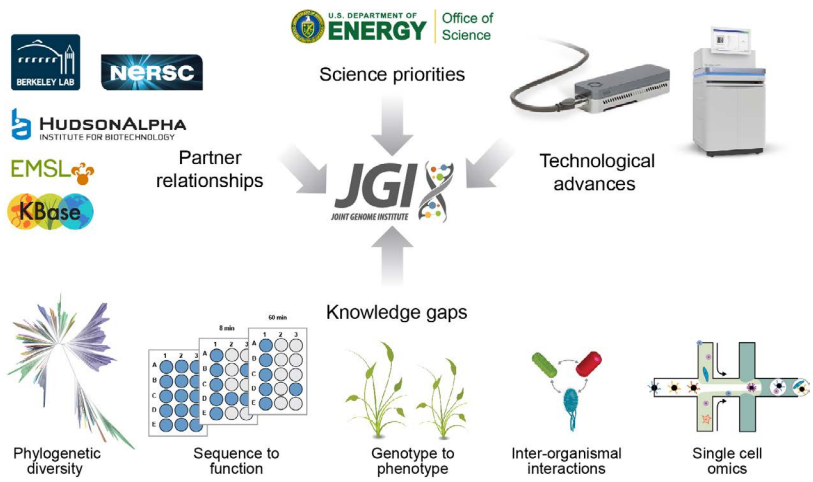


Figure 2.1. Selected drivers of strategic planning at the JGI.

3. Identification

A major focus of the JGI will continue to be advancing scientific discovery connected to the mission areas of the DOE. These scientific insights will be driven by linkages between the JGI data-generating platforms, the Science Programs, and continued growth in data science. In this section, we focus on **Identification** as the first of five components of the JGI's **I5 Strategic Framework**.

Data Generation and Future JGI Platforms

The JGI will continue to have a sizable footprint in generating critical omics data for next-generation users for the foreseeable future. This will include growth in biospecimen accessibility, continued development of platforms in sequencing, single cells, DNA synthesis, and metabolomics, and new areas in secondary metabolite discovery and advanced bioassays. Further integration of these capabilities will maximize scientific discovery.

Historically, the JGI's success in developing and deploying cutting-edge technological capabilities and applying them to DOE mission-relevant questions has relied on the coordinated activities of the Genomic Technologies department and the Science Programs department. The former department identifies, establishes, and optimizes new technologies, scales them into robust and cost-effective data production workflows, and performs all data production for user projects. The latter develops and defines the scientific directions of the JGI, identifies general needs for new technologies required to support the DOE research mission, and works closely with users to facilitate their access to and utilization of the capabilities offered by the JGI. As the JGI further evolves its technology portfolio, this successful model will be applied to new platforms within the **I5 Strategic Framework** (Fig. 3.1). In support of **Identification** (this section), the JGI will develop and expand platforms for sample preparation, sequencing, functional genomics assays, and single-cell omics. In support of **Investigation** (page **41**), the existing DNA synthesis,

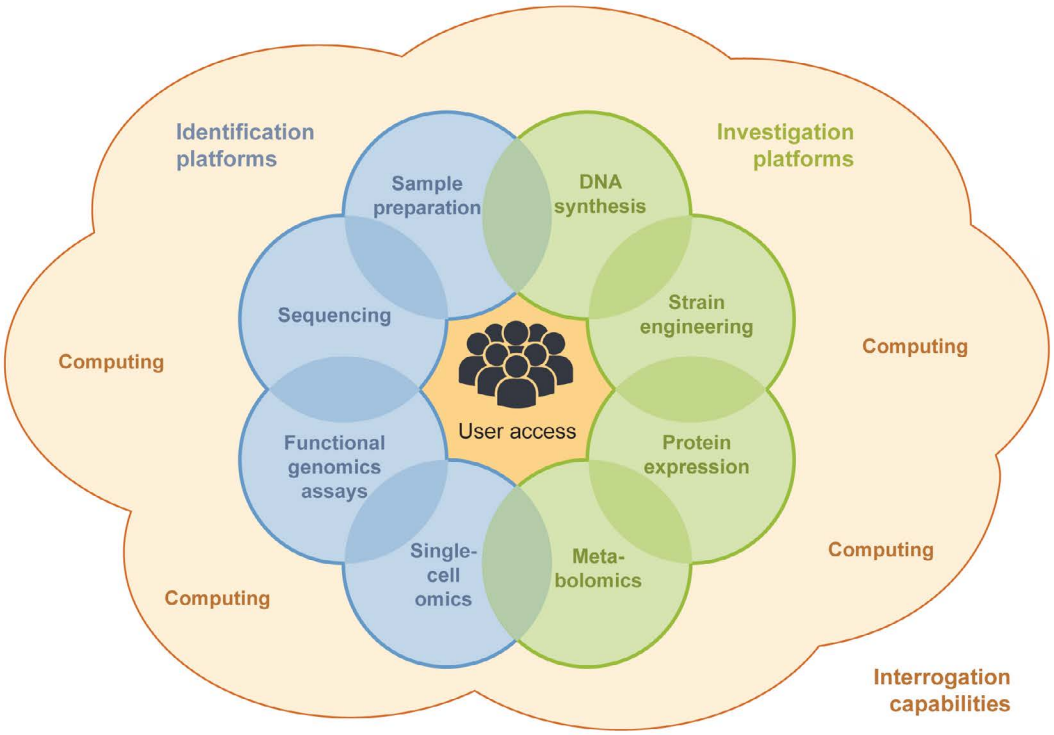


Figure 3.1. Existing and new platforms available to JGI users.

strain engineering, and metabolomics platforms will be further expanded, and new protein expression platforms will be developed. As the types of data produced by these platforms diversify, the evolution of these technology platforms will be accompanied by the development of new computational tools to enable data **Interrogation** (page **31**).

A Critical Role for Biospecimen Accessibility

The increased ability to culture a wide range of plants, microbes, and fungi in-house will be highly beneficial to the JGI's continuing effort to remain at the forefront of energy and environmental genomic research. These benefits include access to living material for methods development, the ability to create mock communities (e.g., microbiome and plant-microbe) to enable benchmarking and hypothesis testing, and development of valuable biological resources for the larger scientific community (e.g., transposon sequencing [Tn-Seq] microbial collections). The JGI's current facilities for growing plants and culturing are primarily designed for model organisms. With the move to the new IGB, a cultivation room and new equipment will facilitate the cultivation of a much expanded repertoire of organisms, providing new capabilities to our user community. For example, the JGI will be in a better position to provide engineered libraries of microbial organisms, particularly those identified as high priority by the BRCs. Additionally, through combining sequencing, metabolic modeling (with KBase), and cultivation, the JGI will be able to explore how to cultivate the "uncultivable."

LIVING CELLS AND TISSUE FOR NEW METHODS DEVELOPMENT

In the past, the JGI has worked almost exclusively with nucleic acid samples (DNA and RNA) provided by our users to construct over 30 different library types with applications including genome assembly, transcriptomics, population resequencing, and epigenomics (small RNA and DNA methylation). However, as sequencing methods continue to evolve, powerful new approaches for understanding the structure of DNA in the nucleus and single-cell transcriptome profiling have been developed that require intact cells or fungal or plant tissues as their starting point. To remain state-of-the art, the JGI has begun to implement several of these approaches to work with user-shipped tissue samples from a limited set of model organisms. In the future, the JGI will expand these methods to a wider range of non-model systems to further optimize protocols to account for organism-specific properties. This will require systematic evaluation of different organisms and tissues to develop robust and generalized protocols and will be facilitated by expanded growth capabilities.

DEFINED MICROBIOMES FOR STANDARDIZATION

Increased efforts in microbial cultivation will improve the JGI's ability to study complex microbial communities and plant-microbe interactions. For example, defined mixtures of known microbial representatives can in principle provide a standard for accurately evaluating laboratory and computation biases associated with new metagenomic methods. However, the phylogenetic diversity of commercially available "defined microbiome" standards is generally too low for realistic tests. Developing complex, defined communities with a broad phylogenetic range will allow us to make much more suitable proxies for benchmarking that will enable the JGI to develop high-accuracy approaches for advanced genome and transcriptome assembly. Combined with expanded growth and culturing facilities, these defined microbiomes will enable controlled experiments to better understand the dynamics of interactions in microbial and plant-microbe communities. The ability to culture a broad spectrum of microbes will also allow the JGI to leverage complementary functional genomics techniques, such as the testing of Tn-Seq insertional mutant collections, within the context of more complex communities.

JGI Sequencing Capabilities

Over the past decade, massive decreases in sequencing cost along with increases in throughput and read lengths have revolutionized genome science, and have allowed the JGI to support ever more ambitious and innovative user projects. Broadly, DNA sequencing can be categorized into short- and long-read sequencing technologies; both play important and complementary roles in the JGI's mission. Short-read sequencing technologies are currently dominated by sequencing-by-synthesis platforms and produce read lengths of 50 to 300 base pairs at very low cost and high throughput. These are ideal for applications such as transcriptome and epigenome profiling or variant calling when a reference genome is available. Single-molecule sequencing technologies produce much longer reads (>30kb), though at higher cost per nucleotide and at a lower throughput, and are ideal

for assembly of high-quality reference genomes, a prerequisite for all downstream analysis. Potential future disruptive technologies are likely in the area of nanopore sequencing with lower input material requirements, the ability to directly sequence RNA to detect RNA-specific modifications, and ultra-long read contributions to complex genome assembly.

While we expect that both short- and long-read platforms will continue to be used extensively at the JGI in the future, over the past five years we have seen a continually growing demand for long-read sequencing (Fig. 3.2). This demand reflects the JGI's unique position as a facility capable of both high-throughput long-read sequencing and high-quality reference genome assembly using efficient computational pipelines. Based on this trend, we expect to continue to aggressively invest in long-read sequencing to support user demand for high-quality reference and structural variant detection for DOE-relevant environmental species. See Milestones **GNT02** and **GNT03**.

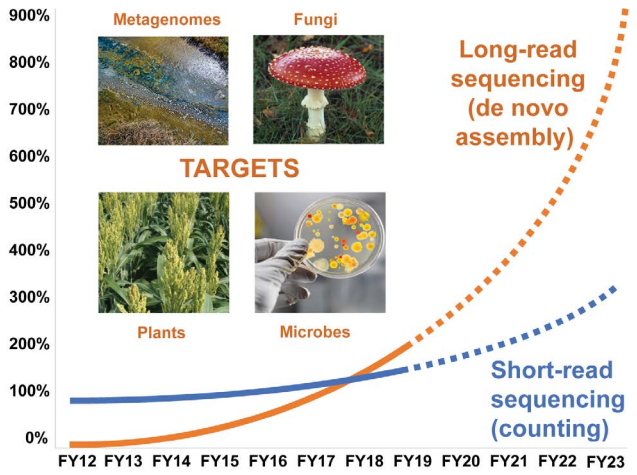


Figure 3.2. Anticipated growth in short-read (blue) versus long-read (orange) sequencing output at the JGI. While short-read sequencing will continue to generate more overall base pair output, we expect long-read sequencing output to grow more dramatically.

FUNCTIONAL GENOMICS CAPABILITIES FOR UNDERSTANDING GENE REGULATORY NETWORKS

Gene regulatory sequences determine when and where genes are expressed within multicellular organisms, which allows individual cell types to execute unique genetic programs, and organisms to respond to internal or external stimuli. In the past few years, new sequencing-based approaches have emerged that enable the characterization of gene regulation. These methods can reveal the cell type-specific modifications to DNA, such as DNA methylation (bisulfite sequencing), the accessibility of DNA to interact with regulatory proteins (ATAC-seq), or the long-range looping of DNA in the nucleus that fine-tunes gene regulatory programs in individual cell-types (HiC). The JGI has already implemented bisulfite sequencing and ATAC-seq, and is currently developing HiC for both plants and fungi to allow users to better understand the underlying properties of the nuclear genome that regulate gene expression. In addition, we have implemented a method for high-throughput characterization of transcription factor (TF) binding sites in the genome (DAP-seq) that has the potential for massive scaling. DAP-seq is a low-cost high-throughput method capable of quantifying all potential TF binding sites for a large proportion (35–65%) of all TFs within a species. The combination of high-resolution maps of TF binding sites, DNA modifications, genome accessibility, and 3D-structure will offer a unique opportunity to investigate the cell- and tissue-specific gene networks that underlie traits relevant to DOE BER scientific goals. See Milestones **GNT08** and **GNT09**.

STABLE ISOTOPE PROBING (SIP) FOR MAPPING FUNCTIONAL ACTIVITY IN MICROBIOMES

Our ability to recover the genomes of uncultivated microbes has improved dramatically in recent years, allowing us to study the metabolic potential of thousands of species. However, measuring the actual metabolic activity of these microbes is challenging, which limits our ability to incorporate the specifics of microbial community composition into conceptual and predictive models. The JGI will develop a pipeline for SIP metagenomics to reveal the metabolic activity of uncultured microbes. With this approach, we will identify microbial groups actively assimilating various isotopically labeled compounds by selectively sequencing isotopically labeled DNA, thus providing quantitative insights into in situ metabolic activity, which can potentially be tested using SIP-metabolomics analysis. We also plan to adapt this approach to RNA

as input, i.e. SIP-metatranscriptomics, which will dramatically shorten incubation times and provide an even clearer picture of "who" is doing "what" in natural microbial communities. See Milestone **GNT04**.

A Future in Single-Cell Omics

SINGLE-CELL TRANSCRIPTOMICS OF COMPLEX EUKARYOTIC TISSUES

Phenotypic differences among tissues, which are composed of genetically identical cells, arise from differences in gene expression. However, the specific expression patterns differentiating tissues are typically obscured by traditional methods that homogenize all cell types/tissues and produce a single expression pattern. Thus, the expression data recovered from a structure like a root or leaf is an amalgamated signal, and none of the cells comprising the structure actually display the measured expression profile. Single-cell transcriptomics methods like Drop-seq have been applied to animal systems to distinguish expression patterns of various tissues and cell types within complex structures. Briefly, this approach leverages the throughput of microdroplet emulsions to encapsulate and uniquely barcode tens of thousands of single cells prior to sequencing in order to preserve the expression profile of individual cells. Using single-cell transcriptomics, it is possible to dissect the expression patterns of tissues and cell types comprising complex structures. The JGI has established proof of principle for the application of Drop-seq to plant tissues and anticipates expanding this approach to tissues from several DOE-relevant plant systems, and perhaps microbial communities, which will provide single-cell resolution transcriptomes and aid in elucidating gene regulatory networks and cellular identities.

Beyond single-cell transcriptomics, several methods originally developed for bulk tissues have recently been adopted for single-cell applications to characterize, at single-cell resolution, properties such as open chromatin, long-range DNA looping, and DNA methylation. We will also explore the application of these approaches to DOE-relevant species to provide insight into the genomic mechanisms underlying cell-type specific gene regulation. See Milestones **GNT01** and **GNT06**.

IMAGING GENE EXPRESSION WITHIN A SPATIAL AND BIOLOGICAL CONTEXT

While single-cell transcriptomics can preserve distinct expression profiles of individual cells making up a structure, the spatial and biological context of these cells is lost. For example, a Drop-seq experiment may reveal that a leaf is composed of cells displaying any one of a dozen different expression patterns, but establishing strong links between particular expression patterns and a particular tissue is challenging. Developing new methods that preserve the spatial context, e.g. sequencing gene transcripts from laser-microdissected tissues, is needed to translate expression data into insights regarding plant physiology and function. Another attractive area of future study is fluorescent in situ sequencing (FIS-seq) where RNA expression is performed directly on thin sections of tissues rather than on extracted RNA. With this, RNA sequencing happens within the cells under a microscope, and the expression patterns of specific cells are maintained within the spatial context of a tissue structure. Building on proof-of-principle demonstrations in human cells and fruit fly embryos, the JGI will explore the possibility of developing this approach in DOE-relevant organisms for user access as a stretch goal. See Milestone **GNT07**.

DE NOVO GENOME ASSEMBLY OF UNCULTURED, UNICELLULAR EUKARYOTES

Most species of unicellular eukaryotes, including fungi, protists, and microalgae, have not been cultivated. Thus their genetic diversity and metabolic potential is unknown. To provide a technology platform to address this issue, we will adapt methods of single-cell whole genome amplification typically used for bacteria to challenges associated with unicellular eukaryotes. Most of our initial efforts will focus on fungi, and in future years we will expand to other types of unicellular eukaryotes. More details on the science drivers for this technology can be found in the "**Eukaryote Intracellular Associations**" and "**Expanding the Fungal Tree: Sequencing Unculturables**" sections, see pages **24** and **25**.

Leveraging Microbial Type Strains to Enable Systems Biology

Under the umbrella of the Genomic Encyclopedia of Bacteria and Archaea (GEBA), the JGI is sequencing thousands of bacterial and archaeal isolate genomes from diverse branches of the microbial tree of life, most recently focused on type strains. This phylogeny-driven approach is of great value in multiple areas of scientific interest. It provides a foundation for improved annotation of other microbial genomes and the phylogenetic anchoring of metagenomic data, and facilitates novel gene discovery and an increased understanding of the processes underlying the evolutionary diversification of microorganisms. Over the past few years, the JGI has sequenced more than 3,000 type strains through multiple community-driven projects, establishing itself as the world leader in bacterial and archaeal type strain genome sequencing (Fig. 3.3). Type strains represent living cultures that serve as a fixed reference point for the assignment of bacterial and archaeal names, and are accessible from public culture collections for experimental work. In a new phase of the GEBA initiative, the scientific research community together with the JGI aims to further expand the phylogenetic coverage of cultured microbes by sequencing 10,000 new type strains. In addition, the JGI plans to expand the diversity and genome-centric understanding towards systems biology by leveraging the JGI's high-throughput RNA-Seq and metabolomics capabilities. Approximately 500–1,000 type strains will be selected for functional characterization using transcriptome and metabolome analyses under different strain-specific growth conditions, uniting genomics with biological properties of these bacteria and archaea and proving a rich substrate for data integration and modeling approaches through DOE KBase. See Milestone **MIP01**.

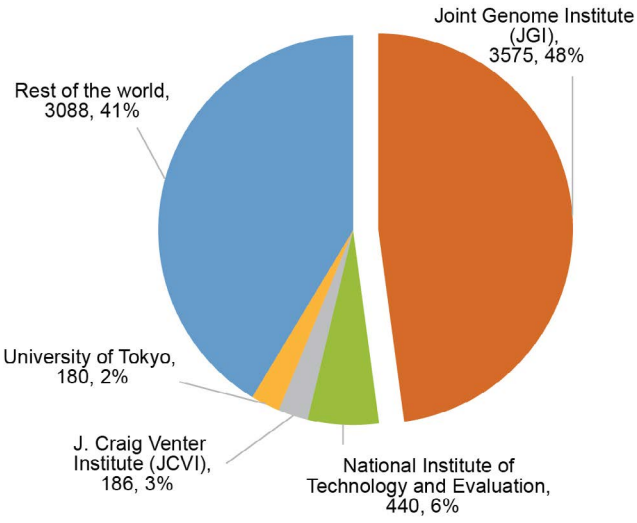


Figure 3.3. The JGI's leadership position in sequencing unique bacterial and archaeal type species.

Unlocking Phylogenetic and Functional Diversity of Environmental Microbiomes

Advancing Phylogeny-Driven Single-Cell and Genome-Resolved Metagenomics

Microorganisms are major drivers of global biogeochemical cycles, affecting the environment, animal and plant health, and the evolutionary trajectory of life on Earth. They encompass a tremendous diversity, much of which has remained uncatalogued and underexplored, in part due to the continued cultivation bottleneck. Deeply characterizing organisms across the tree of life will enable discovery of new enzymes, pathways, and metabolites; enhance our understanding of evolution; and define roles in the flow of nutrients through our environment. Through a targeted focus on phylogenetic diversity, using cultivation-independent approaches, the JGI aims to continue exploration of bacterial and archaeal diversity and their pangenomes.

Our view of microbial diversity has greatly expanded, primarily through the wide application of shotgun metagenome and single-cell genome sequencing (Fig. 3.4). The JGI is among the few research centers that pioneered and scaled the application of high-throughput sequencing of single amplified genomes (SAGs) for studying uncultivated bacterial and archaeal environmental lineages. The rapid pace of improvements in metagenome assembly and binning methods now enables the high-throughput reconstruction of population genomes called metagenome-assembled genomes (MAGs). The JGI aims to further advance high-throughput single-cell genomics and genome-resolved metagenomics through the 16S-independent sequencing of single cells and the generation of MAGs for all publicly

available metagenomes, to facilitate interrogation of uncultivated taxa across diverse ecosystems. Analysis of the predicted metabolic and functional features associated with these uncultivated microbes will further expand our view into the biology of microbial dark matter. See Milestones **MIP02**, **MGP04**, and **PKI05**.

Expanding Function-Driven Genomic Approaches

Most single-cell and metagenomic sequencing projects have relied on random recovery of microbial genomes. As technologies mature, the targeted selection and capture of cells and/or DNA for sequencing is becoming a tractable and highly desirable approach. Such selection, which has been demonstrated at both the taxonomic and function-based levels, allows the enrichment of taxa and/or functions of interest and relevance, based on **who they are** or **what they do**. The JGI aims to further develop, scale, and implement experimental methods for function-driven single-cell genomics and metagenomics for the targeted identification and interrogation of uncultivated taxa and to link sequence to function (Fig. 3.4). This will enable focused investigation of activities and pathways critical to biogeochemical cycling, such as biomass degradation, methanotrophy, and carbon fixation. Ultimately these data will feed into community metabolic models to better predict community behavior and response to perturbation. See Milestones **MIP03** and **MGP01**.

The application of labelled substrates and SIP provide means to expand the toolkit of **sequence-to-function** approaches for uncultivated microbial taxa (technical details on SIP can be found in the section "**Stable Isotope Probing [SIP] for Mapping Functional Activity in Microbiomes**" see page **19**). Similarly, yet with single-cell resolution, fluorescently-labeled substrates can provide an assay to capture microorganisms that bind and putatively degrade a target substrate. Labeled cells can be isolated through fluorescence-activated cell sorting, followed by genome amplification and sequencing. The JGI aims to further develop such functional approaches to elucidate functional properties of microbiomes, particularly focusing on inter-organismal interactions and plant-microbiome. This will enable the targeted enrichment of DOE-relevant functional microbial guilds, the directed discovery of

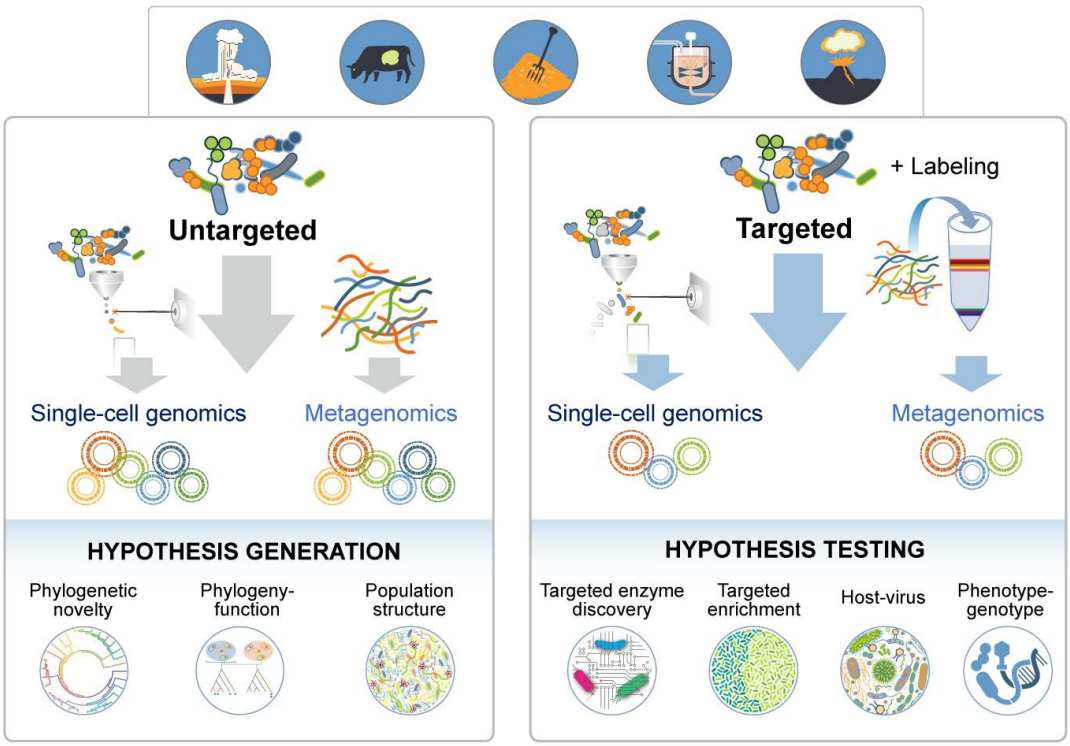


Figure 3.4. Targeted and untargeted approaches for microbial genomics and metagenomics. Traditional approaches to single-cell genomics and metagenomics provide high-throughput pipelines for the generation of reference genomes from uncultivated taxa, primarily for discovery-based science (left-hand side). Targeted genomics (right-hand side) through labeling and SIP enables the enrichment and analysis of microbial subpopulation of interest and provides a link between genotypes and phenotypes.

enzymes with desired properties, the targeted discovery of novel host-virus associations, and the establishment of linkages between phenotype and genotype, ultimately feeding into community metabolic models and ecosystem understanding. See Milestones **GNT04**, **MIP03**, **MTB04**, and **MGP02**.

Exploring Inter-Organismal Interactions

Research over the last two decades has revealed that virtually no organism exists in isolation in its natural habitat. Not only do most microbes exist as members of diverse consortia, but macroorganisms from macroalgae to land plants and from insects to mammals interact with and depend on diverse communities of symbiotic microorganisms, which in turn interact with viruses. Single-celled eukaryotes house bacterial endosymbionts and giant viruses, which are amenable to genomic interrogation when coupled with the JGI's unique single-cell sequencing capabilities. Understanding these associations is essential to fully characterize the behavior and evolutionary history of the organisms we study. What mediates these interactions is often unknown, but likely includes a spectrum of physical and chemical signals. The JGI will focus on using experimental approaches to tackle questions related to plant-microbe interactions, virus-host interactions, and eukaryote intracellular associations (Fig. 3.5).

PLANT-MICROBE INTERACTIONS

The microbiome growing on and in plants has a large effect on plant productivity. Hence there is untapped potential to harness beneficial interactions to sustainably increase yields to help meet the growing demand for energy crops and agricultural products. Despite this potential, with the exception of nitrogen-fixing bacteria and mycorrhizal fungi, few microbes or mixtures of microbes are currently used in large-scale agriculture due to our poor understanding of plant-microbe interactions. It is currently impossible to predict if the addition of a microbe or a mix of microbes to a particular field with its own microbial community, soil type, plant genotype, and environmental conditions will be effective. To fully harness the microbiome requires deep knowledge of the rules that govern microbial colonization, how plants control microbiome composition, and the mechanisms of growth promotion, microbial competition, and cooperation. The JGI is uniquely positioned to address these questions using its portfolio of genomics tools for studies of microbes, plants, and fungi in conjunction with complementary DNA synthesis and metabolomics capabilities.

We aim to better understand the microbial factors influencing microbiome assembly using isolate and metagenome sequencing to assess the functional potential of plant-associated microbes and

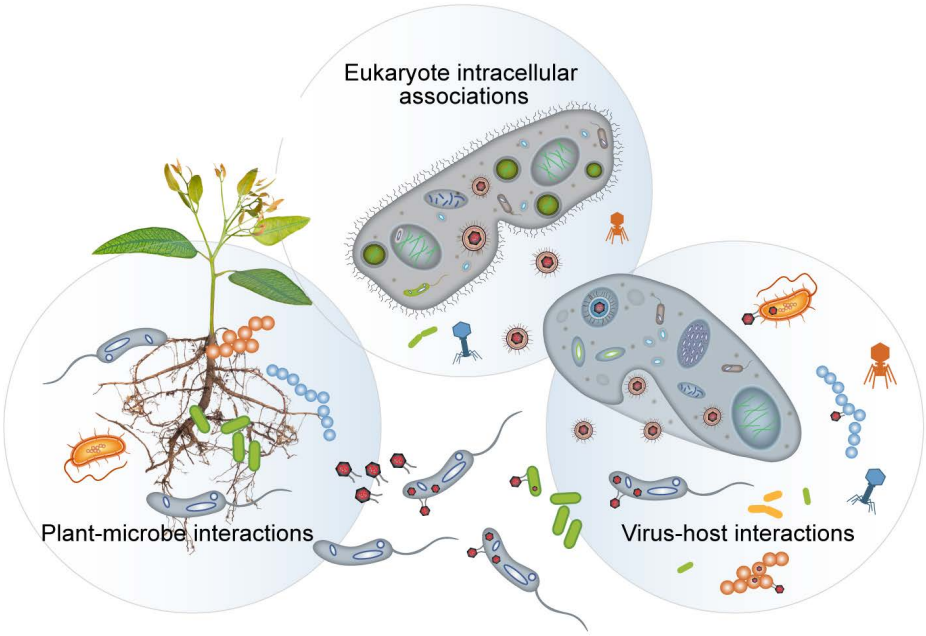


Figure 3.5. Different types of inter-organismal interactions in environmental systems. The study of inter-organismal interactions will focus on using experimental and computational approaches to tackle questions related to plant-microbe interactions, virus-host interactions, and eukaryote intracellular associations.

transcriptomics to characterize their interactions with plants. To accomplish this goal, improved methods to target endophytes for nucleic acid sequencing will be needed, and we will actively pursue technologies to reduce host representation in metagenomes and metatranscriptomes. Further, exchange of primary and secondary metabolites is a major driver of plant-microbe associations. Exo-metabolite profiling of both microbes and plants, together and in isolation, will enable the construction of community metabolic models and identify key signaling molecules mediating the symbiotic relationship. Finally, manipulative experiments in defined ecosystems will be used to both generate and test hypotheses relating to the functional roles of specific organisms, genes, and metabolites in microbiome assembly and function. See Milestones **PLP09** and **GNT11**.

VIRUS-HOST INTERACTIONS

Viruses exert significant impacts on microbial population dynamics, long-term evolution, and ecosystem functions. Recent advances in computational methods to predict viruses directly from sequence data have enabled exploration of diversity at an unprecedented scale and level of detail. Uncovering the viral sequence space and collaboratively establishing standards to describe uncultivated viruses has been a major thrust of the JGI's work in this area over the past few years. Moving forward, the JGI aims to develop new computational and experimental approaches to accelerate the discovery and validation of viral-host linkages in the wild, shedding light on virus evolution and dynamics in nature.

While computational methods have enabled exploration of global viral diversity, a significant fraction of these uncultivated viruses lack association with their respective microbial hosts. A major bottleneck currently exists to experimentally link viruses to their hosts in a high-throughput manner, hindering our ability to understand virus-host interactions in complex communities. To complement the robust computational approaches described in the previous paragraph, the JGI aims to develop or adopt methods for high-throughput identification of viral-host linkages. Promising examples include viral tagging, which uses fluorescently-labeled viruses to tag infected host cells and isolate them with flow cytometry, and epicPCR (Emulsion, Paired Isolation, and Concatenation PCR), which physically links genomic DNA from specific uncultivated viruses to that of their given hosts at high throughput. These and related experimental approaches will enable validation of computational predictions, as well as reveal new virus-host linkages and dynamics occurring in nature. See Milestones **MIP02**, **MIP04**, and **MGP05**.

EUKARYOTE INTRACELLULAR ASSOCIATIONS

Endosymbiotic associations, the phenomenon of one organism living within another, greatly shape the evolutionary trajectory and ecology of life and have been found within different habitats and across the tree of life. Furthermore, phototrophic protists, for example, are important contributors to primary production and thus the global carbon cycle. Understanding associations of algae with other entities will add to our fundamental knowledge of algal biology and how the biomass-to-bioenergy supply chain and algal cultivation strategies might be affected by bacterial endosymbionts and viral infections (see also the "**Algal Genomics**" section, page **26**). Single-cell genomics provides the unique opportunity to link extrachromosomal DNA to the main chromosome by amplifying and retrieving the sequence information of all DNA entities contained in an individual cell. This enables the association of a recovered cell with elements such as plasmids, organelles, intracellular symbionts, and/or viruses. The JGI aims to use both cultivation-independent (single-cell) and cultivation-dependent methodologies to interrogate small eukaryote intracellular associations to better understand functional microbial interdependencies and relationships and their impact on ecology and DOE-relevant geochemical cycles. Single-cell approaches will be leveraged to initiate the new Genomic Encyclopedia of Small Eukaryote Intracellular Associations (GESEIA) project, and cultivation approaches will be implemented for experimentation and hypothesis-testing of novel host-endosymbiont model systems. See Milestone **MIP05**.

Exploring the Vast Phylogenetic, Ecological, and Functional Diversity of Fungi

Since 2009, the JGI has had a dedicated Fungal Program which has produced over 1,000 fungal reference genomes covering the entire Fungal Tree of Life and functional genomics resources for in-depth exploration of model fungal species. Focus areas of the Fungal Program include plant interactions (pathogens, symbionts,

and endophytes) and biotechnological applications (lignocellulose degradation and sugar fermentation) in which the JGI has provided an extensive collection of reference genomes for comparative genomics to enable new discoveries like DNA methylation, new genetic code, and fungal cellulosomes. With over a billion years of fungal evolution, the biological innovations found in fungal genomes are immense and provide the substrates for continued study at the JGI. Fungal diversity is being actively explored using genomics tools in projects like the 1000 Fungal Genomes to produce a reference genome for every known family of fungi. With active participation of the international community of researchers, significant progress is being made in sequencing nucleic acids samples from both culture collections and individual labs. This massive volume of data open new opportunities to study biology and evolution across eukaryotes.

Expanding the Fungal Tree: Sequencing Unculturables

Since the majority of fungal species are uncultivated, new approaches will be developed to extract their genetic information, such as single-cell genomics, sequencing of co-cultures, DNA extracted from laser capture microdissection, and fruiting bodies. For single-cell genomes, we will optimize protocols for cell lysis, reducing amplification biases, and linking to long-read sequencing platforms. For complex communities, new computational methods for binning, assembly, and annotation of metagenome-extracted eukaryotic genomes will be developed. To date, the JGI has conducted limited studies of uncultured fungi in contrast to substantial studies in bacteria. We will leverage our expertise with single-cell genomics in prokaryotes to increase the throughput of fungal single-cell genomics and transcriptomics. These approaches will allow us to sequence both single-cell early diverging fungi and spores of multicellular biotrophic fungi to embrace a larger diversity of fungi. Single-cell transcriptomics will also enable studies of fungal development. See Milestones **GNT05** and **FGP01**.

Improvement of Reference Genome Quality for Key Fungal Species

A substantial number of fungal genomes were sequenced early on in the genomic era using older technologies. As a result, these genomes lack parts of the full sequence due to cloning or sequencing biases, typical for early sequencing approaches, or miss genes due to limited transcriptomics studies. With recent breakthroughs in genome sequencing, long-read sequencing platforms, and development of optical and other mapping strategies, we will revisit those genomes in order to improve quality of assemblies and annotations of the key references for the species which are used in a large number of laboratories around the world. Technologies like HiC and PacBio sequencing have demonstrated their potential to produce higher quality fungal references in comparison to short-read short-range genome assemblies. Near complete reference genomes will be produced across the Fungal Tree of Life and provide a stronger basis for system biology studies of model organisms and systems for biotechnological applications and studies of plant-microbial interactions. Sequencing multiple strains or isolates from the same species will inform researchers about levels of intraspecies variation and create a basis for developing pangenome analysis tools. See Milestones **FGP02** and **EKI01**.

Multi-omics Data Production for Key Reference Species

A high-quality reference genome is just the first step toward the understanding of the biological potential encoded in DNA. Following the successful JGI Fungal ENCODE project, the JGI will produce at scale large multi-omics datasets, including transcriptomics, epigenomics, and metabolomics data. These will be integrated with phenotypic data to enable understanding of gene function and regulation through spatial and temporal patterns of gene expression, translation, and post-translational modifications, building gene networks and metabolic models. These large volumes of data for the first time will enable effective use of machine learning and new modeling techniques to find genetic determinants of fungal lifestyles and interactions between different fungal species and plants and animal hosts, bacterial symbionts, and viruses. These tools, workflows, and pipelines will be integrated with data and made available to researchers for interactive analysis on the JGI and KBase platforms. This will enable data exploration for a larger user community on various public and user-produced datasets to answer a broad spectrum of biological questions. See Milestones **FGP02**, **FGP03**, and **MTB05**.

Functionally Characterize Fungal Conserved Genes of Unknown Function

The JGI fungal genome portal (MycoCosm) is the largest fungal genome collection in the world. It contains over 1,000 sequenced and annotated genomes and continues to grow. However, on average, 40% of genes in fungi are of unknown function, ranging from 15% in yeast species studied for several decades to 70% in genomes of rusts and mycorrhizae, critical in complex interactions with plants. New scalable approaches need to be developed for experimental gene and protein functional characterization, like targeted transcriptomics, phosphoproteomics, or secretomics; DAP-Seq for DNA binding; structural genomics; and other methods. In addition, high-throughput mutagenesis using Tn-Seq or CRISPR-Cas9 may identify essential genes and point to the function of others. Multiple DOE user facilities provide access to experimental capabilities that are normally not available for a single lab or core facility. The JGI will work closely with the fungal research community to prioritize, adopt, and characterize families conserved across a large number of species using traditional biochemical approaches. Thus, this goal spreads over all five Is: **Identification** of gene clusters, **Interrogation** of their functional and evolutionary importance and conservation, **Investigation** of function using new approaches, **Integration** of experimental data and spreading annotations across the Tree of Life, and **Interaction** for community coordination and participation. See Milestones **FGP04** and **FGP05**.

Algal Genomics

Algae are a phylogenetically diverse group of eukaryotes that are collectively responsible for 50% of photosynthesis on Earth and represent major targets for third-generation biofuel and bioproduct development. In the last few years, the JGI has sequenced, annotated, and published a number of high-profile algal genomes, several of which were the first sequenced representatives of the corresponding phyla. However, the enormous phylogenetic and functional diversity of algae require a more systematic exploration to better understand their complex evolution, mechanisms of photosynthesis, and potential for bioenergy, bioproducts, and biomanufacturing (Fig. 3.6). Currently, the research community has a limited understanding of the breadth and depth of algal diversity and physiology, impeding the use of algae for basic research, bioenergy, or biomanufacturing. To address this, the JGI will develop a dedicated Algal Genomics Program for high-throughput sequencing and analysis of reference algal genomes and multi-omics data sets for functional genomics and metabolic modeling. This will be conducted in close interaction with algal research labs around the world for DOE mission-relevant foundational, as well as use-inspired, science. The tools to developing a successful program will include JGI leadership in organizing workshops and other events to bring together communities of interest (see the chapter "Interaction"), the initiation and scaling of algal data production efforts, and the development of a web-based algal multi-omics resource to facilitate community access to algal data produced by the JGI.

Organize Researchers around Algal Genomics Projects

A key role the JGI has played in the existing Science Programs is to develop and bring together communities of interest to study key organisms. In the same vein, the JGI will organize and participate in algal genomics workshops in conjunction with major community events, such as meetings of the Phycological Society of America, Algal Biomass Summit, the Algal Biomass, Biofuels, and Bioproducts conference and other symposia. The purpose of these workshops is to organize research labs around large-scale most-needed genomics projects to develop genetic blueprints for diverse algae to enable comparative and functional algal genomics. These workshops will also help the JGI develop long-term program goals. See Milestone **FGP07**.

Develop JGI Algal Products and Scale Data Production

The breadth of biological questions involving algae will require the development of a spectrum of omics products and tools for their analysis. The JGI will engage the algal research community to aid in prioritizing the most important scientific targets across several directions: identify new targets for bioenergy and bioproduct applications; further explore and develop new algal model systems; enable comparative genomics across the phylogenetic and ecological diversity to study algal diversity, evolution, and interactions; and discover new biology. We will scale-up sequencing and annotation of reference genomes and develop pangenomes across multiple strains with diverse phenotypes. Using transcriptomics, metabolomics,

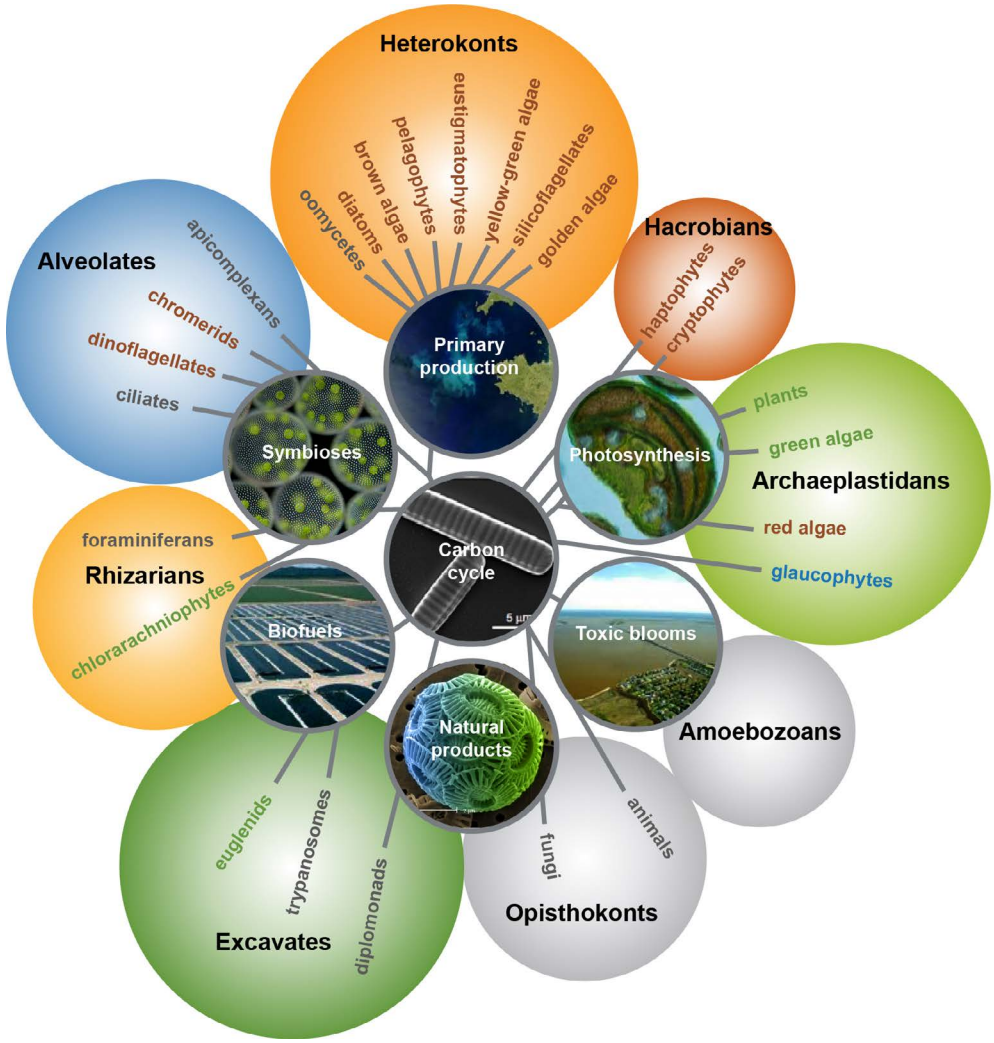


Figure 3.6. Diversity of algae and algae-related applications.

synthesis, and new in vitro approaches being developed at the JGI, as well as proteomics and other multi-omics capabilities of other user facilities through FICUS, we will expand functional characterization of algal strains. We will also target microbial consortia, combining single-cell and meta-genomics and transcriptomics, where for each type of product we will develop pipelines for data production, QC, annotation, and analyses. See Milestones **GNT05**, **MIP05** and **EKI01**.

Develop JGI Web-Based Algal Multi-omics Resources

The JGI has a long history of hosting genomic datasets. Consistent with this experience and benefiting from the scalable MycoCosm platform, the JGI will develop web-based portals for all algal projects. Sequenced genomes and multi-omics datasets will be easily accessible to the research community and equipped with interactive web-based tools for comparative and functional genomics analyses. These tools will also encourage community participation in data curation, sharing, and nominating new species for sequencing as was successfully developed in the fungal program. These resources will be cross-referenced and linked to other JGI data portals, KBase, and other genomic resources. See Milestones **FGP06** and **MTB01**.

Understanding the Diversity and Function of Plant Genomes

Plants are of critical importance to the DOE missions in energy and environmental research because of their central role as sources of biomass for energy production, as well as their impact on biogeochemical cycles and ecosystems. To support this mission, the JGI has developed extensive plant genomic resources. The JGI leads the coordination of large-scale genomic projects across multiple research groups as a user facility, but also serves as a hub to integrate data and analysis for multi-plant genomic comparisons. Critically, all data is publicly released via Phytozome, the JGI portal for plant data, which provides a comparative genomics platform for additional investigation by the scientific community. The JGI continues to support large-scale projects that intersect automated phenotyping experiments with plant diversity to identify beneficial agronomic traits. The JGI also anticipates expanding projects that seek to understand plant stress responses in controlled experiments with high-resolution genomics tools. To anchor inter-species comparative analyses, we have defined a series of "flagship" species that span the evolutionary history of plants and play crucial roles in the DOE science mission. In addition to the JGI flagship species, it is essential for the JGI to contribute to the larger community by generating and analyzing data that integrates well-established plant model systems, such as *Arabidopsis thaliana*, maize, rice, and medicago, to take advantage of the synergistic opportunities for cross-species comparison and functional analysis with JGI flagships. We have adopted one of these flagships, *Brachypodium distachyon*, as an in-house fast cycling grass model to develop functional genomic assays for use by the JGI user community. As we move into the next phase of genome-enabled plant science, we will continue to develop new reference genomes and genomic infrastructures in DOE focus areas. This includes developing additional flagships spanning plant evolutionary history and using new functional techniques to identify and uncover the causes of functional diversity in plants.

Improvements to Genomes: Towards References for Function

Reference plant genomes continue to be the foundation for plant genomics science. Advances in sequencing technology and computational methods have improved our ability to generate reference genome assemblies and annotations from more complex genomes. We will develop methods to sequence and assemble polyploid genomes using long-read single molecule sequencing and HiC technology and apply these methods to the largest plant genomes. For structural annotations, we will add resources such as long-read Iso-Seq (full length mRNA sequencing) to capture new classes of genes and annotate novel transcribed elements such as non-coding RNAs (ncRNAs), and ATAC-Seq, DAP-Seq, and HiC to annotate cis-regulatory elements including promoters, TF binding sites, and enhancers. We will create de novo reference assemblies for multiple genotypes that represent unique subpopulations of flagship species and will allow direct computation of pangenomes. As the gene content within a species varies across genotypes, we will construct pangenomes by integrating sequences from multiple diverse accessions to generate a more complete representation of the gene diversity of the species. Combined, these advances will enable us to more accurately represent and investigate the entire gene content of a species, and to document structural and presence-absence genetic diversity. See Milestones **GNT03**, **GNT08**, **PLP01**, **PLP02**, and **PLP03**.

Comparative Framework for Plant Gene Function and Regulation

Although many complete high-quality genomes across the phylogeny of plants are now available, the overrepresentation of modern crop plants among these genomes limits our ability to perform comparative genomics to understand gene function. In particular, our ability to extend the experimentally validated functional annotations from model species (e.g., *A. thaliana* and *B. distachyon*) depends on knowledge of the evolutionary origin and relatedness of gene sequences between models and new genomes. In collaboration with the larger plant research community, we will sequence additional high-quality genomes as phylogenetic reference points to infer function and identify conserved and regulatory elements. These genomes will allow us to track individual genes and gene families across evolutionary space and provide the framework for understanding how function has evolved with

changes to the genome. We have made strides towards identifying cross-species gene function networks. For example, the JGI Plant Gene Atlas provides an experimental and analytical pipeline that documents the environmental and biological context of gene expression across multiple plant species. However, the Plant Gene Atlas is constrained by available species and can be improved with new functional techniques. First, new experimental techniques for identifying open chromatin (ATAC-Seq), TF binding sites (DAP-Seq), and chromatin conformation (HiC) will allow us to link the regulatory and expression data across all of the plant kingdom. Second, we will leverage single-cell RNA-seq techniques (Drop-seq) to study the multicellular nature of plants. By surveying cell-specific expression in correlated tissues, we will advance our functional knowledge of genes and pathways at a cellular resolution. See Milestones **GNT06**, **PLP04**, **PLP05**, **PLP10**, and **EKI02**.

Common Gardens, Pangenome Diversity, and Improved Phenotyping

The economic and environmental sustainability of plant biofuel feedstocks rests on the production of biomass, both under optimal and stressed conditions. A major research goal for plant biofuel feedstock improvement is therefore to connect DNA sequence variants to plant growth and stress tolerance phenotypes. One approach to this is to plant and grow diverse germplasm in a single environment (common garden) and perform phenotyping for traits that can then be linked to the underlying genomic region responsible. The JGI supports the generation and analysis of common garden experiments conducted with users and BRCs. To improve our knowledge of genetic diversity in these experiments, we will construct and utilize draft pangenomes, built direct from resequencing data, for diversity analysis in several of our flagship species. Also, improved and automated phenotyping methods will be critical to the success of common garden efforts. Such approaches, performed either in a growth chamber or in the field, will allow for inexpensive and accurate assays of plant architecture and growth traits. Improved phenotyping accuracy increases the resolution of association studies and provides insight into intermediate phenotypes that are not observable with conventional phenotyping. We will collect molecular phenotypes including gene expression, microbiome components, and metabolomics in common garden experiments. These high-dimensionality traits can be analyzed in quantitative trait locus (QTL) experiments and integrated with whole-plant traits to increase our ability to identify causal elements and variants for plant improvement. We will develop improved statistical and computational methods that take advantage of the full phenotypic and temporal breadth of these datasets to associate genotypic variation with multiple phenotypes and give us access to unknown genes underlying important agronomic traits (Fig. 3.7). See Milestones **PLP06**, **PLP07**, and **MTB02**.

Understanding Plant Responses to Stress: Identifying Functional Elements

The extent and efficacy of plant responses to stress — limited soil nutrients or water, temperature, pests, and disease — are crucial elements of crop plant growth and yield. The plant physiological response to these stresses depends on the interactions among regulatory elements and their target genes, which collectively constitute a set of complex gene networks. We now have the functional genomic tools to begin to dissect these molecular responses. We will collect standardized genomic functional data for a series of stresses with time courses across the flagship genomes. We will apply molecular phenotyping with gene expression, open chromatin assays, and histone tag captures to identify the changes during the application of stress. Combined with the ability to easily create mutant lines, through large-scale resequencing efforts such as our *Brachypodium* mutant collection or through directed modifications with CRISPR-Cas9, we will validate the identified candidate loci and their effect on plant environmental responses. We expect that with consistent, coordinated experiments across model and crop plants, we can collect standardized genomic functional data that will enable greater understanding of how these genetic elements interact to produce beneficial plant stress responses that will directly aid in crop breeding and improvement. See Milestone **PLP08**.

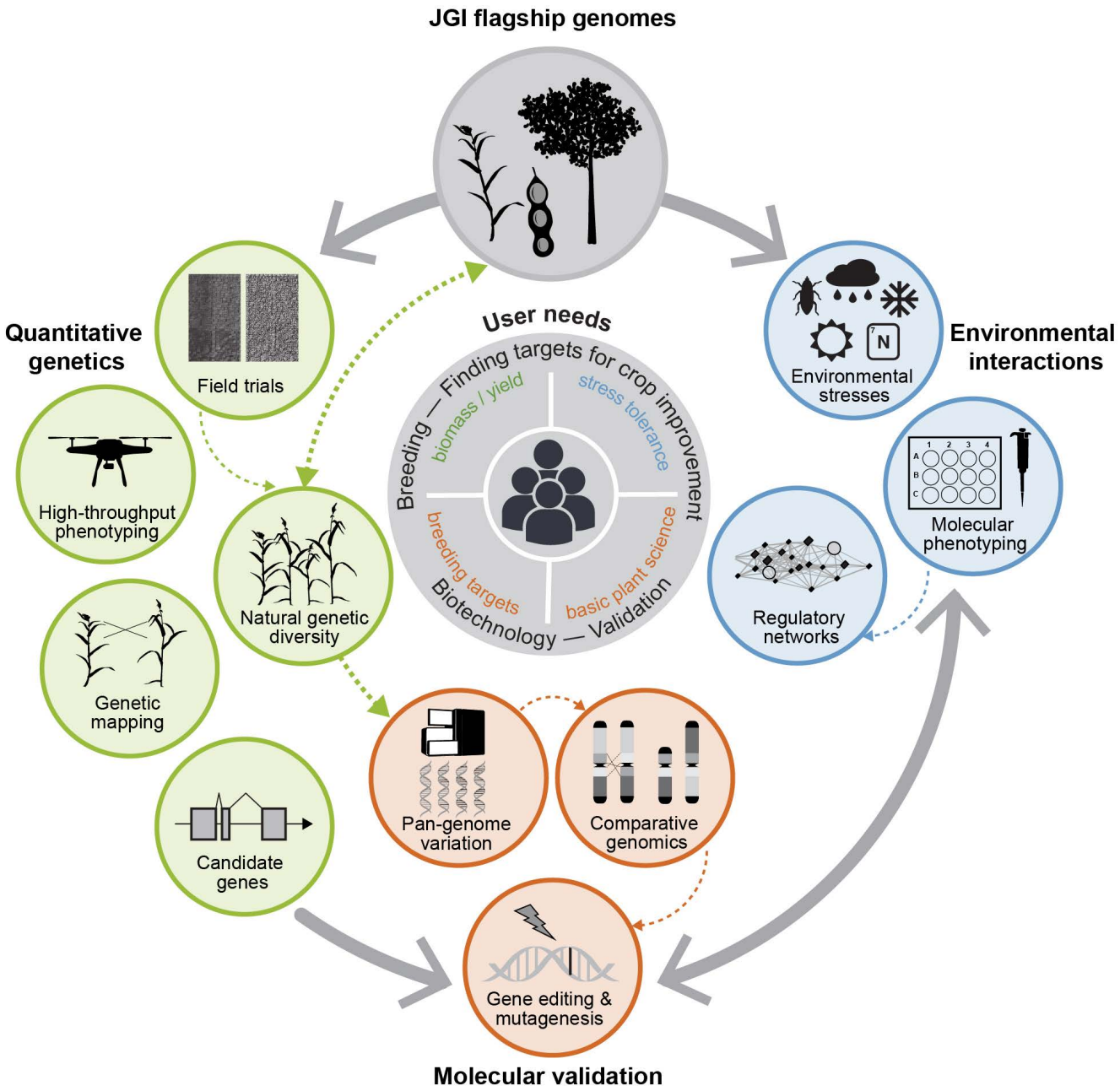


Figure 3.7. JGI user science-focused integrated discovery for plants. Genomes to phenotyping to analysis and molecular validation. Plant flagship genomes: https://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/

4. Interrogation

We envision a connected JGI that provides seamless access to data and tools through the broader biological computing ecosystem. To support this vision, we will adopt FAIR data practices, ensure all software is both portable and scalable, and engage in user-centered design for our systems.

Data Science and Informatics

The research done by the JGI and its scientific collaborators necessitates an understanding of data across different scales and dimensions, from the chemistry within cells to biological systems, across many different species and ecosystems and many different experimental conditions. While the biology underpinning these research questions is intrinsically connected, data are often collected and stored in siloed technology or domain-specific repositories. Advances in the field will be driven by mathematical and statistical machine learning approaches that bring together these disparate data repositories. The JGI produces a large number of high-quality DOE mission-relevant datasets that are foundational for these efforts.

The Integrated Microbial Genomes and Microbiomes (IMG/M), Genomes OnLine Database (GOLD), Phytozome, and MycoCosm data systems have established the JGI as a leader in the development of tools for comparative genomics, as well as metadata collection. As smaller labs get access to large data and compute resources, the JGI must push its capabilities further. There is an increasing demand for platforms that allow the scientific community to ask complex research questions across different data resources at the JGI and beyond. The JGI will help lead a transformative change in biological data science through the deployment of new tools, data, and standards through partnerships with KBase, DOE user facilities, and other genomics centers. Also, adoption of FAIR data practices will enhance the JGI's efforts to build and support data access and integration efforts.

The JGI has access to enormous computing resources through the Advanced Scientific Computing Research- (ASCR) funded compute facilities. The close partnership with NERSC began in 2010 and has led to a number of collaborations including the JGI-NERSC FICUS, ExaBiome, and HipMer projects. The JGI will leverage these unique national resources to run analyses and generate derived data products that would not be created otherwise.

KBase is a powerful platform for genomic data analysis and metabolic modeling. The JGI and KBase are developing complementary but integrated high-performance tools providing users with the ability and infrastructure to explore complex and diverse datasets to extract deeper biological insights. The goals of this partnership are to create a JGI presence within KBase, build a diverse, engaged joint user community, and enable scientific discovery.

Data Science Facility Partnerships

The JGI formed a valuable partnership with NERSC that has helped support our growing computational demands. In 2010, the JGI moved all of its scientific computing resources to NERSC. NERSC and the JGI collaborated to revamp the JGI's hardware and software infrastructure, enabling large increases in both analysis and R&D throughput on the high-performance cluster, Genepool. NERSC staff have worked closely with the JGI teams to improve operational efficiency, explore new methods for analysis on high-performance computing (HPC) resources, and deploy an effective hierarchical data management system.

Recently, the JGI has explored a partnership with the Leadership Computing Facility (OLCF) at Oak Ridge National Laboratory (ORNL) to host a copy of JGI's data as an off-site backup, and to enable easier access for ORNL research staff and any research scientist with access to OLCF. The JGI has provided data to the computations for the exascale runs of the CoMET software. The algorithm in this software takes advantage of machine learning approaches and large volumes of data to explore genotypic and phenotypic patterns across the tree of life. The JGI will continue this partnership with ORNL in an effort to develop and apply new methods for determining the connection between

gene sequence and function across our Science Programs. See Milestone **DSI02**.

Scalability

The global output of omics data is growing at a rate outpacing Moore's law. In order to meet the demand to process, interpret, and interact with these data, the JGI must rework its data platforms, analysis techniques, and integration methods. Our data platforms must be responsive and more modular. In addition, these systems must meet user demands for additional features. The JGI can leverage advances in data science, specifically machine learning and AI, to increase the speed of data analysis and enable discovery. In addition, the ExaBiome Project will provide the JGI and the scientific community with high-performance tools for fundamental analyses, such as assembly, clustering, and alignment, as well as new approaches for the integration of data from different experimental systems, including light sources and distributed sensors, as well as omics data. These developments will be supported by the JGI's efforts in the FICUS program and initiatives in the Biosciences Area at Berkeley Lab.

Scalable Data Processing

The JGI will build on the success of the AutoQC tool developed in the Genome Analysis Group to reduce the number of manual interventions needed to assess the quality of the data from the sequencers. The team will take advantage of statistical methods and unsupervised learning techniques to automatically flag the subset of data that requires further assessment from a JGI analyst. These methods work well because the JGI has a huge amount of operational data for successes and failures that can be used to train different classifiers.

Scalable Data Analysis

The size of JGI data sets is increasing, particularly as interest grows in understanding spatial and temporal effects in metagenomes. Efficient analyses of these data require enormous computational power that is provided by NERSC as well as the ASCR compute facilities. Bioinformatics software must be redesigned and algorithms developed to take advantage of supercomputing resources, and this work is underway as a component of the ExaBiome project. The JGI will support further efforts in this area through collaborations with the CRD at Berkeley Lab.

ExaBiome: Scalable Methods for Complex Genome and Metagenome Assembly

A persistent challenge in genomics has been the rapid and accurate assembly of complete genomes from high-throughput shotgun datasets derived from both individual organisms and microbiomes. As a principal producer of plant, fungal, and metagenome datasets, the JGI must keep pace with the growing demand for genome and metagenome assembly. To this end, JGI data scientists have joined with computational scientists from the Berkeley Lab CRD and NERSC on the ExaBiome project, which will develop the next generation of scalable genome and metagenome assembly and analysis methods. There are two frontiers for this effort. See Milestone **DSI02**.

EXASCALE METHODS FOR EUKARYOTE GENOMICS

The first frontier is the need to assemble multiple individuals from a species to capture its full "pangenome." With projects involving hundreds or thousands of plants, fungi, and algae, JGI scientists and users need scalable methods for rapidly and accurately reconstructing and representing related collections of genomes to capture the possibly rare variants that include novel genes not represented in the "reference" genome, or which have disrupted or deleted genes relative to this reference. As part of the ExaBiome project, JGI scientists and collaborators will develop and implement new tools to address this challenge. See Milestone **EKI03**.

EXASCALE METHODS FOR METAGENOMICS

The second frontier is high-throughput metagenomics, which has unique potential to benefit from the synergistic characterization of microbiome communities in space and time if these datasets can be converted from terabytes of short-read data into annotateable genome sequences. Data volume in the National Center for Biotechnology Information short-read archive is doubling every 11 months and is expected to reach about a million metagenomes in 2020. Many of these datasets have never been assembled, even partially, due to the memory requirements of current production assemblers. The quality of the current assemblies is often quite poor, in part because datasets must be reduced to manageable size before assembly. Capitalizing on expertise in genome assembly and parallel computing algorithms, the JGI and the ExaBiome project are building scalable tools for three core computational problems in metagenomics: (1) metagenome assembly,

(2) protein clustering, and (3) signature-based approaches to enable scalable and efficient comparative metagenome analysis. See Milestone **PKI03** and **EKI03**.

Scalable Data Access

Data are of little value if they cannot be found, accessed, and manipulated by the scientific community. We will improve the design of our data systems and engage in efforts to ensure the JGI adheres to FAIR principles. For example, we plan to stand up endpoints that are similar to the Global Alliance for Genomic Health "beacons" that are standardized application programming interfaces (APIs) for determining if a data set contains a particular gene.

USER-CENTERED DESIGN

The JGI's user community is at the heart of our mission, and "data users" who take advantage of the JGI's data portals are an increasingly important part of this community. Focused efforts on designing systems that work well for the users, who are primarily not geolocated with the JGI, have the potential to amplify our scientific impact. We envision a mode of operation where users come to JGI platforms for the data, but stay on the platforms because the tools for comparative analysis and exploration are high-quality and enhance their productivity. Specifically, the JGI will engage in a usability campaign to determine high-priority targets for the KBase collaboration, to improve the JGI search in the Genome Portal, and to identify the different features that may enhance the utility of our data platforms. We will achieve this understanding through the development of surveys, participation in the KBase User Working Groups, and engaging usability experts in the Data Department in CRD at Berkeley Lab. In five years, we envision connected, responsive JGI data platforms that serve the broader scientific community. See Milestones **DSI05** and **DSI06**.

DISTRIBUTED DATA ACCESS

As the JGI faces increased demand for access to large fractions of its data corpus, there must be corresponding infrastructure in place to support downloads and data movement. The JGI is fortunate to have access to ESnet, which provides 100Gb/s connectivity within the DOE national lab complex and Europe. This will be further enhanced by the deployment of ESnet6, a highly optimized networking infrastructure that will enable large amounts of data to stream between facilities. To better serve the community, the JGI will make a mirror of its data at the OLCF; if there is demand, the JGI will make additional mirrors at partner facilities such as Pacific Northwest National Laboratory (PNNL), or Argonne National Laboratory (ANL). The JGI will rewrite its data management system tool, the JGI Archive and Metadata Organizer (JAMO), to facilitate seamless data movement between computing resources across the DOE complex. See Milestone **DSI03**.

FAIR Data Principles at the JGI

One key element to making datasets, context, or genomic features findable in any system is a powerful search capability. A corollary to search is the ability to integrate and visualize data in a meaningful way. Search and visualization enables scientists to explore data systems interactively, a critical component of any data integration approach.

FINDABILITY: EXPLORATION AND INTERPRETATION

The exploratory power of a single interface from which many data sets can be searched has the possibility to dramatically advance large-scale, integrative analyses. In order to build such a system, the data must be described by rich metadata, have persistent unique identifiers, and be indexed in such a way that a search returns sensible results. These efforts will be coordinated with existing work in the broader community through the JGI's partnership with KBase, JGI staff participation in the Genomics Standard Consortium, and the Institute of Electrical and Electronics Engineers (IEEE) BioCompute Object collaboration. The JGI has a solid foundation for this effort with the GOLD database, JAMO, and the Data Warehouse. The JGI will also issue a unique Digital Object Identifier (DOI) for each user proposal describing all data products associated with that proposal, and will explore the use of DOIs for large data sets and queries. See Milestones **USP03** and **DSI03**.

ACCESSIBILITY: OPEN AND FREE

The JGI must track the access and use of its data resources; however, it does not need to build its own infrastructure for doing so. The JGI is collaborating with other DOE user facilities to take advantage of a common identity management system. A lot of work and focus has gone into the investigation of the Open Researcher and Contributor Identifier (ORCiD) as a platform for tracking user identities, affiliations, and publications. In 2018, the JGI began a renewed effort to define and track metrics associated with so-called "data users" that access only the JGI's digital resources. Going forward, the JGI

will include these users in the annual report of user data to DOE, and engage data users directly to ensure their needs are being met by the JGI's systems.

INTEROPERABILITY: AMPLIFY BY LEVERAGING STANDARDS

The European Bioinformatics Institute and others have adopted the World Wide Web Consortium resource description framework model to represent data and SPARQL to enable queries across siloed data infrastructure. The JGI will leverage this model when refactoring and rebuilding data infrastructure, as it has not only been shown to work well for multi-omics data, but also provides access to a large number of query and analysis tools that have been built by the bioinformatics and semantic web communities.

REUSABILITY: WRITE ONCE, RUN ANYWHERE

In 2015, the JGI made a concerted effort to migrate analysis pipelines to containers. In principle, it is now possible for any member of the scientific community to download and execute JGI tools and pipelines that are open source (i.e., do not have any hidden dependencies). The JGI user community can access these tools through the KBase platform today. In the future, we expect to make all pipelines available from public-facing repositories. The JGI data and pipelines will have a clear policy and usage agreement. See Milestone **DSI01**.

Reproducible Analysis

The JGI will leverage technologies that enable pipelines and analytical tools to run on JGI-owned systems, as well as the commercial cloud, laptops, and other DOE compute facilities. In five years, the JGI will be able to execute computations and analyses on any available computational resource from DOE HPC facilities across the United States to the commercial cloud. The JGI envisions the seamless exchange of data and pipelines between scientific organizations and research groups.

JGI users will be able to access all JGI-developed pipelines and tools through the KBase platform, the JGI-NERSC-KBase FICUS, and existing web systems. The JGI will make all high-value software and pipelines available through open source repositories, and will provide support for these tools through standard tickets and issue tracking. See Milestone **DSI01**.

Opportunities in Machine Learning

The JGI has a wealth of scientific and operational data that have been collected since its establishment as a user facility. With recent advances in machine learning and AI, the JGI can harness these data to address the knowledge gaps that are of critical importance for energy and environmental science. In order to understand and therefore manipulate or engineer biological systems, we must gain a comprehensive understanding of the higher order molecular interactions that occur within and across all omic layers, as well as the environment, that lead to emergent behaviors and organismal- and/or population-scale phenotypes. This is a computationally intractable problem if we fail to account for the physical, chemical, and evolutionary constraints that apply to these systems. However, we lack the necessary mathematical models to apply these constraints and utilize machine learning techniques effectively.

Bioinformatics requires a sound foundation in statistics and experimental design, two critical components for the effective use of machine learning methods. The JGI is uniquely positioned to leverage the skills of our staff, our wealth of scientific data, and computational resources, combined with partnerships with applied mathematicians and computer scientists, to build the tools needed to enable systems level understanding. See Milestones **DSI03**, **DSI05**, **DSI06**, and **DSI07**.

The JGI will use operational data and machine learning tools to address questions about data usage and our platforms, and ensure we continue to meet the needs of our user community, now more broadly defined to include our many data users. We will also use data collected in a central workflow management system to characterize and predict memory, wall time, and computational resource usage to both avoid failure and waste. This will be a valuable tool for staff and possibly our KBase collaborators. This will also be an excellent way for staff to gain the necessary skills and understanding to tackle more complex statistical problems. See Milestones **DSI04** and **JLT07**.

We will partner with the Computer Science and Applied Mathematics departments at Berkeley Lab, as well as other institutions, to develop new AI techniques to produce explainable models. We envision working with our partners to develop a new workflow for biological data science depicted in Fig. 4.1. The models generated by this effort can then be applied to the scientific data at the JGI.

The output of these efforts will be made available in a number of machine learning-based microservices. The JGI will provide access to common algorithms applied to the JGI's data sets that can be used by the community to answer common scientific questions, ranging from protein clustering to gene homology. The backbone of these tools will be high-performance implementations of the software that leverages DOE ASCR compute resources at NERSC and other facilities (for example, HipMCL). See Milestones **DSI04** and **DSI05**.

In five years, we hope to have models for use in machine learning methods that allow us to explain underlying biological mechanisms (Fig. 4.1).

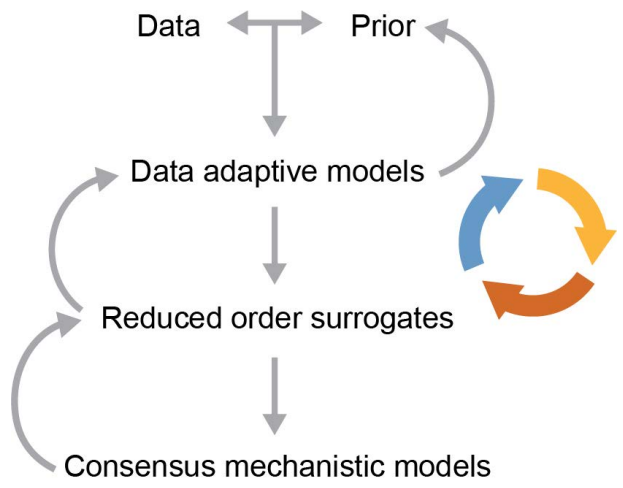


Figure 4.1. Proposed method for developing explainable AI models where it is currently difficult to develop mechanistic models.

Data Platforms

The JGI's data platforms are a key resource for the broader scientific community and require constant developments to meet the ever-changing demands of the users. See Milestones **DSI03** and **DSI08**.

A Next-Generation IMG/M System to Provide Petabytes of Microbiome Data to Thousands of Data Users

Key to achieving our microbiome data science goals is the continuation of our efforts in large-scale microbiome data processing and integration. Over the last 10 years, we have shown through a number of seminal publications that breakthrough discoveries from sequence data in microbiome research depend on comparative analysis against a large number of uniformly processed datasets, which has been the cornerstone of IMG/M's success. In the next two to five years, we plan to expand IMG/M's capabilities through continuous upgrading of all of its components, including the speed and accuracy of the data processing pipelines, the efficiency of data integration, and the friendliness and scaling of its user interface. Moreover, new database technologies that would allow us to continue integrating an increasing amount of data will be explored. Of special importance in this direction is the adoption of community-agreed standards as well as development of new genomic standards in coordination with the community. New key directions include the scaling of binning and phylogenetic classification of MAGs, development and/or integration of important statistical analysis methods, and exploration of new machine learning approaches. We anticipate that the KBase-JGI partnership will further contribute to the development of IMG/M, and that a number of new analytical and visualization tools will be jointly developed between the two teams. See Milestones **FGP01**, **PKI04**, and **PKI05**.

In addition to the central IMG/M system, a number of smaller data marts have been recently developed to support new strategic directions and attract new user communities. First, the JGI recently launched IMG/VR, the largest publicly available database with isolate reference DNA viruses and uncultivated viral genome fragments from thousands of ecologically diverse metagenome samples. This platform facilitates annotation and advanced analysis capabilities for viral sequence data for the broader research community. IMG/M also houses the IMG-ABC, Atlas of Biosynthetic Clusters, that serves as a powerful resource for the natural products community. As described in the "**Secondary Metabolites**" section (page **47**), we will overhaul and expand this resource to incorporate new biosynthetic gene cluster tools and pipelines. See Milestones **GNT11** and **PKI06**.

MycoCosm: Expanding and Enriching the JGI Fungal Genomics Resource

The MycoCosm Fungal Genomics Portal is the world's largest collection of sequenced and annotated fungal genomes, offering over 1000 genomes across the Kingdom Fungi and multi-omics datasets equipped with web-based interactive tools for comparative

genomics, functional analysis, nominating new species for sequencing, and community annotation. Over the next five years, the JGI will double the number of reference genomes to explore larger fungal diversity for DOE science and applications, enrich selected reference genomes with multi-omics datasets including epigenomics, transcriptomics, proteomics, metabolomics and additional new type of data, cross-reference data in public databases such as KBase and other specialized fungal or functional data repositories, and further develop APIs and interactive tools for phylogenetic analysis, comparative and functional genomics, metabolic modeling, and gene network analysis. We will actively utilize new information technologies and modular data organization and software architecture for customizable and efficient large-scale user-driven data analysis. This platform will also serve as a basis for developing a new algal resources portal. See Milestones **FGP01**, **FGP02**, **FGP03**, **FGP04**, **FGP05**, and **FGP06**.

Phytozome: Delivering Data to Plant Communities

The Phytozome Plant Comparative Genomics Portal will continue to evolve to meet the increasing scale and variety of plant genomics data and post-sequencing analyses. For the JGI, the main drivers of plant genomic data scale over the next two to five years are surveys designed to fill out the land plant phylogeny, and deep sampling of closely related species (e.g., the *Brachypodium* Pangenome project). These projects can each yield tens to hundreds of high-quality plant genomes and associated omics datasets. We will therefore continue to develop the informatics infrastructure of Phytozome to rapidly and efficiently handle the integration, retrieval, analysis, and display of hundreds of plant genomes. This will involve an increased reliance on horizontally scalable backend data systems and lightweight server-side containers, combined with efficient client-side visualization components, as are being initially deployed with Phytozome v13. We will also expand the types of data and analyses available in Phytozome to include chromatin accessibility, variant impact on both genic and regulatory elements, and combined metabolomic and transcriptomic support for inferred gene function. Finally, with the goal of increased integration of JGI data with other platforms (like KBase) as well as increased FAIR compliance, we will continue to expose more Phytozome data to programmatic access via well-documented and highly performant data APIs which will enable tighter integration and data sharing with other analysis platforms. See Milestone **PLP11**.

Scientific Strategic Thrust: Microbiome Data Science

The exponential growth of microbiome data, coupled with recent technology advances in processing, analysis, and visualization of "big data," is creating new opportunities for breakthrough discoveries (Fig. 4.2). Traditionally, most microbiome studies have relied on generating new sequence data, which are typically analyzed in isolation or occasionally in the context of a few related studies. While this approach has worked well for developing a better understanding of specific microbiomes, it has limitations for developing generalized models of global scale. The cross-cutting nature of microbiome research in health, agriculture, bioenergy, and the environment necessitates the development of new solutions and community coordination to tackle grand challenges that will accelerate basic discovery and lead to transformative advances. The JGI's IMG/M system was created with the goal of providing an unparalleled integration of high-quality processed microbiome data. Through leveraging IMG/M's compendium of integrated data, processed in a high-quality and standardized manner, the JGI is well positioned to expand research efforts in collaboration with the broader scientific community in the nascent field of microbiome data science.

Over the past few years, the Prokaryote Super Program has demonstrated the power of microbiome data science to tackle grand challenges. The JGI and its users have uncovered the global diversity of alternative genetic codes, unearthed hundreds of thousands of new virus genomes and their predicted hosts, discovered genes with important biotechnological applications (e.g., new CRISPR-Cas9 variants), and recovered many thousands of genomes from uncultivated microbes. A number of big data analysis efforts in microbiome research have already been launched and are expected to be completed in the next two years, while several new ones will be initiated. Moving forward, the JGI aims to maintain and further expand its established leadership role in microbiome data science by tackling five major opportunity areas to support the community (Table 4.1).

Table 4.1. Opportunities in Microbiome Data Science

Grand Challenge	Opportunities for the JGI
Establish a central comparative analysis resource in microbiome data science	Integrate critical datasets and data types in IMG/M to support comparative analysis at a global scale Maintain state-of-the-art data management systems to support the user community
Maintain curated, high-quality microbiome data	Develop and implement quality control tools and measurements for microbiome data
Successfully scale computational capabilities at the Exascale level of data generation	Develop and implement new state-of-the-art data processing and analysis methods
Provide access to microbiome data science capabilities to a global user community	Develop new or expand the JGI's existing user programs to support microbiome data analysis Connect various Berkeley Lab microbiome data science resources (i.e., KBase, NERSC, IMG/M, GOLD, etc.) to provide unparalleled data analysis capabilities to the user community
Support the community to develop and implement microbiome data and metadata standards	Expand the JGI's contributions to international community-standards efforts, such as the Genomics Standards Consortium

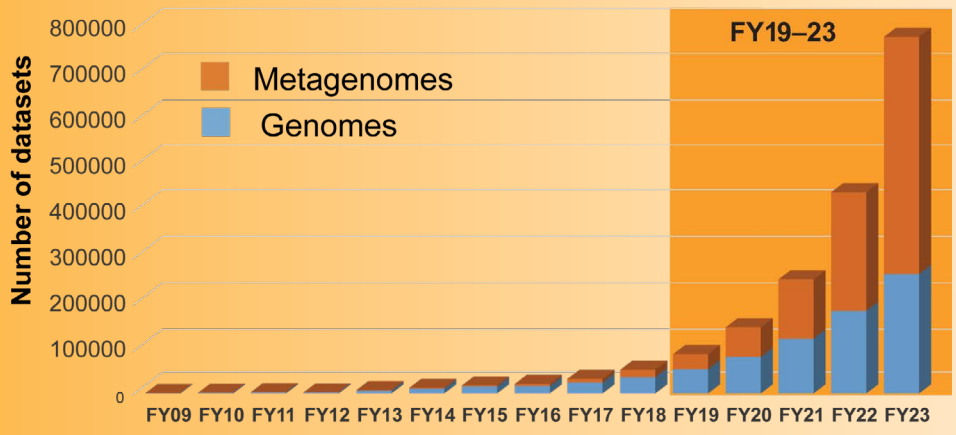


Figure 4.2. Growth in microbiome data. Anticipated microbiome data growth from 2018—2023, following a linear increase based on historic growth.

Beyond the JGI's commitment to infrastructure support for microbiome data science, we highlight key scientific drivers that push the frontiers of microbiome data science (Fig. 4.3).

CHARACTERIZE THE FUNCTIONAL DARK MATTER WITHIN MICROBIOMES

While improvements in metagenome assembly and binning methods have enabled new avenues to genome-centric analysis and interpretation, there remains significant "functional dark matter" encoded within environmental microbiomes. This functional dark matter can be defined as novel protein families with no or minimal sequence similarity to genes found in known isolate genomes or represented in current protein family databases. To tackle the annotation gap for these novel, unknown protein families, the JGI has undertaken a large-scale protein clustering approach linked with de novo structure prediction calculations in a unique marriage of microbiome data science and structural biology. These approaches hold potential to generate new insights into functional dark matter through identification of novel metabolic properties and prospective new enzymes with important biotechnological applications. See Milestone **PKI01**.

EXPLORING METATRANSCRIPTOMES FOR NOVEL RNAS

New research directions are currently under design for exploration of the RNA world, where metatranscriptomes can be leveraged to explore and characterize RNA viruses, regulatory RNAs, and transfer RNAs. Small regulatory RNAs have traditionally been characterized within model microbes, and are known to be involved in important functional processing including photosynthesis, quorum sensing, and biosynthesis of amino acids. This presents an opportunity to expand the diversity of these regulatory components through querying microbiomes and associating expression and putative ecological function at a global scale. Similarly, large-scale data mining efforts to search for transfer RNAs or new translational components will contribute new information on an expanded understanding of genetic code. Furthermore, new opportunities are emerging for global mining of metatranscriptomes as a result of the rapid growth of these datasets. Specifically, the identification of expressed genes, pathways, and biosynthetic clusters for important metabolites under specific environmental conditions can now be explored at a massive scale.

IDENTIFY NEW VIRUSES FROM METAGENOMES AND PREDICT THEIR HOST INTERACTIONS

The JGI has made substantial progress over the past few years to uncover a wealth of uncultivated viral genomes using shotgun metagenomes and metatranscriptomes. New computational approaches now hold the potential to reveal a more expansive diversity of viruses. Given the divergent nature of many uncultivated environmental viruses, new search methods are required to further capture diversity, as well as computationally predict microbial hosts. The JGI aims to leverage machine learning algorithms to more fully capture the virosphere, including both DNA and RNA viruses, and integrate with experimental approaches to validate divergent viruses and their predicted linkages to microbial hosts. Dissecting viral co-infection networks at a global scale, through the help of CRISPR data, will also elucidate the mechanisms and patterns of viral infection. Additionally, the JGI aims to assess giant virus distribution, diversity, and coding potential on a global scale that will redefine shared gene content and phylogenetic markers across all giant virus clades. See Milestones **MIP06** and **MGP05**.

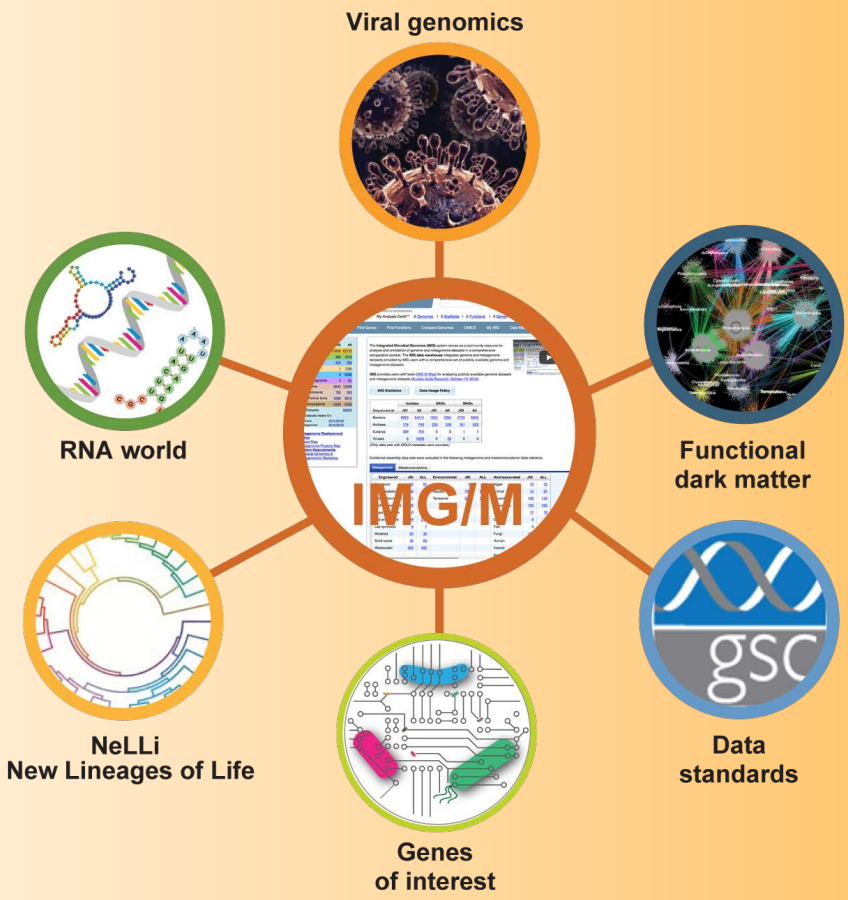


Figure 4.3. Strategic directions in microbiome data science. Scientific directions for the next five years. These directions include (1) significant expansion in viral genomics with the development of new methods for the identification and classification of viruses, as well as viral host prediction, identification of RNA viruses, and development of viral co-infections networks; (2) new approaches for the identification and characterization of "functional dark matter;" (3) development of new genomic standards in collaboration with the larger scientific community and organizations such as the Genomics Standards Consortium (see Milestone **PKI02**); (4) identification of new variants and members of known protein families with important biotechnological applications (e.g., identification of novel Cas9 variants, or identification of novel subfamilies with distinct functionality); (5) identification and characterization of new phylogenetic lineages (see Milestone **MGP04**); and (6) exploration of the RNA world with identification and characterization of novel RNA gene families, as well as exploration of metatranscriptomic data for the identification of habitat (or other physicochemical specific) expression patterns.

INTEGRATING MACROECOLOGICAL THEORY AND MICROBIOME DATA SCIENCE

Applying and adapting macroecological and historical biogeographic principles to microbial systems shows great promise for developing an improved understanding of the underlying ecological processes and importantly, forecasting the behavior of microbial systems (e.g., responses to disturbances). Five core ecological processes are generally understood to underpin the assembly of all ecological communities: (1) dispersal limitation, (2) selection, both habitat filtering and biotic interactions, (3) historical contingency, (4) mutation, and (5) neutral processes. While a current major focus of microbial ecology is to understand the relative roles of these processes, there is tremendous potential to leverage macroecological and historical biogeographic theory as a hypothesis-driven approach to microbial ecology. As part of a new focus area, the JGI will apply classical macroecology theory to large-scale microbiome data to uncover principles of assembly, patterns, and processes in microbial systems. Specifically, the JGI aims to explicitly incorporate spatial, temporal, and phylogenetic scale to evaluate mechanisms of microbial community assembly through the application of new computational methods and statistical approaches. See Milestone **MGP06**.

Eukaryote Data Science

Representation of Eukaryotic Pangenomes

The standard paradigm for working with genetic variation focuses on aligning resequencing data to a single reference genome, cataloguing single-nucleotide variants, and attempting to associate these sequence differences with phenotypic variation. It is increasingly recognized, however, that the genomes of individuals within a single species are structurally variable, leading to gene presence/absence polymorphisms that can have major phenotypic impact. This is especially true for plant genomes, which often encode notable partial gene redundancy due to polyploid ancestry, and whose extensive complement of active transposable elements provides a mechanism for the introduction of gene deletions, duplications, and translocations far beyond what is found in human genomes (Fig. 4.4). For these systems, the proper representation of genetic information is not a single linear sequence, but rather a graph that captures diverse structural and sequence features that can be recombined through breeding and/or engineering to produce individual genomes with desirable features.

Current genomic data systems are designed around linear reference genomes, but it is clear that these need to be adapted and extended to represent genes and regulatory elements on branching structures that embody the full variation found in eukaryotic species. This is a rapidly developing area, and JGI data scientists will take advantage of the worldwide community of bioinformaticists to test and adopt scalable solutions to these challenges. See Milestone **EKI01**.

An Integrative Toolkit for "Actionable Genetics"

JGI scientists and collaborators are producing an increasing number of large and complex datasets that decorate genomes (and pangenomes) of plants, algae, and fungi with diverse types of complementary functional information. These include catalogues of genetic sequence and structural variation; gene expression levels across development, tissue types, and under diverse environmental conditions, both in bulk and at single-cell resolution; and a rapidly growing set of datasets that measure epigenetic states. With genomes of hundreds of plants and algae, and thousands of fungal genomes expected in the next five years, we anticipate that cross-species and/or cross-clade comparisons will allow rapid transfer of functional information from model systems to aid in the design of eukaryotes with agricultural and industrial applications.

Over the next five years, JGI data scientists will extend the capabilities of Phytozome and MycoCosm to include toolkits that allow users to interactively integrate such diverse datasets from the JGI and elsewhere. These tools will enable researchers to rapidly develop testable hypotheses and begin to design organisms with predictable combinations of traits. Given the JGI's position as a collaborative research hub, JGI data scientists will be at the forefront of testing the value of new combinations of data types working with users. To maximize the synergies with KBase, the JGI will provide fully documented APIs providing programmatic access to JGI data, including gene annotations, gene families, orthology, reference sequences, diversity datasets, and expression datasets. See Milestone **EKI02**.

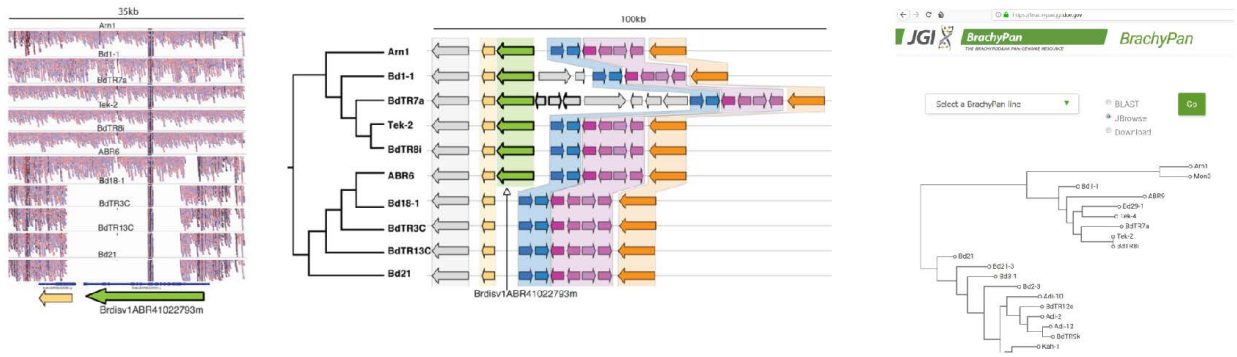


Figure 4.4. Analyzing pangenomes at scale. Left: Pangenomes of a species are much larger than the genomes of reference individuals due to extensive presence-absence variation (PAV). Center: Analyzing PAV using existing tools is laborious and cumbersome. Right: Current databases that present pangenomes are little more than aggregations of individual genomes. The JGI will develop new assembly, variation analysis, and functional annotation approaches to effectively utilize pangenomes for plant breeding and research. *Modified from Gordon et al. 2017 Nature Communications and* **https://brachypan.jgi.doe.gov**.

5. Investigation

As a result of the tremendous advances in DNA sequencing, we have, and continue to accumulate, vast amounts of sequence information from all types of organisms and environments. Functional genomic technologies, including gene expression, epigenetic profiling, and large-scale protein and metabolite, measurements, as well as powerful imaging modalities, are providing unprecedented views into the activities encoded in DNA. However, we are far from a predictive level of understanding. Currently, we do not know the functions for about half of the genes in the genomes of most environmental microbes (Fig. 5.1). Even less is understood about how to scale from the genes in a genome to the phenotypes and activities of an organism and then, in-turn, move from individual organisms and genomes to highly interdependent communities of organisms. Presumably a large fraction of the poorly annotated genes in environmental organisms are only functional under specific environmental conditions or in the context of other species with which they interact. Hence, discovering the function of such genes will require dedicated efforts and studying them in an ecological context.

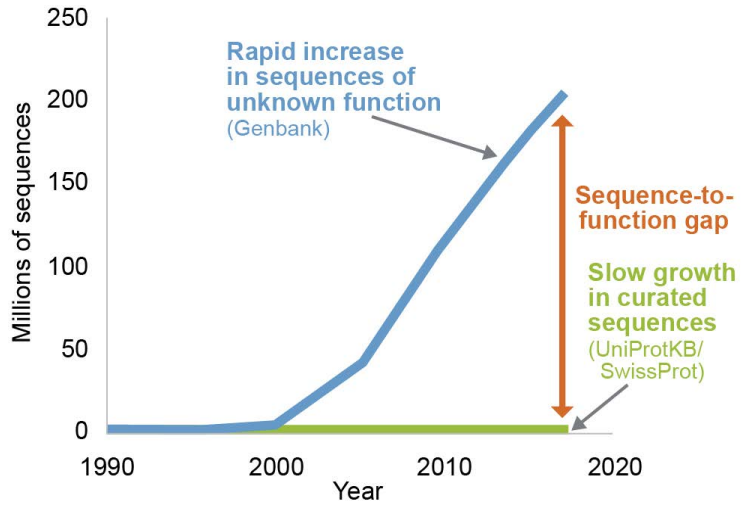


Figure 5.1. A widening sequence-to-function gap. The growing disparity between the rates of gene discovery and functional protein characterization results in a pressing need for scalable functional annotation methods.

From Sequence to Function: Leveraging DNA Synthesis

DNA synthesis affords the ability to precisely probe gene, pathway, and organism function by bringing engineering principles to biology through the design-build-test-learn (DBTL) cycle. As the technologies for designing functional constructs, generating synthesized DNA parts, accurately assembling these into long fragments, and expressing these in controlled fashions have rapidly evolved, this has paved the way for experimentally validating gene and pathway function in both isolated and genome-wide contexts. At the JGI, we have established a powerful DNA Synthesis and Strain Engineering platform to enable users to explore and validate hypotheses and an impactful DNA Synthesis Science Program that harnesses this platform for discovery and applications relevant to the DOE mission (Fig. 5.2).

DNA Synthesis and Strain Engineering Platforms

EXPECTED GROWTH IN DNA SYNTHESIS

The JGI DNA Synthesis platform provides end-to-end capabilities to enable users to validate hypotheses generated from the mining of genomic data, by designing and building physical DNA constructs that can be functionally characterized in vivo or in vitro. This platform has grown nearly fourfold since 2014 reflecting (1) the increasing demands from JGI users; (2) the drop of DNA synthesis cost due to technology advancement; and (3) the implementation of new design tools that can turn hundreds of users' genes and pathways into synthesizable, overlapping DNA in matter of minutes, a process that used to take weeks. Moving forward, we expect these advances to continue, which will expand our

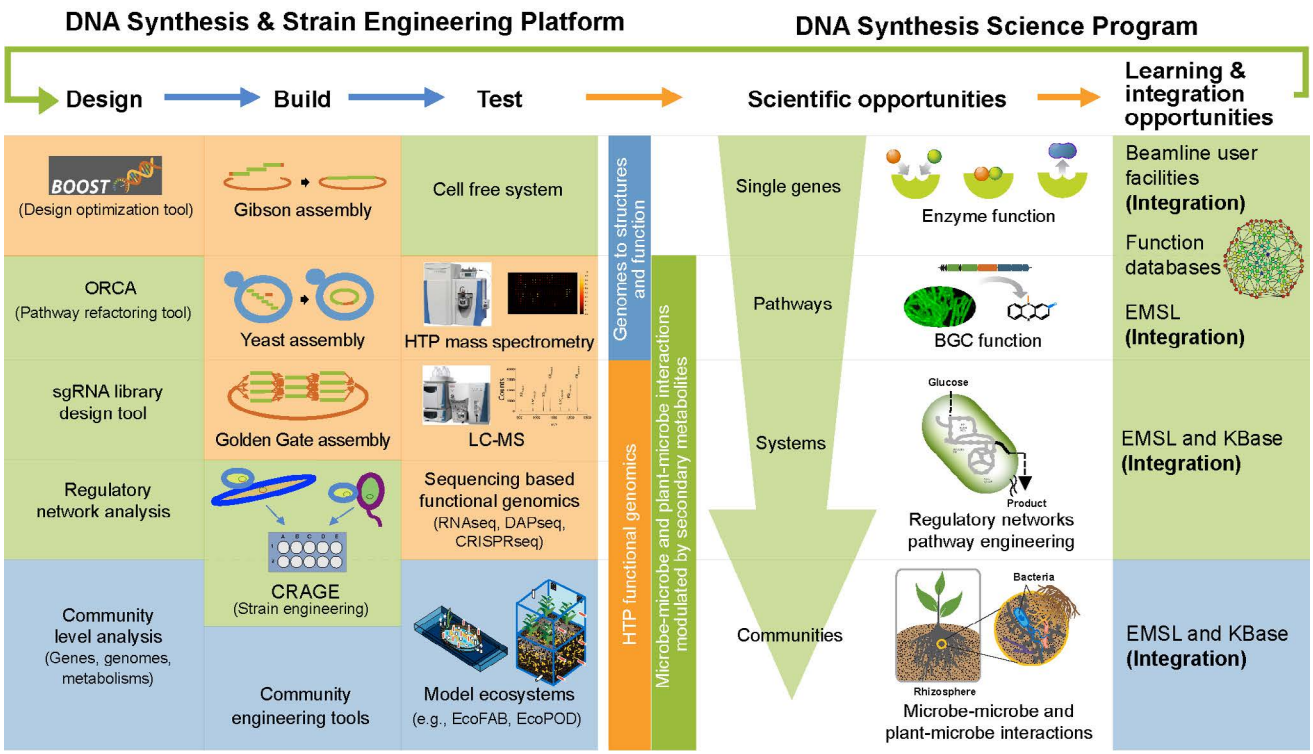


Figure 5.2. Overview of the JGI DNA Synthesis and Strain Engineering Platform and the DNA Synthesis Science Program (Tan: capabilities currently offered to users, Green: capabilities we will offer to users in two years, Blue: capabilities we will offer to users in five years).

output and facilitate contributions to an increasingly diverse portfolio of user-accessible capabilities. This will be accomplished by continued close interactions with commercial DNA synthesis vendors to gain access to advanced synthesis technologies. DNA synthesis has evolved from microtiter plate to silicon chip base operation, which resulted in a significant cost reduction. Another new advancement on the horizon is enzymatic DNA synthesis, which can achieve better accuracy and longer length than the current phosphoramidite chemistry. The JGI will evaluate and, where appropriate, adopt these new types of synthesis technologies and products as they emerge.

Another opportunity for growth is provided by the adoption of increasingly automated processes for DNA assembly and cloning. The JGI's current DNA construct assembly and cloning processes use a series of independent robotic platforms. This modularity in design offers flexibility for a process that is constantly improving, but requires manual interfaces between modules. To reduce labor costs and standardize output construct quality, we envision a fully automated process for the most commonly used, multi-step Gibson assembly and cloning procedures. The fully automated process will be executed through a central scheduler that controls various tasks on an integrated multiple-robotic platform. These advances will continue to position the JGI at the forefront of DNA synthesis capabilities and throughput. See Milestones **BDT04**, **GNT12**, and **GNT13**.

DNA SYNTHESIS INFORMATICS OUTLOOK

Our DNA synthesis informatics efforts over the past two years have been focusing on establishing an infrastructure to support the design and build aspects of the DBTL cycle. In the short term, we have three strategic directions. First, we will continue developing design tools based on our existing Build Optimization Software Tools (BOOST) platform to support more complex designs, such as pathway refactorings and combinatorial designs. Second, we will enhance our DNA synthesis and assembly workflow tracking database SynTrack to support the integrated, multi-platform assembly and cloning operations. Third, we will integrate our tools to streamline the data flow from the design tools to the SynTrack database to data repositories for data release using a standardized data exchange format, such as the

Synthetic Biology Open Language (SBOL). Our long-term strategic goal is developing software-based solutions to enable the Learn aspect of the DBTL cycle. The JGI is well positioned to gather test results from users who perform functional studies using the JGI-built constructs. In order to do that, we plan to develop a large-scale experimental and computational data hub that links (1) sequences to biological insights, (2) gene constructs to High-Throughput Functional (HTP) protein expression and structure data, and (3) CRISPR targets to phenotypic changes. Data quality will be monitored and improved through capturing data from consortium-type projects and the development of standards for functional data, such as standards for interpreting metabolomics data. This Learn capability will become increasingly valuable, especially from a design and enabling functionality perspective.

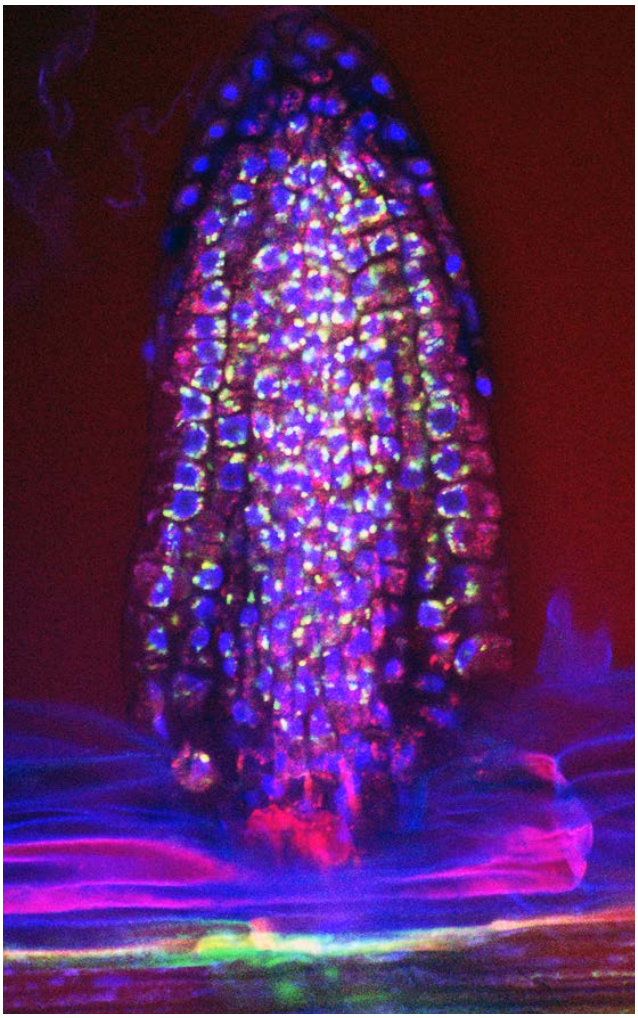
EXPANDING THE PORTFOLIO OF DNA SYNTHESIS PRODUCTS

A new focus area of the DNA Synthesis platform, in support of the strategic directions of the DNA Synthesis Science Program and the JGI Secondary Metabolites Group, is to enable users to discover secondary metabolites and understand how they mediate interactions between plants, fungi, and microbes.

Another growth area for DNA synthesis will be constructs with combinatorial design, in which genetic modules are used repeatedly in assembling multi-genic constructs to create a wider diversity of gene-to-gene, promoter-to-gene, or ribosome binding site-to-gene designs. This approach is also used in exchanging active domains of single genes, such as polyketide synthase genes, to expand their functional diversity and produce new small molecules. Finally, a new product type beyond enzymes and pathways is complex libraries with a high degree of variation, derived from commercial long oligonucleotide pools, with up to a million variants per pool of 200mers. Such libraries enable the construction of whole genome sgRNA libraries for CRISPR-Cas9 studies, which can be used to turn off or turn on gene expression gene-by-gene to determine gene function in a population of cells. We also anticipate coupling this with the use of transposon mutagenesis approaches (Tn-Seq) to enable the en masse recovery of individual gene knockouts for downstream functional investigation.

STRAIN ENGINEERING

The ability to genetically manipulate a broad range of organisms including those previously believed to be intractable is expected to revolutionize functional studies in vivo. Towards this goal, we have recently prototyped a generalizable strain engineering approach called chassis-independent recombinase-assisted genome engineering (CRAGE). This technology employs a transposon to insert a landing pad that contains two mutually exclusive *loxP* sites and a Cre recombinase gene into the recipient cell genome (Fig. 5.3). Thereby, individual genes, pathways, or genome editing or regulation tools can be integrated into novel genetic backgrounds to explore function or control expression. We anticipate this strain engineering technology to be applicable to a wide variety of downstream studies. For example, microbes can be engineered to produce novel fuels or chemicals, or to colonize roots to study plant-microbiome interactions. Strategically, we expect to pilot access to this nascent technology with JGI users to assess the impact of this capability and to scale accordingly. See Milestones **BDT04**, **GNT14**, and **GNT15**.



Sarah Lebeis, University of Tennessee, Knoxville (formerly at University of North Carolina at Chapel Hill)

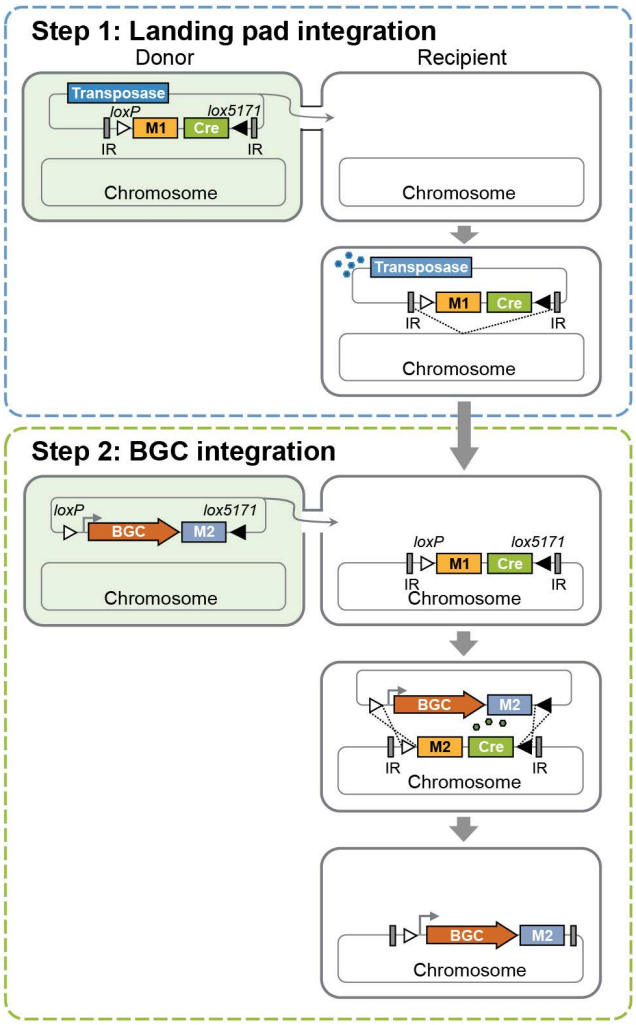


Figure 5.3. CRAGE technology for engineering diverse genetically-intractable organisms.

MAINTAINING A STRONG AND FORWARD-THINKING BIOSECURITY PROGRAM

As a good steward of foundational research and early-stage technology development, the JGI has made biosecurity best practices a core component of its synthetic biology operations. This includes, for instance, implementing the U.S. Department of Health and Human Services guidelines for providers of synthetic double-stranded DNA. The JGI will continue to make its own contributions to the continuous improvement of these biosecurity practices (e.g., by implementing screens against the JGI's extensive viral sequence databases). The JGI will also collaborate with other groups (e.g., Agile BioFoundry, Lawrence Livermore National Laboratory, Battelle, and the International Gene Synthesis Consortium) as well as other federal agencies (e.g., Intelligence Advanced Research Projects Activity [IARPA] and the Department of Homeland Security) who are themselves developing and maintaining complementary biosecurity approaches and infrastructure in order to integrate best practices regarding DNA synthesis biosecurity across organizations.

DNA Synthesis Science Program

The DNA Synthesis Science Program harnesses the DNA Synthesis platform for focus areas including (1) genomes to structure and function (see Milestones **SSP01**, **SSP02**, and **SSP03**), (2) high-throughput functional genomics (see Milestone **SSP04**), and (3) microbe–microbe and plant–microbe interactions (see Milestone **SSP05**). The program currently offers users the ability to perform collaborative research projects requiring large-scale DNA synthesis and construct assembly, combinatorial pathway assembly, and CRISPR library assembly, as well as opportunities to use integrated genome and metabolomics technologies and the JGI's informatics pipelines. As outlined in Fig. 5.2, the program will evolve from characterizing single genes and pathways to understanding biology at system levels over the next two years, and at community levels over the next five years. Additionally, we envision further integrating the capabilities of other user facilities to facilitate learning activities for program users.

GENOMES TO STRUCTURE AND FUNCTION

Advances in DNA synthesis technologies provide researchers unprecedented opportunities to mine the massive sequence data spaces provided by large-scale genome and metagenome studies and to identify and access genes of interest for heterologous expression and characterization. Since its inception, the DNA Synthesis Science Program has been supporting more than 50 user projects for large-scale characterization of enzymes involved in biomass degradation (e.g., glycosyl hydrolases and lignases), fuel and chemical synthesis (e.g., lyases and oxidoreductases), bioremediation (e.g., nitrate reductases), and carbon, nitrogen, and phosphorus cycling. To further support users' projects, the program is developing cell-free transcription-translation systems in collaboration with users through an ETOP project. The function of some of these enzymes may be characterized using the high-throughput mass spectrometry capabilities (NIMS and MALDI-TOF) we offer in collaboration with the Metabolomics platform group. Because structural information provides mechanistic and evolutionary

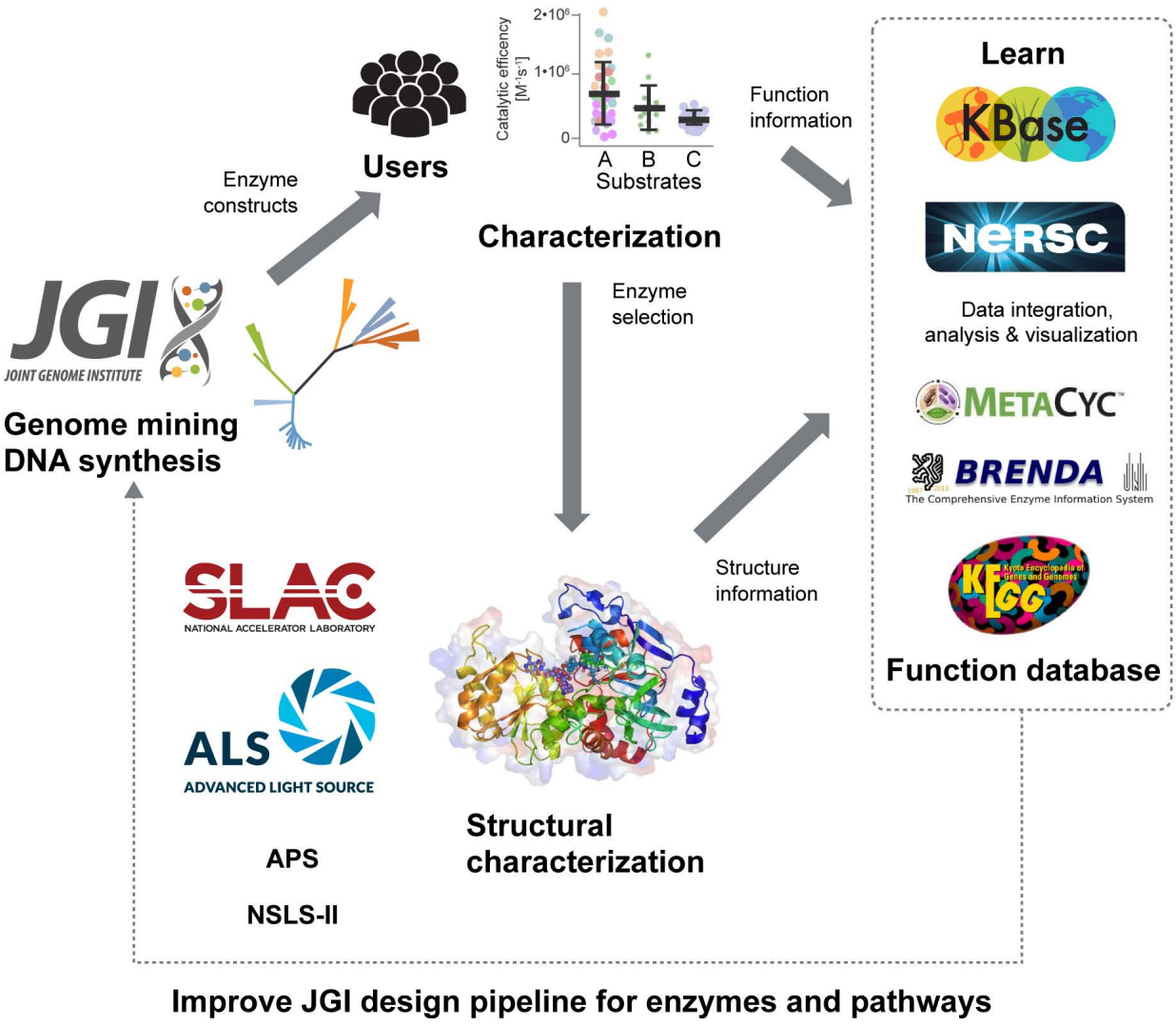


Figure 5.4. Overview of the genomes to structures and function program.

insights into these enzymes, the DNA Synthesis Science Program recently initiated several pilot crystallization projects with the beamline user facilities, including the Advanced Light Source (ALS), SLAC, Advanced Photon Source (APS), and National Synchrotron Light Source II (NSLS-II). See the section "**Genomes to Structure: Cross-Facility Opportunities for Structural Biology**" on page **55**. These efforts have already elucidated more than 10 structures with or without substrates, cofactors, and/or products. Our goal for these collaborations is to develop a new FICUS program focusing on elucidating biological structures and to offer our users the ability to streamline their work while moving from large-scale genomics to DNA synthesis and characterization of function and structure (Fig. 5.4). See Milestones **USP01**, **USP02**, **MTB01**, **MTB02**, **MTB05**, **MTB07**, **FGP04**, and **GNT10**.

Our goals include the development of a database to store function information for all characterized enzymes (within two years, 50 families of enzymes) and continued expansion of this database (within five years, 200 families of enzymes). The function information will have many downstream applications, including pathway design and engineering. We will also continue to analyze these data and develop algorithms to help select genes of interest based on (for instance) sequence similarity network analysis. In addition, we will develop a platform to characterize structures and functions of novel protein families

found in environmental metagenome samples in a high-throughput manner. See Milestones **SSP01**, **SSP02**, **SSP03**, and **MTB07**.

HIGH-THROUGHPUT FUNCTIONAL GENOMICS

Systems-level understanding of key microbes in energy production and environmental studies is a long-standing goal of the DOE. Some of these microbes serve as chassis for biofuel and renewable chemical production from lignocellulosic sugars or CO_2 via photosynthesis, and others are involved in the global cycling of carbon, nitrogen, and other elements. The DOE has already invested significantly in developing platform technologies to study these microbes as models, and has made them available to user communities. Among those communities, the JGI offers capabilities in transcriptomics, other sequencing-based functional genomics, DNA synthesis, and production metabolomics. EMSL offers capabilities in proteomics and systems metabolomics. KBase is currently building a platform to enable integrative analyses of all these omics data. We anticipate growth in JGI functional genomics capabilities in areas such as DAP-seq, sgRNA amplicon sequencing, FACS-seq, and combinatorial pathway design ("**Functional Genomics Capabilities for Understanding Gene Regulatory Networks**" section, see page **19** and Milestone **GNT09**). These capabilities collectively support users to improve their understanding of microbes at the systems level. The DNA Synthesis Science program will continue to develop new capabilities in this area and actively collaborate with EMSL and KBase to develop a new FICUS program to support integrative omics studies of microbial systems. See Milestone **USP02**.

Our goals in this area include further developing new capabilities to study functions of genes and pathways and regulatory networks (e.g., network regulations by transcription factors, posttranslational modifications, and allosteric regulations) of microbial systems. We will develop sets of HTP libraries (e.g., sgRNA libraries, promoter libraries, and TF libraries) for 10 species within two years and 50 species within five years. See Milestones **BDT04** and **SSP04**.

MICROBE-MICROBE AND PLANT-MICROBE INTERACTIONS

Plant-associated microbes play important roles in modulating plant systems. For example, some of these microbes are known to produce plant hormones (e.g., indole acetic acid, salicylic acid, and gibberellin) to promote plant growth and anti-bacterial and fungal agents to protect plant systems from infection by pathogens. Some of these microbes can also fix atmospheric nitrogen (converting it to ammonia) and mobilize recalcitrant soil phosphorus to increase these macronutrients' availability to plant systems. Although research communities are increasingly finding evidence that secondary metabolites derived from plants and plant-associated microbes modulate plant systems, communications, and environments, our understanding is still very limited. Addressing these questions will require the ability to (1) characterize the molecular structures of these secondary metabolites, (2) perform loss-of-function studies by eliminating biosynthetic pathways for these secondary metabolites from native organisms, and (3) perform gain-of-function studies by expressing these pathways in heterologous host systems (see "Scientific Strategic Thrust: Secondary Metabolites"). Additional efforts include the deployment of CRISPR-Cas9 genome editing and Tn-seq technologies for whole genome analyses of the roles of genes and pathways. Coupled with appropriate test systems, these approaches would allow users to directly explore these roles in situ. Furthermore, we are looking to integrate omics capabilities among the JGI, EMSL, and KBase, which would allow us to improve our understanding of microbe–microbe and plant–microbe interactions at a community level.

Our goals in this area include development of engineerable synthetic root-associated microbial communities comprising at least 25 species to study how gene and pathway function (10 plant growth promoting traits) contributes to the plant growth promoting activity of microbial communities, and to develop a design principle to engineer microbial communities to promote plant growth. See Milestones **BDT04** and **SSP05**.

Scientific Strategic Thrust: Secondary Metabolites of Bacteria, Fungi and Plants

Secondary metabolites are those molecules produced by organisms that are not required for growth per se, but which provide significant advantages to those organisms producing them. These molecules are synthesized in response to environmental cues and facilitate nutrient acquisition, defense mechanisms against predatory organisms, communication between symbiotic and mutual hosts, and the ability to resist toxic compounds. Thus, identifying secondary metabolites and understanding their biosynthesis and roles in the environment can enable both new routes to bio-based molecules and also new methods to improve bioenergy crop yields. Genes responsible for production of these secondary metabolites are often clustered in their genomes to form biosynthetic gene clusters (BGCs) which contain all genes required for biosynthesis of precursors, assembly of the compound scaffold, tailoring of the scaffold, and often additional genes for resistance, export, and regulation. Even in bacteria, only 10 to 20% of secondary metabolite BGCs are expressed when organisms are grown in isolation under standard laboratory conditions. Therefore, a tremendous amount of biochemical diversity remains untapped. The expertise and capabilities that the JGI has to explore bacteria, fungi, and plants will position the JGI to functionally explore the diversity of secondary metabolites across these organisms, and expand the repertoire of secondary metabolites and the tools for characterizing them (Fig. 5.5).

EXPANDING THE DIVERSITY OF SECONDARY METABOLITE BGCS

Systematic discovery and characterization of novel secondary metabolites is an essential step toward better understanding of diverse ecosystems and exploiting their chemical diversity for new applications in health, food, agriculture, and the environment. The first sequences of Actinomycete genomes revealed a previously undiscovered richness of biosynthetic potential that was not apparent from previous analyses of culture broths. Indeed, the number of BGCs reported was three to 10 times that of the number of secondary metabolites found in broth extracts. Application of next-generation sequencing technologies has revolutionized novel secondary metabolite BGC identification. Through projects such as the 1000 Actinos and the 1000 Fungal Genomes, the JGI is making significant contributions to this expansion of secondary metabolite diversity. It is expected that the JGI will grow these contributions

Recognizing the opportunities associated with secondary metabolite discovery and characterization, the JGI has already initiated or will conduct several targeted efforts addressing the following grand challenges (Table 5.1).

Table 5.1. Opportunities in Secondary Metabolites

Grand Challenge	Opportunities for the JGI
Expanding the diversity of secondary metabolites	Sequence large collections of secondary metabolite producers
Accurate prediction of secondary metabolite BGCs and their products	Conduct comparative benchmarking of existing tools, develop new tools and pipelines, and further develop IMG-ABC
Rapid large-scale assembly, cloning, and expression of BGCs	Increased construction, assembly, and capture of large BGCs and their expression in novel hosts and cell-free systems
Detection, identification, and structural characterization of secondary metabolites	Metabolomics and nuclear magnetic resonance spectroscopy (NMR) of culture extracts and isolated secondary metabolites
Functional characterization of secondary metabolites	Conduct functional assays using fabricated ecosystems and microbes known to produce secondary metabolites and measure phenotypes Use of engineered microbes to ascertain effects of secondary metabolites Partner for additional functional assays

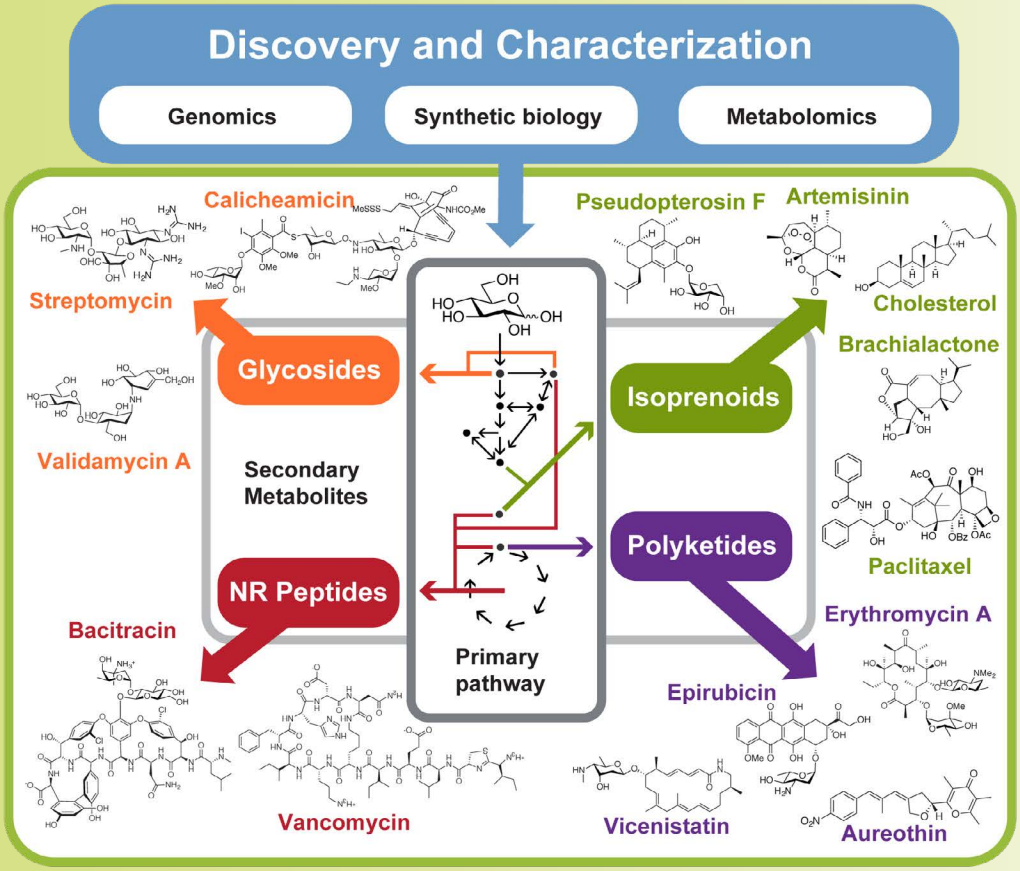


Figure 5.5. The JGI is well positioned to explore the galaxy of secondary metabolites. Illustration courtesy of King JR et al., © 2016.

through collaborations with institutions housing large collections of organisms with the potential to produce diverse secondary metabolites (e.g., the Tiny Earth Project). See Milestones **BDT04**, **FGP02**, **FGP04**, **MIP07**, **MGP07**, and **PKI06**.

ACCURATE PREDICTION OF SECONDARY METABOLITE BGCS AND THEIR PRODUCTS

As the wealth of sequence information grows, so does the need to accurately predict secondary metabolite BGCs and their products. Over the last few years, a series of computational tools have been developed that enable computation of different BGC types, such as polyketide synthases (PKS), non-ribosomal polyketide synthase (NRPS) hybrids, terpenes, and ribosomally synthesized and post-translationally modified peptides (RiPPs). These tools predict BGC boundaries, provide information on possible chemical modifications, and predict the total coding capacity for BGCs of a particular microbe. Several tools now enable insights into possible chemical structures of the products of BGCs and determine linkages between metabolites and genes (see the "Illumination of Dark Biochemistry" section). The JGI will conduct comparative benchmarking of existing tools, develop new tools to fill gaps, develop a new secondary metabolites pipeline, and further develop the IMG-ABC system as a major resource for the secondary metabolites scientific community. See Milestones **MTB01** and **PKI06**.

RAPID LARGE-SCALE ASSEMBLY, CLONING, AND EXPRESSION OF BIOSYNTHETIC GENE CLUSTERS

To access novel BGCs, we require methods to functionally express BGCs at a scale that befits the JGI's role as a large-scale genome science user facility. Firstly, we require increases in the construction of large biosynthesis gene clusters and the expansion of host systems from *E. coli* to yeast, where homologous recombination is available as an engineering tool. In addition, we anticipate expanding plate-base DNA assembly processes using yeast-compatible vectors to increase the throughput of cloning large multi-gene constructs. We will also explore long fragment DNA capture technologies from high molecular weight DNA,

which is a nascent technology for large BGCs which can then be directly cloned or expressed.

To functionally express BGCs, we will take advantage of the previously described CRAGE and cell-free technologies. With CRAGE, sets of BGCs can be integrated into a panel of microbial chassis and, through cultivation and RNA-seq, expression and metabolites can be measured to correlate BGC expression and metabolite identification. Cell-free transcription-translation technologies are gaining in prominence as approaches that circumvent cellular processes such as transformation, substrate and product transport, and gene regulation. The JGI will utilize existing and develop new cell-free expression cassettes that will drive expression of BGCs to generate enough material for metabolite detection and identification using mass spectrometry. See Milestones **MTB07**, **SSP06**, **GNT12**, and **GNT14**.

DETECTION, IDENTIFICATION, AND STRUCTURAL CHARACTERIZATION OF SECONDARY METABOLITES

The exploration of novel secondary metabolites expressed from BGCs requires the ability to accurately and with high sensitivity detect and identify these molecules. This will be accomplished by employing the JGI's state-of-the-art Metabolomics platform to analyze both polar and non-polar metabolites. The metabolite features will be compared to the JGI's metabolite database and structural predictions performed. For structural characterization, the JGI will partner with EMSL to utilize its NMR expertise to determine structures. See Milestones **MTB01**, **MTB02**, **MTB04**, **MTB05**, and **USP02**.

FUNCTIONAL CHARACTERIZATION OF SECONDARY METABOLITES

Secondary metabolites have broad roles in the environment and have commercial application as antibiotics, therapeutics, insecticides, and fungicides and in other uses. The JGI and its users are interested in the roles these metabolites play in governing microbe-microbe and plant-microbe interactions in the environment. The JGI will conduct functional assays using fabricated ecosystems and microbes known to produce secondary metabolites and measure phenotypes associated with plant/microbe growth and metabolite production/consumption. Engineered microbes and microbiomes with differing abilities to produce secondary metabolites can be employed to ascertain effects of secondary metabolites. Further types of functional assays for secondary metabolites outside of JGI's mission space will be conducted in collaboration with suitable partners.

EARTH'S SECONDARY METABOLOME

The JGI Earth's Secondary Metabolome Project will bring together these capabilities to systematically explore novel secondary metabolites and their biosynthesis, using state-of-the-art workflows that integrate large-scale bioinformatic mining, DNA synthesis, metabolomics, high-performance computing, chemoinformatics, and bioassays (Fig. 5.6). This will greatly expand our understanding of these important molecules and their roles. See Milestones **MTB04**, **MTB05**, **SSP06**, **MIP07**, and **MGP07**.

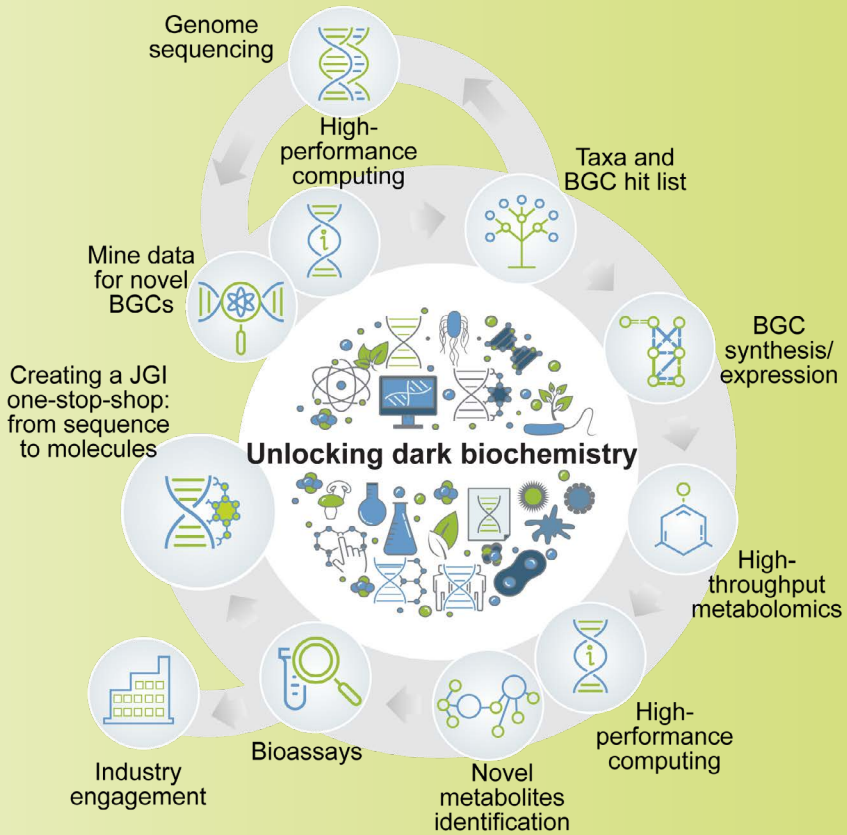


Figure 5.6. Integrating JGI capabilities for exploration of Earth's secondary metabolome.

OTHER OPPORTUNITIES

The DNA Synthesis Science program continues to seek new research focus areas. Biomaterials (materials made by organisms, which may be based on peptides [e.g., silk and leather], sugars [e.g., insect exoskeletons], silica [e.g., diatom shells], proteins or other biopolymers [e.g., adhesives], or other molecules) is an emerging research area in biology that could benefit from the JGI's unique large-scale DNA sequencing and synthesis capabilities. Recent advances in DNA sequencing technology continue to reveal the enzymes and pathways responsible for synthesis of these useful biomaterials. Our DNA Synthesis and Strain Engineering platform capabilities may help researchers understand biosynthesis of these biomaterials and synthesize them more efficiently and cost-effectively. Over the long term, our goal will be to develop new capabilities to help our users design and synthesize novel materials with desired properties. Meeting this goal may be enabled by closer collaboration with DOE's Nanoscale Science Research Centers, particularly the Molecular Foundry at Berkeley Lab, to which the IGB will be in close proximity.

High Throughput Metabolomics for Functional Genomics

Metabolomics Platforms

Technical advances in DNA sequencing and related technologies have led to an explosion in the number of sequenced genomes. As understanding metabolism is central to understanding biogeochemical cycles, plant-microbe interactions, and enabling sustainable bioenergy, metabolomics is a powerful functional complement to sequencing efforts. The JGI develops high-throughput metabolomics for functional genomics by effectively integrating metabolomics with DNA sequencing and synthesis to provide users with new functional insights into DOE biological mission areas. Since metabolomics was launched at the JGI in 2016, there has been a strong demand for this capability, and approximately 20% of recently accepted large-scale CSP proposals include metabolomics. Currently, our core capabilities are in liquid chromatography tandem mass spectrometry- (LC-MS/MS) based analyses of polar and non-polar metabolites, and these two products account for the majority of user demand.

Over the next five years, we anticipate significant investments will be made in further productionizing metabolomics and developing the accompanying cheminformatics and bioinformatics tools to enable users to visualize, analyze, interpret, and integrate their data. Productionization of workflows will primarily focus on increasing the number of analytical standards analyzed; standardizing extraction protocols for processing using liquid handling robots; and implementing a Laboratory Information Management System (LIMS) that will include QA/QC capabilities to improve standardization, facilitate metadata capture and management, and minimize re-work. Similarly, we will be making major efforts to automate standardized data analysis and release on the JGI Portal. In order to facilitate effective production of metabolomics at the JGI, the current metabolomics program will be split into a Metabolomics platform for production aligned with the Genomics Technologies department, and a Metabolomics Science program aligned with the Science Programs. This will allow for leveraging systems, e.g., LIMS, for production metabolomics, closer alignment and integration of applied metabolomics to scientific projects, and clearer communication. See Milestones **BDT04**, **MTB01**, **MTB02** and **MTB03**.

EXPECTED GROWTH IN METABOLOMICS CAPACITY

In FY17 we increased our capacity by 30% through reduced cycle times for polar metabolite analyses and then by another 50% through the addition of a second LC-MS/MS system. Our continuing development of production high-throughput metabolomics capabilities along with powerful bioinformatic tools will be a unique user capability for functional genomics. Hence, we anticipate that demand for these high-throughput capabilities will dramatically increase over the next five years. This will be driven by increased user awareness and the development of bioinformatic tools that facilitate user investigation, interrogation, and integration. In addition to CSP and FICUS users, we anticipate continued strong demand from bioenergy research centers and industry for this capability. See Milestone **MTB02**.

FACILITATING INTERROGATION AND INTEGRATION OF METABOLOMICS AND GENOMICS DATA

Our long-term goal is to provide users tools that enable those untrained in metabolomics to correctly interpret and utilize our outputs through the adoption and development of the necessary cheminformatics tools. These analyses include targeted data analysis using our large library of standards to identify a limited set of metabolites with a high degree of confidence, and untargeted approaches that enable users to explore metabolites not within these libraries, including novel metabolites. This requires advanced cheminformatic tools. To integrate metabolomics and genomics data, we will continue to invest in the development of nascent tools connecting genes and metabolites through non-rigid biochemical reactions (e.g., **https://magi.nersc.gov**) (Fig. 5.7). We will invest in the further development of capabilities for compound classification, identification, gene-function discovery, integration of genomics and metabolomics, and visualization on biochemical and chemical networks. In addition, we plan to offer machine learning algorithms that enable users to identify undiscovered molecules and pathways from mass spectrometry signals. See Milestones **BDT04**, **MTB01**, and **MIP01**.

Rapidly Determining the Activities of In Vitro Expressed Proteins

An emerging opportunity for the JGI is to complement DNA synthesis capabilities with protein expression, thus enabling users to directly characterize the functions of proteins. The JGI has already been using cell-free and cell-based expression platforms for the production of a small number of proteins. However, the recent advances in cell-free transcription-translation technologies afford the ability to scale protein production such that hundreds, if not thousands, of proteins can be produced in parallel. The JGI is working via an ETOP project to develop high-throughput and optimized cell-free systems, and the resulting cell lysates can then be used to prototype individual genes and metabolic pathways. We expect the technology being developed in this project to be transferred to the JGI and productionized for our users. We will further continue to develop the technology and optimize it for additional applications (e.g., discovery of novel secondary metabolites). We anticipate that JGI metabolomics will prove to be a very important component of these studies by enabling users to interrogate putative enzyme activities and metabolite binding. LC-MS/MS methods would be used to characterize these activities in complex mixtures, and

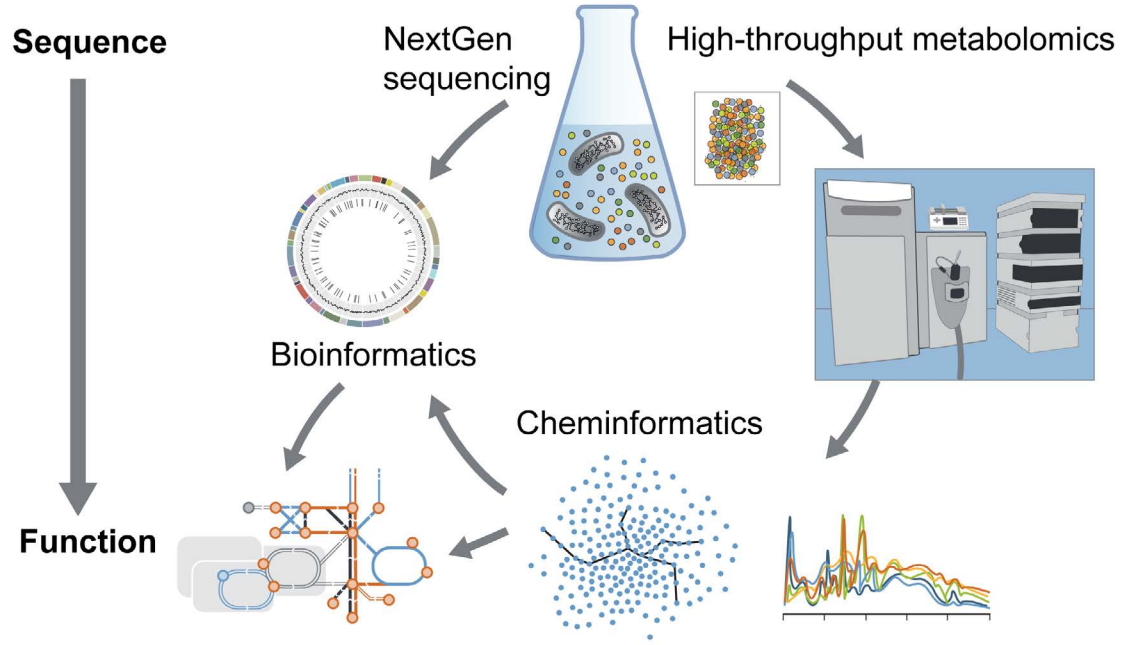


Figure 5.7. Integrated experimental, cheminformatic, and bioinformatic workflow for integrating metabolomic and sequencing for functional genomics.
Mass spectrometry signals are connected to compounds via methods such as accurate mass searching, fragmentation pattern matching, and spectral networking. Compounds are connected to reactions directly or via the chemical similarity network, which effectively expands the reaction database to similar reactions.

our high-throughput laser desorption platform can be used for large targeted enzyme activity screens. Where applicable, time-series enzyme incubations can be used to measure kinetic parameters. Towards this goal, we will perform a pilot project integrating our high-throughput metabolomics pipeline to characterize important proteins produced through in vivo and in vitro expression, including analysis of lysates, crude protein preparations, and purified proteins. See Milestone **MTB07**.

Illumination of Dark Biochemistry

Metagenomics and single-cell sequencing have enabled, for the first time, glimpses into the vast metabolic potential of Earth's collective biological systems. Yet, for the most part we can't accurately predict or identify the products of most genes or biosynthetic pathways. Most of what we know of microbial biochemistry is based on the characterization of a few model microorganisms, and these findings have been extended through sequence correlations to the rest of sequence space. Unfortunately, these extrapolations have questionable validity for the vast majority of environmental microbes that are distant from cultured isolates and therefore require fundamentally different approaches for directly linking novel sequences to their biochemical functions.

The JGI will shed light on this unexplored metabolic potential through deploying its core capabilities to mine genomic data through to functional characterization. Our approach will begin with bioinformatics mining of the over 50 billion unique genes in our systems to prioritize high-novelty genes and candidate biosynthetic clusters. Synthetic biology approaches will be used to refactor and express candidate genes and pathways in model organisms and cell-free systems for characterization of the resulting biochemical activities and products with mass spectrometry. When integrated with novel chemoinformatic algorithms, this will create a powerful closed-loop cycle of design-build-test-learn for systematically mapping biochemical space. Systematically testing sequences for their metabolic products will be an iterative process, through which subsequent steps are informed by the previous work and guided in a way to maximize novelty. This strategy will extend our knowledge of the biochemical space, enabling testing of a wider variety of organisms, conditions, environments, phenotypic responses, and biological activities. This approach is expected to have broad impact on application areas involving microbiomes, plant-microbe interactions, and ecosystem functions. Importantly, we will complement these efforts by partnering with other institutions to perform rapid phenotypic screens to identify biological functions for novel metabolites. Functional annotation of a wide variety of genomic data will provide a greater understanding of biology across all spectrums of life. The tools and biological knowledge developed through this effort will have material impact on systems biology and biomanufacturing research for years to come.

To enable these studies and user investigations, the JGI metabolomics platform will continue to invest in LC-MS/MS-based platforms for the analysis of novel polar and non-polar metabolites. Primary efforts will focus on expanding the range of metabolites that can be identified and providing users with tools that facilitate their ability to gain new insights. To achieve these goals, it will be important for the JGI to remain focused on LC-MS/MS technologies. To expand our ability to identify novel metabolites, the JGI will develop a new generation of advanced computational analysis methods combined with state-of-the-art physical chemistry models to map the unknown biochemistry we seek to understand. These algorithms will be made to efficiently score large collections of spectra in parallel and comprehensively capture potential fragmentation fingerprints to predict a parent molecule's identity. We expect that these algorithms will learn from measured signals and quantum mechanical simulations to quickly generalize principles guiding molecular identification. See Milestone **MTB04**.

Evaluating New Product Offerings in Glycomics and Lipidomics

Glycans and lipids are a central component of plants and microbes. The glycan structure and composition of plants is an important consideration for biomass-based biofuels approaches. Glycans are also critically important to microorganisms. It has been estimated that 40% of all predicted bacterial gene clusters encode saccharides. These glycans have diverse functional roles: for example, they constitute the glycocalyx "coats" of cells and are a major determinant of cell-cell interactions. Similarly, exopolysaccharides (EPS) are a critical environmental adaptation for many microbes and in many ecosystems drive the formation of biofilm microbial communities. Recently, LC-MS/MS-based methods and standard databases have been developed that facilitate routine

glycan profiling. Given that these types of analyses are extremely important to the BER scientific community and yet not currently available, the JGI will evaluate the possibility of establishing them at the JGI as part of an ETOP project.

Lipids play critical roles in membrane organization, structure, and signaling; responding to environmental conditions; and production of biofuels. The JGI provides lipidomic analysis as part of our existing non-polar metabolomics capabilities. However, our ability to identify lipids is limited to algae and is based on several JGI CSPs focused on algal lipids. We plan to extend these capabilities to plants and other microbes as part of a science project. Based on the impact and demand for the resulting capabilities, we will launch this as a specialized product offering for JGI users. See Milestone **MTB06**.

Harnessing In-House Plant Growth Capabilities

The development and optimization of new functional assays for plants, including technologies focused on plant-microbe interactions, requires numerous design-test-learn cycles and control over all aspects of the experimental system. The JGI has developed in-house plant growth capabilities and extensive plant genetic resources to provide this capability. Both the model dicot *Arabidopsis thaliana* and the model grass *Brachypodium distachyon* are utilized for research and development and Program science. For example, both species are being used to develop Drop-Seq for single-cell transcriptomics from protoplasts and isolated nuclei. See Milestone **GNT06**.

To enable detailed examination of plant-microbe interactions, *Brachypodium* is being used to develop controlled ecosystems that can be visually examined over time and interrogated with various omics technologies. In addition to developing new assays, in-house growth capabilities will be used to conduct user-directed experiments at the JGI for applications that require fresh tissue and/or extensive immediate post-harvest processing of tissues (e.g., tissue fixation or creation of protoplasts). Similarly, the JGI will continue to provide *Brachypodium* germplasm (e.g., seeds from sequenced mutants) to the research community. See Milestones **GNT07** and **GNT11**.

6. Integration

Cross-Facility Opportunities

Similar to other DOE user facilities, the JGI offers a diverse and unique suite of cutting-edge research capabilities to its users that would otherwise not be accessible to individual investigators. However, it is now increasingly recognized that many of the most challenging research problems in the DOE mission space are best addressed by inter- and transdisciplinary teams and through study designs that combine methods and expertise beyond those offered by any single user facility. To maximize the potential for synergies and facilitate user access to the orthogonal capability portfolios available across user facilities as we move into the next phase for the JGI, we expect to increasingly work with partner user facilities to generate and integrate powerful datasets for understanding the natural world (Fig. 6.1).

While it is expected that DOE-supported user facilities be open to all investigators, securing the resources for a large-scale, multidimensional research project could involve navigating a series of uncoordinated proposal processes, deadlines, and requirements. In 2014, the JGI and the EMSL joined forces to simplify this process and allow investigators to write a single proposal to access capabilities at both facilities. In the ensuing years, proposals to this joint user initiative, now named FICUS, have increased. Accepted proposals have led to high-impact science, validating the benefits of such partnerships. Based on this successful experiment, we have started to expand and improve cross-user facility collaborations (Fig. 6.2). See Milestone **USP01**.

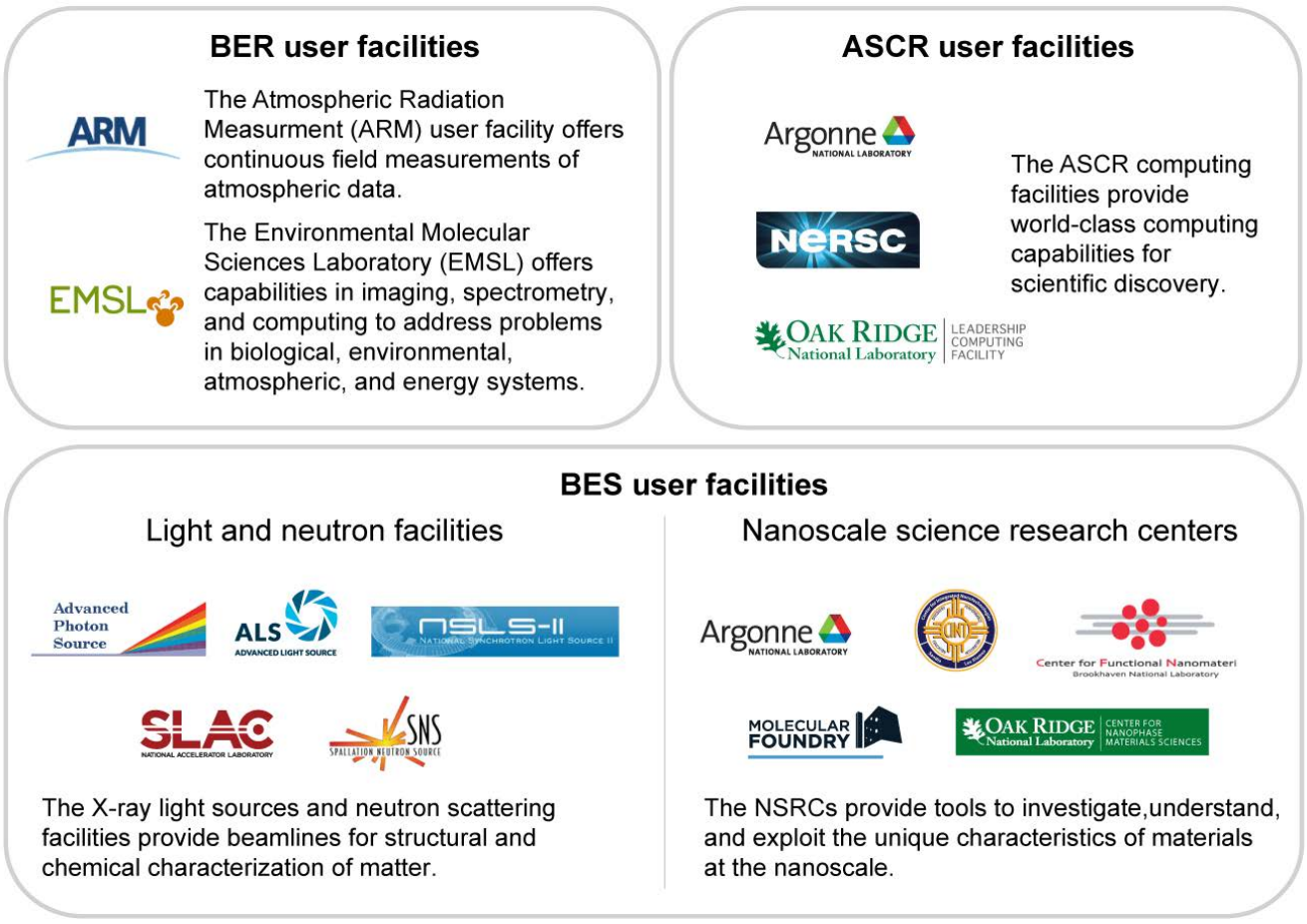


Figure 6.1. Overview of DOE Office of Science user facilities offering capabilities complementary to the JGI's.

COMPLEMENTARY VIEWS OF THE METABOLOME: INTEGRATION WITH METABOLOMICS AT EMSL

The JGI capabilities in metabolomics are complementary to those being developed at EMSL. Specifically, the JGI is focused on productionizing standard metabolomics pipelines and tools for integration with DNA sequencing and synthesis to create high-throughput metabolomics for functional genomics, the goal being to achieve a comparable capacity as the JGI sequencing platforms. EMSL's metabolomics capabilities complement these efforts by developing unique and powerful advanced technologies to enable structurally and spatially resolved metabolomics informing biological-ecosystem dynamics. Thus, where the JGI will have the capacity to perform over ten thousand metabolomic analyses per year, EMSL will have the capabilities to work with users to characterize metabolites with a spatial and structural resolution that requires specialized instrumentation. These integrated capabilities will be especially powerful in the context of the FICUS initiative, where users can request initial analysis of hundreds to thousands of samples at the JGI and then follow up on interesting findings using the advanced technologies at EMSL. See Milestones **MTB02** and **USP02**.

ADVANCED ANALYSIS STRATEGIES: CROSS-FACILITY OPPORTUNITIES FOR DATA SCIENCE

Since 2010, the JGI has worked closely with NERSC at Berkeley Lab to support data-intensive DNA sequence analysis. JGI computing resources are almost exclusively located and managed at NERSC. All JGI users therefore rely indirectly on NERSC to complete their projects. Given the high-quality analytical tools, high-performance computing resources, and sequence analysis expertise found at the JGI and NERSC, outside investigators have expressed an interest in collaborating with the JGI on large-scale data analysis projects. This motivated the first JGI call focused exclusively on data analysis, the JGI-NERSC Microbiome Data Science call launched in 2017, which resulted in six successful projects focused on improved annotations and comparisons of metagenomes as well as improved structural and functional predictions for genes. This successful program has been continued and expanded to a JGI-NERSC-KBase call for biological data science proposals with an increased emphasis on algorithm development and machine learning. KBase provides a platform for users to access NERSC's HPC resources, JGI data, and analysis tools in a user-friendly environment.

INTEGRATIVE BIOENGINEERING: CROSS-FACILITY OPPORTUNITIES FOR BIOPRODUCT DEVELOPMENT

Significant barriers stand in the way of more effectively harnessing biological systems to produce fuels, chemicals, therapeutics, food, and feed. To overcome these challenges, Berkeley Lab has developed a set of comprehensive and complementary capabilities to accelerate the discovery and development of bio-based products while diminishing risk for industry. This Integrative Bioengineering initiative combines the multiple omics technologies of the JGI with the robust predictive biology toolbox for multiple hosts and pathways and scale-up capabilities of the Agile BioFoundry, along with the fermentation and chemical process, product recovery and purification, and techno-economic modelling and analytical chemistry capabilities of the Advanced Biofuels and Bioproducts Process Development Unit to rapidly transition from biological hypotheses to pre-commercial applications.

GENOMES TO STRUCTURE: CROSS-FACILITY OPPORTUNITIES FOR STRUCTURAL BIOLOGY

A large fraction of DNA Synthesis Science Program users are interested in large-scale mining of enzyme genes from genome and metagenome dataspaces, paired with characterizing the function of the enzymes they encode. These projects often aim to discover enzymes with novel catalytic activity, explore the evolutionary correlation between primary sequence and function, and identify design principles to engineer enzymes with the desired function. Structural characterization is an important next step toward understanding the function of these enzymes or other proteins at the molecular level (e.g., see the section "**Functionally Characterize Fungal Conserved Genes of Unknown Function**", page **26**).

BER sponsors beamlines and capabilities at the DOE Office of Basic Energy Sciences (BES)-supported synchrotron and neutron facilities for protein structural characterization and other biological applications. Collaborations between the JGI and these facilities offer new and powerful opportunities for structural biology, possibly combining computational and experimental approaches as well as leveraging JGI capabilities in DNA synthesis and metabolomics for parallel structural and functional studies. We have already conducted several successful pilot structural biology studies leveraging capabilities from these facilities and the JGI DNA Synthesis Science Program. We plan to pursue joint FICUS opportunities with light sources in the near term, and

potentially develop new mechanisms for streamlining projects requiring capabilities at multiple facilities.

To this end, we established partnerships with DOE's X-ray light source user facilities (including SLAC, ALS, NSLS-II, and APS) and began collaborating on several selected DNA synthesis CSP projects as a pilot program. Additionally, we are investigating collaborations involving the combinatorial assembly and strain engineering capabilities of the DNA synthesis science program that may be useful for structural characterization of enzymes that comprise multiple subunits and are difficult to express in ordinary heterologous hosts. Through these collaborations, we are building towards a seamless cross-facility pipeline for structural biology to be offered through FICUS. See Milestones **FGP04** and **USP01**.

GENOMES TO FUNCTION: INTEGRATIVE COMPUTATIONAL AND EXPERIMENTAL OMICS ANALYSES

One focus of the DOE BER is developing non-model organisms as chassis for producing biofuels and bioproducts. The JGI and EMSL collectively offer transcriptome, proteome, metabolome, and other functional genomics capabilities (e.g., DAP-seq and CRISPR genome editing) through our joint FICUS program. These capabilities will be important for helping scientists improve systems-level understanding of non-model organisms. However, this understanding needs to be captured and presented through tools and models accessible to synthetic biologists. DOE KBase could provide a forum for a broader community of scientists to analyze and interpret large, complex integrative omics datasets and build predictive models enabling further development of industrially relevant organisms. See Milestones **FGP02**, **FGP03**, **GNT04**, and **USP02**.

HARNESSING NATURE: CROSS-FACILITY OPPORTUNITIES FOR NANOSCIENCE

Nature has evolved to generate highly sophisticated and intricate structures that, for example, power motion (flagellae), protect contents (viral capsids), and enable highly efficient chemical reactions while protecting the cell against toxic intermediates (microcompartments). These structures can serve as new materials themselves or can inspire the generation of new materials, using biology as a blueprint or as a biofactory. Genomes contain the instructions for making and assembling these structures, and through mining specifically for these the JGI can enable biomaterial discovery and development. New collaborative approaches will be explored between the JGI and DOE's Nanoscale Science Research Centers that can expand nanobiology and drive advances in DOE-relevant biomaterials.

SPATIALLY RESOLVED OMICS: CROSS-FACILITY OPPORTUNITIES IN IMAGING

While most omics data are based on "bulk" samples, the DOE-BER *Grand Challenges* report stressed the importance of smaller scales and improved spatial resolution. Numerous capabilities at current or future partner facilities enable interrogation of biological structures or processes within their three dimensional context. Techniques like cryo-electron microscopy or soft x-ray tomography enable new types of phenotyping to assess gene function, and mass spectroscopy-based imaging could enable correlation of metabolites with gene expression in eukaryotic tissues or cultivated microbes, or even correlation with phylogenetic identity in microbial communities. Collaborative FICUS calls will aim to leverage these capabilities to achieve the cross-scale science envisioned by DOE. See Milestone **USP01**.

Opportunities in Data Integration

LINKING AND SHARING DATA

Currently, the process for developing FICUS calls, soliciting and receiving proposals, technical and scientific review, and proposal acceptance are highly coordinated between facilities, but management of the work and provision of the data are largely managed at the facility responsible for each component (Fig. 6.2). The JGI is working with EMSL to better coordinate on project and data tracking, including cross-linking of websites, standardizing and sharing of sample metadata, and ultimately linking or sharing datasets to facilitate integrated analysis, for example through KBase (Fig. 6.2). As new FICUS partnerships are built, appropriate mechanisms for data linkage and integration will be developed. Over the long term, we aim to make the data from cross-facility proposals available to collaborators and the public in an intuitive and interpretable way regardless of where the data were generated. See Milestone **DSI07**.

THE JGI AND KBASE: COLOCATION, COORDINATION, COMPLEMENTATION

The JGI and KBase will be colocated in the IGB starting mid-2019. This physical proximity will be instrumental

in further expanding the already ongoing successful interactions and co-developments across the JGI and KBase. Our vision for the JGI-KBase partnership is that the JGI and KBase will continue to develop complementary and integrated high-performance tools to provide users with the ability and infrastructure to explore complex and diverse datasets to extract deeper biological insights. The goals of this partnership are to create a JGI presence within KBase, build a diverse, engaged joint user community, and enable scientific discovery. See Milestones **DSI04**, **DSI05**, and **DSI06**.

Opportunities in Cooperative Technology Development

The Emerging Technologies Opportunity Program (ETOP), launched in 2013, created a mechanism for exploration of new technologies in partnership with domain experts. Through this program, the 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 JGI. Depending on the project and technology, capabilities are either transferred to the JGI or can potentially be made available to JGI users through longer-term ETOP partnerships with external groups.

As technological methods advance and new areas of genomic science emerge, JGI scientists constantly evaluate potential capabilities for user benefit. When the technologies of interest require significant expertise beyond that available in-house, one option is issuing a call for ETOP proposals with a description of the desired technology or technologies. Through a competitive application and selection process, the JGI then identifies and funds investigators outside the JGI to develop the capability and adapt it to a user facility environment. These productive partnerships provide an agile mechanism to explore and test the potential for diverse technologies to contribute to the JGI strategic vision.

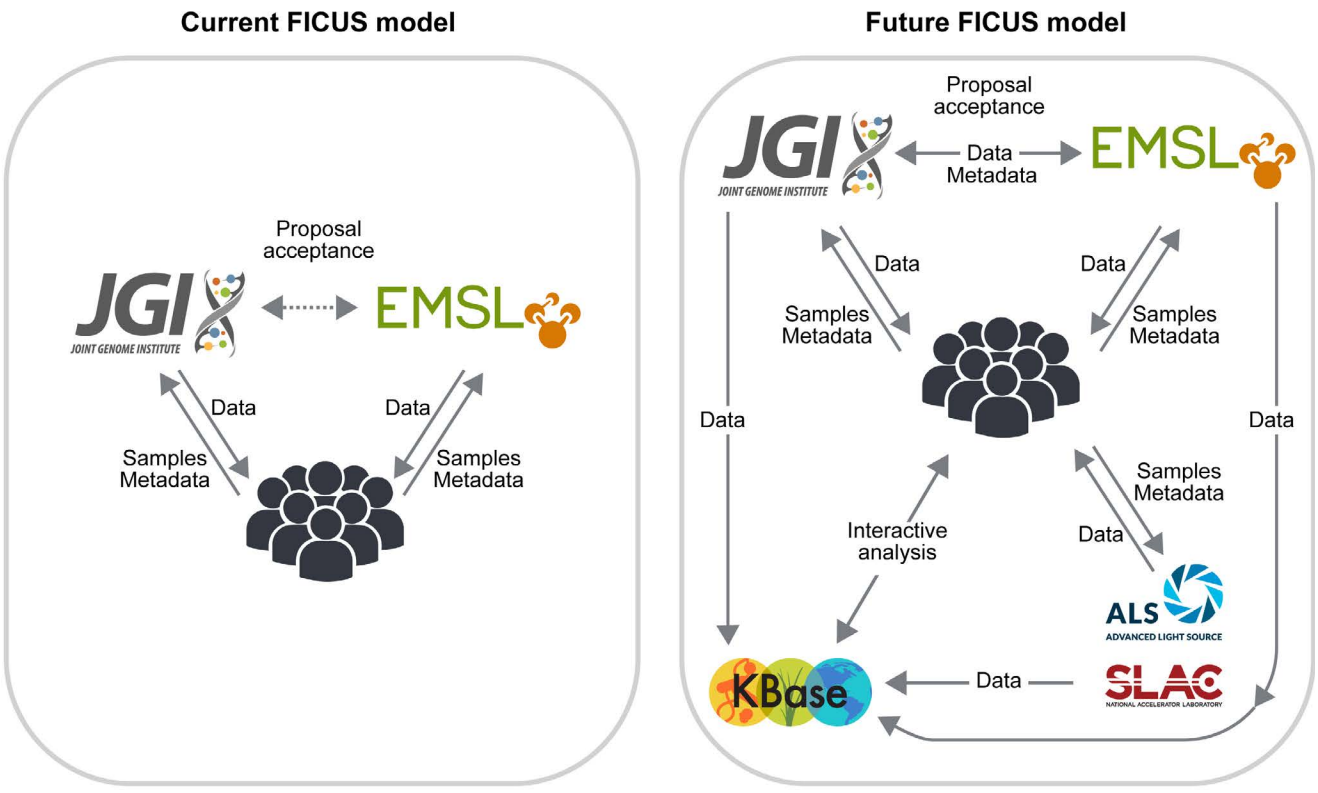


Figure 6.2. Current and future models for FICUS coordination.

7. Interaction

Communications and Outreach

The Communications and Outreach (C&O) team provides services internally, in support of staff, and externally to diverse sets of stakeholders, delivering relevant, coherent messages and products that establish best practices and ensure the integrity of the the JGI brand in support of the JGI's role as a key player in the worldwide environmental and energy genomics research ecosystem. To maximize impact, the JGI's effectiveness in communicating and conducting outreach is predicated on partnership. That is, every individual at the JGI shares responsibility for communicating effectively, with the onus on the JGI leadership and the C&O team to reinforce this sense of responsibility through our internal communications campaigns (Intranet, monthly and weekly e-newsletters, and all-hands meetings). To ensure the JGI's continued success, communication needs to be considered in all of the JGI's operational planning and decision-making. With this in mind, these guiding principles reinforce the qualities of the JGI's communication strategy:

- **Inclusive:** to engage the diversity of perspectives.
- **Proactive:** to recognize the appropriate messages for the target audience at the right time.
- **Credible:** to instill the essential elements of truthfulness, respect, and trust.
- **Unifying:** to integrate and coordinate efforts to engage and serve stakeholders.
- **Consistent:** to build a reliable network for sustainable relationships.
- **Service-oriented:** to embrace that the communications function of the JGI operates in support of the JGI (and by extension, Berkeley Lab, UC, and DOE) mission, priorities, initiatives, and best interests of staff and greater community.

Understanding and adopting this interlocking set of attributes is key to the "OurJGI" vision in which we all have a shared commitment and future.

Assuring/Sustaining Brand Integrity

The JGI is now in its 20th year residing in Walnut Creek, California. In advance of the move to the IGB at Berkeley Lab, it is critically important that JGI staff understand the JGI brand hierarchy, and that the JGI's brand is understandable to a broad spectrum of external stakeholders.

The JGI is supported through the DOE Office of Science BER, and visibility of this important stakeholder is to be acknowledged in all of our communications. Likewise, as a national user facility of Berkeley Lab, the Berkeley Lab logo and that affiliation as we move to "The Hill," need to be seamlessly incorporated in our communications, both visual and verbal. The C&O team is also actively engaged with Berkeley Lab Biosciences Area communications staff and the Berkeley Lab Strategic Communications Office to develop communication materials related to accomplishments of the JGI, other Biosciences Area Divisions (primarily Environmental Genomics and Systems Biology [EGSB]) where JGI staff have appointments and separately funded research, and cross-cutting work with other areas of Berkeley Lab. As University of California employees, there is much benefit accrued in terms of evoking pride and respect from acknowledging this alliance as well. Over the last year, we have made a deliberate effort to consistently highlight the opportunities arising from the future colocation of the JGI and KBase in all communications about the IGB. Computation and data management underpin all of the JGI's products, and we strive to acknowledge the contributions provided to the JGI by NERSC and ESnet, working with the communications staff resident in those areas.

The JGI is also actively advancing mechanisms to integrate the resources offered to the user community with complementary capabilities, such as metabolomics, proteomics, and imaging through the FICUS initiative. As FICUS diversifies from the original partnership with EMSL at PNNL, and now NERSC and KBase, we will continue to work closely with communications colleagues at the other national laboratories so that the JGI can most effectively co-promote collaborative contributions as new partnering facilities are brought under the FICUS

umbrella. The C&O team works proactively with the JGI BER program manager and the communications and public affairs staff in the DOE Office of Science so that the JGI's messages are appropriately calibrated with DOE's. To establish a common understanding about the attributes that the JGI brand ideally conveys and how this profile fits with the C&O goals, refreshing the JGI brand will be guided by internal and external perceptions gleaned from the following objectives.

Targeting Audiences

Knowing the educational background, technological experience, and what matters to an audience — among other considerations — is essential for tailoring our communications efforts to accurately hit their mark. C&O has sought to identify, understand, and network with a prioritized roster of audiences. This list continues be refined as the JGI responds to opportunities for developing new networks that will amplify the JGI brand, establishing and sustaining robust relationships with stakeholders on the basis of trust. The C&O team, working with JGI leadership, has identified a network of audiences for our communications efforts (Fig. 7.1):

- Worldwide scientific user community: academic, non-profits/other government agencies, and industry.
- Program managers in the DOE Office of Science, Office of BER.
- Internal audiences, including the JGI and Berkeley Lab staff.
- The DOE-supported research community, including the BRCs, national lab-based scientific focus areas, and university principal investigators (PIs) with DOE grants.
- Other audiences of influence, including elected officials and staffers (federal, state, and local), media (scientific and popular), professional societies, local community groups (benevolent/service organizations, e.g., Rotary Club, Lions, etc.), educators and students — with an emphasis on high schools and community colleges.

Crafting Relevant Messages

The C&O team supports the JGI's efforts in conceiving, refining, and propagating messages that enhance common understanding of the JGI's strategy and its

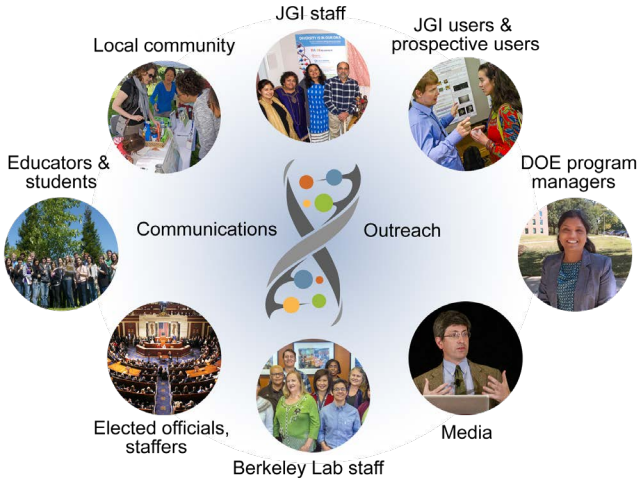


Figure 7.1. The JGI's audiences.

culture. By working with JGI staff, the C&O team ensures that our external messages promoting JGI-enabled achievements, user resources, and how the JGI fits in the research ecosystem are aligned. The C&O approach focuses on capturing the words and voice of the people behind values and achievements as the vehicle for delivering the message. The C&O team repurposes content, tailoring it for particular audiences to make it understandable outside of the JGI, and deploying it through channels deemed most effective — e.g., the JGI's public website and social media presence and/or working with partners or the media for third-party validation and amplification of messages. Through a series of surveys and focus group discussions with the JGI's user community, partners, and program manager, several attributes have been identified to help guide and underscore the JGI's messages: leadership, inclusiveness, coordination, transparency, responsiveness, accountability, and respect. These attributes are collectively umbrellaed under the OurJGI initiative. To help enable JGI staff and other stakeholders to be effective in conveying the JGI's messages, C&O have put forward activities targeting specific audiences. Through this coherent, transparent, collaborative approach, the C&O team will serve in support of the goals in the JGI strategic plan to help clarify and amplify the JGI's vision for the future, through activities described in the following paragraphs. See Milestone **BDT02**.

Conducting Brand Audit/Assessment

In order to reinforce and sustain the JGI's brand integrity, a baseline inventory and assessment of the JGI existing brand will be conducted. This includes logos, icons, symbols, images, acronyms, terms, and extant marketing collateral. This effort will also entail surveying staff, users, reviewers, and advisory committee members as well as engaging a set of other independent outside perspectives to evaluate the effectiveness of these vehicles for communicating the JGI's identity, relevance, responsiveness, and value to stakeholders.

Following this assessment and a report of the findings, recommendations will be assembled to formalize the approval process of naming conventions and style guidelines for all new and emerging programs, products, and their deliverables, including portals, software, data management platforms, and resource offerings.

Based on the feedback from the first year of implementing more rigorous branding standards and surveying the effectiveness of this guidance and whether it has resulted in an increase in positive visibility for the JGI, the C&O team will embark on a brand recalibration. This will take into account the realignment of the JGI's partnerships and programs that may be necessitated based on relocation to Berkeley Lab and colocation with KBase and proximity to the other national user facilities and prospective resources to be housed on the adjoining development sites (i.e., the "Biosciences Campus"). See Milestone **CMO01**.

Refreshing Onboarding

Revisiting and refreshing the JGI's onboarding procedures in preparation for the IGB migration will enable the JGI leadership to establish clear, consistent messages that evoke what it means to contribute to OurJGI, with an emphasis on safety, sustainability, and social/personal responsibility. This will provide a more deliberate alignment with the motivating principles of the JGI Diversity & Inclusion Working Group and compatibility with Berkeley Lab diversity, equity, and inclusion principles. In this way, the JGI is not only perceived in the recruitment process as an employer of choice but can retain those staff who have committed to a career at the JGI and Berkeley Lab. See Milestone **CMO02**.

Enhancing Effective Communications

For the staff to be effective at conveying the JGI brand, modules will be designed and offered to staff to hone their presentation skills, which will enable them to better engage prospective industry and academic partners, the media, elected officials, and staff. See Milestone **CMO03**.

Engaging User Communities

New and Expanded Scientific User Communities

The JGI strives to maximize its impact by supporting a large and diverse user community, and therefore continually reaches out to new users and new user communities who can benefit from JGI capabilities. One of the JGI's continued strengths lies in the assembly of research communities centered around DOE-relevant topics.

ENGAGE WITH USERS AND PROSPECTIVE USERS AT THE ANNUAL USER MEETING

Every spring, the JGI user meeting brings together hundreds of the JGI's users and potential users to hear talks on the cutting edge of environmental and energy genomics, present and view posters on DOE mission-relevant science, and interact with JGI personnel and fellow investigators. Plenary speakers are chosen to both showcase exciting JGI science and highlight areas of strategic interest to the JGI, attracting attendees with aligned interests with the potential to become JGI users. Future user meetings and associated workshops will aim to advance the strategic directions identified in this document.

ENGAGE THE MICROBIAL AND VIRAL ECOLOGY RESEARCH COMMUNITY

In April 2017, the JGI hosted its first New Lineages of Life workshop in Walnut Creek, California, entitled "NeLLi 2017: From New Lineages of Life To New Functions." This oversubscribed meeting featured an outstanding roster of speakers and provided ample opportunity for networking and the encouragement for community science in the area of phylogenetic and functional microbial diversity. The continuation of this workshop will be held in 2019 in San Francisco as the "NeLLi 2019: From New Lineages of Life To New Functions" Symposium. In a parallel effort, the JGI convened the first Viral EcoGenomics and Applications (VEGA) Symposium in 2018, during the JGI's 13th Annual Genomics of Energy &

Environment Meeting. With these and future efforts, the JGI aims to bring together the microbial ecology and "viral ecogenomics" community to foster discussions on how to capture and phylogenetically and functionally characterize microbial and viral diversity, understand the role of microbes and viruses in natural ecosystems, and functionally explore genetic diversity toward innovative biotechnological and industrial applications. Through community building, the JGI aims to continue to engage the research community to drive large projects that are greater than the sum of their parts. See Milestone **MGP03**.

ENGAGE THE FUNGAL AND ALGAL COMMUNITIES

Over the past 10 years, the JGI Fungal Program has transformed projects from single-genome single-PI projects to large-scale multi-genome multi-lab projects in all focus areas of the program. We will continue our successful user recruitment efforts through workshops and jamborees at JGI User Meetings or in conjunction with large fungal conferences, such as the Asilomar Fungal Genetics Conference, European Fungal Genetics Conferences, and others. These workshops usually combine presentation of program focus areas and capabilities, MycoCosm tutorials, and presentations from PIs of successful CSP projects.

Following this model, the JGI started active participation in algal conferences and will also organize algal workshops to attract new users to the JGI and consolidate the research community around large algal genomics projects. See Milestones **FGP01**, **FGP04**, and **FGP07**.

ENGAGE THE PLANT COMMUNITY

Phytozome will continue to train and engage new and existing plant genomics users via JGI User Meeting training workshops and the trainings held biannually at the HudsonAlpha CROPS conference. We are a frequent presence at major plant genomics meetings and hold a JGI workshop at the Plant and Animal Genome Meeting. This workshop provides an opportunity to present JGI user science and advances in JGI capabilities for plant research. New initiatives for plant science at the JGI typically arise in the context of multi-PI, multi-year CSP projects. For example, the Open Green Genomes Initiative aims to greatly improve comparative plant analyses by ensuring broad coverage of all major evolutionary lineages of land plants. We anticipate continuing to have focused meetings on specific topics including expanding phylogenetic breadth of plants and improving plant functional assays.

Strengthen JGI Interactions with Other Groups and Initiatives in Data Science

During the last few years, one of the most important partnerships for the JGI has been its close interaction with NERSC. Moving forward, we anticipate that this relationship will be strengthened through the continued evolution of the JGI-NERSC FICUS program, the first round of which supported advanced projects in microbiome data science. In addition, the strategic partnership with KBase will enable JGI users to explore new capabilities in the area of data science, not previously available through the JGI. Accordingly, we plan to further strengthen the JGI's interactions with KBase, NERSC, and EGSB and where applicable engage in joint developments, particularly in the area of analytical tools for large-scale data comparison. In addition to the engagements with NERSC and KBase, the JGI will continue to explore opportunities for collaboration with CRD at Berkeley Lab. Over the past five years, the JGI and CRD staff have collaborated on projects like ExaBiome. Moving forward, the JGI plans to form deeper connections in the areas of user-centered design, science search, and data management.

JGI ENGAGEMENT IN KBASE USER WORKING GROUPS

Over the next two years, KBase will be establishing and coordinating user working groups (UWGs) on three topics: (1) metabolism, (2) microbiome analysis, and (3) functional genomics. KBase will need to lead these particular UWGs because the primary goal for these groups is to organize efforts to integrate the data types, data, tools, and analyses from DOE-sponsored groups operating in each topic area into either the KBase platform itself, or shared infrastructure run by KBase, the JGI, and other user facilities (e.g., NERSC and EMSL). However, these UWGs will also (1) obtain feedback from users on services provided by DOE user facilities; (2) generate user-driven/science-driven designs for new tools and services (user-driven design); (3) develop data standards to facilitate data sharing and comparison among various scientific efforts; and (4) catalog all ongoing scientific activities in each area and aid in establishing priorities for future tools, data types, and services.

The KBase UWGs will have a significant role in dictating priorities and design within KBase, and given the desire for a strong and seamless collaboration between KBase and the JGI, it is very important that the JGI have a strong presence in each KBase UWG. Thus, minimally, all KBase UWGs will include JGI personnel as members, and in some cases, JGI personnel will aid in leading and coordinating sub-topics within the KBase UWGs.

Industry Engagement Program

MISSION AND STRATEGIC GOALS

Since the JGI opened its doors as a user facility in 2004, the vast majority of its users (~98%) have come from academic institutions. A key element of the continued evolution of the JGI is the inauguration of an Industry Engagement Program (IEP). As the JGI extends its efforts from basic discovery to translational integrative genome science, it will be critical to proactively reach out to strategic industrial partners to (1) further develop existing JGI capabilities; (2) bring in resources for discovery that would not otherwise be available; (3) maximize the impact of JGI science in the U.S. bioeconomy; and (4) explore and co-develop new pathways for discovery science complementing the DOE Office of Science mission space. To this end, the JGI has initiated an aggressive program for reaching out into the private sector. The JGI's IEP launched in 2017 with a team comprised of the JGI Director, Senior Manager of C&O, Deputy Director of User Programs, and a newly created role of Business Development Leader. The strategic goals of this initiative are to (see Milestone **BDT02**):

1. Sustain the relevance of the JGI by directing some of the JGI's discovery research towards use-inspired science.
2. Enable industry access to the JGI's cutting-edge technologies through CSP or sponsored research that will benefit the U.S. economy.
3. Encourage sponsored and collaborative research opportunities to develop state-of-the-art integrative genome science technologies.
4. Build new industry user communities.

PROGRAM STRUCTURE

In the first year of the program, the team has developed a narrative around the JGI "brand" to communicate to users unfamiliar with the JGI. This process included identification of the JGI's strengths, capabilities, and differentiators. During this time, the value proposition of the JGI to the private sector was further defined during conversations with over 100 companies. Inputs on the value, strategic objectives, and limitations of industrial engagement were also solicited from the JGI's Scientific Advisory and Informatics Advisory committees. It should be noted that engagement efforts exceeded the initial Year 1 goal to identify 40 firms for discussion of collaboration opportunities, suggesting a very large untapped reservoir of potential new private partnerships. A formalized outreach plan and implementation strategy is now in development as a result of these combined efforts. See Milestones **BDT02** and **BDT03**.

ENGAGEMENT MECHANISMS

Berkeley Lab provides for three broad mechanisms by which entities may work with the JGI and other DOE user facilities (Fig. 7.2). Traditionally, the JGI has focused on in-kind support in the form of sequencing or other services via the Community Science Program. This program is also accessible to industrial users and work is bound by the criteria and requirement for DOE mission relevance. Data generated under this program are publicly released. Understandably, this may be of limited appeal to companies interested in protection of intellectual property or trade secrets. Two sponsored research mechanisms, Strategic Partnership Projects (SPPs) and Collaborative Research and Development Agreements (CRADAs), offer an alternative by which a corporate partner may maintain privacy of data and ownership of intellectual property (IP). In the case of SPP projects, the industrial collaborator is required to fully fund the project within the agreed-upon scope. CRADAs differ in that there may be sharing of both costs and IP by the JGI and the partner. In the case of work performed under a CRADA, there is also the additional opportunity for Berkeley Lab to benefit from IP generation or ownership. Both SPP and CRADA projects are carefully evaluated for scientific impact and DOE mission relevance and approved by the JGI Director before any agreement is executed. In all cases, the expectation is that the work helps propel discovery and innovation forward in a way that would not be possible without this public-private partnership. The ALS at Berkeley Lab has an established record of working successfully with commercial entities to both support ALS infrastructure and enable scientific discovery. The JGI will benefit from the ALS's experience in development of a roadmap for industry engagement. See Milestone **BDT05**.

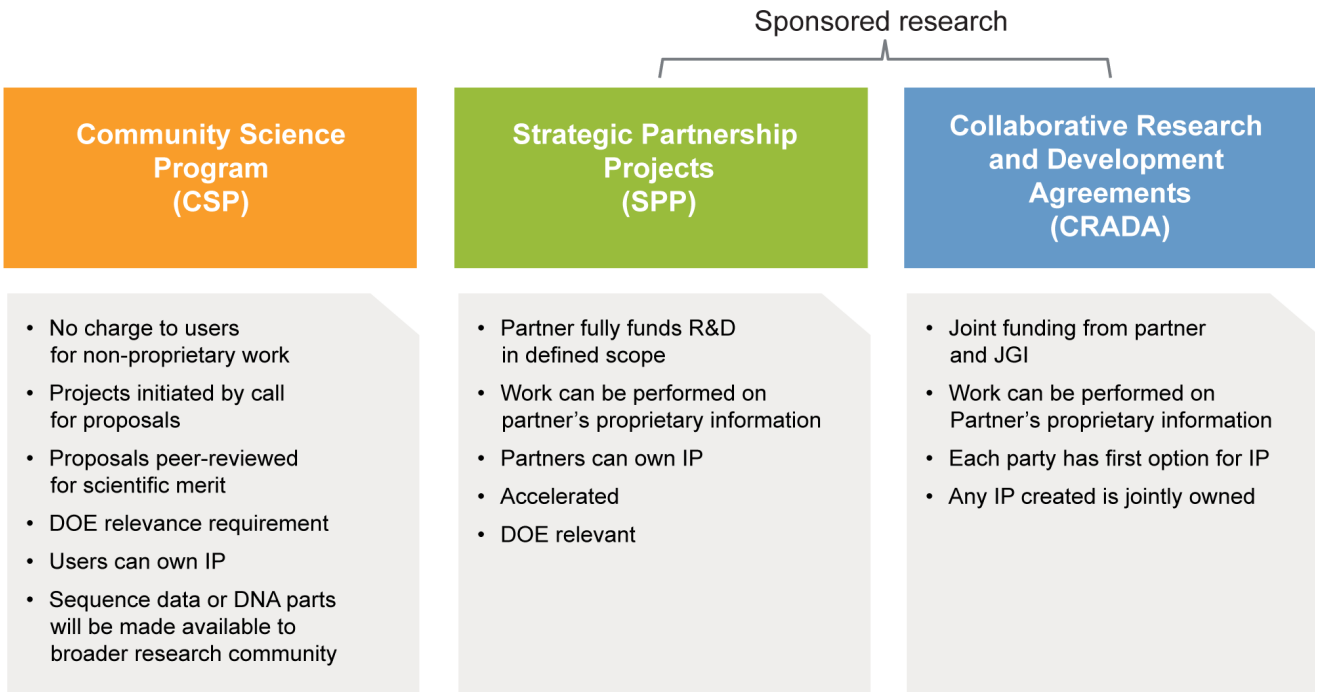


Figure 7.2. Engagement models for industry partners. The vast majority of the JGI's ~1,600 users come from the public sector (academic and government institutions). Through the IEP, companies can readily access the JGI's capabilities and resources while enhancing the relevance of the JGI's discovery research towards use-inspired science that will benefit the U.S. economy and maintain the nation's competitiveness.

The IEP additionally will develop a mix of proprietary and pre-competitive programs to support industrial users. This engagement strategy may be tailored to different industry segments. For example, small companies may see value in access to standard services that depend upon capital-intensive equipment. Large companies may see different value, such as access to the JGI as an expert partner in bioinformatics and high-throughput biology. Hence, highlighting these different modes of engagement is expected to drive increased engagement. During this process, different success metrics will be established that do not emphasize publication as a primary outcome. The impact of a relatively small amount of work by the JGI for a DOE mission-relevant start-up could produce a disproportionate amount of value for that startup and for the DOE more generally. The IEP will capture several facets of measurable impact and continually evaluate the portfolio of potential work, taking these into consideration. Examples of this include discovery leading to new product development, generating understanding of biologic modes-of-action, enabling product regulatory approval via provision of data, testing and validation of new technologies to support product launch, and many others. Recently, the JGI has collaborated on a strategic partnership project that touches on many of these measures of impact. In addition, this project represents a collaboration between the company, the JGI, and other groups and capabilities within Berkeley Lab. We are uniquely positioned to provide a portfolio of capabilities to interrogate, support, and develop commercial scientific endeavors. It is expected that collaborations leveraging capabilities from across Berkeley Lab will become a commonplace occurrence. See Milestones **BDT04**, **BDT05**, and **BDT06**.

As the IEP matures, we hope that the value the JGI brings to industrial collaborators grows. It is conceivable that intellectual property stemming from JGI collaboration with industrial partners could be commercialized. Additionally, as we grow the relationships, it is possible that we could secure additional funding for the JGI through SPP or CRADA agreements that could grow towards 5% of the overall JGI operating budget. See Milestone **BDT02**.

Measures of Impact

The JGI is supported by the DOE to advance energy and environment science, and it is important to assess whether this objective is being addressed. Scientific publications are the primary metric by which scientific productivity has been judged, and the JGI typically enables more than 150 publications per year (Fig. 7.3). However, the simple number of publications produced is only a very rough representation of impact, since the impact of individual citations on the field varies widely, which is evident from substantial differences in the rate at which individual papers are cited. Thus, the JGI also tracks the citations that individual JGI manuscripts accumulate over time. Finally, since eventual citation rates for a given article are unknown at the time of publication, for recently published articles the JGI uses the impact factor of journals in which articles are published as a proxy to categorize their expected impact.

Digital Object Identifiers

The JGI has tracked JGI-enabled publications since its inception, both those on which JGI staff are authors and those enabled through user proposals. The outreach team compiles these publications, based largely on self-reporting by users and JGI staff as well as periodic literature searches, and assigns them to science programs based on topic. However, as the number of proposals, users, and publications has increased, this manual tracking has become labor-intensive and error-prone.

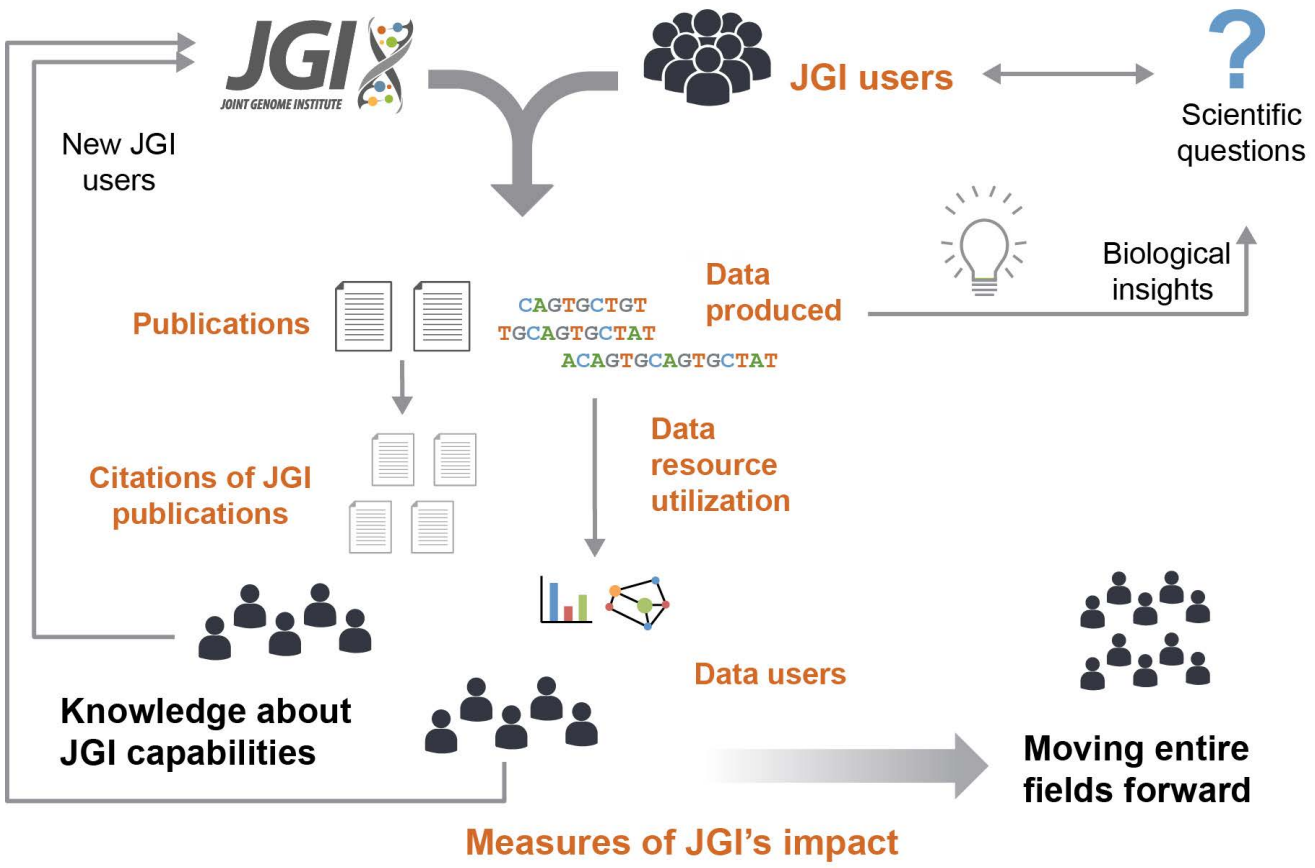


Figure 7.3. Some of the JGI's measures of impact. Major measures of the JGI's impact (shown in orange) include the number of JGI users, the number of publications authored by JGI users and staff, the number of citations of these JGI papers, the amount of data produced by the JGI and its users, and the utilization of JGI-provided data resources by data users. Both primary and secondary publications raise awareness in the community about the availability of unique resources at the JGI, thereby recruiting new users, and they provide a measure of the extent to which JGI-supported science answers important scientific questions and moves entire fields and communities forward.

Currently, users and staff are expected to acknowledge JGI contributions with a standard "auspice statement" noting the support from the DOE Office of Science and the Berkeley Lab contract number, which does not allow for linking of publications to specific proposals or programs. To facilitate retrieval of publication as well as linkage to programs and proposals, we plan to issue unique DOIs for each proposal and ask that all publications cite the relevant DOI. We also plan to facilitate self-reporting of publications by creating a user interface for users to report, particularly when submitting new proposals. See Milestone **USP03**.

Acknowledging Both Data Generation and Data Interrogation by Users

Another measure of our impact is the number of users whose research we facilitate through user projects. Each year we compile data on all investigators associated with user projects active in the previous fiscal year and provide their details in a report to DOE, along with a tally of the total number, which has steadily grown over time. However, this count only accounts for "primary" users directly associated with user proposals, and does not include the large community of secondary "data users" who analyze and download data from various JGI portals.

In the coming years, we plan to upgrade our user counting with improved data collection and curation as well as expand our efforts to include data users. Several years ago, the JGI developed a single sign-on (SSO) system that allowed investigators to create a single account that provides access to our proposal and sample submission systems as well as data downloads and custom tools in JGI portals, IMG/M, Mycocosm, and Phytozome. This allows more accurate collection and storage of user data, and also identifies data users and the systems they use. The JGI has informally tracked and counted data users who download data from JGI portals for the last six years, but was recently encouraged to include them in the annual primary user count and will do so starting in FY19. See Milestone **USP04**.

Unique User Identifiers

As a first step towards accurately tracking unique users, the JGI has implemented tools to collect unique digital ORCiD identifiers for researchers, when creating or updating JGI SSO accounts. This is currently optional but we plan to require ORCiD identifiers for all lead PIs on submitted proposals, as well as enable an option to sign in with an ORCiD identifier to encourage investigators to provide their IDs. In the long term, this is expected to reduce the effort needed to accurately track and dereplicate users who have multiple affiliations and email addresses (the current default "unique identifier" for DOE reporting).

Beginning in FY19, we will add data users to our annual user reporting, as well as collect information on the systems they use and the data they download, to see which portals and data are most heavily utilized. In addition to tracking data downloads, the JGI must understand which web-based pages and tools are accessed frequently. This data will provide insight into the data sets and tools that are of greatest value to the user community. The JGI will engage in a restructuring of the analytics systems across our flagship portals to enable this analysis. As part of the JGI's collaborative efforts with Berkeley Lab, we will work with the CRD and Names4Life to employ data mining approaches in order to better understand the broader impact of JGI data on the community by finding publications beyond those generated by the JGI users that have leveraged our resources. See Milestone **USP04**.

Another reflection of our impact on the scientific community and beyond is our brand recognition and social media presence. The JGI actively disseminates information on JGI projects and accomplishments through multiple social media channels, and the growing number of followers indicates both the growing recognition of the JGI and the increased reliance on these channels for accessing information. Going forward, we plan to continue to leverage these platforms for promoting our calls for proposals, sharing our science, and highlighting our user community.

Measures of JGI Impact on Industry

As the interactions and partnerships between the JGI and industry partners grow, it is important that we define the impact of the translational research and development performed within the IEP. While it is hard to predict owing to both the early-stage discovery work and the nature and number of company engagements, some examples of potential high-value impact are (see Milestone **BDT01**):

- Discovery research leading to product development (e.g., secondary metabolites as agricultural products, biotherapeutics, and novel antibiotics).
- Generating understanding of mode-of-action of microbes, microbiomes, and molecules.
- Supporting regulatory data generation for crop protection and trait development.
- Testing and validation of new technologies.
- Filing of patents and licensing of new technologies.

Education and Training

Since the era of the Human Genome Project, the JGI has been looked upon by students, educators, community groups, government agencies, and other research institutions as a resource for information about the burgeoning field of genomics. In the intervening years, the JGI has sought to meet the demand in K–12 science, technology, engineering, and mathematics (STEM) engagement through targeted outreach at the middle-school level and facility tours for high schools and community colleges, marshalling a corps of volunteer goodwill-education ambassadors resident at the JGI. For hands-on research experiences, the JGI hosts high school students, undergraduates, and graduate student internships and works with educators for better articulation within classroom curricula. These activities target those preparing on a trajectory to enter the workforce or for a shift in career path.

Workforce Training for Educators and Students

Through early introduction to the JGI, the goal is to inspire more students to enter the STEM workforce, developing a more diverse and inclusive pool of talent. These engagements seek to inspire an understanding of the JGI's role in advancing the energy and environmental genomics research frontiers though educational opportunities such as internships, school tours, and working closely with educators to calibrate JGI education "products" with classroom curricula and industry best practices. The C&O team currently coordinates school tours, organizing visits to the JGI, and sends JGI staff to off-site locations for school fairs or community events (Fig. 7.5).

Many of the JGI's education activities evolved independently from those administered through Berkeley Lab's Workforce Development and Education (WD&E) office. Most notably, the JGI has hosted University of California, Merced, graduate students through the Genomics Distinguished Graduate Internship Program (Fig. 7.4). Undergraduate UC Merced students who are part of the California Alliance for Minority Participation (CAMP), funded by the National Science Foundation in support of underrepresented minorities in STEM fields, were accepted into the internship program starting in 2015. Over the duration of the collaboration, about 20 students total have participated in both the undergraduate and graduate components of the JGI-UC Merced Genomics Internship Program. As the cohort of Merced students continues to grow, so too does the JGI-UC Merced leadership/faculty connection, assuring continuity and stability of the program, which will be necessary as the JGI moves to Berkeley Lab. As evidence of a strong commitment to the future of the partnership, a memorandum of understanding was signed by the two institutions to facilitate closer ongoing collaboration. The JGI will continue to nurture the relationship by offering UC Merced undergraduate and graduate internships after relocating into the IGB while exploring the capacity to develop additional internships. See Milestone **CMO07**.

To ensure a consistent, enriching experience for those students and to position the JGI to explore new partnerships with UC Berkeley (UCB), the local community colleges, and high schools, the JGI will need to prepare new mentors. As the JGI will be co-locating next year with KBase in the IGB and situated closer to other Berkeley Lab Biosciences research activities, the

Figure 7.4. Building the next generation of talent. The JGI-UC Merced Five-Year Program Celebration was held July 26th, 2018.

JGI will take the lead in designing a training module to enhanced mentoring skills to be made available to all prospective hosts for research internships.

Aligning Education Programs with Diversity, Equity, and Inclusion Core Values

With a more robust pipeline of mentors, the JGI will be better positioned to formalize additional educational linkages that include internships with UCB undergraduate and graduate students. In particular two UCB programs have missions that align with the JGI Diversity, Equity & Inclusion Working Group's stated interest in providing educational opportunities for communities traditionally underrepresented in science and engineering. The Biology Scholars Program [**http://bsp.berkeley.edu/**] aligns well with what Biotech Partners is accomplishing at the high school and community college level. Cal NERDS [**http://calnerds.berkeley.edu/**] offers "New Experiences for Research and Diversity for Science" for non-traditional undergraduate and graduate students in STEM fields. Their community reflects the full diversity of the Bay Area and includes LGBTQ, low-income, first-generation, transfer, disabled, foster, underrepresented minority and undocumented students, student parents, and women. There have already been preliminary discussions with the Biology Scholars Program leadership and enthusiasm expressed in formalizing a pilot internship program with the JGI.

At the high school level, the JGI and Biotech Partners — a non-profit school-to-career program for at-risk youth — have coordinated internships for three to five students per year from Antioch High School since 2016. When the JGI moves to Berkeley Lab, there will be direct connection with the program through its long-established biotechnology academies at Berkeley High School and Oakland Technical High School. Biotech Partners offers the JGI a "turn-key" connection with nearby East Bay high schools. The JGI and the Joint BioEnergy Institute (JBEI, one of the four DOE-supported Bioenergy Research Centers) already have positive name recognition in the community as there has been coordinated promotion of internship placements. The Berkeley Lab K–12 program management shifting October 1, 2018, from WD&E to the Government and Community Relations (G&CR) Office affords the JGI a seat at the organizing table for high school partnerships

Berkeley Lab-wide going forward. Biotech Partners' high school cohort have the option of pursuing year-round internships through the Peralta Community Colleges: Berkeley City, Laney, and Contra Costa. This conduit provides students additional training toward proficiency, enabling students at an associate degree level to compete favorably for skilled entry-level technician or research associate positions at Berkeley Lab. Bayer Corporation, over the 25-year development agreement with the City of Berkeley through Biotech Partners, has hired dozens of program graduates. The agreement is in the process of being renewed. The JGI is positioned to work with Berkeley Lab's G&CR Office to develop an analogous agreement with the City of Berkeley and Biotech Partners to start training minority and female students from challenging economic circumstances to compete for career positions at the JGI and the Biosciences Divisions. See Milestone **CMO06**.

In 2018, the JGI accommodated 27 interns. By developing, formalizing, and communicating criteria for each level of internship opportunity, implementing the proposed intern supervisor training module for supervisors, and centralizing the administration of all educational partnerships, the JGI will be better positioned to sustain, track, and grow student engagement. One such mechanism would be to provide tutorials to educators whose classroom curriculum can accommodate the use of the suite of tools and boundless data resources within the IMG/M system. The use of IMG/M paired with access to the JGI's sequencing capabilities could provide the framework for a high school-level environmental microbiome discovery program. A pilot, to be supported under the JGI Director's Discretionary Program, is already being conceived in partnership with a high school teacher in Florida. The project, starting in the fall of 2018, will focus on characterizing a DOE-relevant environment of mangrove soils. Twenty juniors in an accelerated biology class will each collect a sample (20 metagenomes in all), process these samples, and send the DNA to the JGI where the sequence will be generated. The class then, using IMG/M, will analyze/annotate their respective metagenomes. At the end of the semester, if the engagement with the JGI is deemed edifying by the students and mutually beneficial for the institutions, a plan will be developed between the Metagenome Program and C&O to scale-up in 2020 to five more schools. See Milestone **CMO04**.

Longer term, as the curriculum-development/classroom articulation expertise resident at the JGI is limited, it is recommended that the JGI engage an independent consultant who can conduct an evaluation of the education plans that the JGI proposes to put forward. The goal is to align the JGI's education efforts with the Graduate STEM Education for the 21st Century (**http://sites.nationalacademies.org/pga/bhew/graded/**) report. These areas include new initiatives and models that are influencing graduate education, including massive open online courses (MOOCs) and other digital learning programs, allowing the JGI to stay abreast of the national goals for undergraduate, graduate, and high school STEM best practices.

Figure 7.5. Outreach and education. For the second consecutive year, 21 JGI volunteers participated in the organization and implementation of a JGI module at the Annual Foothill Middle School STEAM (Science, Technology, Engineering, Arts, and Mathematics) Day at Walnut Creek's Arbolado Park. Over 300 6th graders participated in the event and assembled simulated DNA sequence strands comprised of four-color beads that can be worn as bracelets. These sequences were derived from actual sequenced JGI genomes.

Training the Next Generation of Users

In 2004, when the JGI opened its doors to the worldwide scientific community as a national user facility, the definition of the JGI's education efforts expanded to embrace "User Training." The JGI Science Programs led outreach efforts aimed at facilitating access to JGI data through hands-on trainings to hone skills on analytical tools hosted at IMG/M, Phytozome, and MycoCosm, enabling users to advance their research.

Moving forward, the JGI intends to further amplify the impact of these education efforts in parallel with training the next generation of data users and

prospective members of its own workforce through coordinating with user ambassadors and through virtual outreach training campaigns. Goals for each are outlined in the implementation milestones, and several focus on developing formal training modules for staff who participate in tours and other outreach efforts, as well as for those who wish to serve as mentors. See Milestone **CMO05**.

Finally, by establishing an organizational structure with a single point of contact for coordinating all the JGI's education activities, including internship opportunities internally and with Berkeley Lab's WD&E Office and other partner institutions, communications, outreach, and tracking of outcomes will be streamlined. This function will be administered out of the JGI's C&O office, which is already overseeing all tours and most high school and community college interactions. See Milestone **CMO04**.

The strongest, most visible, and sustained JGI education effort, as defined by training data users on JGI analytical tools, has been championed by the JGI Prokaryote Super Program's Microbial Genomics & Metagenomics (MGM) Workshops (Fig. 7.6). For over 10 years, JGI staff have offered five-day hands-on workshops, two to three times per year onsite at the JGI or associated with professional society meetings, for groups of 40 to 50 participants from dozens of different countries. So far, more than 1,100 scientists from more than 55 countries have received training in the use of JGI data analysis capabilities. We envision that as microbiome data generation increases in the near future, so too will the demand for training of researchers previously uninitiated in the analysis of massive, complex data sets. Condensed variations of this workshop have been held off-site, often in conjunction with meetings of professional societies, such as the American Society for Microbiology. Analogous, but shorter, training sessions are hosted by the JGI Plant Program on the resources of Phytozome, typically at the annual Plant and Animal Genome Meeting in San Diego, California, and by the Fungal Program on MycoCosm, every other year in coordination with the Fungal Genetics Conference, in Asilomar, California, as well as at other relevant meetings. The JGI Annual User Meeting provides a venue for overviews of these data repositories and tools as well.

With limited staff resources available to support outreach campaigns, to scale-up significantly will necessitate a plan to development and deputize JGI data ambassadors who can be enlisted to introduce the JGI's data portals and tools to new user communities. In partnership with the JGI Science Programs, Data Sciences and Informatics, and consultation with outside training professionals, the C&O team aims to deliver an immersive experience through more scalable channels.

The implementation milestones center on standardizing rigorous usability testing of JGI-supported web-based assets, formalizing ambassador training modules, and coordinating outreach activities to maximize access to JGI data and tools for new and returning user communities.

Looking ahead, the JGI will continue to refine expectations and goals of workforce education and user training opportunities that uniquely suit the JGI's expertise, resources, ability to play a role that aligns with Berkeley Lab and Biosciences Area objectives, and the DOE Office of Science mission. As the JGI relocates to Berkeley Lab, it will work more closely with WD&E, G&CR, and other Berkeley Lab areas and user facilities to set common metrics for gauging the impact of outreach to prospective users, educators, and students.

Figure 7.6. Hands-on workshops for current and future data users. Members of the JGI Prokaryote Super Program conduct an all-day workshop on Microbial Genomics and Metagenomics at the 2016 American Society for Microbiology Annual Meeting in Boston, Mass. This hands-on workshop is designed to familiarize users with the IMG/M data and workflows for computational analysis and interpretation of sequence data.

8. Stewarding the Strategic Plan

The JGI, as a national user facility, serves a broad scientific community. As part of Berkeley Lab, the JGI is committed to conducting groundbreaking research in genomic science and technology, focusing on DOE mission-relevant topics. As leaders and employees, we are charged to serve as stewards of this enterprise, responsible for ensuring that it remains a valuable national asset. Our stewardship responsibilities extend to the JGI's research, people, and resources, and we are committed to the following outcomes:

- We perform research of the highest scientific and ethical quality.
- We make our data available to the broad scientific community adhering to FAIR principles of data sharing.
- We thoughtfully manage and lead our talented staff, who are central to achieving our mission.
- We ensure a safe working environment and processes.
- We foster a culture of respect and collaboration for everyone in our community and advance principles of diversity, equity, and inclusion.
- We exercise the highest standards of financial accountability and transparency.
- We responsibly manage our infrastructure and assets.

In order to steward the success of our strategic plan, we are committed to these principles and have defined milestones that drive and allow us to measure progress in these areas.

Talent is Central to Achieving Our Mission

People are our most valuable resources. The talented staff at the JGI are central to achieving our mission, no matter whether they are administrative, operational, engineering, scientific, or managerial staff. To enable the best user science and efficient operation of the JGI and to meet our institutional goals, we must retain and develop our existing talent, and attract new talent. We must ensure that we cherish and value our diverse and inclusive culture. In 2017, the JGI Director launched the OurJGI initiative to complement the scientific and technical strategy of the JGI (Fig. 8.1). The OurJGI initiative encompasses converging bottom-up and top-down activities to focus efforts on evolving the culture of the JGI to increase inclusivity, skill and career development, teamwork, communication, respect, and scientific excellence.

This initiative strongly aligns with Berkeley Lab's renewed emphasis on the stewardship of people and resources, including funding and property (see **https://stewardship.lbl.gov**). Leaders have a huge impact on their organizations, and the development of our leaders is a crucial element in stewarding Berkeley Lab into the future. Leaders in Berkeley Lab's scientific divisions and operations areas are encouraged to develop deep expertise in managing people, and Berkeley Lab

Fig. 8.1. The OurJGI initiative.

is equipping leaders with the necessary skills to be successful. Berkeley Lab is committed to recruiting and developing a diverse workforce. Berkeley Lab's leaders take an active role in retaining our talented staff. Managers assess employees' strengths and areas of development annually and offer training and educational assistance to grow skills. Mentoring is an important way for Berkeley Lab leadership to foster early career staff development. The Biosciences Area mentorship program was launched in 2017, and many JGI employees are active as both mentors and mentees.

At the JGI, we have recently launched a new Talent Management Strategy (TMS) to focus our efforts on ensuring that we have the right people in the right roles to meet the needs of the organization. Continued implementation of this TMS will be critical to the successful implementation of the overall JGI strategic plan, since the JGI's strategic milestones and its continued evolution can only be accomplished through a sustained effort of its exceptional workforce (Fig. 8.2). To enable the TMS, we are empowering employees and supervisors to actively manage performance and development.

Fig. 8.2. Talent management is aligned with the overall JGI strategy.

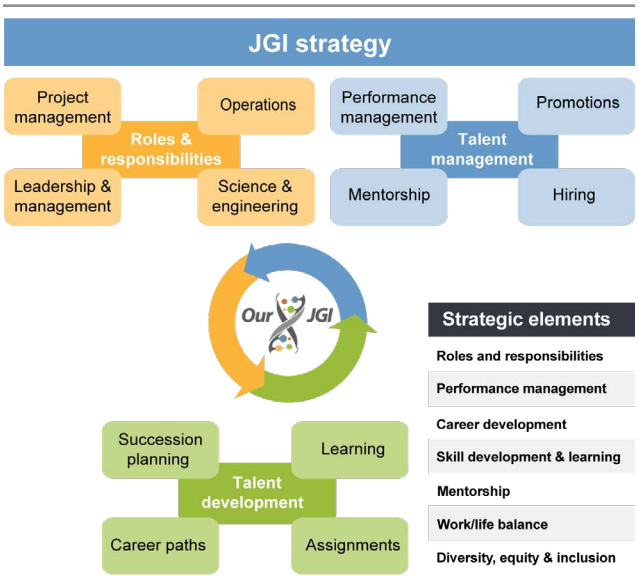


Fig. 8.3. JGI's talent management strategy and tools.

Talent Management Strategy

The JGI TMS consists of seven distinct elements (Fig. 8.3): (1) Roles and Responsibilities, (2) Performance Management, (3) Career Development, (4) Skill Development and Learning, (5) Mentorship, (6) Work/Life Balance, and (7) Diversity, Equity, and Inclusion. Each of these strategic elements fulfills an important function in support of developing and enhancing the JGI's talent pool.

ROLES AND RESPONSIBILITIES

To ensure that everyone at the JGI knows what they are expected to deliver and what constitutes a success, roles and responsibilities need to be clearly defined and understood. This will bring clarity to what people need to focus on and how they need to conduct their activities to meet their individual goals and the goals of the organization. It is important to match talent to the task, not the other way around, to better execute projects as well as recognize and utilize talent effectively. Strategic activities include to:

- Develop and communicate role cards for all roles at the JGI. Include responsibilities for training and mentorship. See Milestone **JLT01**.
- Identify and address gaps in roles and responsibilities with cross-training, stretch assignments, and exposure to new activities as members of multi-group and multi-disciplinary teams.

- Highlight the value that each staff member brings to the JGI in communications materials, presentations, and panel discussions.

PERFORMANCE MANAGEMENT

A transparent, equitable, and defined process for setting goals, monitoring progress, providing feedback and feedforward, and recognizing exceptional performance is a valuable means for aiding people to meet their goals and for their development. Performance feedback should not be punitive, but rather should meaningfully focus on what staff, supervisors, and management should improve on to be more successful. Strategic activities include to:

- Continue to move towards a continuous performance management culture built around feedforward. See Milestone **JLT02**.
- Administer the performance management process consistently and fairly in the organization with more discussion between management, supervisors, and employees on both successes and missed opportunities.
- Provide ongoing training and coaching to supervisors and employees to enhance the planning and feedback/feedforward discussions between supervisors and employees.
- Set challenging stretch goals to grow employee skills and ensure employees are supported.
- Develop mechanisms and processes to deliver effective recognition to employees so that they feel adequately rewarded and valued.

CAREER DEVELOPMENT

Defined career paths, career plans, opportunities for skill development, and support of management for career development not only grow staff but also provide opportunities to retain staff. Transparent promotion criteria and a stewardship process provide staff the opportunity to advance. Identifying suitable roles and/or assignments for staff to move into upon incumbents leaving their current role provides succession opportunities within the organization, and beyond in the Berkeley Lab environment. Strategic activities include to:

- Develop and communicate clear career paths. See Milestone **JLT03**.
- Include career paths as a touch point into the new hire orientation process.
- Partner with the Berkeley Lab Postdoctoral Scholar Program Manager to proactively develop resources for skill development and career advancement for postdocs, including:
  - Develop a JGI sponsor program to prepare postdocs for careers in industry, the DOE/national labs, entrepreneurship, and academia that comprises workshops, networking events, and career coaching opportunities to help advance postdocs.
  - Develop a bi-annual Lab Career Family information session including Lab panelists. See Milestone **JLT06**.
- Implement a transparent and equitable hiring process for all roles, with more transparent communication of job openings and associated benefits. Prioritize diversity in hiring and onboarding. See Milestone **JLT05**.
- Implement an easier-to-navigate job opening portal and application process for new positions.
- To attract new talent, the JGI should communicate a powerful value and opportunity proposition that is inclusive of all roles.
- Develop, communicate, and steward a transparent promotion process. See Milestone **JLT04**.
- Implement the new succession planning process. See Milestone **JLT04**.
- Hold and participate in career development workshops.

SKILL DEVELOPMENT AND LEARNING

Our journey of learning never ends, regardless of our role and level. Equipping staff with the skills, both technical and soft, necessary for them to fully succeed in their current roles and expanding their skills for future roles or assignments is a must. This can take the form of formal training and learning, on-the-job assignments, job rotation, and job shadowing. The JGI should consider all these as possibilities and seek opportunities for employees to engage in such activities while maintaining the JGI's operations and products. Strategic activities include to:

- Develop learning curricula based on current and future organizational needs. See Milestone **JLT07**.
- Develop leadership curricula for current and future leaders and encourage leadership development opportunities, including:
  - Serving as a supervisor or a mentor.
  - Gaining leadership experience via organizing events or leadership of internal JGI groups (e.g., the JGI Safety and Wellness [SWELL] team, the JGI Diversity, Equity, and Inclusion Working Group [DE&I], or as the JGI representative at the Berkeley Lab Women Scientists and Engineers Council [WSEC]).
  - Presentations (talks, posters, etc.) at international meetings. Presentation skills can be enhanced through the JGI Toastmasters Club.
- Provide opportunities for attending training and workshops.
- Use the proximity of IGB to other parts of Berkeley Lab/the UC Campus to develop new opportunities for staff to work on projects in other areas/facilities in addition to the many resources available for staff in DE&I, WSEC, Employee Resource Groups (ERGs), etc. See Milestone **JLT08**.

MENTORSHIP

Effective mentors can provide situational, emotional, and career guidance that can help manage situations, solve challenges, and aid personal development. While the JGI itself does not have a formal mentorship program, many JGI staff are participating in the Biosciences Area mentorship program as mentors and mentees. Strategic activities include to:

- Encourage participation in a mentorship program. See Milestone **JLT09**.
- Continue with team building activities/events and make conferences/talks available for all JGI staff to participate.
- Identify JGI sponsors who can share their knowledge/experiences in specific HR areas, such as tuition assistance vs. certification or taking a one-time class. Engage in employee activities associations.

WORK/LIFE BALANCE

Our staff at the JGI are no different from any other organization in that they have commitments and responsibilities outside of the JGI. It is vital that we strive for achieving the right balance between work and outside commitments, and that we provide our staff with the means to balance their work and personal life responsibilities. This becomes ever more important as the JGI moves from Walnut Creek, its home for the past 20 years, to the IGB on the Berkeley Lab campus. The JGI will work with Berkeley Lab to explore options to enhance work/life balance activities. Strategic activities include to:

- Determine what work/life balance means to JGI staff, and what the work/life balance priorities are for JGI staff. See Milestones **JLT10** and **JLT11**.
- Provide more transparent communication of existing programs to support work/life balance at Berkeley Lab.
- Evaluate flexible and compressed working arrangements, including expanded telecommuting options.
- Explore options for managing family responsibilities.

DIVERSITY, EQUITY, AND INCLUSION

A diverse and inclusive staff and culture is paramount to the continued success of the JGI. The details of the JGI Diversity, Equity, and Inclusion (DE&I) strategy will be developed later, along with tools and mechanisms to ensure implementation. See Milestone **JLT12**.

Operational Excellence Enables Scientific Excellence

Successful Transition to the Integrative Genomics Building (IGB)

In the summer of 2019, the JGI will move from the Walnut Creek, California, campus to the new state-of-the-art IGB funded by DOE at Berkeley Lab (Fig. 8.4). This new facility is designed to support two DOE research programs, the JGI and KBase. The move to the Berkeley Lab campus will facilitate an increase of the intellectual exchange and collaboration between these two programs and also enhance synergies with other existing science programs and user facilities on the campus, including the Molecular Foundry, the ALS, and NERSC. Properly stewarding all aspects of the construction, move, and transition to operations with the IGB will be critical to the JGI achieving most of the **I5** goals in the Strategic Plan.

CONSTRUCTION

Construction of the IGB began in 2017 and is projected to be completed in the summer of 2019. During the construction phase, JGI management has been actively involved in regular construction meetings to ensure features in the design phase are properly implemented and that any facility changes that may be required due to new mission needs or technology upgrades are successfully incorporated into the building. In addition, JGI Management is an active participant in partnership with the architectural and engineering firm, construction team, Berkeley Lab project team, and procurement to ensure both JGI building requirements are incorporated and construction milestones are met to successfully enable the move to the new facility in the summer of 2019. Participation in regular IGB Owner, Architect, Construction weekly meetings ensures the JGI properly stewards this phase of the project.

MOVE

The JGI's move to the Berkeley Lab campus will involve moving approximately 250 staff members and all of the office and lab equipment located in four buildings and over 80,000 ft^2 of office and lab space at the Walnut Creek location. Additionally, ~15 staff from KBase will move from their current location in West Berkeley. In order to maximize scientific and operational productivity in 2019, it will be essential that the JGI and KBase develop a detailed and well-coordinated move plan to minimize

Figure 8.4. The IGB, the new home of the JGI and KBase.

downtime with compute and scientific lab equipment during the move phase of the project. Key activities and stakeholders were identified and incorporated into the move project plan in November of 2017. Through the construction phase of the project, ongoing move meetings with vendors, Berkeley Lab subject matter experts, and JGI and KBase management are being conducted to implement a floor-by-floor phased approach to successfully move to the new facility. The goal in this specific phase of the project is to have office staff up and running 24 to 48 hours after the move of the compute equipment and for most mission-essential lab equipment to be operational within two weeks of the move to the IGB.

TRANSITION TO OPERATIONS

The move to the new IGB will be a significant change to the way the JGI has conducted science support at the Walnut Creek location. Moving to the Berkeley Lab campus will include a significant shift in roles and responsibilities with Facilities, Safety, Administration, Procurement, IT, and Safety. Support functions traditionally managed by the JGI Operations Department and supported by contract vendors will now be supported and managed out of the Berkeley Lab centralized facilities organization. In addition, new Berkeley Lab campus business processes will have to be socialized and effectively implemented in virtually every area of our Operations Department. To achieve the JGI's mission goals in the next two to five years, it will be paramount for all new Operations support roles and responsibilities to be documented and to ensure both the Operations support team and JGI staff understand how to operate efficiently and effectively on the Berkeley Lab campus. The JGI move and transition team will work closely with the various Berkeley Lab Operations Departments and the Facilities Area Manager to ensure all aspects of the JGI's science support activities are documented and briefed to JGI staff before the move to the Berkeley Lab campus. See Milestones **OPS01**, **OPS02**, **OPS03**, **OPS04**, and **OPS06**.

Safety Stewardship

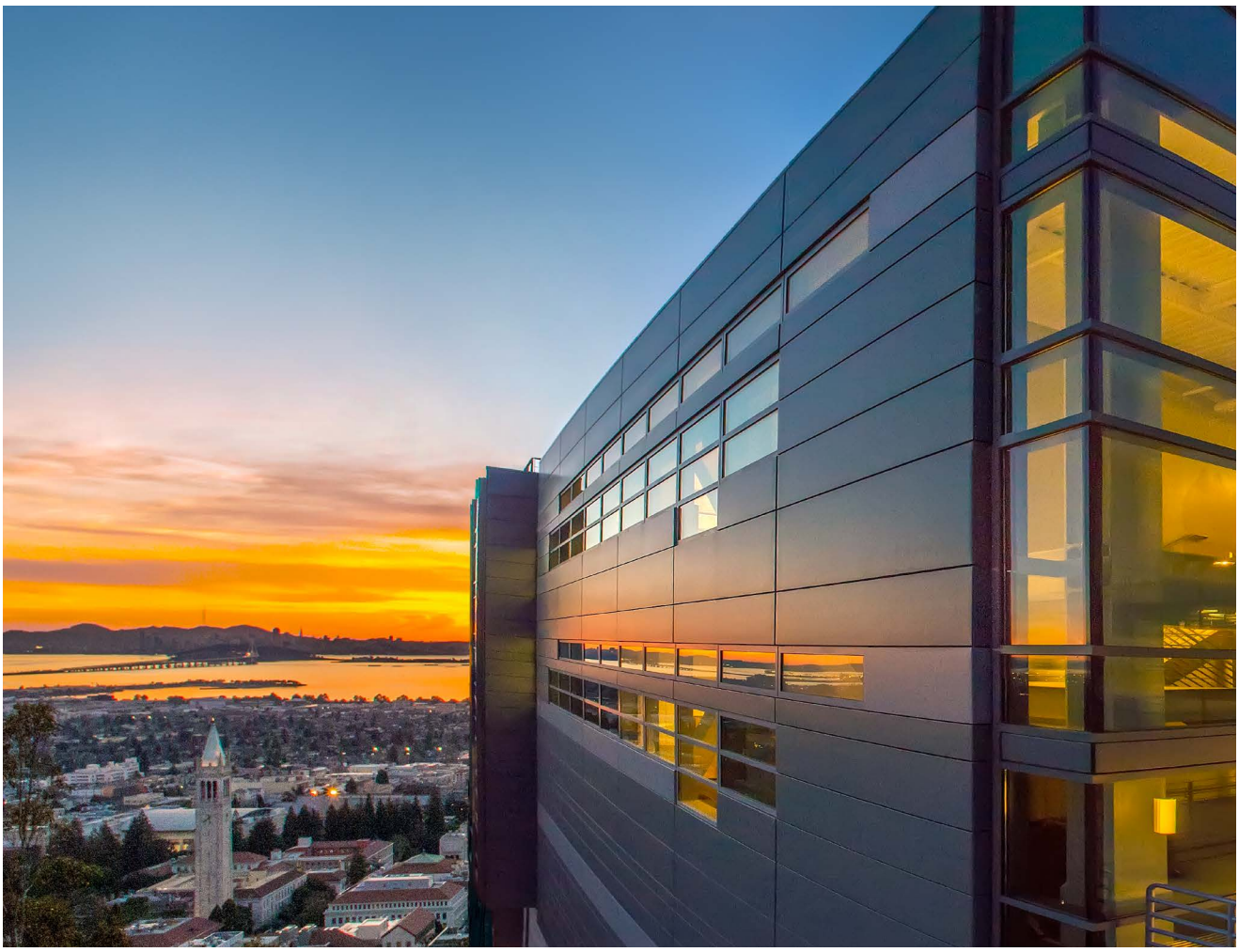
Over the past 10 years, the JGI has built a strong and diligent safety culture through implementation of active volunteer safety teams that include the Emergency Response Team, Area Safety Leads (ASL), and the SWELL Team. In addition, there has been a continued emphasis by the JGI management team on improving safety awareness through dedicated safety discussions (Safety Minute) in department meetings, including safety-relevant safety topics as a standing agenda item for each JGI All Hands meeting, and through "Potty Training" with safety flyers posted in all restroom stalls. Leadership engagement in safety has increased with regular safety walkthroughs by JGI senior leadership and the JGI Director, and the re-establishment of the JGI Safety Committee comprising the JGI Leadership Team, ASLs, and the JGI Safety Coordinator. One particular strength of the safety program has been the focus on lab and office ergonomics where the JGI has won national awards (Two Time Ergo Cup Winner-Applied Ergonomics Conference) for being an industry leader in implementing innovative ergonomic solutions in the workplace. All of these efforts have resulted in the JGI having some of the lowest injury, Total Recordable Cases, and Days Away, Restricted or Transferred rates when compared with rates at other DOE Laboratories.

While the basics of safety do not change, the environment around the JGI workforce will be significantly changing in 2019 and beyond. With the move to the Berkeley Lab campus, it will be incumbent on the JGI management team to reassess the highest safety risks at the new site and ensure appropriate awareness is raised through the mechanisms already put in place to minimize the chances of an injury occurring at the new site. As the team prepares for the move, it will be critical that the management team prepares emergency response plans and incorporates appropriate safety orientations to the staff prior to the move. See Milestone **OPS05**.

Financial Stewardship

SHORT- AND LONG-TERM FISCAL PLANNING

The move to the Berkeley Lab Campus will have an impact on the overall operating costs to support the JGI mission. With the move, there will be a significant increase in JGI labor costs due to the additional overheads that will be incurred to pay for site support activities. While the JGI has appropriately done long-term financial planning to support projected costs through FY20, it will be necessary to continue to identify the most important mission priorities and to take appropriate actions as needed to ensure that these priorities can be funded with the resources provided by DOE. Effective financial stewardship activities will include long-range budget planning (four year), regular



communications, and budget updates through monthly financial reporting and executive management calls between BER and the JGI, annual budget planning to prioritize and fund the most impactful science as a user facility, and submission of Field Work Proposals to identify levels of funding required to continue to support the global user community.

FINANCIAL ANALYSIS

With the increased labor costs associated with the move to the Berkeley Lab campus, it will be even more important to continually analyze and explore alternative ways to provide services to JGI users. The initial focus would be on the higher-cost elements of the JGI's business with computing and storage, sequencing (in house vs. outsource), shared services, and automation. Financial activities would not be limited to these specific areas, but the JGI would strive for success with an overarching organizational goal of finding the most cost-effective ways to conduct JGI business to free up valuable resources to perform high-impact science. See Milestones **OPS08** and **OPS09**.

OPERATIONAL SAVINGS THROUGH AUTOMATION, TECHNOLOGY IMPROVEMENTS, AND ALTERNATIVE SOURCES

In the future, there will be a continued focus on finding innovative automation and technology solutions to provide higher outputs at reduced costs. With labor costs representing a significant portion of the JGI budget, it will be imperative that the JGI continue to find ways to produce higher levels of output at reduced costs. Targets for these types of cost savings will include exploring lab automation solutions for user facility production activities, implementation of new technologies to reduce costs of sequencing and synthesis, exploring alternative software and computing solutions, and taking advantage of other DOE-funded resources to accomplish the JGI mission. See Milestone **OPS07**.

Appendix I: Implementation Milestones

Overview

Milestones are organized by the program or department facilitating their implementation. Many milestones require collaboration across multiple programs and departments. The "facilitating" program/department is accountable for overall progress and reporting on this milestone. All collaborating programs, groups and departments are responsible for implementation progress.

Acronyms

BDT	Business Development Team
CMO	Communications and Outreach
DSI	Data Science and Informatics Department
EKI	Eukaryote Informatics
FGP	Fungal and Algal Genome Science Program
GNT	Genomic Technologies Department
JLT	JGI Leadership Team
MGP	Metagenome Science Program
MIP	Microbial Genome Science Program
MTB	Metabolomics Group
OPS	Operations Department
PKI	Prokaryote Informatics
PLP	Plant Genome Science Program
SSP	DNA Synthesis Science Program
USP	User Programs Department

Genomic Technologies Department (GNT)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Biospecimens	**GNT01-2:** Establish nuclei preparation from plant tissues and fungi for chromatin and single-cell genomics	**GNT01-5:** Enable user access to molecular studies that require the JGI to have direct access to tissues and cells from plants and fungi (i.e., single-cell isolation for single-cell RNA sequencing)	FGP, PLP	p20				
Sequencing technologies	**GNT02-2:** Scale long-read sequencing technology output to 25Tb/annually	**GNT02-5:** Establish the ability to routinely generate 100kb sequence reads	-	p19				
	GNT03-2: Enable direct RNA sequencing for eukaryotic genome annotation	**GNT03-5:** Establish the ability to directly detect modified RNA bases	-	p19 p28				
	GNT04-2: Provide user access to SIP-metagenomics capabilities	**GNT04-5:** Establish SIP-metatranscriptomics capabilities	MGP	p20 p23			p56	

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Single-cell omics	**GNT05-2:** Develop eukaryotic de novo assembly for uncultivated fungi	**GNT05-5:** Expand unculturable eukaryotic de novo assembly to more complex genomes, including protists and/or algae	FGP	p25 p27				
	GNT06-2: Develop a robust Drop-seq end-to-end pipeline for single-cell transcriptome profiling as a user product	**GNT06-5:** Scale and apply Drop-seq for use on JGI plant flagships	PLP	p20 p29		p53		
	GNT07-2: Explore an approach for imaging gene expression patterns of complex tissues within a spatial context	**GNT07-5:** Combine spatial imaging transcriptomics and single-cell transcriptomics methods in plant tissue studies	PLP	p20		p53		
Experimental genome annotation	**GNT08-2:** Expand DAP-seq up to 1,000 TFs across 7 plant and 5 fungal species	**GNT08-5:** Apply DAP-seq to TFs across eukaryotes.	FGP, PLP	p19 p28				
	GNT09-2: Automate and scale existing complementary epigenomic and chromatin characterization assays for plants and fungi	**GNT09-5:** Develop eukaryotic synthetic biology design strategies that incorporate epigenomic/chromatin knowledge	FGP, PLP	p19		p46		
	GNT10-2: Transfer the cell-free transcription-translation protein expression technology being developed under the current ETOP and implement in production at the JGI	**GNT10-5:** Further develop the cell-free transcription-translation platform for additional applications, e.g., discovery of secondary metabolites	SSP, MTB			p45		
	GNT11-2: Develop experimental and computational tools for PCR-free high-throughput microbiome profiling	**GNT11-5:** Develop methods targeting endophytes for metagenomics and/or metatranscriptomics	MGP	p24	p35	p53		
DNA Synthesis and Strain Engineering platforms	**GNT12-2:** Develop a high-throughput yeast platform with assembly capacity for up to 1,000 pathways/year to remove workflow bottlenecks for budgeted assembly projects	**GNT12-5:** Further increase DNA synthesis capacity to enable assembly of up to 5,000 pathways/year	SSP			p42 p49		
	GNT13-2: Increase DNA synthesis capacity to enable cloning of up to 50Mbp/year	**GNT13-5:** Further increase DNA synthesis capacity to enable cloning of up to 300Mbp/year	SSP			p42		
	GNT14-2: Domesticate three major bacterial phyla (Proteobacteria, Actinobacteria, and Firmicutes)	**GNT14-5:** Domesticate 10 additional bacterial phyla	SSP			p43 p49		
	GNT15-2: Offer bacterial strain engineering product to users	**GNT15-5:** Develop fungal and algal landing pad technology for future user access	SSP			p43		

Microbial Genome Science Program (MIP)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Phylogenetic diversity	**MIP01-2:** Pilot integration of genomics with RNAseq and metabolomics for type strains	**MIP01-5:** Build KBase models for pilot data for 20 type strains integrating genome, transcriptome and metabolome data and validate a subset of models experimentally	DSI, MTB	p21		p51		
	MIP02-2: Conduct comparative genomic analysis of 1000 single-cell genomes from candidate phyla	**MIP02-5:** Experimentally validate the activity of several predicted enzymes from bacterial/archaeal candidate phylum-level lineages using DNA synthesis and functional assays	SSP	p22 p24				
Function-driven genomics	**MIP03-2:** Expand function-driven single-cell genomics to new approaches, such as the fluorescent labeling of specific enzyme targets or gene-FISH, to capture microorganisms based on a functional trait of interest	**MIP03-5:** Pilot novel function-driven single-cell genomics on environmental samples	GNT	p22 p23				
Inter-organismal interactions	**MIP04-2:** Develop and implement viral tagging for genomic interrogation of host-virus associations in the wild	**MIP04-5:** Apply viral tagging to a broad range of environments to create genomic host-virus catalog	GNT, MGP	p24				
	MIP05-2: Pilot single-cell sequencing of 20 protists	**MIP05-5:** Perform scale-up of protist single-cell sequencing by one order of magnitude	FGP	p24 p27				
Microbiome data science	**MIP06-2:** Complete global survey of giant virus MAG dataset, encompassing analysis of phylogenetic diversity and coding potential	**MIP06-5:** With research community, establish new Nucleo-Cytoplasmic Virus Orthologous Groups for giant virus phylogenomics	MGP, DSI		p38			
Earth's Secondary Metabolome Project	**MIP07-2:** Identify new sources of isolates containing target pathways from the GEBA project	**MIP07-5:** Enable access to novel secondary metabolite producers	MGP, MTB, FGP			p48 p49		

Metagenome Science Program (MGP)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Function-driven genomics	**MGP01-2:** For multiple metagenome-assembled genomes, perform experimental characterization of predicted metabolic activities	**MGP01-5:** Build community-level metabolic models from reconstructed genomes with characterized activities in partnership with KBase	MTB	p22				
Inter-organismal interactions	**MGP02-2:** Pilot SIP metagenomics from viral-targeted fractions	**MGP02-5:** Investigate virus activity and the potential impact on elemental fluxes from a specific environment	GNT	p23				
Phylogenetic diversity	**MGP03-2:** Include an evolving set of viral-centric and diversity-focused emphasis areas in the annual CSP call and convene the VEGA and NeLLi Symposia	**MGP03-5:** Partner with new users to develop diversity-focused and host-virus grand challenges, leveraging other capabilities (e.g., EMSL, NERSC, and KBase)	MIP					p61
Microbiome data science	**MGP04-2:** Enable phylogenome exploration and reconstruction of uncultivated microbial diversity from tens of thousands of microbiome samples	**MGP04-5:** Develop new visualization and analysis approaches for large-scale characterization of uncultivated lineages	PKI	p22	p39			
	MGP05-2: Develop new machine learning-based approaches for unsupervised identification of new viruses and prediction of viral-host interactions	**MGP05-5:** Experimentally validate computational predictions of viral-host interactions	MIP, PKI	p24	p38			
	MGP06-2: Develop new tools for analysis of microbial community assembly from microbiome data that incorporate macroscale ecological process models	**MGP06-5:** Develop, apply, and make available to the user community new computational methods and statistical approaches to test macroecological hypotheses	PKI		p39			
Earth's Secondary Metabolome Project	**MGP07-2:** Develop new scalable approaches to recover and perform comparative analysis of novel biosynthetic gene clusters from microbiome samples	**MGP07-5:** Expand functionalities to MycoCosm and Phytozome	MIP, FGP, DSI, MTB		p48 p49			

Fungal and Algal Genome Science Program (FGP)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Unculturable fungi	**FGP01-2:** Develop eukaryote-targeted binning and assembly methods for metagenome-extracted eukaryotic genomes	**FGP01-5:** Scale up single-cell genomics, analysis of co-cultures, and laser capture microdissection; develop the corresponding production capabilities	GNT, MIP	p25	p35 p36			p61
Enriching the Fungal Tree of Life data	**FGP02-2:** Improve assemblies and annotations of 10 to 20 key reference genomes across the Fungal Tree of Life	**FGP02-5:** Develop multi-omics datasets for these references and integrate them in MycoCosm	GNT, DSI, MTB	p25	p36	p48	p56	
Genome-scale modeling of fungi	**FGP03-2:** Enable metabolic modeling for fungi on the KBase platform and link curated models to 10 to 20 key reference genomes in MycoCosm	**FGP03-5:** Develop interactive tools for multi-omics data integration and analysis in MycoCosm	DSI, MTB	p25	p36		p56	
Functional characterization of fungal protein families	**FGP04-2:** Identify conserved gene families across fungi to apply experimental techniques for targeted functional annotation	**FGP04-5:** Mobilize the community to functionally characterize 100 new fungal gene families using high-throughput and targeted capabilities including those offered through FICUS	-	p26	p36	p45 p48	p56	p61
MycoCosm development	**FGP05-2:** Develop scalable pipelines to facilitate MycoCosm-wide updates	**FGP05-5:** Double the number of genomes in MycoCosm within the same database size or computational resources footprint	DSI	p26	p36			
Algal diversity	**FGP06-2:** Develop and release a comparative algal genomics resource	**FGP06-5:** Produce 100 algal reference genomes	DSI, PLP	p27	p36			
Algal user community	**FGP07-2:** Unite algal researchers around JGI-provided genomes, annotations, and web resources	**FGP07-5:** Develop algal grand challenges for broader community participation	USP, PLP	p26				p61

Plant Genome Science Program (PSP)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Improvements to plant reference genomes	**PLP01-2:** Sequence and assemble high-quality genomes >5GBs using single-molecule sequencing (i.e., PacBio). Sequence and assemble a high-quality hexaploid genome.	**PLP01-5:** Make routine sequencing tetraploid, hexaploid, and outbred complex genomes by integrating single-molecule sequencing with dense genetic mapping and HiC for chromosome construction	GNT	p28				
	PLP02-2: Improve reference annotations of flagships with new full-length cDNA sequencing and additional algorithmic improvements	**PLP02-5:** In flagship species, annotate new elements: ncRNA, promoters, TF binding sites, and enhancers	GNT, EKI	p28				
	PLP03-2: Develop a complete pangenome from high-quality references for one key JGI flagship, sorghum	**PLP03-5:** Develop complete pangenomes from references for additional 4 flagship plants	EKI	p28				
Comparative genome platform for plant gene function	**PLP04-2:** Develop additional 10 high-quality references across phylogeny that expand references into early angiosperms and early plants	**PLP04-5:** Complete reference genomes spaced ~50 MYA across angiosperm evolution and expand early plant genomes	EKI	p29				
	PLP05-2: Obtain single-cell plant transcriptomes of roots in 3 diverse plants	**PLP05-5:** Obtain single-cell plant transcriptomes for root, stem, leaf across 7 plants	GNT	p29				
Common gardens and phenotyping	**PLP06-2:** Develop draft pangenomes for DOE JGI flagships (*Populus*, *Brachypodium*, *Sorghum*, *Panicum*) suitable for association analysis	**PLP06-5:** Use these draft pangenomes in combination with common garden experiments to improve genotype to phenotype discovery	EKI	p29				
	PLP07-2: Collect baseline phenotyping data across several DOE JGI flagship populations	**PLP07-5:** Collect and analyze molecular, metabolomics, and microbiome phenotypes for common garden experiments	MTB, MGP	p29				
Understanding abiotic stresses in plants	**PLP08-2:** Develop a system and collaborations to perform consistent functional experiments for drought, nitrogen, temperature, and salt stress across plants	**PLP08-5:** Collect phenotypic, expression, and molecular functional data across stress conditions for comparisons of 5 to 7 species	-	p29				
Inter-organismal interactions	**PLP09-2:** Identify plant genes controlling microbial associations	**PLP09-5:** Build predictive models for plant-microbiome systems	EKI	p24				

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Harnessing in-house plant growth capabilities	**PLP10-2:** Leverage the JGI's in-house plant growth capabilities and new genomics approaches, including Drop-seq, ATAC-seq and Hi-C, including to characterize cell- and tissue-specific regulatory networks involved in abiotic stress. Explore using in-house plant growth resources to study metabolism and plant-microbe interactions relevant to bioenergy.	**PLP10-5:** Utilize the JGI's in-house plant growth capabilities and genetic resources to enable the development of reproducible model ecosystems that can be used by users to explore the effects of plant-microbiome interactions on plant productivity and health	GNT	p29				
Plant data delivery	**PLP11-2:** Provide fully documented API providing programmatic access to all Phytozome data including gene annotations, gene families, orthology, reference sequences, diversity datasets, and expression datasets	**PLP11-5:** Develop Phytozome pipelines, backend infrastructure, and interfaces to provide timely, performant, and intuitive end-user access to more than 250 genomes, 5 flagship pangenomes, and their associated omic datasets	EKI		p36			

Prokaryote Informatics (PKI)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Microbiome data science	**PKI01-2:** Develop methods and approaches to identify and characterize functional dark matter. Identify and characterize novel protein families from metagenomes.	**PKI01-5:** Include "dark matter" protein families in annotation pipeline. Identify and characterize novel protein families from metagenomes, and include them in the annotation pipeline of new metagenomes.	MGP		p38			
	PKI02-2: Implement newly developed standards for genomes of uncultured organisms and viral genomics in IMG/M	**PKI02-5:** Develop new standards for improved characterization of genome sequences in collaboration with the Genomics Standards Consortium and implement in IMG/M	MIP, MGP		p39			
	PKI03-2: Develop and implement in production protein clustering methods at the exascale level	**PKI03-5:** Successfully scale computational capabilities at the exascale level of data generation. Develop and implement new exascale data processing and analysis methods.	DSI		p33			
	PKI04-2: Complete the redevelopment of IMG/M's user interface for querying genes and genomes	**PKI04-5:** Scale IMG/M to support hundreds of thousands of genomes and metagenomes. Co-develop a number of analysis tools and applications in collaboration with KBase.	DSI		p35			
	PKI05-2: Establish automated bin/MAG identification from metagenomic datasets	**PKI05-5:** Implement large-scale identification of bins, automatic submission of high-quality metagenome bins as genomes, and complete development of new analytical and visualization tools for comparative analysis of bins	MTB	p22	p35			
	PKI06-2: Expand the repertoire of computational tools and pipelines for analysis of Secondary Metabolites in IMG-ABC	**PKI06-5:** Establish IMG-ABC as the central resource for the natural products community for computation and data management of BGCs for prediction of clusters, their regulation, and the products they encode	MTB, DSI		p35	p48		

Eukaryote Informatics (EKI)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Pangenome assembly	**EKI01-2:** Construct shotgun pangenomes for two flagship species from hundreds of accessions	**EKI01-5:** Develop a scalable method for robust assembly of eukaryotic pangenomes for both plants and fungi	PSP, FGP	p25 p27	p40			
Flagship plant data integration	**EKI02-2:** Prototype tools for integrating three distinct data modalities for a flagship genome	**EKI02-5:** Implement tools for "actionable" genetics in Phytozome	PSP	p29	p40			
Exascale	**EKI03-2:** Implement a robust scalable method for large-scale metagenome assembly	**EKI03-5**: Assemble five multiterabase metagenomes with MetaHipMer	PKI, MGP, DSI		p32 p33			

Metabolomics Group (MTB)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Metabolomics	**MTB01-2:** Pilot online tools enabling users to visualize and integrate metabolomic data with genomic information	**MTB01-5:** Release online tools that enable users to analyze, visualize, and integrate metabolomic data with genomic and proteomic information	GNT, DSI	p27		p45 p48 p49 p50 p51		
	MTB02-2: Increase metabolomics capacity to 8,000/yr non-polar sample assays	**MTB02-5:** Increase metabolomics capacity to 12,000/yr non-polar sample assays	GNT	p29		p45 p49 p50	p55	
	MTB03-2: Implement a LIMS for sample tracking	**MTB03-5:** Implement a complete end-to-end LIMS system for tracking and managing production metabolomics	GNT, DSI			p50		
Earth's Secondary Metabolome Project	**MTB04-2:** Characterize secondary metabolites from >250 diverse bacteria	**MTB04-5:** Offer targeted single-cell approaches for 10 bioenergy-relevant substrates as a productionized user capability	PKI, MIP, MGP, SSP, DSI, GNT	p23		p49 p52		
	MTB05-2: Identify metabolite products of target pathways from existing isolates	**MTB05-5:** Identify metabolite products from host strains containing novel biosynthesis clusters	-	p25		p45 p49		

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Lipidomics evaluation project	**MTB06-2:** Perform a pilot project to define user needs, demand, and resource requirements for offering a general lipidomics capability	**MTB06-5:** Decide if there is sufficient user demand and scientific impact to launch a lipidomics product	-			p53		
High-throughput functional analysis of proteins	**MTB07-2:** Pilot integrated high-throughput protein expression and metabolomic analysis	**MTB07-5:** Establish integrated robotic protein expression and metabolomic analysis for direct functional annotation	-			p45 p46 p49 p52		

DNA Synthesis Science Program (SSP)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Genomes to structure and function	**SSP01-2**: Develop a database to store and share function information for all enzymes characterized as part of the DNA Synthesis Science Program by its users. Enter data for at least 10 enzyme families.	**SSP01-5:** Expand to data for at least 50 enzyme families	GNT, DSI, MGP			p44 p46		
	SSP02-2: Develop an algorithm to help select genes of interest for characterization	**SSP02-5:** Develop an algorithm to help select genes to assemble pathways	GNT, DSI			p44 p46		
	SSP03-2: Develop a platform for high-throughput characterization of structures and functions of novel protein families found in environmental metagenome samples (10 novel protein families)	**SSP03-5:** Develop a platform to high-throughput characterize structures and functions of novel protein families found in environmental metagenome samples (50 novel protein families)	GNT			p44 p46		
High-throughput functional genomics	**SSP04-2:** Develop sets of HTP libraries (e.g., sgRNA libraries, promoter libraries, and TF libraries) for 10 species	**SSP04-5:** Develop sets of HTP libraries (e.g., sgRNA libraries, promoter libraries, and TF libraries) for 50 species	GNT			p44 p46		
Microbe-microbe and plant-microbe interactions	**SSP05-2:** Develop at least 12 engineered strains for use in medium-complexity defined plant-associated microbiomes consisting of up to 25 species total	**SSP05-5:** Engineer the plant-associated microbiome and test 10 plant growth-promoting traits	GNT, PLP			p44 p46		

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Earth's Secondary Metabolome Project	**SSP06-2:** Synthesize novel biosynthesis clusters and introduce into diverse host strains	**SSP06-5:** Develop new host strains belonging to at least five phyla and use to express novel pathways	SSP, GNT, MTB			p49		

Data Science and Informatics Department (DSI)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Portability	**DSI01-2:** Leverage container technology and Common Workflow Language in all JGI production pipelines	**DSI01-5:** Achieve the ability to execute all JGI production pipelines across DOE compute facilities as well as the cloud	PKI, EKI		p34			
Scalable analyses	**DSI02-2:** Make ExaBiome tools available to the user community	**DSI02-5:** Integrate data from across all JGI programs for functional analysis	PKI, EKI		p32			
Search capabilities	**DSI03-2:** Provide a scalable (handle hundreds of requests per second and billions of objects) query interface to JAMO and Data Warehouse; support complex queries	**DSI03-5:** Provide a common query interface for all JGI portals	PKI, EKI		p33 p34 p35			
KBase codevelopment	**DSI04-2:** Co-develop and implement a joint infrastructure for homology microservices with KBase	**DSI04-5:** Co-develop and implement a joint infrastructure for additional computationally intensive analyses with KBase	KBase		p34 p35		p57	
KBase user support/ engagement	**DSI05-2:** Complete a usability study of the user-facing JGI-KBase codevelopment effort	**DSI05-5:** Establish a diverse joint JGI-KBase user community that will be tracked through the number of joint KBase and JGI users	KBase		p33 p34 p35		p57	
KBase User Working Groups	**DSI06-2:** Participate in, lead, and coordinate subtopics within the UWGs	**DSI06-5:** Establish expanded UWGs for shared infrastructure with the JGI, KBase, and NERSC	KBase		p33 p34		p57	
FICUS data integration	**DSI07-2:** Work with EMSL to coordinate project and metadata tracking across organizations such that linkages are visible to external users and implement for FY20 call	**DSI07-5:** Provide data management and analysis tools for data integration across multiple facilities	USP		p34		p56	

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
User and prospective user training	**DSI08-2:** Develop and implement rigorous usability testing of JGI portals and tools	**DSI08-5:** Design and implement MOOC/webinar curriculum to increase the scale of outreach efforts in support of JGI data and tools	CMO, USP		p35			

User Programs Department (USP)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
FICUS development	**USP01-2:** Offer capabilities from an additional DOE user facility in an FY20 FICUS call	**USP01-5:** Work with other user facilities to streamline cross-user facility requests	-			p45	p54 p56	
Cross-facility collaboration	**USP02-2:** Expand the existing FICUS program with EMSL to include secondary metabolite characterization	**USP02-5:** Expand the existing FICUS program with EMSL to include KBase to help enable omics integration studies	MTB, DSI, SSP			p45 p46 p49	p55 p56	
User program impact	**USP03-2:** Use DOI citations to link publications to user programs (e.g., Community Science Program and JGI-EMSL FICUS) and thereby assess the impact of each program	**USP03-5:** Develop strategies to track secondary data use	DSI		p33			p65
Data impact	**USP04-2:** Begin officially reporting data users to DOE and address their needs with an annual data user survey	**USP04-5:** Grow the data user community by 20%	DSI					p65

Communications and Outreach (CMO)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Sustaining brand and message integrity	**CMO01-2:** Complete brand audit/assessment and initiate brand renewal effort	-	JLT					p60
	CMO02-2: Complete and evaluate refreshed onboarding process	-	OPS					p60
	CMO03-2: Design, implement, and assess effective communications module for staff	-	-					p60
Workforce training for educators and students	**CMO04-2:** Establish organizational infrastructure to support a single point of contact in CMO for all of the JGI's education activities	**CMO04-5:** Develop, test, and evaluate a high school classroom data portal and tool training module	MIP, MGP, FGP, PLP, SSP, MTB					p68 p69
	CMO05-2: Design and implement a "host" training module to develop new JGI intern supervisors	-	OPS					p69
	CMO06-2: Establish East Bay high school, community college, and UCB internship alliance priorities	-	MIP, MGP, FGP, PLP, SSP, MTB					p68
	CMO07-2: Conduct a review of the 5-year progress of the JGI-UC Merced partnership	-	-					p66

Business Development Team (BDT)

O C: Other Contributors; ID: Identification; IR: Interrogation; IV: Investigation; IG: Integration; IX: Interaction

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
IEP impact	**BDT01-2:** Develop a strategy and process for tracking impact and use of JGI products generated for industrial users	**BDT01-5:** Tabulate examples of product development and scientific discovery enabled by JGI products generated for industrial users	USP					p66

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	ID	IR	IV	IG	IX
Engaging new user communities	**BDT02-2:** Develop a standalone communications plan for industrial users	**BDT02-5:** Secure annual external funding totaling 2% or more of the overall JGI operating budget	CMO					p59 p62 p63
	BDT03-2: Complete targeted outreach to 200 companies for potential collaborations	**BDT03-5:** Develop a dedicated JGI scientific staff to support industry collaborations	JLT					p62
Translation to practice	**BDT04-2:** Complete 5 targeted SPPs	**BDT04-5:** Complete 20 targeted SPPs	MIP, MGP, FGP, PLP, SSP, MTB, GNT			p42 p43 p46 p48 p50 p51		p63
	BDT05-2: Establish a documented and reproducible process within the JGI for interacting with SPO and IPO groups at Berkeley Lab	**BDT05-5:** Expand industry engagement activities to include the larger Biosciences Area	OPS					p62 p63
	BDT06-2: Implement a data management infrastructure to support private data storage and computation	-	DSI					p63

JGI Leadership Team (JLT)

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	Stewardship
Roles and responsibilities	**JLT01-2:** Develop and communicate role cards for all roles at the JGI	**JLT01-5:** Identify gaps in roles and responsibilities and address these with cross-training, stretch assignments, and exposure to new activities	OPS (HR), all JGI	p71
Performance management	**JLT02-2:** Establish a continuous performance management culture built around feedforward and fair and consistent documented practices	-	OPS (HR), all JGI	p72

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	Stewardship
Career development	**JLT03-2:** Develop and communicate clear career paths for all job classes	-	All JGI	p72
	JLT04-2: Enhance, communicate, and deploy promotion and succession planning processes	-	OPS (HR)	p72
	JLT05-2: Implement a transparent and equitable hiring process for all roles	-	OPS (HR)	p72
	JLT06-2: Develop resources for postdocs' skill development and career advancement that prepare them for academic or industry roles (e.g., a mentorship/sponsor program and networking events)	-	MIP, MGP, FGP, PLP, SSP, MTB, JGI Postdocs & Early Career Scientists Association	p72
Skill development and learning	**JLT07-2:** Develop and deploy learning curricula based on current and future organizational needs	**JLT07-5:** Use the proximity of the IGB to other parts of Berkeley Lab/UC campus to develop new opportunities for staff to work on projects in other areas/facilities	-	p34 p73
	JLT08-2: Develop and deploy leadership curricula for current and future leaders	-	-	p73
Mentorship	**JLT09-2:** Implement and encourage participation by all JGI staff in a mentorship program	**JLT09-5:** Achieve participation of at least 25% of JGI staff in a mentorship program	OPS (HR)	p73
Work/life balance	**JLT10-2:** Organize workshops and activities to promote a healthier working environment and work-life balance	-	JGI SWELL team, JGI DE&I Working Group	p73
	JLT11-2: Develop a plan to address work/life balance concerns based on the priority needs of the JGI	**JLT11-5:** Implement measures to support work/life balance in alignment with Berkeley Lab policies	JGI SWELL team, JGI DE&I Working Group	p73
Diversity, Equity, and Inclusion	**JLT12-2:** Develop a DE&I strategy, along with tools and mechanisms to ensure implementation	-	JGI DE&I Working Group	p73

Operations Department (OPS)

AREA	2-YEAR MILESTONE	5-YEAR MILESTONE	O C	Stewardship
Transition of operations to the IGB	-	**OPS01-5:** Identify operational efficiencies where possible to maximize use of General and Administrative (G&A) support at Berkeley Lab & minimize operational costs to the JGI	-	p75
IGB transition — administration	**OPS02-2:** Revise administrative processes and the employee onboarding program to enable employees to operate effectively in the Berkeley Lab campus environment	-	CMO	p75
	OPS03-2: Transition procurement activities to G&A support vs. the dedicated buyer support program at the Walnut Creek Facility	-	-	p75
IGB transition — shipping and receiving	**OPS04-2:** Successfully integrate and document new shipping and receiving procedures at the IGB	-	-	p75
IGB transition — safety	**OPS05-2:** Transition all safety and emergency programs, processes, and documentation successfully to be in compliance with Berkeley Lab site processes and procedures	-	-	p75
IGB transition — all operations	**OPS06-2:** Update all JGI intranet content to reflect new procedures and policies to operate effectively at Berkeley Lab/IGB	-	CMO	p75
Explore and implement automation, new technology, and alternative sourcing solutions	**OPS07-2:** Implement an integrated automation solution that results in enhanced throughput and/or cost reductions to one of the JGI's production pipelines	**OPS07-5:** Implement 2 to 3 automation solutions that reduce operational costs	GNT, Science Programs	p76
	OPS08-2: Identify and implement new technology solutions in sequencing, synthesis, and metabolomics	**OPS08-5:** Reduce sequencing and synthesis costs by 10 to 25% at FY19 throughput levels	GNT	p76
	OPS09-2: Identify and implement less costly compute, storage, and software resources	**OPS09-5:** Reduce computational and storage support costs by 10 to 25% due to cost savings through alternative sourcing	DSI	p76

Appendix II: Contributors

JGI Staff

April 2018 Strategic Planning Retreat Participants

Adam Arkin, Lawrence Berkeley National Laboratory, DOE Systems Biology Knowledgebase (KBase), University of California, Berkeley

Shannon Bennett, California Academy of Sciences

Matthew Blow, JGI User Support Analysis Group Lead

Harvey Bolton, Environmental Molecular Sciences Laboratory (EMSL)

Benjamin Bowen, JGI Metabolomics Software Development

Siobhan Brady, University of California, Davis

James Bristow, Lawrence Berkeley National Laboratory, Biosciences Area

Eoin Brodie, Lawrence Berkeley National Laboratory, Earth and Environmental Sciences Area

Jan-Fang Cheng, JGI Functional Genomics Group Lead

Katy Christiansen, Lawrence Berkeley National Laboratory, Biosciences Area

Alex Copeland, JGI Genome Assembly Group Lead

Christopher Daum, JGI Sequencing Technology Team Lead

Samuel Deutsch, JGI DNA Synthesis Science Platform Lead

Emiley Eloe-Fadrosh, JGI Metagenome Program Head

Kjiersten Fagnan, JGI CIO / Data Science & Informatics Leader

David Gilbert, JGI Senior Manager of Communications & Outreach

Louise Glass, Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology Division, University of California, Berkeley

Igor Grigoriev, JGI Fungal Program Head

Nancy Hammon, JGI Workflow Planning Management

Peter Karp, SRI International

Linda Kinkel, University of Minnesota

Nikos Kyrpides, JGI Prokaryote Super Program Head

Wendi Lynch, JGI Administration

Rex Malmstrom, JGI Microscale Applications Group Lead

Nigel Mouncey, JGI Director

Vincent Noireaux, University of Minnesota

Trent Northen, JGI Metabolomics Group Lead

Ronan O'Malley, JGI Sequencing Technologies Group Lead

Len Pennacchio, JGI Deputy of Genomic Technologies

Daniel Rokhsar, JGI Eukaryote Super Program Head

Jeremy Schmutz, JGI Plant Program Head

Ben Shen, Scripps Research Institute

Gary Stacey, University of Missouri, DOE Biological and Environmental Research Advisory Committee

Shawn Starkenburg, Los Alamos National Laboratory

Owatha (Tootie) Tatum, JGI Business Development and Project Management Office Leader

Susannah Tringe, JGI Deputy of User Programs

Ray Turner, JGI Deputy of Operations

Axel Visel, JGI Deputy of Science

John Vogel, JGI Plant Functional Genomics Group Lead

Zhong Wang, JGI Genome Analysis Group Lead

Sean Ward, Synthace

Tanja Woyke, JGI Microbial Program Head

Yasuo Yoshikuni, JGI DNA Synthesis Science Program Head

Participants in the April 2018 Strategic Retreat in Napa, California

Members of JGI Advisory Committees

JGI Scientific Advisory Committee

Mark Adams, J. Craig Venter Institute

Brenda Andrews, University of Toronto

Harvey Bolton, Environmental Molecular Sciences Laboratory (EMSL)

Steve Briggs, University of California, San Diego

Carol Bult, Jackson Laboratory

Barry Canton, Ginkgo Bioworks

Jeff Dangl, University of North Carolina

Ashlee Earl, Broad Institute

Paul Flicek, European Bioinformatics Institute (ad hoc)

Jeff Gilbert, Corteva Agriscience

Greg May, Corteva Agriscience

Trina McMahon, University of Wisconsin-Madison

Dianne Newman, Caltech

Stefan Schillberg, Fraunhofer Institute of Molecular Biology and Applied Ecology

Kathy Yelick, Lawrence Berkeley National Laboratory

Informatics Advisory Committee

Toby Bloom, New York Genome Center

Amber Boehnlein, Jefferson Lab

Austin Che, Ginkgo Bioworks

David Dooling, Atomist

Paul Flicek, European Bioinformatics Institute (EBI)

Dan Jacobson, Oak Ridge National Laboratory (ORNL)

Nadim Jessani, Genedata AG

David Lewis, Illumina

Anne Deslattes Mays, The Jackson Laboratory for Genomic Medicine

Katherine Riley, Argonne National Laboratory (ANL)

Fungal Program User Advisory Committee

Scott Baker, Pacific Northwest National Laboratory

Randy Berka, Archer Daniels Midland

Ronald de Vries, Westerdijk Fungal Biodiversity Institute

Audrey Gasch, University of Wisconsin-Madison, Great Lakes Bioenergy Research Center

N. Louise Glass, Lawrence Berkeley National Laboratory, University of California, Berkeley

Stephen Goodwin, Purdue University

David Hibbett, Clark University

Francis Martin, Institut National de la Recherche Agronomique (INRA)

Michelle O'Malley, University of California, Santa Barbara

Joseph Spatafora, Oregon State University

Kathleen Treseder, University of California, Irvine

Adrian Tsang, Concordia University

Prokaryote Super Program User Advisory Committee

Jill Banfield, University of California, Berkeley

Ed DeLong, University of Hawaii at Manoa

Jonathan Eisen, University of California, Davis

George Garrity, Michigan State University

Steve Hallam, University of British Columbia

Phil Hugenholtz, University of Queensland

Janet Jansson, Pacific Northwest National Laboratory

Kostas Konstantinidis, Georgia Tech

Trina McMahon, University of Wisconsin-Madison

Monica Medina, Penn State University

Mary Ann Moran, University of Georgia

Nancy Moran, University of Texas at Austin

Jennifer Pett-Ridge, Lawrence Livermore National Laboratory

Rich Roberts, New England Biolabs

Lynn Schriml, University of Maryland

Ramunas Stepanauskas, Bigelow Laboratory for Ocean Sciences

Matt Sullivan, The Ohio State University

Plant Program User Advisory Committee

Siobhan Brady, University of California, Davis

Gloria Coruzzi, New York University

Jeff Dangl, University of North Carolina, Chapel Hill

Joe Ecker, The Salk Institute for Biological Studies

Samuel Hazen, University of Massachusetts, Amherst

Tom Juenger, University of Texas, Austin

Toby Kellogg, Danforth Center

Sabeeha Merchant, University of California, Los Angeles

Stephen Moose, University of Illinois

Sue Rhee, Carnegie Institution for Science

Bob Schmitz, University of Georgia

Gary Stacey, University of Missouri, DOE Biological and Environmental Research Advisory Committee

DNA Synthesis Science User Advisory Committee

Richard Bailey, independent consultant

Doug Cameron, Firstgreen Partners

Sunil Chandran, Amyris Inc.

James Flatt, Synthetic Genomics

Jay Keasling, Lawrence Berkeley National Laboratory

Megan Palmer, Center for International Security and Cooperation (CISAC), Stanford University

Elizabeth Sattely, Stanford University

Elizabeth Shank, University of North Carolina, Chapel Hill

David Weller, USDA-ARS

Other Contributors

Shane Canon, National Energy Research Scientific Computing Center (NERSC) and DOE Systems Biology Knowledgebase (KBase)

Robert Cottingham, Oak Ridge National Laboratory and DOE Systems Biology Knowledgebase (KBase)

Paramvir Dehal, DOE Systems Biology Knowledgebase (KBase)

Diane Dickel, Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology Division

Timothy Donohue, University of Wisconsin-Madison, Great Lakes Bioenergy Research Center

Daniel Drell, DOE Office of Science, Office of Biological and Environmental Research

Roy Kamimura, DOE Systems Biology Knowledgebase (KBase)

Michael Koepke, LanzaTech

Robert Landick, University of Wisconsin-Madison, Great Lakes Bioenergy Research Center

Ramana Madupu, DOE Office of Science, Office of Biological and Environmental Research

Mary Maxon, Lawrence Berkeley National Laboratory, Biosciences Area

Tina Nielsen, University of Wisconsin-Madison, Great Lakes Bioenergy Research Center

Dirk Norman, University of Wisconsin-Madison, Great Lakes Bioenergy Research Center

Phil Robertson, Michigan State University, Great Lakes Bioenergy Research Center

Appendix III: Abbreviations

AI	artificial intelligence
ALS	Advanced Light Source
ANL	Argonne National Laboratory
API	application programming interface
APS	Advanced Photon Source
ARM	Atmospheric Radiation Measurement and Climate Research Facility
ASCR	Advanced Scientific Computing Research
ASL	Area Safety Lead
BER	DOE Office of Biological and Environmental Research
BES	DOE Office of Basic Energy Sciences
BGC	biosynthetic gene cluster
BONCAT	bioorthogonal noncanonical amino acid tagging
BOOST	Build Optimization Software Tools
BRC	Bioenergy Research Center
C&O	Communications and Outreach
CRADA	Collaborative Research and Development Agreement
CRAGE	chassis-independent recombinase-assisted genome engineering
CRD	Computational Research Division
DBTL	design-build-test-learn
DE&I	Diversity, Equity, and Inclusion
DOE	U.S. Department of Energy
DOI	Digital Object Identifiers
EGSB	Environmental Genomics and Systems Biology Division
EMSL	Environmental Molecular Sciences Laboratory
epicPCR	Emulsion, Paired Isolation, and Concatenation PCR
ESnet	DOE Energy Sciences Network
EPS	exopolysaccharides
ETOP	Emerging Technologies Opportunity Program
FACS	fluorescence-activated cell sorting
FAIR	findable, accessible, interoperable, and reusable
FICUS	Facilities Integrating Collaborations for User Science
FIS-seq	fluorescent in situ sequencing
G&A	General and Administrative
G&CR	Government and Community Relations Office
GEBA	Genomic Encyclopedia of Bacteria and Archaea
GESEIA	Genomic Encyclopedia of Small Eukaryote Intracellular Associations
GOLD	Genomes OnLine Database
HPC	high-performance computing
HTP	high-throughput
IEEE	Institute of Electrical and Electronics Engineers
IEP	Industry Engagement Program
IGB	Integrative Genomics Building
IMG-ABC	Integrated Microbial Genomes — Atlas of Biosynthetic Clusters
IMG/M	Integrated Microbial Genomes and Microbiomes
IP	intellectual property
JAMO	JGI Archive and Metadata Organizer
JBEI	Joint BioEnergy Institute
JGI	Joint Genome Institute
KBase	DOE Systems Biology Knowledgebase
LC-MS/MS	liquid chromatography tandem mass spectrometry
LIMS	Laboratory Information Management System
MAG	metagenome-assembled genome
MOOC	massive open online course
ncRNA	non-coding RNA
NeLLi	New Lineages of Life

NERSC	National Energy Research Scientific Computing Facility
NMR	nuclear magnetic resonance spectroscopy
NRPS	non-ribosomal polyketide synthase
NSLS-II	National Synchrotron Light Source II
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
PAV	presence-absence variation
PI	principal investigator
PKS	polyketide synthases
PNNL	Pacific Northwest National Laboratory
QTL	quantitative trait locus
RiPP	ribosomally synthesized and post-translationally modified peptide
SAG	single amplified genome
SBOL	Synthetic Biology Open Language
SIP	stable isotope probing
SPP	Strategic Partnership Projects
SSO	single sign-on
STEM	science, technology, engineering, and mathematics
SWELL	JGI Safety and Wellness team
TF	transcription factor
TMS	Talent Management Strategy
Tn-Seq	transposon sequencing
UCB	UC Berkeley
UWG	User Working Groups
VEGA	Viral EcoGenomics and Applications
WD&E	Workforce Development and Education
WSEC	Women Scientists and Engineers Council

Photo Credit: Moira Hough

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Photo: Model of DNA. Photograph taken February 19, 1965. John H. Lawrence Collection-4821.

U.S. DEPARTMENT OF
ENERGY
Office of Science

UNIVERSITY
OF
CALIFORNIA

A U.S. Department of Energy
National Laboratory
Managed by the University of California

18-JG-5163