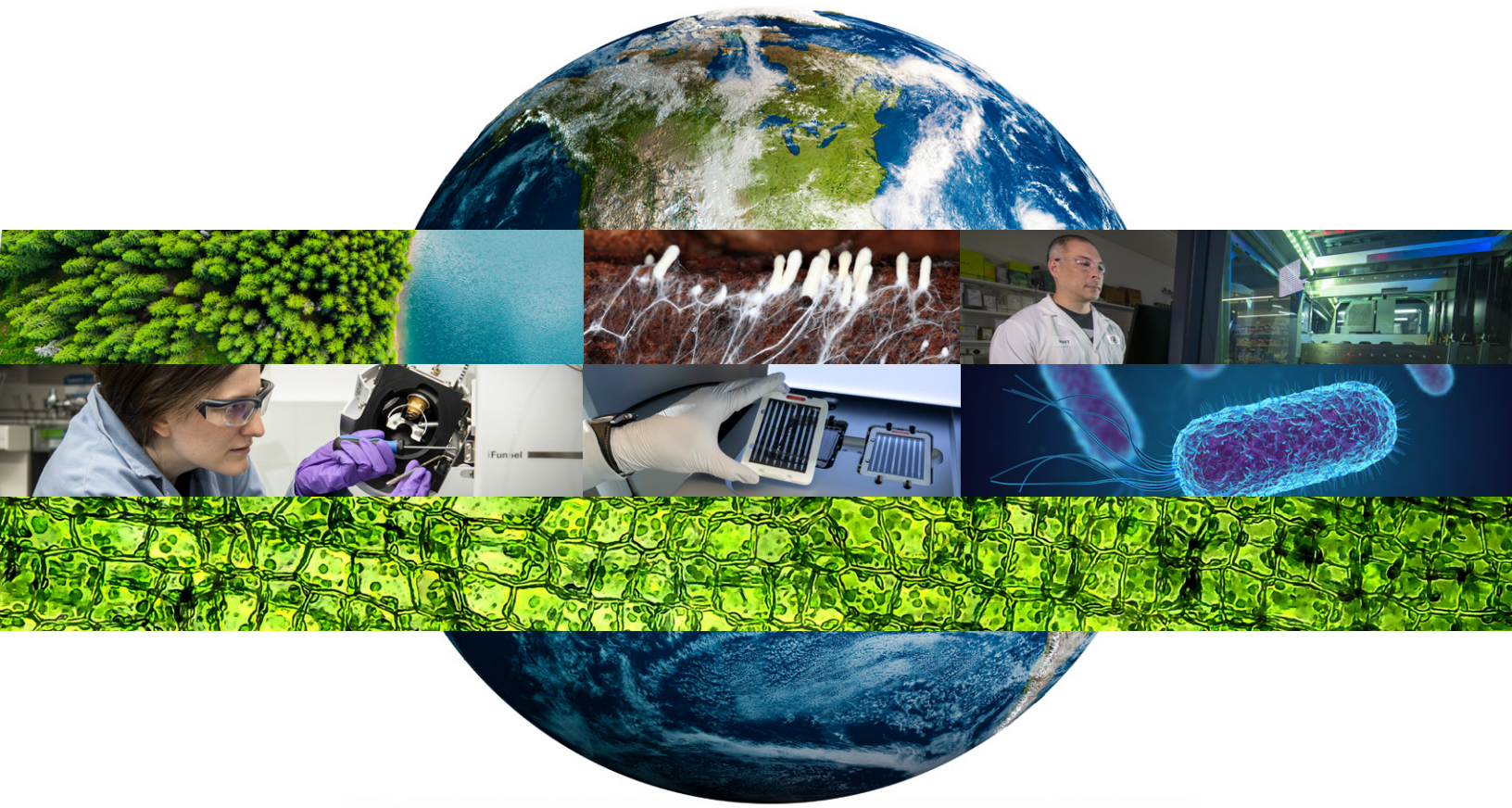
5-Year Strategic Plan

U.S. Department of Energy
Joint Genome Institute

April 2024

Innovating Genomics to Serve the Changing Planet



JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

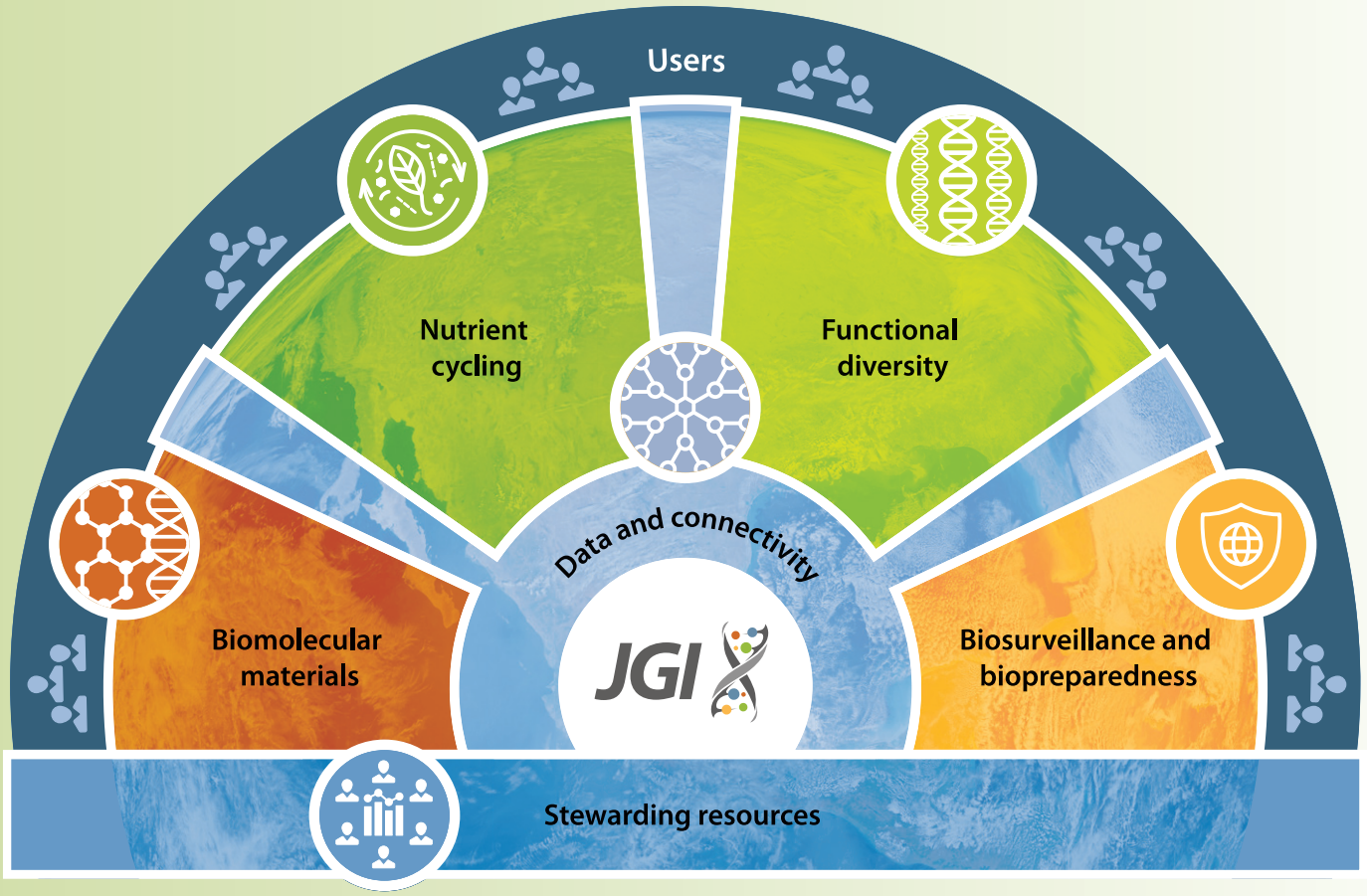

Users
Nutrient cycling
Functional diversity
Data and connectivity
JGI
Biomolecular materials
Biosurveillance and biopreparedness
Stewarding resources

Table of Contents

Vision

Lead genomic innovation for a sustainable bioeconomy

Mission

As a U.S. Department of Energy Office of Science user facility, we provide advanced genomic capabilities, large-scale data, and professional expertise to support the global research community in studies of complex biological and environmental systems. We optimize our service to the community through responsibly managing our people and resources.

Executive Summary

At the frontier of science, large-scale genomics research unveils life's dynamic intricacies through sequencing, analyzing, understanding, and harnessing DNA, the universal code of life. The U.S. Department of Energy (DOE) Joint Genome Institute (JGI) leads the integration and application of genomics for energy and environmental research to drive transformative innovation in pursuit of a sustainable bioeconomy for tomorrow's challenges.

Originally founded as a sequence production facility for the Human Genome Project, over the past 26 years the JGI has evolved into the world's only genome center exclusively dedicated to helping scientists unravel the mysteries of biology for environmental and energy challenges. Among the 28 DOE Office of Science (SC)-supported user facilities, the JGI is unique in providing access for researchers and communities of users to high-throughput (HTP) and state-of-the-art sequencing technologies, advanced capabilities in functional genomics, DNA synthesis and metabolomics, and extensive computational analysis tools and data portals. The JGI supports more than 2,300 primary users with active project proposals, over 20,000 secondary users interacting with JGI-provided systems and tools to analyze JGI-produced data, and countless others who advance the DOE science mission through further analyses building upon the work of JGI users. Funded through the DOE SC Biological and Environmental Research (BER) program, the JGI aligns its capabilities and research directions with those of the DOE BER genomics science mission[1] and has become an essential cornerstone of the global research endeavor in energy and environmental genomics. In particular, the JGI mission supports and aligns with the primary research areas of BER's Biological Systems Science Division (BSSD), including bioenergy research, environmental microbiome research, and biosystems design. The JGI's work in basic science lays the foundation for application-focused DOE BER programs and projects. That the strategic vision of the JGI advances continuously to stay aligned with the evolving DOE BER research mission is therefore essential.

Since 2011, the JGI has used a systematic and rigorous strategy process, which involves the development of long-term visions in close and inclusive coordination with JGI users and stakeholders, as well as the development of milestones and implementation plans to guide the realization of the institutional vision. Building on the successful development and implementation of two previous JGI strategic plans (*Forging the Future of the JGI*, 2012;[2] *Beyond Basepairs*, 2018[3]), we present here our vision for the next five years

1 US Department of Energy, 2021, *Biological Systems Science Division Strategic Plan*, Washington, D.C.: Department of Energy, **https://science.osti.gov/-/media/ber/pdf/bssd/BSSD_Strategic_Plan_2021_HR.pdf**.

2 Joint Genome Institute, 2012, *Forging the Future of the JGI*, Berkeley, California: Joint Genome Institute, **https://jgi.doe.gov/about-us/10-year-jgi-strategic-vision**.

3 Joint Genome Institute, 2018, *Beyond Basepairs – A Vision for Integrative and Collaborative Genome Science*, Berkeley, California: Joint Genome Institute, **https://jgi.doe.gov/latest-jgi-strategic-plan-released**.

in the evolution of the JGI. The title of this new plan, *Innovating Genomics to Serve the Changing Planet*, reflects the JGI's leadership in evolving capabilities to serve a broad user base to address Earth's issues. Like our 2018 plan, it is accompanied by an extensive set of more than 130 milestones that we will deliver to meet our strategic objectives (see **Appendix I: Implementation Milestones**). These milestones will be carefully tracked throughout the next five years by JGI management, to ensure progress toward implementing our vision.

The new vision and mission and this strategic plan outline JGI's planned advancements over the next five years in enabling our users to address environmental and energy challenges to support a sustainable bioeconomy. In particular, we will focus our efforts on how genomes translate into the functions of organisms and biological systems (e.g., methane emission, carbon sequestration) and how these allow organisms to adapt and respond to changing environmental conditions (e.g., drought-impacted feedstocks). The biological principles learned from these efforts can be used to design and engineer new biological systems for beneficial purposes in support of the bioeconomy. We will continue to innovate and optimize the capability portfolio we offer to our users based on their needs and those of the evolving DOE BER research mission.

Our plan goes beyond outlining the JGI scientific priorities for the next five years to also include an institutional vision that carefully considers the opportunities arising from leveraging laboratory and computational technologies at scale, improving connectivity across the DOE data ecosystem, enhancing the JGI's impact through communication, and maximizing stewardship of DOE resources. These overarching goals are reflected in the **new vision statement** of the JGI: "*lead genomic innovation for a sustainable bioeconomy*."

With extensive input from a diverse set of stakeholders, the JGI has identified four **strategic themes** to realize this vision. **Nutrient Cycling** addresses the critical need for better comprehension of the carbon cycle and other biogeochemical cycles of crucial importance for understanding and mitigating climate change, and for developing environmentally sustainable solutions for bioproducts to support the bioeconomy. **Functional Diversity** encompasses approaches for identifying, studying, and harnessing the vast range of biological functions that have evolved on Earth, encoded in the genomes of plants, fungi, algae, protists, bacteria, archaea, and viruses. Sequencing and advanced data mining efforts allow us to tap into this virtually unlimited range of functions and to use them in

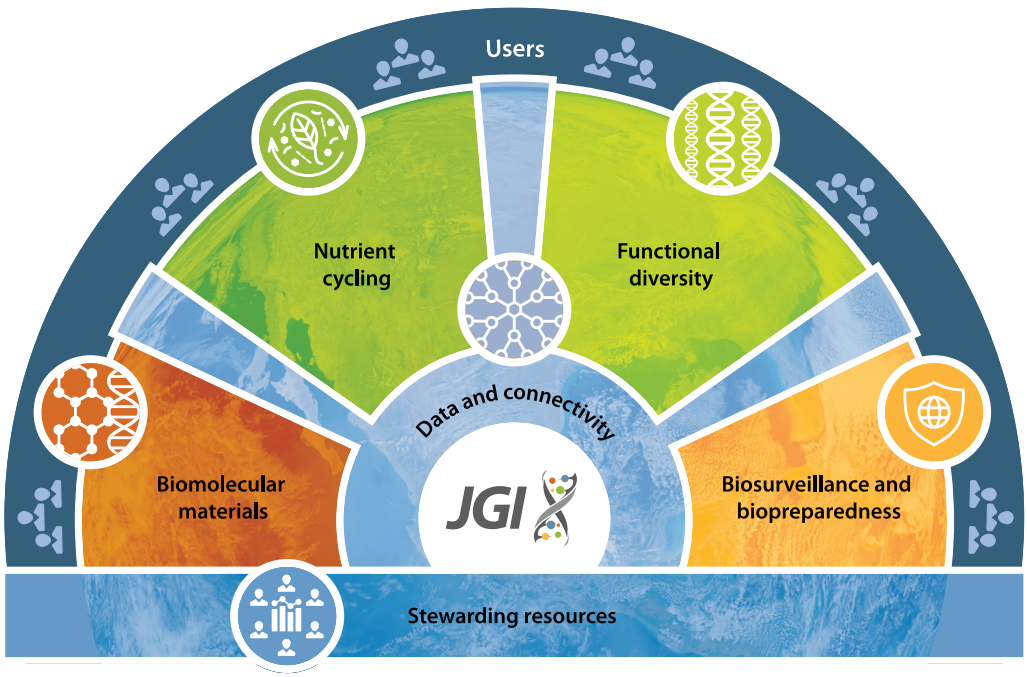


Fig. 1. Strategic themes and initiatives. The new JGI roadmap encompasses four strategic themes: Nutrient Cycling, Functional Diversity, Data and Connectivity, and Stewarding Resources. Additionally, it includes two new strategic initiatives: Biomolecular Materials and Biosurveillance and Biopreparedness.

engineering applications in support of the bioeconomy. **Data and Connectivity** addresses the challenges and opportunities associated with ever-increasing amounts of sequence data, as well as the growing diversity of data types that can be produced using contemporary and yet-to-be-developed technologies. Along with further scaling of JGI data production capabilities, it will be critical to develop systems for the analysis, storage, and dissemination of the resulting massive-scale datasets in partnership with other facilities and resources. In parallel, development and adoption of data and metadata standards will be critical to enable reuse and interoperability of these datasets by JGI's data users and the broader community. **Stewarding Resources** provides the essential foundation for the JGI's ability to support its vision. The unique expertise and diversity of the JGI workforce are critical to the scientific progress, technological innovation, and support for users that define the JGI's success in achieving its mission. Continued evolution of this workforce is essential. Clearly communicating JGI capabilities, user projects, and successes, along with investing in workforce development and efforts to make JGI processes more efficient, will help the JGI support progress toward a sustainable bioeconomy.

As a leader in genomic innovation, the JGI is continually seeking new ways and areas in which to enable scientific breakthroughs in alignment with our mission to extend our impact. From discussions with stakeholders, we have identified two new **strategic initiatives** in which to apply JGI technologies. In collaboration with the Molecular Foundry (TMF), another DOE user facility at Lawrence Berkeley National Laboratory (LBNL), the JGI will develop a **Biomolecular Materials** science capability. This capability will enable researchers to develop a mechanistic understanding of the molecular underpinnings of how these materials are synthesized and regulated, and how they interact with other cellular processes and the external environment. It will also explore opportunities to harness these findings to produce advanced biomaterials with users, aligned with the vision of an environmentally sustainable bioeconomy. In a second strategic initiative building on JGI contributions during the response to the COVID-19 pandemic, the JGI will engage in activities to address national needs in **Biosurveillance and Biopreparedness**. This will leverage the vast and unique data resources of the JGI to discover relevant genes and genomes across suitable sample matrices, develop advanced wet-lab and computational methods for biodetection and biosurveillance, and enable users to perform studies in this area through scaled DNA synthesis capabilities.

Through all this work, that the JGI leverages its existing partnerships and develops others providing complementary expertise to advance the science and serve as impact multipliers is vital. The JGI has a long-term partnership with the HudsonAlpha Institute for Biotechnology for our Plant Program, and is further leveraging the services of the Arizona Genomics Institute for DNA and RNA extractions to support plant and other efforts. The JGI works closely with DOE computational facilities (e.g., the National Energy Research Scientific Computing Center [NERSC][4]) and data infrastructure projects (the DOE Systems Biology Knowledgebase [KBase][5] and the National Microbiome Data Collaborative [NMDC][6]) to provide access to datasets, workflows, and tools to seamlessly expand data functionality. Close partnerships with other user facilities, especially the Environmental Molecular Sciences Laboratory (EMSL),[7] enable the ability to leverage their capabilities through the Facilities Integrating Collaborations for User Science (FICUS) program. The JGI has multiple partnerships with educational institutions to develop the next generation of the genomics workforce. The JGI is also conducting outreach with companies as part of our Industry Engagement Program to facilitate transformational research.

Over the next five years, the JGI will continue to stand at the forefront of scientific innovation, using large-scale genomics research to unravel the complexities of life by serving our users and focusing efforts on the strategic themes of Nutrient Cycling, Functional Diversity, Data and Connectivity, and Stewarding Resources, to address pressing environmental and energy challenges. Additionally, the JGI is launching two new initiatives, Biomolecular Materials and Biosurveillance/ Biopreparedness, while fostering partnerships with institutions, computational facilities, and educational organizations to support transformative research in genomics and biotechnology. Together these efforts will allow the JGI to realize its vision to "*lead genomic innovation for a sustainable bioeconomy*."

4 https://www.nersc.gov
5 https://www.kbase.us
6 https://microbiomedata.org
7 https://www.emsl.pnnl.gov

Strategic Drivers

Toward a Sustainable Bioeconomy

Biology holds the power to create a multitude of solutions to address some of the most pressing challenges faced by human society. Innovation in genomics science, and the resulting opportunities to drive progress toward a sustainable national and global future, have been clearly recognized as strategic focus areas by the DOE BSSD.[8] The bioeconomy encompasses all economic activity derived from biotechnology and biomanufacturing and is commonly defined as "the knowledge-based production and utilization of biological resources, innovative biological processes and principles to sustainably provide goods and services across all economic sectors."[9] The power of the bioeconomy has been impressively exemplified by the rapid development and production of diagnostics, vaccines, therapeutics, and protective devices during the COVID-19 pandemic. However, its scope reaches far beyond human health applications. The use of biotechnology and biomanufacturing can also drive our abilities to meet our climate and energy goals, improve food security and sustainability, develop new materials, and develop the bioeconomy into a core manufacturing base. In recognition of these opportunities, there is now a renewed and strong focus on using biotechnology and biomanufacturing to create new services and products, and to mitigate the effects of climate change.[10]

The roots of biotechnology lie in powerful biological and biochemical functions encoded in the genomes of organisms. These naturally occurring genomic functionalities can be used for a wide range of applications, either directly, engineered, or as inspiration for engineering, to drive the development of bio-based processes from sustainable feedstocks. In coupling sequencing, DNA synthesis, and other omics technologies with computational analysis and prediction, the JGI is a leader in harnessing genomes and genomic information to drive innovations that can power the bioeconomy. The JGI, along with its users, can continue to develop tools and technologies that unlock biological data and advance scientific discovery and development to enable the acceleration and escalation of bio-based product commercialization.

In 2023, the Office of Science and Technology Policy (OSTP) released a report, *Bold Goals for U.S. Biotechnology and Biomanufacturing*,[11] describing a wide-ranging set of activities required to mitigate the effects of climate change and build a strong and resilient bioeconomy. The activities of the JGI align to these goals. Examples include the following:

- **Climate**: In plant and microbial biology, the JGI supports its users by providing the technologies, capabilities, and expertise to increase understanding of biological processes' role in climate change. JGI users also explore how this knowledge can be harnessed to mitigate climate effects through developing increased plant resiliency, increased carbon sequestration, and sustainable routes to convert bio-based feedstocks into recyclable-by-design polymers to replace existing plastics and for other bioproducts.
- **Food and agriculture**: JGI work in understanding nutrient cycling processes in terrestrial environments can aid in improving plant health of food and energy crops, reducing methane emissions, and optimizing the use of waste feedstocks for food production. The JGI (and DOE BER) focus on energy crops, and use of marginal lands for their cultivation provides feedstock supplies that do not interfere with food and feed crops.
- **Supply chain**: By discovering novel genome-encoded metabolic pathways for converting

8 DOE BER, 2021, *Biological Systems Science Division Strategic Plan*, Washington, DC: DOE BER, **https://genomicscience.energy.gov/doe-ber-biological-systems-science-division-strategic-plan**.
9 Global Bioeconomy Summit, 2015, *Communiqué of the Global Bioeconomy Summit 2015: Making Bioeconomy Work for Sustainable Development*, Berlin: Global Biosecurity Summit, **https://gbs2020.net/wp-content/uploads/2021/10/Communique_final_neu.pdf**.
10 Executive Office of the President, September 12, 2022, Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe and Secure American Bioeconomy, Washington, D.C., **https://www.federalregister.gov/d/2022-20167**.
11 Released March 2023, available at **https://www.whitehouse.gov/wp-content/uploads/2023/03/Bold-Goals-for-U.S.-Biotechnology-and-Biomanufacturing-Harnessing-Research-and-Development-To-Further-Societal-Goals-FINAL.pdf**.

sustainable feedstocks into diverse chemicals, the JGI can enable new production routes for resilient supply chains.

- **Health**: While the JGI does not focus directly on health applications, many of the tools and resources for systems biology research it develops have direct secondary applications in human health. For example, its efforts in secondary metabolites can lead to novel therapeutics, and its work on environmental microbiomes provides tools, resources, and conceptual insights that can inform our understanding of human microbiomes.
- **Cross-cutting advances**: Across sectors of research and manufacturing, genome-enabled biology can seed new innovations, products, and applications. The JGI will continue to discover new genes, pathways, and organisms, and provide the means to explore their respective functions.

The JGI's contribution to a sustainable bioeconomy will be further enhanced through our role as an SC-supported user facility. We not only provide an array of genomics products but also educate our users on how to effectively use these products to answer their questions, gain new insights beyond those questions, and share these approaches with others. In addition to data generation for our users, our data policies prioritize enabling the global research community to access these data. By continuously expanding, improving, and refining our data portals and pipelines, we are making it easier for those working to foster the bioeconomy to find and leverage data produced by the JGI for their innovations.

An Evolving Scientific User Facility Mission

As an SC-supported user facility, the JGI supports DOE's mission in basic research to address fundamental biological questions by providing advanced genomics capabilities to our users. At the heart of this is the relationship between genomic information and observable phenotypes, including the physiology and metabolism of cells and organisms, and the influence of these phenotypes within their respective biological systems. The JGI also supports studies of how these systems interact with each other and their surrounding environments. Through this knowledge, the understanding of biological principles and processes can be harnessed to develop biology-based and -inspired solutions to environmental and energy challenges, in particular nutrient cycling and carbon sequestration.

Fig. 2. Authors citing or collaborating with the JGI, 2018 to 2023. The map illustrates the affiliations of the approximately 300,000 authors who have collaborated with JGI personnel or primary users, or who have cited JGI publications, data, or data systems.

Driven by rapid technology developments and an evolving scientific mission over the past 25 years, the JGI has moved beyond simply cataloging life's diversity through sequencing to functionally understanding the dynamics of biological systems, organisms and metabolic pathways, and the molecules that underpin these processes. Combining computational and experimental platforms has enabled us to conduct large-scale data generation, processing, and analysis. The advent of new machine-learning (ML) approaches will power new types of analyses, such as those we are already seeing in the realm of structural biology. We foresee that increasing connections between computation and datasets with experimental platforms for validation will further drive new scientific inquiry across disciplines.

Through the strategic themes and initiatives articulated in this new strategic plan, the JGI will be well positioned to continue to enable new understanding of genomics-based discovery but also serve as a conduit for translation to applied research and development toward the growing bioeconomy. Our geographical location in the San Francisco Bay Area gives us a lens into the biotechnology industry given the proximity to many companies, helping us play a role in addressing their challenges. Genome-enabled science is well poised to spur on the bioeconomy, and JGI leadership in genomics innovation will enable biotechnology advancements.

Technology Drivers

The main focus of the JGI since its inception as a user facility has been supporting its users through (1) experimental data generation and (2) associated computational analysis.

Data generation occurs through platforms dedicated to sequencing, synthesis, single-cell analyses, and metabolomics. In the next five years, we will continue to support these highly successful areas while expanding our capabilities to include imaging-connected methods, such as spatial transcriptomics in plant tissues. The biggest footprint of data generated has and will continue to be in the traditional genome center area of sequencing and our development of an advanced next-generation Production Sequencing Platform. Recent and foreseeable commercial product releases will provide a threefold increase in both short- and long-read sequencing output in the immediate future, creating a critical strategic need to capitalize on this vast increase in potential output through matched major facility upgrades to enable the associated need for increased sample handling. Similarly, the DNA Synthesis, Single-Cell, and Metabolomics Platforms are targeted for substantial growth with expanding community interest as costs continue to decrease, enabling additional output and unprecedented abilities to probe biology. We also anticipate further integration of these datasets through multi-omic data generation on identical samples to explore genomes, transcriptomes, and metabolomes as functionally connected units. On top of this, our cellular resolution will grow through large-scale application of single-cell gene expression capabilities to a focused set of JGI flagship plant genome species as feedstocks and for environmental processes. Important related technology development will target imaging-associated, single-cell-resolved transcriptomes to relate cell type functions to their organization within multicellular tissues. Finally, we will seek to expand our capabilities in plant synthetic biology to enable advancements in harnessing plant data mining, DNA synthesis, and ultimately plant genome engineering to convert sequence data into a functional assessment of plant biology.

JGI data production and analysis capabilities have led to exciting scientific breakthroughs by primary and secondary users. Our **data strategy** will support the facile access and reuse of the high-quality, standardized data produced in support of our user community. We anticipate increased interest in large quantities of data for use in developing artificial intelligence (AI) or ML methods for biology, as efforts like AlphaFold illuminate the value of these techniques. The JGI has deployed modest increases in computing capacity over the past decade. The biggest computational shifts have come through partnerships with the Computing Sciences Area at Berkeley Lab. The JGI has engaged in collaborations with NERSC, the ExaBiome project with the Applied Mathematics and Computational Research Division, and Science Search activities with the Scientific Data Division. We will build on these partnerships to maintain state-of-the-art computational capabilities.

Development of the JGI Strategic Plan

The JGI is part of the national ecosystem of scientific user facilities supported by the DOE SC. Given the complex scientific, policy, and technology drivers described previously, the JGI uses a rigorous and inclusive visioning process to tap into the diverse expertise of its users, funders, stakeholders, staff, and partners to develop a trajectory for continued success aligned with the evolving research mission of the DOE Office of Science. The JGI is located within the Biosciences Area of Berkeley Lab; therefore, alignment with the area's strategic directions[12] is critical to maximize synergies.

The current strategic plan emerged out of a year-long visioning process initiated and coordinated by JGI leadership. In an initial series of internal workshops, the JGI leadership team focused on context discovery, data gathering, and retrospective analysis of the previous strategic plan and its implementation. Building on themes and directions from these workshops, the JGI held a three-day strategic planning retreat in March 2023 that brought together JGI staff and external stakeholders to discuss core topics (see **Appendix II: Contributors**). In plenary and breakout sessions, retreat participants discussed topics including (1) bioenergy, bioproducts, bioeconomy; (2) global element cycles; (3) biosystems design; (4) biopreparedness and biosurveillance; (5) data and metadata; (6) technology drivers to support science drivers; (7) impacts as measures of success; and (8) communication with multiple audiences and communities. Beyond this retreat, JGI leadership also received input on the strategic plan from DOE BER management and the JGI advisory committees. Input from the strategic workshop and advisory committees culminated in the development of the final vision, mission, strategic themes, and strategic objectives by the JGI leadership team. An initial draft of the full strategic plan was shared with Berkeley Lab Biosciences Area leadership, advisory committees, retreat participants, and all JGI staff for input. Extensive comments and suggestions provided by these groups were incorporated into the final plan prior to release.

Overview of the JGI Strategic Themes

In this plan, we describe four strategic themes that will guide JGI activities over the next five years, with the goal to support our vision of leading genomic innovation for a sustainable bioeconomy. The next four subsections provide an overview of the motivations driving each of these themes, as well as the main activities we will pursue under each of them. All four themes are described in detail throughout the remainder of this strategic plan.

Understanding and Using Biomolecular Mechanisms of Nutrient Cycling

In natural ecosystems, carbon and other nutrients flow in a continuous cycle between inorganic and organic pools. Humans have altered these cycles intentionally in agricultural settings to increase yield and unintentionally by burning fossil fuels. Human-induced imbalances in these cycles often result in negative environmental impacts, including eutrophication of surface waters, algal blooms, contamination of groundwater, and climate change. Finer control of these cycles could minimize negative environmental impacts, reverse existing environmental damage, and provide new ways to sustainably increase plant productivity. A deeper understanding of the biological drivers of these cycles is essential to develop methods to achieve this goal, and recent advances in genomics, metabolomics, synthetic biology, and experimental systems have given us the tools to obtain this knowledge. As a leader in these fields, the JGI is ideally positioned to support users studying the role of natural genomic diversity in nutrient cycling using cutting-edge genomic and metabolomic technologies, determine the cast of species and biological pathways relevant for the major biogeochemical cycles, and determine how interactions between microbes, plants, and the environment influence nutrient cycling. By developing and employing new experimental systems across scales, we will enable the research necessary to test hypotheses generated from the mining of enormous datasets, including the efficacy of interventions designed from these data. Finally, we will support our users in exploring how carbon from plants can be used for bioproducts, essentially creating a new carbon pool and reducing the release of carbon from fossil fuels.

12 Berkeley Lab, 2019, *Biosciences Strategic Plan*, Berkeley, California: Berkeley Lab, **https://biosciences.lbl.gov/strategic-plan**.

Fig. 3. Overview of the JGI strategic themes and objectives.

Understanding Functional Diversity across the Domains of Life

Biodiversity on Earth is immense and with it the morphological and physiological traits of species that contribute to the maintenance of ecosystem processes and functioning. To leverage what nature has evolved over billions of years to move us toward an efficient bioeconomy, it is critical to discover, understand, and characterize functional diversity across the domains of life. Capitalizing on the robust foundation the JGI has built over the last two decades in genome sciences, we will support our users in characterizing functional traits in plants, fungi, algae, bacteria, archaea, and viruses across DOE-relevant ecosystems. We will work with our users to identify mechanisms and processes that govern ecosystem communities, including community assembly and antagonistic and mutualistic interactions. We will expand sequence-based discovery, generating a diversity of quality reference genomes and pangenomes from cultivated and uncultivated organisms. We will integrate these data using conventional computational approaches as well as nascent advanced AI-enabled strategies for gene function discovery and prediction. These in silico methods will guide and facilitate experimental efforts for the validation of plant and microbial functions and phenotypes, including multi-omics, HTP measurements through genome-scale approaches, and phenotypic analyses in uncultivated systems. An emphasis will be on empowering JGI users to expand functional insights through tool development, large collaborative projects, and user-centric prioritization for wet-lab functional genomic capabilities. The JGI will also begin to use these functional insights for biosystems design. Genome-enabled science is at the center of basic bioeconomy research. Decoding nucleic acids while also beginning to validate the function of the different building blocks encoded by the genomic data will keep the JGI at the forefront of this endeavor.

Leveraging Scale in Data and Connectivity

Progress toward a sustainable bioeconomy is driven by access to high-quality data for exploration and discovery. Patterns are discoverable only with large amounts of high-quality data from well-designed experiments. Biological technology has progressed to a point where more data are produced than can be analyzed in aggregate without assistance from AI. The best results are obtained from training models on datasets that are of high and consistent quality, well-curated, and consistently annotated. The JGI recognizes that a failure to develop and evolve such data resources will create

insurmountable hurdles for applying current and emerging advanced analysis strategies in this area.

There is an implicit trade-off between standardized and bespoke processes for data generation. Standards allow for process optimization and production of large amounts of data, whereas bespoke experiments demonstrate the utility of new technologies or capabilities. User facilities are a centralized resource capable of generating unique, standardized data for a large number of users through economies of scale. The JGI will focus its efforts on systems and experiments that create or contribute to the data resources needed for DOE and the bioeconomy. We will ensure that we are engaging a diverse primary and secondary user community through an emphasis on improved accessibility.

Enhancing the JGI's Impact through Nurturing Its People, Systems, Processes, and Communications

The foundation of a house is critical because it must support and distribute the full weight of the building and provide stability. The foundation of the JGI is its people—the current and future workforce—and the network of systems and processes for daily operations, mortared in place by the stories of and about them shared with a variety of audiences.

JGI scientific, technological, and computational capabilities all rest on this base, allowing the institute to serve as a significant upstream contributor of genomic products and data that are in turn harnessed toward the development of a sustainable bioeconomy. By placing a focus on cultivating an excellent workforce, thorough and targeted communications, optimized processes and systems, and opportunities for working cross-functionally, the JGI will amplify its impact toward its mission to support the global research community in the field of genomics.

Strategic Theme 1:

Nutrient Cycling

Understanding and Utilizing Biomolecular Mechanisms of Nutrient Cycling

Background

Understanding the global cycles of carbon and other nutrients is necessary to maximize plant productivity while minimizing external inputs, to increase carbon storage in soils, manage nutrient metabolism, conversion, and leaching, and identify pools of raw materials for the emerging bioeconomy. The global element cycles are driven by interactions between plants, microbes, and the environment, yet most of the organisms and interactions in these extremely diverse communities are unknown or poorly understood. As an SC-supported user facility, the JGI is well positioned to leverage advancements in data science, new sequencing technologies, novel research techniques, improvements in sample throughput, and relationships with other research facilities in the DOE ecosystem to facilitate improved understanding of the biomolecular mechanisms of carbon and other nutrient cycling under changing environmental regimes. By working with users from the academic, government, and commercial sectors, we will help develop a foundational understanding of plant and microbial metabolic pathways, the interactions between plants, microbes, and the environment, and how those interactions may vary in response to changing environmental conditions. This understanding will allow the research community to work toward solutions to pressing global energy, environmental, and resource challenges in support of a growing bio-based economy.

Opportunities

Advances in sequencing, metabolomics, and experimental systems have enabled researchers to understand the intricate relationships between plants, microbes, and the environment that underpin complex nutrient cycles at a molecular level. As a pioneer in this area, the JGI is poised to take advantage of emerging technologies and enable its users to increase our understanding of the diverse organisms and compounds mediating environmental nutrient and carbon cycles and their linkages to specific metabolic pathways. These urgently needed insights into the biological and environmental factors governing the stability of specific nutrient pools within ecosystems will inform the modeling and implementation of effective nutrient sequestration approaches. Research in this area will also lead to the identification and characterization of primary and secondary metabolic pathways that can be harnessed for carbon sequestration and bioproduct discovery. These data will help the emerging bioeconomy optimize plant-microbe-environment interactions to improve soil health, resilience, and plant productivity in an environmentally sustainable way.

Strategic Objective 1: Illuminating the Role of Genomic Diversity in Nutrient Cycling

HTP sequencing transformed our collective ability to explore and describe diversity across ecosystems. A comprehensive understanding of nutrient cycling requires systematically and robustly connecting these vast (meta)genomic datasets to metabolic and ecosystem processes. We plan to develop new multi-omics and analytics frameworks to address several major knowledge gaps, including limitations in current genome annotation databases and tools, and enable further characterization of inter-organismal interactions and their impact on energy and nutrient flow (**Fig. 4**).

By leveraging these resources and capabilities, our users develop the foundational knowledge necessary to improve models of nutrient cycling and sequestration in biomass and soils, discover more efficient nutrient management practices, and pave the way for a bioeconomy that reduces anthropogenic environmental impacts.

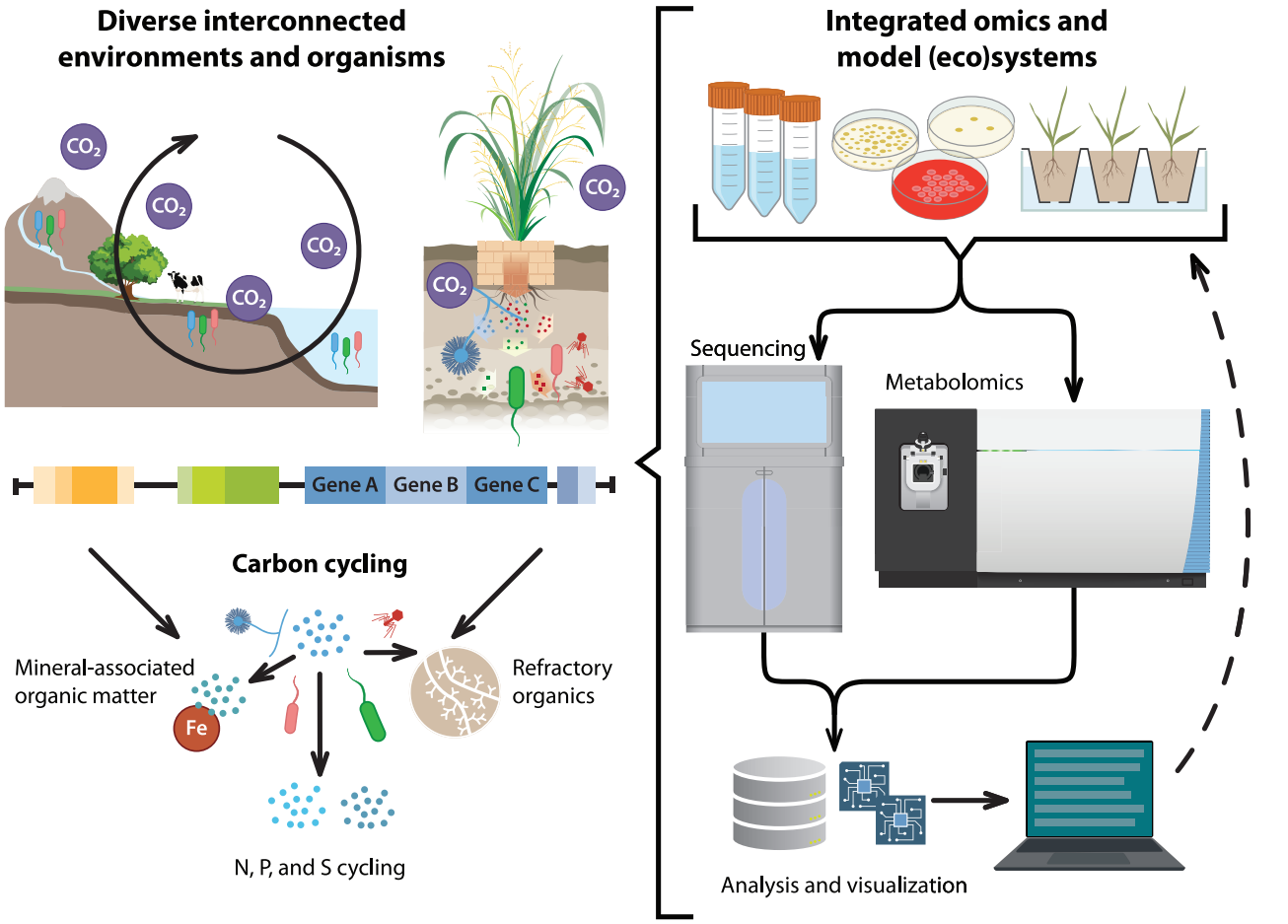


Fig. 4. Nutrient cycling across scales and ecosystems. Fully understanding nutrient cycling requires designing experimental systems at scales ranging from model organisms to ecosystems. Iterative multi-omic strategies applied to these experimental systems allow JGI users to design the experiments necessary for developing the sustainable environmental and agricultural solutions to maximize the potential benefits of the emerging bioeconomy.

Develop a Genomics-Driven Framework for Analyzing Environmental Nutrient Cycling

As (meta)genomic datasets are being generated at an increasing pace and scale, connecting these data to biogeochemical cycling remains a challenge due to limited annotation for metabolic pathways for which marker genes are lacking, and limited tools available to robustly integrate gene- and genome-level data into pathway-centric analyses. The JGI will build on ongoing efforts toward community-curated annotation sets by combining the scientific expertise of JGI staff, large-scale benchmarking of existing tools, and additional curation efforts in partnership with our users, along with institutional partners like EMSL, to improve annotation for key microbial metabolisms and pathways across taxa and ecosystems (**MG2-2**, **MG2-5**, **FA5-2**). These efforts will be first directed toward curated nutrient cycling functional databases across diverse ecosystems and spanning all domains of life. Automated JGI tools, such as the Integrated Microbial Genomes and Microbiomes IMG/M, MycoCosm and PhycoCosm, and Phytozome annotation pipelines, can leverage these resources and contribute to ecosystem modeling efforts (**PI1-5**, **PI4-2**, **FA4-2**).

The JGI will enable the use of paired metagenomes and metatranscriptomes for sample sets with rich metadata to better identify connections between taxa, key metabolic gene clusters, their primary and secondary metabolisms, and environmental factors. Specifically, additional insights on environmental (micro)organisms will be gained, and novel functions will be discovered by developing a robust computational framework to conduct differential abundance testing or correlation

analysis on these increasingly common datasets (**PI2-2**, **MG1-2**, **MG5-2**, **MG5-5**, **FA1-2**, **FA4-2**). In addition to improving community-curated annotation databases, these analytical frameworks will help translate the vast sequencing datasets generated by JGI users into usable catalogs of interconnected genomes and functional traits. Genes and gene clusters of interest can be synthesized and their functions experimentally validated/characterized by users leveraging JGI DNA synthesis capabilities (**GT8-2**, **MC4-2**, **SM4-2**, **SS3-2**). These efforts will allow JGI researchers, our users, and the broader research community to link specific plant, fungal, algal, bacterial, archaeal, and viral lineages to the metabolic cycling of carbon and other nutrients under different and changing environmental regimes.

Link Nutrient Cycles to Genes and Metabolites

As JGI users' research questions evolve, so must our sample processing pipelines. While the unique single-cell and stable-isotope probing (SIP) metagenomic sequencing pipelines will continue to provide valuable data for our users (**GT1**, **GT2**, **MC1**, **MG3-2**, **MC5**, **FA1-2**), moving forward the JGI will aim to develop additional activity-based, functionally targeted sequencing assays (**GT5**, **MC4**) and improve the integration of sequencing and metabolomics data (genometabolic analysis; **ML2**, **MG3-5**, **FA5-2**).

Techniques for monitoring low-diversity microbial communities at the single-cell level in a spatiotemporally resolved manner provide a critical foundation for strain-level classification and prediction of genomic features relevant to niche selection, nutrient preferences, and sequestration. Current functional assays available to JGI users enable the identification of metabolically active microbes via the uptake of non-canonical amino acids or isotopically labeled substrates. Sample barcoding techniques currently under development, in conjunction with additional protocol and analytical optimizations, will increase sample throughput, reduce costs, and allow for more complex and nuanced experimental designs (**GT6**, **MG3-2**). To offer more ways of targeting organisms of interest, existing microdroplet-based single-cell sorting technologies will be coupled with fluorescent sensors (e.g., biosensors) and probes (e.g., small molecules), allowing us to better identify microbes with functional activities our users are interested in (**MC4**, **SS1**, and **SS2**). Finally, we will pursue opportunities combining these independent techniques. For instance, targeted functional assays can be coupled with strain domestication protocols to insert or modify specific genes and study the resulting effects on specific environmental isolates and combinations of isolates demonstrated to be active under specific environmental conditions. Through these efforts, we envision improved coupling of sequence-based characterization of environmental samples with downstream engineering of community members (**MC4**, **SS1**, and **SS2**).

In parallel to activity assays, integrating metabolomics approaches with sequencing data using the JGI metabolomics compute analysis infrastructure has the potential to better connect key metabolites to genes and pathways associated with their catabolism and biosynthesis, including genes with limited or no functional annotation (**FA3-2**, **FA5-2**, **ML4**, **MG3-5**). The JGI will place specific emphasis on user projects designed to resolve, at the molecular level, metabolites likely to represent microbial growth substrates, root exudates, and microbial secreted products. Results will be integrated with multi-omics data from samples collected from distinct ecological niches and at distinct seasonal time points across multiple ecosystems of interest, including soil, rhizosphere, and microbial biofilms and biocrusts. Comparative analysis of these data within and between user studies will allow us to establish new connections between key (novel) metabolic genes and nutrient cycles across diverse taxa, functionally link microbial community structure to environmental chemistry, and deepen understanding of how specific evolutionary pressures shape microbial traits across environments (**MC4**). These projects will provide valuable insights into the interplay between microbial genetics and metabolism in a variety of ecological niches and provide our users and the broader research community with a deeper mechanistic understanding of these systems.

Identify and Integrate Key Microbial Interactions into Metabolic Models

Beyond the metabolic capability of each individual organism, inter-organismal interactions are critical drivers of nutrient cycling in all ecosystems. Spatial and short- and long-term temporal sampling series offer an unprecedented opportunity to characterize the impact of microbe-microbe and virus-microbe interactions by evaluating co-occurrence and co-activity of predicted partners at a community scale. To maximize the insights

users can gain from JGI-produced data, we will explore and develop innovative computational approaches to infer and characterize microbe-microbe interactions from multi-omics data, including interactions between hosts and viruses, plasmids, and other mobile genetic elements (MGEs; **MG1-2**, **PI2-2**). These new tools and the resulting data will eventually be integrated and shared with users via the IMG/M platform (**PI3-5**). The approaches will also be applied to fungal-bacterial interactions, especially early diverging fungi and *Burkholderia*-related endobacteria, as these endosymbionts impact reproduction, lifestyle, and lipid production (**FA4-2**).

In parallel, we will evaluate experimental approaches to infer microbe-microbe and virus-microbe interactions in vitro, such as the use of proximity ligation or co-localization via digital droplet polymerase chain reaction (PCR). These methods can be useful to validate and refine predictions of interactions based on sequencing data, and we envision targeting applications of these approaches to key ecosystems and taxa previously identified from multi-omics analyses (**MG1-5**). Approaches proven amenable to HTP application for some or all JGI user samples will be offered as a user capability.

New model systems enabling controlled studies of inter-organismal interactions with protists, algae, and fungi to better understand their role in nutrient cycling are also urgently needed. The JGI will develop new tools and methods aimed at understanding long-term inter-organismal interactions involving giant viruses, symbionts, and organelles (**FA4-2**, **PI3-2**). The JGI will also specifically encourage user projects focused on the gene flow between symbionts or viruses and their respective hosts, and the expression of complementary or stress-related metabolic genes encoded by a symbiont or virus to enhance or support host physiology under adverse conditions. For these, users will be encouraged to leverage JGI capability in single-cell sequencing (**MC5**) as well as metagenomic characterization of laboratory incubations, such as microcosm experiments. The data generated from these newly established model systems and in-depth characterization of the symbiont and viral impacts on host nutrient cycling will be integrated into KBase to generate high-quality, genome-scale metabolic reconstructions of these organisms, thereby improving flux-based models of microeukaryotes used to predict growth and behavior under varying nutrient regimes.

Characterize How Plant-Microbe Interactions Impact Nutrient Cycling and Plant Productivity

JGI users study interactions between plants and microbes (including bacteria, viruses, fungi, protists, and archaea) that can be beneficial, neutral, or detrimental to plant growth and productivity. Additionally, they study interactions between plants and the microbial communities living on, around, and in them (the microbiome), resulting in a range of outcomes more complex than the sum of the interactions between plants and individual microbial strains alone. An increasing global population necessitates dramatic increases in agricultural yields, for not just food or feed, but also biomass-derived products, such as biofuels and biomaterials. User-designed studies of the mechanisms underlying plant-microbe/microbiome interactions are necessary to sustainably increase plant productivity and resilience. While the complexity of the soil microbiome and related environments typically preclude controlled experiments in the field, the JGI will employ a battery of emerging techniques and model systems to study microbiomes in controlled systems (**PL5**). EcoFABs, EcoPODs, EcoBoxes, and EcoBOTs (**Fig. 5**) allow users complete control of their experimental environment down to microbial composition. Additionally, single-cell and spatial transcriptomics will provide the resolution necessary to study interactions at a cellular level (**PL5**, **GT3-5**, **GT4-5**). This is particularly important for plant-microbe interactions, where only a small subset of plant cells directly interact with microbes, as these methods enable detection of changes in expression lost when entire plants are homogenized prior to molecular analysis.

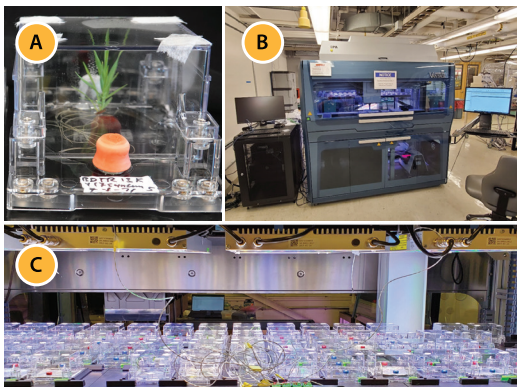


Fig. 5. State-of-the-art platforms developed at Berkeley Lab for studying rhizosphere communities. (A) EcoFAB is a model laboratory ecosystem enabling reproducible and sterile investigation of soil-plant-microbe systems. (B, C) EcoBOT is a custom-built robotic handling system allowing autonomous HTP EcoFAB growth, sampling, and imaging. Photo credit: Albina Khasanova

Plants influence the composition and function of the microbiome by releasing molecules that promote, prevent, or otherwise influence microbial growth and function by various mechanisms. Thus, many plant genes influence microbiome function and composition. Identifying and characterizing these genes provides a foundation for genetic engineering and chemical interventions to modify the microbiome and increase plant yield and resilience. Extensive sequencing of plant populations, including the creation of high-quality pangenomes (see **Functional Diversity**), provides our users with an unprecedented opportunity to use genome-wide association studies (GWAS) and similar genetic approaches to identify candidate genes. These efforts are complemented by measurements of small molecules released or consumed by plants and microbes and by emerging technologies like single-cell and spatial transcriptomics. Considering these multi-omic datasets phenotypes when looking for genetic associations will allow us to help users identify candidate genes involved in plant-microbe interactions, as well as genes involved in all aspects of plant metabolism (**PL4**). This approach requires highly controlled systems. The model grass *Brachypodium distachyon*, grown in EcoFAB devices, allows for studies of this type (**Fig. 5**). Functional characterization of candidate genes will benefit from the tools described in **Functional Diversity**, including the sequencing of plant mutants.

To fully understand plant-microbe interactions, we must also understand the impacts of specific microbes, microbial communities, metabolites, and microbial genes on plant growth, productivity, and microbiome function. We will enable and encourage user studies designed to understand the eco-evolutionary drivers of plant-microbe interactions and their impact on plant growth. This will include investigations based on the different types of plant-fungal interactions (e.g., growth promotion, mycorrhizae formation, endophytic colonization, and pathogenesis; **PL5**). We will also help our users identify and characterize plant growth-promoting microbes and synthetic microbial communities, especially those mediated through small molecule metabolites (**ML3**, **PL5**). To enable research on nitrogen fixation by *Bradyrhizobium*, a high-quality pangenome will be constructed as part of an ongoing Community Science Program (CSP) project, allowing the correlation between natural genome variation and phenotypes relevant to nitrogen fixation.

Strategic Objective 2: Understanding Biological Drivers of Carbon Capture and Sequestration

Soil carbon is critical for soil health because it affects nutrient and water retention, as well as microbial diversity. Soil is also a major global carbon sink. However, an estimated 50% of soil carbon has been lost due to extractive land-use practices. This makes soil carbon restoration an important goal. Plants naturally sequester a large fraction of their photosynthates into soils as rhizodeposits. Unfortunately, some 95% of these inputs are respired as CO_2 by soil microbial communities.[13,14] Thus, altering carbon cycling through plant breeding and engineering along with interventions to optimize microbial communities can increase soil carbon. Generating and harnessing genomic and experimental knowledge is necessary to develop new plant lines and associated microbiomes that increase the amount of carbon that persists in soils. This will help achieve the major potential benefits of the bioeconomy and enable, for example, bioenergy crops to be used to store carbon while restoring the fertility of soils.

Identify the Factors Governing Carbon Persistence in Soil

Soil carbon is beneficial for soil health and serves as a medium- to long-term carbon sink. While persistence in soil is key for both advantages, we neither fully understand which carbon-containing molecules persist in the soil nor how environmental factors influence the half-life of these molecules in soil. Optimizing soil systems and plants to increase soil carbon content requires an understanding of biological recalcitrance, mineral associations, aggregate entombment, root exudate priming, coupling of high diversity and low abundance, and oxidative exoenzyme presence. We are integrating metabolomic technologies with sequencing, cheminformatics, and bioinformatics to develop and support testing of user-driven hypotheses addressing factors governing soil carbon persistence to rapidly advance understanding of the chemical, biological, and environmental controls on soil carbon cycling. This will require the development of improved liquid chromatography tandem mass spectrometry (LC-MS/

13 Jansson, C., S. D. Wullschleger, U. C. Kalluri, and G. A. Tuskan, 2010, "Phytosequestration: Carbon Biosequestration by Plants and the Prospects of Genetic Engineering," *Bioscience*, 60, 685–696.
14 Sanderman, J., T. Hengl, and G. J. Fiske, 2017, "Soil Carbon Debt of 12,000 Years of Human Land Use," *Proc Natl Acad Sci USA*, 114, 9575–9580.

MS) and cheminformatic methods for metabolite identification (**ML1**, **ML5**); stable-isotope-enabled approaches for quantifying the persistence of specific compounds (**ML3-5**), and bioinformatic tools for linking to sequencing data and models (**ML2**, **MG3-5**). These experiments will be enabled by the development of automated fabricated ecosystem capabilities optimized for the investigation of molecular mechanisms and environmental factors affecting the persistence of soil organics. This will include development of sterile habitats (building on current EcoFAB devices) and synthetic microbial communities that capture relevant aspects of natural communities and processes in controlled laboratory conditions to enable testing using diverse plant lines, microbial strains, and genetic engineering tools (**ML3**, **PL5**).

Develop Model Organisms and Consortia to Maximize Carbon Capture and Sequestration

The JGI will support user projects to identify and use model organisms and consortia to explore and enhance the effectiveness of carbon capture and sequestration. The potential of both microeukaryotes and plants for carbon sequestration will be assessed by metrics including biomineralization, biomass accumulation, and the accumulation of lipids and other highly reduced carbon molecules (e.g., **FA2-5**). Additionally, pangenome analysis and population genomic studies will allow us to better understand the genes and pathways associated with carbon capture potential (**FA5-5**, **PL1-5**, **PL2-5**, **MG4-5**).

We will also study rumen microorganisms, such as those from the Hungate1000 project.[15] This project was initiated as a JGI community science proposal to generate quality reference genomes of bacteria and methanogenic archaea isolated from a broad range of ruminant animal hosts, with an overall goal to help curb methane emissions from these systems. A recently renewed collaboration with users will allow us to complete the project's next phase by sequencing new bacterial and archaeal isolates in coordination with the Global Methane Hub[16] (**MC3-5**). High-quality reference genomes are an important resource for investigators exploring ways to inhibit methanogens (e.g., vaccines, rumen modifiers). Comparative genomics and annotation will be employed to explore new strategies for modification of the rumen microbiome for reduced emissions (e.g., methanogen viruses, synthetic communities). By combining these approaches, we will better understand the carbon sequestration mechanisms in model organisms and gain insights that can be translated to non-model systems.

Identify Pathways for Biosynthesis and Degradation of Recalcitrant Carbon Molecules

Recalcitrant carbon is the subset of soil organic matter that is resistant or protected from microbial degradation and may therefore serve as long-term carbon storage. Recalcitrance can, for example, result from the association of organic compounds with mineral soil surfaces. In collaboration with the user community, the JGI will conduct large-scale metabolomics projects and generate datasets providing molecular-level insights into soil organics and the pathways that microorganisms use to consume persistent soil organic compounds. By linking soil metabolites to genomic potential (**ML1-2**, **ML2-5**, **MG3-5**), the JGI will support an integrated genometabolic mechanistic understanding of the interplay between plants, microbes, and the stability of soil carbon pools (**ML2**, **ML3-5**). Additionally, the JGI MycoCosm database of wood-decayer genomes will be used and expanded to explore metabolic pathways involved in converting recalcitrant carbon, such as lignin, into biomass and other biotechnologically relevant compounds. This knowledge is crucial for developing sustainable biofuels and bioproducts. Understanding the enzymes responsible for the deconstruction of polyphenolic compounds in anaerobic ecosystems is of particular interest because these types of compounds are known to be a major component of recalcitrant carbon (**ML4**).

Climate-driven shifts in plant and algal species are another significant factor in global carbon cycling and climate regulation. High-quality plant and algal reference genomes, along with comprehensive analysis products, will enable JGI users and the broader research community to study biodiversity and carbon sequestration across plant and algal species (**FA1-5**, **FA2-5**, **PL2-2**). This may include the genomic characterization of seaweeds, which is currently limited to a few species only and is expected to further our understanding of the biosynthetic pathways involved in carbon sequestration (**FA2-5**, **MG4-2**). This genomic information will facilitate the identification of critical pathways and mechanisms related to carbon capture

15 Seshadri, R. et al., 2018, "Cultivation and Sequencing of Rumen Microbiome Members from the Hungate1000 Collection," *Nat. Biotechnol*, 36, 359–367.
16 **https://www.globalmethanehub.org**

and storage in algal and plant species, thus advancing our ability to mitigate climate change impacts.

Predict Ecological Drivers of Carbon Sequestration at the Ecosystem Scale

Biological CO_2 assimilation by agricultural and ecological systems is one mechanism for humanity to reduce atmospheric carbon. Working with our user groups, the JGI will coordinate the discovery and prediction of genetic mechanisms that may improve carbon capture by plant, fungal, and microbial communities. To accomplish this we will (1) leverage natural diversity among DOE-relevant plants to understand how carbon capture and sequestration can be manipulated, including the exploitation of interactions with the microbiome and environmental factors (**PL2-5**); (2) identify persistent plant inputs and their genetic controls to inform development of carbon sequestration crops through manipulation (e.g., EcoFAB-like systems, EcoFAB3.0, and field experimentation with collaborators; **ML3**, **PL2-5**) and observation of plant-environment interactions; and (3) develop advanced approaches to track carbon flux through systems and fully characterize microbiome activities to more effectively link microbiome activity to carbon sequestration potential (**PL5-5**). We will accomplish these objectives by gathering existing and generating additional data on environmental and biological carbon cycling systems to understand the patterns and process of carbon flow through ecosystems (**ML2-5**). These data will allow us to increase carbon sequestration in the environment using plant genetics, microbes, environmental manipulations, and altered agricultural practices.

Explore the Role of Viral Lysis of Microbes and Fungi in Carbon Cycling

Viral infections and subsequent cell death substantially impact carbon cycling in many ecosystems. In the ocean, 25% of the carbon is estimated to flow through a "viral shunt" and is redirected to the pool of dissolved organic carbon instead of being transferred to higher trophic levels. Recent studies suggest that a similar phage-driven carbon shunt may also occur in soils, especially through the induction of prophages. The JGI and its users can address two critical knowledge gaps to improve our collective understanding of the viral impact on carbon cycling. First, there is a need to characterize the impact of phage lysis of bacterial cells on the available pool of carbon in different soils. We will explore how combining multi-omics analyses and microscopy observations from standardized incubation devices, including EcoFAB and EcoBOT (**Fig. 5**), may enable users to better characterize compounds released following phage infection, and understand how other community-associated microbes leverage this pool of carbon (**ML3**, **MG1-2**, **MG1-5**). Taken together with the multi-omics framework proposed to characterize phage-infected virocells (cells undergoing an active viral infection, see **Functional Diversity, From In Silico Predictions to Functions and Phenotypes**), these studies will help establish a robust framework to interpret viral signals in metagenomics and metatranscriptomics data (see **Nutrient Cycling**).

The JGI will work with users to progressively expand these studies to include microeukaryotic organisms (fungi and microalgae), some of which play a significant role in carbon cycling. Specifically, we will focus on linking the recently discovered diversity of DNA and RNA viruses in soil to their microeukaryotic hosts. RNA viruses have been predominantly associated with animal and plant hosts; however, these organisms represent a minor fraction of the eukaryotic tree of life. Given the vast unexplored eukaryotic diversity, particularly in microeukaryotes, it is plausible to posit that many of the newly discovered viruses may inhabit these understudied hosts. A comprehensive exploration of microeukaryotes as potential hosts will significantly contribute to our understanding of the RNA viral landscape and may reveal novel unexpected features of RNA viruses that directly link to nutrient cycles, such as viral capability for metabolic host reprogramming (**FA4-2**, **MC5-5**). DNA viruses of microeukaryotes, in particular giant viruses, frequently encode genes that may enable metabolic reprogramming of microeukaryotic hosts. Currently, there is a paucity of experimental evidence to substantiate these interactions. Employing the aforementioned methodologies becomes crucial to shed light on the role that giant viruses-induced virocells play within host populations and nutrient cycles.

Strategic Objective 3: Carbon Utilization for Bioproducts

Creating conventional petroleum-derived products requires substantial energy input and causes substantial emissions. The carbon footprint of plastics alone, 2.2 billion tons of CO_2 equivalent in 2015, shows the need for carbon-neutral replacements. Therefore, understanding how plants and microorganisms process carbon and CO_2 and convert them into bioproducts is essential for the emerging bioeconomy and a DOE priority.
The JGI can pave the way by supporting users in the identification of pathways for new potential bioproducts in emerging model systems, and enable the understanding of metabolic pathways to explore novel bioproducts with diverse applications and renewable materials. Functionally characterizing secondary metabolite pathways, including their regulation, across prokaryotes and eukaryotes at scale is crucial to harnessing the potential of these compounds for various applications. Finally, leveraging large-scale metagenomic data produced by the JGI for enzyme discovery and optimization is a promising avenue to discover novel bioproducts. This objective will combine the exploration of new analysis and visualization strategies, including the potential of large language models (LLMs) to embed pathway-level information.

Identify Novel Bioproduct Candidate Pathways in Emerging Model Experimental Systems

The ability to manipulate microbial biosynthetic pathways and metabolism using synthetic biology provides unprecedented opportunities to address a wide range of topics related to the DOE mission in sustainable bioenergy development. An integrated approach combining various research areas can significantly advance our understanding of carbon utilization pathways and biotechnological applications of emerging model systems. One approach involves lipid metabolism, which encompasses molecules of the cellular metabolome that have high energy density and hold great potential for bioproduct development beyond just being used as fuels (**FA2-5**). Therefore, to meet user needs we will continue to improve the performance of our nonpolar metabolomics products for scalable analysis of lipids (**GT12**). This will enable broad exploration of metabolic pathways to unlock novel applications for lipids in biotechnology.

We will also help users explore the degradation of cellulosic substrates by heterotrophic protists. While lignocellulose-degrading protists have been observed within consortia found in the guts of wood-consuming insects like cockroaches and termites, their complex nature limits their direct applicability in industrial settings. Therefore, identifying free-living protists capable of degrading organic plant materials, such as litter, holds promise for industrial utilization and advancements in processing organic plant material (**MC5**, **MG4-5**). This could lead to innovative approaches and drive progress in industrial applications. Additionally, algal and other protistan bioproducts, such as those derived from *Haematococcus*, *Nannochloropsis*, and *Schizochytrium*, offer potential in terms of production and various applications (**FA2-5**). Exploring these and other protist species and their metabolic pathways can pave the way for developing sustainable and economically viable bioproducts. Furthermore, the discovery of pathways for silica biomineralization in diatoms, achieved through comparative genomics and transcriptomics analyses of related groups, provides insights into the mechanisms underlying the synthesis of biomineral structures (see **Biomolecular Materials**).

Functionally Characterize Secondary Metabolite Pathways at Scale

In addition to the search for bioproducts, our goal is to functionally characterize secondary metabolite pathways across prokaryotes and eukaryotes at a large scale to improve our understanding of the molecules produced and the roles that these metabolites play in environments. Secondary metabolites are molecules that provide advantages to the organisms that produce them, such as nutrient acquisition, defense mechanisms, communication, and toxin resistance. Genes responsible for their production are often clustered in genomes as biosynthetic gene clusters (BGCs). Only a small percentage of BGCs are expressed under standard laboratory conditions, leaving a large amount of untapped and uncharacterized biochemical diversity. The current JGI expertise and capabilities to explore this diversity put it in a uniquely favorable position to dramatically expand the repertoire of secondary metabolites in unexplored lineages. The development of improved computational and experimental techniques will enable the identification of BGCs and gene cluster families likely to encode novel biochemical reactions or understudied subclasses of secondary metabolites

(**ML1**, **SM1-2**, **SM2**, **FA5-2**). We will continue to develop the Secondary Metabolism Collaboratory (SMC) as a key resource for the community through expanding the database across kingdoms, bringing in and developing new prediction tools and new ways to standardize the description of BGCs (**SM1**, **SM2**, **SM3**).

An additional approach will be to continue to scale our abilities to synthesize, clone, and express BGCs in either heterologous hosts or cell-free systems, facilitating the identification of their products, with a view to being able to offer an on-demand service to users with a high rate of success (**SM4**). These discoveries will be added to databases such as Minimum Information about a Biosynthetic Gene cluster (MIBiG)[17] to improve the accuracy and quality of future predictions. Another area to be explored is a broad characterization of unusual and overlooked secondary metabolite pathways, exploring clusters lacking canonical core biosynthetic genes. To enable accurate predictions of BGCs, we plan to design and develop algorithms for predicting BGCs that other tools currently do not support (**SM2-5**). These new algorithms will be validated through the heterologous expression of relevant BGCs, and we plan to continually scale our experimental platforms working with users to facilitate the expression of BGCs and the characterization of their products (**ML3**, **ML5**, **SM4**).

Furthermore, we will adopt existing and develop new tools to study the characterization of secondary metabolites and their synthesis. We will extend nascent efforts to dissect the complex regulatory networks controlling secondary metabolite biosynthesis by employing DNA affinity purification sequencing (DAP-seq), multi-DAP-seq, and regulon identification by in vitro transcription-sequencing (RIViT-seq) across model and non-model secondary metabolite producers (**SM5**). These technologies will also be applied to characterization of understudied families of transcription factors (TFs) predicted to be involved in secondary metabolism (**SM5**). We will also explore the use of various imaging techniques for structural studies of these metabolites through collaborations with facilities including EMSL, the Biological and Environmental Program Integration Center (BioEPIC), and the National Center for Electron Microscopy (NCEM), leveraging their capabilities, such as microcrystal electron diffraction (MicroED), to enhance our research efforts (**SM6**).

Leverage Large-Scale Metagenomic Data for Enzyme Discovery and Optimization

Beyond the secondary metabolites and novel bioproducts mentioned previously, the vast amount of (meta)genomic data available in JGI systems represents a treasure trove of novel enzymes and pathways of potentially major bioeconomic value. One challenge, however, is enabling nonexpert users to systematically mine these data for a specific enzyme or pathway or interest. Given the diversity of JGI users and the integrated work across JGI programs including genome science, metagenome science, metabolome science, secondary metabolites, and synthetic biology, we believe the JGI is ideally positioned to design, develop, and offer this type of framework to the community.

First, and connected to overall database improvements (see **Illuminating the Role of Genomic Diversity in Nutrient Cycling**), we intend to build capability for JGI users to search for specific enzymes or combinations of enzymes of interest. Users would then be assisted in the processing of the search results, including informed selection of representative sequences across the existing diversity based on, for example, phylogenetic diversity, changes in key residues (if known), or changes in predicted structure (**PI4-2**). This will require new analysis and visualization features, along with new workflows available to IMG/M users. In parallel, we will also leverage these data to explore the potential of LLMs to embed pathway-level information. LLMs were recently shown to be able to accurately map genes to a "functional space" where genes that are functionally related are close together, enabling systematic identification of functional modules and pathways. An LLM properly trained on the vast and diverse IMG/M metagenome data could, in principle, improve users' ability to systematically search for metabolic pathways, including incomplete or novel ones, and potentially build pathways de novo following constraints provided by the user (**PI4-5**).

17 https://mibig.secondarymetabolites.org

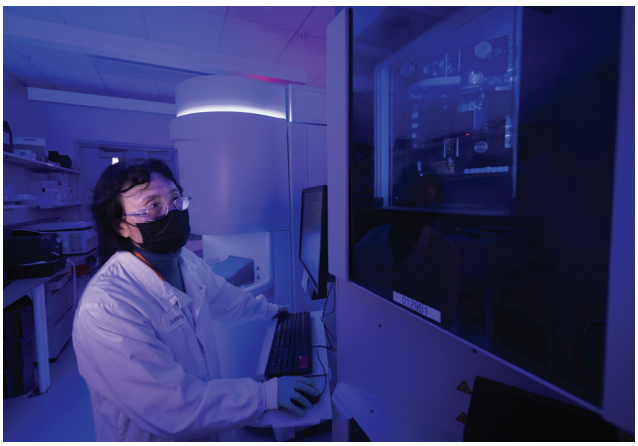
Strategic Theme 2:

Functional Diversity

Characterizing Functional Diversity across the Domains of Life

Background

Biology has evolved over billions of years across Earth's diverse biomes, offering a vast array of metabolic capabilities and adaptive traits. This functional diversity encompasses the biological mechanisms, structures, and processes encoded in the genomes of organisms, enabling them to interact with their environment and occupy specific ecological niches through biochemical functions. The JGI, as an SC-supported user facility within Berkeley Lab's Biosciences Area, recognizes the significance of functional diversity, and aims to advance foundational science that can translate into data-driven and experimentally validated discoveries by empowering users worldwide to study functional diversity. The JGI does this through providing access to cutting-edge sequencing, DNA synthesis and metabolomics technologies, data science pipelines, and computational infrastructure to characterize functions encoded in genetic sequences across Earth's environments. These collective efforts will continue to facilitate progress in understanding biodiversity to address energy and environmental challenges and help sustain the bioeconomy.



Opportunities

Founded shortly after completion of the Human Genome Project, the JGI has established itself as a world leader in genomics, gaining worldwide recognition for producing high-quality reference genomes for plants, fungi, and microbes. This long history of leadership has allowed us to forge relationships with a massive and diverse scientific community, build platforms to support innovative science, and push the boundary of what is possible within the field of environmental genomics. This history also provides us with an excellent opportunity to grow our capabilities for exploring life's functional diversity and maximizing the impact of this growth through sharing new capabilities with our user community. Toward this end, we continue to pioneer the development of HTP synthetic biology techniques for autonomous protein, pathway, and strain engineering at scale, approaches to study transcription regulation to advance functional insights, and fabricated ecosystems to support controlled plant-microbiome research. Furthermore, the JGI has for years been cultivating a talented computational workforce and powerful computational infrastructure to establish the resources and the expertise to bring bioinformatics squarely into the new AI-era. Within this theme, we describe our plans for how we will enable our users to (1) advance sequence-based discovery and gene function prediction, (2) take those predictions and translate them into robust, functionally characterized proteins and phenotypes, and (3) begin to leverage those functional insights to enable biosystems design. The findings and resources derived from omics and functional experimentation will serve as a crucial cornerstone for the bioeconomy. The resulting insights will support the enhancement of production strains and crops, inform the development of processes to harness these strains and crops for energy and environmental applications, and facilitate the discovery of novel bioproducts.

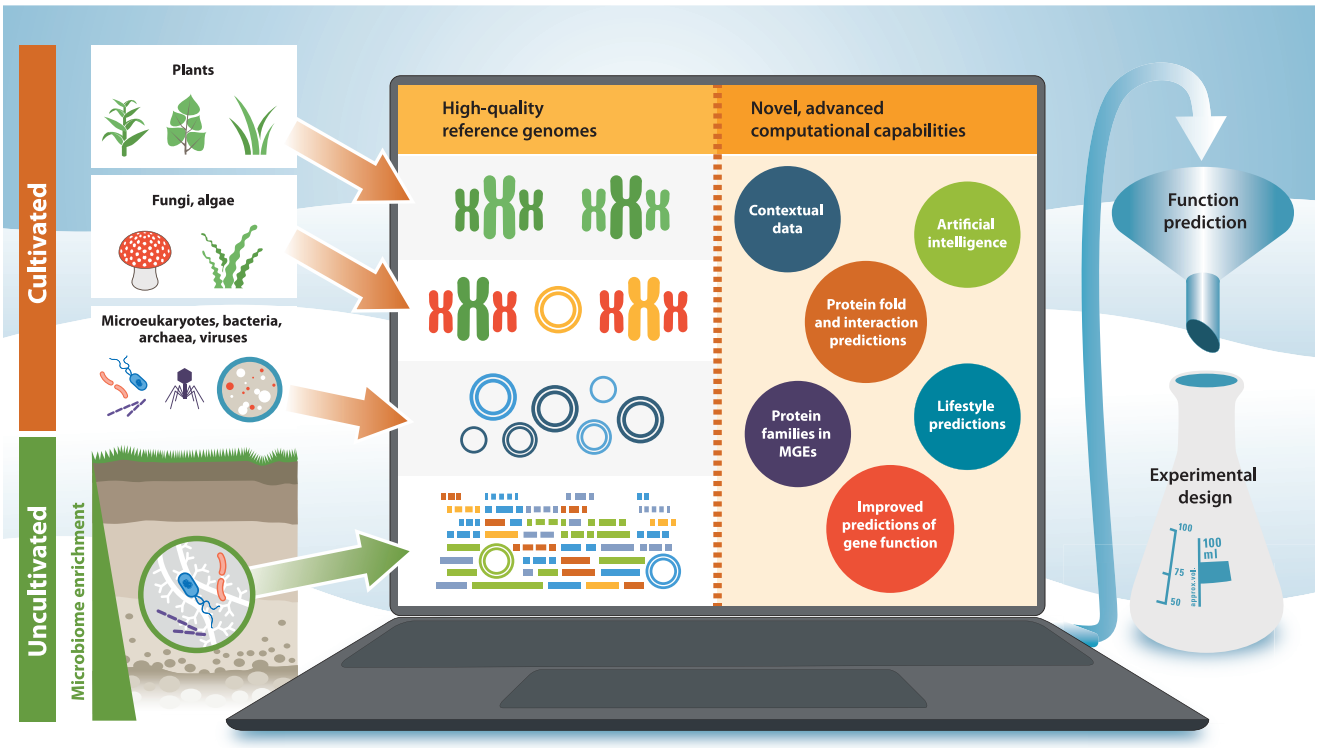


Fig. 6. Integration of experimental and computational capabilities. Integrating quality genomic data with advanced computational capabilities for sequence-based discovery and prediction.

Strategic Objective 1: Sequence-Based Discovery and Prediction

Over the past decade, the generation and availability of genomic sequence data have increased exponentially. The resulting information has revolutionized the fields of genetics and genomics and enabled unprecedented scientific breakthroughs. With continually increasing throughput and reduced costs of sequencing technologies, the JGI is destined to remain on its path of high-quality sequence data generation for DOE-relevant systems studied by our users. By developing and scaling advanced, new computational analyses, the JGI together with its user community will focus on function prediction to facilitate experimental design and provide a robust foundation for the bioeconomy.

Build Quality Reference Genomes as a Foundation for Exploring Functional Diversity

High-quality genomes are the bedrock on which complex multi-omics data types are anchored and from which gene function predictions are generated to understand functional diversity. The JGI's long-standing reputation for constructing high-quality reference genomes is a result of combining robust, scalable wet-lab workflows and state-of-the-art computational technologies for sequence generation. The high quality of JGI-generated genomes is in large part responsible for attracting and retaining a vast user community. Moving forward, we will continue along this path but expand in several crucial areas. Specifically, we will leverage advanced sequencing platforms to construct more and even higher-quality genomes for plants, microeukaryotes, bacteria, and archaea for our users at scale (**GT1, GT2**).

PLANT REFERENCE GENOMES AND PANGENOMES

Plant genomes are characterized by highly repetitive sequences, variable ploidy, recurrent hybridization, and inbreeding. While these factors make assembly of large plant genomes particularly difficult, recent advances in single-molecule, long-read sequencing (e.g., PacBio HiFi) and bioinformatic pipelines have enabled the JGI to accurately assemble increasingly complex plant genomes. For example, in collaboration with our users we have resolved individual haplotypes for the DOE

bioproduct feedstocks sugarcane, switchgrass, poplar, and miscanthus. Analysis of these genomes has revealed substantial large-scale sequence inversions, deletions, and duplications. These variants cannot be genotyped through traditional short-read-based methods, demonstrating that a single reference genome is not sufficient to fully support genome-informed breeding efforts. To overcome this challenge, we must build more reference genomes, both within and across species, and develop more nuanced data analysis and exploration methods.

The goal of improving the number and quality of reference genomes for DOE feedstocks has historically been limited by both the quality of sequence data sources and the massive personnel effort involved in computationally assembling a genome. We envision significant improvements in both areas will soon occur. We currently have a higher throughput and accuracy long-read sequencing platform and are developing more automated genome assembly pipelines. In the short term, we will continue to make large eukaryotic genome production more cost effective and incorporate new technologies into our pipelines as they become available (**PL1-2**). These efforts will let us (1) update genomes built on older sequencing technology, (2) sequence a much broader diversity of species, and (3) begin to build out multiple reference pangenome resources for key DOE feedstocks. As technologies mature, we envision scaling up these efforts to produce, in collaboration with the Bioenergy Research Centers (BRCs) and CSP investigators, annotated reference genomes for 50 or more genotypes of critical DOE species. This will also cover a much broader phylogenetic distribution of species, including crop relatives (**PL1-5**). Combined, these resources will complement our ongoing phylogenetic sampling efforts through the Open Green Genomes CSP project[18] and achieve a more complete view of plant gene evolution.

While this burgeoning diversity of genome resources offers an opportunity to better define variation, methods to access and analyze pangenomes are not fully mature. Our past efforts to integrate multiple reference genomes with breeding goals helped found the pangenomics field, but nuanced exploration of pangenomes remains piecemeal at best. We will continue to develop novel bioinformatics methods to analyze, visualize, and explore multiple reference genomes to enhance discovery efforts for the community (**PL2-2**). These efforts will focus in part on the integration of multiple data sources (**PL4**), especially between reference genomes, contig-level HiFi assemblies, and short-read data. Our long-term goal is to enable a multi-reference genome analysis paradigm for plants (**PL2-5**) and to use these pangenomes to achieve DOE goals related to gene discovery, function, plant engineering, and breeding. To accomplish this, we will construct new production workflows that integrate innovative pangenome methods to identify all functional variation, including large-scale rearrangements, so these variants can be linked to functional traits (**PL2-5**). Along with these workflows, we will develop new, community accessible methods for comparative genome visualizations based on our work to date with GENESPACE (**PL2-2**) and improve mechanisms for user interaction (**PL3-2**) by improving integration of pangenomic information into Phytozome (**PL3-5**, **PL6-2**). Combining this information with existing JGI-produced reference-based population genetics data (single nucleotide polymorphisms, SNPs) will allow for the development of novel, comprehensive maps that link genetic variation with trait variation across BRC and CSP efforts.

GENOME SEQUENCING TO REVEAL FUNCTIONAL DIVERSITY OF FUNGI, ALGAE, AND OTHER MICROEUKARYOTES

The 1000 Fungal Genomes Project has brought together over 300 researchers and has to date yielded more than 700 annotated fungal genomes. Furthermore, several genus-focused genome sequencing efforts have emerged, such as those for *Aspergillus* and *Trichoderma*, which provide insights into genetic variations and functional diversity. The resulting resources also set the foundation for further characterization with multi-omics exploration. These initiatives have transformed the field of mycology into a genome-based science and enabled the JGI and its user community to embrace fungal diversity and build valuable catalogs of fungal genes, enzymes, and pathways. While numerous genome projects under this initiative are still in progress, the vast unexplored fungal diversity necessitates further efforts.

The JGI and the fungal user community aim to expand their efforts by targeting 10,000 fungal genomes as a grand challenge. Drawing upon integrated genomics and multi-omics data (**FA5-5**) in the JGI fungal resource MycoCosm[19] and in collaboration with research labs and

18 https://phytozome-next.jgi.doe.gov/ogg
19 https://mycocosm.jgi.doe.gov

culture collections, this ambitious goal seeks to build the most comprehensive collection of fungal genomes, starting with 5,000 genomes in five years (**FA1-5**). With a focus on functional diversity, these endeavors will open new scientific avenues and applications for the future bioeconomy. Genomes from DOE-relevant groups of fungi involved in plant interactions and plant biomass degradation will aid in promoting and controlling plant growth and yield building blocks for new compounds and materials (see **Biomolecular Materials**). Additionally, in-depth exploration of mycobiomes associated with bioenergy crops and from diverse environments will lead to the discovery of new species, enabling their description alongside their genomes and contributing to more sustainable growth and biomass production (**FA1-2**, **FA4-5**). These molecular data will enhance our understanding of these species' unique biology, ecology, and adaptation. The integration of externally sequenced genomes from individual labs and public resources into MycoCosm and linking to culture collections where these strains are available will further facilitate DOE-relevant research and applications (**FA1-5**). This large, community-wide effort also provides us with renewed opportunities to expand our user community and build relationships with currently underrepresented populations. We will seek partnerships with university educators (including at minority-serving institutions [MSIs]) to explore workforce-development opportunities for students to contribute samples for genome sequencing (**FA4-5**) and train them on fungal comparative genomics techniques and platforms.

Another important group of microbial eukaryotes, algae, is even more diverse than fungi. Algae have evolved multiple times across the eukaryotic tree of life. They capture CO_2 and synthesize organic compounds at a magnitude that constitutes the majority of global primary production, occur in nearly every ecosystem on Earth, and provide a deep diversity of potentially useful platforms, pathways, and products. By studying the full genomic repertoire of algae, we can harness their capabilities for sustainable applications in biology, biotechnology, and the bioeconomy. Building upon the success of early JGI algal genome projects, the JGI established algal genomics as a focus area of research five years ago. Over 150 algal genomes have since been integrated into PhycoCosm,[20] the JGI algal genomics resource, setting the stage to develop the Algal Genome Annotation Encyclopedia (ALGAE) to provide reference genomes for all high-level taxonomic groups of algae and their non-photosynthetic protistan relatives and ecological partners (**FA2-5**). As with JGI's previous experiences with MycoCosm and Phytozome, the consolidation of the ALGAE genomes into PhycoCosm will enable collaboration between phycologists in disparate research areas, including photosynthesis, nutrient cycling, toxic blooms, symbiosis, bioenergy, and biomaterials (see **Biomolecular Materials**). Partnerships with algal culture collections around the world and the Arizona Genomics Institute will further advance the ALGAE project by acquiring samples and extracting high-quality nucleic acids for sequencing. These partnerships will facilitate the inclusion of diverse species in ALGAE, enabling a comprehensive representation of high-level taxonomic groups. In short, the ALGAE project will bring together diverse algal genomes and other -omics data into one resource, and thus help attract and bring together diverse algal researchers (**FA2-2**).

LEVERAGING BACTERIAL AND ARCHAEAL ISOLATE GENOMES TO ADVANCE FUNCTIONAL INSIGHTS

The JGI aims to enable its user community to discover and characterize genomes of novel microbial isolates, including those of bioeconomic value, spanning the entire tree of life (**PI4-2**, **MC2**, **MC3**). This objective builds upon the long-term success of the Genomic Encyclopedia of Bacteria and Archaea (GEBA),[21,22] launched approximately 15 years ago, which allowed the JGI to maintain global leadership in sequencing unique bacterial and archaeal species, particularly type species. This new effort includes sequencing yet-unexplored organisms at higher taxonomic ranks, as well as identifying new genus and species-level groups. This is well aligned with the recent OSTP report *Bold Goals for U.S. Biotechnology and Biomanufacturing*,[23] which identified the ambitious goals of genome sequencing one million microbial species and advancing their functional characterization. Top priority will be given to organisms that have remained relatively underexplored, for example those from select environments with fewer representative genomes based on a global genome

20 **https://phycocosm.jgi.doe.gov**

21 Wu, D, P. Hugenholtz, K. Mavromatis, K. R. Pukall, E. Dalin, N. N. Ivanova, et al., 2009, "A Phylogeny-Driven Genomic Encyclopaedia of Bacteria and Archaea," *Nature*, 462:1056. doi: 10.1038/nature08656.

22 Mukherjee, S. et al., 2017, "1,003 Reference Genomes of Bacterial and Archaeal Isolates Expand Coverage of the Tree of Life," *Nat. Biotechnol*, 35, 676–683.

23 March 2023, Washington, D.C., **https://www.whitehouse.gov/wp-content/uploads/2023/03/Bold-Goals-for-U.S.-Biotechnology-and-Biomanufacturing-Harnessing-Research-and-Development-To-Further-Societal-Goals-FINAL.pdf**.

census (**PI4-2**). This might also include microorganisms found in extreme environments and habitats that are generally challenging to access. Additionally, emphasis will be placed on studying organisms exhibiting environment-specific or niche-specific functional traits that hold potential value for applications in the bioeconomy. By targeting these underrepresented organisms, we aim to uncover novel biological resources and gain insights into unique adaptations and functional capabilities that can be harnessed for various biotechnological and industrial purposes. Quality reference genomes of bacterial and archaeal isolates will not only provide invaluable datasets for continued phylogenetic anchoring of metagenomic data, sequence-based function predictions, and robust expansion of proteome diversity, but will ultimately be instrumental for experimental cultivation-based functional verification and phenotypic characterization.

Tap into the Coding Potential of Life's Dark Matter through Cultivation-Independent Genomics

While generating more and higher-quality genomes for isolates is foundationally important and provides a strong genomic backbone for the tree of life, most life on Earth has not been cultured in the laboratory. Tapping into this less-explored sequence space across different taxonomic levels will enable the discovery of novel functional capabilities and more efficient metabolic pathways, and might transform how we understand host-microbe interactions and complex communities. Over the past several years, we have established robust platforms and workflows for exploring microbial dark matter via microbial metagenomic and single-cell genomic sequences and contributed to building a collaborative ecosystem for exploring such data (NMDC). Over the next five years, we will further expand these efforts in collaboration with our user community.

OPTIMIZE WORKFLOWS FOR EXTRACTING GENOMES OF EUKARYOTES AND THEIR ORGANELLES FROM COMPLEX ENVIRONMENTAL SEQUENCES

Metagenomics is a powerful tool for describing bacterial, archaeal, and viral communities. While continuing efforts to deepen the knowledge of these lineages are well under way, these studies often overlook or miss eukaryotes, which are another major component of communities. Despite their often-lower abundance, microeukaryotes frequently play outsized roles. For example, in ruminant guts, anaerobic fungi (Neocallimastigomycota) break down plant biomass and shape communities through secretion of antimicrobial compounds.[24] Consequently, it is imperative that methods are developed enabling the identification and characterization of microeukaryote genomes to produce a holistic picture of functional diversity within communities and ecosystems.

Recently, the JGI has developed a new workflow to mine metagenomic data and identify eukaryotic bins, which provides JGI and IMG/M users with a broader picture of diversity within their samples. We will annotate the highest-quality bins and make them available to users via the MycoCosm and PhycoCosm web portals (**FA1-2**). Over the next five years, the JGI aims to scale this process and take a leading role in developing new experimental and computational techniques geared toward characterizing eukaryotic diversity within complex communities (**MG4-2**, **MG4-5**). In parallel, we will continue to grow our single-cell eukaryotic sequencing capabilities (**MC5-2**), enabling our users to capture organisms recalcitrant to culture or enrichment strategies, leading to a more comprehensive understanding of the deep branches of the eukaryotic tree of life as well as a catalog of unstudied and novel enzymes for biotechnological innovation. We will also optimize sample preparation techniques specifically for eukaryotic metagenomics (**MC5-2**). This will include selective filtration steps and manipulating growth conditions. Our users will therefore be able to access a broader diversity of eukaryotic taxa with improved genome completeness.

Organellar genomes can be easily captured from metagenomic samples and serve as a robust proxy for eukaryotic diversity within communities. The Fungal and Algal Program has developed novel annotation strategies and pipelines to address the nuances of organellar genomes, such as introns and repetitive content. Applying these approaches to metagenome-derived organelles will aid in capturing eukaryotic diversity across environments, identify potential mutualistic and antagonistic interactions between eukaryotes and other microbes within these environments, and provide resources to enable our user community to search available metagenomic datasets for eukaryotic signatures (**FA4-2**).

24 Swift, C. L. et al., 2021, "Cocultivation of Anaerobic Fungi with Rumen Bacteria Establishes an Antagonistic Relationship," *mBio* 12:e0144221.

LEVERAGE LONG-READ AND SINGLE-CELL SEQUENCING TO STUDY BLIND SPOTS AND MICRODIVERSITY

Our understanding of Earth's microbial and viral genome diversity has been transformed by advances in (meta) genomics. Yet the improved recovery of genomes offered by these approaches is not evenly distributed across the tree of life, as "taxonomic blind spots" are known to exist. The challenge of blind spots disproportionately affects rare microbes and viruses because their sequence coverage in short-read metagenomes tends to be too low to yield high-quality assemblies. Further, MGEs, such as viruses and hypervariable genomic islands, are commonly not or poorly assembled from short-read metagenomes due to the population-level variability of these regions, which results in unresolved ambiguities in the assembly graph (**PI3-2**).

To begin to fill in these microbial and viral blind spots, we will first build a global census of microbial (including viral) phylogenetic diversity based on IMG/M data for cultivated and uncultivated lineages (**PI1-2**). Based on these data, we will establish a hit list for targeted genome recovery (**MC2-2**). In collaboration with JGI users, we will target underrepresented taxa and genome elements leveraging long-read metagenomics and single-cell sequencing for complex and heterogeneous genome assemblies. This also entails the improved recovery of MGEs, as well as performing extra-large-scale combined assemblies of metagenomes using the recently developed MetaHipMer tool, a collaborative effort between the JGI and the Computational Research Division at Berkeley Lab. Together with the JGI user community, we will conduct large-scale single-cell sequencing, analyzing approximately 1,000 individual bacterial or archaeal cells per environmental sample (**MC1-2**) to capture genomic snapshots of microorganisms from distinct environmental niches including microbial mats, biofilms, and sediment and water samples, and observe temporal variations across different seasons. We expect that these targeted and complementary efforts will aid in saturating the microbial and viral genome sequence space, and bring us closer to the goal of providing at least one high-quality representative genome for all major clades to the research community (**MG4-5**).

Beyond taxa that are challenging to assemble, another aspect of microbial diversity poorly captured by shotgun metagenomics is population-level microdiversity. Fine-scale genetic diversity has been observed in different lineages of bacteria, archaea, and microeukaryotes, and single-cell genomics and long-read metagenomics have emerged as valuable tools for studying this heterogeneity in natural environments. Using single-cell sequencing, we will provide population structure information via linkage of SNPs among closely related single-cell genomes. Such information will allow us to assess niche partitioning and ecological niche adaptations of closely related strains (**MC1-5**), enhancing our understanding of microbial diversity and function in the wild. Furthermore, in collaboration with our BRC partners and entities like the Advanced Biofuels and Bioproducts Process Development Unit (ABPDU), we aim to investigate microbial heterogeneity in an applied setting, primarily bioreactors, using both single-cell sequencing (**MC1-2**) and long-read metagenomics, which is particularly important since population heterogeneity in bioprocesses has been observed but remains poorly understood and characterized.

Harness AI for Prediction of Gene Function and Experimental Design

To date, the JGI has sequenced over three petabases of nucleic acids across all domains of life. These projects range from resequencing individual species to deciphering expressed genes through transcriptomics and building reference genomes across entire kingdoms. Hosting such vast amounts of user data, combined with the computational resources and expertise to analyze them in collaboration with the users who generated the data, constitutes a unique capability that the JGI possesses and sets us apart from other scientific environments. It also positions us to make substantial contributions toward integrating AI and other advanced computational techniques to predict traits and gene functions, and to use these emerging technologies to facilitate experimental design for functional validation.

LEVERAGE EMERGING TECHNOLOGIES TO DISCOVER GENES AND PREDICT THEIR FUNCTION

Many JGI projects are anchored in genes as the central unit of analysis, either as a singular entity within one genome, or as a set of orthologs across species or pangenomes. The proliferation of genome resources across the tree of life has dramatically expanded the encyclopedia of gene sequences. Concurrently, new AI-powered computational tools have the potential to enable the inference of causal networks that connect DNA sequence variation to protein function and whole-

organism phenotypes. Despite these extensive resources and tools, gene discovery, validation, and putative functional prediction remain challenging in all but the simplest systems.

While superficially simple, the task of determining gene sequences in a single genome is a complex and computationally intensive process. Indeed, gene prediction remains a major bottleneck in JGI production pipelines. We are currently undertaking efforts to accelerate gene prediction without incurring losses in sensitivity or accuracy. Over the next five years, we will implement these new speed and accuracy improvements in JGI gene prediction pipelines (**PL6-2**), bringing annotation methods in line with our planned improvements in genome assembly production throughput (**PL1-2**).

Ongoing efforts across JGI genome annotation groups have dramatically improved the quality of single-genome gene prediction. Despite these efforts, both false-positive "artifact genes" and false-negative "unannotated" gene sequences regularly occur even in the highest-quality reference genomes. We have begun extensive testing of novel methods, including those using AI and integration of multiple data sources, to resolve both issues. We will apply methods that significantly reduce the number of artifact genes, such as the incorporation of known presence-absence variation, in the vetting of pangenome gene sets (**PL6-2**, **FA3-5**, **FA5-5**). Protein folding and deep-learning-based gene calling algorithms offer additional fertile avenues for gene prediction improvement, even in the absence of closely related, previously characterized proteins. We will directly integrate protein-assisted methods with our existing transcriptomic and homology support to improve gene model prediction across all new JGI plant genomes released in Phytozome (**PL6-5**) as well as fungal and algal genomes in MycoCosm and PhycoCosm, respectively (**FA1-5**). These efforts will ensure the JGI retains its position as a purveyor of gold-standard gene annotations.

Gene annotation also includes functional predictions that represent the potential actions of a protein in the system, including molecular interactions and phenotypic effects. Functional annotations are typically assigned based on a combination of protein sequence homology, domain, and pathway information from related model species with experimentally validated gene-trait associations. However, this approach is limited when working with genomes of species distantly related to any model organisms. Because of advances in deep neural networks for protein folding, predicted protein fold structures can be used as an additional line of evidence to identify and transfer function between species. Beyond aiding in characterizing protein function through improved deep orthology detection, protein folding and other AI-based gene functional resources facilitate predictions of protein interactions and partners. In the long term, we will apply these and other methods to improve gene functional prediction (**PL6-5**, **FA3-5**).

Using standard similarity-based approaches, on average more than 20% of all predicted microbial proteins cannot be associated with any putative function. This increases to 50 to 70% when analyzing viruses and microeukaryotes, such as fungi and algae. The JGI will create cross-team projects focused on developing and validating new AI-based approaches, which will be applied to improve protein function prediction and solve other important mission-relevant challenges (**FA3-5**). Leveraging the massive amount of data now available in some clades, we will create new genome-based methods that can link genomic features with important traits and associated genes. Together with our user community, we will then validate those predictions using HTP synthetic biology approaches (e.g., approaches for overexpression or knockdown of genes; see **From In Silico Predictions to Functions and Phenotypes**), opportunistically leveraging the Microbial Molecular Phenotyping Capability (M2PC) at EMSL through the existing FICUS program and future opportunities within the M2PC operational framework. Such capabilities will allow us to predict traits in newly sequenced lineages, assign function to potentially thousands of genes simultaneously, and propagate those findings to related species. In the process, we will employ a Design-Build-Test-Learn (DBTL) cycle for efficiently developing and validating new AI technologies.

AI is also a promising tool to predict host-microbe interactions and lifestyle for symbionts and uncultivated microbes. Traditional cultivation methods have limitations in capturing the full diversity of microorganisms; this restricts our knowledge of their metabolisms, evolutionary processes, and lifestyles. AI approaches in genomics and microbiology have potential in overcoming these limitations. By leveraging our new computational system (JGI-Dori) and graphics

processing unit (GPU) acceleration (NERSC's Perlmutter), JGI researchers and users can process large genomic datasets and apply AI algorithms to analyze and interpret the information, allowing for the identification of complex relationships between microorganisms, their hosts, and their environments. "Big data" integration, modeling, and analysis will enable a deeper understanding of symbiont-host interactions. Specifically, the exploration of new lineages of life from extreme environments, such as hydrothermal springs and arid desert regions, presents intriguing possibilities, as they often harbor underexplored microbes. These environments pose challenges for traditional cultivation methods, hampering the study of the organisms thriving there. The JGI will deploy AI approaches to gain insights into evolutionary processes and adaptation strategies within these systems (**MC5-5**). These efforts will expand our understanding of the functional diversity on Earth and pave the path for exploiting microbial functions for biotechnological and other bioeconomy-relevant applications.

CAPITALIZING ON CONTEXTUAL DATA TO FACILITATE CHARACTERIZATION OF FUNCTION AND DIVERSITY

Computational methods for characterization of conserved protein families of unknown function heavily rely on "guilt-by-association" techniques, such as patterns of co-occurrence in different taxonomic lineages, co-localization on the chromosome, and correlated expression patterns. Some of these techniques are implemented in IMG/M, a web-accessible platform for users that provides data and tools to enable the exploration of protein family co-occurrence and co-localization. While these tools offer insights into the potential functions of uncharacterized protein families and support predictions, a unified framework that combines these multiple lines of evidence and assigns statistical significance based on the entirety of the data is lacking. We will implement methods and analysis tools for exploration of all available contextual evidence for functional predictions for uncharacterized protein families of bacteria, archaea, and viruses and develop a framework for assessment of statistical significance of functional predictions (**PI2-2**).

AI methods such as AlphaFold2 can also be used to buttress and refine functional predictions based on chromosomal, phylogenetic, and expression patterns. For some microbial and microbiome data, we will add structural predictions from AlphaFold2 to aid with characterization of protein function through improved deep orthology detection. We will also benchmark the extensions of AlphaFold2 to see if they can help predict interactions with small molecules and infer interactions between proteins. While such predictions for all isolate proteomes and metaproteomes are computationally prohibitive, the selection of candidate interaction partners can be narrowed by application of JGI tools for the analysis of co-occurrence and co-localization mentioned previously. These new tools and data will allow us and our users to generate functional predictions for currently uncharacterized protein families, such as those encoded by MGEs (see the next section), and rank these predictions by their reliability and specificity of predicted function (**PI2-5**). We will maintain data provenance and transparency by indicating the source evidence for these predictions.

LARGE-SCALE IDENTIFICATION AND CHARACTERIZATION OF PROTEIN FAMILIES FROM MOBILE GENETIC ELEMENTS

MGEs, including viruses and plasmids, are abundant in all domains of life, displaying an impressive range of genetic diversity. Despite their pivotal role in horizontally transferring genetic information between unrelated organisms, the functional repertoire of MGEs remains largely unexplored. A significant portion of their genes lack known functions, and the dynamics of gene gain, loss, or exchange mediated by MGEs have not been systematically studied across large datasets.

To enable comprehensive evaluations of MGE gene repertoires, we will organize the protein space of viruses and plasmids through large-scale protein clustering. This clustering will result in a dataset consisting of MGE-encoded protein families. By grouping diverse MGE proteins into families, we can use this dataset to gain valuable insights into their function and evolution. Specifically, this will include (1) protein structure prediction, (2) identification of gene co-occurrence modules, (3) correlation between gene families and phenotypic traits, (4) generation of pangenomes, and (5) taxonomic assignment of viruses. This approach will leverage the existing infrastructure of the Integrated Microbial Genomes and Microbiomes (IMG/M) platform. Specifically, we will use virus and plasmid sequences from the IMG Viral Resources (IMG/VR) and IMG Plasmid Resources (IMG/PR), respectively. Users will be able to explore protein families through the IMG/M web interface and correlate them with rich metadata, including host taxonomy, geographical location,

and ecological variables. This integration of diverse data sources will provide a comprehensive resource for studying MGEs and their associated protein families (**MG4**, **PI3-2**).

DIVE INTO THE UNTAPPED RESERVOIR OF SECONDARY METABOLITES

Organisms have many ways to thrive and survive in the environments they inhabit. Essential functionalities include the production of metabolites, cellular and higher order structures, motility, and a host of responses to stimuli. A key class of metabolites that mediate a myriad of processes are secondary metabolites. While not required for growth per se, secondary metabolites allow organisms to scavenge nutrients, kill predators, and communicate with each other, and confer advantages to those organisms possessing the ability to produce these molecules.

Over the last few years, the JGI has focused its efforts on providing new and expanded capabilities for users on secondary metabolites through creating a new Science Program for Secondary Metabolites. A central asset for this program is the SMC, which is already the most comprehensive publicly accessible resource for secondary metabolite BGC information. The SMC is a repository that adheres to findable, accessible, interoperable, and reusable (FAIR) data principles, with convenient web and application programming interface (API) access for users and the scientific community, and currently contains data for nearly thirteen million BGCs from over 1.1 million bacterial genomes. We are already working to further populate this resource with data from a broader set of prokaryote and eukaryote genomes generated by the JGI or provided by the scientific community. As more sequence information becomes available, we will update the predictions and continue to expand SMC (**SM1-2**, **SM1-5**).

We will further develop the SMC through the import and development of new tools for BGC prediction, especially for new classes of secondary metabolites that have so far been overlooked, and include tools desired or recommended by our users (**SM2-2**). We will add functionalities to help users navigate and better understand the content of SMC, as well as for experimental purposes, such as for gaining deeper understanding of how BGCs relate to the host's niche (**SM2-5**). AI techniques will be applied in a bgc-Chat mode to give users a new way of querying the data in SMC and deriving novel insights from it (**SM3-5**). This latter aim will be aligned to collaborative efforts to create a standardized language (biosynthetic gene cluster query language, BGC-QL) to describe and query genetic and enzymatic content of BGCs (**SM3-2**).

Strategic Objective 2: From In Silico Predictions to Functions and Phenotypes

While computational methods provide valuable insights and predictions about functions and phenotypes in silico, confirming these predictions through laboratory experiments is essential. This is often accomplished by studying mutations in candidate sequences or over- or mis-expression of genes in the host or heterologous systems. The moderate throughput of these approaches has limited the rate of gaining definitive functional knowledge about individual genes. This bottleneck has in turn limited the development of approaches to improve desired phenotypes through transgenic, gene editing, and synthetic biology approaches. Advances in genome sequencing and population genomics have created a monumental backlog of candidate genes and sequences that are awaiting this final, but necessary, step before they can be used to improve phenotypes of interest. Thus, developing and applying more efficient methods to functionally characterize sequences at a higher scale is essential. To best support its user community, the JGI will develop an experimental toolkit for probing functional diversity to translate in silico predictions into actionable biological knowledge, supporting a growing and sustainable bioeconomy.

Characterize Functional Diversity through Multi-omics Integration

Leveraging the ongoing advancements in generating high-quality genome sequences at the JGI in conjunction with our additional omics capabilities, we will generate and integrate multi-omics user datasets to deepen our understanding of functional diversity across diverse organisms. Connecting different layers of biological information, including TF binding, epigenetic markers, expression patterns, and metabolomes, will create a more complete picture of an organism's functional dynamics. Such a more holistic approach to our understanding of biology has great value for the

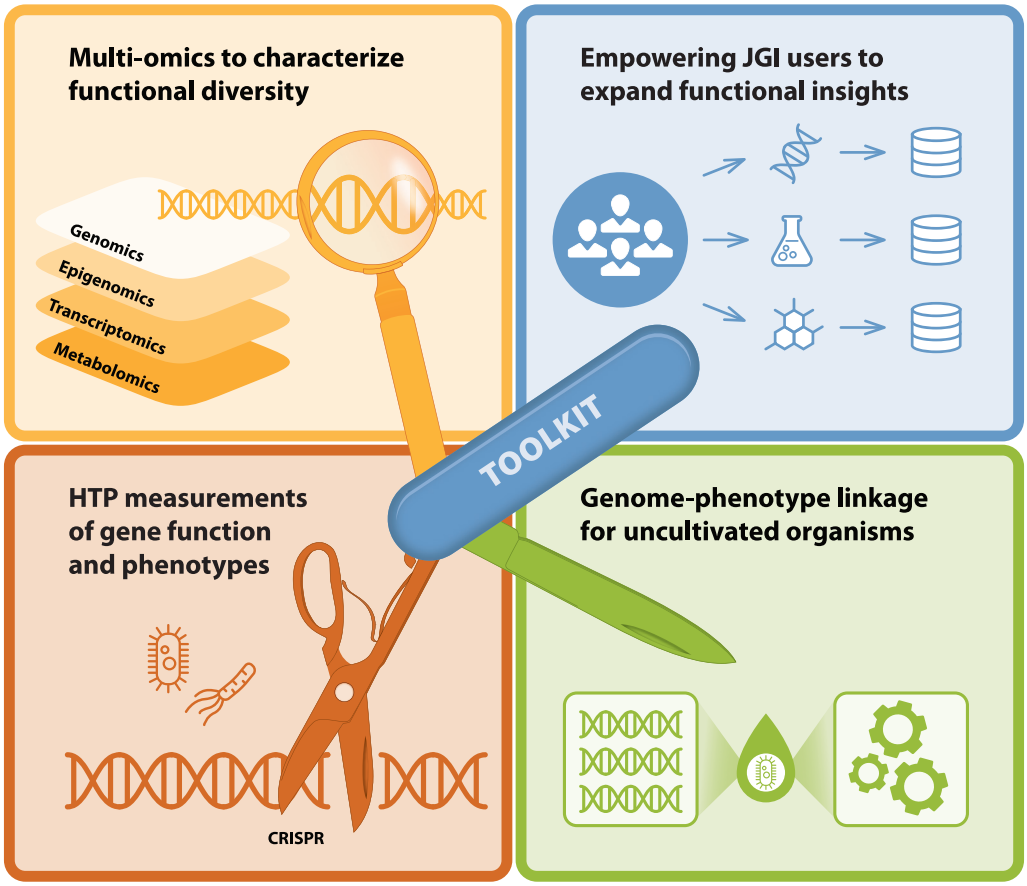


Fig. 7. Probing functional diversity. The JGI toolkit for probing functional diversity will facilitate moving from in silico predictions to functions and phenotypes.

bioeconomy, for example by facilitating increased yields of bio-based products. Overlaying sequenced genomes with these multi-omics datasets will allow new insights into functional diversity and ultimately pave the way toward predictive biology.

INTEGRATE GENE REGULATORY NETWORKS, TRANSCRIPTOMES, AND METABOLOMIC DATA

A spectrum of multi-omics capabilities is already available to JGI users through the user programs. Integration of this data can provide information about gene function and enable reconstruction of metabolic pathways and gene networks. A set of tools for visualizing and analyzing diverse omics data is growing and being integrated into JGI portals for user analysis (**FA5-5**). The latest addition to this suite of capabilities is DAP-seq for determining the sequence motifs bound by TFs and mapping the genome-wide distribution of their binding sites. We will improve upon existing experimental DAP-seq capabilities to further scale this method and enable the investigation of many TFs simultaneously. This will also require expanded computational analyses and tools, including databases of known TFs and their respective binding motifs and visualization capabilities to analyze DAP-seq experiments within an interactive network analysis framework. Combined with various multi-omic measurements of TF targets (e.g., gene expression or DNA methylation), such visual capabilities will enable JGI users to identify TFs regulating important traits and metabolic functions more easily. Combined analysis of TFs, gene target orthologs, and binding sites can further expand the impact of DAP-seq data through the propagation of results to related genomes. Discovery through visualization will be enabled by overlaying the information over pathways and interactive network analysis (**FA5-5**).

We will also continue to improve the performance of JGI metabolomic capabilities to meet JGI user needs and the DOE mission. Central to this strategic objective will be metabolomic analyses that illuminate the currently dark

biochemistry occurring within plants, microbes, and their communities to discover novel genes, proteins, and microbes by associating novel metabolites, genes, and proteins. To accomplish this, we will develop, test, and deploy integrated and experimental and cheminformatic tools for identifying novel metabolites and integrating them with sequencing data to provide an integrated genometabolic understanding (**ML1**, **ML2**, **ML5**).

DEVELOP SINGLE-CELL TRANSCRIPTOME SEQUENCING IN MICROBES AND MICROBIAL COMMUNITIES

Connected with efforts to establish single-cell RNA sequencing (RNA-Seq) on microbial cultures and enrichments, an important capability for the field would be methods enabling researchers to measure transcriptional activity at the single-cell level and in large communities (**MG5**). Considering the potentially high impact of such methods for our understanding of microbiome processes, we will explore several promising techniques in this area and evaluate whether they may be applicable to JGI user samples. We will build on existing droplet PCR capacities at the JGI and other recently proposed approaches to test the targeted capture, barcoding, and sequencing of prokaryotic transcripts at the single-cell level. We will also evaluate the potential application of spatial transcriptomics using spatial barcodes, combinatorial mRNA labeling, or sequential fluorescence in situ hybridization (FISH) to prokaryotic communities. The ultimate goal of these efforts is to enable single-cell RNA-seq in microbial communities at scale and to offer this capability to JGI users.

Enable High-Throughput Measurements of Gene Function and Phenotypes

Interpreting biological meaning from large omic datasets is often impeded by the lack of functionally verified knowledge of basic gene functions. This knowledge gap, in turn, limits our ability to engineer organisms and to predict the outcomes of engineering efforts, especially in systems beyond model organisms and of relevance to DOE. While traditional approaches to determine gene function are non-scalable and laborious, more modern genome-scale methods can be performed at a scale and pace approaching that of genome sequencing.

PROVIDE CRISPR SGRNA LIBRARIES FOR GENOME-WIDE FUNCTIONAL SCREENS

Genome-scale clustered regularly interspaced short palindromic repeats (CRISPR) interference (CRISPRi; i.e., gene knockdown) and CRISPR activation (CRISPRa; i.e., gene upregulation) libraries allow rapid sequence-based linkage of genotype to phenotype and provide the foundation for accelerating HTP gene function assays. Specifically, the contribution of each targeted locus to any environmental or genetic perturbation can be quantitatively assessed. Having already generated hundreds of CRISPR libraries, the JGI has the capability to rapidly advance those technologies from model organisms to diverse environmental samples. Along with these computational and molecular approaches, we are building a web-accessible database to house all captured, normalized data, and providing the analytical tools for cross-experimental comparative investigations. Specifically, this will enable identification of genes responding to given screen conditions across organisms, and add confidence to function predictions via guilt-by-association principles (e.g., gene co-occurrence or gene expression analyses). This new tool will advance community access to data, providing tools for analysis and visualization, and facilitate distribution of the physical libraries to users (**GT8**).

SUPPORT FUNCTIONAL SCREENS FOR ENZYME ACTIVITY AND PROTEIN INTERACTIONS

Mass spectrometry is well suited for HTP characterization of enzyme activities. For example, the metabolomics team has developed a number of enzymatic assays for glucoside hydrolases, lignin active enzymes, and amino transferases that can be made available to JGI users. In addition, JGI LC-MS/MS methods can be optimized for additional types of enzymatic assays. Therefore, we will evaluate potential user demand for enzymatic activities highlighted in the strategic plan, including against polyphenolic compounds (**ML4**). Some of these enzymes might find utility in bio-based processes.

Link Genomes to Phenotypes for Uncultured Microbes and Consortia

While our understanding of isolated organisms has grown immensely, a wealth of biological diversity and intricate inter-organismal interactions in microbial ecosystems remains unexplored. Approaches for sequencing, assembling, and binning metagenomes have advanced greatly, but methods for assessing

phenotypes in microbial communities and uncultured microbial species are currently limited. We will harness single-cell and SIP approaches, establish new experimental model systems, and explore the transformative impact of viral infections on microbial cells to connect genotypes and phenotypes and to shed light on interactions between organisms and/or viruses. This strategy is poised to reveal the functional traits of uncultured organisms and microbial ecosystems and shed light on the complex interactions within these communities.

DEVELOP SINGLE-CELL AND STABLE-ISOTOPE PROBING (SIP) APPROACHES TO LINK PHENOTYPES OF UNCULTIVATED MICROORGANISMS TO TAXONOMY

Most microorganisms across the tree of life remain uncultivated. Thus, a scarcity of information regarding their phenotypes is available, including morphological, physiological, and other functional traits. As a result of these gaps, our textbook knowledge regarding the phenotypic features of microorganisms is likely heavily skewed and incomplete.

To elucidate phenotypes of microbial dark matter, the JGI aims to develop novel cultivation-independent, single-cell approaches to connect the phenotypes of uncultivated microorganisms with their taxonomy (**MC4**). Approaches currently under investigation include imaging combined with laser microdissection (**GT5-2**), as well as Raman-activated cell sorting and flow cytometry combined with whole-genome amplification and sequencing. We also aim to further develop probe-labeling approaches to target environmental microorganisms that are, as one example, involved in the production of secondary metabolites (**MC4-2**), as well as other DOE-relevant pathways and processes. If any of these newly explored approaches are successfully validated, we plan to offer them as additional products to JGI users (**GT5-5**). Similarly, the JGI will keep developing its quantitative SIP metagenomics capacity, first by increasing sample processing throughput and establishing standardized pipelines to robustly link uncultivated taxa to specific metabolisms, and later to connect these lineage-specific metabolic measurements to metabolomics data and metabolic models (**MG3**).

ESTABLISH NEW MODEL SYSTEMS TO STUDY INTER-ORGANISMAL INTERACTIONS

Our current understanding of the mechanisms of microbe-host interactions and the consequential impact of these associations on ecosystem structure and function is limited. Reasons include the paucity of symbiont-host systems that are experimentally accessible and the focus of existing model systems on microorganisms of medical relevance.

The JGI will work with its user community to facilitate the development of an experimental framework for the systematic study of inter-organismal interactions. Applying correlative microscopy will allow us to visualize the three-dimensional distribution of the interacting partners (**MC4-5**), and reveal the location of an endosymbiont in relation to the surrounding host cellular compartments. Combining such data with genomics and transcriptomics is expected to reveal putative virulence factors and other genes required for symbiotic interactions and genes that may alter host metabolic pathways. To study microbe-host interactions under controlled conditions, we will evaluate microcosm setups in EcoFABs (**ML3**, see **Fig. 5**), which allow controlled experimental manipulations of the system, expected to provide novel insights into the impact of microbial symbionts and their hosts on ecosystems. If EcoFABs can be validated for the reproducible and standardized study of microbial symbiosis model systems relevant to the DOE mission, the JGI will offer this capability to its user community.

UNDERSTAND THE IMPORTANCE, SPECIFICITY, AND PROPERTIES OF VIROCELLS

Microbial phenotypes can be deeply impacted by viral infections. While viral infection may eventually culminate in the destruction of the host cell as the most obvious and consequential impact, the metabolism and behavior of a microbial cell changes substantially throughout earlier stages of an infection. These changes can persist over extended periods of time and cause the host cell to be very different from an uninfected cell. The concept of a "virocell" was established to describe these cells undergoing an active viral infection, and thus far has been studied mostly in aquatic environments, using established models such as cyanophages-cyanobacteria. Comparable information is not available for soil microbes.

To improve our collective understanding of virus-host interactions and virocell metabolism in soil, the JGI will work with the user community to formalize and standardize this virocell characterization approach and apply it to an expanded set of virus-host pairs, with a focus on those relevant for soil ecosystems. We

will leverage the latest advances in RNA-Seq and HTP metabolomics, as well as transcription regulation analysis using DAP-seq, to identify virus-driven regulation of host operons and to compare infected to uninfected cells for different infection types and stages (**MG1-5**). We will also explore how the robotic capabilities of the EcoBOT and the standardized incubation device EcoFAB can be used to perform these studies across different environmental conditions, and eventually address questions about the interaction between hosts, viruses, and environments (**ML3**). Once established, these combined approaches and associated analysis strategies (see **Nutrient Cycling, Understanding Biological Drivers of Carbon Capture and Sequestration**) could become a new JGI product enabling users to characterize a virus-host pair they already have in culture and identified as potentially important in their ecosystem of interest.

Empower JGI Users to Expand Functional Insights

As an SC-supported user facility, the JGI is built on a strong foundation of user support, collaborative science, and fostering scientific user communities, which brings together expert researchers from around the world. User-driven inquiry has formed the foundation and continues to be a mainstay for basic and applied discovery from sequencing projects. In the next five years, we will further promote these values and pursue efforts to facilitate functional discoveries by the JGI user community.

ENABLE FUNCTIONAL DISCOVERY IN PLANT GENOMES THROUGH TOOL DEVELOPMENT AND COMMUNITY PROJECTS

The JGI is leading the generation of pangenome data for DOE-relevant plant species. However, the paradigm of pangenomes brings with it significant complexities. The analyses required to access even simple data remain challenging and require specialized data science expertise. We have piloted a program in which we pair a JGI analyst with users. The resulting mentoring relationship reduces total analyst time investment and grants independence to users. We plan to advance this program and apply it to many ongoing projects over the next two years (**PL3-2**). We will simultaneously work to develop better data access protocols and user-facing analysis methods (**PL3-5**), which will allow all users, regardless of coding experience, to interact with and explore the complex pangenome data structures.

While genetic models offer a foundation to better understand agricultural and ecological systems, their impact is limited without methods to easily translate information across species. The JGI has initiated and will continue to lead this effort by integrating quantitative experimental design with comparative genomics, producing tools and data resources that will improve the ability for users to translate functional information for regions of interest across genomes within species (pangenomes) and between model and experimental systems (**PL2-5**, **PL3-5**, **PL4**). These efforts will expedite biotechnological improvement of emerging and established DOE species of interest, including switchgrass, miscanthus, sorghum, poplar, pennycress, camelina, and models Brachypodium and *Panicum hallii*. Combined, these efforts will (1) increase knowledge about functional impacts of genetic variation and its links to phenotypic applications (**PL4-2**), (2) improve gene function identification (**PL4-5**, **PL6-2**), and (3) permit fine-scale comparisons of variation within a species for engineering pathways (**PL2**, **PL4-2**). At a finer scale, discovery and analysis of contrasting molecular evolutionary patterns in conserved regulatory elements and genes within regions of interest across and within taxa (e.g., grasses, eudicots, or poplars), will enable a priori exploration of high-value targets for biotechnology (**PL2-5**).

ENABLE USER-CENTRIC PRIORITIZATION AND ACCESS TO PLANT, FUNGAL, MICROEUKARYOTE, BACTERIAL, AND ARCHAEAL FUNCTIONAL GENOMIC CAPABILITIES

The definitive assignment of function to sequence requires experiments to test the role of candidate sequences. The prioritization and access of our users to resources for engineering is, therefore, instrumental.

For plants, we will facilitate access and develop three types of resources for genetically and experimentally tractable plants. First, since even the most efficient methods to functionally characterize plant genes are labor intensive, we will employ a multifaceted strategy to select the most promising candidate genes based on integrating as many data types as possible (e.g., metabolomic, co-expression, pathway construction, single-cell transcriptomics, population genomics, and comparative genomics, **PL4**). Second, we will sequence collections of chemically or radiation-induced mutants to identify millions of mutations to enable forward- and reverse-genetic experiments (**PL5**). We will work

with user communities to organize these projects and facilitate access to the mutants and data.

For microbes, we currently focus on type strains from species that are highly DOE mission related. However, users often have their own field isolates tied to specific environments with unique physiological properties. Fortunately, many genetic tools are transferable with minimal modifications at the species and genus levels, and sometimes even higher taxonomic levels. We will develop generalizable workflows for microbial strain engineering, starting from sequencing, to replicable vector construction and DNA transformation, to genetic parts characterization and archiving, and finally to building genome editing tools (**GT9**). We will pair type strains with user-submitted strains to test the application of those tools. We eventually would like to build a comprehensive toolkit to enable user-centered microbial engineering capabilities.

ENABLE THE RESEARCH COMMUNITY TO ADDRESS MAJOR CHALLENGES IN FUNCTIONAL ANNOTATION

Advanced annotation approaches such as the ones described earlier will identify a number of proteins, enzymes, and pathways that are likely to be novel and should be characterized experimentally. While some of these experiments will be conducted at the JGI, the number and diversity of elements to be characterized will be much larger than can be reasonably addressed by a single institution. The development of specific analyses, visualization tools, and user interfaces to explore the global diversity of novel protein sequences will help enable the community at large to perform these experiments in an informed and targeted way (**MG4**). Based on the development of an integrated framework for enzyme discovery described in **Nutrient Cycling, Carbon Utilization for Bioproducts**, we will expand this to enable (1) the development of probes (e.g., PCR primers, FISH probes) for targeted detection of a specific group of genes or organisms based on metagenome sequences, and (2) the development of metrics and analytics to establish lists of "most wanted" uncultivated organisms and novel genes that could be used as a starting point for community-wide collaborative efforts, including the prediction of potential cultivation condition and in vitro verification of these predictions (**MC2-2**).

Strategic Objective 3: Leveraging Functional Insights to Enable Biosystems Design

Technologies we use to modify and harness biological systems are a driving force for the bioeconomy. However, successful adoption and improvement of biological systems typically requires multiple rounds of the DBTL cycle (**Fig. 8**) due to the inherent complexity of such systems and our limited ability to predict the consequences of targeted changes. Over the next five years, JGI researchers will keep building their knowledge of plant and microbial systems to enable biosystems design using big data for sequence-function relationships and advanced computing to predict fruitful genetic modifications, in combination with automation to move toward a self-driving platform. These efforts will help to reduce the risks of developing and scaling up these new processes.

Collect Large Datasets for Analyzing Sequence-Function Relationships

AI is increasingly used in biosystems design to develop more useful enzymes, pathways, and strains. However, the lack of large-scale, reliable datasets for sequence-function relationships limits the use of AI. To improve the ability to generate these datasets, the JGI will work with its users to curate functional data on natural and synthetically created sequence diversity and leverage generalizable HTP capabilities for functional characterization of diverse biosystems.

CURATE FUNCTION DATA ON NATURAL SEQUENCE DIVERSITY

To design and build desired biosystems, useful biological parts for target chassis are needed, such as enzymes and transporters with optimal properties (e.g., expressibility, solubility, stability, specificity, and activity). Through its Genomes to Structure and Function efforts,[25] the JGI works with users to characterize diverse protein families. These datasets may be used to train AI models to better correlate sequence-function relationships and improve selection of useful biological parts. The JGI is interested

25 https://www.berstructuralbioportal.org/genomes-to-structure-and-function-workshop-report-2022-released

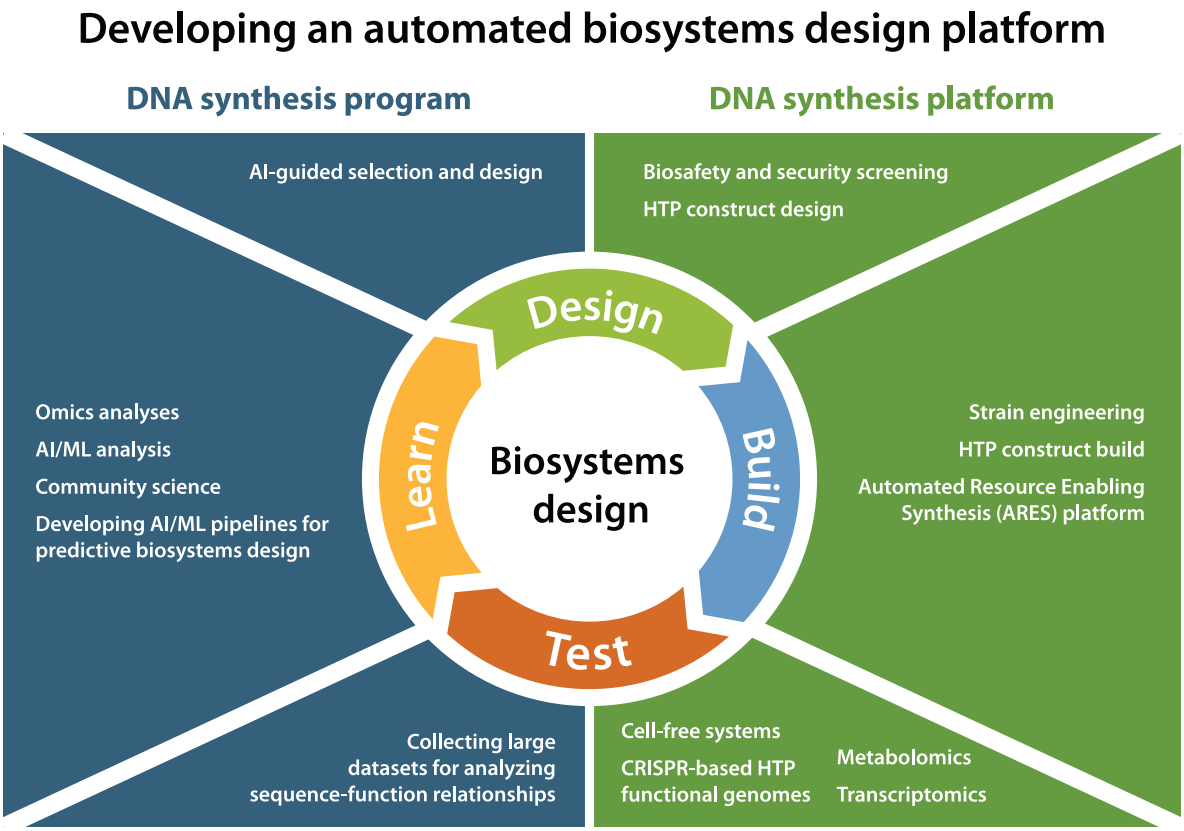


Fig. 8. Accelerating the DBTL cycle. Advancing the pace of the Design-Build-Test-Learn (DBTL) cycle is critical for enabling effective biosystems design.

in creating an efficient feedback loop, curating user data, and creating a data repository (**SS3-2**, **SS4-2**).

IMPROVE OUR ABILITY TO EXPAND ENGINEERING SEQUENCE DIVERSITY

To generate large-scale sequence-function relationship data, we must improve our ability to create diverse sequences. The JGI has adapted diverse capabilities for protein and pathway engineering. We can also synthesize single-guide RNA (sgRNA) libraries for CRISPR-based strain engineering (e.g., SNPs, insertions and deletions, cassette exchange). We have also developed chassis-independent recombinase-assisted genome engineering (CRAGE) for rapid integration of genetic payloads in diverse bacterial species. The JGI will continue to expand its ability to build constructs and strains with high-sequence diversity (**GT8**).

DEVELOP GENERALIZABLE PLATFORMS FOR HTP FUNCTION CHARACTERIZATION

The JGI will explore generalizable platforms for HTP function characterization of proteins, pathways, and strains (e.g., cell-free expression, cell-surface display, microdroplet, microfluidics biosensors, imaging) coupled with next-generation sequencing as a core readout (**SS1** and **SS2**). The sequence modifications important for strain performance will be identified on a massively parallel scale, and datasets will be created that can be used for analyses and correlative phenotype-genotype associations as well as in training more accurate AI models. Additionally, the JGI will explore the opportunities to collaborate with microbial phenotyping centers at EMSL. For example, EMSL has developed NanoPOTS, which are designed to obtain a maximum amount of information (e.g., proteomics and metabolomics) from the smallest amount of sample. These technologies are expected to enable further characterization of microbial strains.

Develop AI Pipelines for Predictive Biosystems Design

To design and build stress-tolerant microbes, the JGI has developed ML-assisted engineering of stress

tolerance rational optimization (MAESTRO), which efficiently correlates target traits with their responsible genetic variants. We have already validated some of these predictions experimentally. MAESTRO has several improvements over traditional GWAS approaches and is more applicable for biosystems design. For example, MAESTRO facilitates prediction and prioritization of engineering targets from a much smaller data size than usually required for traditional GWAS. In the next five years, we will further improve MAESTRO's prediction accuracy and expand its use in protein and strain engineering to improve bioproduct production (**SS3**).

CREATE A MULTI-OMICS DATA REPOSITORY AND REFINING FEATURE SELECTION

Sequence-function relationship datasets often have deep genotype data resulting from natural or artificially created diversity, but limited depth-associated phenotype data. To make these datasets more useful, we must reduce the genetic variants to consider. We currently use linear (i.e., GEMMA) or nonlinear (i.e., HSIC-Lasso) regressions. The omics data may be more effectively used to further improve the feature selection. For example, these data could suggest genes that are co-regulated under given conditions. Genetic variations which occur in these co-regulated genes may be prioritized (**SS3**).

EXPAND THE UTILITY OF MAESTRO IN PROTEIN AND PATHWAY ENGINEERING

Metabolic engineering requires enzymes and transporters with high activity and specificity, and their expression must be balanced to maximize flux through the pathways. Flux among central metabolism must also be rewired to increase flux through the immediate precursors of the target pathways. MAESTRO is currently limited to GWAS to improve the stress tolerance of strains. Therefore, the JGI will modify MAESTRO to accept protein and pathway sequences as input data. In this way, we can extend the utility of MAESTRO for protein and pathway engineering. Additionally, the JGI plans to explore different AI methods to improve MAESTRO's predictive accuracy (**SS3**).

EXPAND COLLABORATION TO DEVELOP NEW AI TOOLS FOR STRAIN DEVELOPMENT

The JGI will strengthen collaboration with its users, partners, and stakeholders to develop new AI abilities (see **Stewarding Resources, Evolve the JGI Workforce**) to predictively design strains with desired functions. We will continue to build on existing connections and expand these collaborations to facilitate further progress in this dynamic field. For instance, the Joint BioEnergy Institute (JBEI) and the Agile BioFoundry developed the Automated Recommendation Tool (ART), which uses ML and probabilistic modeling to guide strain development systematically, even without full understanding of the biological system. The Center for Advanced Bioenergy and Bioproducts Innovation (CABBI) developed an evolutionary context-integrated neural network (ECNet), which uses a deep learning (DL) algorithm that exploits evolutionary contexts to predict functional fitness for protein engineering. JGI users have developed AI tools that can design high-efficiency sgRNA for genome, enzyme, and pathway engineering. We will further encourage users to use the JGI CSP to develop new AI capabilities and adapt their tools to further support other users (**SS3-2**). We will initiate a competition similar to the critical assessment of protein structure prediction (CASP) to accelerate AI development through community-based approaches (**SS3-5**). In this competition, the JGI will provide the participants with large datasets (e.g., enzyme sequences and their corresponding activities) and ask the participants to design variants with desired activities. The JGI will then synthesize and characterize the activities of these variants. We will collaboratively publish the outcomes of these studies and make the design tool available to users. We will work with KBase to make some of these tools available for scientists to use.

Develop an Automated Biosystems Design Platform

Biosystems design allows us to engineer proteins, pathways, and strains with desired characteristics. However, engineering within the DBTL cycle is often ad hoc, and therefore is costly in terms of funding, human resources, and time. Complementary to the development of AI pipelines as described in the previous section (**Develop AI Pipelines for Predictive Biosystems Design**) the JGI aims to accelerate the DBTL cycle by enhancing automation for each phase of the cycle.

DEVELOP ADDITIONAL CAPABILITIES IN THE DESIGN AND BUILD PHASES

The JGI has established HTP pipelines for designing and building synthetic DNA constructs. The design pipeline includes computational tools such as Biosecurity List Sequence Screening (BLiSS) and Build Optimization

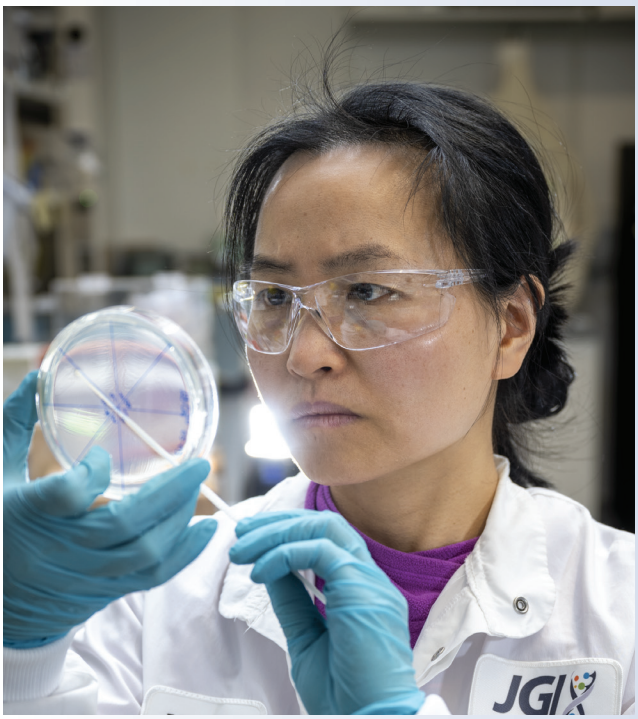
Software Tools (BOOST). These software tools enable biosafety and biosecurity screening as well as streamlining of DNA synthesis and construct assembly. The JGI also developed an automated platform for DNA construct building called Automated Resource Enabling Synthesis (ARES).[26] This platform enables the use of diverse modern cloning technologies to build constructs of various sizes. In the next five years, we will further develop BOOST and related capabilities to support acceleration of the design and building of synthesis constructs as well as explore additional options for biosecurity (**GT-8**).

DEVELOP AN AUTOMATED PROCESS IN THE TEST PHASE

Transforming the libraries of synthesized constructs into recipient strains and identification of high-performing engineered strains among the resulting transformants represent principal bottlenecks of the test phase within the DBTL cycle. Both provide opportunities for substantial acceleration through automation. We envision the use of robotics to integrate the capabilities discussed in the section **Develop Generalizable Platforms for HTP Function Characterization** with the sequencing analyses. This integrated process will enable classifying the engineered strains into different populations depending on their levels of target trait performance (**SS4-5**). Additionally, it will provide the sequence information needed to understand key genetic changes for further exploration. This integrated process will be automatically iterated to continually engineer the strain toward higher performance potential and gain insights into positive and negative contributions to performance.

DEVELOP COMPUTATIONAL TOOLS TO HELP BRIDGE THE GAP BETWEEN THE LEARN AND DESIGN PHASES

Significant gaps remain between the Learn and Design phases of the DBTL cycle. Integration of knowledge gained from large-scale datasets, as described previously, provides opportunities to improve our design principles (**SS4-2**). For example, MAESTRO proposes a number of genetic variations that may improve the target traits. Currently, we manually design constructs to test the effects of these genetic variations. Our goal is to add a new function to MAESTRO so that it can provide basic construct design information, which could be fed to the JGI existing construct design pipeline.



Aspirational Goal: Developing Holistic Cell Models for Biosystems Design

The ability to design and build commercially viable biosystems to produce biofuels, bioproducts, and biomaterials from renewable resources has relied on limited information about cellular systems. To improve these abilities, the community must increase knowledge of the biological, physical, and chemical processes of biomacromolecules at the systems level. Knowledge of nano- to mesoscale organization of cellular assemblies is critical for understanding interactions and processes in cellular metabolism that will lead to the development of new products, such as biofuels, bioproducts, and advanced bio-based materials. As a stretch goal, we aim to use the approaches linking genetic methods with various analytical tools, such as spectroscopy, microscopy, and tomography techniques, through inter-facility collaborations with BER structural biology and bioimaging resources and to study environmentally relevant genes in the context of their cellular environment. These studies will lead to new knowledge of the functions of microbial organisms and of biomolecules in their native biosystems that can be used to accelerate biosystems design applications.

26 https://jgi.doe.gov/about-us/progress-report/2022-progress-report

Strategic Theme 3:

Data and Connectivity

Standardize and Streamline JGI Data, Systems, Tools, and Resources to Enable Scale

Background

Since the JGI was established as an SC-supported user facility in 2005, it has continuously adapted its user programs and proposal calls to meet the evolving demands of the broad scientific community and align with technological advancements and scientific priorities of BER. Capabilities offered have expanded and scales have increased, going hand in hand with higher machine capacities and laboratory automation, as well as new science drivers. For projects that require standard products at scale, such as time-resolved metagenomics studies with thousands of samples, this will enable a shift in JGI proposal calls toward encouraging significantly larger-scale community proposals that require the capacity, logistics support, and expertise of a major genome center and cannot be handled by smaller institutional core sequencing facilities. As the JGI continues this growth, there is a crucial need to expand both experimental and computational capacities to meet these increasing demands.

For more than 10 years, the JGI partnered with NERSC for its computing infrastructure and to gain access to world-class supercomputers. In 2022, the JGI shifted to a geographically distributed infrastructure to meet demands for data access, integration, and generation. In addition to a large allocation at NERSC, JGI computing resources now include a new midrange cluster, Dori; web services and instrument data storage at the Integrative Genomics Building (IGB); computing and archive capacity at our partner facility, EMSL; and surge capacity in the commercial cloud. The JGI has deployed complementary software infrastructure to enhance the usability of the infrastructure, and in the future will look to partner with KBase to offer interactive data and computing capabilities for our users.

Science is entering a new era where agencies are emphasizing data integration to support exploration of new research questions that span domains and may incorporate new paradigms like AI. The JGI will leverage its current partnerships with EMSL, KBase, NMDC, the National Center for Biotechnology Information (NCBI), NERSC, and the National Science Foundation's National Ecological Observatory Network (NEON) to offer new capabilities and resources to users. Some of the new resources will be tools like Branchwater that enable search of petabytes of sequence data in seconds. Others will be education and training materials co-developed with KBase, NMDC, and EMSL. Unifying data infrastructure in partnership with the scientific community is the first step in establishing a new data ecosystem for biological and environmental research.

Opportunities

The JGI supports the entire data life cycle from experimental design to reuse, creating opportunities to influence and improve the user experience throughout (**Fig. 9**). Metadata or contextual information regarding sample collection, data generation, and analysis must be available to understand and reuse scientific data. Ideally, metadata also conforms to community standards. However, this has proven too high a bar for many individual labs, as evidenced by the metadata available in repositories like the Sequence Read Archive (SRA) at NCBI. The JGI excels at providing diverse scientific communities with large amounts of standardized high-quality data and metadata, as well as data generated from emerging new technologies, because we support the entire data life cycle and embed metadata and its curation in our core processes through our Genomes OnLine Database (GOLD) system. JGI scientists work closely with users to design experiments that make the best use of the technology and capabilities the JGI offers. Every digitized sample is processed through robust quality control, assembly, annotation, and analysis pipelines, so users can put their data into context with

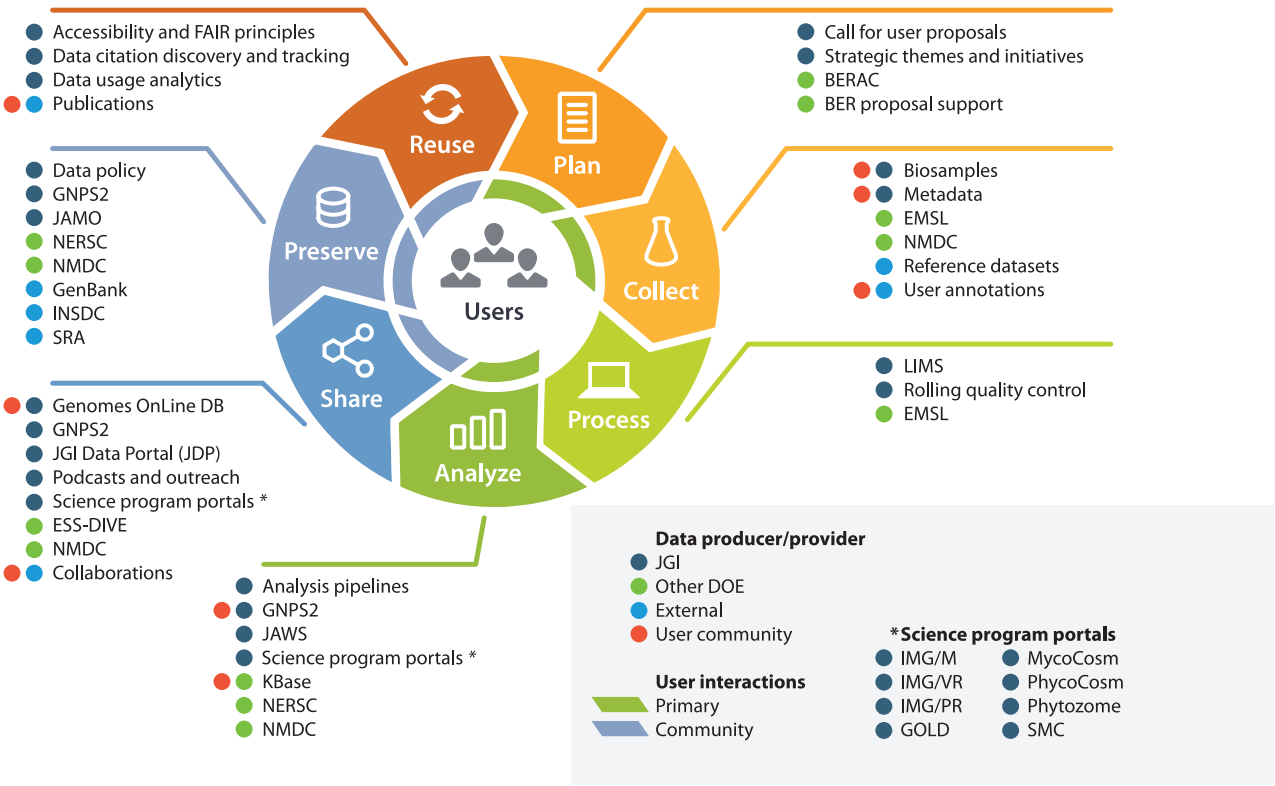


Fig. 9. The user-centric data use cycle and relevant JGI services in the context of the cross-BER data ecosystem. The segmented outer circle represents stages of the data use cycle, along with existing JGI services and relevant resources across the BER data ecosystem. The JGI science program portals (IMG/M, IMG/VR, IMG/PR, GOLD, MycoCosm, PhycoCosm, Phytozome, SMC) are grouped into a single science program portals item. BER supports user programs including BRCs and BERSS.

other experiments via JGI flagship portals (IMG/M, MycoCosm, Phytozome, PhycoCosm, SMC). The JGI also deposits metadata and data into national repositories on behalf of the users, further enhancing the findability, accessibility, and reusability of that data.

The JGI data stewardship expertise and infrastructure make it possible for a broad community of scientists and engineers to exploit our data to develop new algorithms, models, and tools in support of DOE mission priorities and the bioeconomy. For example, data generated by the JGI was instrumental in modeling the impact of a large-scale global ocean warming event on microbial communities and understanding how climate change may affect the ability of oceans to function as a carbon sink.[27,28] In another instance, JGI users developed a semi-automated HTP stable-isotope probing (HT-SIP) technology to uncover novel interactions between bacteria and arbuscular mycorrhizal fungal (AMF) that play crucial roles in terrestrial nutrient cycling processes.

Initiatives like these are central to DOE's Energy Earthshots programs to drive us toward our goal to achieve a clean energy economy.

Strategic Objective 1: Evolving a Cross-BER Data Ecosystem

A data ecosystem is a connected resource where data from different organizations is discoverable and accessible. Linking the resources of powerful institutions like the JGI, KBase, NCBI, NMDC, EMSL, and the Protein Data Bank (PDB) will enable predictive ecosystem biology. The collaborations present exciting possibilities for scientific exploration, granting access to diverse datasets that deepen our understanding of how organisms interact with each other and with their environment and help unravel the complexities

27 Traving, S. J., C. T. E. Kellogg, T. Ross, et al., 2021, "Prokaryotic Responses to a Warm Temperature Anomaly in Northeast Subarctic Pacific Waters," *Communications Biology*, 4:1217, https://doi.org/10.1038/s42003-021-02731-9.

28 Nuccio, E .E., S. J. Blazewicz, M. Lafler, et al., 2022, "HT-SIP: a Semi-automated Stable Isotope Probing Pipeline Identifies Cross-Kingdom Interactions in the Hyphosphere of Arbuscular Mycorrhizal Rungi," *Microbiome* **10**, 199, https://doi.org/10.1186/s40168-022-01391-z.

of ecosystems that remain unexplored. By combining omics and climate data, we gain valuable knowledge about ecosystem responses to environmental changes. Similarly, combining omics and structural data represents a promising avenue for further characterization of novel functional diversity.

As a starting point, the initial integration of genomics data from KBase, NCBI, and the JGI will reveal molecular mechanisms that could be discovered only through the analysis of DNA sequences at the global scale, giving new perspectives on genetic and functional traits still hidden in the sequencing data collected in the past two decades. Interconnected NCBI and JGI databases will allow the reconstruction of comprehensive global ecological networks to enhance our understanding and predictive power of the role of organisms on climate and ecosystem dynamics, while nurturing the future management and development of sustainable ecosystems through simulations and real-time analysis of ecological disturbances. The incorporation of three-dimensional protein structures and structure predictions from databases like PDB will provide us with novel insights into the molecular drivers of ecosystems and their evolution. Based on this new scientific knowledge, we can identify precise intervention points to safeguard populations and ecosystems and enhance their resilience, address climate change, ensure food security, and drive innovation in biotechnology to power the bioeconomy. The interconnected future will also foster interagency and, ultimately, global collaboration, unveiling planet-level ecological patterns and enabling the creation of large-scale eco-biological models. Emphasizing open access and reproducibility, we will empower everyone to contribute to the collective knowledge base.

Data discoverability and accessibility are critical requirements for amplifying the JGI's impact. Data generated from a single investment will be reused for further science. Knowing how JGI data are reused is crucial to understanding the changing directions of research, fulfilling the evolving needs of the user community, and determining the value of JGI products. Knowing how existing datasets have been used in the past can facilitate their discovery and reuse in the future. Rather than create an entire data ecosystem from scratch, we can nurture one into existence by leveraging existing systems and applying lessons learned from user experience testing and interagency initiatives, such as the DOE and National Institutes of Health Petabyte-Scale Sequence Search, to improve the openness and interconnectedness of JGI data. Successful strategies should be transferable to or implemented by other repositories both within and outside of BER.

Improve Integration with Data Generating Partners

JGI partners HudsonAlpha and Arizona Genomics Institute have been long-standing key liaisons to the JGI, providing users with unique and complementary capabilities not available elsewhere in the JGI ecosystem. These include specialized eukaryotic genome competencies and genomic DNA and RNA extractions. To ensure that the resulting sequence data are easily findable and reusable on the JGI Data Portal, the JGI will work closely with its partners to develop integrated workflows that facilitate data and metadata sharing and more streamlined communications with users.

Enhance Data Interoperability with KBase

Biological data are complex and multiscale. To reason about the influence of microbes on nutrient cycling and link that model through multiple scales to investigate how microbial communities may impact climate change over time, we need a common conceptual model for biology. **Fig. 10** contains an example conceptual model that connects observations to chemical, genomic, environmental, and other relevant processes. Some of these processes have mathematical representations, but others are connected through defined relationships. What this model elucidates is not just descriptions of the data, but also the links between the data and the process that the scientist was investigating, thereby adding crucial context when considering data for reuse. The JGI will work with KBase and the scientific user community to align data to the conceptual model, and associated underlying ontologies, to create a repository of interoperable experimental data (**DS1-5**).

The structure of the repository will support ordered reasoning, where the ordering or relationships among measurements and assertions has semantic meaning. Many modern AI tools rely heavily on that ordering to perform effectively. For example, LLMs rely on the highly

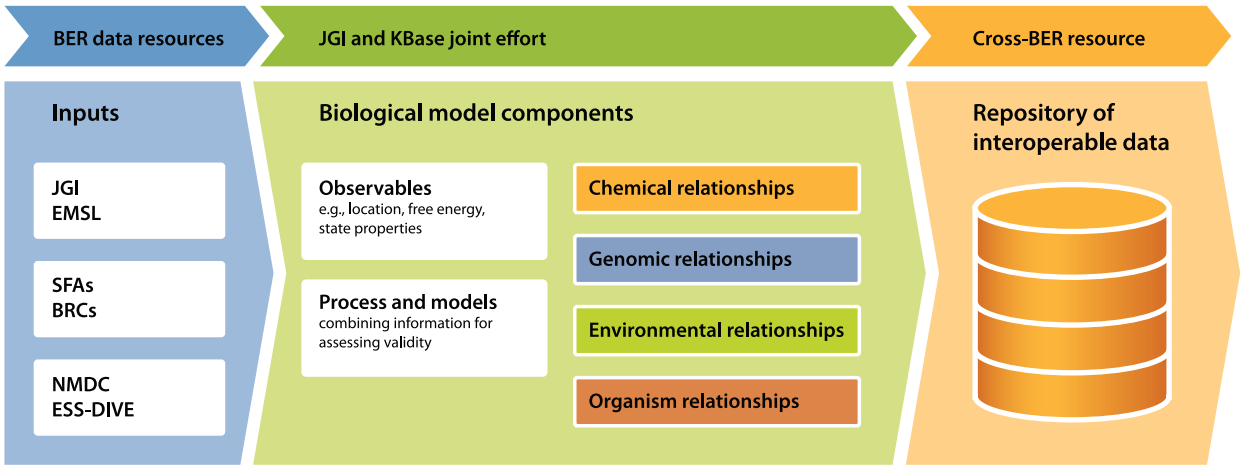


Fig. 10. Enhancing data interoperability with KBase. Components of a conceptual biological model being developed by the KBase team and scientific community that can be used to relate data from different experiments by connecting observations to the underlying process that was investigated. A collaborative harmonization effort will generate a repository of interoperable data. The JGI and EMSL are user facilities that generate more public data and metadata than the others and will serve as foundational elements to this effort. NMDC and the Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) are data aggregators and metadata generators. Science Focus Areas (SFAs) and BRCs represent experiments integrating diverse scientific data.

structured semantic ordering of text or sequence data to generate meaningful responses. Encoding data in this way ensures good harmonization, better human understanding, and a powerful organization of data such that these AI systems can better perform and produce interpretable answers.

Create an Advanced Search Interface that Links National Repository Identifiers to JGI Identifiers

The JGI will build upon our existing collaborations with the computational biology community to incorporate the Petabyte-Scale Sequence Search[29] efforts into our software infrastructure. We have a strong foundation with the Data Citation Explorer, GOLD, NamesforLife intellectual property, and our data management system that, when combined with our user community, creates a unique opportunity for the JGI to develop the core of a genomics-linked data ecosystem that overcomes the barriers within BER and creates reliable, traversable bridges to large-sequence repositories and publication repositories outside of BER.

Co-develop a BER Metadata Submission System

The heart of linked data is the sample the data came from, and multiple data types can be generated from a single sample. The JGI has partnered with EMSL through the FICUS program to produce multi-omics from the same physical samples for years. Increasingly, scientists want to be able to link these data to sensor, simulation, or other geolocation-based data generated by other projects at the same site. While doing this linking for a given project is possible, it is a time-consuming process because samples are not identified or tracked the same way across experiments. Addressing this core issue eliminates a major challenge in scientific data management and has the potential to improve reusability because data from different samples can be linked through a common globally unique sample identifier. ESS-DIVE has been leading this effort for environmental samples, while NMDC, EMSL, the JGI, and KBase have been collaborating to address this for biological samples.

A common sample identifier is the first step to collection of more robust metadata through systems that can offer the same user experience. Momentum is growing for the use of the underlying technology for the NMDC submission portal, which relies on standardized templates and will be adopted for JGI microbiome samples (**DS3-2**). To move toward standardized metadata submission for JGI users beyond microbiome studies, the JGI will support the development of this same toolkit across the large portfolio of JGI sequence and metabolomics products (**DS3-5**). We will collaborate across BER facilities to build resources at our respective organizations that have the same look and feel, so a user

29 https://ncbiinsights.ncbi.nlm.nih.gov/2021/12/17/psss-codeathon-2021

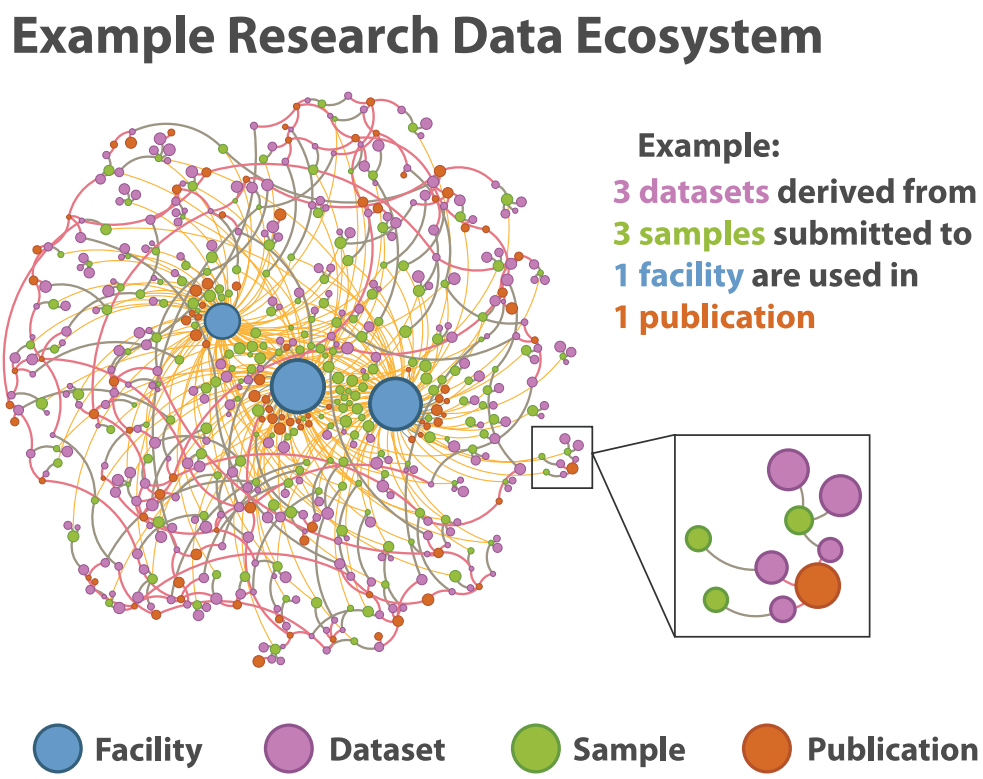


Fig. 11. Links in a research data ecosystem rely upon identifier citations. Citations are crucial for connecting samples, datasets, and publications to create a richer understanding of facility impact. Citations of persistent unique identifiers connect publications back to the facility or project that generated the data and highlight the impact of the data, samples, and facility. Developing an understanding of these links is a critical part of JGI ongoing efforts to understand our impact.

can enter and manage their metadata anywhere in the same manner. Once the user has completed their work, they will be able to return to the metadata system to submit their project data to national repositories and obtain the identifiers required for publication.

Contribute to a Common Search Entry Point for BER Data

The power of LLMs like GPT has shifted the way we can approach linking data systems across BER resources. These resources can be leveraged to build intuitive query interfaces to metadata repositories. While users would still be expected to examine results critically to decide whether they are relevant for their work, we aim to enable querying via natural language. For example, a user could simply write "show me all metagenomes, single-cell genomes, and isolate genomes from soil samples across North America that have paired metabolomics data" into a single text prompt analogous to chatGPT and receive results for further exploration of the data (**DS4-2**). This approach would improve the experience for novice users by reducing the burden of learning a particular search interface. It would also reduce the time from asking a question to receiving results for all users and especially for novice users. Several resources could be used to prototype and expand this type of search capability, including the JGI, NMDC, KBase, and ESS-DIVE (**DS4-5**). Collaboration beyond the JGI may require additional resourcing to fund collaborative effort on other projects.

Automate Discovery of JGI Data Use and Citation

In the absence of data provenance and proper citation of data in the literature, specialized systems must employ heuristics based on citation surrogates such as accessions, metadata, and practitioners (authors and institutions) to infer the use of specific datasets by published research. The JGI has the Data Citation Explorer resource and the expertise to expand the connections between data cited in publications to better understand data reuse and our evolving user

community. An automated process could generate publications that could be linked back to JGI search results and create positive feedback loops for consistent citation of a scientist's work. We will create an interactive interface for program managers, scientists, and JGI staff that aids in visualizing connections between proposals, publications, data, and reuse by the community. The aggregated information can help us better understand the impact of the JGI and multi-omics research, as well as support our efforts to connect with scientists who use JGI data but are not necessarily directly connected to the JGI (**UP-2**).

Strategic Objective 2: Understand the JGI User Communities to Enhance Accessibility and Diversity

A JGI key asset is its user community, comprising a broad range of researchers and trainees who use JGI resources and expertise. These collaborations have yielded numerous, high-quality outputs, such as data, publications, and software. On an annual basis, the JGI serves more than 2,000 primary users, who gain access to JGI capabilities through the JGI user programs. While most users are from North America, they are found across the world. Most JGI users are from academic institutions, with an increasing proportion of them from MSIs. In addition, the JGI serves a secondary user community of tens of thousands of users, which consists of direct downstream users of public JGI data, systems, and tools. Despite this broad user base, ample opportunity exists to further diversify the JGI user community.

To broaden our impact and reach as an SC-supported user facility, the JGI will begin to identify, characterize, and understand the part of the research community not yet leveraging our resources or expertise. By doing so, we can identify challenges and barriers specific to those potential users, which will lead us to set criteria for removing or lowering these barriers. The insights gained from better recognizing and understanding changes in the needs of the user community can be used to inform the planning, development, and refinement of user-facing services.

For example, recent user interviews have established a desire within the community to identify potential collaborators for community science proposals. These collaborations can be encouraged and facilitated via improved user-facing features on the JGI Data Portal and JGI web site. Furthermore, several JGI users who have had multiple successful proposals have generously volunteered to act as mentors for principal investigators (PIs) from MSIs whose JGI CSP or FICUS proposals have not yet been successful. The JGI will facilitate this mentorship initiative, serving as a matchmaking entity to pair the user volunteer mentors with prospective JGI users, thereby assisting them in achieving successful new proposal submissions. These efforts will benefit JGI users and support the goal of having proposals with greater numbers of participants, while at the same time promoting access for new investigators and MSIs (**UP1-2**).

To further diversify our user community, we will seek to engage new users from other scientific disciplines through cross-facility collaborations. In 2013, to harness the combined unique capabilities of the JGI and EMSL, both facilities initiated a joint access program through a single proposal call, the FICUS call. This annual call has since become a crucial component of the JGI User Program portfolio and has expanded to include additional complementary resources. This includes access to the Biological Small-Angle Neutron Scattering instrument at the Center for Structural Molecular Biology at Oak Ridge National Laboratory and, more recently, the Advanced Photon Source at Argonne National Laboratory. Furthermore, access to samples and specimens from terrestrial and aquatic sites is provided through NEON. Over the next five years, the JGI plans to further evolve these cross-facility collaborations through joint user program calls (**UP1-5**). The inclusion of additional light source user facilities could introduce new structural biology-oriented tools to further complement JGI DNA synthesis and other functional genomics capabilities. Another natural partnership in the FICUS-like model would leverage the M2PC at EMSL. TMF at Berkeley Lab is an additional candidate for such joint initiatives, as further discussed in **Biomolecular Materials**. These cross-facility collaborations effectively create a "one-stop shop" for users for cutting-edge, multidisciplinary DOE-focused science.

These activities for broadening our user base will need to go hand in hand with formalized agreements with non-DOE funding agencies. A memorandum of

understanding (MOU) was recently signed between the Biological and Physical Sciences Division at the National Aeronautics and Space Administration (NASA) and BER. This MOU establishes a framework for cooperation between the NASA-funded GeneLab, NMDC, KBase, and the JGI, marking a significant first step into this direction.

Understand Our Users by Characterizing the Current and Possible JGI User Population

Analytics are at the core of understanding the JGI user community, and collecting appropriate data is the enabling first step of performing useful analytics. We will use analytics information to improve the user experience for our systems, as well as to inform the prioritization for community engagement and development of new services or features. The JGI has started to analyze available user data to characterize user relationships, demographics, and connections between proposals and downstream outputs. We have realized there are gaps in our understanding that we can address through additional data collection and analysis. To best leverage JGI publication data, we will improve our publication database through migration and integration with other JGI systems (**UP2-2**). Importantly, we will investigate means for leveraging publication data and other linked data for understanding the JGI primary and secondary user bases and characterizing their activities (**UP2-5**). Additional effort will be focused on user metadata and creating links between JGI systems to reduce ambiguity in the analysis. We will also engage directly with the community through user interviews and surveys to collect information from people who are either no longer JGI users or have never used the JGI (**DS2**). This information will help the JGI form a strategy for diversifying and engaging the broader community.

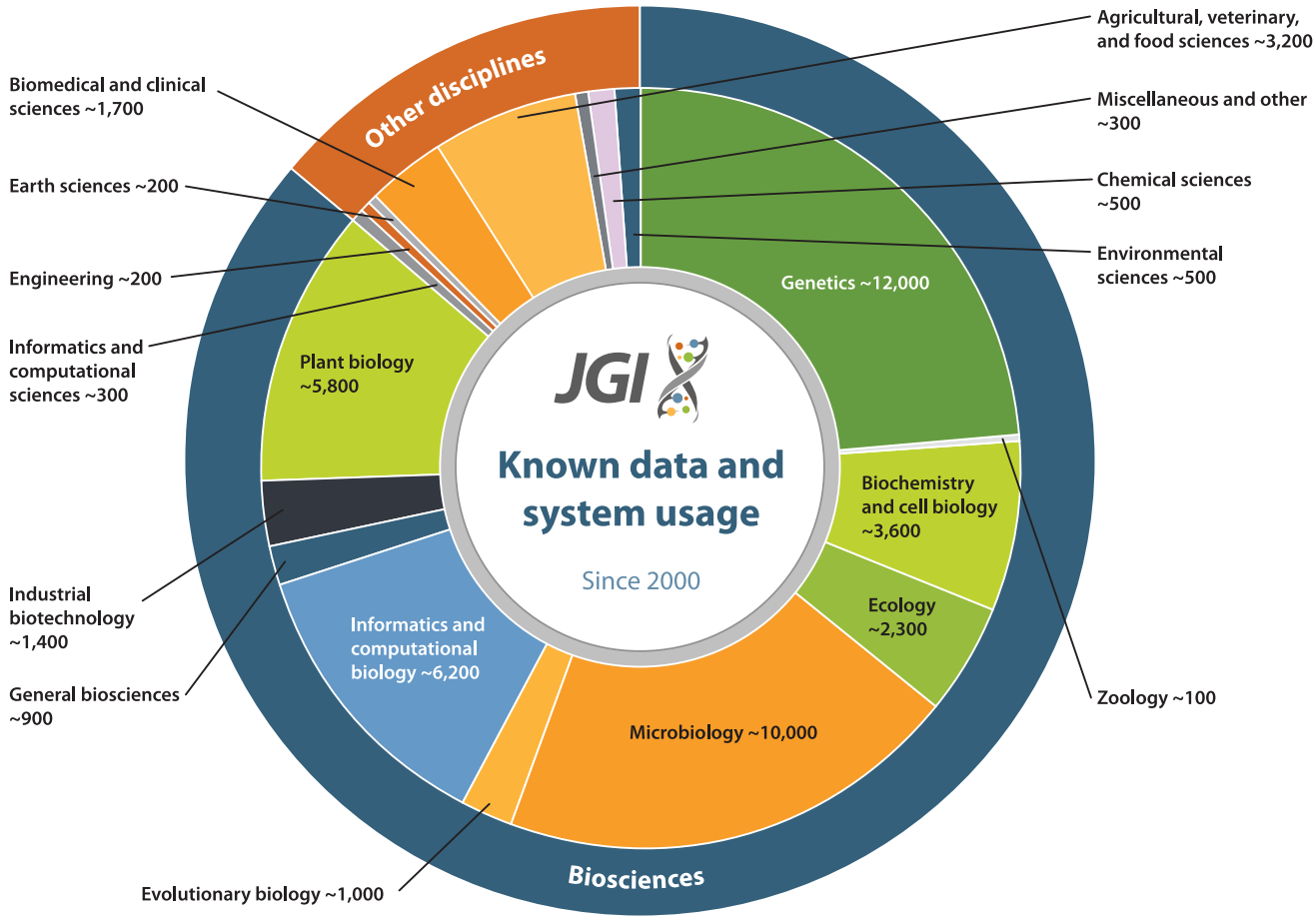


Fig. 12. Research areas of the approximately 29,000 secondary user studies enabled in part by public JGI data and data systems since 2000. Individual studies can belong to multiple research areas.

Strategic Objective 3: Scaling and Optimization of Processes and Data Generation

The demand from JGI users for larger amounts of sequencing continues to grow. In addition, many users rely on a multifaceted combination of genomic, transcriptomic, proteomic, phenotypic, metabolomic, imaging, sensor, and other data produced at scales amenable to statistical analysis. To meet these needs, the JGI will deploy state-of-the-art sequencers, expand current capabilities in strain engineering and functional assays, support new methodologies such as single-cell transcriptomics, and enhance the integration of various data types.

Scale Nucleic Acid Sequencing

The JGI processes approximately 30,000 samples per year for sequencing and is already onboarding next-generation long- and short-read sequencing platforms to support even broader access to these capabilities. Recent instrument upgrades have decreased the per-base-pair cost of both short-read and long-read sequencing by nearly a third. However, with this great opportunity to scale sequencing output come major requirements for workflow upgrades. The next-generation JGI Sequencing Platform will undergo a transformation in how we handle samples, build corresponding libraries, load sequencers, and ultimately track and computationally process the resulting output. We have phased goals to increase output 30 to 50% in the next two years (**GT1-2**, **GT2-2**) and more than double the capacity in the next five years (**GT1-5**, **GT2-5**). These developments will go hand in hand with the JGI aim to evolve its user programs by shifting toward larger-scale community proposals (**UP1-2**). Community proposals of a large scale and multifaceted nature will allow the JGI to tackle big questions in the DOE mission space and keep the JGI unique and desirable. For example, it will facilitate massive-scale efforts, such as time-resolved metagenomics in Long-Term Ecological Research (LTER) sites.

Generate Single-Cell-Resolved Atlases

Complementary to processing more samples is the development of new sequencing products for DOE users. During the past strategic plan implementation, a major effort was focused on single-cell transcriptomic technologies. This approach moves research beyond whole tissues to the ultimate level of the biological unit of life: the individual cell. Moving forward, we will focus on an extension of the JGI Plant Gene Atlas,[30] which was generated from bulk tissues and will next be extended to single-cell data generation (**GT3-2**). Importantly, we will also have a related research and development target of spatial transcriptomics, which will not only provide single-cell RNA profiles of cells but will also resolve them spatially through molecular and imaging approaches (**GT4-2**).

Use Functional Microbial Approaches

Microbiome research remains a critical area of JGI focus with now-routine capabilities to generate genome references for both cultured and uncultured species. New opportunities exist to link uncultured microbes to their cognate cells through molecular and imaging techniques, which will be a new area of study for our institute (**GT5-2**). In addition, the JGI helps users gain functional insights into microbiomes through SIP metagenomics, a sequence-based approach to identify metabolically active species living in complex communities. As part of our strategic plan, we will scale up this unique and sought-after capability to meet user demand (**GT6-2**).

Integrate Experimental Data

Technological advances are also providing entry points for integrating experimental data (commonly referred to as multi-omic or multimodal studies). For example, a user may be interested in not only having genome and transcriptome data from a given sample but also complementary information, such as data describing the epigenome, proteome, metabolome, or microbiome of the same sample. These products exist through the JGI or partnerships such as FICUS. Our next goal is to focus on their better integration (**GT7-2**) with a stretch goal to develop similar products but at single-cell resolution (**GT7-5**).

30 https://plantgeneatlas.jgi.doe.gov

Expand DNA Synthesis and Strain Engineering

As part of the implementation of the previous strategic plan, we expanded our DNA Synthesis Platform, and we plan to continue this scaling. An important application will be to support end-to-end efforts in plant engineering where construct design and manufacturing can be performed at the IGB with existing relevant expertise and infrastructure (**GT8-2**). Further targets will be growing reagents for CRISPR screens to either activate or repress genome-wide catalogs of genes to determine their functional properties (**GT9-2**). Additionally, efforts will continue to expand user access to engineered strains as an orthogonal approach to link sequence to function (**GT10**).

Diversify Metabolomics

The Metabolomics Platform was founded as part of the implementation of our previous strategic plan and continues to grow as a user capability. We currently offer untargeted analysis for polar and nonpolar metabolites, with a targeted analysis product for polar metabolites available as a service for users. Our goal in the next two years is to formally launch a targeted analysis product for nonpolar metabolites as well (**GT11-2**). In addition, we will continue to improve the performance of our nonpolar metabolomics capabilities for lipids relevant to energy, environmental biology, and the bioeconomy (**GT12-2**). Finally, the JGI will continue to expand annotations of metabolomic standards to increase the breadth of molecules identifiable for users (**GT11-5**).

Strategic Theme 4:

Stewarding Resources

Enhance JGI Impact through Nurturing Its People, Systems, Processes, and Communications

Background

The DOE established the JGI more than two decades ago to advance genomic contributions to a national effort. That mission remains unchanged even as the institute's capabilities offered to users have evolved and expanded in response to scientific and technological advances. Similarly, the heart of the JGI remains its people, even as workforce skills and support structures have adapted to maintain and increase the core services that keep the JGI operational. The JGI must continually foster all components of its infrastructure to ensure optimal impact for stakeholders and succeed in its vision to support innovation for a sustainable bioeconomy.

The JGI must have a competent, diverse, and dynamic workforce with an eye on consistent improvement in all areas. The JGI must strategically evolve its workforce to meet its short- and long-term goals. The institute will accomplish this through strategic hiring for future needs, community efforts, and especially development opportunities for current staff.

The work done by JGI staff in collaboration with JGI users is translated and shared through various communications and outreach efforts, amplifying the JGI impact. A strategy for telling the JGI story involves enhancing and focusing messages to those who already know the JGI, expanding reach to communities who can benefit from the JGI but are not yet doing so, and enabling an ecosystem of advocacy among JGI stakeholders to optimize the conduits for showcasing its efforts and to justify its existence and federal funding.

Finally, a network of systems and processes provides the underlying structure that enables JGI day-to-day work and discoveries. In an era of escalating costs, unpredictability of funding and budgets, and increased scrutiny and regulation, the JGI must strategically evaluate, re-engineer, and manage the effects of changes to its existing business and other processes to maximize the impact of every dollar of funding granted.

Opportunities

The JGI is constantly balancing making key contributions to science in support of a sustainable bioeconomy with efforts to improve efficiency in existing processes and systems to counteract increasing operational costs. Upskilling and education of staff to support new capabilities for our users and targeting communications to encourage researchers to partner with the JGI and the broader community to advance their science are ways the JGI maintains this balance. Additionally, as part of Berkeley Lab's Biosciences Area, the JGI and its staff have opportunities to unite with our colleagues and work together to further develop scientific endeavors to support the bioeconomy.

The JGI talent management team will continue to provide opportunities to evolve the workforce. We will conduct a review of current and future programmatic needs and goals to assess our existing staff, identify talent gaps, and align our workforce with current and future demands. In addition to developing our existing staff, we will identify the mix of future talent required to match programmatic deliverables. This process will involve identifying the types of employees needed (career, term, limited, contingency staff, etc.), and laying out the timing for these staff needs. This knowledge will provide the foundation for attracting and hiring new staff to fill talent gaps and allow us to proactively plan for attrition and knowledge transfer across the JGI.

Strategic Objective 1: Evolve the JGI Workforce

In 2017, we launched the OurJGI culture initiative (**Fig.13**) to support the scientific and technical strategy of the JGI.

This initiative includes bottom-up and top-down efforts and activities and aligns with Berkeley Lab's emphasis on the stewardship of people and resources, including funding and property.[31] JGI culture will continue to evolve alongside the JGI strategic plan. The organization and its users perform research of the highest scientific and ethical quality, make data available to the broad scientific community, and adhere to the FAIR data sharing principles. We thoughtfully manage and lead our talented staff, who are central to achieving our vision, ensure safe working environments and processes, and foster a culture of respect and collaboration for everyone in our community. The JGI is committed to advancing the principles of inclusion, diversity, equity and accountability (IDEA), exercising the highest standards of financial accountability and transparency, and responsibly managing its infrastructure and assets.

Collaborate with Berkeley Lab

The JGI location on the Berkeley Lab main campus offers opportunities to collaborate and coordinate with Lab-wide efforts to adopt best practices and approaches to bolster the JGI workforce. These activities include increased interactions with Berkeley Lab's Inclusion, Diversity, Equity, Accountability (IDEA) and Learning and Culture Offices. Through targeted communication with JGI staff, we will ensure that opportunities offered by these Lab-wide offices are available for everyone (**OP1-5**, **OP2-5**).

Fig. 13. Growing OurJGI culture and workforce objectives.

Develop the JGI Workforce

We will bolster development and mentoring opportunities and seek out opportunities for upskilling (learning to improve current work), reskilling (learning to do new types of work), and cross-training. Examples of areas in which we expect an increased need for staff expertise include AI, advanced statistical analysis, and AI-optimized study design (**OP1-2**). We will evaluate and balance staffing resource needs when onboarding new programs and projects. We will also expand our recruitment and outreach efforts and partner with Berkeley Lab's outreach team and Biosciences Communications to identify career fairs and student learning opportunities. We will bolster recruitment advertising to new spaces in all recruitment activities (**OP2-5**).

Foster Opportunities for Engagement, Collaboration, and Connection

The JGI will expand and create hybrid venues and forums for future workforce engagement as a significant proportion of our workforce continues to work remotely or in hybrid work arrangements. We will continue to use the JGI IDEA engagement survey to gauge employee satisfaction and identify areas of improvement to address topic areas that scored unfavorably (**OP2-2**, **OP2-5**).

We will introduce new mechanisms and expectations that JGI staff will devote a portion of their time to larger strategic institutional initiatives and priorities (e.g., accelerating workflows, development of a new laboratory information management system [LIMS]) in addition to their departmental, platform, or program responsibilities. This strategy will provide new opportunities for individual employees to have a greater impact within the organization by participating in new cross-department teams and accepting new leadership roles (e.g., product manager, technical lead; **OP1-5**).

31 https://stewardship.lbl.gov

Strategic Objective 2: Share the JGI Story

The Communications and Outreach team engages a range of audiences, including potential users, academic partners, DOE stakeholders, and the public. The team leverages a portfolio of products to do so: the website, featured science stories, podcasts, videos, newsletters, and social media. These products showcase the breadth of capabilities available to users, primarily by highlighting successful proposal outcomes that draw attention to opportunities such as the CSP and FICUS. These stories demonstrate the value of cultivating partnerships and how successive studies build on foundational efforts. The editorial strategy incorporates diverse voices, including early career and senior scientists. Storylines reflect both the lab work and computational analyses applied toward user research. Measuring how audiences engage with these stories on various platforms helps develop content targeted toward unique stakeholder behaviors.

Improve the JGI Institutional Website

As a communication tool, **https://jgi.doe.gov** should be effective as a critical point of contact for potential users. It also serves as an important source of information for existing users, academic partners, and other stakeholders, including those at the DOE.

To continually serve the user community and reach a wider pool of researchers, **https://jgi.doe.gov** will be scalable and accessible enough to meet visitors where they are: desktop, mobile, or tablet. After updates to its messaging, navigation, and infrastructure (**CO1**), the JGI website will better align with the needs of these diverse primary audiences. Even a brief visit to the site will convey to visitors that the JGI supports user science with innovative capabilities. Design, text, and navigation will engender visitors' confidence in both JGI activities and in their own ability to leverage JGI resources to advance research in the energy and environment landscape.

Grow the Cohort of JGI Ambassadors

While the Communications and Outreach team is the vanguard, its members are not the only storytellers at the JGI. Each staff member is a JGI representative; every time they answer the question, "What do you do?" within their social circles, they relay their understanding of the JGI work culture and their evaluation of the JGI as an employer. Post-pandemic, transitioning from an all on-site workforce to hybrid teams means JGI representatives are farther flung than before. The geographic distribution of our workforce offers new possibilities for increasing hiring diversity, outreach, collaboration, and awareness of JGI services and resources. The JGI is cultivating a new cohort of staff as representatives who can serve as ambassadors at events ranging from facility tours to career panels and conference booths when the need arises (**CO2-5**).

JGI users are proven, effective advocates for partnering with the facility. A more concerted effort to increase the efficacy of the JGI User Executive Committee members as JGI advocates is in development (**CO2-2**).

Develop the Next Generation of Colleagues and Collaborators

Over a decade, the JGI-University of California (UC) Merced Internship Program has slowly but steadily grown from two UC Merced graduate students with two JGI mentors into a group of undergraduate and graduate students who now comprise half of the JGI annual summer intern cohort with nearly a dozen mentors.

The success of the JGI-UC Merced program has encouraged the JGI to cultivate a similar long-term partnership with the Alabama Agricultural and Mechanical University (AAMU). AAMU faculty at the Agricultural Research Station have sent interns to nearby HudsonAlpha Institute for Biotechnology (a JGI partner) and run feedstock crop projects in collaboration with the Great Lakes Bioenergy Research Center (GLBRC, **CO3**). The DOE launched the Justice40 program in 2021 to bolster underrepresented minorities in science, technology, engineering, and mathematics (STEM) careers. The funding initiatives from Justice40 programs underlie the continuing evolution of the partnership with UC Merced and the nascent collaboration with AAMU.

JGI outreach efforts extend beyond recruiting and training interns by pairing them with staff. Efforts to broaden the data user community continue through workshops that provide attendees access to the data portals and training in the available analytical tools so they can incorporate the information into their curricula

and, by extension, amplify this knowledge by training additional people at their respective home institutions.

For example, JGI staff have conducted workshops aimed at California State University (CSU) system faculty educators. These workshops familiarized nonexpert faculty with the suite of comparative genomics tools available through the IMG/M data portal and discussed ways to develop lesson plans around these topics. CSU now offers a handful of bioinformatics courses at multiple campuses. Based on these discussions, the "adopt-a-genome" initiative has emerged involving faculty-led student groups analyzing microbial genomes. An initial cohort of CSU faculty is currently piloting this effort. The publication of genome reports will provide students first-hand experience with the peer review process. The JGI will expand the project's scope as more faculty become involved, drawn in by the success of the initial group. We plan to foster and support these activities as they generate crucial manual assessments of genome quality, enhance the visibility and utility of individual "orphan" genomes, and boost JGI data utilization metrics.

A longer-term approach could seek to expand workforce development by incorporating JGI resources and programmatic efforts into curricula across K–12 and university education programs. An exemplary effort is an existing partnership with the Tiny Earth project, in which college students enrolled in a Tiny Earth research course discover antibiotics from soil bacteria in their backyards through activity screening, culture isolation, sequencing, and metabolomics. Today, the Tiny Earth network has 14,000 students participating annually from 517 global institutions and has successfully isolated 19,000 microbes with 125 genome sequences completed by the JGI. Similar JGI partnerships with projects and educators will enable broad and diverse educational experiences for students.

The JGI is also looking at developing the next generation of genome scientists abroad through MOUs with institutions including the Nara Institute of Science and Technology in Japan (NAIST) and the University of Galway in Ireland. Through such partnerships, JGI researchers aim to cultivate training and internship opportunities and collaborate with faculty members on shared research topics of interest, expanding the JGI user community.

Strategic Objective 3: Maximize JGI Efficiency

As a federally funded institute, the JGI must maximize the use of every dollar granted by U.S. taxpayers. The scale of the JGI is large compared to other DOE-funded enterprises; this minimizes resource constraints that other smaller projects may have. However, resource constraints still exist, including relatively flat funding and a fixed amount of lab space. In addition, specific challenges are associated with working within a national lab ecosystem, such as relatively high indirect costs to ensure full cost recovery and the complexity of the combined UC and DOE policy frameworks, which can increase the cost of business processes. Maximizing the efficiency of existing processes and systems is mandatory if we are to maintain or increase our scientific output in the face of rising labor and material costs.

As a national laboratory, we serve a broad spectrum of scientific disciplines across a wide range of funding levels using systems built in ways that require continuous optimization to best serve our user community. The JGI must constantly evaluate its operational needs and create or modify processes to minimize redundancies. As the JGI evolves, it must adapt the fixed physical space in the IGB to accommodate changing workflows and requirements. Logical and well-planned workspace evolution is essential for supporting the institute's continued growth and productivity.

Leverage Existing or New Business Systems to Eliminate Redundancy

Many JGI processes and systems evolved out of the need to operate autonomously when the institute was located off the Berkeley Lab main campus. Our central location on Berkeley Lab's main site now allows direct access to additional systems and processes, meaning many JGI internal processes and systems can be updated to leverage this proximity. Through the evaluation and careful evolution of JGI systems and internal processes, we can reduce the time and effort involved in completing business processes (**OP3-2**, **OP3-5**).

Currently, the JGI employs a home-grown procurement process to track various types of orders. Members of the JGI Operations Team will critically evaluate the end-to-end acquisition process, assess

alternative systems, consider leveraging Berkeley Lab systems, and re-engineer the procurement processes and systems as necessary (**OP3-2**). Additionally, the team will identify other systems or processes at the JGI that existed based on the remote nature of the prior Walnut Creek facility. Once these are identified, we will evaluate and prioritize opportunities to update them to fit our current needs, or leverage Berkeley Lab or other off-the-shelf systems to reduce double work or other inefficiencies (**OP3-5**). Such systems might include the JGI stockroom and the administration of travel services for the JGI staff.

Rethink the Budget and Planning Process

As labor and materials costs increase, and become a larger proportion of the JGI's overall cost profile, it is necessary to more critically evaluate the existing budget structure and make near-term decisions to help alleviate longer-term potential budget issues. We will adjust the JGI annual budget and capacity planning process to require budget holders to more actively monitor and manage their costs and provide options to JGI management on how to most effectively use the budget allocations (**OP4-2**).

This process will involve a re-evaluation of the components of the JGI budget at the group level that are essential to monitor, measure, and make decisions on. Using that evaluation, modifications will be made to the budget process that focus on those components and involve the group leads taking an active role in prioritizing and making decisions on what is included in the budget (**OP4-5**). This activity will inform updates to the process by which group leads monitor and report quarterly on performance to budget.

Evolve the JGI's Space to Accommodate the Strategic Needs of the Institute

The COVID-19 pandemic changed how space is used globally, not just at the JGI. The evolution of a more hybrid environment for those whose work does not require a full-time presence in the building has resulted in the availability of additional office and cubicle space. Moving forward, we must evaluate the lab and office space profiles of the IGB and make necessary plans and investments in the building to accommodate the needs identified throughout this strategic plan.

We will evaluate and invest in necessary facility modifications to match the laboratory, office, and conference space needs of the IGB to support the new working environment. Consideration for potential modification will include post-pandemic response, additional non-JGI residents in the building, and the evolution of the JGI workforce profile (**OP5-2**). With cross-functional teams, we will also evaluate the profile of technical space in the IGB and other JGI facilities and forecast the needs based on changes in process, technology, and science to proactively alter space or allocation methodologies to ensure efficient use of limited space (**OP5-5**).

Strategic Objective 4: Amplify the JGI's Impact

In addition to the efforts mentioned previously, the JGI can further amplify its impact through mechanisms involving coordination across groups within the institute. By leveraging the expertise of experienced staff across the organization (some with over 20 years of experience serving a diversity of users and scientific communities), it is possible to realize impact amplifications with minimal additional monetary investment. Engaging temporary outside expertise or creating task forces or working groups that cross organizational boundaries will help us achieve these results.

Further Develop Industry Partnerships

The JGI engaging with industry partners on projects relevant to the DOE mission has many benefits. These collaborations provide JGI staff with perspectives from the more applied industry research, helping connect the basic science supported at the JGI to the broader bioeconomy. At the same time, the unique capabilities and expertise of the JGI benefit our industry partners through unique discovery that might otherwise not have been possible, potentially exposing the JGI to a broader audience and a new user base.

To achieve the amplified impact of these partnerships, the JGI must refine how projects are tracked and managed through the various pipelines. Sponsored projects have more stringent time, financial, information security, and tracking constraints, and thus these projects may need to be monitored more closely throughout the process. This additional monitoring will

require identifying critical components of the JGI sequencing and data pipelines for which increased visibility and control would be necessary to accommodate and attract industry projects. The JGI will determine how to meet information security needs within its infrastructure and data policies and devote the personnel and resources required to implement those changes to the JGI systems (**OP6-2**).

To increase the effectiveness of our industry engagement efforts, we will create a standard operating procedure for forming partnerships with potential industry partners that rely on the JGI's normally user-focused processes and resources. While every industry project will have unique considerations and needs, developing such a framework is expected to minimize the iteration of plans once a potential project is identified. To shape the initiation and execution of future projects, we intend to collect information and data from completed industry partnerships and use it to guide further improvement (**OP6-5**).

Seek Automation Solutions for the Full Breadth of JGI Processes

To continue to increase productivity and broaden the portfolio of JGI services, it is essential to continuously seek opportunities to automate tasks throughout JGI processes, whether lab based or computational. The JGI will create a cross-functional group of staff members to evaluate and provide recommendations for improvements to JGI processes to achieve the goal of increased productivity.

We will establish a cross-functional team, potentially including an external observer or consultant, to design a system for evaluating processes and systems at the JGI and identifying targets for reengineering. We will then create a prioritization scheme to determine the most impactful changes based on metrics related to cost, efficiency, throughput, or customer satisfaction (**OP7-2**). Finally, we will implement an organizational change management process that will enable proper preparation for and communication of significant changes to the staff impacted by the changes (**OP7-5**).

Enhance "Inreach" to JGI Staff to Connect Daily Work to the Mission

An informed staff enables teams to work together more effectively to achieve the organization's goals and better understand how individual tasks connect to the overarching mission. Multiple communication streams are available but not all of them are used effectively: information can be overlooked when delivered through parallel and repeated communications or in long communications with too much information.

We will assess current communications to JGI staff, both internal to the JGI and from the institutions that employ JGI staff, to assess the level of overload and identify the different paths these communications follow. Working with a representative cross section of JGI staff, we will identify the areas where the most gaps exist and receive feedback on ways to close those gaps (**OP8-2**). This process has already begun but will require a sustained effort moving forward.

We will also use due diligence, with a focus on not increasing anyone's long-term workload, to identify communication channels or resources that should be supported, as well as those that can be eliminated or replaced with new ones that close the gaps identified in the assessments. Finally, we will establish a plan and process for routine review of channels and resources for communication to internal JGI stakeholders and updates to those resources to better connect information to intended audiences (**OP8-5**).

Strategic Initiative A:

Biomolecular Materials

Background

Materials from Biology

Biomaterials are materials produced by organisms, such as plants, fungi, algae, animals, and microbes, which are or can be used to make macroscale polymers and composite materials, such as plastics, rubber, wood, coatings, and fabrics or nanomaterials with defined properties, including bioadhesives and magnetosomes. This definition includes recombinantly produced materials and precursors, which have become possible with advancements in genetic engineering. We also include biominerals that can be used in the replacement of materials currently made with unsustainable practices, such as in commercial cement production, and natural biomineralization processes for patterning, templating, and self-organization, as exemplified by diatom silica frustules.

Understanding the Synthesis and Control of Biomaterials for a Profitable, Secure, and Environmentally Sustainable Bioeconomy

Evolution has enabled biological systems to explore various biomolecular material synthesis landscapes. The resulting biomaterials constitute the bulk of the biomass on the planet, are synthesized under physiological (relatively mild) conditions, and exhibit extraordinary properties that are, in most cases, unmatched by synthetic materials. Interest in such materials goes beyond the unique properties or applications of the material. As most materials of biological origin are carbon and/or mineral rich, they also play an important role in sequestration of carbon and nutrient cycling on the planet. They act as concentrators and reservoirs, accumulating carbon and nutrients, and regulate their release back into the environment.

Dual-purpose technologies are desperately needed to supply economically viable products while mitigating anthropogenic climate change and environmental degradation. As an example, atmospheric carbon dioxide (CO_2) levels hit a milestone in 2021, reaching a 50% increase over the pre-industrial baseline. Although CO_2 absorbs less heat than other greenhouse gasses, such as methane, it is more abundant, stays in the atmosphere longer, and is responsible for most of the atmospheric heating imbalance. Various man-made carbon capture and storage technologies are available, but they are expensive and inefficient. Biological carbon capture and conversion to bio-based materials has the potential to not only mitigate anthropogenic CO_2 emissions but to also support growth of the burgeoning bioeconomy.

The Systems Biology of Biomaterials

To harness and accelerate the development of various biomaterials, a mechanistic understanding of the molecular underpinnings of how these materials are synthesized and regulated and interact with other cellular processes and the external environment is needed. Due to the expansion of technologies that facilitate genome sequencing from a diverse range of organisms and complex environments, we can now access a rich collection of biomaterial blueprints. However, our ability to read these blueprints is still lacking.[32] A focused effort to understand the genomics and the function of genes and proteins involved in the systems-wide processes responsible for the synthesis of biomaterials is required.[33]

Beyond the proteins and enzymes responsible for catalyzing synthesis, various biological processes are tightly orchestrated to produce biomaterials, and include (1) the transport of needed building blocks and protein cofactors, such as ions and metabolic precursors, (2) a confined space within the cell or extracellular space to concentrate and combine precursors to produce defined properties, (3) the ability to control the growth of the

32 U.S. DOE, 2019, *Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa Workshop Report*, DOE SC-0199, U.S. Department of Energy Office of Science. **https://genomicscience.energy.gov/genefunction**.

33 U.S. DOE, 2019, *Genome Engineering for Materials Synthesis Workshop Report*, DOE SC-0198, U.S. Department of Energy Office of Science. **https://genomicscience.energy.gov/wp-content/uploads/2021/09/GEMS_Report_2019.pdf**.

biomaterial or break it down to prevent unnecessary or potentially deleterious overproduction, and (4) the ability to regulate the cellular production of the biomaterial in response to various nutrients and external stimuli. These processes must now be understood for organisms that synthesize DOE-relevant biomaterials.

A New Partnership with the Molecular Foundry

Biology operates at the nanoscale to catalyze the chemistry and orchestrate the synthesis of not only essential materials, such as proteins and DNA, but also more robust materials like nacre and bone. Innovative nanotechnologies for predictive design, genome and protein engineering, biosecurity, and bioimaging are needed for the discovery, development, and deployment of scalable strategies in biomanufacturing. A substantial amount of work remains to be done to make biomaterial production both effective and sustainable. Genomic information provides a critical foundation to do so, but in isolation is insufficient. Where, how, and when specific functions generate a biomaterial with defined properties is critical orthogonal information required to harness naturally occurring processes in applications. This problem needs to be solved at scale before predictive control over biomaterial production is achievable. Breakthrough inventions will require the united effort of multidisciplinary teams that can seamlessly merge genomics, nanoscale science, and computing.[34] To build these multidisciplinary teams and expand synthesis, production, and characterization capabilities for biomaterials, the JGI will partner with TMF to provide users with important tools in biomaterials characterization (see **Highlight Box: A New Partnership for Biomolecular Materials**).

Alignment with JGI Vision

The goal of the Biomolecular Materials Strategic Initiative is to enable the JGI user communities to make genome-grounded discoveries that supply the bioeconomy with strategies and resources for the sustainable and affordable biomanufacturing of dynamic and complex materials that will address current challenges in energy and the environment.

34 U.S. DOE, 2017, *Technologies for Characterizing Molecular and Cellular Systems Relevant to Bioenergy and Environment*, DOE SC-0189, U.S. Department of Energy Office of Science. **https:// science.energy.gov/ber/community-resources**.

JGI Goals for Biomolecular Materials

To build a scalable and actionable understanding of biomaterial synthesis processes and a usable library of biomaterial-genetic-parts-list for the design of biosystems, the Biomolecular Materials Strategic Initiative aims to address the following challenges:

- Accelerate the discovery of genome-encoded biomaterials.
- Understand the systems biology of biomaterial production.
- Enable the development of living foundries for advanced material biosynthesis, replacing more energy-intensive, less environmentally sustainable, manufacturing methods.
- Establish a Genome-Encoded Biomaterials User Program which will enable users to access multi-institutional capabilities in nanoscience and genomics through a single user proposal by leveraging the FICUS call to seamlessly integrate expertise and instrumentation at the JGI, TMF, EMSL, and synchrotron facilities (**UP1-5**).
- Build experimental workflows that provide current and new communities of users seamless integration of expertise and capabilities at TMF and the JGI to accelerate discovery.

Activities

To capture the complex genomic underpinnings of biomaterial synthesis, understand how these materials can be produced in a more economically sustainable and environmentally acceptable way, and redesign material biosynthesis to achieve tailored properties, the JGI will enable users to perform a wide range of studies:

- Sequence the genomes and transcriptomes of diverse organisms that synthesize DOE-relevant biomaterials or encode critical functionalities for the development of novel biomaterials, such as carbon sequestration, photosynthesis, mineral capture and assimilation, and recycling of critical resources and precursors.
- Understand the complex genome-based dynamics that are responsible for the control and regulation of biomaterial synthesis, such as chromatin structure, protein-DNA interactions, regulatory networks, and localization of pathways and sites of synthesis.
- Explore the diversity of biomaterial biosynthesis and share genome-grounded discoveries and hypotheses. A biomaterials-centric computational platform and collaboratory will accelerate discovery by the community and disseminate genome-based knowledge of biomaterial synthesis and the systems biology of material biosynthesis. This platform will provide a curated repository of parts lists, comparative genomic tools and alignment-free methods for gene function discovery, and functionality for analyzing data and sharing discoveries.
- Experimentally test organism-, process-, and gene-level functions responsible for biomaterial synthesis and control of material properties. The JGI DNA Synthesis Platform is routinely designing and assembling thousands of DNA-based molecules that can be used for experimental validation of predicted functions via appropriate genetics- or biochemical-based assays either at the user's home institution, at the JGI through development of HTP-functional assays, or at our partnering user facility TMF for structural biology, bioimaging, materials characterization, and science-question focused experimentation (see **Highlight Box: A New Partnership for Biomolecular Materials**).
- Engineer diverse chassis for heterologous production of materials. Around 50 diverse microbes, containing universal components, have been generated by the JGI, enabling heterologous expression of synthetic genes and genome editing. We expect this list of available strains to grow further in the next five years (see milestone **GT10**). Future research and development efforts may focus on the identification and rapid prototyping of carbon-negative "living foundries," such as plants and algae, organisms that are of direct relevance to carbon capture and biomaterial production. One area of focus may include the design and assembly of genetic components of relevance to biomaterials, such as organelle-localized signal peptides for engineered eukaryotes, promoter libraries, and experimentally tested synthetic components.

Highlight Box: A New Partnership for Biomolecular Materials

The Molecular Foundry (TMF) is a user facility at Berkeley Lab supported by the DOE Office of Basic Energy Sciences (BES) through its Nanoscale Science Research Center (NSRC) program. TMF provides users from around the world access to expert staff and leading-edge instrumentation to enable the characterization and control of matter at the nanoscale. By providing access to TMF capabilities in biomolecular materials research as part of the FICUS call, users will be empowered to combine state-of-the-art nanoscale materials research at TMF with cutting-edge genomics at the JGI. Like the JGI, TMF is open to any interested researcher through a competitive peer-reviewed proposal process. Users travel to Berkeley Lab where they are trained on-site and work side-by-side with TMF scientists to perform multidisciplinary research beyond the reach of their own laboratory. With TMF's broad spectrum of core capabilities and expertise, users increase the scope, technical depth, and impact of their research.

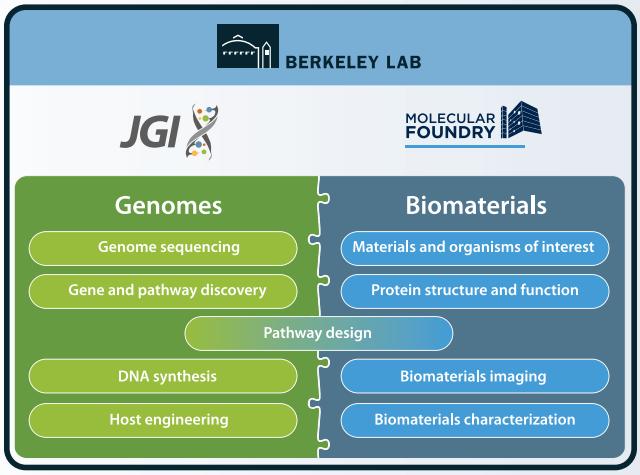


Fig. 14. Partnership between the JGI and TMF.

TMF provides a range of capabilities to users for defining and understanding the emergent nanoscale properties of biomaterials and the biology responsible for synthesis of those biomaterials. A key challenge is that biomaterial synthesis requires the dynamic assembly of proteins to coordinate the accumulation and transformation of dilute precursor molecules and ions into polymers and composite nanoscale materials. The identity and function of these proteins is often unknown, requiring predictive capacity and experimentation to test candidate proteins and decipher molecular-level functional properties. Complementing the JGI's strengths in the genome-based prediction of gene and protein function, TMF has capabilities and expertise for the expression, purification, and characterization of proteins and protein complexes. Tailored biophysical and kinetic assays can be developed with TMF's scientists and combined with structural characterization of macromolecules to understand the function of individual proteins and protein complexes and the materials these synthesize.

Another key challenge is visualizing and characterizing the hierarchical organization of biomaterials across length scales (nanometers to millimeter) and time. Even relatively simple monomers, for example those constituting rolled peptoid sheets or R-bodies, form widely varying macroscopic structures due to slight changes in intermolecular forces in the crystal unit cell at the nanoscale. The co-location of multiple microscopes in TMF provides users the opportunity to leverage multiple imaging modes. Optical microscopy, atomic force microscopy, and electron microscopy may be combined for high-resolution, multimodal characterization of the physical, electrochemical, and mechanical properties of biological materials. Developing HTP optical imaging for subcellular localization of biomaterial synthesis and correlative imaging of soft materials by scanned probe and electron microscopy will be needed to gain a predictive understanding of the connection between genotype and material composition and characteristics.

Strategic Initiative B:

Biosurveillance and Biopreparedness

Background

Importance of Biosurveillance and Biopreparedness

Throughout history, human civilizations have experienced the emergence and spread of infectious diseases, such as bubonic plague, which historians estimate claimed between 75 and 200 million lives during the Black Death pandemic, the 1918 influenza pandemic, which killed roughly 50 million, and the COVID-19 pandemic, which has so far resulted in almost seven million deaths. As the world grapples with climate change, rapid population growth, increased urbanization, and a drive for improved quality of life, the risk of future outbreaks with the potential to severely affect human health, the food supply, and global stabilization increases. The recent COVID-19 pandemic highlighted the need for (inter)national biopreparedness and effective response capabilities to quickly determine threat level, mode of transmission, clinical symptoms, and molecular mechanisms of pathogen invasion and infection that together can inform effective containment, treatment, and prevention strategies. Other examples include foodborne outbreaks, such as the recent *Escherichia coli* infection of California lettuce and animal and crop infections. New pathogens can arise that have not been seen previously, rendering current biodetection modalities ineffective and undermining our ability to rapidly assess whether a substance poses a threat to the host. Increased biosurveillance is required to better anticipate outbreaks and prevent health crises before they begin.

Genomics played a never-before-seen and fundamental role in the COVID-19 response and will likewise be crucial in addressing other emerging pathogens. Genome sequencing revealed the causative agent of COVID-19, SARS-CoV-2, and subsequent emergence of variants of concern. PCR analyses allowed for the near real-time tracking of the spread being extended to environmental surveillance. Genome information is also crucial to guide therapeutic interventions, such as the development of vaccines for SARS-CoV-2, flu, and other viruses, to combat variants that evade the immune response due to mutations. Similar genomics approaches were employed for the recent mpox outbreak and will be a cornerstone of the response to future pandemics.

With future pandemics, the need for rapid and accurate identification of the causative agent(s), followed by comprehensive surveillance, will be crucial to understand pathogens and develop effective spread mitigation and treatment strategies. Genomics will be central to these efforts to identify potential virulence factors, signatures that can be used for detection and diagnosis and in comparative studies to determine pathogen evolution, and for assessing host biomarkers that can serve as hallmarks for infection.

JGI Contributions to Biosurveillance and Biopreparedness

The JGI, as part of DOE SC's National Virtual Biotechnology Laboratory (NVBL), contributed to the COVID-19 pandemic response by conducting several projects, namely (1) studying potential recombination events between SARS-CoV-2 open reading frames (ORFs) ORF7a and ORF8; (2) developing new algorithms for the design of primers with higher specificity that are agnostic to many variants; (3) conducting metagenomic sequencing of the nasal microbiome to identify potential correlations between nasal microbiome composition and COVID-19 disease severity; and (4) identifying quasispecies patterns for SARS-CoV-2 that could influence infection severity. The JGI was also fortunate to receive philanthropic funding enabling the creation of a dedicated laboratory. This laboratory developed pooling protocols and conducted research and development into sample inactivation and more cost-effective diagnosis protocols, as well as new

long-read strain-resolved sequencing protocols for pathogen-agnostic surveillance, detection, and genomic characterization.

Alignment with JGI Vision

The renewed focus on preparing for the next biothreat, whether it be directly in humans, food sources (both plants and animals), or our environment or infrastructure, offers the JGI an opportunity to leverage its capabilities in genomics and data science to address fundamental scientific challenges. The detailed genomic surveillance and detection methodologies applied to SARS-CoV-2 have not yet been widely adopted for use in the agricultural and livestock sectors where they could aid in preventing outbreaks affecting food supply and feedstocks for biofuels and bioproducts. Nor have these methods been fully applied to newly emerging pathogens in ecosystems after changing environmental conditions (e.g., through climate change). Additionally, our expertise in microbial genomics, virus discovery, and metabolomics can be used, for example, to better understand the influence of viruses of microbes on plant-pathogen systems, and characterize the molecular mechanisms associated with these tripartite interactions. Furthermore, JGI expertise in developing and facilitating large-scale data systems for comparative analyses can aid in identifying novel pathogens and understanding their virulence factors, and tracking these pathogens' spread.

JGI Goals for Biosurveillance and Biopreparedness

The 2022 DOE SC virtual roundtable Foundational Science for Pandemic Preparedness identified the most important biopreparedness research areas for future pandemics, including improved understanding of the

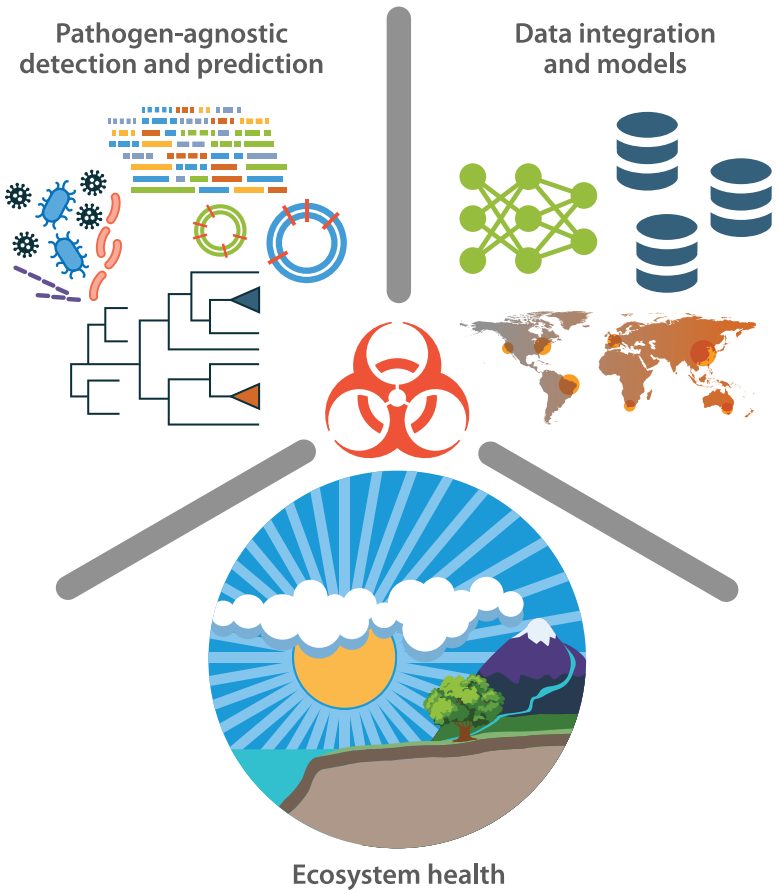


Fig. 15. JGI's biosurveillance and biopreparedness focus areas to identify and mitigate biothreats.

molecular mechanisms that lead to pathogenesis (i.e., the complex physical, chemical, and biological dynamics that occur when a pathogenic microbe interacts with a susceptible host). Through this strategic initiative, the JGI aims to address the following challenges:

- Detect all potential pathogens within complex sample matrices using unbiased and pathogen-agnostic approaches that obviate the need for a priori knowledge of a pathogen's genome. This approach will include (a) detection of low-abundance pathogens that may be highly virulent, using deep sequencing or innovative enrichment methods, such as those that deplete host or other abundant nucleic acids; and (b) identification and characterization of pathogens associated with underexplored environmental vectors.
- Assess the impact of a changing environment (e.g., desertification, melting permafrost) on microbiome resilience and ecosystem health (e.g., introduction of new pathogens, harmful algal blooms) for enhanced biopreparedness (**ML3-2**, **ML3-5**).
- Improve data analysis to integrate sequencing data with phenotypic data, sensing data, and imaging data with comparisons to high-quality reference datasets that, together, can be used to generate models of pathogen emergence, spread, and disease severity, and share these data broadly and rapidly such that other entities can make use of them (**MC4-2**).

Addressing these challenges together will contribute to One Health, an approach calling for "the collaborative efforts of multiple disciplines working locally, nationally, and globally, to attain optimal health for people, animals and our environment," as defined by the One Health Initiative Task Force (OHITF).[35] These efforts also align with DOE's Biopreparedness Research Virtual Environment (BRaVE) initiative, which seeks to provide the underpinning science to enable DOE's strategy for biopreparedness and response. The genomics monitoring system described previously can be applied to human, animal, algae, and plant settings considering viruses, bacteria, and eukaryotic pathogens and disease vectors.

Activities

Building on the efforts funded through the NVBL and philanthropy, the JGI proposes to accomplish the following through DOE's BRaVE program or via alternative non-DOE funding sources:

- Develop a comprehensive collection of genomes of viral, bacterial, and eukaryotic pathogens for plants, algae, humans, and animals, including uncultivated predicted pathogens, and augment existing resources (GOLD, IMG/M) to make the database publicly available.
- Continue developing sensitive pathogen-agnostic sequencing approaches that provide sequence resolution at the strain level.
- Establish enrichment metagenomics to identify and characterize unexplored pathogens and their associated mobile genetic elements (plasmids, phages, etc.; **PI3-2**, **PI3-5**).
- Develop ML models for predicting pathogenicity and virulence factors in environmental sequence data through anomaly learning and reporting.
- Develop new synthetic biology capabilities to detect and monitor pathogen outbreaks in environmental and host systems, such as cell-free circuits to detect pathogens or host responses to pathogens.

35 https://onehealthinitiative.com/brief-definitions-of-one-health-and-one-health-approach

Appendix I: Implementation Milestones

MILESTONE ABBREVIATIONS

CO	Communications and Outreach
DS	Data Science and Informatics
FA	Fungal and Algal Genome Science
GT	Genomic Technologies
MC	Microbial Genome Science
MG	Metagenome Science
ML	Metabolomics
OP	Operations
PI	Prokaryote Informatics
PL	Plant Genome Science
SM	Secondary Metabolites Science
SS	DNA Synthesis Science
US	User Programs

Milestones: Communications and Outreach (CO)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Improve JGI institutional website	**CO1-2:** Refine home page messaging, restructure website navigation of **https://jgi.doe.gov** to serve prospective users and partners and DOE stakeholders as primary audiences	**CO1-5:** Work with web developer to implement improved back-end infrastructure and site architecture; will include mobile responsiveness and updated page indexing for better metrics tracking and accessibility
Cultivate JGI advocates	**CO2-2:** Incorporate User Executive Committee members into JGI advocate ranks to reach target audiences	**CO2-5:** Cultivate JGI advocates from staff and rebuild corps who can serve as representatives
Expand outreach	**CO3-2:** Establish a JGI-AAMU pilot internship	**CO3-5:** Grow JGI-AAMU relationship beyond the summer internship

Milestones: Data Science and Informatics (DS)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Enhance metadata via curation and AI	**DS1-2:** Establish priorities and identify requirements for centralizing and consolidating JGI internal and external metadata	**DS1-5:** Develop and implement systems and practices for ongoing metadata centralization and retroactive backfilling of historical metadata gaps
Capture and analyze citation data	**DS2-2:** Investigate data citation practices across all science programs and their respective communities	**DS2-5:** Develop and implement systems and practices for capturing, categorizing, and displaying JGI data citations; link information to the JGI Data Portal
Centralized metadata capture	**DS3-2:** Metadata submission portal that captures NMDC-compliant metadata in production for the Metagenome Program	**DS3-5:** Metadata submission portal in production that captures standardized, validated metadata for all JGI Programs
Intuitive search across resources	**DS4-2:** A search interface spanning the JGI, KBase, and NMDC that supports semantic queries	**DS4-5:** A search interface that spans the JGI, KBase, NMDC, ESS-DIVE, and EMSL and supports semantic queries

Milestones: Fungal and Algal Genome Science (FA)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Advance toward 10,000 fungal genomes	**FA1-2:** Streamline annotation and analysis of fungal single-cell, MAG, and co-culture genomes	**FA1-5:** Optimize analytical workflows to integrate up to 5,000 fungal genomes in MycoCosm
Lead algal genomics community development	**FA2-2:** Launch a regular algal genomics meeting to accelerate innovation in algal genomics and foster the growth of a collaborative algal user community	**FA2-5:** Engage research labs and culture collections to produce up to 300 diverse algal genomes in PhycoCosm
Enable functional predictions	**FA3-2:** Develop genome-centric AI tools to predict gene functions and traits in biotechnologically relevant yeasts	**FA3-5:** Predict function of fungal conserved hypothetical proteins using 3D structure and large protein-language models
Support environmental genomics	**FA4-2:** Enable eukaryotic microbiome exploration with catalogs of gene markers, organellar and nuclear genomes	**FA4-5:** Produce environmental genomes from educational and research initiatives
Support pangenomes and multi-omics	**FA5-2:** Integrate tools for exploration of multi-omics in MycoCosm and PhycoCosm	**FA5-5:** Build first fungal and algal pangenomes in MycoCosm and PhycoCosm

Milestones: Genomic Technologies (GT)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Upgrade short-read sequencing platform	**GT1-2:** Build capacity to process 30,000 samples per year	**GT1-5:** Build capacity to process 40,000 samples per year
Upgrade long-read sequencing platform	**GT2-2:** Build capacity to process 2,500 samples per year	**GT2-5:** Build capacity to process 4,000 samples per year
Develop plant single-cell atlas	**GT3-2:** Generate single-cell transcriptome data from multiple tissues for two plant flagship species	**GT3-5:** Generate single-cell transcriptome data from multiple tissues for four plant flagship species
Develop spatial transcriptomic platform	**GT4-2:** Test nascent spatial transcriptomics technologies on same samples as plant single-cell atlas	**GT4-5:** Explore feasibility of spatial transcriptomic user program
Link imaging and sequencing of uncultured microbes	**GT5-2:** Develop strategy to link cell imaging with genome sequencing of uncultivated microbes	**GT5-5:** Provide combined imaging and genome sequencing of uncultivated microbes to JGI users
Expand isotopic labeling for functional characterization of microorganisms	**GT6-2:** Increase the scale of SIP metagenomics to 650 samples per year	**GT6-5:** Examine strategies for isotopic labeling and genome characterization of uncultured microbes at the level of single cells
Enable experimental data integration	**GT7-2:** Collect multimodal data from bulk tissues (e.g., transcriptomes, epigenomes, methylomes, etc.)	**GT7-5:** Collect multimodal data from single cells (e.g., transcriptomes epigenomes, methylomes, etc.)
Increase DNA synthesis to enable future capabilities	**GT8-2:** Increase capacity to allow 12 Mb of DNA synthesis to enable plant engineering	**GT8-5:** Further increase potential capacity to 16 Mb of DNA synthesis as required by other platforms/programs
Develop infrastructure for CRISPR screens	**GT9-2:** Construct publicly accessible database for distribution of CRISPR screen data and libraries	**GT9-5:** Develop ability to comparatively analyze screen data across condition and organism
Enable user access to engineered strains	**GT10-2:** Increase user access to strain engineering to three phyla	**GT10-5** Increase user access to strain engineering to seven phyla, three kingdoms (i.e., yeast, algae, and bacteria)
Increase rate of metabolite identification	**GT11-2:** Launch targeted nonpolar metabolomics analysis product	**GT11-5:** Increase database of annotated metabolomics standards to >10,000 compounds
Expand nonpolar platform development	**GT12-2:** Develop expanded nonpolar product covering at least three lipid species	**GT12-5:** Broaden expanded nonpolar metabolite product to cover a broader range of species

Milestones: Microbial Genome Science (MC)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Dissect microbial population structure	**MC1-2:** Generate large-scale, quality bacterial and archaeal single-cell genomic datasets to study population genetics in the wild	**MC1-5:** Link strain-level diversification in uncultivated microorganisms to ecological niche adaptations
Establish microbial genome hit lists	**MC2-2:** Establish data-driven hit list for targeted bacterial and archaeal cultivation and genome recovery, based on the global census data	**MC2-5:** Establish data-driven target list for the prioritization of underexplored bacterial and archaeal genome sequencing targets from culture collections and assess their value for the bioeconomy
Enable large-scale microbial isolate genomics	**MC3-2:** Explore new partnerships with public and private isolate collections to access tens of thousands of bacterial and archaeal isolates for genome sequencing	**MC3-5:** Complete the genome sequencing for large-scale isolate projects (10,000 Actinomycetota, 10,000 type strains, Hungate1000 collection)
Characterize phenotypic traits	**MC4-2:** Investigate and test novel approaches to link phenotypes of uncultivated microorganisms to their taxonomy	**MC4-5:** Experimentally investigate morphological or other traits of at least three microbial clades
Decipher inter-organismal interactions	**MC5-2:** Generate at least 1,000 microeukaryote single-cell and enrichment genomic datasets, to cover both broad phylogenetic, as well as population-level diversity	**MC5-5:** Exploit microeukaryote single-cell and metagenome sequencing datasets for the discovery of novel clades of microbial symbionts and associated viruses, and employ AI to predict microbial lifestyles

Milestones: Metagenome Science (MG)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Characterize microbial interactions over time and space	**MG1-2:** Benchmark and onboard new tools and methods to understand microbe-microbe and virus-microbe interactions from time- and spatial-series studies	**MG1-5:** Leverage detailed characterization of model systems combined with large-scale, time-series metagenome data to understand assembly processes and virus-microbe interactions over ecological and evolutionary time scales
Connect microbes to biogeochemical trait catalogs	**MG2-2:** Benchmark and compare catalog(s) of curated biogeochemically relevant genes and pathways for metagenome bin annotation	**MG2-5:** Develop and provide computational tools and visualization to support user analyses of biogeochemically relevant pathways and enable trait-based modeling of microbiomes

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Develop quantitative SIP toolkit	**MG3-2:** Formalize a standard qSIP analysis pipeline to functionally associate viruses and microbial genomes to metabolisms	**MG3-5:** Robustly link community-scale measurement of microbial activity, including qSIP, to metabolomics data and metabolic processes
Establish comprehensive genome resources across Earth's biomes	**MG4-2:** Expand the metagenome toolkit to enable JGI users to identify, analyze, and submit to public databases viral, microbial, and microeukaryote population genomes recovered from Earth's microbiomes	**MG4-5:** Establish a global scale catalog of viral, microbial, and microeukaryote population genomes connected to custom interfaces and visualization resources for user analyses
Advance analytics and visualization for multi-omics of microbiomes	**MG5-2:** Develop a plan for integrated analysis and visualization tools enabling interpretation of metatranscriptome and paired metagenome data	**MG5-5:** Leverage paired metatranscriptomes and metagenomes to characterize microbial and viral activity, transcription regulation mechanisms, and noncoding RNAs in microbiomes at scale

Milestones: Metabolomics (ML)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Advance cheminformatics for improved metabolite identification	**ML1-2:** Develop and test integrated experimental and cheminformatic tools for identifying novel metabolites with a focus on soils	**ML1-5:** Deploy integrated experimental and cheminformatic tools for the analysis of diverse metabolites with a focus on soils
Integrate metabolomics with gene expression	**ML2-2:** Develop technologies for in situ analysis of metabolic activities and gene content and expression for an integrated genometabolic understanding of environmental systems	**ML2-5:** Demonstrate technologies for in situ genometabolic analysis of metabolic activities with related expressed genes especially to understand controllers of soil carbon persistence
Harness fabricated ecosystems for mechanistic ecology	**ML3-2:** Pilot approaches using fabricated ecosystems technologies to test predicted plant and microbial interactions and activities including those mediated by secondary metabolites	**ML3-5:** Harness EcoFAB and EcoBOT capabilities to enable users to test predicted metabolic activities and interactions of plants and microbes with a focus on those impacting soil carbon and nutrient cycles
Advance mechanistic understanding of key soil carbon cycling processes	**ML4-2:** Develop JGI capabilities for studying the transformations of phenolic compounds in soil	**ML4-5:** Use metabolomics capabilities to gain new insights into the plant and microbial enzymatic transformations of soil phenolics
Enhance user data analysis tools	**ML5-2:** Deposit raw data, both private and public, for web-based access by users	**ML5-5:** Enable users to mine their raw data using queries, data exploration, and networking with web-based tools

Milestones: Operations (OP)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Evolve JGI workforce	**OP1-2:** To guide future development of the JGI workforce and strategic initiatives, develop methodologies to assess current skills of JGI staff to identify core underutilized skills and determine skill inventory gaps that need to be addressed with cross-training, knowledge sharing, stretch assignments, and general learning opportunities	**OP1-5:** Deploy two learning programs and partner with Berkeley Lab's Learning and Organizational Development for the entire JGI workforce; expand learning opportunities with one external collaborating institution (e.g., UC system, national labs)
	OP2-2: Continue to conduct the annual JGI IDEA engagement survey to gauge employee satisfaction and identify two areas of improvement each year to address topic areas that scored unfavorably	**OP2-5:** Deploy and formalize communication plans/approaches that foster connection, engagement, and inclusion for new hires; an updated approach for new hire onboarding will be established
Maximize JGI efficiency	**OP3-2:** Critically evaluate the end-to-end acquisition process, evaluate alternative systems or leverage Berkeley Lab systems, and re-engineer the processes and systems as necessary for the order through delivery pipeline	**OP3-5:** Identify at least one additional system or process, such as travel or the IGB stockroom, to evaluate for the opportunity to either update it to fit the current needs or leverage Berkeley Lab or other off-the-shelf systems to reduce duplicate efforts or other inefficiencies
	OP4-2: Optimize the JGI annual budget process to critically evaluate the staffing level of each department and priorities for the work of that staff	**OP4-5:** Implement more transparency and control for individual budget holders, including decision-making when resource constraints exist
	OP5-2: Evaluate and invest in necessary facility modifications to match the office and meeting space needs of the IGB to support the current working environment	**OP5-5:** Evaluate the profile of technical/lab space in the IGB and modify space and allocation mechanisms to adapt to needs established by the JGI strategic plan

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Amplify the JGI's impact	**OP6-2:** Identify the most valuable components of the JGI sequencing, synthesis, metabolomics, and data pipelines that industry needs; increase the visibility of these and facilitate two new industry projects	**OP6-5:** Create and maintain a standard operating framework for the contact, initiation, and formulation of partnerships with potential industry partners to facilitate five new industry projects
	OP7-2: Establish a cross-functional team to evaluate and prioritize opportunities for business process reengineering or optimization	**OP7-5:** Implement an organizational change management process that will enable proper preparation for and communication of process and systems changes
	OP8-2: Assess gaps, inefficiencies, and redundancy in current internal communications channels, identify and implement highest-impact change opportunities	**OP8-5:** Establish a routine process for the evaluation and refresh of internal communications mechanisms to maximize effective knowledge of staff

Milestones: Prokaryote Informatics (PI)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Explore functional diversity	**PI1-2:** Complete global census of microbial (including viral) phylogenetic diversity based on IMG/M data for cultivated and uncultivated lineages	**PI1-5:** Develop computational tools for identification of metabolic complementarities between uncultivated lineages to understand cultivation bottlenecks and help guide targeted cultivation efforts
	PI2-2: Develop computational tools/ approaches and data to enable the analysis of co-occurrence, co-localization, and co-expression patterns to facilitate characterization of functional diversity	**PI2-5:** Develop AI methods for structural prediction to enable the characterization of functional diversity, including interactions with small molecules and protein-protein interactions
	PI3-2: Identify MGE functional diversity based on viral and plasmid protein families, associated with MGE host-sharing networks with custom visualization and analysis tools	**PI3-5:** Identify new MGE lineages with complete and nearly complete genomes from MGE gene-sharing networks, and develop new methods for connecting new viruses and MGEs to their hosts, including potential collaboration with PhageFoundry
Support studies of nutrient cycling	**PI4-2:** Enhance IMG/M and GOLD to support large-scale metagenomic data mining for enzyme discovery to enable DNA synthesis target prioritization	**PI4-5:** Develop tools and approaches for characterization of protein families of unknown function, including their participation in known and novel metabolic pathways

Milestones: Plant Genome Science (PL)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Advance DOE plant genome resources	**PL1-2:** Continue to develop and optimize methods to make production more cost effective; apply these methods to build genomes both across the plant phylogeny and within the diversity of JGI flagship species, including updating older-technology-based genomes to state-of-the-art builds	**PL1-5:** Build pangenome resources for most important DOE plant species, with integrated panannotations of gene models, and complete phylogenetic diversity sampling of key species across the plant kingdom
Develop and improve pangenome analyses	**PL2-2:** Build a platform to integrate multiple reference genomes, contig-level assemblies, and resequencing; apply these methods to switchgrass, eucalyptus, sorghum, and other important DOE species	**PL2-5:** Fully integrate pangenomes, phylogenetically dispersed references, and model species genomes to better characterize molecular and phenotypic evolution across plant diversity, and accelerate crop improvement efforts
Improve accessibility and interoperability of pangenome data	**PL3-2:** Continue to develop mentoring capacity and accompanying analytical tools that will permit users to independently query multiscale pangenome data and analyses, including building tools to undertake large collaborative projects that let users detect regions of interest	**PL3-5:** Continue to expand and develop this analytical toolbox, and provide user access to the toolbox methods via query and visualization tools integrated into Phytozome
Improve and use tools for multi-omic data integration to aid annotation and selection of candidate sequences	**PL4-2:** Develop methods for the integration of disparate data types (e.g., reference genomes, comparative genomic, metabolomic, expression, epigenetic) to improve annotation and identification of candidate sequences controlling phenotypes of interest	**PL4-5:** Continue to develop these methods and apply them to user projects
Develop mechanistic understanding of plant-microbe interactions	**PL5-2:** Apply emerging techniques to identify genes and metabolites that influence the composition and function of plant-associated microbiomes	**PL5-5:** Develop a broad understanding of plant-microbiome interactions, including the roles of stress and microbiome composition, which can be used to design interventions that increase plant growth and productivity
Improve single-reference and pangenome gene prediction	**PL6-2:** Improve the throughput and sensitivity of single-reference annotation pipelines to dovetail with increased genome assembly production; optimize pangene representation and discovery	**PL6-5:** Incorporate protein folding, deep learning-based gene calling algorithms, and gene presence-absence scores to improve gene annotation and functional prediction; integrate these approaches with transcriptomic and homology evidence

Milestones: Secondary Metabolites Science (SM)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Dive into the untapped reservoir of secondary metabolites	**SM1-2:** Conduct user-centered feedback studies on SMC 1.0 release design and functionality; create software development and design a plan for future versions of SMC, and address SMC users' most urgent needs	**SM1-5:** Integrate the SMC data analysis pipeline and database import process with JGI metadata coordination efforts (**DS1-5**) to allow BGC identification from all available genomes and metagenomes with direct connections to experimental metadata
	SM2-2: Develop and import new prediction tools for BGC prediction and clustering	**SM2-5:** Develop new prediction tools to meet gaps for novel secondary metabolite BGCs and accessory genes
	SM3-2: Develop BGC-QL to standardize description of BGCs	**SM3-5:** Develop bgc-Chat as a new methodology for users to query secondary metabolites data
Functionally characterize secondary metabolite pathways at scale	**SM4-2:** Successfully synthesize, clone, express, and identify products from 10 BGCs that were specified by users	**SM4-5:** Successfully synthesize, clone, express, and identify products from 100 BGCs that were specified by users
	SM5-2: Conduct DAP-seq, multiDAP, and RIViT-seq on one secondary metabolite producer and one TF family to build out a comprehensive gene regulatory network	**SM5-5:** Conduct DAP-seq, multiDAP, and RIViT-seq on five secondary metabolite producers and five TF families specified by users to build out comprehensive gene regulatory networks
	SM6-2: Establish collaboration to explore use of imaging technology(ies) for structural characterization of secondary metabolites from one producer	**SM6-5:** Implement imaging technology(ies) for structural characterization of secondary metabolites via a collaborator as a user offering

Milestones: DNA Synthesis Science (SS)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Collect large datasets for analyzing sequence-function relationships	**SS1-2:** Develop a strategy to rapidly identify new biosensors useful for detection of diverse bioproducts (e.g., biofuels, biochemicals)	**SS1-5:** Using the strategy developed in **SS1-2**, help users identify biosensors of their interests
	SS2-2: Test the utility of microdroplet, imaging, biosensors, or their combinations to build a pipeline for evaluation of biosystems with desired function	**SS2-5:** Use the pipeline developed in **SS2-2** for screening of biosystems with enhanced capabilities of producing biofuels and bioproducts
Develop AI pipelines for predictive biosystems design	**SS3-2:** Identify AI experts to collaborate with and develop data resources and AI-assisted tools to study sequence-function relationships	**SS3-5:** Host a CASP-like competition for development of AI tools to study sequence-function relationships based on these data resources and validate the prediction
Develop a biosystems design platform that is increasingly self-driving	**SS4-2:** Develop computational tools to help bridge the gap between the Learn and Design phases of the DBTL cycle	**SS4-5:** Using the capabilities developed in **SS1** and **SS2**, establish an automated process for biosystems design

Milestones: User Programs (UP)

TOPIC	2-YEAR MILESTONE	5-YEAR MILESTONE
Evolve user program call	**UP1-2:** Evolve user programs toward larger-scale community proposals, while promoting access for new investigators and MSIs	**UP1-5:** Expand cross-facility collaborations through joint user program calls
Improve impact tracking	**UP2-2:** Migrate JGI publication database and associated infrastructure to IGB servers and integrate with other JGI systems	**UP2-5:** Investigate means for leveraging publication data and other linked data for understanding the JGI primary and secondary user bases and characterizing their activities

Appendix II: Contributors

Strategic Retreat Participants

JGI internal participants: Massie Ballon, Leo Baumgart, Ian Blaby, Justina Clarke, Crysten Blaby-Haas, Matthew Blow, Neil Byers, Stephen Chan, Chris Daum, Emiley Eloe-Fadrosh, Nicholas Everson, Kjiersten Fagnan, Igor Grigoriev, Nathan Hillson, Nikos Kyrpides, Katherine Louie, Rex Malmstrom, Nigel Mouncey, Trent Northen, Ronan O'Malley, Len Pennacchio, Georg Rath, Dan Rokhsar, Simon Roux, Jeremy Schmutz, Frederik Schulz, Atif Shahab, Grace Sprehn, Axel Visel, John Vogel, Steven Wilson, Tanja Woyke, Yasuo Yoshikuni, Zhong Wang

JGI external participants: Paul Adams, Lawrence Berkeley National Laboratory; John Archibald, Dalhousie University; Rodney Brister, National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information; Shreyas Cholia, Lawrence Berkeley National Laboratory; Timothy Donohue, University of Wisconsin, Madison; Bruce Hungate, Northern Arizona University; Lauren Jabusch, Lawrence Berkeley National Laboratory; Elizabeth Murphy, University of Illinois; Michelle O'Malley, University of California, Santa Barbara; Stavros Papadopoulos, TileDB; Teresa Pawlowska, Cornell University; Lita Proctor, National Institutes of Health; Seung Yon (Sue) Rhee, Carnegie Institution for Science; Linda Setchell, Product Ecology Group; Patrick Shih, University of California, Berkeley, JBEI; Suzanne Sindi, University of California, Merced; Mingxun Wang, University of California, Riverside

Writing Team

Overall coordination and general sections: Axel Visel, Nigel Mouncey

Theme 1, Nutrient Cycling: John Vogel (Lead), Berkeley Kauffman, Simon Roux, Andrei Steindorff

Theme 2, Functional Diversity: Tanja Woyke (Lead), Leo Baumgart, Hualan Liu, Stephen Mondo, Juan Villada

Theme 3, Data and Connectivity: Kjiersten Fagnan (Lead), Len Pennacchio, Steven Ahrendt, Supratim Mukherjee, Charles Parker, Daniela Cassol

Theme 4, Stewarding Resources: Nicholas Everson (Lead), Justina Clarke, Massie Ballon, TBK Reddy

Initiative A, Biomolecular Materials: Crysten Blaby-Haas (Lead), Natalia Molchanova (TMF), Bishoy Kamel

Initiative B, Biosurveillance and Biopreparedness: Nigel Mouncey (Lead), Frederik Schulz, Harshini Mukundan (Biosciences Area, Berkeley Lab)

Other JGI internal writing and review contributors: Matthew Blow, Neil Byers, I-Min Chen, Emiley Eloe-Fadrosh, Robert Evans, Igor Grigoriev, Nancy Hammon, Nathan Hillson, Rene Perrier, Jeremy Schmutz, Dan Udwary, Yasuo Yoshikuni

JGI External Contributors

Strategy development consultant: Linda Setchell, Product Ecology Group

JGI advisory committee members, representatives of JGI partner organizations, stakeholders, and other external contributors: Mark Adams, The Jackson Laboratory; Paul Adams, Lawrence Berkeley National Laboratory; Todd Anderson, U.S. Department of Energy; John Archibald, Dalhousie University; Adam Arkin, University of California, Berkeley; Emily Aurand, Engineering Biology Research Consortium; Scott Baker, Pacific Northwest National Laboratory; Gwyn Beattie, Iowa State University; Gregg Beckham, National Renewable Energy Laboratory; Kirsten Benjamin, PivotBio; Shannon Bennett, California Academy of Sciences; Randy Berka, Archer Daniels Midland Company; Siobhan Brady, University of California, Davis; J. Rodney Brister, NCBI, National Institutes of Health; Shreyas Cholia, Lawrence Berkeley National Laboratory; Katy Christiansen, Lawrence Berkeley National Laboratory; Jonathan Conway, Princeton University; Gloria Coruzzi, New York University; Jeffery Dangl, University of North Carolina,

Chapel Hill; Kristen DeAngelis, University of Massachusetts Amherst; Douglas Michael Densmore, Boston University; Timothy Donohue, Great Lakes Bioenergy Research Center, University of Wisconsin-Madison; Elizabeth Dumont, University of California, Merced; Joseph Ecker, The Salk Institute for Biological Studies; Carrie Eckert, Oak Ridge National Laboratory; Joanne Emerson, University of California, Davis; Paul Flicek, European Bioinformatics Institute, University of Cambridge; Jack Gilbert, University of California, San Diego; Louise Glass, University of California, Berkeley; Stephen Goodwin, Purdue University; Rebecca Goss, University of St. Andrews; Kathleen Greenham, University of Minnesota; Susan Gregurick, National Institutes of Health, Office of Data Science Strategy; Steven Hallam, University of British Columbia; Roland Hatzenpichler, Montana State University; Samuel Hazen, University of Massachusetts Amherst; Matthias Hess, University of California, Davis; David Hibbett, Clark University; Kirsten Hofmockel, Pacific Northwest National Laboratory; Philip Hugenholtz, University of Queensland; Bruce Hungate, University of Northern Arizona; Lauren Jabusch, Lawrence Berkeley National Laboratory; Janet Jansson, Pacific Northwest National Laboratory; Thomas Juenger, University of Texas, Austin; Jay Keasling, University of California, Berkeley; Elizabeth Kellogg, Danforth Center; Kostas Konstantinidis, Georgia Tech; Daniel van der Lelie, Gusto LLC; Ramana Madupu, U.S. Department of Energy; Douglas Mans, Pacific Northwest National Laboratory, Environmental Molecular Sciences Laboratory; Francis Martin, Institut National de la Recherche Agronomique, France; Christopher Mason, Cornell; Trina McMahon, University of Wisconsin, Madison; Sean McSweeney, Brookhaven National Laboratory; Sabeeha Merchant, University of California, Berkeley; Stephen Moose, University of Illinois; Mary Ann Moran, University of Georgia; Nancy Moran, University of Texas, Austin; Elizabeth Murphy, University of Illinois Urbana-Champaign, CABBI; Irina Novikova, Pacific Northwest National Laboratory; Michelle O'Malley, University of California, Santa Barbara; Howard Ochman, University of Texas, Austin; Stavros Papadopoulos, TileDB, Inc.; Teresa Pawlowska, Cornell University; Graham Peers, Colorado State University; Jennifer Pett-Ridge, Lawrence Livermore National Laboratory; Mircea Podar, Oak Ridge National Laboratory; Juergen Polle, Micro Bioengineering; Lita Proctor, National Institutes of Health; Betsy Reed, California State University, San Marcos; Seung Yon (Sue) Rhee, Carnegie Institution for Science; Richard Roberts, New England BioLabs; Bob Schmitz, University of Georgia; Lynn Schriml, University of Maryland; Ashley Shade, Centre National de la Recherche Scientifique; Patrick Shih, University of California, Berkeley; Suzanne Sindi, University of California, Merced; Joseph Spatafora, Oregon State University; Gary Stacey, University of Missouri; Jason Stajich, University of California, Riverside; Ramunas Stepanauskas, Bigelow Laboratory for Ocean Sciences; Matt Sullivan, Ohio State University; Kathleen K. Treseder, University of California, Irvine; Adrian Tsang, Concordia University, Canada; Jana U'Ren, University of Arizona; Ophelia Venturelli, University of Wisconsin; Ronald de Vries, Westerdijk Fungal Biodiversity Institute, Netherlands; Momchilo Vuyisich, Viome Life Sciences, Inc.; Setsuko Wakao, University of California, Berkeley; Mingxun Wang, University of California, Riverside; Ian Wheeldon, University of California, Riverside; Philipp Zerbe, University of California, Davis; Huimin Zhao, University of Illinois

Creative Services: Eduardo de Ugarte, Caitlin Youngquist, Meghan Zodrow, Jessica Scully (JS Communication Consulting), Thor Swift

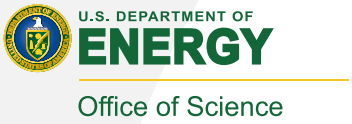
Photo credit: Thor Swift.

Appendix III: Abbreviations

AAMU	Alabama Agricultural and Mechanical University
ABPDU	Advanced Biofuels and Bioproducts Process Development Unit
AI	artificial intelligence
ALGAE	Algal Genome Annotation Encyclopedia
AMF	arbuscular mycorrhizal fungi
API	application programming interface
ARES	Automated Resource Enabling Synthesis
ART	Automated Recommendation Tool
BER	DOE Program of Biological and Environmental Research
BERAC	Biological and Environmental Research Advisory Committee
BES	DOE Program of Basic Energy Sciences
BGC	biosynthetic gene cluster
BGC-QL	biosynthetic gene cluster query language
BioEPIC	Biological and Environmental Program Integration Center
BLiSS	Biosecurity List Sequence Screening
BOOST	Build Optimization Software Tools
BRaVE	Biopreparedness Research Virtual Environment
BRC	Bioenergy Research Center
BSSD	Biological Systems Science Division
CABBI	Center for Advanced Bioenergy and Bioproducts Innovation
CASP	Critical Assessment of Protein Structure Prediction
CRAGE	chassis-independent recombinase-assisted genome engineering
CRISPR	clustered regularly interspaced short palindromic repeats
CRISPRa	CRISPR activation
CRISPRi	CRISPR interference
CSP	Community Science Program
CSU	California State University
DAP-seq	DNA affinity purification sequencing
DBTL	Design-Build-Test-Learn
DEI	diversity, equity, and inclusion
DL	deep learning
DOE	U.S. Department of Energy
ECNet	Evolutionary Context-Integrated Neural Network
EcoFAB	fabricated ecosystem
EMSL	Environmental Molecular Sciences Laboratory
ESS-DIVE	Environmental System Science Data Infrastructure for a Virtual Ecosystem
FAIR	findable, accessible, interoperable, and reusable
FICUS	Facilities Integrating Collaborations for User Science
FISH	fluorescence in situ hybridization
GEBA	Genomic Encyclopedia of Bacteria and Archaea
GLBRC	Great Lakes Bioenergy Research Center
GNPS2	Global Natural Product Social Molecular Networking
GOLD	Genomes OnLine Database
GPU	graphics processing unit
GWAS	genome-wide association study
HTP	high throughput
HT-SIP	high-throughput stable-isotope probing
IDEA	Inclusion, Diversity, Equity, Accountability
IGB	Integrative Genomics Building
IMG/M	Integrated Microbial Genomes and Microbiomes
IMG/PR	Integrated Microbial Genomes Plasmid Resources
IMG/VR	Integrated Microbial Genomes Viral Resources
INSDC	International Nucleotide Sequence Database Collaboration
JAMO	JGI Archive and Metadata Organizer

JAWS	JGI Analysis Workflow Service
JBEI	Joint BioEnergy Institute
JDP	JGI Data Portal
JGI	Joint Genome Institute
KBase	DOE Systems Biology Knowledgebase
LC-MS/MS	liquid chromatography tandem mass spectrometry
LIMS	laboratory information management system
LLM	large language model
LTER	Long-Term Ecological Research
M2PC	Microbial Molecular Phenotyping Capability
MAESTRO	ML-assisted engineering of stress tolerance rational optimization
MGE	mobile genetic element
MIBiG	Minimum Information about a Biosynthetic Gene cluster
MicroED	microcrystal electron diffraction
ML	machine learning
MOU	memorandum of understanding
MSI	minority-serving institution
NAIST	Nara Institute of Science and Technology in Japan
NanoPOTS	nanodroplet processing in one pot for trace samples
NASA	National Aeronautics and Space Administration
NCBI	National Center for Biotechnology Information
NCEM	National Center for Electron Microscopy
NEON	National Ecological Observatory Network
NERSC	National Energy Research Scientific Computing Facility
NMDC	National Microbiome Data Collaborative
NSRC	Nanoscale Science Research Center
NVBL	National Virtual Biotechnology Laboratory
OHITF	One Health Initiative Task Force
ORF	open reading frame
OSTP	Office of Science and Technology Policy
PCR	polymerase chain reaction
PDB	protein data bank
PI	principal investigator
RIViT-seq	regulon identification by in vitro transcription-sequencing
RNA-seq	RNA sequencing
SC	DOE Office of Science
SFA	Science Focus Area
sgRNA	single-guide RNA
SIP	stable-isotope probing
SMC	Secondary Metabolism Collaboratory
SNP	single nucleotide polymorphism
SRA	Sequence Read Archive
STEM	science, technology, engineering, and mathematics
TF	transcription factor
TMF	The Molecular Foundry
UC	University of California


MCMXCVII



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

23-JGI-23146