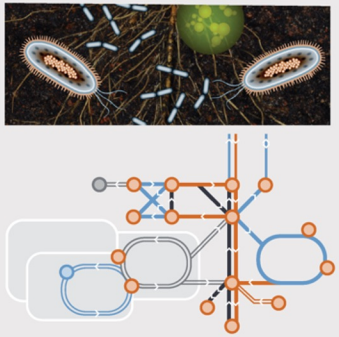
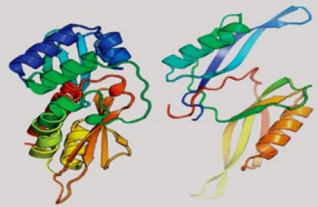
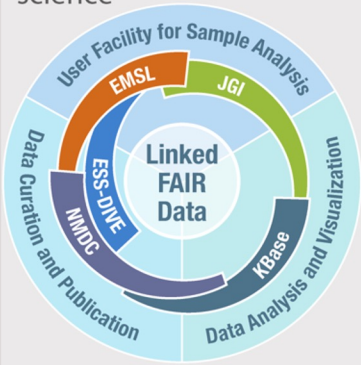
JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

JGI-NERSC Partnership: lessons in data-intensive computing at scale

NERSC Celebrates 50 years
July 22, 2024

Continued Evolution into an Integrative and Collaborative User Facility

JGI JOINT GENOME INSTITUTE

	1997	2004	2012	2018	2021…
Structure	The Human Genome Project	A Production Sequencing User Facility	A Next-Generation Genome Science User Facility	An **Integrative** Genome Science User Facility	An **Integrative and Collaborative** User Facility
Science	Humans & animals 	Plants & microbes 	Organismal interactions, DNA synthesis, metabolism 	**INTEGRATIVE OMICS**  	Leverage key partnerships for multidisciplinary science 
Mission	Sequence, assemble, and annotate **GENOMES**		Understand gene and genome **FUNCTION**	Integrate sequence and functional capabilities for systems-level **BIOLOGY**	Expand interactions to characterize **BIOLOGICAL PROCESSES**

User Programs and Science Programs

SCIENTIFIC PEER REVIEW

40%
Community Science Program (CSP)

10%
Facilities Integrating Collaborations for User Science (FICUS)

30%
Bioenergy Research Centers (BRC)

10%
Director's Science

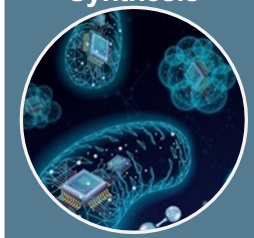
10%
Biological and Environmental Research Support Science (BERSS)


Plant


Fungal, Algal


Metagenome


Microbial


DNA Synthesis


Metabolomics


2° Metabolites

2023: JGI by the Numbers

259 Staff · 25 New Hires
20 Grad Students · 18 Postdocs

Total Funding

U.S. DEPARTMENT OF ENERGY | Office of Science

DOE BER $89.4M

2,373 Primary Users

22,262 Secondary Users

238 Publications

- 716.9 Terabases sequence generated
- 11 Megabases DNA synthesized
- 11.56K metabolomics analyses runs
- 161 proposals submitted
- 64 proposals approved

- Total files requested: 7.9M
- JGI Archive size grew to:
  15.2 million file records
- 15.95 Petabytes (PB) of data

GENOME INSIDER | Natural Prodcast

8.2K total podcast downloads

Engaged Reach: 24.5K 3.3K

Mission

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

2,243 Primary Users leveraging JGI data generation capabilities in FY22

15,219 Secondary Users that engaged with JGI science gateways

14PB JGI Data Repository size as of December 2020

2,862 Publications

JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

Office of Biological and Environmental Research

Data Generation and Reuse

Primary Users provide **unique samples** from **fungi, plants, algae, bacteria, archea, and communities** as part of their studies

Samples become data

Primary and Secondary Users **leverage data** through JGI Flagship Science Gateways

JGI Data Portal
JGI PhycoCosm — THE ALGAL GENOMICS RESOURCE
JGI MycoCosm — THE FUNGAL GENOMICS RESOURCE
JGI Phytozome 13 — THE PLANT GENOMICS RESOURCE
JGI IMG/M — INTEGRATED MICROBIAL GENOMES & MICROBIOMES

FY22 Downloads

4.1M Files

1.2PB Data

NERSC National Energy Research Scientific Computing Center
BioSciences
BERKELEY LAB
U.S. DEPARTMENT OF ENERGY | Office of Science

JGI User Groups

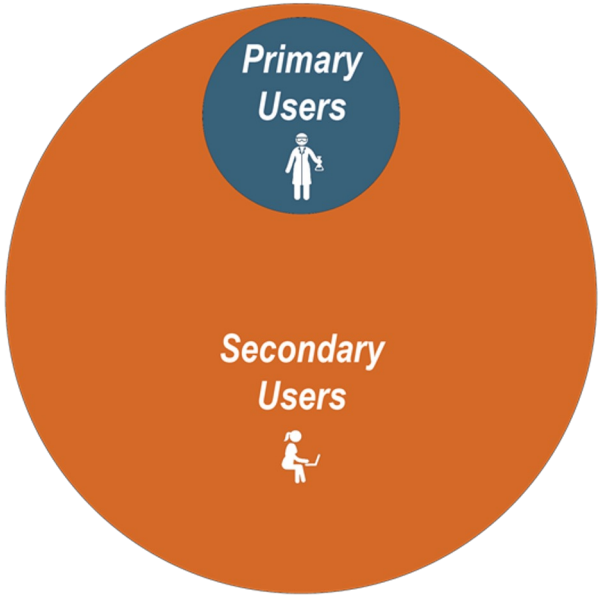
Primary and Secondary Users

Primary Users

are associated with one or more JGI **User Program proposals**

Secondary Users

build on the work of JGI personnel and primary users through **direct downstream use** of JGI data, systems, and tools.



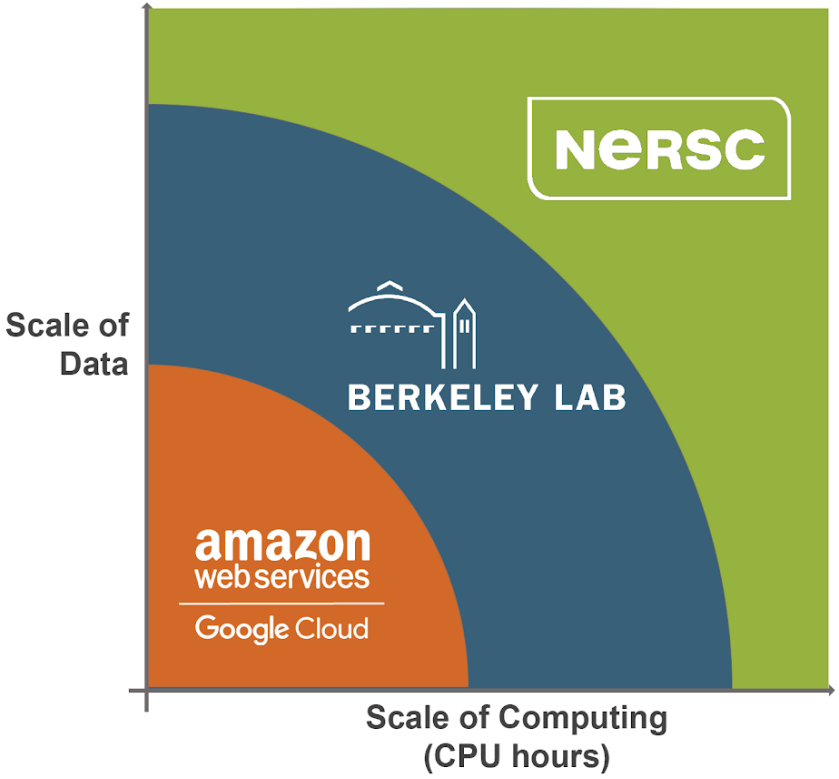
Example Outcomes

- Publications
- Patents
- Software Adaptations
- New Technologies
- Marketable Products
- Methods & Standards
- Start-ups
- Grant Funding

JGI and NERSC, a brief history

- 2010 – JGI data and computing was growing too fast for the Walnut Creek facility
- New hardware was redirected to NERSC facility in Oakland
- 2011 - JGI hardware unified to become the Genepool system
- 2012 – Consultants hired to help support the JGI use of shared storage and computing resources
- 2013 – Mendel deployed (consolidate Genepool and PDSF)
- 2019 – Mendel retired (moved to LBL IT), JGI had a cabinet of Cori
- 2023 – Cori retired, JGI stands up Dori at LBL IT

JGI's Computing Infrastructure Spectrum



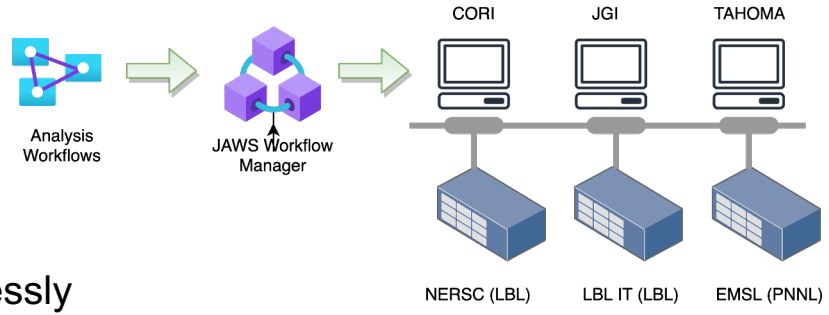
JGI JOINT GENOME INSTITUTE

Binning JGI Compute Infrastructure Requirements

- Prototyping, exploratory analysis, small-scale production work
- Large-scale data or compute needs (>100,000 CPU hours)

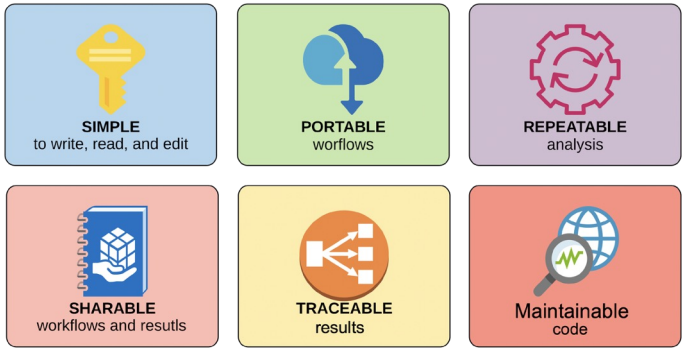
Unifying Workflow Execution Across JGI Resources

- Developed a workflow manager called **JGI Analysis Workflow Service (JAWS)** to run complex computational workflows with support for distributed computation across multiple HPC enabled sites.



- Provides a **user-friendly common interface** to seamlessly route jobs and data across multiple sites.

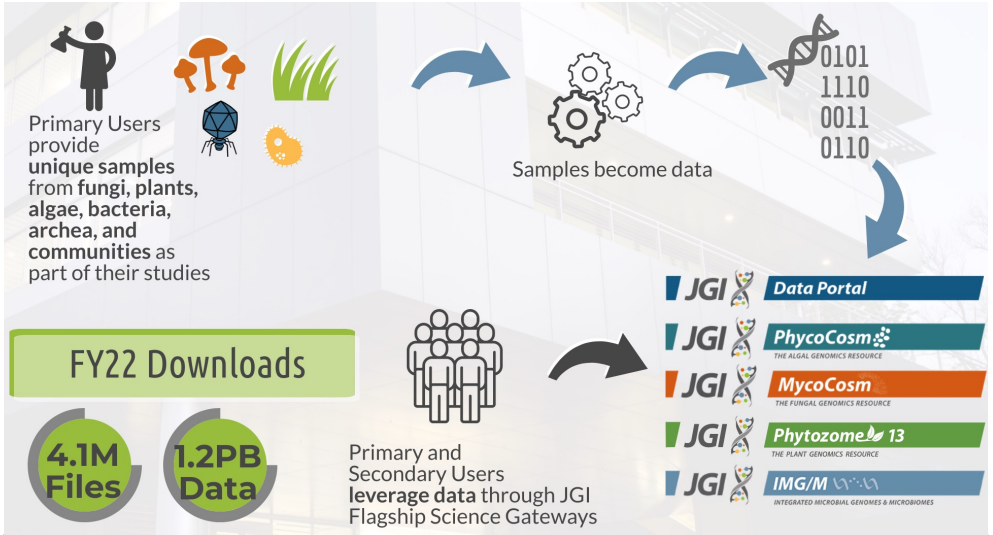
- Uses **Cromwell** to execute workflows in a common **Workflow Description Language (WDL)**, standardizing the workflow language.



- Improves the **reusability** and **robustness** of bioinformatics workflows in evolving and/or diverse high-performance computing (HPC) and cloud environments.

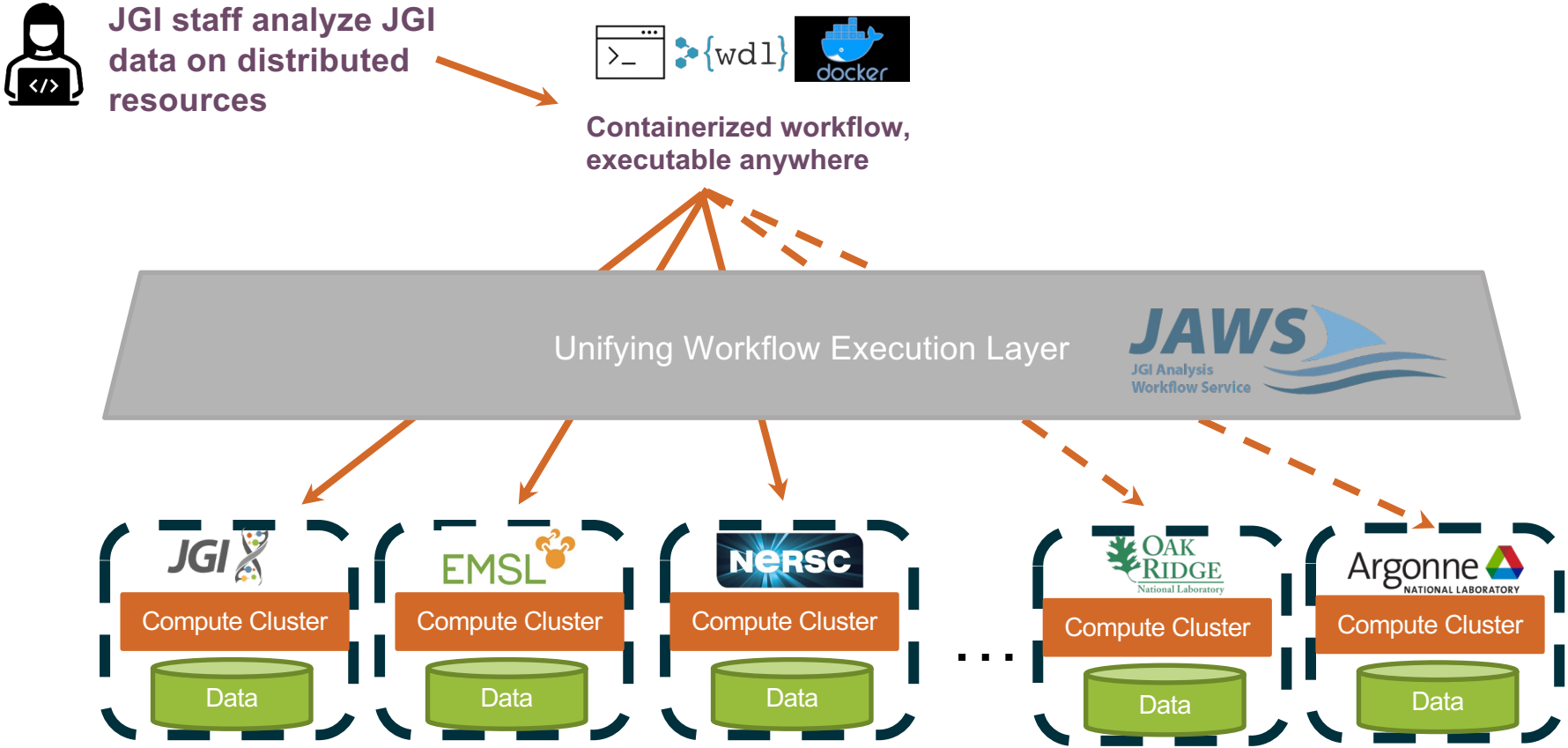
Centralized Data Access and Movement across Distributed Resources

- The JGI Archive and Metadata Organizer **(JAMO)** deployed in 2013
- Holds the **metadata and locations** for data produced by JGI
- **Powers data distribution** across JGI storage systems (file system and archives)
- Makes **centralized search** possible
- Supports **reusability research**

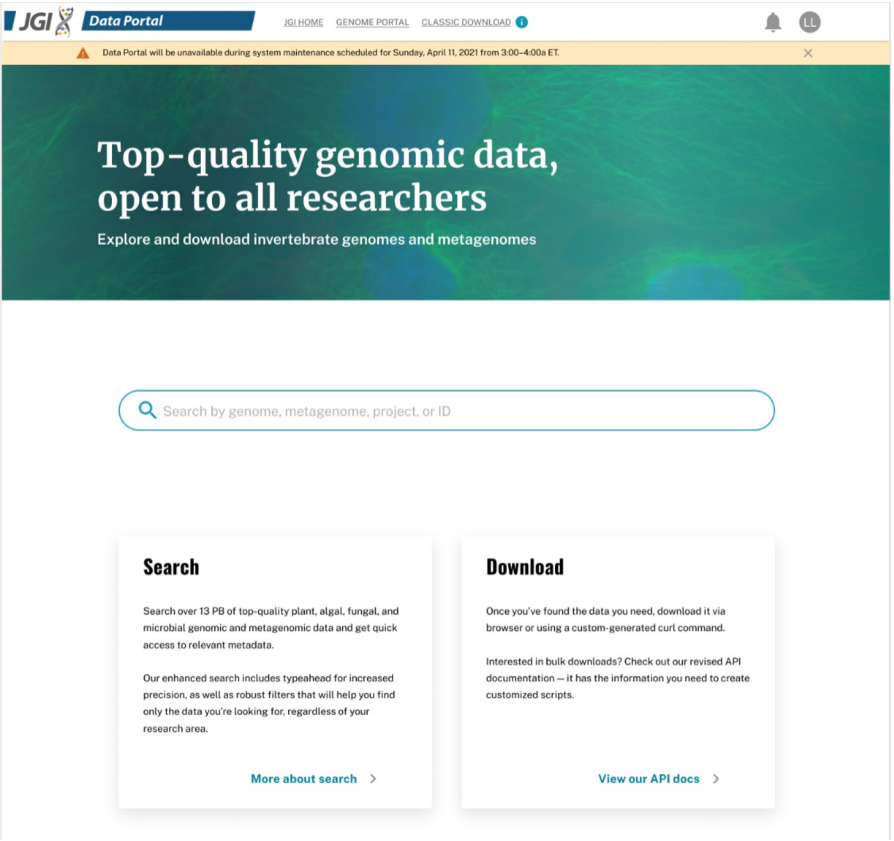


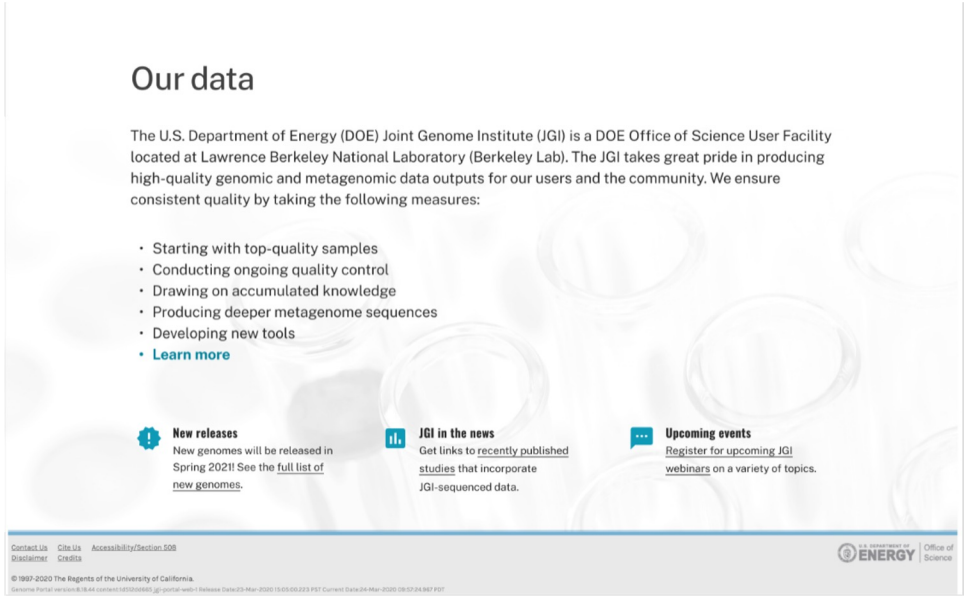
JAMO

JAWS: Write Once, Run anywhere



The JGI Data Portal – Access to all Public JGI Data

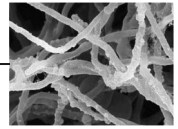




GUI: https://data.jgi.doe.gov
API: https://files.jgi.doe.gov/apidoc

Impact Narrative Examples

How does 'impact' begin?



JGI sequences *P. chrysosporium* fungal genome

JGI sequences *P. placenta* fungal genome

JGI launches its Genome Portal

2005 2006 2011 2014

Berkeley UNIVERSITY OF CALIFORNIA

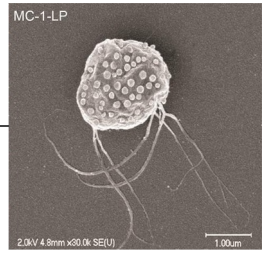
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

bp

Using data from the above genomes retrieved from JGI's Genome Portal, university and BP researchers patent processes to improve cellular sugar transportation for **production of biofuels** like ethanol.

Impact Narrative Examples

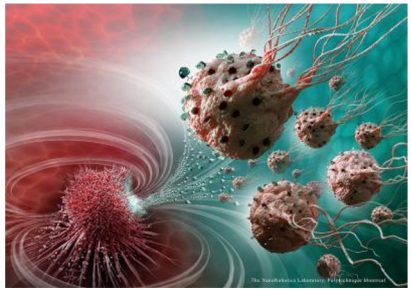
How does 'impact' begin?



JGI sequences the genome of *Magnetococcus* MC-1, a bacterium with special mobility traits that thrives in low-oxygen marine environments

2007

2016



Aided by available genomic information, researchers determine that this bacteria is a very effective **medication delivery tool** for tumors in hard-to-reach areas of the brain.

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Impact Narrative Examples

How does 'impact' begin?

JGI releases Phytozome and MycoCosm

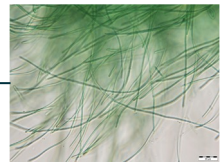
JGI sequences Shiitake genomes and integrates them with external genomes in Mycocosm

2008 2010 2016 2019 2020 2021

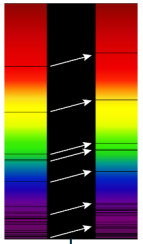
Utilizing Shiitake genome data from MycoCosm and protein data from Phytozome, Mycotechnology, Inc. files two patents involving novel methods for the **production of sustainable meat substitutes**.

Impact Narrative Examples

How does 'impact' begin?



JGI sequences the genome of bacterium *Leptolyngbya sp.* JSC-1 for users



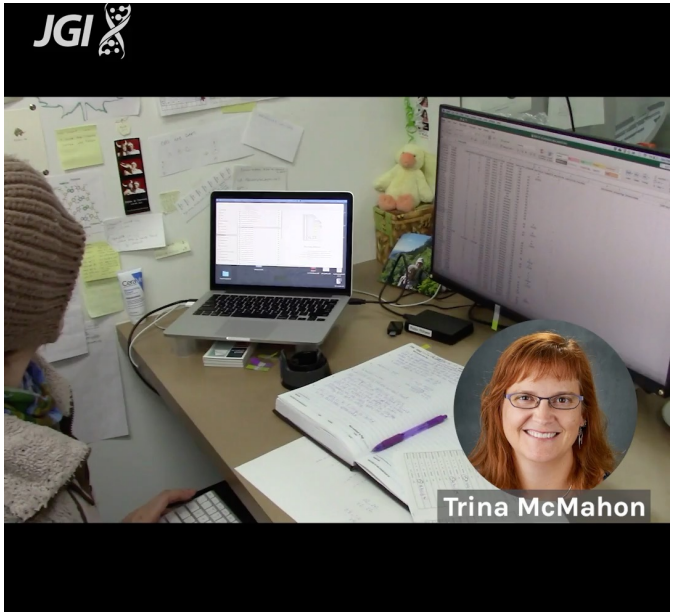
JGI users leverage genomic data received from JGI to determine how this bacterium conducts photosynthesis using far red-shifted light

2007 2009 2014 2018

Downstream researchers leverage the findings of JGI's users to determine that the mechanisms in Cyanobactera could be used to **produce oxygen for humans on Mars**.

Highlight: The Megadata of Lake Mendota

- 3-part podcast arc released November-December 2023
- 20-year field samples from UW-Madison
- Highlighted JGI, NERSC, ExaBiome resources and capabilities harnessed to sequence and assemble 25Tb of metagenome data
- Episode interludes highlight that these are projects approved through JGI proposal calls



GENOME INSIDER

Host: Menaka Wilhelm

Amazing Podcast - Genome Insider

Host: Menaka Wilhelm

Season 4 Eps 6-8 highlight the combination of supercomputing and sequencing to understand environmental microbiomes!

Supercomputing, MetaHipMer and the Mega Dataset of Lake Mendota

David Buoy sits at the deepest part of Lake Mendota, where microbial samples have been collected for over two decades.

Sequencing technology has changed dramatically over the last 25-plus years since the JGI's inception, making it possible for researchers to get a close look at more ecosystems and organisms than ever before. In 2006, the JGI produced 33 billion base pairs of sequence; by 2023, that number was almost 717 trillion. Last year, the JGI surpassed three Petabases of data sequenced — that's three-quadrillion base pairs of DNA sequence!