JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

JGI Data Capabilities and Infrastructure

BERAC
April 21, 2023

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

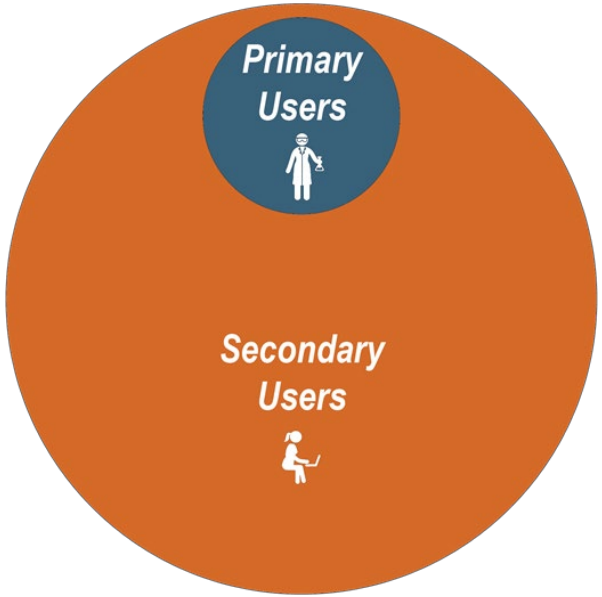
U.S. DEPARTMENT OF ENERGY | Office of Science

JGI User Groups

Primary and Secondary Users

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

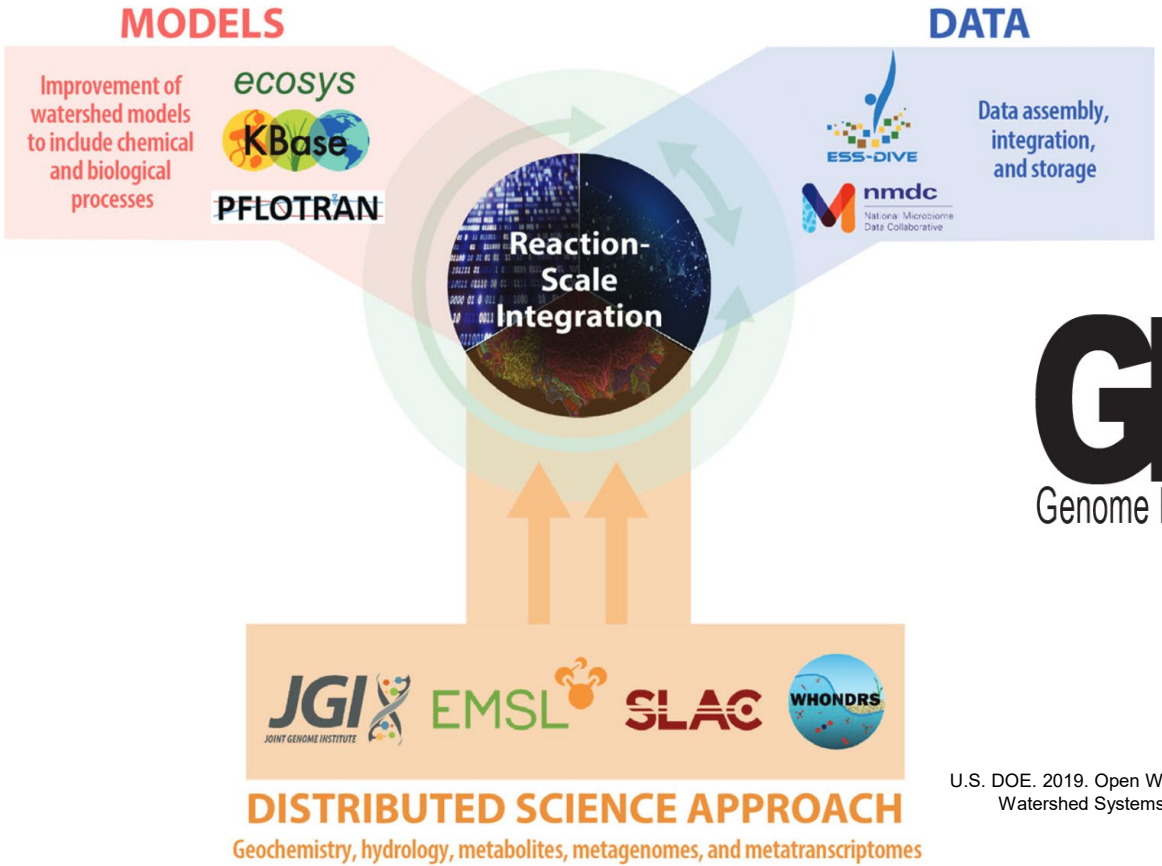
Secondary Users
build on the work of JGI personnel and primary users through **direct downstream use** of JGI data, systems, and tools.



Example Outcomes

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

JGI Contributes Data to Integrated Projects



GROW
Genome Resolved Open Watersheds

Kelly Wrighton

Mikayla Borton

U.S. DOE. 2019. Open Watershed Science by Design: Leveraging Distributed Research Networks to Understand Watershed Systems Workshop Report, DOE/SC-0200, U.S. Department of Energy Office of Science.

JGI Public Resources for Data, Metadata, & Analysis

Data & Metadata

Data & Metadata

JGI Systems

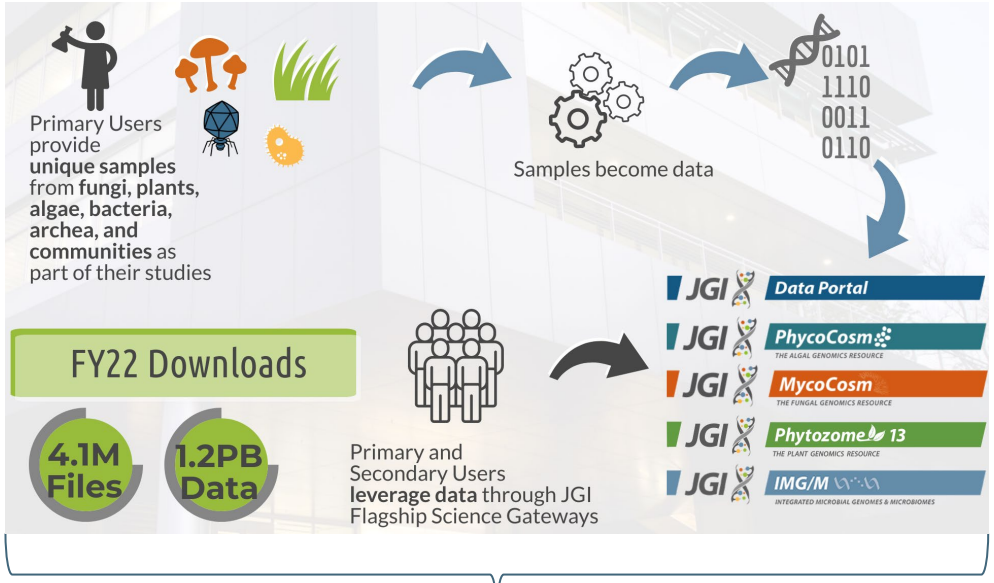
- PhycoCosm — THE ALGAL GENOMICS RESOURCE
- IMG/M — INTEGRATED MICROBIAL GENOMES & MICROBIOMES
- Phytozome 13 — THE PLANT GENOMICS RESOURCE
- MycoCosm — THE FUNGAL GENOMICS RESOURCE
- Data Portal
- GOLD — GENOMES ONLINE DATABASE

External Systems

- National Center for Biotechnology Information (NCBI)
- ENA — European Nucleotide Archive

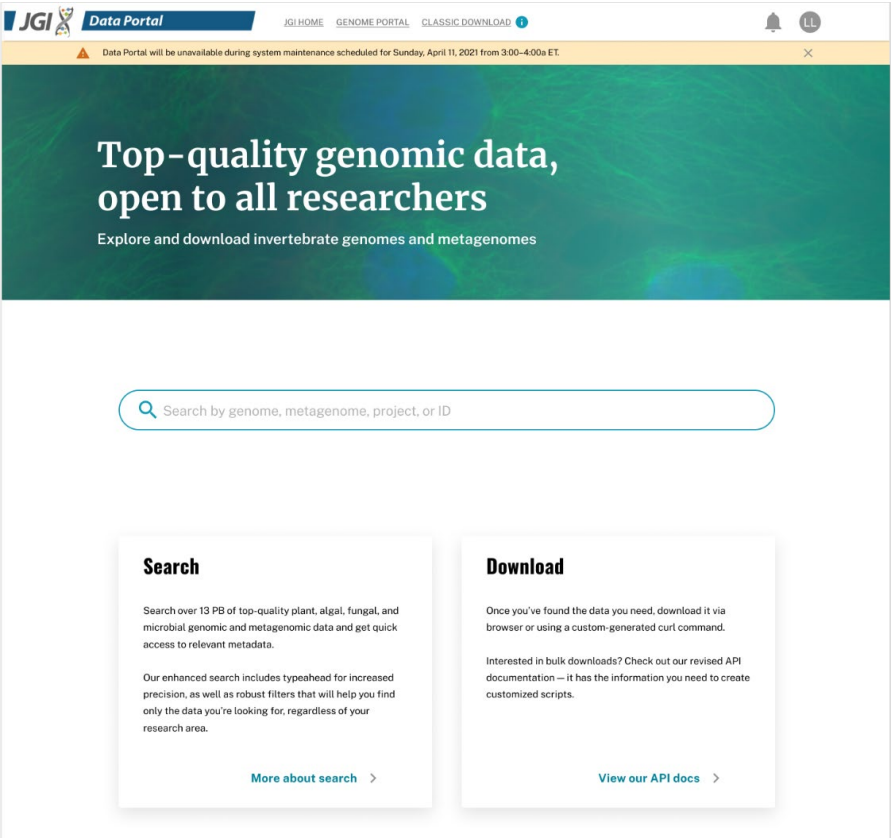
Centralized Data Access and Movement across Distributed Resources

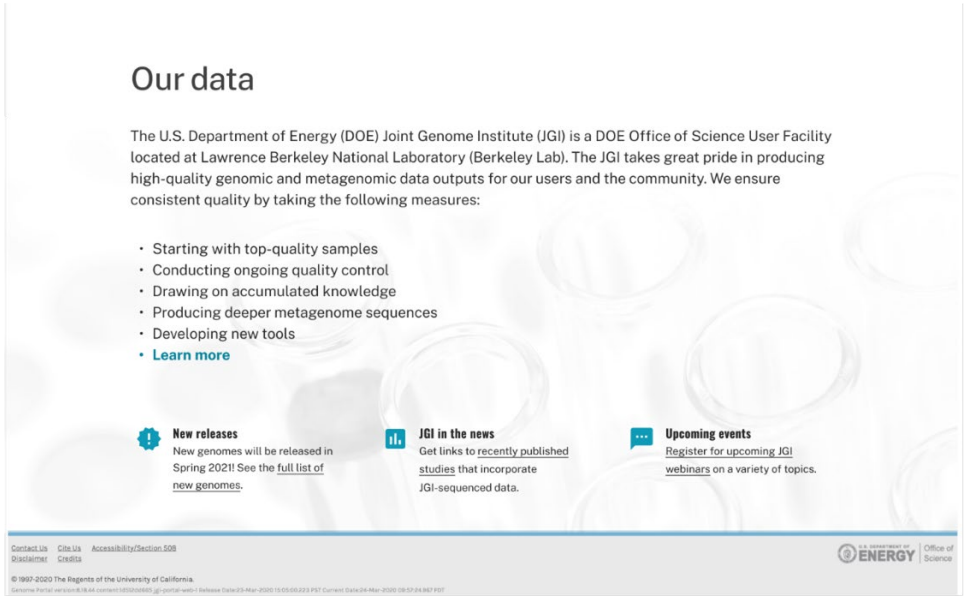
- The JGI Archive and Metadata Organizer **(JAMO)** deployed in 2013
- Holds the **metadata and locations** for data produced by JGI
- **Powers data distribution** across JGI storage systems (file system and archives)
- Makes **centralized search** possible
- Supports **reusability research**



JAMO

The JGI Data Portal – Access to all Public JGI Data



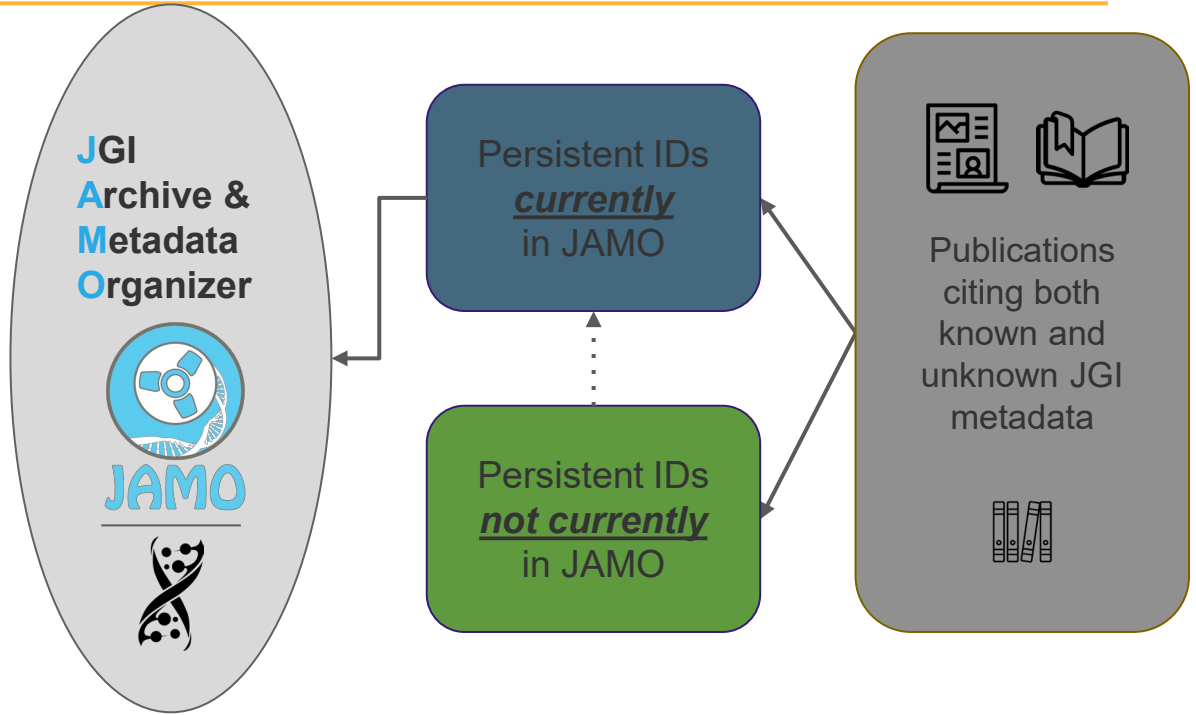


GUI: https://data.jgi.doe.gov
API: https://files.jgi.doe.gov/apidoc

Identifying Downstream Data Use

The Data Citation Explorer (DCE) Connects Data Use to User Projects

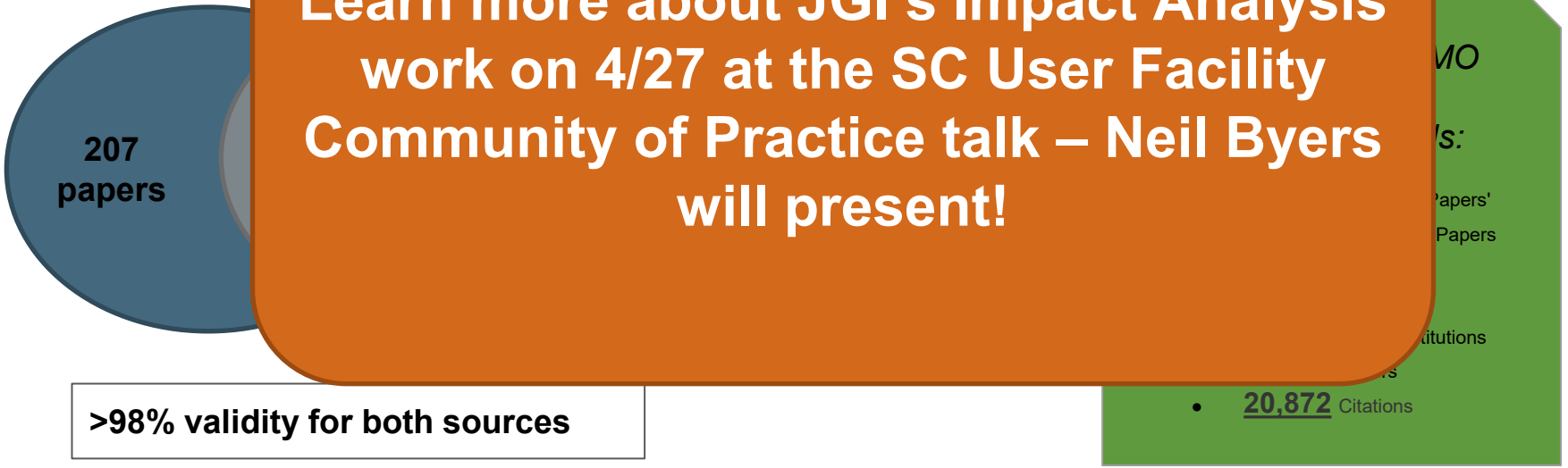
1. We know that many publications exist which cite JGI data
2. These publications cite both cataloged and uncataloged metadata
3. The DCE finds new metadata + citing publications and stores them in JAMO



Expanding Our Reach with Dimensions

How much does Dimensions expand JGI's search capabilities?

Using a sample… …DCE), we captured 118… …e been possible with…

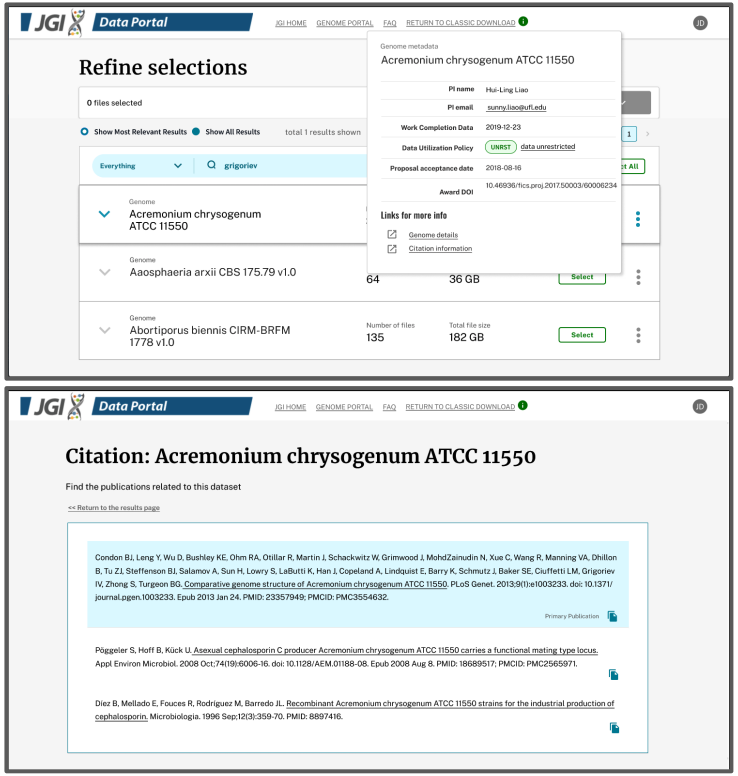


Learn more about JGI's Impact Analysis work on 4/27 at the SC User Facility Community of Practice talk – Neil Byers will present!

>98% validity for both sources

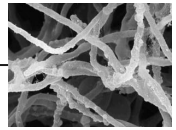
Data Citation Explorer – Augmenting the Data Portal

- Tie literature to data and display to data users
- Increase JGI understanding of institutional and individual impact
- Provide model for identifying data citations at other DOE repositories



Impact Narrative Examples

How does 'impact' begin?



JGI sequences *P. chrysosporium* fungal genome

JGI sequences *P. placenta* fungal genome

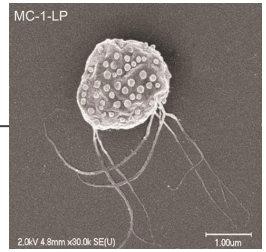
JGI launches its Genome Portal

2005 2006 2011 2014

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

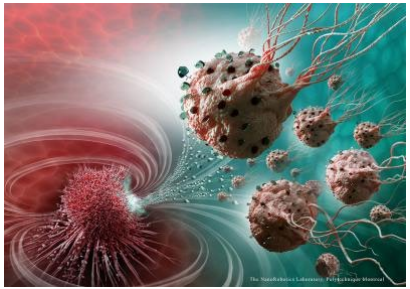
How does 'impact' begin?



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

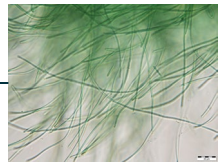
JGI sequences Shiitake genomes and integrates them with external genomes in Mycocosm

2008 2010 2016 2019 2020 2021

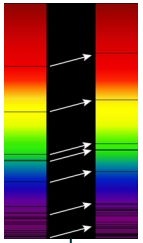
Utilizing Shiitake genome data from MycoCosm and protein data from Phytozome, Mycotechnology, Inc. files two patents involving novel methods for the **production of sustainable meat substitutes**.

Impact Narrative Examples

How does 'impact' begin?



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



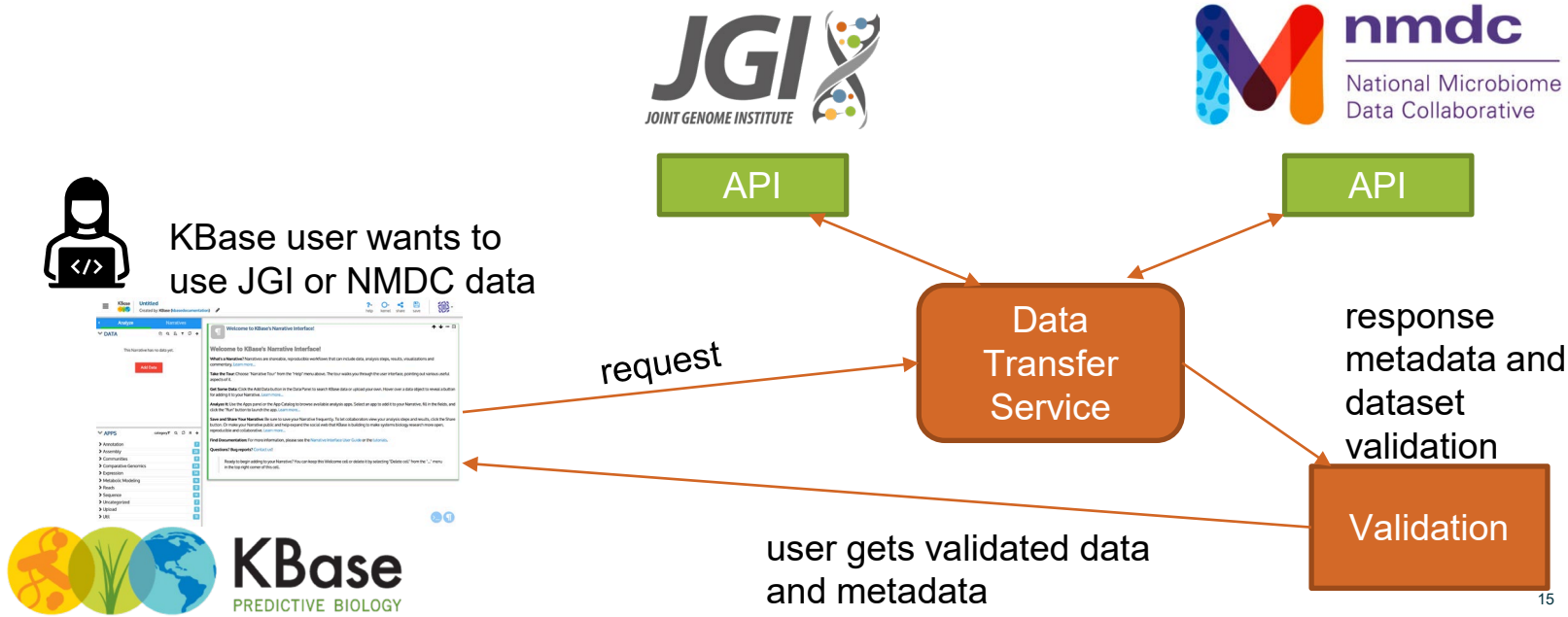
JGI users leverage genomic data received from JGI to determine how this bacterium conducts photosynthesis using far red-shifted light

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**.

Work in Progress: Data Transfer Service w/ KBase & NMDC

- **Validated Data Transfer**
  - Make sure data moves between resources with citation and contextual information
- **Clear citation guidelines in Credit metadata**
  - Consistent citations will aid in identifying data use in publications



Sample Metadata Wrangling is Challenging

Link Journal Publication mSystems

nmdc
National Microbiome Data Collaborative

ABOUT PRODUCTS COMMUNITY RESOURCES CONTACT MONTANA SMITH

Home
Begin or resume a submission.

Submission Context
Input form

Study Information
Input Form

Multiomics Data

Information about the type of samples being submitt

Alternative Names

Project, study, or sample set names the are also associated with this submission or other names / ident

GOLD Study ID

Provide the GOLD study IDs associated with samples for this study.

NCBI BioProject Accession

Provide the NCBI BioProject Accession Number associated with the listed NCBI BioProject Title.

Data types *

Check all -omics data types associated with samples collected for this study.

Other Non-DOE

- [ ] Metagenome
- [ ] Metatranscriptome
- [ ] Metaproteome
- [ ] Metabolome
- [ ] Natural Organic Matter (FT-ICR MS)

*** indicates required field**

GO TO PREVIOUS STEP

Data types *

Check all -omics data types associated with samples collected for this study.

Joint Genome Institute (JGI)

- [x] Metagenome
- [ ] Metatranscriptome
- [ ] Metabolome

JGI Proposal ID/Study ID *

This is the 6 digit ID assigned to your JGI Proposal and is required when completing metadata for samples to be sent to JGI for sequencing.

Environmental Molecular Science Laboratory (EMSL)

- [x] Metaproteome
- [ ] Metabolome
- [ ] Natural Organic Matter (FT-ICR MS)

EMSL Proposal / Study Number *

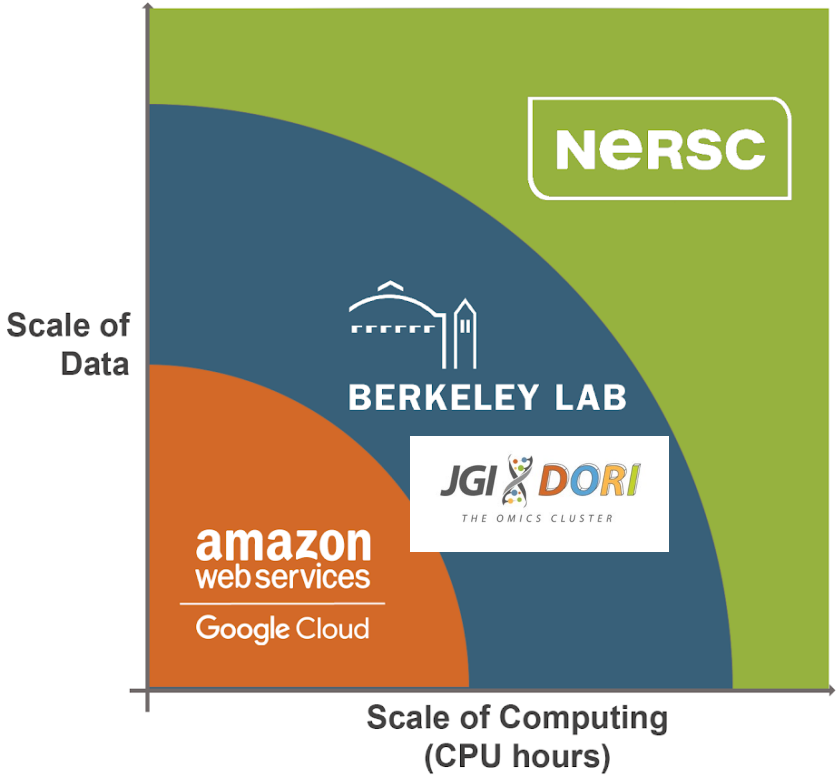
EMSL Study Number is required when processing was done at EMSL

Other Non-DOE

- [ ] Metagenome
- [ ] Metatranscriptome
- [ ] Metaproteome
- [ ] Metabolome
- [ ] Natural Organic Matter (FT-ICR MS)

ESS-DIVE
Deep Insight for Earth Science Data

JGI's Computing Infrastructure Spectrum



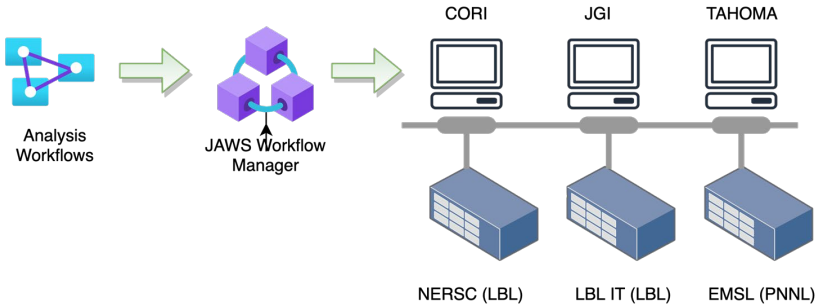
JGI JOINT GENOME INSTITUTE

Binning JGI Compute Infrastructure Requirements

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

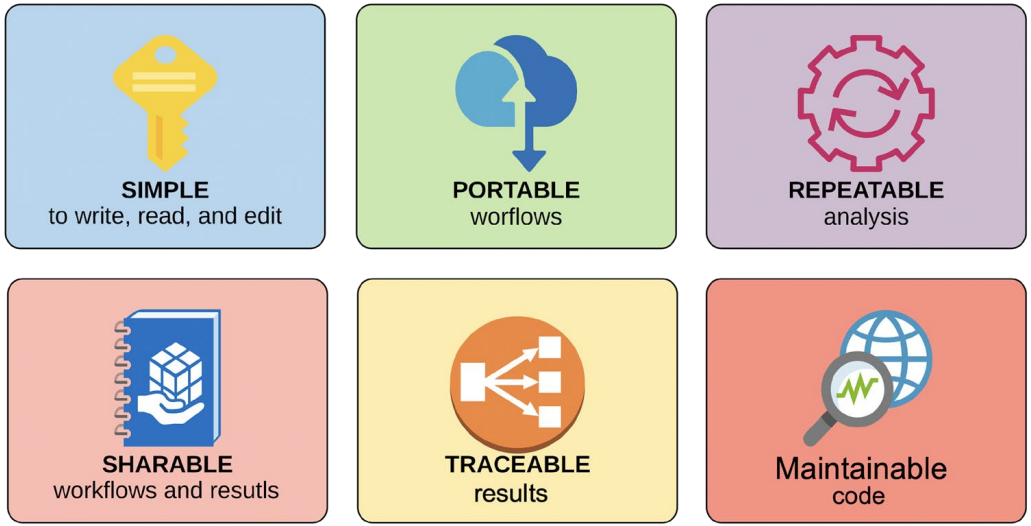
Our Solution to Unify Workflow Execution Across JGI

- 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.

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



JGI Software Infrastructure Reuse with NMDC

- Consolidated workflow efforts between JAWS and NMDC team
- JAWS provides the overall framework for submitting, running and collecting results
- JAWS team will provide operational support for core JAWS service (initially for NERSC workflows)
- NMDC will be able to customize their cluster configurations and work with JAWS team on new features to support NMDC requirements
- JAWS team supporting NMDC workflow team

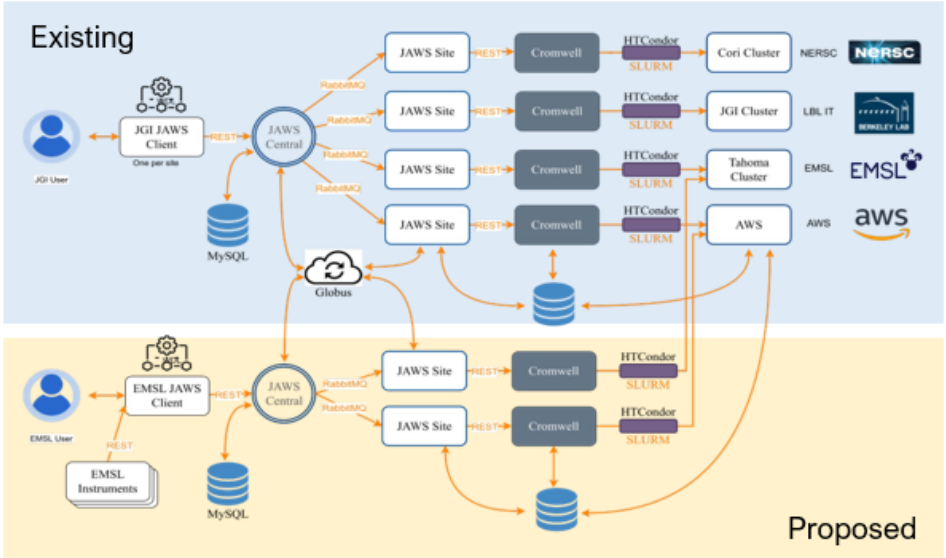
JGI Software Infrastructure Reuse with EMSL

Accelerators

KBase

nmdc
National Microbiome Data Collaborative

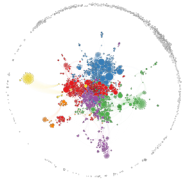
ESS-DIVE



Future Work and Opportunities

- **Build stronger connections between data resources and computing platforms**
  - Submission portal to harmonize sample, environment, and process metadata
  - Sample registration for all BER samples
  - Collaborative User-centered Design

- **More analysis of the Data Citation Explorer results**
  - Explore differences in data citations (e.g. methods vs background citation)
  - Expose the Data Citation Explorer as a public-facing resource
  - Add support for other identifiers (e.g. IGSNs or unique IDs from other fields)

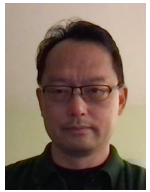


- **Shared software and hardware infrastructure**
  - Expand the number of sites JAWS supports
  - JAMO infrastructure available to partners

Thank you

Staff who make this happen

JAWS
JGI Analysis Workflow Service



Steve Chan
Advanced Analysis GL

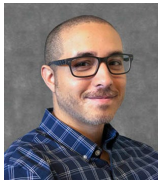
- Dani Cassol
- Jeff Froula
- Ed Kirton
- Angie Kollmer
- Ramani Kothadia
- Mario Melara
- Seung-Jin Sul
- Stephan Trong
- Nick Tyler

JGI DORI
THE OMICS CLUSTER

Georg Rath
Systems Infrastructure Team Supervisor

- Harper Mann
- Matt Dunford
- Brian Yumae
- John White
- Karen Fernsler
- Wei Feinstein

JGI Data Portal



Steven Wilson
Advanced Analysis GL

- Tatyana Smirnova
- Shijie Yao
- Alexandre Poliakov

The Data Citation Explorer

Neil Byers
Impact Analyst Data Scientist

Chuck Parker
Computer Systems Engineer

JAMO

Chris Beecroft
Staff Data Scientist

- Ed Lee