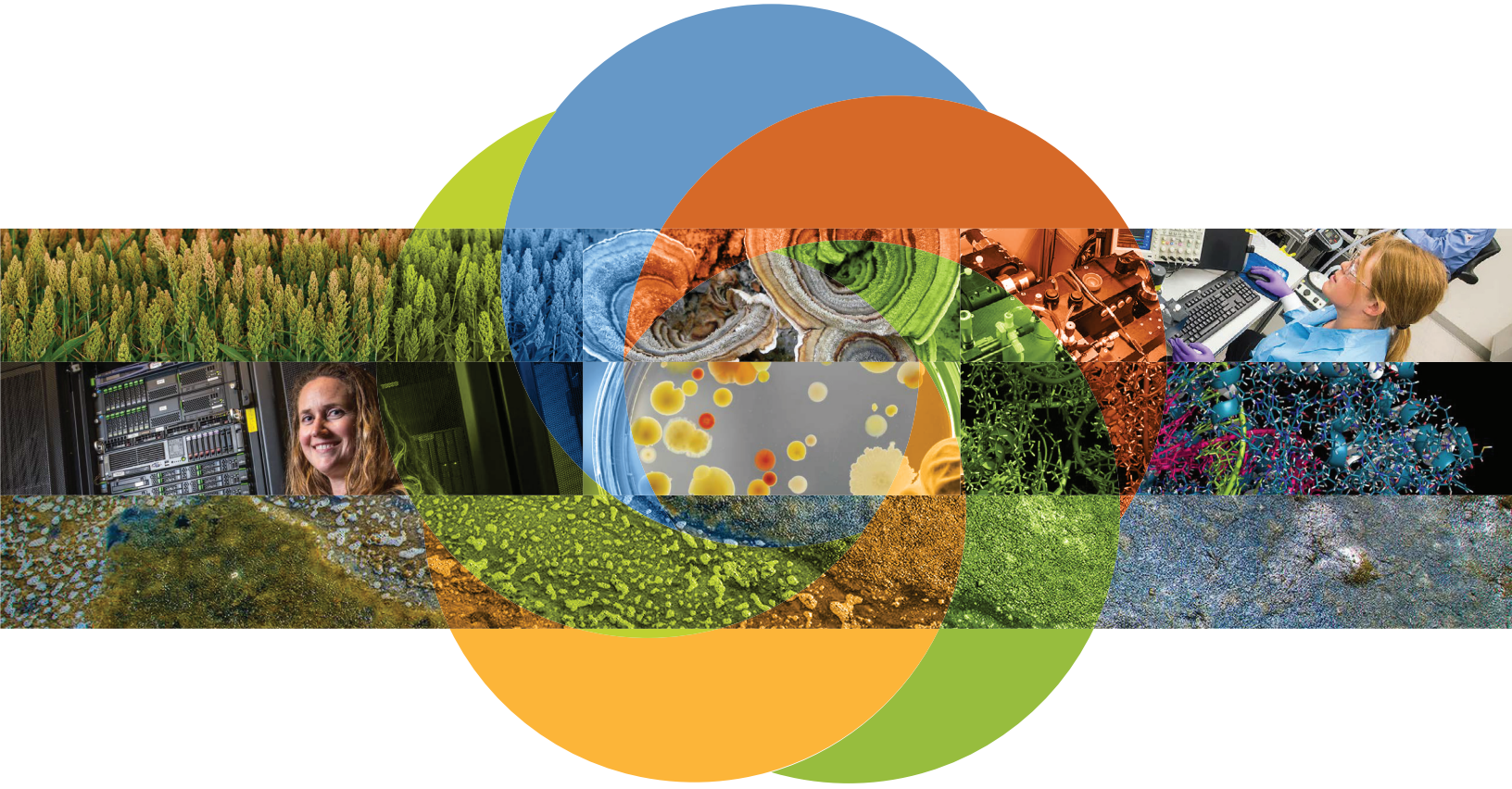
5-Year Strategic Plan

U.S. Department of Energy
Joint Genome Institute

Beyond Basepairs

Implementation Report
February 2024



JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

Table of Contents

Summary

In 2018, the U.S. Department of Energy (DOE) Joint Genome Institute (JGI) embarked on an ambitious journey with the launch of its strategic plan, "Beyond Basepairs - A Vision for Integrative and Collaborative Genome Science." In this plan, the JGI described a five-year vision, covering years 2019-2023 and encompassing the transformation of the JGI into an “Integrative Genome Science User Facility”. The work conducted under this plan was supported by the DOE Office of Science Biological and Environmental Research (BER) Program.

The vision was guided by the I5-framework, which outlined planned efforts to: 1. Support continued genomic data generation and discovery (Identification); 2. Establish advanced data science strategies (Interrogation); 3. Develop new tools for the exploration of genome function (Investigation); 4. Foster collaborations across the DOE research ecosystem and beyond (Integration); 5. Engage new user communities (Interaction). The plan also described critical aspects of stewardship to enable these activities, including operational excellence and talent management.

To ensure measurable progress toward this vision, the strategic plan featured more than 200 specific milestones, providing a granular roadmap for the implementation of the vision and describing specific activities and goals for individual groups and departments across the JGI. Over the past five years, these milestones have provided focus for JGI contributors and have been used routinely to guide strategic management decisions. Progress toward implementation of each milestone was captured in real time in an “Implementation Dashboard” to provide JGI’s senior management and stakeholders with an instantaneous view of implementation status.

As we conclude this five-year period, the JGI is sunsetting the 2018 strategic plan and transitioning into a new phase of its journey with the release of the new 2023 strategic plan *“Innovating Genomics to Serve the Changing Planet”.* In the present report, we provide a final overview of implementation progress made over the past five years, and illustrate across all five “I”s how the strategic plan has been instrumental in guiding the directions of the JGI. We also highlight how the plan contributed substantially to JGI’s continued ability to support its users in performing impactful science in energy and environmental genomics. Overall, the JGI accomplished 90% completion of the two- and five-year milestones defined in 2018. This remarkable success in accomplishing a long-term vision through sustained, focused efforts on implementing long-term goals underscores the effectiveness of JGI’s strategic planning and implementation processes and reinforces JGI’s plans to rely on a similar implementation approach for its upcoming 2023 strategic plan.

Overall Progress

Approach to Tracking Implementation

The JGI used a continuous tracking system and visual dashboard to measure the implementation of all milestones over the past five years. JGI leadership, Department and Program Heads reported on a monthly basis for each milestone its overall status (e.g., not started, in progress, complete) and, for milestones that were in progress, the estimated percent completion. Depending on the milestone, completion was based on quantifiable deliverables (e.g., 50 of 100 genomes completed) or, for qualitative milestones (e.g., establishment of a new method) based on the expected remaining effort required to reach the endpoint. The completion rate across all milestones was measured as the proportion of fully completed milestones and completed portions of partially completed milestones, relative to the total number of milestones.

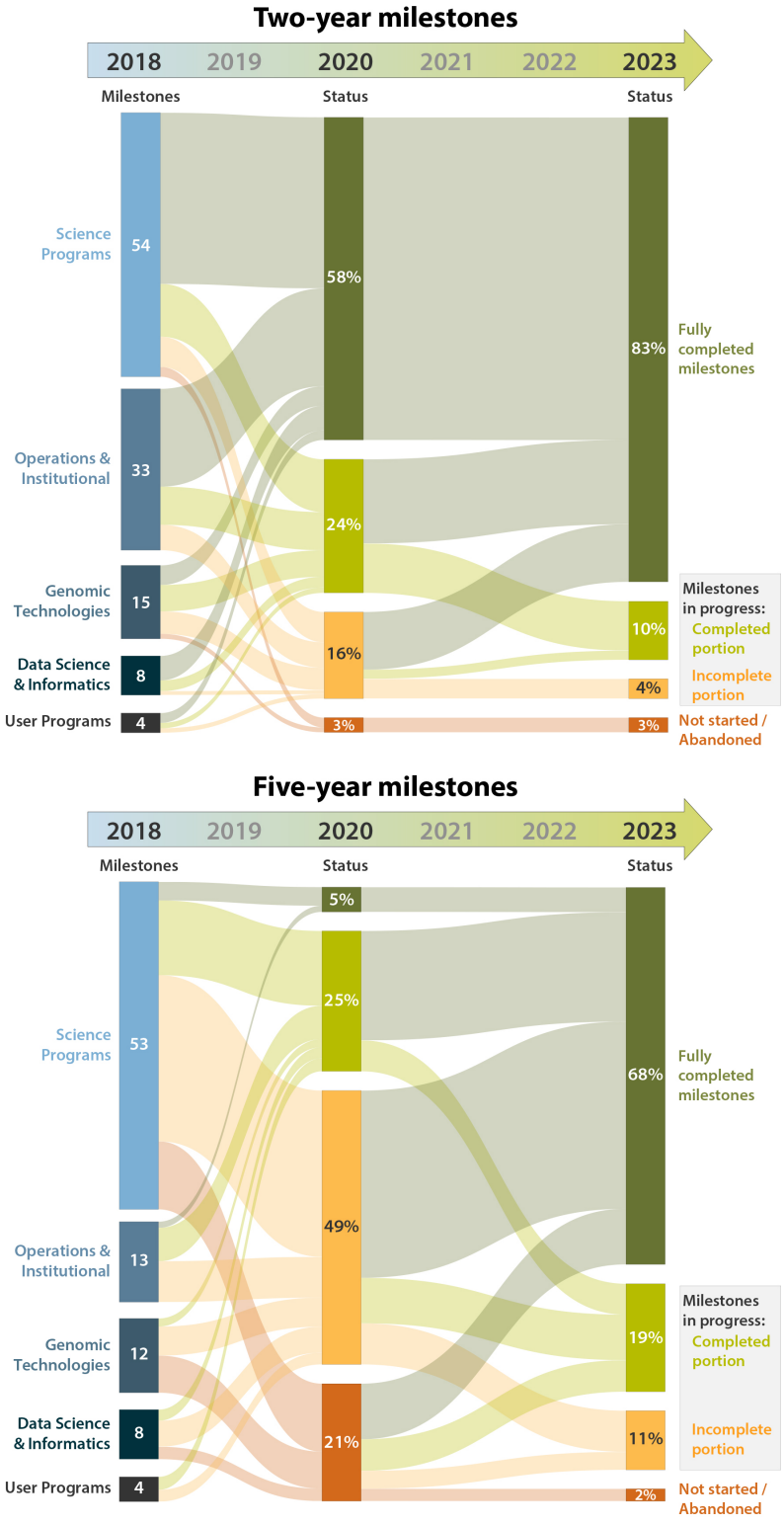


Fig. 1: Implementation of two- and five-year milestones.

Throughout the five-year implementation period, JGI senior management performed regular reviews of milestones that required adjustments. For example, in some cases milestones were no longer considered worth pursuing because their potential impact had been diminished by technological or scientific advancements in the field. In other cases, milestones were no longer considered attainable due to external dependencies. Finally, there were cases in which milestones were adjusted quantitatively or qualitatively. In total, fewer than 5% of milestones (10 of 208) underwent modifications and fewer than 2.5% (5 of 208) were abandoned. Any decisions to modify or abandon milestones were made only after detailed discussion between milestones owners and JGI leadership.

Implementation of Milestones over Time

The Sankey plots in **Fig. 1** summarize the implementation status of two-year (**top**) and five-year (**bottom**) milestones over time. During the initial implementation phase (October 2018 - October 2020), the main focus was on implementation of two-year milestones, which reached 82% completion by October 2020, while many five-years milestones had not yet been started (e.g., because they built on related two-year milestones) or were in early stages of implementation. In the following three years, focus shifted toward completion of five-year milestones, while completing remaining tasks for two-year milestones. By October 2023, we reached 93% completion of two-year milestones and 87% completion of five-year milestones (including completed portions of milestones still in progress).

Overall, JGI leadership considers this progress in implementation (90% completion across all milestones) very successful. With respect to the remaining ~10% of implementation steps, it is important to consider that many milestones were set intentionally to be very ambitious (high-risk, high-reward). Work on several milestones for which implementation progress has been slower than anticipated will be continued under the next strategic plan, where remaining work has been incorporated into our new sets of milestones.

Implementation of the I5 Framework

The 2018 Strategic Plan was built on the I5 framework (**Fig. 2**), which outlined the vision of an integrative genome science user facility that supports identification, interrogation, investigation, integration, and interaction, enabled by strong steward support in areas including talent management and operational excellence. Below, we describe selected implementation accomplishments, organized by the topic areas of the I5 framework.

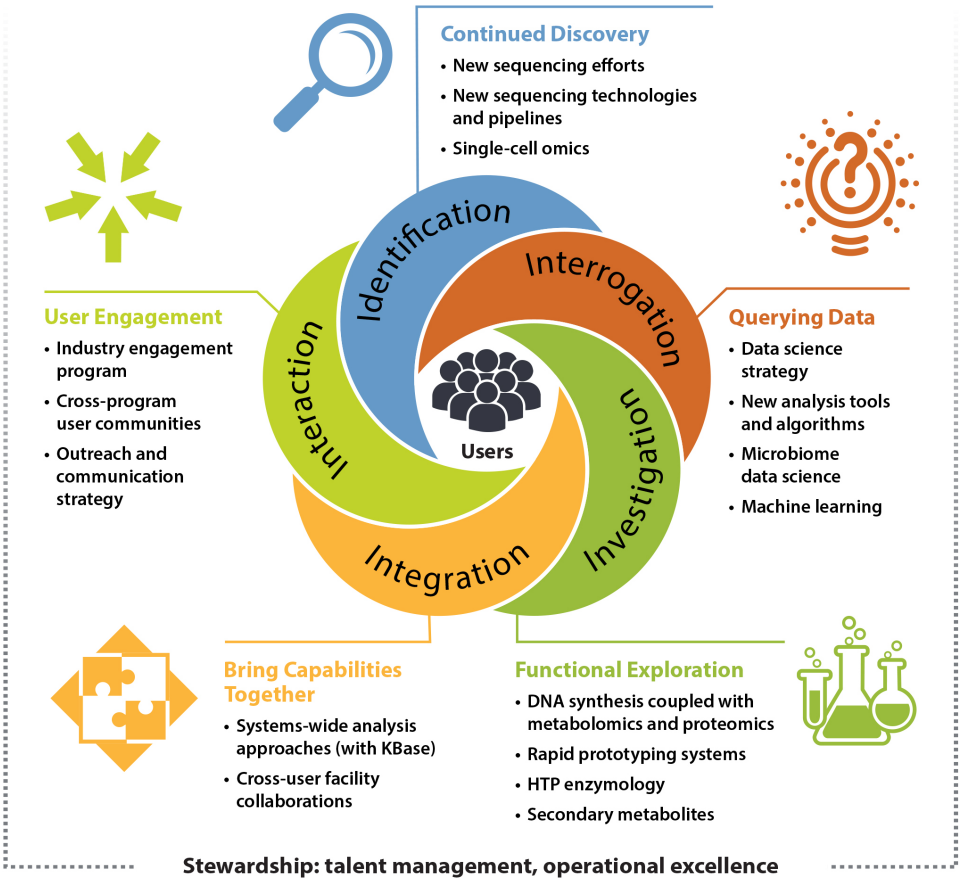


Fig. 2: The I5 framework

Identification - Continued Discovery

- **Expansion of DNA sequencing:** To enable broader user access to sequencing, we proposed to increase our DNA sequencing output to >25 Terabases (Tbs)/annually (**GNT02**). This target has been exceeded by more than an order of magnitude, with the generation of >700 Tbs of sequence in fiscal year 2023 alone. This capacity represents both short- and long-read technologies and provides a powerful foundation for addressing outstanding energy and environment questions at a scale not previously possible.
- **Single-cell transcriptome capabilities:** The ultimate resolution of biology is the single cell. The JGI has now established technological advances that enable the identification of transcriptomes at single cell resolution (**GNT01** and **GNT06**). This work provides a foundation for expanding projects like the Plant Gene Expression Atlas to single-cell resolutions.
- **Metabolomics:** The JGI has made a major investment in analyzing authentic metabolite standards to provide users with highest-confidence metabolite identification using high throughput metabolomics workflows (**MTB02**). In total, over 4,000 standards have been analyzed. To enable JGI users and the larger scientific community to leverage these data, we have deposited the corresponding spectra in Global Natural Products Social Molecular Networking (GNPS). To date, approximately 100,000 GNPS analysis jobs using these data have been completed.
- **DNA synthesis science:** The DNA Synthesis Science program in collaboration with the Data Science and Informatics (DSI) department and the Prokaryotic Informatics Group established computational genome mining capabilities for both genes and pathways of interest to our users (**SS02**). Users can now request genome mining support through our community science programs (e.g., Community Science Program (CSP), CSP Functional Genomics, Facilities Integrating Collaborations for User Science (FICUS)).
- **Stable isotope probing (SIP) metagenomics:** This newly implemented approach combines genome sequence analysis with functional labeling assays to reveal the activities of uncultivated microbes. With support from the JGI Emerging Technologies Opportunity Program (ETOP), JGI staff collaborated with Lawrence Livermore National Laboratory to develop a high-throughput pipeline for quantitative SIP metagenomics. Starting in 2020, the JGI provided SIP metagenomics capabilities to users seeking accurate estimates of in situ metabolic rates of microbiome community members and has since seen growth of this capability to 21 proposals (**GNT04**).
- **Improved reference plant genomes and pangenomes:** Technologies for long and high quality sequence reads permit both more accurate and higher throughput genome assemblies. Recently optimized sequencing and assembly pipelines have allowed us to produce over 100 plant genomes a year (**PLP01-2**), including those from plants with large, complex, and polyploid genomes (**PLP01-5**). The improved capacity for and accuracy of assembly has enabled us to build multiple high quality reference genomes for DOE species (**PLP02-2**). To aid scientific discovery, we have integrated our multiple Sorghum references and their gene annotations into a pangenome (**PLP03-2, PLP06-2**) and have developed the computational methods to complete a similar pangenomic resource for Poplar, Switchgrass, and several other DOE flagship species (**PLP03-5**). Our ongoing downstream analytical method development (see below) has allowed us to access structural variation and link molecular to phenotypic variation (**PLP06-5**). Combined, these resources and discoveries provide a comprehensive resource for DOE plant improvement efforts.

- **Single cell genomics of microeukaryotes:** To study uncultivated microorganisms and inter-organismal interactions between hosts, symbionts, and viruses, we proposed to pilot (**MIP05-2**) and scale up (**MIP05-5**) genomic sequencing of single cells from protists and other microeukaryotes. This was accomplished through collaborations with several JGI CSP investigators yielding more than 200 single amplified genomes (SAGs) from protists and more than 100 SAGs from chytrids. These efforts provide a foundation for microeukaryote single cell sequencing and data processing as a future user capability and application to other groups of microeukaryotes.

Interrogation - Querying Data

- **Genomes of unculturable eukaryotes extracted from metagenomes.** Given that the majority of fungi and algae cannot currently be cultured, extracting eukaryotic genomes from already sequenced metagenomes was pursued as a potentially scalable approach to access fungal and algal genomes (**FGP01**). In collaboration across several groups and programs, we benchmarked, optimized, and integrated methods and tools for binning, classification, assessment, and annotation of fungal and algal genomes. Several eukaryotic bins representing near-complete fungal and algal genomes have been annotated in MycoCosm and PhycoCosm to expand the diversity of sequenced genomes and enable high-impact publications.[1]
- **Enabling large-scale detection of viruses and other mobile genetic elements.** To enhance our ability to identify new viruses (**MGP05-2**) and to provide better annotation of uncharacterized protein families (**PKI01-5**), we developed geNomad, a new machine-learning tool for virus sequence detection. geNomad leverages innovations in autoencoders and neural networks along with a large-scale curated database of protein families, which enabled detection of a broad range of viruses across prokaryotic and eukaryotic hosts as well as the identification of other mobile genetic elements, such as plasmids. geNomad is more versatile and computationally efficient than previous state-of-the-art tools. Applying geNomad at scale across all public genomes and metagenomes in Integrated Microbial Genomes and Microbiomes (IMG) led to the detection of more than 5 million confident virus sequences, now integrated in the IMG Viral Resources (IMG/VR) database.[2]
- **Resources for viral, prokaryote, and eukaryote genomes from metagenomes.** We have advanced recovery of tens of thousands of genomes spanning all domains of life with improvements in scaling metagenome binning efforts and developing new approaches to identify virus and eukaryotic genomes (**FGP01-2; FGP01-5; MGP04-2; MGP04-5; MGP05-2; PKI05-2; PKI05-5**). To facilitate analysis of these recovered genomes, IMG/M, IMG/VR, MycoCosm, and PhycoCosm have expanded to support search functionality, improved annotations, and comparative analysis. Together, these data and new analysis capabilities are supporting views into the diversity and functional capacity of uncultivated viruses, prokaryotes, and eukaryotes[3,4].

---

[1] e.g., Nelson, A *et al.*, 2022, “Wildfire-dependent changes in soil microbiome diversity and function,” *Nature Microbiology*, 7, 1419–1430.

[2] Camargo, A. P. *et al.*, 2023, “IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata,” *Nucleic Acids Research*, 51:D733-D743.

[3] Nayfach, S. *et al.*, 2021, “A genomic catalog of Earth's microbiomes,” *Nature Biotechnology*, 39,499-509.

[4] Camargo, A. P. *et al.*, 2023, “Identification of mobile genetic elements with geNomad,” *Nature Biotechnology*, advance online, doi: 10.1038/s41587-023-01953-y

- **Expanded resources for plant reference genomes and pangenomes:** The Phytozome database added new capabilities to allow users to interrogate new data types at larger scale and to help organize user communities (**EKI02, PLP11**). Users can now interact, for example, with plant pangenomes (multi-scale synteny viewer[5]) and single-cell transcriptomic data.[6] To help organize user communities and facilitate access to multiple data types and metadata from large projects, Phytozome began creating websites tailored to specific user groups and projects, for example the Open Green Genomes project.[7] Phytozome currently hosts 375 annotated and assembled plant genomes, spanning 163 Archaeplastida species, and includes pangenomes for *Brachypodium distachyon, Sorghum bicolor*, and *Camelina sativa*.
- **MycoCosm:** To support the growing user community demand in fungal genomes within the constraints of available computing resources, we doubled the number of fungal genomes in MycoCosm[8] in five years and within the same compute footprint. This goal (**FGP05**) was achieved ahead of schedule through optimization of databases, pipelines, and portal infrastructure, which also identified additional areas for optimization to support further growth of the fungal genome collection. Currently over 2,500 fungal genomes are available from MycoCosm for interactive multi-omics and comparative genomics analyses.
- **PhycoCosm:** To advance algal genomics, we proposed to develop and release a comparative algal resource (**FGP06-2**) to support data distribution and analysis of 100 algal reference genomes expected to be sequenced (**FGP06-5**). PhycoCosm[9] has been developed as the new JGI algal genomics resource. The resource was released, published, and presented to the algal research community through a series of webinars, workshops, and major community meetings (**FGP07-2**). PhycoCosm has already grown to include more than 150 annotated algal genomes and is widely used by algal genomics researchers. It provides a critical foundation for advancing our understanding of the biology of algae and for improving them in support of DOE mission applications.
- **Metabolomics:** To advance our understanding of metabolism and empower users to directly analyze their data we have developed online tools (GNPS) for metabolomic data analysis (**MTB01-2, MTB01-5**). These tools have now become JGI's primary workflow for releasing data to users and for enabling untargeted data analysis.
- **Understanding plant pangenomes**: While high quality reference plant genomes (**PLP04**) and pangenomes (**PLP03**) already constitute a crucial resource for plant biology and breeding, deeper biological discovery from JGI's ever-expanding set of plant genomes requires innovative tools for comparative and integrative analysis. We developed the GENESPACE software to connect phenotype-genotype associates (e.g. Quantitative Trait Loci, QTL) across related species and genetic models (**PLP06-5**). GENESPACE has been integrated into Phytozome and has become a crucial tool for comparative genomics across disciplines. Furthermore, GENESPACE allows direct comparative analyses across DOE plant genomes to identify distantly related sequences that aid our understanding of gene function in plants.

---

[5] http://phytozome.jgi.doe.gov/tools/synteny

[6] http://phytozome-next.jgi.doe.gov/tools/scrna

[7] http://phytozome.jgi.doe.gov/ogg/

[8] https://mycocosm.jgi.doe.gov

[9] https://phycocosm.jgi.doe.gov

- **Genomic resource for uncultivated giant viruses**: We proposed to perform a global survey of giant virus metagenome assembled genomes (MAGs, **MIP06-2**), provide these data to the research community, and implement tools in IMG/VR to enable the research community to identify and explore giant virus MAG data (**MIP06-5**). An in-depth analysis of global giant virus MAGs was completed and published[10]. The *gvclass* tool, which uses giant virus orthologous groups to assign lineage information to giant virus contigs or MAGs, was developed and released through IMG/VR v4. These datasets and the tool are now easily accessible to the user community for sequence-based exploration of giant viruses.
- **Expansion of IMG/VR and development of IMG/PR:** As part of the development of IMG, we proposed to develop new visualization and analysis approaches for large-scale exploration and characterization of uncultivated viruses and plasmids (**MGP04-5**). We also proposed to develop and implement new standards for these mobile genetic elements (MGEs) in collaboration with the Genomics Standard Consortium and the National Microbiome Data Collaborative (NMDC) (**PKI02-2** and **PKI02-5**). New standards for virus genomes (MIUViG) were implemented in IMG/VR versions 3 and 4, alongside new treemap-based browsing and search capabilities.[11,12] Meanwhile, the new IMG Plasmid Resources (IMG/PR) portal was developed to host and provide to users a new large-scale database of plasmid sequences, extracted from genomes and metagenomes in IMG.[13] This new resource complements IMG/VR by providing an unprecedented view of global plasmid diversity, and will serve as a basis for defining new approaches and standards to identify and classify plasmids.

Investigation - Functional Exploration

- **DNA Synthesis Scaling:** To enhance JGI's user capabilities for functional characterization of genes and pathways, we proposed to increase the throughput of the DNA synthesis platform, allowing for an increase in the capacity for DNA design and assembly (**GNT13**). Optimized molecular processes, implementation of newly acquired laboratory automation (purchased with Coronavirus Aid, Relief, and Economic Security (CARES) Act funds) and updated DNA-design software have enabled the platform to increase output from 5.2Mb to 10.7Mb delivered during the 5 year reporting period, with the capability to scale further as user demand increases.
- **Pathway characterization:** We proposed to develop the capacity to design and assemble 1000 pathways per year (**GNT12-2**). To meet this goal, we developed Bio-CAD solutions for the automated design of sequential yeast assembly methods. We also developed protocols for 96-well format automated conjugation for the introduction of pathways into microbial hosts.

---

[10] Schulz, F. *et al.*, 2020, "Giant virus diversity and host interactions through global metagenomics," *Nature*, 578, 432–436.

[11] Roux, S. *et al.*, 2021, "IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses," *Nucleic Acids Research*, 49, D764-D775.

[12] Camargo, A.P. *et al.*, 2023, "IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata," *Nucleic Acids Research*, 51, D733-D743.

[13] Camargo, A.P. *et al.*, 2023, "IMG/PR: a database of plasmids from genomes and metagenomes with rich annotations and metadata," *Nucleic Acids Research*, advance online, doi: 10.1093/nar/gkad964

- **Development of non-model microbial chassis strains:** To enhance our user capabilities for characterizing and identifying the products of pathways of unknown function, we proposed to domesticate multiple phyla of microbes (**GNT14**) and make these bacterial strain engineering products available to users (**GNT15-2**). By developing these genetic toolkits and chassis strains, users are now able to express genes and pathways of interest in systems besides model organisms, elevating the likelihood of expression.
- **Controlled experimental ecosystems:** To advance plant-microbiome science, we adopted fabricated ecosystems, EcoFABs as a platform to conduct reproducible experiments (**PLP08**). This system allows control of all variables (plant genetics, microbiome community members, environmental conditions) and allows non-destructive sampling of liquid growth media for metabolomic analysis as well as non-destructive imaging of living roots. This capability is now available to JGI users through the CSP.
- **Scaled characterization of protein structure and function:** The JGI proposed to develop a platform for high-throughput characterization of structures and function of novel protein families found in environmental genomes and metagenomes (**SSP03**). To date, the JGI and its users, in collaboration with DOE structural and bioimaging resources, have structurally characterized representatives of more than 25 protein families. These new capabilities enable BER researchers to combine genomics, functional, and structural approaches to advance the understanding of protein functions underlying biological and environmental systems.

Integration - Bringing Capabilities Together

- **Cross-facility collaborations for genome-to-structure-to-function:** To support the ambitious goal of the JGI in enabling its users in high-throughput characterization of structures and functions of novel protein families (**SSP03**), the JGI led community efforts exploring the need for BER researchers to combine genomics, functional, and structural approaches offered by the JGI and other DOE Office of Science structural and bioimaging resources to advance their research. The JGI led the organization of a series of “Genomes to Structure and Function” workshops[14] to support further expansion of inter-facility programs.
- **FICUS cross-facility collaborations**: With the goal of broadening JGI’s user base while fostering cross-collaborations with other DOE user facilities, we proposed to first offer capabilities from an additional DOE user facility in the JGI-Environmental Molecular Sciences Laboratory (EMSL) FICUS call (**USP01-2**), then work with other user facilities to streamline cross-user facility requests (**USP01-5**). Since the FY21 FICUS call, several additional user facilities have been included: the X-ray Fluorescence Spectroscopy beamline at the National Synchrotron Light Source II (NSLS-II) in FY21, the Biological Small-Angle Neutron Scattering (Bio-SANS) instrument since FY22, and additionally the Advanced Photon Source (APS) since FY24. Furthermore, a joint strategic hire between the Molecular Foundry (TMF) and the JGI has been established, marking a crucial step in expanding and streamlining cross-user facility requests involving TMF. This collaborative effort ensures that the DOE user base can optimally leverage JGI’s capabilities alongside complementary capabilities offered by other user facilities.

[14] Adams, P. et al., *Genomes to Structure and Function Workshop Report 2022,* United States. https://doi.org/10.2172/1959294

- **NMDC collaboration on data integration**: In a collaboration with the NMDC, we aimed to expand the JGI-EMSL FICUS program to include the NMDC, facilitating data integration across the user facilities through the NMDC (**USP02-05**). Users of both the JGI and EMSL can now express their interest in collaborating with the NMDC during proposal submission. Additionally, the JGI has developed support for handling persistent sample identifiers provided by users. This initial effort will pave the way for establishing sample links across various data types generated at different DOE user facilities.
- **Multi-omics data integration.** To enable systematic functional exploration of genes, the Fungal and Algal Program pursued opportunities for multi-omics data integration, visualization, and analysis (**FGP02-5** and **FGP03-5**). JGI- and EMSL-produced omics data can be integrated with the corresponding genomes in MycoCosm and PhycoCosm. In addition, the JGI developed scripts to detect, import, load and visualize data from public repositories like the Proteomics Identification Database (PRIDE) or National Center for Biotechnology Information (NCBI) to bring over 300 omics datasets to MycoCosm and PhycoCosm. These integrated datasets and new sequencing platforms enabled genome improvement for key fungal genomes broadly used by larger user communities (**FGP02-2**). It is now also possible to publish metabolic models derived from JGI data into the DOE Systems Biology Knowledgebase (KBase) and a KBase modeling app was developed in collaboration with the Argonne National Laboratory (ANL) team (**FGP03-2**).
- **Findable, Accessible, Interoperable, and Reusable (FAIR) data:** JGI invested in the development of a cross-portal search capability (**DSI03-05**) to make JGI data more *findable* and *accessible*. The community can search across plants, metagenomes, microbes, fungi, and algae through the new JGI Data Portal[15] and API. In addition, the JGI worked with KBase to deploy common microservices for LAST and identifier mapping, making JGI data more *interoperable* (**DSI04-05**). Both of these efforts created the foundation for enhanced *reuse* of JGI data through KBase or partner resources.

Interaction - User Engagement

- **Microbiome users.** Engaging the microbial and viral ecology research communities has been a focused effort over the past five years under the umbrellas of the New Lineages of Life (NeLLi) and Viral EcoGenomics and Applications (VEGA) efforts. The JGI has successfully hosted the NeLLi and VEGA Symposia to convene leading scientists to share their cutting-edge research and foster networking opportunities to build these communities (**MGP03-2**). These efforts have been complemented with evolving viral-centric and diversity-focused emphasis areas in the annual CSP call, along with partnering and leveraging other capabilities to tackle grand challenges (**MGP03-5**). In particular, JGI scientists have engaged with members of the International Committee for Taxonomy on Viruses (ICTV) and European Virus Bioinformatics Center (EVBC) to develop new resources for automated taxonomy that could be used by ICTV members, establish new standard and guidelines for the detection and report of RNA viruses, and design a community challenge to systematically evaluate the different viral taxonomy classifiers currently available.
- **Fungal and algal users.** Building on JGI's long history of successfully engaging a fungal genomics user community, we continued for fungi (e.g., **FGP04-5**) and initiated for algae (e.g.,

[15] https://data.jgi.doe.gov

FGP07-2) broad engagement in major user community meetings with workshops, helpdesks, talks, posters, printed materials, sometimes in partnership with EMSL, often inviting CSP Principal Investigators to talk about their projects and share their experiences. The JGI also developed webinars for both MycoCosm and PhycoCosm, in partnership with University of California Cal-Teach students added video tutorials targeted for the next generation of scientists, trained ~100 undergraduates, graduate students, and postdocs in both classrooms and during JGI internships. The results of these efforts are reflected in numbers of publications, new CSP proposals, and MycoCosm and PhycoCosm users.

- **Metabolomics**: The metabolomics program is now providing Users hands-on experience in using GNPS as part of an annual workshop in conjunction with the JGI User Meeting (**MTB01-5**).
- **JGI brand integrity:** To reinforce and sustain the JGI's brand integrity, a baseline inventory and assessment of the JGI existing brand was conducted (**CMO01-2**). The assessment led to the development of a style guide for all new and emerging programs, products, and their deliverables, including portals, software, data management platforms, and resource offerings. The style guide and associated templates for materials such as slides and posters are available to all staff.
- **Communications & outreach**: Ahead of the move to the IGB, there was a need to refresh the onboarding process, specifically formalizing how the JGI is introduced to new hires (**CMO02-2.**) The resulting onboarding overview slide deck is now on the Hiring Supervisor Orientation checklist for employees and supervisors. The overview deck is currently complemented by a video of a JGI tour.

Stewardship

- **Role cards and gap analysis:** As part of JGI's talent management strategy, we identified an opportunity to enhance the effectiveness of our workforce through clarifying roles and responsibilities, as well as systematic analysis of skill gaps that can be filled through cross-training, mentoring and stretch assignment opportunities (**JLT01**). We completed an institutional effort aimed at creation of role cards for all roles, and the majority of JGI employees completed individual role cards. JGI leadership also developed a "Supervisor-Manager Competency Matrix" that was used to identify competency gaps and provide individualized training recommendations to managers and supervisors. All staff are encouraged to develop Individual Development Plans with their supervisors, and cross-training, stretch assignments and new opportunities are now commonplace at JGI.
- **Transparent and equitable hiring processes:** To build a strong and diverse future workforce, the JGI set a goal of revising hiring processes for all roles to become more transparent and equitable (**JLT05**). Several changes have been made to our processes to reach this goal. To help minimize any implicit biases during the recruitment efforts, JGI search committee members are required to watch implicit bias videos to improve their self-awareness and minimize overall bias to ensure fair diversity and inclusion efforts and principles of equal opportunity are considered throughout the interview process. We also implemented a new set of interview and hiring guidelines in alignment with the Berkeley Lab Biosciences Area.
- **Develop a DE&I strategy and mechanisms to ensure implementation:** The JGI set a goal of developing a formal Diversity, Equity, and Inclusion (DE&I) strategy (**JLT12**). A first version of this strategy was developed by the JGI DE&I Working Group in 2018, with annual updates in following

years. The strategy includes initiatives related to outreach to the scientific community, the JGI work environment, and employee development and training. It also includes specific actions and goals that serve as performance indicators. The JGI DE&I Working Group also led JGI-wide employee engagement surveys for the last three years and worked with the JGI leadership team on actions following on these surveys. The JGI also developed strong partnerships with the Biosciences DEI Committee and the Lab's Inclusion, Diversity, Equity, and Accountability (IDEA) Committee and shared the results of JGI engagement surveys, best practices and lessons learned that have been invaluable to the IDEA Council as they develop the Berkeley Lab-wide engagement survey to launch in 2024.

- **Automate lab processes for efficiency and throughput improvements:** The JGI constantly strives to seek automation opportunities to increase throughput and efficiency through the use of robotics and technology solutions. The library creation and gene synthesis processes were upgraded through the purchase of robotic platforms that significantly reduced the per-unit cost of processing these samples (**OPS07**). The JGI Automation Team also leveraged their programming skills to expand library creation from 96- to 384-well processes, greatly increasing the capacity and reducing the cost of the process. Finally, the JGI leveraged the newest sequencing technologies by adopting the Pacific Biosciences Revio and Illumina NovaSeq X instruments, reducing the cost of sequencing by approximately 60% (**OPS08**).

Strategic Implementation Highlights

In this section, we provide a more detailed overview of selected implementation highlights from the past five years. These examples were chosen to illustrate how directions defined by the strategic planning process result in the implementation of overarching efforts and initiatives.

Highlight 1: Secondary Metabolites

Secondary metabolites play critical roles in environmental processes including nutrient acquisition, cell-host and cell-cell communication, antagonism, and mediating symbiotic relationships. Recognizing that JGI is in a prime position to further advance research in this area, secondary metabolites were the subject of a "strategic thrust" in the 2018 strategic plan. In 2021 the JGI launched the Secondary Metabolites Science Program under the direction of the JGI Director, Nigel Mouncey, leveraging his extensive experience in research and commercialization of secondary metabolites. The Secondary Metabolites Science Program has three focus areas: 1. Discovery of novel secondary metabolites by combining innovative computational genome mining with functional expression and molecular profiling, 2. Functional characterization of secondary metabolites taking advantage of host engineering and metabolomics capabilities at JGI, and 3. Providing support for a new initiative, the Earth's Secondary Metabolome Initiative (ESMI, **Fig. 3**).

Working with collaborators, most notably Professor Ben Shen from the Natural Products Discovery Center[16] at the University of Florida-Scripps, the JGI has expanded the diversity of secondary metabolites through sequencing culture collections that harbor organisms that are prolific producers of secondary metabolites. The JGI has developed a new pipeline to predict biosynthetic gene clusters (BGCs) for secondary metabolites through incorporating existing prediction tools (**MGP07**). This pipeline is now part of

[16] https://shen.scripps.ufl.edu/natural-product-discovery-center-npdc/

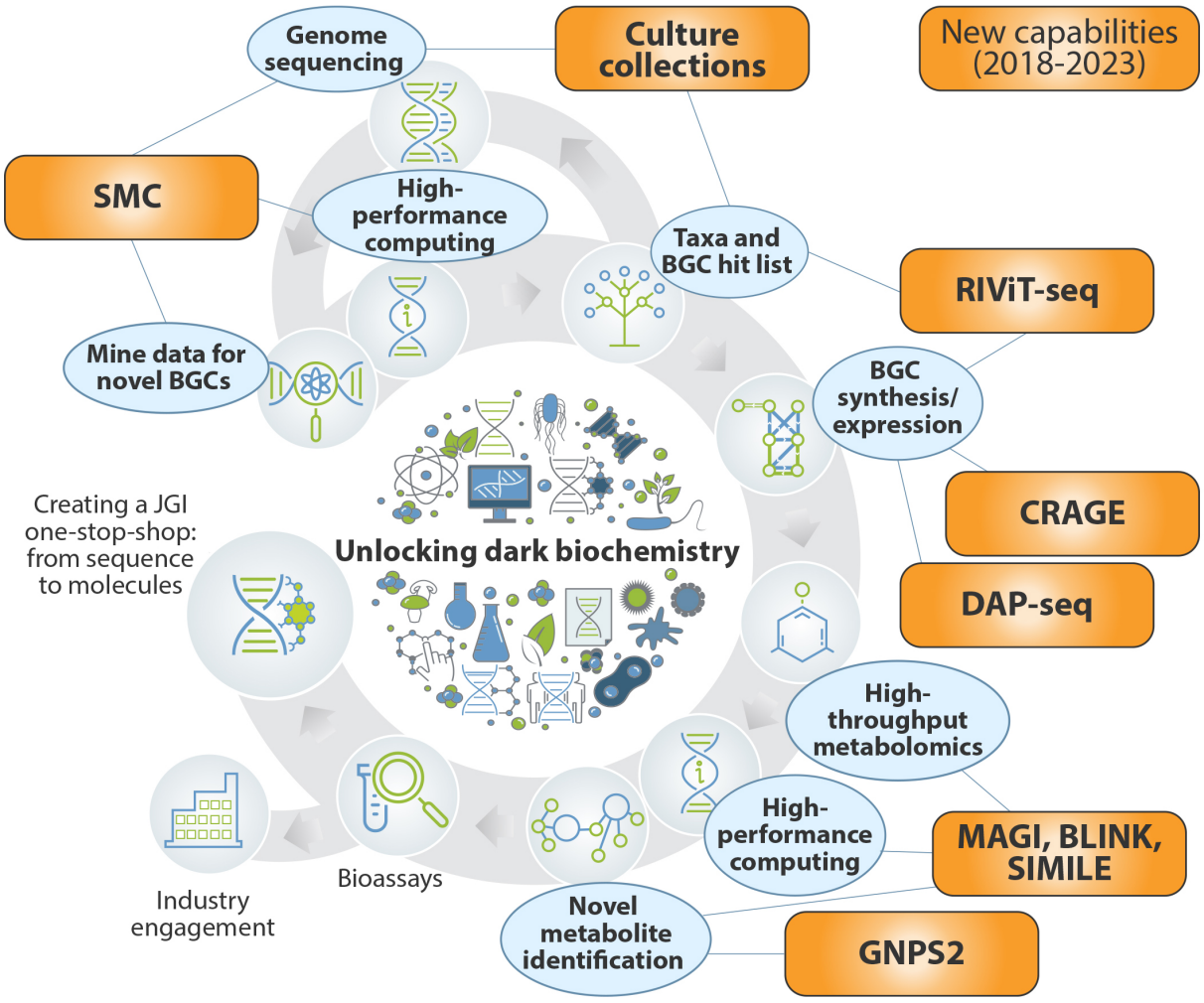


Figure 3: Integrating JGI capabilities for exploration of Earth's secondary metabolome. Center: The initial vision for the Earth's Secondary Metabolome Initiative (ESMI) as outlined in the 2018 Strategic Plan. Orange boxes: Selected new resources and capabilities in support of JGI's "Secondary Metabolites" strategic thrust.

a new JGI data portal for BGCs, the Secondary Metabolism Collaboratory (SMC)[17]. This unique resource represents the world's largest openly accessible repository of BGCs and currently houses more than 13 million BGCs from over 1.3 million bacterial genomes (**PKI06**). In SMC, users can search for BGCs of interest but also upload their own genomes and run these through the SMC prediction pipeline. Additionally, through user testing, features for users to curate BGCs and hold discussions have been included. With the official launch of SMC in 2023, the previous JGI system for BGC analysis (IMG-ABC[18]) has now been sunsetted.

The Secondary Metabolites Program has partnered with JGI's DNA Synthesis and Metabolomics Science Programs to offer a suite of capabilities that aid in the detection, identification, and characterization of BGCs and their metabolites. The JGI has employed multi-chassis engineering using the chassis-independent recombinase-assisted genome engineering (CRAGE) technology for rapid large-scale assem-

[17] https://smc.jgi.doe.gov/

[18] Palaniappan, K. *et al.*, 2020, "IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase", *Nucleic Acids Research* 48:D422-D420

bly, cloning and expression of BGCs (**GNT12, GNT13, GNT14, GNT15**). CRAGE enabled rapid integration of complex genetic constructs directly into the chromosome of diverse microbial species. This multi-chassis approach enabled rapid characterization of secondary metabolite BGCs and identification of chassis strains favorable for production of diverse chemicals[19,20]. The utility of CRAGE was extended by combining CRISPR technologies such as CRISPR-based activation and inhibition of target gene expression as well as CRISPR-based gene deletion. These technologies were used to perform gain- and loss-of-function studies, and the teams were able to efficiently characterize the function of diverse secondary metabolite BGCs[21] (**MIP07**, **MTB04**, **MTB05**, **SSP02**, **SSP03**, **SSP06**). Additionally, the utility of the CRISPR technologies was scaled. In collaboration with users, the JGI established an efficient way to design and build sgRNA libraries using oligonucleotide pool library technologies. Combining amplicon sequencing and sequence analysis, the JGI was able to demonstrate CRISPR-based high throughput functional genomics tools to enable rapid characterization of gene functions and engineering of production strains for target metabolites (**SSP04, SSP05**)[22]. The Metabolomics Program has deployed high throughput metabolomics analysis, fabricated ecosystems, and cheminformatic tools to provide important insights into secondary metabolites and their function in biological and ecological processes (**MTB05-2, MTB05-5**). This has resulted in numerous publications and powerful new tools for secondary metabolite discovery and integration with genomic information. Highlights include new cheminformatic and bioinformatic tools led by the Metabolomics Program: MAGI[23], SIMILE[24], and BLINK[25] as well as GNPS2 capabilities[26,27]. Together, these tools increase the throughput for comparing spectra approximately 1000-fold,

---

[19] Yang, G. *et al.*, 2019, “CRAGE enables rapid activation of biosynthetic gene clusters in undomesticated bacteria”, *Nature Microbiology*, 4, 2498–2510.

[20] Ke, J. *et al.*, 2022, “Development of platforms for functional characterization and production of phenazines using a multi-chassis approach via CRAGE”, *Metabolic Engineering*, 69, 188–197.

[21] Ke, J. *et al.*, 2022, “CRAGE-CRISPR facilitates rapid activation of secondary metabolite biosynthetic gene clusters in bacteria”, *Cell Chem Biol*, 29, 696–710.

[22] Schwartz, C. *et al.*, 2019, “Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast *Yarrowia lipolytica*”, *Metabolic Engineering*, 55, 102–110.

Bowman, E.K. *et al.*, 2020, “Bidirectional titration of yeast gene expression using a pooled CRISPR guide RNA approach”, *PNAS*, 117, 18424–18430.

Baisya, D. *et al.*, 2022, “Genome-wide functional screens enable the prediction of high activity CRISPR-Cas9 and -Cas12a guides in *Yarrowia lipolytica*”, *Nature Communications*, 13, 922.

Lupish, B. *et al.*, 2022, “Genome-wide CRISPR-Cas9 screen reveals a persistent null-hyphal phenotype that maintains high carotenoid production in *Yarrowia lipolytica*”, *Biotechnology and Bioengineering*, 119, 3623–3631.

Ramesh, A. *et al.*, 2023, “acCRISPR: an activity-correction method for improving the accuracy of CRISPR screens”, *Communications Biology*, 6, 617.

[23] Erbilgin, O. *et al.*, 2019, “MAGI: A Method for Metabolite Annotation and Gene Integration”, *ACS Chem Biol*, 14, 704–714.

[24] Treen, O. *et al.*, 2022, “SIMILE enables alignment of tandem mass spectra with statistical significance”, *Nature Communications*, 13, 2510.

[25] Harwood, T. *et al.*, 2023, “BLINK enables ultrafast tandem mass spectrometry cosine similarity scoring”, *Scientific Reports*, 13, 13462.

[26] Nothias, L. *et al.*, 2020, “Feature-based molecular networking in the GNPS analysis environment”, *Nature Methods*, 17, 905-908.

[27] Petras, D. *et al.*, 2022, “GNPS Dashboard: collaborative exploration of mass spectrometry data in the web browser”, *Nature Methods*, 19, 134-136.

greatly increase the number of compounds that can be matched to experimental spectra, and integrate genomic and metabolomic data to annotate both genes and metabolites.

The JGI further realized that in order to successfully express BGCs for characterizing their activities and products, a need existed to better understand how BGCs are regulated in their native and heterologous hosts. Regulon Identification by *in vitro*-sequencing (RIViT-seq), an *in vitro* transcription methodology to identify the regulons controlled by sigma factors was developed and used to characterize the regulons of 12 sigma factors from *Streptomyces coelicolor*[28]. DAP-seq has been employed for multiple secondary metabolite producers and gene regulatory networks created. Lastly, the JGI is developing cell-free expression systems for BGCs building on earlier cell-free work from an ETOP project with Professors Hal Alper and Michael Jewett (**GNT10**).

Highlight 2: Microbiome Data Science

The rapidly growing field of microbiome data science is poised for transformative advances, fueled by the exponential growth of microbiome data and technological advancements in data processing, analysis, and visualization. While microbiome data science provides major opportunities relevant to health, agriculture, bioenergy, and environmental research, innovative solutions and community-wide collaborations will be required to address the grand challenges in these areas and facilitate groundbreaking discoveries.

Acknowledging the JGI's pivotal role in propelling the field forward, the 2018 strategic plan designated microbiome data science as a second "strategic thrust" along with JGI scientists launching a separate effort to establish the National Microbiome Data Collaborative (NMDC)[29]. The focus on microbiome data science was influenced by the availability and unique strengths of the Integrated Microbial Genomes with Microbiome Samples (IMG/M) system, a pioneering platform that integrates high-quality microbiome data to enhance research and promote collaborations within this nascent field. Driven by this strategic focus, the JGI launched a series of big data-driven initiatives that target areas including the characterization of "functional dark matter" within microbiomes (**PKI01**), the exploration of metatranscriptomes for novel RNAs[30,31,32], the identification of new viruses and prediction of their host interactions (**MGP05**, **MIP06**), and the integration of macroecological theory into microbiome data science (**MGP06**). These efforts aimed to bridge annotation gaps, expand our understanding of RNA diversity, uncover the breadth of viral diversity, and apply macroecological principles to microbial ecology.

The observation that a large fraction of predicted genes in metagenomic data cannot be functionally annotated using standard databases of gene function, such as Pfam and COG, has led to the realization that functional annotation of sequencing data has been lagging significantly behind for metagenomic data. To bridge this gap, the JGI embarked on a large-scale effort to organize and explore the protein

---

[28] Otani, H. *et al.*, 2022, "RIViT-seq enables systematic identification of regulons of transcriptional machineries", *Nature Communications*, 13, 3502.

[29] https://microbiomedata.org

[30] Neri, U. *et al.*, 2022, "Expansion of the global RNA virome reveals diverse clades of bacteriophages", *Cell*, 185, 4023-4037.

[31] Fremin, B. J. *et al.*, 2022, "Identification of over ten thousand candidate structured RNAs in viruses and phages", *Comput Struct Biotechnol J*, 21, 5630-5639.

[32] Lee, B. D. *et al.*, 2023, "Mining metatranscriptomes reveals a vast world of viroid-like circular RNAs", *Cell*, 186, 646-661.

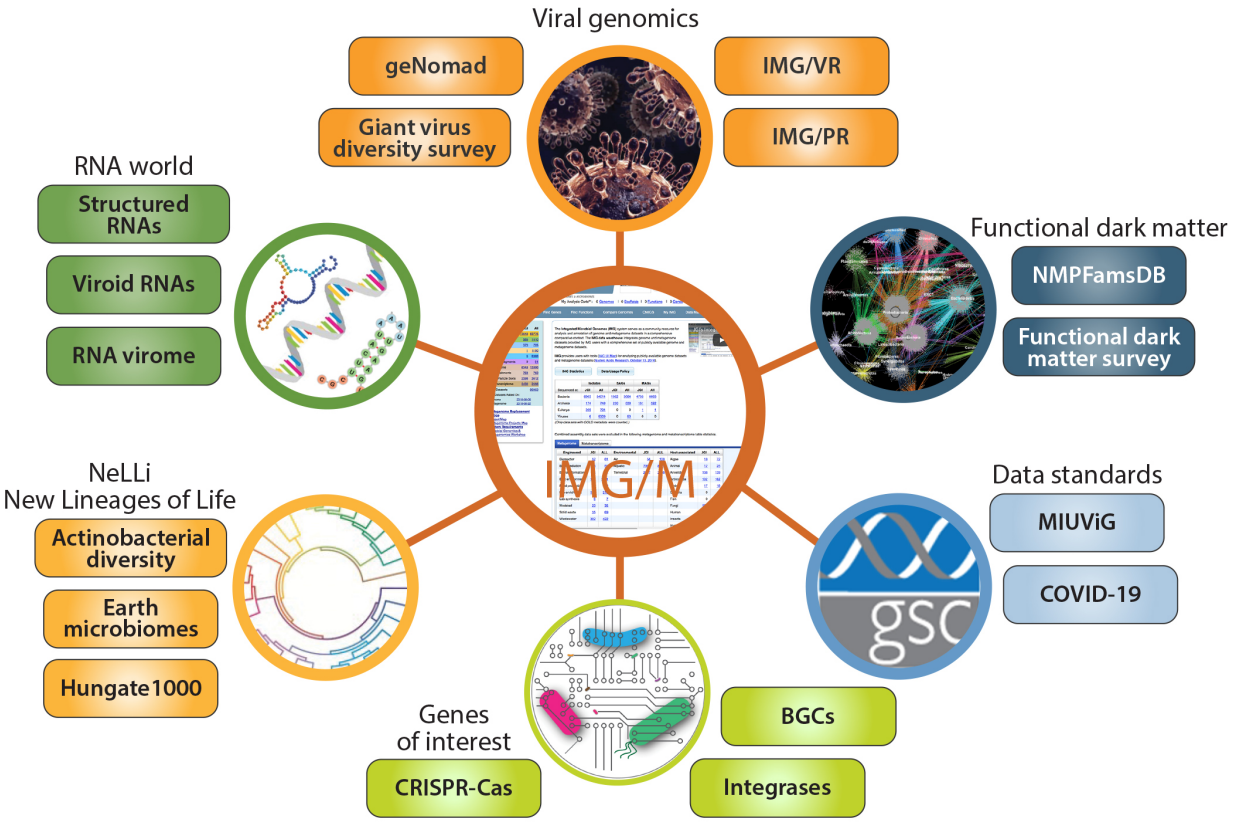


Figure 4: Accomplishments in Microbiome Data Science. Center: Six major areas of interest in microbiome data science, as described in the 2018 Strategic Plan. Boxes: Selected accomplishments in microbiome data science.

sequence landscape in metagenomic data, identifying novel protein families lacking annotation for targeted analysis (**PKI03**). To achieve that, the team focused on exploring the protein diversity in metagenomes through generation of protein families in two independent projects: the novel metagenome protein families (NMPFs) and geNomad **(PKI01, PKI02, PKI03, PKI05**).

In the NMPF project, which constitutes the largest survey of its kind performed to date, the JGI team collaborated with more than 100 users and data contributors who were part of the Novel Metagenome Protein Families Consortium to perform a systematic analysis of nearly 30,000 metagenomes[33]. This effort identified 1.2 billion protein sequences without similarity to previously known proteins and doubled the number of known protein families. From a subset of more than 13,000 high-confidence predictions of three-dimensional protein structures, we found 162 completely novel structural folds, demonstrating that metagenomic data enables novel insights into the fundamentals of protein biology. The novel proteins and protein families data are now available to the community through the Novel Metagenome Protein Families Database (NMPFamsDB) for metagenome- and metatranscriptome-derived protein families[34]. The approaches developed through these efforts are rooted in comprehensive protein clustering and showcase the versatility of metagenomic protein families in unlocking new dimensions of understanding within biological sequencing data.

[33] Pavlopoulos, G. A. *et al.*, 2024, "Unraveling the functional dark matter through global metagenomics", *Nature*, 622, 594-602.

[34] Baltoumas, F. *et al.*, 2024, "NMPFamsDB: a database of novel protein families from microbial metagenomes and metatranscriptomes", *Nucleic Acids Research*, 52, D502-512.

We further explored the utility of protein families during the development of the geNomad tool[35], which enables large-scale discovery of mobile genetic elements (MGEs), namely viruses and plasmids, within sequencing data. Benchmarking showed that geNomad vastly outperforms other tools for virus and plasmid identification. To allow accurate detection of MGEs, geNomad employs a set of 230,000 protein families that are exclusive to viruses, plasmids, or chromosome sequences. The majority of these families were obtained by a large-scale clustering of more than 230 million proteins. To identify viruses and plasmids, geNomad annotates query sequences using its marker dataset and then performs classification based on the specificity of the markers that matched proteins encoded by the query. To facilitate the analysis and promote utilization of the data from the larger user community, the JGI released all new data generated from geNomad through two new IMG data marts, the IMG/VR[36] for viruses and the IMG/PR[37] for plasmids.

Beyond the examples described in detail above, **Figure 4** provides an overview of multiple additional activities and accomplishments at the JGI in microbiome data science thrust. They are organized around six major opportunity areas for JGI related to microbiome data science identified in the 2018 strategic plan, including viral genomics, functional dark matter, data standards, genes of interest, NeLLi, and the RNA world. Boxes show selected accomplishments in each of these areas that were driven by the strategic thrust in microbiome data science.

Highlight 3: Algal Genomics

The 2018 strategic plan identified algal genomics as a significant opportunity area for the JGI. Algae are phylogenetically and functionally diverse and, due to their contributions to total photosynthetic production, they have a major impact on Earth's carbon cycle (**Fig. 5**, center). They have also been identified as promising targets for biofuel and bioproduct development. Driven by the strategic plan, the JGI has launched or completed a wide range of algal user science projects (**Fig. 5**, colored boxes) and developed new resources and capabilities in support of the algal genomics community.

Enabled by resources from across JGI Science Programs, the JGI empowered the algal research community by producing and bringing together algal genomic and other -omic data and tools into a single public resource hosted by JGI, called PhycoCosm[38] (**FGP06-2**). Its release and an associated publication[39] were followed by a webinar and multiple presentations at major user community meetings such as the Phycological Society of America Annual Meetings, the Conferences on Algal Biomass, Biofuels and Bioproducts, and the Algal Biomass Summits (**FGP07-2**). JGI also contributed to the *Interagency Algal*

[35] Camargo, A. P. *et al.*, 2023, "Identification of mobile genetic elements with geNomad", *Nature Biotechnology*, advance online, doi:10.1038/s41587-023-01953-y

[36] Camargo, A. P. *et al.*, 2023, "IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata", *Nucleic Acids Research*, 51, D733-D743.

[37] Camargo, A. P. *et al.*, 2024, "IMG/PR: a database of plasmids from genomes and metagenomes with rich annotations and metadata", *Nucleic Acids Research*, 52, D164-D173.

[38] https://phycocosm.jgi.doe.gov/

[39] Grigoriev, I. V. *et al.*, 2021, "PhycoCosm, a comparative algal genomics resource", *Nucleic Acids Research*, 49, D1004-D1011.

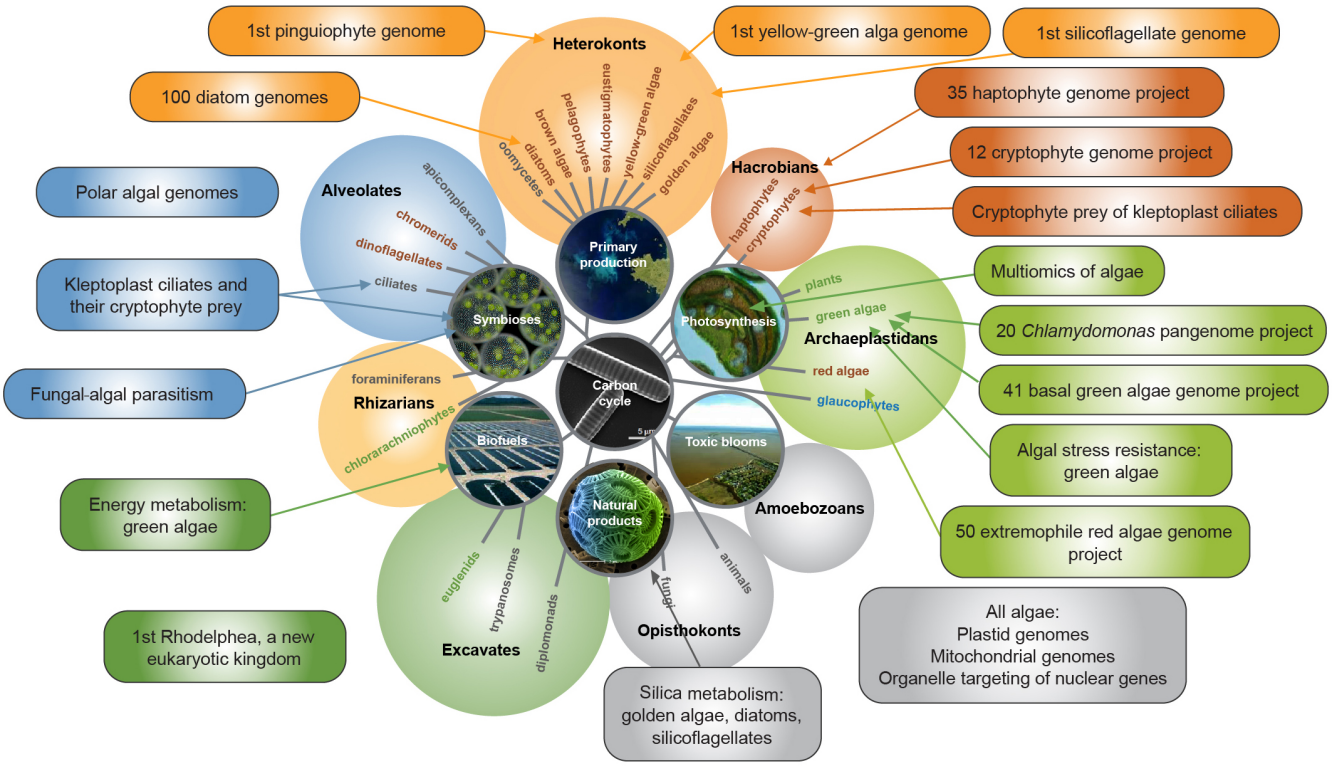


Figure 5: New algal genomics efforts. Center: Taxonomic diversity of algae and resulting opportunities for basic and applied research, as outlined in the 2018 Strategic Plan. Colored boxes: New projects and efforts launched or completed by JGI in support of algal genomics research.

Working Group to produce a "Federal Activities Report on Bioeconomy: Algae" that summarizes available resources and opportunities for algal researchers[40].

These outreach efforts resulted in several new algal CSP projects targeting specific large groups of diatoms, red algae, haptophytes, and others (see **Fig. 5**; **FGP07-5**). Additional CSP projects aim at first sequenced genomes in heretofore empty clades (e.g. silicoflagellates and new kingdom Rhodelphea) or pairing genomes of partners with symbiotic and other ecological interactions of importance to carbon cycling and biofuel production (e.g., ciliate-cryptomonad kleptoplasty and fungus-chlorophyte parasitism). To support these projects, JGI continues to improve algal genome sequencing, assembly, and annotation techniques, as well as to add new genomic and multi-omic analyses including transcription factor binding site (DAP-Seq) and methylation site discovery, pangenome dissection, assembly and annotation of organellar genomes, and prediction of organellar targeting of nuclear gene products. In partnership with Los Alamos National Lab (LANL) and with support of the Bioenergy Technologies Office of DOE, JGI built a multi-omics pipeline to better understand the biology of high-productivity algal strains on a molecular level. The integrated analysis of multi-omics datasets in the context of metabolic pathways and gene networks enables predictive modeling to identify gene targets for strain improvement[41,42].

[40] U.S. Department of Energy. 2020. Federal Activities Report on the Bioeconomy: Algae. Washington D.C.: U.S. Department of Energy. DOE/EE-2009. doi:10.2172/1656710

[41] Calhoun, S., *et al.*, 2021, "A multi-omic characterization of temperature stress in a halotolerant Scenedesmus strain for algal biotechnology", *Communications Biology*, 4, 333.

[42] Calhoun, S., *et al.*, 2022 "Multi-omics profiling of the cold tolerant *Monoraphidium minutum* 26B-AM in response to abiotic stress", *Algal Research,* 66, 102794.

To enable algal phylogenomics and comparative genomics the algal team adds to CSP-driven projects[43,44] by actively importing non-JGI sequenced genomes into PhycoCosm, from both public databases and private collaborations. In addition, metagenomes are now being tapped as a source of uncultivated algal genomes[45] (e.g., Duncan et al., 2022). The result of these combined efforts is a near-doubling of the number of genomes hosted at PhycoCosm since the 2021 paper (**FGP06-5**). Building on these accomplishments, the JGI expects to further scale up genome production of diverse algae from CSPs and other sources, to continue to develop analysis and visualization tools for an increasing array of multi-omics data types integrated in PhycoCosm, and amplify outreach efforts to algal researchers including training of the next generation of scientists.

Highlight 4: Plant Pangenomes

Ongoing technological and computational advances across the genomics community and within the JGI plant program have massively accelerated the production of long and accurate sequencing reads. Over this period, we have scaled long-read sequencing and deployed Pacific Biosciences HiFi technology for highly accurate sequencing. When combined with our optimized bioinformatic assembly pipelines, HiFi sequences have allowed us to produce over 100 genomes a year (**PLP01-02**), including those from plants with large, complex, and polyploid genomes (**PLP01-05**). These genomes include 32 inbred Sorghum accessions (genome size: 800M base pairs, bp), 16 haplotypes from outbred tetraploid switchgrass (genome size: 1.1Gbp), all six haplotypes of outbred hexaploid big bluestem (genome size: 5.3Gbp), the 12Gbp Ceratopteris fern, and perhaps the most complex plant genome sequenced to date, hybrid dodecaploid sugarcane (genome size: 10 Gbp).

These efforts have allowed the JGI to produce a reference quality genome for nearly all major lineages of plants (**PLP04-02**), and ongoing work under the Open Green Genomes project will allow us to fill in the few persisting phylogenetic gaps. Despite advances in bioinformatics, some lineages remain recalcitrant to standard DNA extraction, HiFi sequencing, and genome assembly methods. We are gradually overcoming these challenges and are finally producing assemblies for some of the most problematic plant species. Overcoming these technical issues and accumulating tissue from rare botanical specimens will let us complete a set of reference genomes that evenly span plant evolutionary diversity (**PLP04-05**). Combined, this resource will provide a closely related reference genome for the vast majority of plant genetics projects, which will improve variant detection and accelerate information transfer between genetic models and experimental species and crops.

To complement the phylogenetic distribution of genomic resources, we have sought to better understand genetic variation within species. Such pangenomes simultaneously capture the majority of sequence variation within a species and reduce bias when detecting variants from population-scale genetics experiments. The improved capacity for assembly has allowed us to construct and annotate high quality pangenomes (**PLP03-02**) in Sorghum (32 genomes), poplar (36 genomes, with many more forthcoming), *Camelina sativa* (12 genomes), and *Brassica rapa* (8 genomes; **PLP03-05**). Our ongoing downstream

---

[43] Ye, N. *et al.*, 2022, “The role of zinc in the adaptive evolution of polar phytoplankton”, *Nature Ecology Evolution*, 6, 965-978.

[44] Dorrell, R. G. *et al.*, 2021, “Convergent evolution and horizontal gene transfer in Arctic Ocean microalgae”, *Life Sci Alliance*, 6, e202201833.

[45] Duncan, A. *et al.*, 2022, “Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans”, *Microbiome*, 10, 67.

analytical method development (see below), has allowed us to access structural variation and link molecular to phenotypic variation (**PLP06-02**). Combined, these resources and discoveries will provide a comprehensive resource for DOE plant improvement efforts (**PLP06-05**).

While our many high quality reference genomes (**PLP03-02**, **PLP04-05**) are themselves a crucial resource for plant biology and breeding, biological discovery from this expanding set will come from comparative and integrative analysis (**PLP06-02**, **PLP06-05**). We are tackling this significant task through development of novel publicly available software and nuanced analysis pipelines in collaboration with JGI users and scientists. For example, we developed the GENESPACE software to compare multiple genomes using gene synteny across and within pangenomes[46]. GENESPACE has been integrated into Phytozome and has become a crucial tool for comparative genomics across disciplines. Furthermore, GENESPACE allows direct comparative analyses across DOE plant genomes to identify distantly related sequences that aid our understanding of gene function in plants. Despite the success of GENESPACE, the next step of pangenomic integration remains to be solved for JGI plant program projects, where short read "resequencing" efforts characterize the bulk of the genetic diversity in a species and the goal is to link the pangenome to species wide diversity.

Highlight 5: JGI-UC Merced Internship Program

The JGI is committed to cultivating long term educational partnerships as part of efforts to train the next generation workforce and collaborating researchers. One such partnership is with UC Merced, which began in Summer 2014 with 2 graduate student interns. The students gain hands-on experience in genomic research, applying lab work and computational biology to aspects of energy and environmental challenges. A decade on, the interns comprise half of the JGI's summer intern cohort and come from both undergraduate and graduate programs at UC Merced[47].

Figure 6: Attendees of the 10-year anniversary symposium of the JGI-UC Merced Internship Program.

To evaluate the success of the JGI-UC Merced Internship Program, an independent impact assessment report was generated in 2020, fulfilling one of the strategic plan milestones (**CMO07**). Based on feedback provided by both intern alumni and mentors who had been part of the program's first five years, the assessment highlighted the rapid growth of the program and made recommendations to bolster its efficacy and sustainability. For example, UC Merced students reported seeing a substantial increase in their technical training and communication skills. They also expressed more interest at the end of their internships to pursue graduate studies and consider careers in the national lab ecosystem.

---

[46] Lovell, J., *et al.*, 2022, "GENESPACE tracks regions of interest and gene copy number variation across multiple genomes", *Elife*, 11, e78526.

[47] https://jgi.doe.gov/researching-and-solving-real-world-problems-with-the-2023-jgi-uc-merced-interns/

To date, more than 70 undergraduate and graduate UC Merced students have been paired with JGI mentors, contributing to ~40 JGI projects. In Summer 2023, the internship program marked its 10th anniversary (**Fig. 6**), and now has plans to expand this partnership beyond the summer training experience[48,49]. Through the DOE's Reaching a New Energy Sciences Workforce (RENEW) initiative, an approved proposal spearheaded by the founders of the JGI-UC Merced Internship Program aims to formalize and institutionalize the training components of the program, in line with the recommendations of the impact assessment report.

Highlight 6: Enhanced Data Reproducibility and Resilience

Reproducible research depends on access to both the data and analysis workflows utilized to generate a result. The JGI has taken steps in both of these areas to make reproducible science easier for staff and our User community. In 2020, the JGI kicked off a project to build a new Data Portal (JDP)[50] in order to provide one location to find, access, and download public data across JGI's Science Programs and projects (**DSI03-5**). The JDP team utilizes a user-centered design approach to deploy iterations of the resource that are tested and improved through user feedback (**DSI08-2**). Staff have conducted more than one hundred interviews with stakeholders that informed design, functionality, and prioritization. In addition to a new graphical interface, JGI has deployed an application programming interface (API)[51] where users can query JDP from a script or Jupyter notebook. Since the number of files returned for a given query may change as the JGI generates or updates data, the JDP team provided a mechanism to save query results for a given date. This means that a user can share the queries and results files to facilitate the reproduction of analysis in the future.

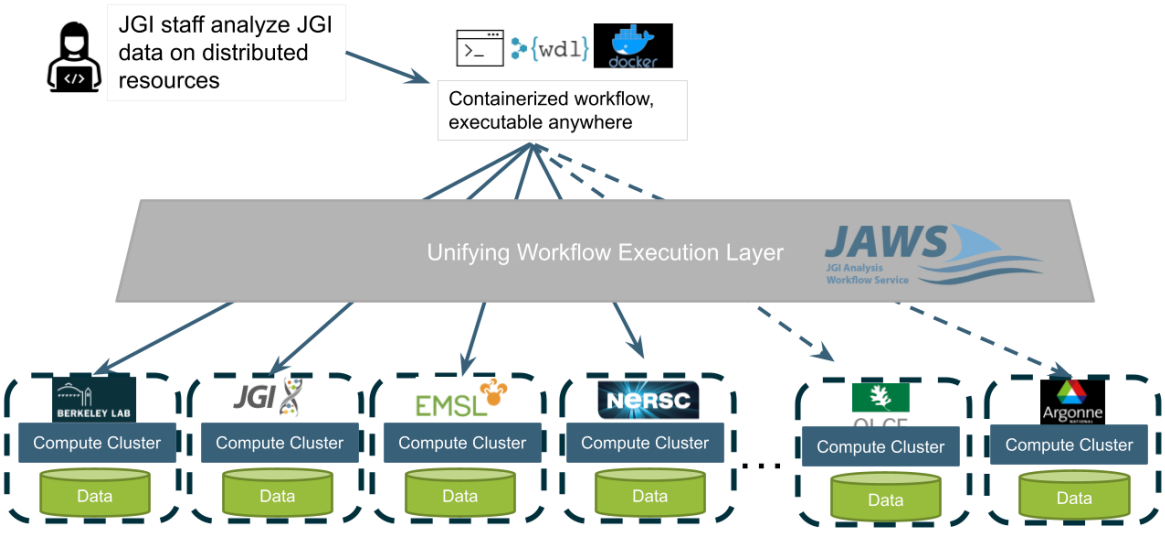


Figure 7: **The JGI Analysis Workflow Service (JAWS)**. JAWS is one entrypoint for five DOE computing resources and one cloud provider. Addition of Oak Ridge and Argonne sites planned for 2024.

[48] https://jgi.doe.gov/celebrating-a-decade-of-science-through-the-jgi-uc-merced-genomics-internship-program/

[49] https://youtu.be/SJyULTfv_dI

[50] https://data.jgi.doe.gov/

[51] https://files.jgi.doe.gov/apidoc/

Reproduction of analysis requires instantiating the same computing environment that was used to run a workflow on a given data set. For many years scientists used virtual machines to share environments, however, Docker containers have emerged as lightweight specifications that are more easily shared through resources like DockerHub. DOE computing facilities and the cloud provide support for Docker, so once a workflow has been ported to Docker, it is possible to run it on multiple computing resources. There are several other hurdles to seamless workflow execution across diverse computing sites, such as scheduler configuration, reshaping the workflow to fit a site's hardware, optimizing for data storage, and data management. To facilitate resilient distributed computing, the JGI built JAWS, the JGI Analysis Workflow Service. JAWS sites are running on NERSC, Berkeley Lab's Lawrencium cluster, JGI's Dori cluster, Tahoma at EMSL, and the cloud (**DSI01-5, Fig. 7**). Staff require a protocol, Docker container, and input data set to use JAWS. JGI staff are leveraging the Workflow Description Language (WDL) as the protocol to describe the steps to execute their analyses. The WDL standard is supported by several workflow management systems and JGI has selected Cromwell, from the Broad Institute, for JAWS.

The JDP API and JAWS software infrastructure are a foundation for reproducible public research that also make JGI more resilient to outages or policy changes at computing sites. These components are key collaboration points with partners including NMDC, KBase, and EMSL.

Highlight 7: DNA Affinity Purification Sequencing

Gene regulation is a major function shared across biology and largely controlled by transcription factors encoded in genomes. We developed DNA Affinity Purification Sequencing (DAP-Seq) as a high-throughput method to define the DNA sequence motifs bound by transcription factors (**GNT08**). This approach is now routinely available in JGI user calls and has been applied to bacteria, fungi, and plants.

Progress towards these milestones included the development of an optimized end-to-end liquid handler method for processing sets of transcription factors (TFs) using 96-well plates in DAP-seq assays. This method allows for efficient handling and analysis of TF sets covering entire species, facilitating large-scale studies of gene regulation and annotation of candidate gene regulatory elements. We optimized two separate workflows, one tailored specifically towards prokaryotic and archaeal TFs, and the second optimized for eukaryotic TFs including plants[52]and fungi[53].

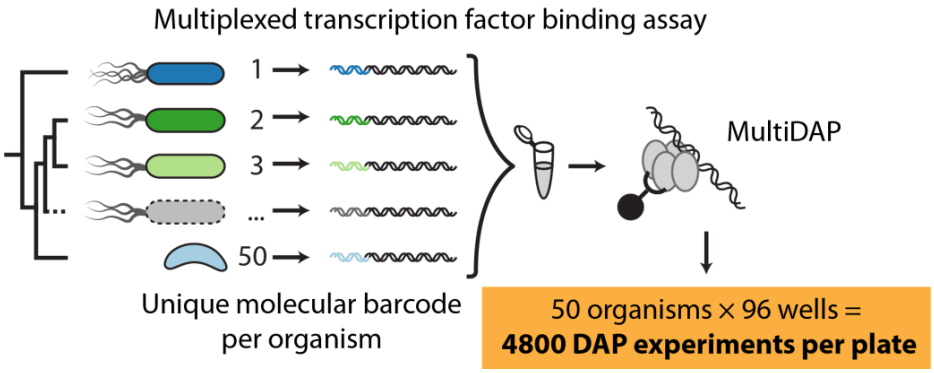


Figure 8: The multiDAP-seq methods provides a major increase in throughput of DAP-seq experiments..

[52] Hyden, B., *et al.*, 2023, "Functional analysis of Salix purpurea genes support roles for ARR17 and GATA15 as master regulators of sex determination", *Plant Direct*, 7, e546.

[53] Huberman, L. B., *et al.*, 2021, "DNA affinity purification sequencing and transcriptional profiling reveal new aspects of nitrogen regulation in a filamentous fungus", *PNAS*, 118, e2009501118.

Further advancements were made with the introduction of a multiplexed method, named multiDAP-seq, which enables the processing of dozens of microbial genomic DNA samples within a single reaction with a minimal increase in reagent costs[54]. This approach can generate the equivalent of 4,800 genome-wide transcription factor binding maps from a single 96-well plate experiment (**Fig. 8**). We demonstrated the utility of this method and how the resulting data can be applied in comparative genomic analysis, revealing conserved gene promoter architectures and aiding in the functional annotation of microbial genomes based on TF binding data. The multiDAP-seq method has been made available to JGI users, attracting interest particularly from researchers studying bacteria and fungi.

To complement the laboratory methods, an automated computational analysis pipeline for DAP-seq data has been developed. This pipeline utilizes the JGI Analysis Workflow Service (JAWS) to enable efficient and parallel processing of DAP-seq and multiDAP-seq datasets to provide JGI users with directly interpretable results including TF binding sites, motifs, and candidate target gene assignments.

Highlight 8: Move to the Integrative Genomics Building

In 2019, Berkeley Lab completed the construction of JGI's new home on the Berkeley Lab site, the Integrative Genomics Building (IGB, **Fig. 9**). The JGI Operations Department planned and executed a successful transition of the JGI's staff, instrumentation, and processes to the new building, while finding ways to leverage the new closer proximity to the Berkeley Lab.

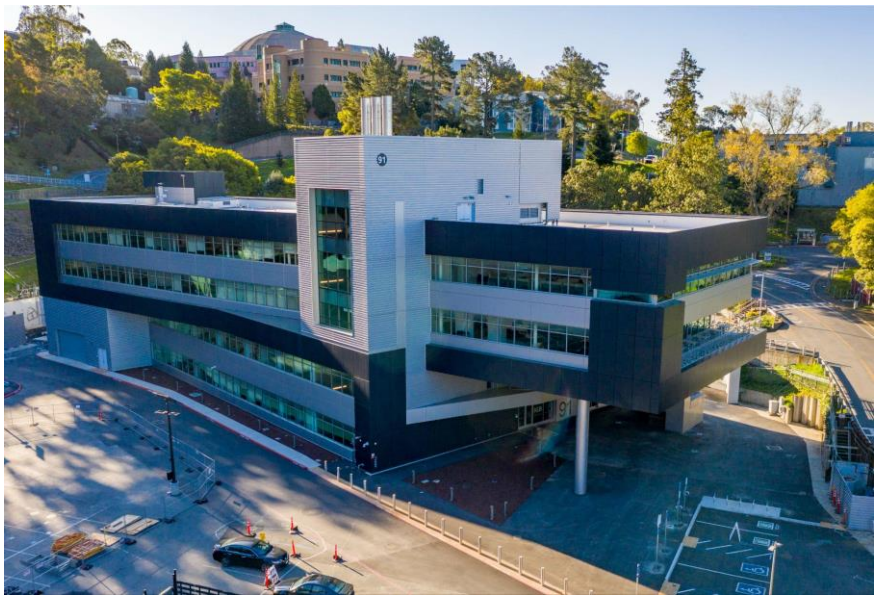


Figure 9: The Integrative Genomics Building. The new home of the JGI is located on the Lawrence Berkeley National Lab main campus in Berkeley, California.

One of the most impactful transitions was the integration of the high-volume shipping and receiving area (**OPS04**). The IGB would be one of highest volume delivery points on the Berkeley Lab campus and the team met with Lab stakeholders to ensure timely, safe, and well-coordinated deliveries several times a day. At the time, most package deliveries at the Lab involved individual packages to offices and lab areas throughout the Lab, so the IGB team assisted Lab Shipping & Receiving personnel in planning for the impacts of delivering a higher volume to one central receiving point at the IGB.

Additionally, new onboarding processes were put in place to help not only new employees, but those who had worked in Walnut Creek for years, to understand the aspects of working on the Berkeley Lab site and enable access to all of the resources necessary to be successful (**OPS02**). On the health and safety

[54] Baumgart, L. A., *et al.*, 2021, "Persistence and plasticity in bacterial gene regulation", *Nature Methods*, 18, 1499-1505.

front, the JGI team coordinated with Berkeley Lab EH&S personnel to ensure a safe and smooth transition into the building and established regular inspections, monitoring, and coordination to ensure the proper safety precautions, chemical processes, and disposal practices were in place (**OPS05**). This also included the training of JGI's volunteer emergency response personnel on local practices to prepare for how to respond to emergencies at Berkeley Lab. Finally, procurement and other administrative practices were updated to leverage the proximity to Berkeley Lab central offices and services (**OPS03**). This included leveraging the additional overhead costs that JGI took on as part of the move to Berkeley to use the general procurement services offered by the Lab's Procurement organization.