LA-UR-24-26605

Title: 2024 NMDC Ambassador Training Materials

Author(s): Kelliher, Julia Mae; Rodriguez, Francisca Ester; Johnson, Leah Young Davenport; Prime, Kaelan Jantira; Roux, Simon; Eloe-Fadrosh, Emiley A.; Smith, Montana; Clum, Alicia; De Santiago, Alejandro; Gajigan, Andrian; Maher, Rebecca; Hanson, Buck Timothy; Robinson, Christopher; Betacurt-Anzola, Daniela; Skoog, Emilie; Cirolia, Giana Teresa; Skeen, Heather; Oduwole, Iyanu; Kajihara, Kacie; Pham, Kent; Camuy-Vélez, Lennel; et al.

Intended for: NMDC Ambassador training materials that will receive a DOI

Issued: 2024-07-02

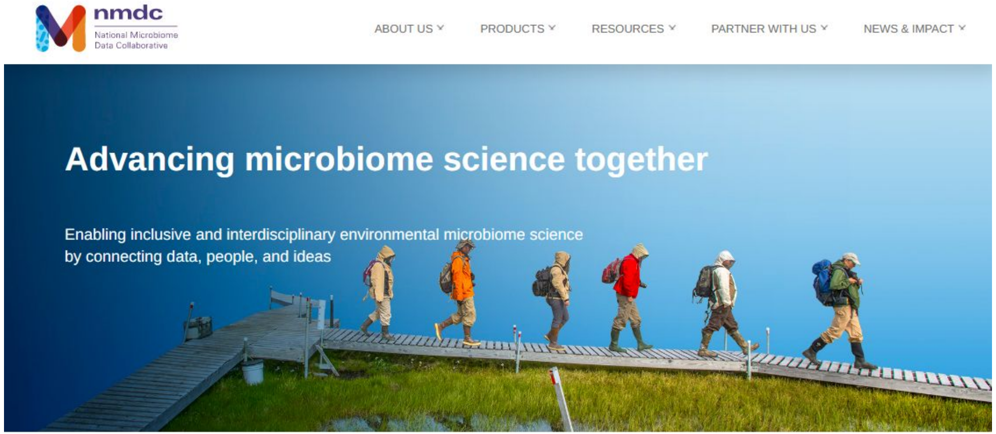
nmdc

National Microbiome Data Collaborative

What is the NMDC?

2024 Ambassador Cohort

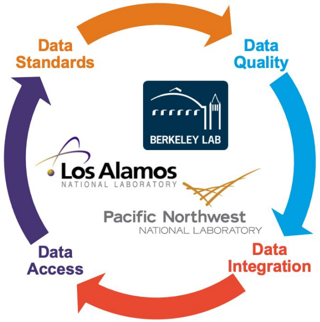
What is the NMDC?



Advancing microbiome science together

Enabling inclusive and interdisciplinary environmental microbiome science by connecting data, people, and ideas

U.S. DEPARTMENT OF ENERGY


Data Standards
Data Quality
BERKELEY LAB
Los Alamos NATIONAL LABORATORY
Pacific Northwest NATIONAL LABORATORY
Data Access
Data Integration

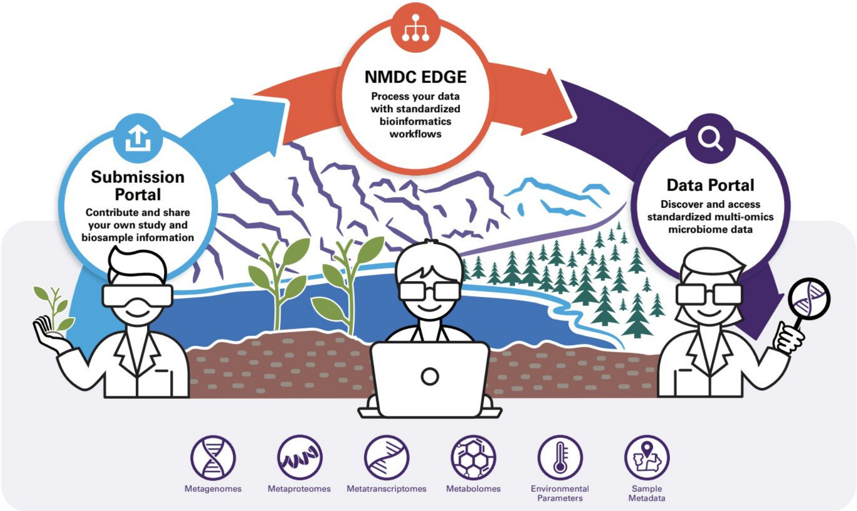
The NMDC is a *sustainable data discovery platform* that promotes open science and shared-ownership across a broad and diverse community of researchers, funders, publishers, societies, and other collaborators. The NMDC aims to enable multi-omic microbiome research to accelerate scientific discovery. The NMDC is a Department of Energy funded program that is a collaboration between 3 National Laboratories: Lawrence Berkeley National Laboratory (LBNL), Los Alamos National Laboratory (LANL), and Pacific Northwest National Laboratory (PNNL)

“Enabling inclusive and interdisciplinary environmental microbiome science by connecting data, people, and ideas”

Why we need standards

nmdc

National Microbiome Data Collaborative


NMDC EDGE
Process your data with standardized bioinformatics workflows
Submission Portal
Contribute and share your own study and biosample information
Data Portal
Discover and access standardized multi-omics microbiome data
Metagenomes
Metaproteomes
Metatranscriptomes
Metabolomes
Environmental Parameters
Sample Metadata

Synthesizing these to gain insights requires:

⇒ *meaningfully* FAIR data

⇒ *computable* standards

⇒ *community* adoption

Findable

Accessible

Interoperable

Reusable

What role do microbes play in the persistence of soil carbon?

How do microbes and viruses mediate watershed scale nitrogen transformations?

How do microbes impact biogeochemical cycling in deep subsurface environments?

Vision and Mission

Vision

To **connect data, people,** and **ideas** to advance microbiome innovation and discovery

metadata

Mission

To support a FAIR microbiome data sharing network, through **infrastructure, data standards,** and **community building**, that addresses pressing challenges in environmental sciences

Product Initiatives

Submission Portal

Lower barriers to collect study and biosample data

NMDC EDGE

Streamline multi-omics data processing

Data Portal & API

Access and discovery of microbiome information

Engagement

User Facilities

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

EMSL

Individuals

The NMDC Champions Program

The NMDC Ambassador Program

Strategic Partners

NSF

neon Operated by Battelle

NASA

GeneLab

Metadata Standards

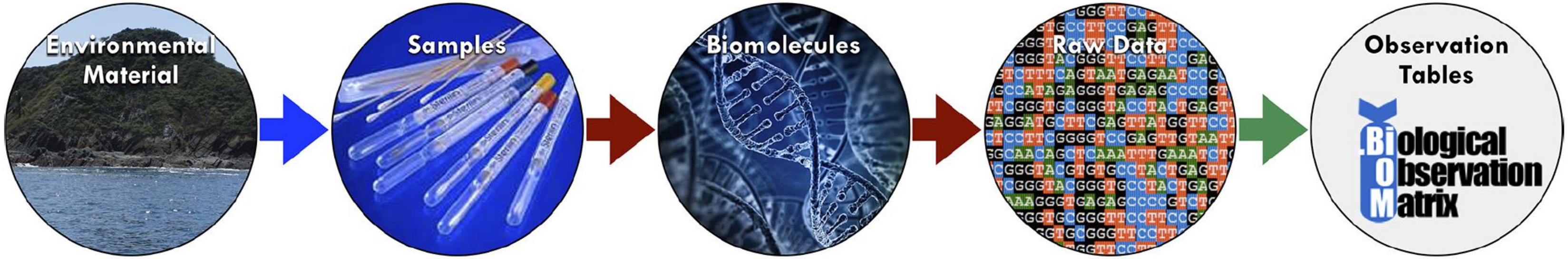
Adopting standards for reporting makes data human and machine readable.

Metadata (data about data):

- Contextual data about your data
- Vital for data
  - Preservation
  - Discovery
  - Access
  - Reuse

Sample metadata includes information about:

- ***When*** it was collected
- ***Where*** it was collected
- ***What*** kind of sample is it
- ***Treatment*** applied during experimentation
- ***Environmental Properties*** from which the sample was taken



NMDC Submission Portal

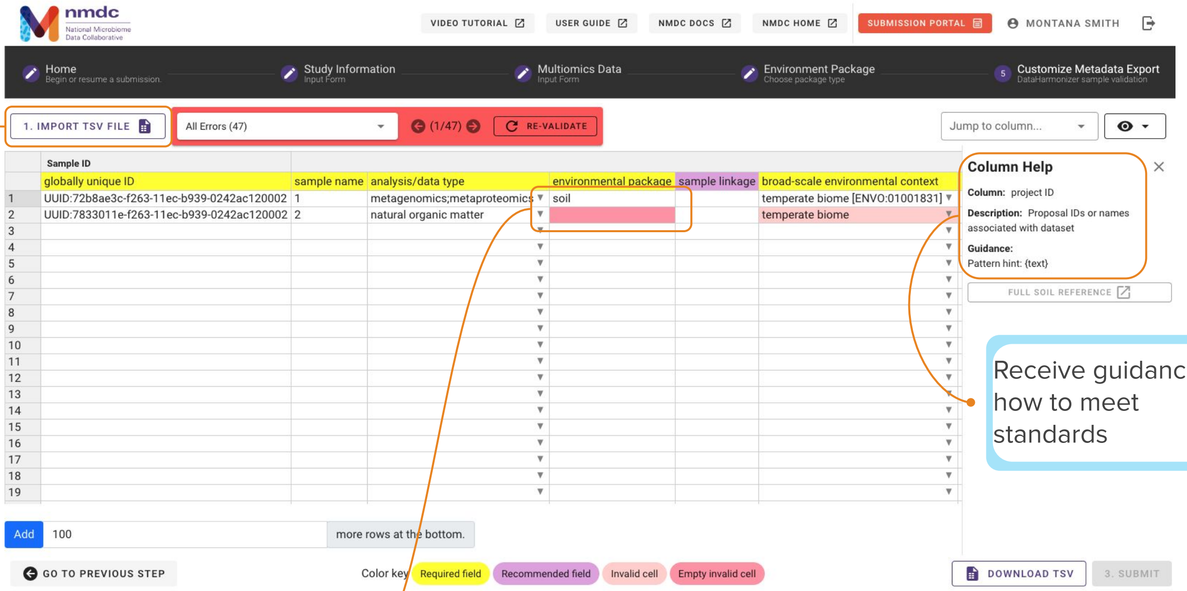
nmdc

National Microbiome Data Collaborative

Submission Portal

Lower barriers to collect study and biosample data

Upload your data



	Sample ID					
	globally unique ID	sample name	analysis/data type	environmental package	sample linkage	broad-scale environmental context
1	UUID:72b8ae3c-f263-11ec-b939-0242ac120002	1	metagenomics;metaproteomics	soil		temperate biome [ENVO:01001831]
2	UUID:7833011e-f263-11ec-b939-0242ac120002	2	natural organic matter			temperate biome
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						

Column Help

Column: project ID

Description: Proposal IDs or names associated with dataset

Guidance:
Pattern hint: {text}

Color key: Required field | Recommended field | Invalid cell | Empty invalid cell

Receive guidance on how to meet standards

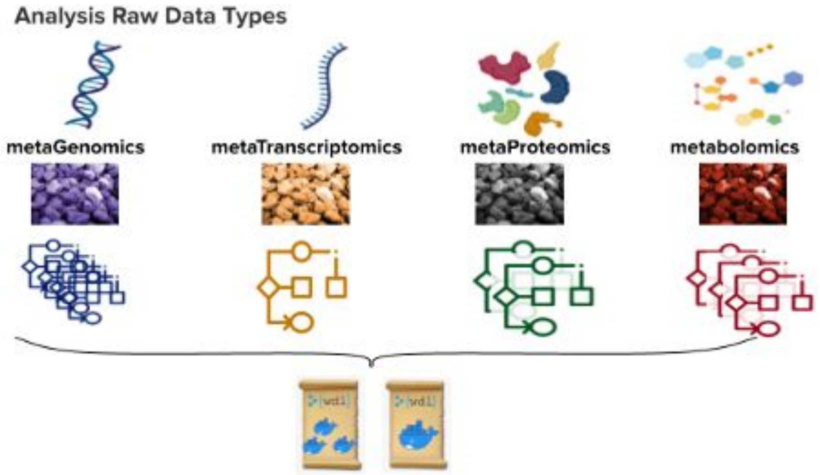
Validate your submission against pre-made MIxS templates

Standardized Bioinformatic Workflows

nmdc
National Microbiome Data Collaborative

The NMDC has integrated state of the art open-source bioinformatics tools into standardized workflows for processing raw multi-omics data to produce **interoperable** and **reusable** annotated data products

- **NMDC Workflows:**
    - Metagenome data
        - ReadsQC
        - Read-based taxonomy classification
        - Assembly
        - Annotation
        - Metagenome assembled genomes (MAGs)
    - Metatranscriptome data
    - Natural organic matter data
    - Metabolome data
    - Metaproteome data
    - Viruses & Plasmids



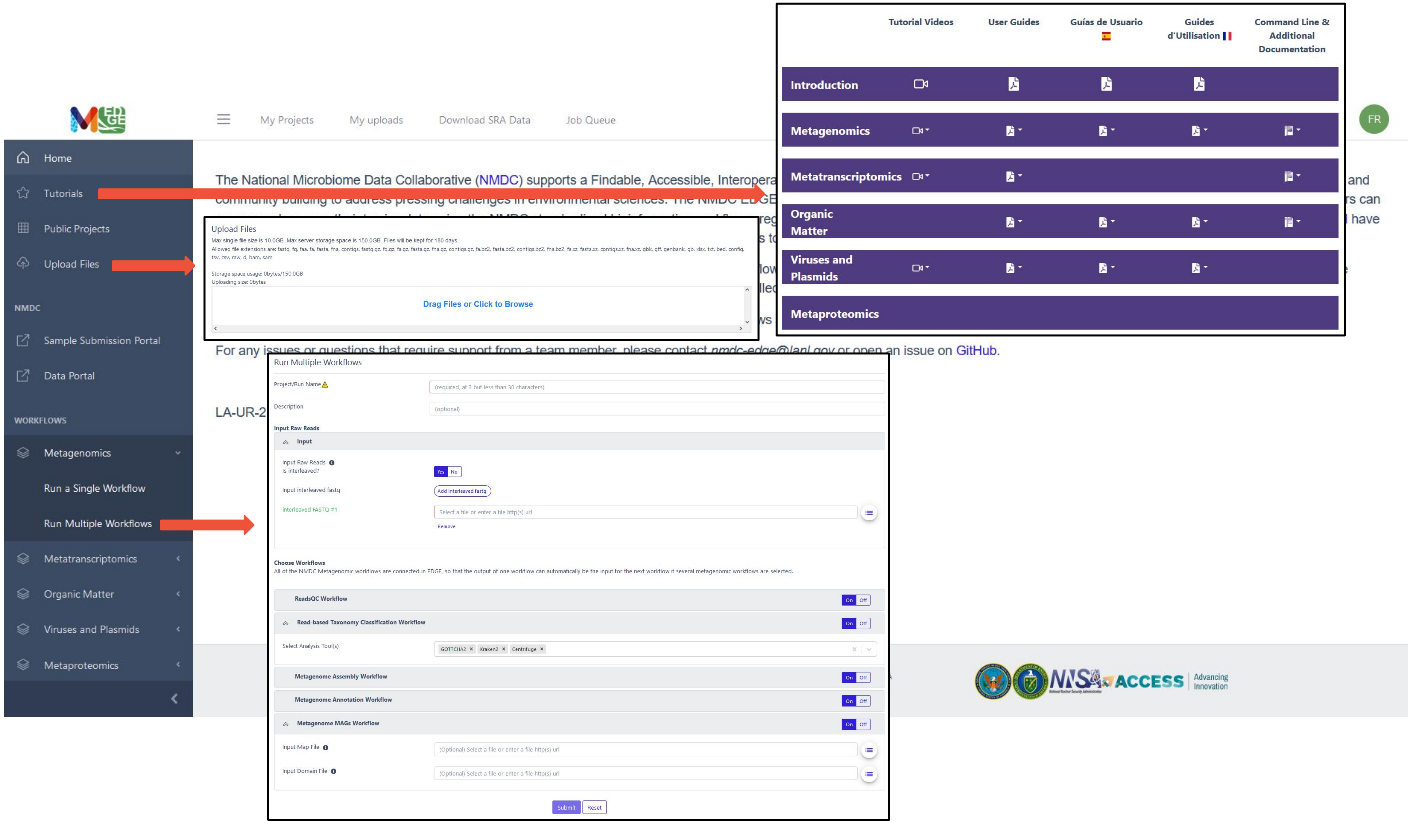
NMDC EDGE

nmdc

National Microbiome Data Collaborative

NMDC EDGE

Streamline multi-omics data processing



	Tutorial Videos	User Guides	Guías de Usuario	Guides d'Utilisation	Command Line & Additional Documentation
Introduction	🎥	📄	📄	📄	
Metagenomics	🎥	📄	📄	📄	📄
Metatranscriptomics	🎥	📄			📄
Organic Matter		📄	📄	📄	📄
Viruses and Plasmids	🎥	📄	📄	📄	
Metaproteomics					

The National Microbiome Data Collaborative (NMDC) supports a Findable, Accessible, Interopera

For any issues or questions that require support from a team member, please contact *nmdc-edge@lanl.gov* or open an issue on GitHub.

LA-UR-2

Upload Files

Max single file size is 10.0GB. Max server storage space is 150.0GB. Files will be kept for 180 days.

Allowed file extensions are: fastq, fq, faa, fa, fasta, fna, contigs, fastq.gz, fq.gz, fa.gz, fasta.gz, fna.gz, contigs.gz, fa.bz2, fasta.bz2, contigs.bz2, fna.bz2, fa.xz, fasta.xz, contigs.xz, fna.xz, gbk, gff, genbank, gb, xlsx, txt, bed, config, tsv, csv, raw, d, bam, sam

Storage space usage: 0bytes/150.0GB

Uploading size: 0bytes

Drag Files or Click to Browse

Run Multiple Workflows

Project/Run Name ⚠: (required, at 3 but less than 30 characters)

Description: (optional)

Input Raw Reads

Input

Input Raw Reads
Is interleaved? [✓ Yes] No

Input interleaved fastq: Add interleaved fastq

interleaved FASTQ #1: Select a file or enter a file http(s) url
Remove

Choose Workflows

All of the NMDC Metagenomic workflows are connected in EDGE, so that the output of one workflow can automatically be the input for the next workflow if several metagenomic workflows are selected.

ReadsQC Workflow [✓ On] Off

Read-based Taxonomy Classification Workflow [✓ On] Off

Select Analysis Tool(s): GOTTCHA2 ×, Kraken2 ×, Centrifuge ×

Metagenome Assembly Workflow [✓ On] Off

Metagenome Annotation Workflow [✓ On] Off

Metagenome MAGs Workflow [✓ On] Off

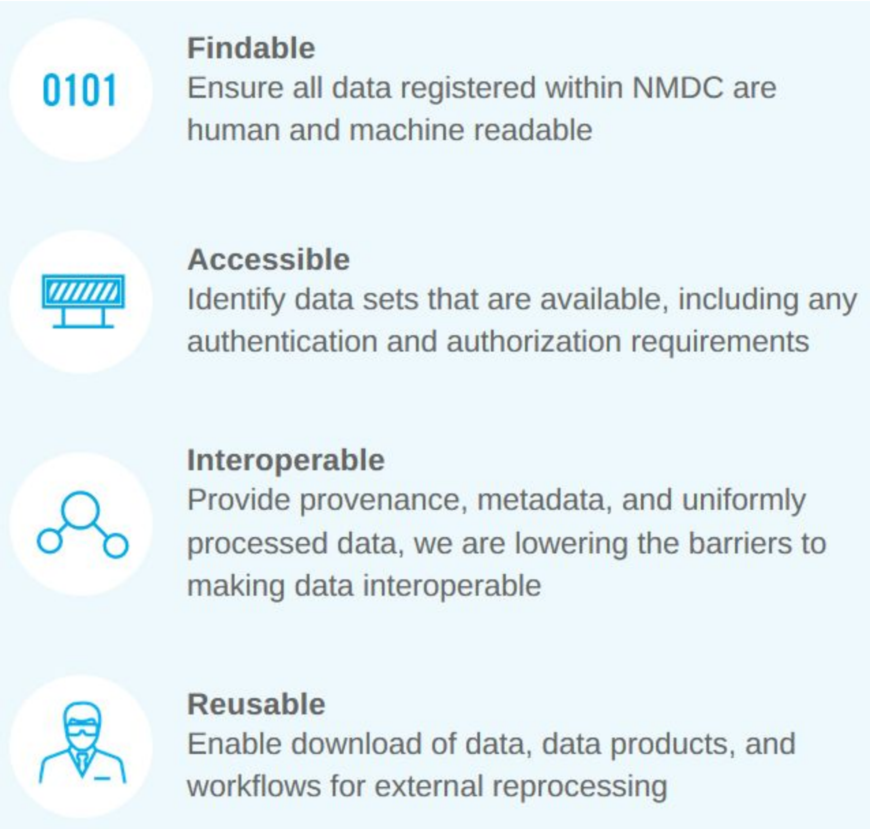
Input Map File: (Optional) Select a file or enter a file http(s) url

Input Domain File: (Optional) Select a file or enter a file http(s) url

Submit Reset

Data Stewardship + FAIR Data

- NMDC is committed to FAIR data principles of making microbiome data Findable, Accessible, Interoperable, and Reusable
- Both raw and processed data should be FAIR
- Processing data in a standardized way makes it increasingly interoperable and reusable



0101 **Findable**
Ensure all data registered within NMDC are human and machine readable

Accessible
Identify data sets that are available, including any authentication and authorization requirements

Interoperable
Provide provenance, metadata, and uniformly processed data, we are lowering the barriers to making data interoperable

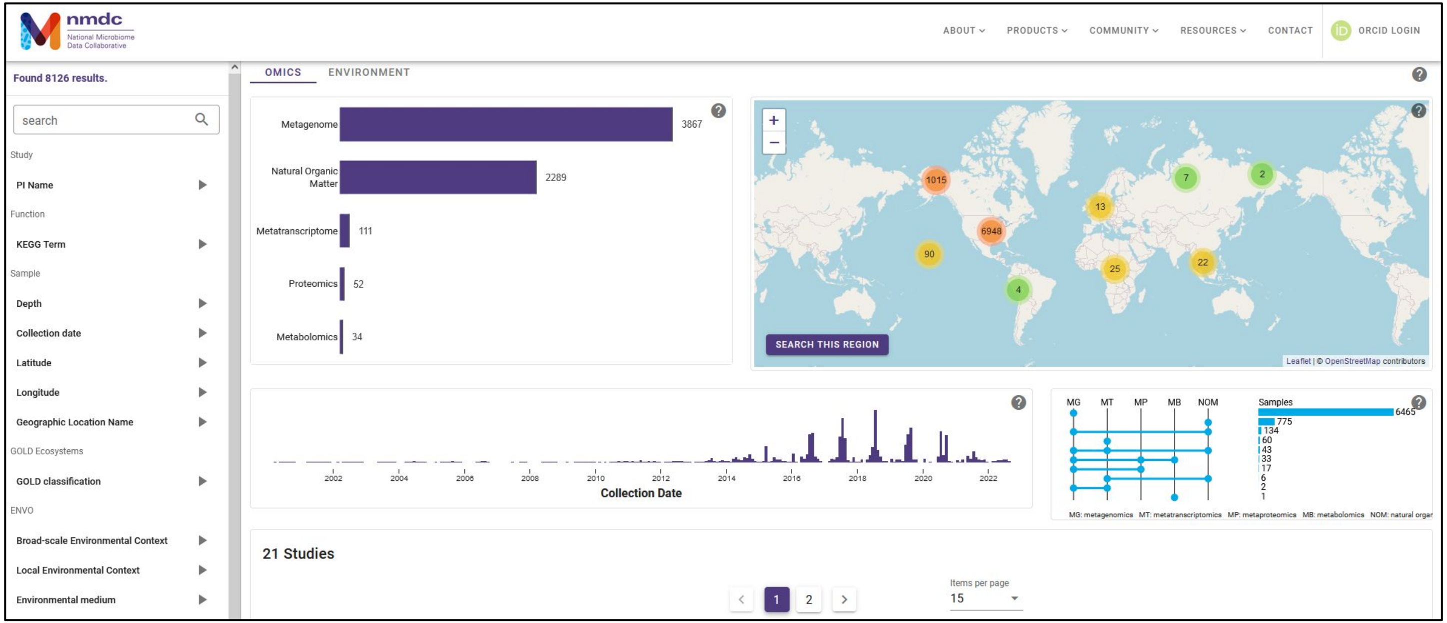
Reusable
Enable download of data, data products, and workflows for external reprocessing

NMDC Data Portal

nmdc
National Microbiome Data Collaborative

Data Portal & API

Access and discovery of microbiome information



Found 8126 results.

21 Studies

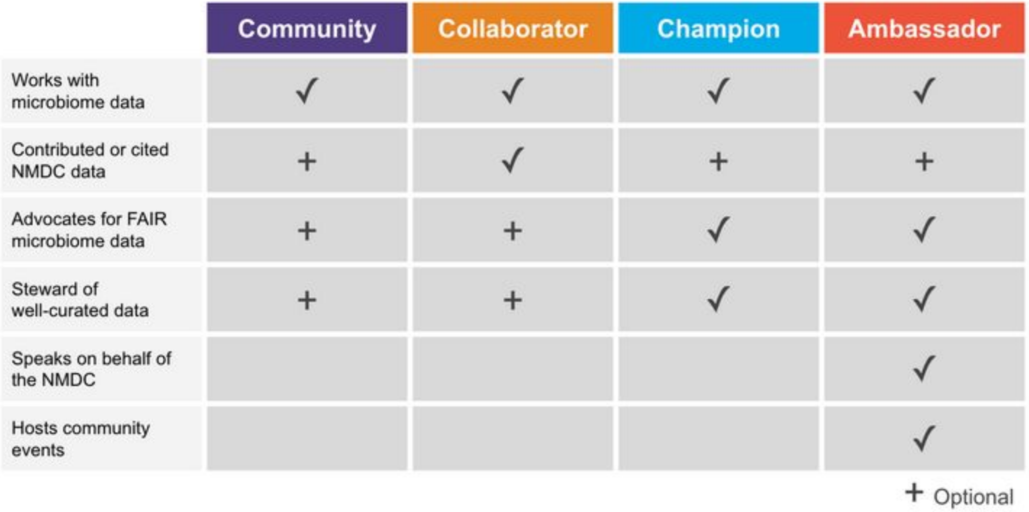
NMDC Engagement

nmdc
National Microbiome Data Collaborative

Goal 1: Recognize and support the diverse research needs and perspectives of the microbiome research community

Goal 2: Promote best practices across the microbiome community, from researchers to funders

Goal 3: Build a microbiome ecosystem that enables scientific discovery and innovation



	Community	Collaborator	Champion	Ambassador
Works with microbiome data	✓	✓	✓	✓
Contributed or cited NMDC data	+	✓	+	+
Advocates for FAIR microbiome data	+	+	✓	✓
Steward of well-curated data	+	+	✓	✓
Speaks on behalf of the NMDC				✓
Hosts community events				✓

+ Optional

Partnerships

IMMSA

JAWS
JGI Analysis Workflow Service

JGI
JOINT GENOME INSTITUTE

ESS-DIVE
Deep Insight for Earth Science Data

AMERICAN SOCIETY FOR MICROBIOLOGY

KBase
PREDICTIVE BIOLOGY

NSF

neon
Operated by Battelle

gsc genomic STANDARDS consortium

EMSL

DMPTool

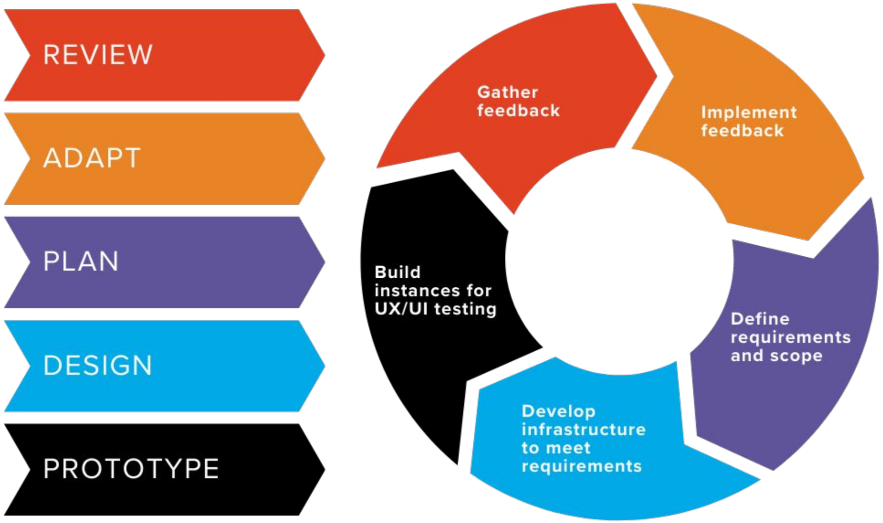
MICROBIOME CENTERS CONSORTIUM

NASA

GeneLab

Community-centered design process

nmdc
National Microbiome Data Collaborative


REVIEW
ADAPT
PLAN
DESIGN
PROTOTYPE
Gather feedback
Implement feedback
Define requirements and scope
Develop infrastructure to meet requirements
Build instances for UX/UI testing

2023-2024 User Research

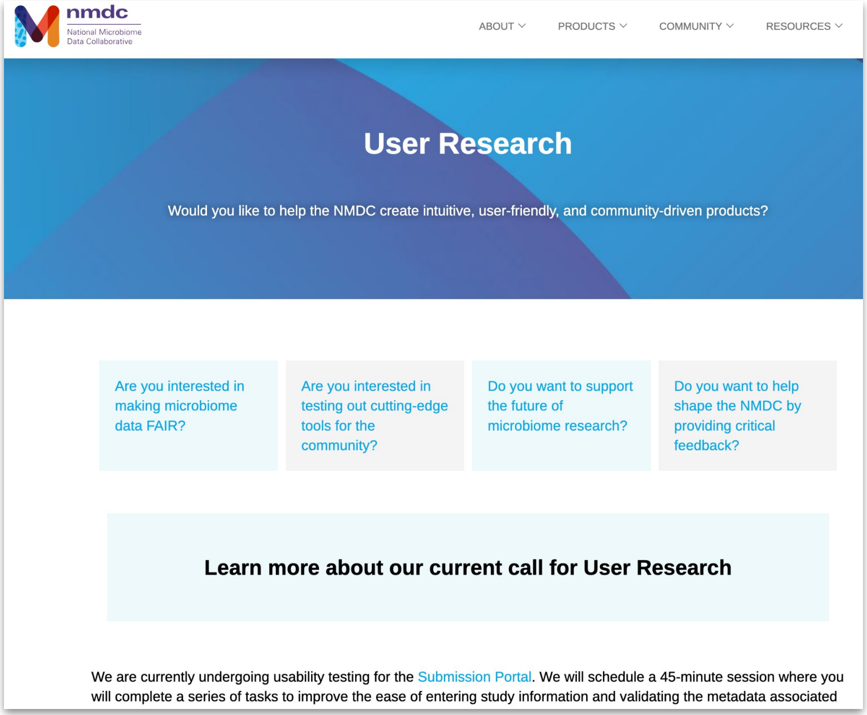
34
User research participants

321
Insights from user interviews

120
Action items from insights

- The NMDC program utilizes a community-centered design approach to its product development process
  - ○ User interviews, usability testing, beta-testing
  - ○ Understand the needs of the community and how researchers use the products
- Workshops and event feedback is critical too
- Want to develop products that are as useful to the community as possible

User Research



User Research

Would you like to help the NMDC create intuitive, user-friendly, and community-driven products?

Are you interested in making microbiome data FAIR?

Are you interested in testing out cutting-edge tools for the community?

Do you want to support the future of microbiome research?

Do you want to help shape the NMDC by providing critical feedback?

Learn more about our current call for User Research

We are currently undergoing usability testing for the Submission Portal. We will schedule a 45-minute session where you will complete a series of tasks to improve the ease of entering study information and validating the metadata associated

Examples of feedback:

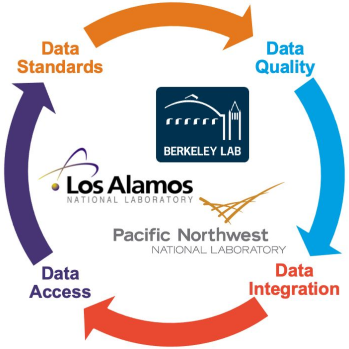
- Updates to interface design, tutorials, and help guidance
- Feature implementation
  - Batch processing in NMDC EDGE
  - Data download updates in the Data Portal
  - Template accessibility updates in the Submission Portal

Volunteer for our user research program!

https://microbiomedata.org/user-research/

How to talk about the NMDC

U.S. DEPARTMENT OF ENERGY


Data Standards
Data Quality
BERKELEY LAB
Los Alamos NATIONAL LABORATORY
Pacific Northwest NATIONAL LABORATORY
Data Access
Data Integration

What is NMDC?

NMDC is a DOE funded *data discovery platform* that promotes open science for multi-omic microbiome research to accelerate scientific discovery.

What we are NOT

- Primary data repository

FAQs available on the NMDC website

Three NMDC products

The people doing the work:
The NMDC team

The organization:
The NMDC

Annual Reports

nmdc
National Microbiome
Data Collaborative

nmdc
National Microbiome
Data Collaborative

National Microbiome Data Collaborative

Progress Report 2023
U.S. Department of Energy

Los Alamos NATIONAL LABORATORY
BERKELEY LAB
Pacific Northwest NATIONAL LABORATORY

Advancing microbiome science for the benefit of all

In January of this year, the White House Office of Science and Technology Policy launched the Year of Open Science with new actions to advance open and equitable science policies across the federal government. The National Microbiome Data Collaborative (NMDC) is committed to open and equitable research in microbiome science. This commitment is the foundation for our infrastructure development activities and has been a prominent driver this past year across our three products: the Submission Portal, NMDC EDGE, and the Data Portal. It has been an exciting and busy time for our team as we work to serve the scientific community in a way that enables microbiome innovation and discovery. I am exceptionally proud of the progress the NMDC team has made this past year in fostering strong community partnerships and advancing our powerful products into tools that drive scientific impact.

We launched the NMDC persistent identifier service in January 2023 and deployed programmatic access to all NMDC data through a public application programming interface (API). Together, these efforts support a larger findable, accessible, interoperable, and reusable (FAIR) data ecosystem to programmatically exchange and link data across resources. The Data Portal now hosts over 7,700 biosamples with a collective nearly 90 TB of multi-omics microbiome data and links across complementary data platforms, including the Integrated Microbial Genomes and Microbiomes (IMG/M) and Genomes OnLine Database (GOLD) of the Joint Genome Institute (JGI), the Department of Energy (DOE) Systems Biology Knowledgebase (KBase), the National Center for Biotechnology Information (NCBI), the Mass Spectrometry Interactive Virtual Environment (MassIVE), the Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), and the National

Emiley Eloe-Fadrosh
National Microbiome
Data Collaborative Lead

A community-driven data infrastructure

The NMDC is tackling existing gaps in microbiome research by using proven approaches and new innovations in distributed data infrastructure and linked data technologies. Our three products — the Submission Portal, NMDC EDGE, and the Data Portal — are driven by community needs. They support data, information, knowledge sharing, and access. This past year, we worked closely with the research community to strengthen our existing infrastructure in ways that will catalyze new research. This included launching a new persistent identifier service, supporting programmatic access to NMDC data, expanding the amount of available high-quality data and workflows, and contributing to major updates of community data standards.

Sevilleta Long Term Ecological Research (LTER) Site in New Mexico. Credit: Buck Hanson

4

https://microbiomedata.org/annual_report/

NMDC Resources

nmdc
National Microbiome Data Collaborative

Website: https://microbiomedata.org/
Data Portal: https://data.microbiomedata.org/
Submission Portal: https://data.microbiomedata.org/submission/home
NMDC EDGE: https://nmdc-edge.org/home
Github: https://github.com/microbiomedata
Docker Hub: https://hub.docker.com/u/microbiomedata
Documentation:
https://nmdc-documentation.readthedocs.io/en/latest/overview/nmdc_overview.html
YouTube: https://www.youtube.com/channel/UCyBqKc46NQZ_YgZIKGYeglw/featured

Get involved!

Sign up for our newsletter
microbiomedata.org

Become a NMDC Champion
bit.ly/championsapp

Find us on X/Twitter
@microbiomedata

Find us on LinkedIn
https://bit.ly/NMDC_LinkedIn

Find us on Instagram
@microbiomedata

Read more about the NMDC

nature microbiology

Kelliher JM *et al.* Cohort-based learning for microbiome research community standards. *Nat Microbiol* (2023). doi.org/10.1038/s41564-023-01361-7.

frontiers

Hu B, Canon S, Eloe-Fadrosh EA, et al.. Challenges in Bioinformatics Workflows for Processing Microbiome Omics Data at Scale. *Front Bioinform*. 1:826370. (2022) doi: 10.3389/fbinf.2021.826370.

Nucleic Acids Research

Eloe-Fadrosh EA *et al*. The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res.* 7;60(D1):D828–D836. (2022) doi: 10.1093/nar/gkab990.

nature REVIEWS MICROBIOLOGY

Wood-Charlson, E.M., Anubhav, Auberry, D. *et al.* The National Microbiome Data Collaborative: enabling microbiome science. *Nat Rev Microbiol* **18,** 313–314 (2020). doi.org/10.1038/s41579-020-0377-0

mSystems

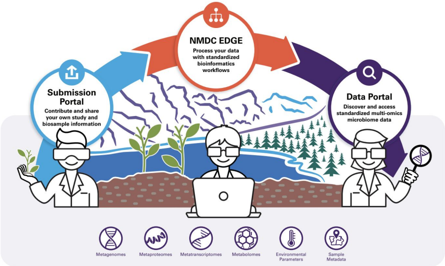
Vangay, P *et al.* Microbiome metadata standards: Report of the National Microbiome Data Collaborative's workshop and follow-on activities. *mSystems* **6**, e01194-20 (2021). doi.org/10.1128/mSystems.01194-20

‘One Slide on the NMDC’ for Presentations

National Microbiome Data Collaborative

Mission

To support a FAIR microbiome data sharing network, through **infrastructure**, **data standards**, and **community building**, that addresses pressing challenges in environmental sciences


NMDC EDGE
Process your data with standardized bioinformatics workflows
Submission Portal
Contribute and share your own study and biosample information
Data Portal
Discover and access standardized multi-omics microbiome data
Metagenomes
Metaproteomes
Metatranscriptomes
Metabolomes
Environmental Parameters
Sample Metadata

Findable
Accessible
Interoperable
Reusable

The NMDC offers 3 Products:

Submission Portal

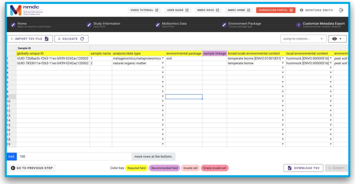
Lower barriers to collect study and biosample data

NMDC EDGE

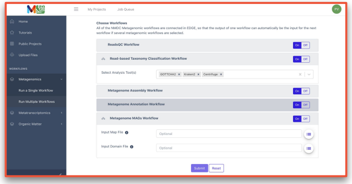
Streamline multi-omics data processing

Data Portal & API

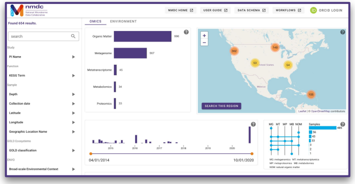
Access and discovery of microbiome information



Submission Portal



NMDC EDGE



NMDC Data Portal

Get involved!

Website/newsletter
microbiomedata.org

Become a Champion
bit.ly/champions-program

Find us on X/Twitter
@microbiomedata

Find us on LinkedIn
https://bit.ly/NMDC_LinkedIn

Find us on Instagram
@microbiomedata

Engagement

User Facilities

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY
EMSL

Individuals

The NMDC Champions Program
The NMDC Ambassador Program

Strategic Partners

NSF
neon Operated by Battelle
NASA
GeneLab

nmdc

National Microbiome Data Collaborative

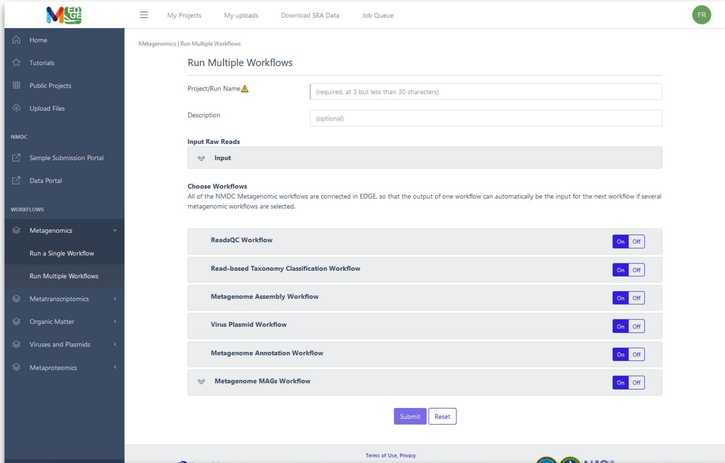
Standardized Workflows and NMDC EDGE

2024 Ambassador Cohort

Purpose of this training

nmdc
National Microbiome
Data Collaborative

- Provide an overview of the benefits of standardized bioinformatics workflows and how they promote FAIR data
- Discuss the importance and benefits of using standardized workflows for the future of microbiome data
- Provide template slides that anyone can use in their own events
- Introduce audience to NMDC EDGE
  - The hands-on activity can be used (or modified and used) for use in your own events too! You can upload or link your own data



Metagenomics | Run Multiple Workflows

Run Multiple Workflows

Project/Run Name ⚠ (required, at 3 but less than 30 characters)

Description (optional)

Input Raw Reads

Input

Choose Workflows

All of the NMDC Metagenomic workflows are connected in EDGE, so that the output of one workflow can automatically be the input for the next workflow if several metagenomic workflows are selected.

ReadsQC Workflow [✓ On] Off

Read-based Taxonomy Classification Workflow [✓ On] Off

Metagenome Assembly Workflow [✓ On] Off

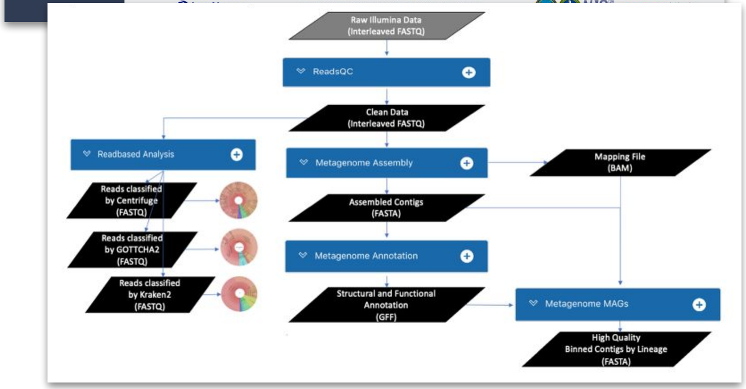
Virus Plasmid Workflow [✓ On] Off

Metagenome Annotation Workflow [✓ On] Off

Metagenome MAGs Workflow [✓ On] Off

Submit Reset

Terms of Use, Privacy


Raw Illumina Data (Interleaved FASTQ)

ReadsQC

Clean Data (Interleaved FASTQ)

Readbased Analysis

Reads classified by Centrifuge (FASTQ)

Reads classified by GOTTCHA2 (FASTQ)

Reads classified by Kraken2 (FASTQ)

Metagenome Assembly

Mapping File (BAM)

Assembled Contigs (FASTA)

Metagenome Annotation

Structural and Functional Annotation (GFF)

Metagenome MAGs

High Quality Binned Contigs by Lineage (FASTA)

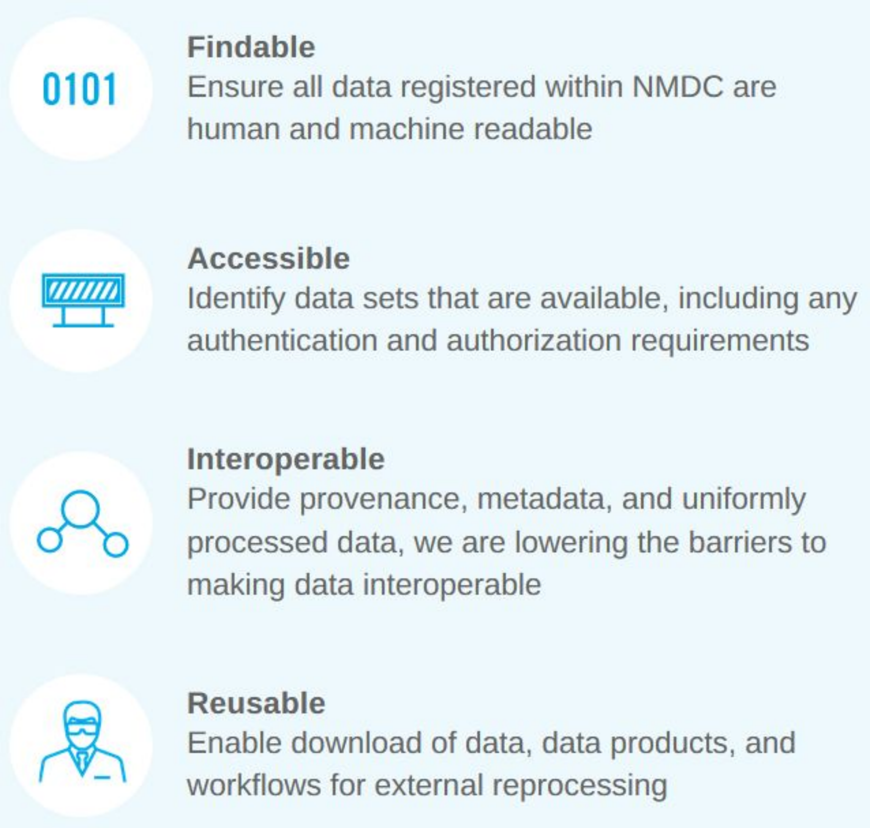
nmdc

National Microbiome
Data Collaborative

Standardized Bioinformatic Workflows

FAIR Data

- NMDC is committed to FAIR data principles of making microbiome data Findable, Accessible, Interoperable, and Reusable
- Raw and **processed** data should be FAIR
- NMDC prioritizes standardization and user-centered design to achieve FAIR and open data



0101

Findable
Ensure all data registered within NMDC are human and machine readable

Accessible
Identify data sets that are available, including any authentication and authorization requirements

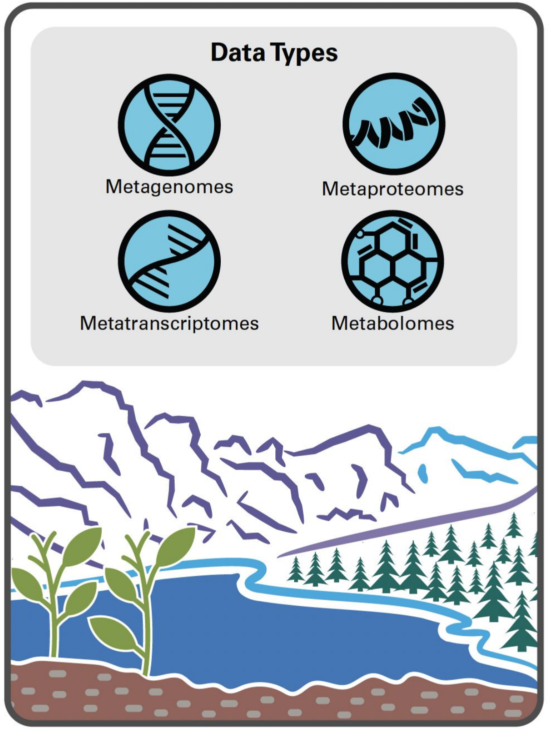
Interoperable
Provide provenance, metadata, and uniformly processed data, we are lowering the barriers to making data interoperable

Reusable
Enable download of data, data products, and workflows for external reprocessing

Intro to multi-omics techniques

- Multi-omics: the integration of data from multiple omics techniques, such as genomics, transcriptomics, proteomics, and metabolomics, to provide a comprehensive assessment of a biological system

- By analyzing various layers of molecular information simultaneously, taxonomic, genetic, and functional information can be used to unravel information about an organism and/or a community



Multi-omics data processing

- Multi-omics data collection is rapidly becoming one of the most effective ways to interrogate microbiomes
  - The infrastructure surrounding these data is not keeping up
    - Large amounts of compute power needed to store and process these data
    - Data not FAIR (not made publicly accessible, not utilizing standards, etc.)
    - Different omics data not connected, data not comparable within an omics type
- Processing these omics data streams is currently very difficult to do in a way that allows for effective cross-study comparisons and data re-use

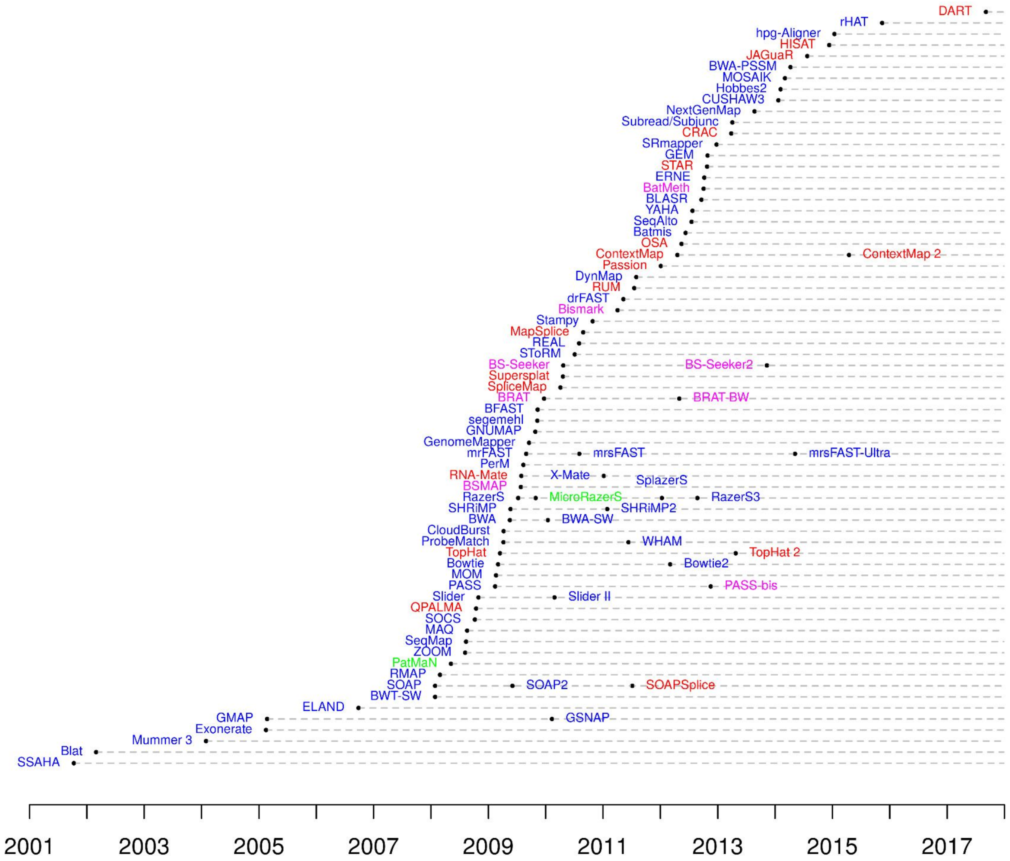
Metagenomes
Metaproteomes
Metatranscriptomes
Metabolomes

Bioinformatics explosion

nmdc
National Microbiome Data Collaborative

The massive explosion in the number of bioinformatics tools and workflows has led to data being processed in many ways, thus **limiting between-study comparability**

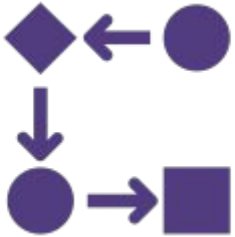
These are ***just*** mapping tools for genomic data developed between 2001-2018.

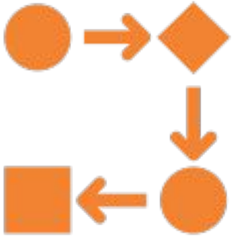


Timeline of NGS read aligners. Image from Nuno Fonseca https://www.ecseq.com/support/ngs/what-is-the-best-ngs-alignment-software

Benefits of standardized workflows

- Help to make data and data products FAIR
- Reproducibility within a lab and between labs
- Better keep track of what was done to data
- Allows for integrations and comparisons with other datasets from other studies
- No need to put together own workflows
  - Existing workflows can include the best tools for omics data processing
  - Saves research, implementation, and testing time





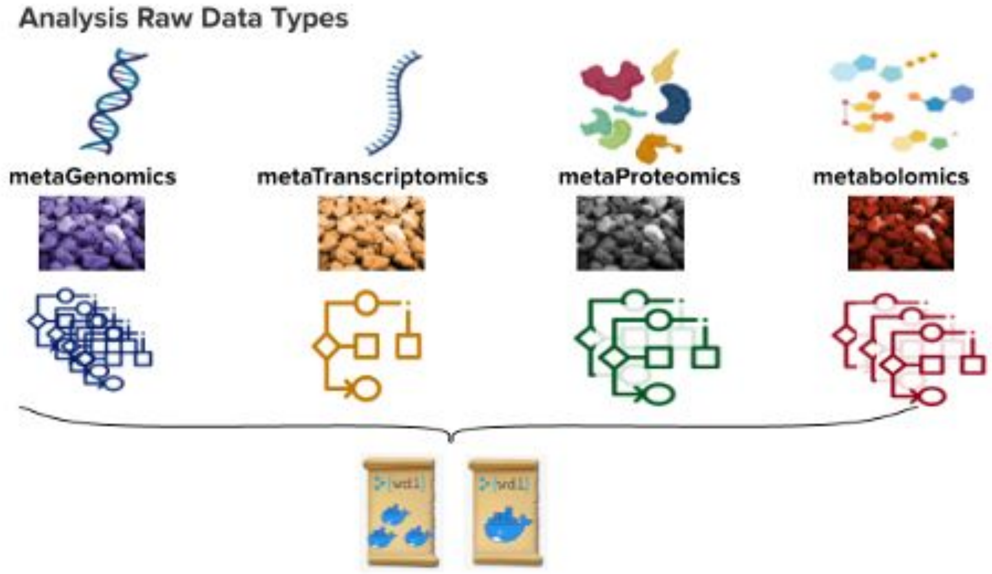
NMDC Bioinformatics Workflows

nmdc
National Microbiome Data Collaborative

The NMDC has integrated state of the art open-source bioinformatics tools into standardized workflows for processing raw multi-omics data to produce **interoperable** and **reusable** annotated data products.

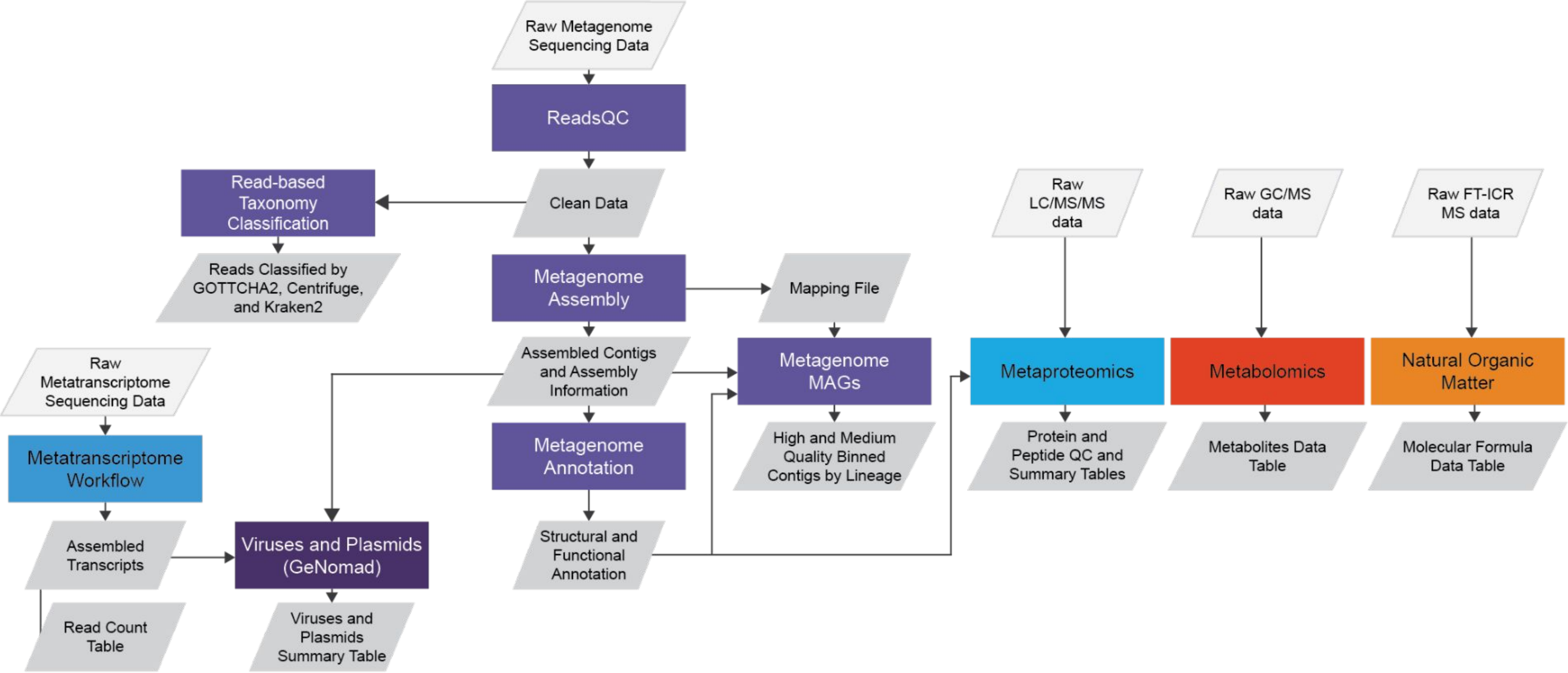
NMDC Workflows:

- Metagenome data
  - ReadsQC
  - Read-based taxonomy classification
  - Assembly
  - Annotation
  - Metagenome assembled genomes (MAGs)
- Metatranscriptome data
- Natural organic matter data
- Metabolome data
- Metaproteome data
- Viruses & Plasmids



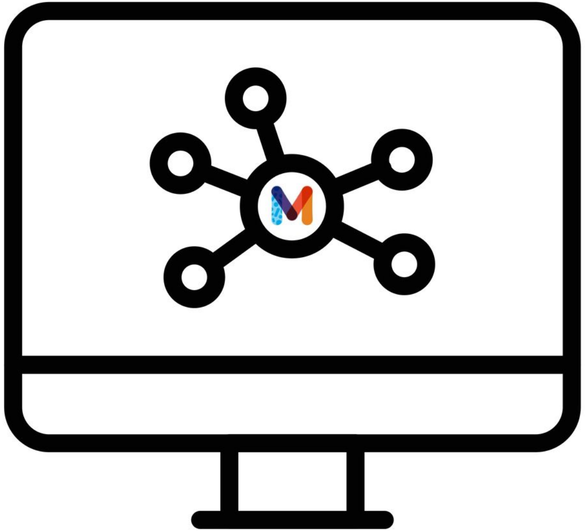
NMDC Workflows

nmdc
National Microbiome Data Collaborative


Raw Metagenome Sequencing Data
ReadsQC
Clean Data
Read-based Taxonomy Classification
Reads Classified by GOTTCHA2, Centrifuge, and Kraken2
Metagenome Assembly
Mapping File
Assembled Contigs and Assembly Information
Metagenome MAGs
High and Medium Quality Binned Contigs by Lineage
Metagenome Annotation
Structural and Functional Annotation
Raw Metatranscriptome Sequencing Data
Metatranscriptome Workflow
Assembled Transcripts
Read Count Table
Viruses and Plasmids (GeNomad)
Viruses and Plasmids Summary Table
Raw LC/MS/MS data
Metaproteomics
Protein and Peptide QC and Summary Tables
Raw GC/MS data
Metabolomics
Metabolites Data Table
Raw FT-ICR MS data
Natural Organic Matter
Molecular Formula Data Table

Why the NMDC workflows?

nmdc
National Microbiome Data Collaborative



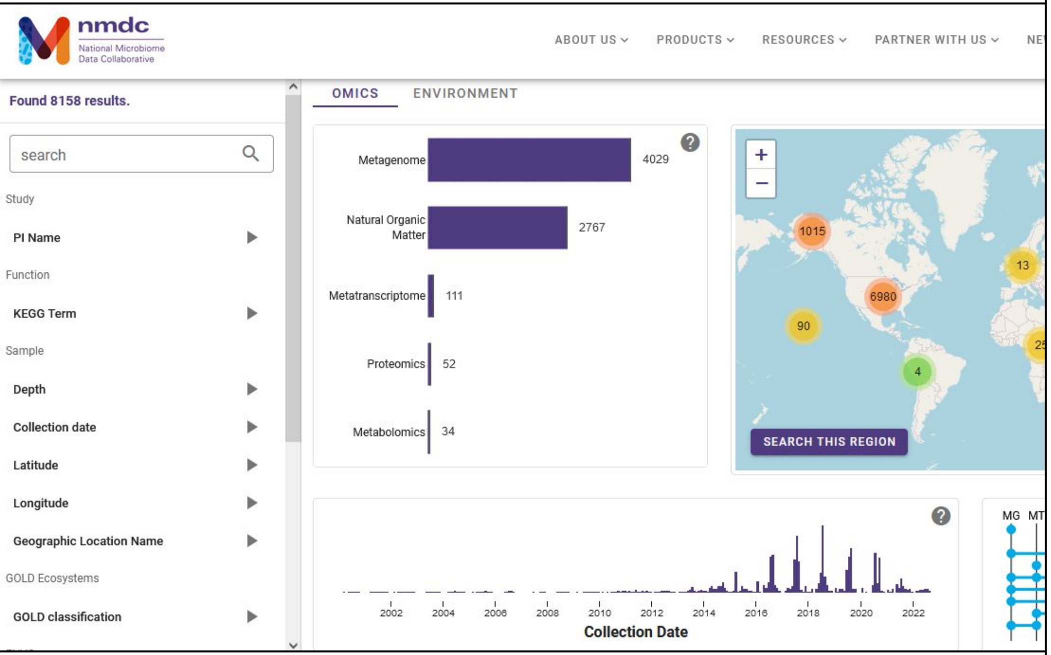
Benefits of using the NMDC workflows:

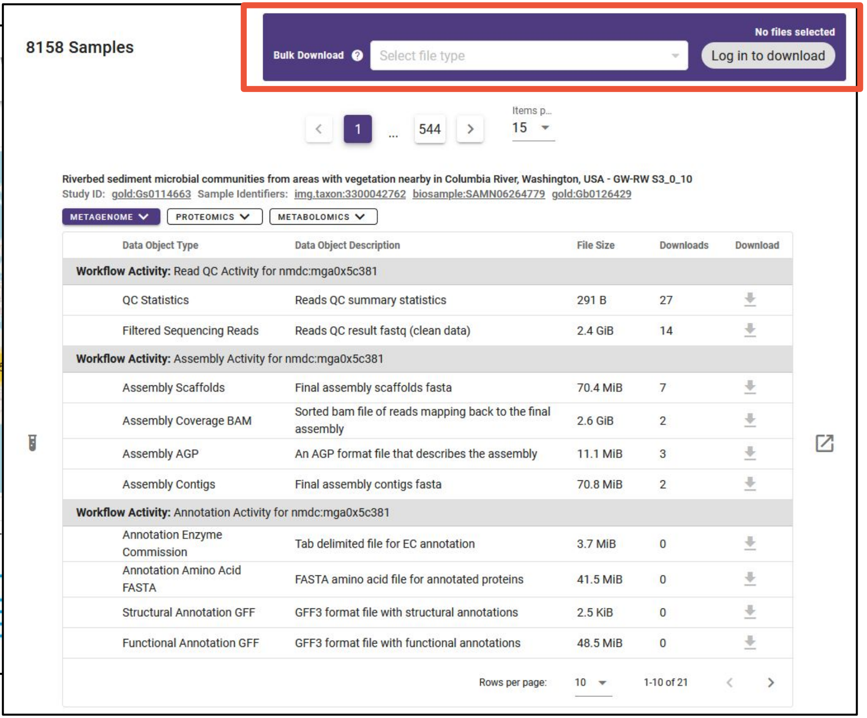
- Tools were carefully researched, selected, and modified for optimal performance
    - Many tools are the tools are production quality, regularly used to process thousands of datasets from DOE user facilities
- Workflows have been extensively tested on data from dozens of institutions and sample types
- Users are able to run these workflows through shared computing resources
    - Users don't need to download these tools or databases, nor have access to their own computing clusters
- The workflows are offered in a user-friendly interface for users with any level of bioinformatics experience
- Open source platform, extensive documentation

NMDC Data Portal

nmdc
National Microbiome Data Collaborative

Processed datasets available on the NMDC Data Portal have all been run through the NMDC workflows allowing for direct comparisons between this data and data processed in NMDC EDGE





Riverbed sediment microbial communities from areas with vegetation nearby in Columbia River, Washington, USA - GW-RW S3_0_10

Study ID: gold:Gs0114663 Sample Identifiers: img.taxon:3300042762 biosample:SAMN06264779 gold:Gb0126429

Data Object Type	Data Object Description	File Size	Downloads	Download
Workflow Activity: Read QC Activity for nmdc:mga0x5c381				
QC Statistics	Reads QC summary statistics	291 B	27	⬇
Filtered Sequencing Reads	Reads QC result fastq (clean data)	2.4 GiB	14	⬇
Workflow Activity: Assembly Activity for nmdc:mga0x5c381				
Assembly Scaffolds	Final assembly scaffolds fasta	70.4 MiB	7	⬇
Assembly Coverage BAM	Sorted bam file of reads mapping back to the final assembly	2.6 GiB	2	⬇
Assembly AGP	An AGP format file that describes the assembly	11.1 MiB	3	⬇
Assembly Contigs	Final assembly contigs fasta	70.8 MiB	2	⬇
Workflow Activity: Annotation Activity for nmdc:mga0x5c381				
Annotation Enzyme Commission	Tab delimited file for EC annotation	3.7 MiB	0	⬇
Annotation Amino Acid FASTA	FASTA amino acid file for annotated proteins	41.5 MiB	0	⬇
Structural Annotation GFF	GFF3 format file with structural annotations	2.5 KiB	0	⬇
Functional Annotation GFF	GFF3 format file with functional annotations	48.5 MiB	0	⬇

Rows per page: 10 1-10 of 21

nmdc

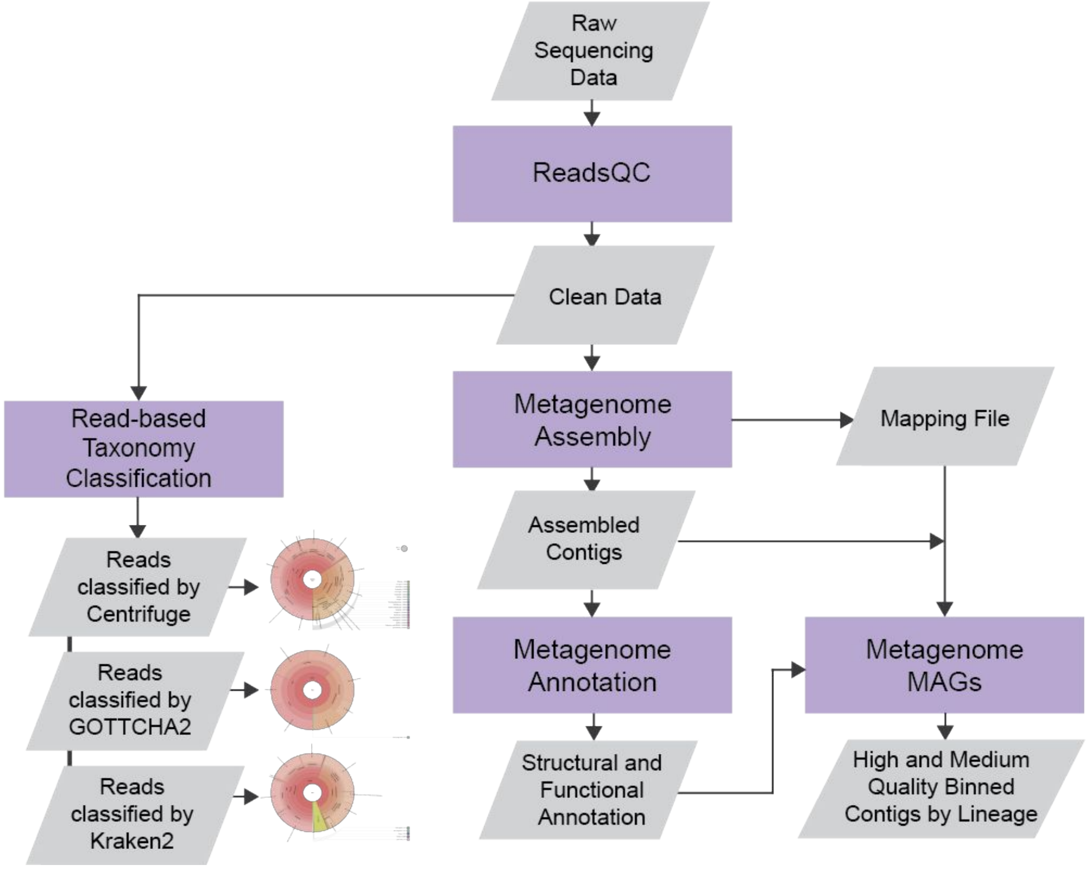
National Microbiome Data Collaborative

Metagenome Workflows

NMDC Metagenomic Workflow

nmdc

National Microbiome Data Collaborative



Raw Sequencing Data

ReadsQC

Clean Data

Read-based Taxonomy Classification

Reads classified by Centrifuge

Reads classified by GOTTCHA2

Reads classified by Kraken2

Metagenome Assembly

Mapping File

Assembled Contigs

Metagenome Annotation

Structural and Functional Annotation

Metagenome MAGs

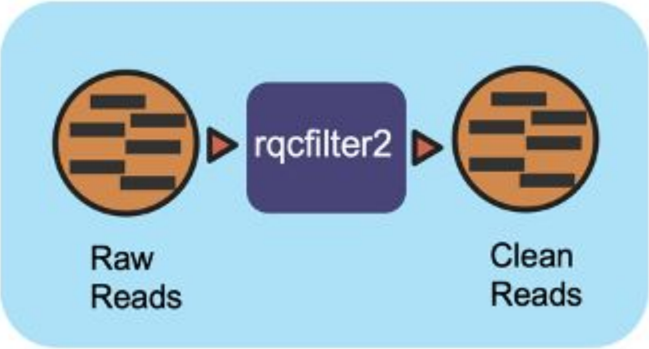
High and Medium Quality Binned Contigs by Lineage

Reads QC

Performs quality control on raw metagenome Illumina reads to trim/filter low quality data and to remove artifacts, linkers, adapters, spike-in reads and reads mapping to several hosts and common microbial contaminants.

Input: Raw Illumina data

Output: File of cleaned reads and QC statistics

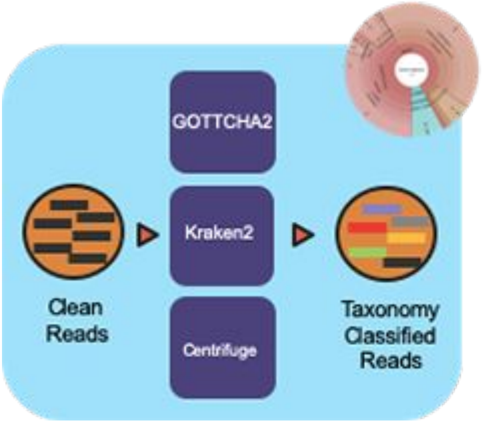


Read-based Taxonomy Classification

Takes in Illumina sequencing files and profiles the reads using 3 taxonomic classification tools (GOTTCHA2, Kraken2, and Centrifuge) with a range of sensitivity and specificity

Input: Illumina data: it is highly recommended to input clean reads from the ReadsQC workflow

Output: Results for each tool at three taxonomic levels (Species, Genus, and Family). Interactive Krona plots are also generated

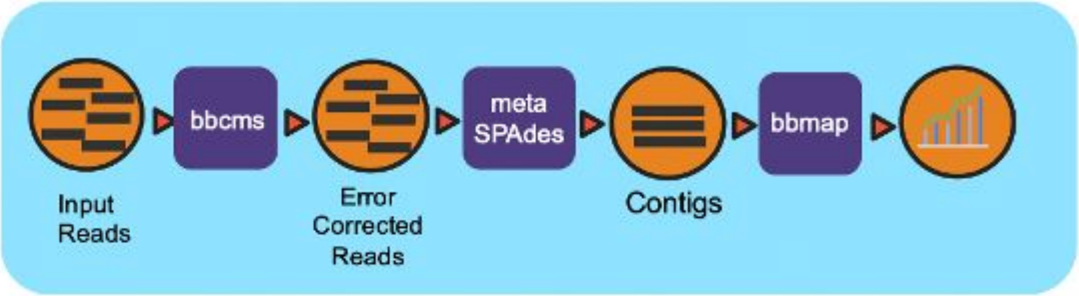


Metagenome Assembly

Takes in Illumina data, runs error correction, assembly, and assembly validation

Input: Illumina data: recommended input is the output from the ReadsQC workflow

Output: File of assembled contigs; assembly statistics

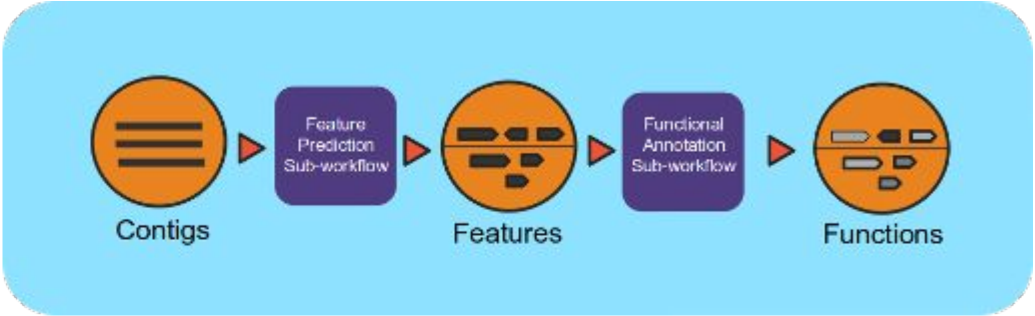


Metagenome Annotation

Takes in assembled metagenomes and generates structural and functional annotations

Input: Assembled contigs: recommended input is the output from the Metagenome Assembly workflow

Output: Structural annotation file, functional annotation file, several summary files

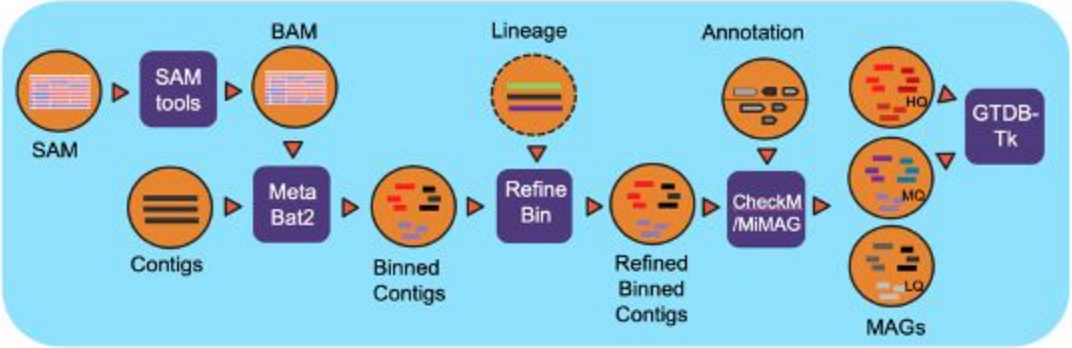

Contigs
Feature Prediction Sub-workflow
Features
Functional Annotation Sub-workflow
Functions

Metagenome Assembled Genomes

Classifies contigs into bins, bins are refined using functional annotation file, bins are evaluated for completeness and contamination. Quality of bins is determined and lineage is assigned

Input: Assembled contigs, read mapping file from the assembly, functional annotation of the assembly

Output: Summary statistics, file of high quality (HQ) and medium quality (MQ) bins



nmdc

National Microbiome
Data Collaborative

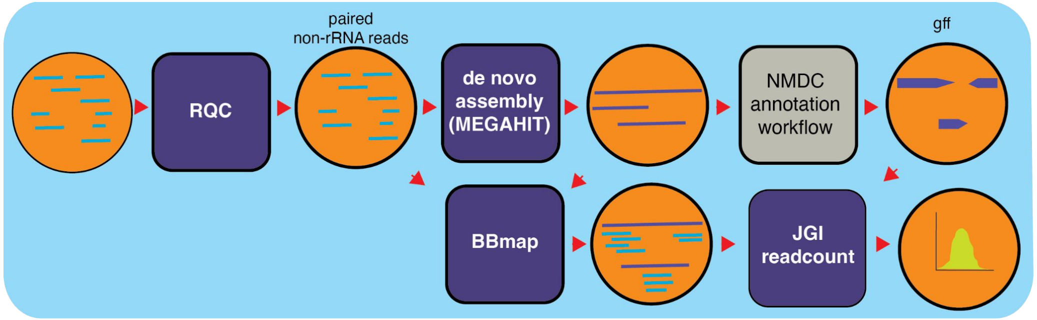
Additional -omics Workflows

Metatranscriptome Workflow

Takes in raw metatranscriptome data, filters data for quality, removes rRNA reads, assembles and annotates the transcripts. Data is mapped back to the genomic features in the transcripts and RPKMs (Reads Per Kilobase of transcript per Million mapped reads) are calculated for each feature in the functional annotation file.

Input: Illumina data

Output: Assembled transcripts, annotated features file, annotation files

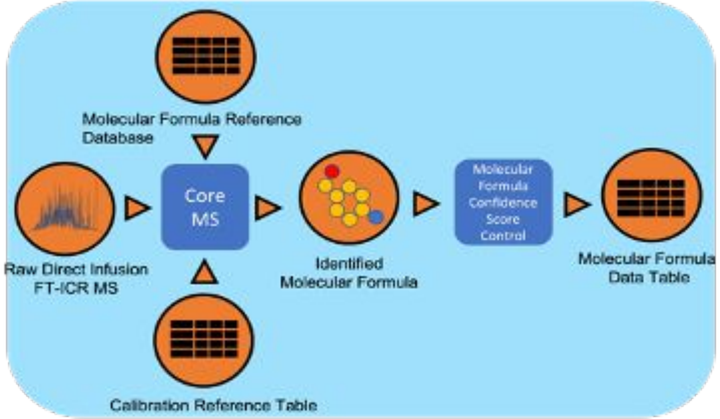


Natural Organic Matter Workflow

Takes mass spectrometry data collected from organic extracts to determine the molecular formulas of natural organic biomolecules in the input sample.

Input: The output from a mass spec experiment; a calibration file of molecular formula references is also required when running via command line

Output: Primary output file is the Molecular Formula Data Table

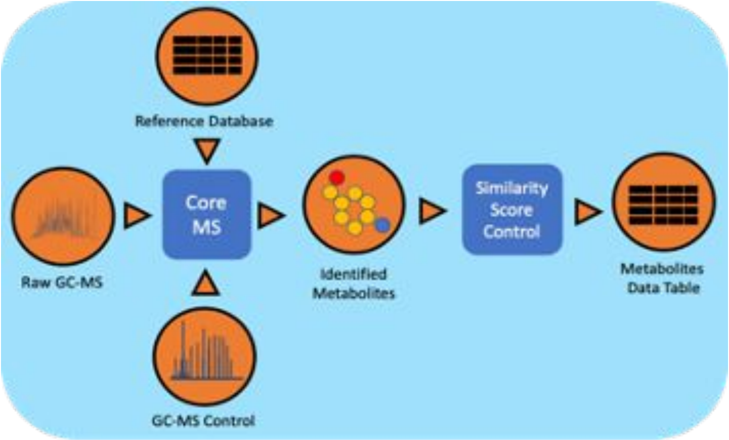

Molecular Formula Reference Database
Raw Direct Infusion FT-ICR MS
Core MS
Identified Molecular Formula
Molecular Formula Confidence Score Control
Molecular Formula Data Table
Calibration Reference Table

Metabolome Workflow

The GC-MS based metabolomics workflow leverages PNNL's CoreMS software framework.

Input: Raw GC-MS data

Output: Metabolites data table

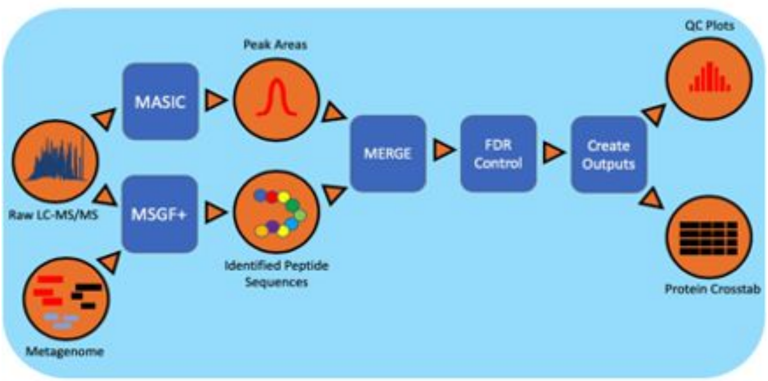


Metaproteome Workflow

This workflow is an end-to-end data processing workflow for protein identification and characterization using MS/MS data

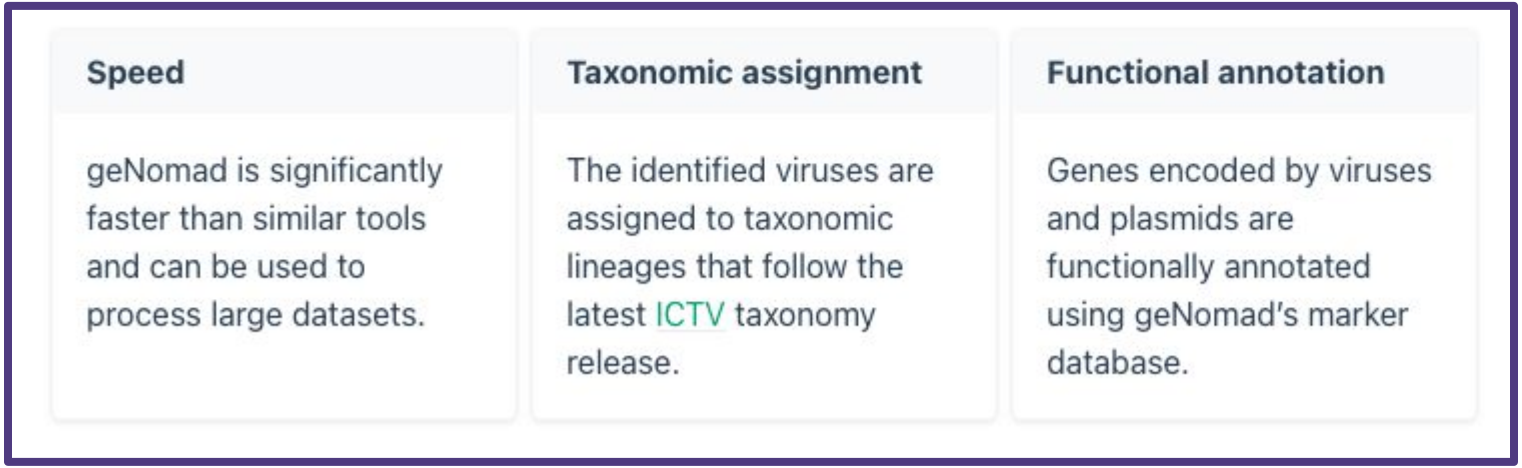
Input: Raw LC-MS/MS data and an associated metagenome file

Output: Protein crosstab; QC plots



Viruses & Plasmids Workflow

geNomad is a tool that identifies virus and plasmid genomes from nucleotide sequences. It provides state-of-the-art classification performance and can be used to quickly find mobile genetic elements from genomes, metagenomes, or metatranscriptomes.



Speed

geNomad is significantly faster than similar tools and can be used to process large datasets.

Taxonomic assignment

The identified viruses are assigned to taxonomic lineages that follow the latest ICTV taxonomy release.

Functional annotation

Genes encoded by viruses and plasmids are functionally annotated using geNomad's marker database.

Workflow Availability

All workflows can be downloaded and run locally on your own computing resources, or can be run through NMDC EDGE

GitHub: **https://github.com/microbiomedata**

Docker Hub: **https://hub.docker.com/u/microbiomedata**

nmdc

National Microbiome Data Collaborative

Questions?

nmdc

National Microbiome
Data Collaborative

NMDC EDGE

Product Initiatives

nmdc
National Microbiome Data Collaborative

Submission Portal

Lower barriers to collect study and biosample data

NMDC EDGE

Streamline multi-omics data processing

Data Portal & API

Access and discovery of microbiome information

Engagement

User Facilities

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

EMSL

Individuals

The NMDC Champions Program

The NMDC Ambassador Program

Strategic Partners

NSF

neon Operated by Battelle

NASA

GeneLab

Product Initiatives

nmdc
National Microbiome Data Collaborative

Submission Portal

Lower barriers to collect study and biosample data

NMDC EDGE

Streamline multi-omics data processing

Data Portal & API

Access and discovery of microbiome information

Engagement

User Facilities

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

EMSL

Individuals

The NMDC Champions Program

The NMDC Ambassador Program

Strategic Partners

NSF

neon Operated by Battelle

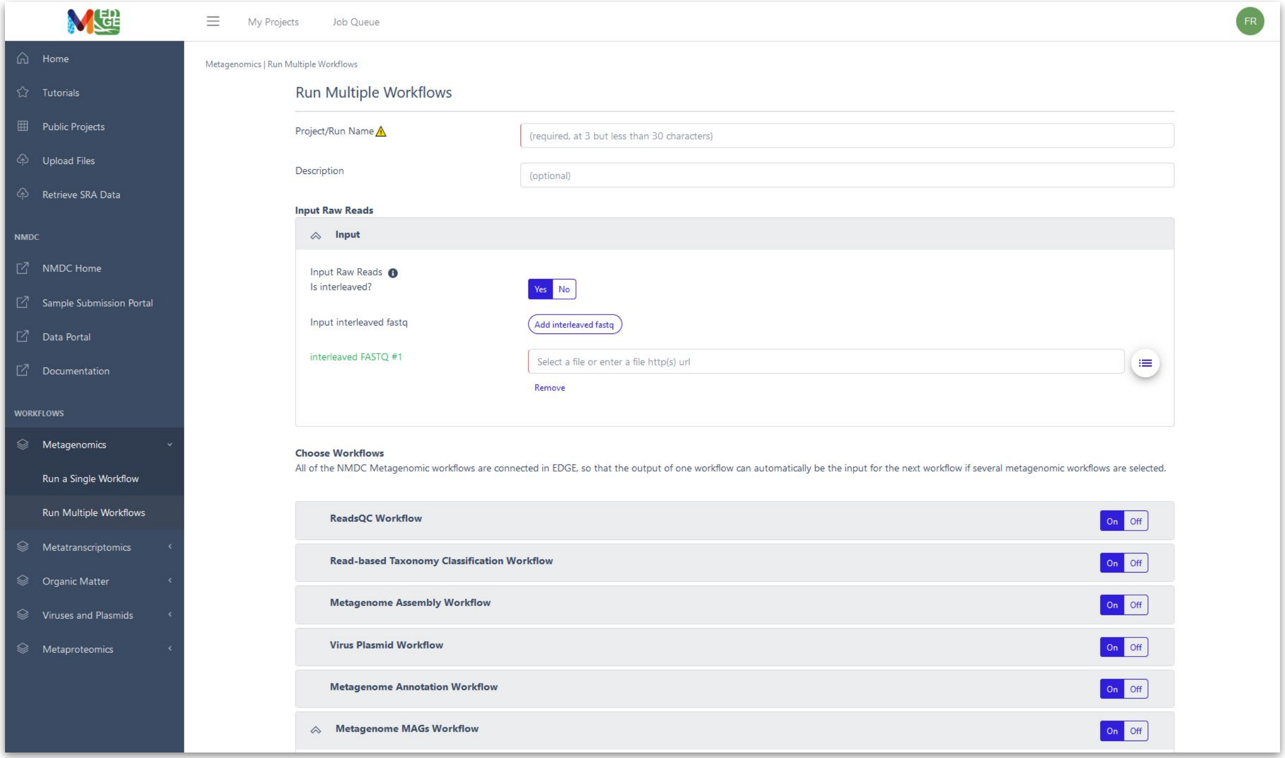
NASA

GeneLab

NMDC EDGE

The workflows can be run in a user-friendly interface: **https://nmdc-edge.org/home**

Designed for bioinformaticians of every level of expertise (including novices)



Metagenomics | Run Multiple Workflows

Run Multiple Workflows

Project/Run Name ⚠ (required, at 3 but less than 30 characters)

Description (optional)

Input Raw Reads

Input

Input Raw Reads
Is interleaved? [✓ Yes] No

Input interleaved fastq: Add interleaved fastq

interleaved FASTQ #1: Select a file or enter a file http(s) url

Remove

Choose Workflows

All of the NMDC Metagenomic workflows are connected in EDGE, so that the output of one workflow can automatically be the input for the next workflow if several metagenomic workflows are selected.

ReadsQC Workflow [✓ On] Off

Read-based Taxonomy Classification Workflow [✓ On] Off

Metagenome Assembly Workflow [✓ On] Off

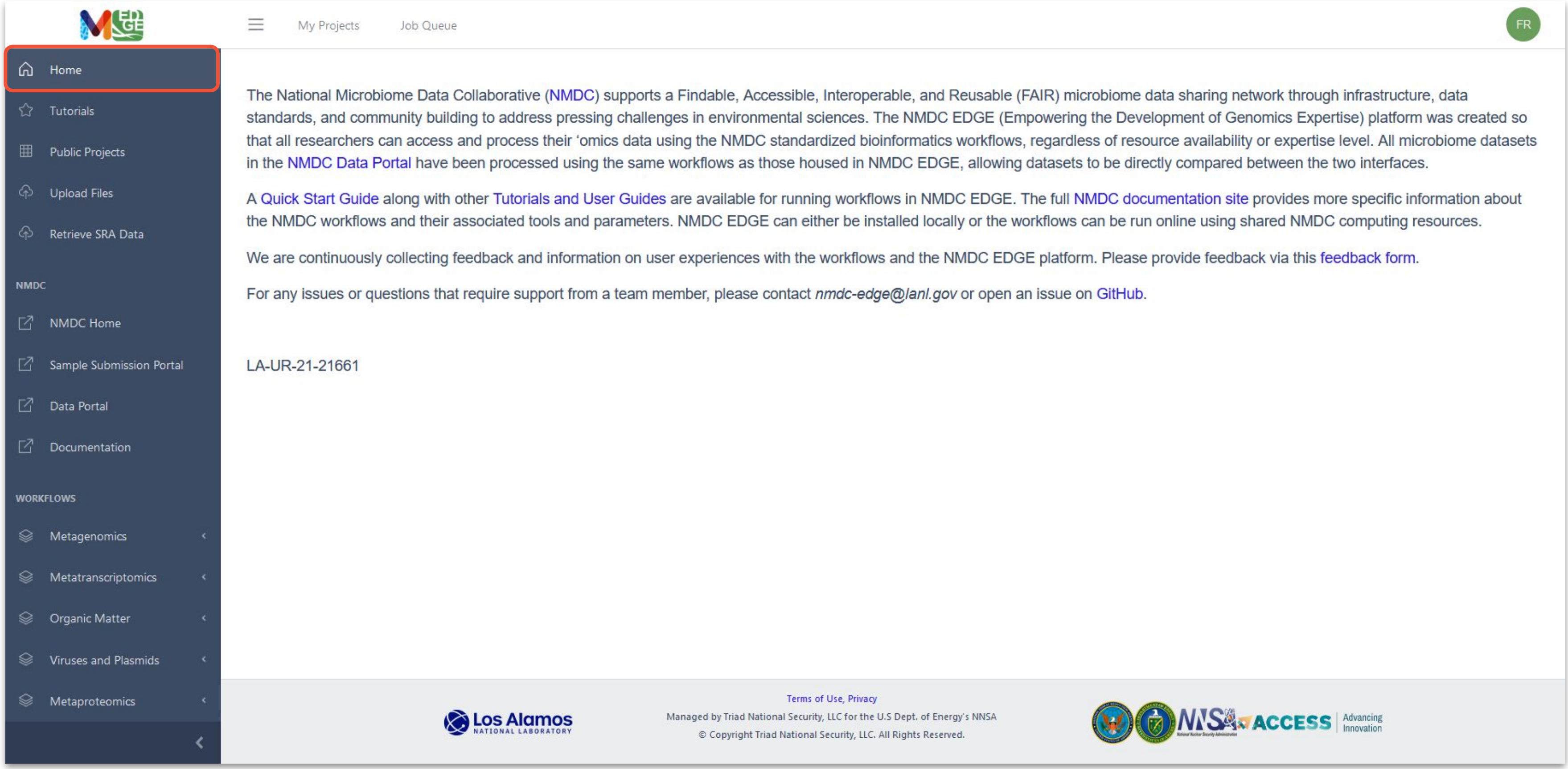
Virus Plasmid Workflow [✓ On] Off

Metagenome Annotation Workflow [✓ On] Off

Metagenome MAGs Workflow [✓ On] Off

NMDC EDGE Walkthrough

nmdc
National Microbiome Data Collaborative



The National Microbiome Data Collaborative (NMDC) supports a Findable, Accessible, Interoperable, and Reusable (FAIR) microbiome data sharing network through infrastructure, data standards, and community building to address pressing challenges in environmental sciences. The NMDC EDGE (Empowering the Development of Genomics Expertise) platform was created so that all researchers can access and process their 'omics data using the NMDC standardized bioinformatics workflows, regardless of resource availability or expertise level. All microbiome datasets in the NMDC Data Portal have been processed using the same workflows as those housed in NMDC EDGE, allowing datasets to be directly compared between the two interfaces.

A Quick Start Guide along with other Tutorials and User Guides are available for running workflows in NMDC EDGE. The full NMDC documentation site provides more specific information about the NMDC workflows and their associated tools and parameters. NMDC EDGE can either be installed locally or the workflows can be run online using shared NMDC computing resources.

We are continuously collecting feedback and information on user experiences with the workflows and the NMDC EDGE platform. Please provide feedback via this feedback form.

For any issues or questions that require support from a team member, please contact *nmdc-edge@lanl.gov* or open an issue on GitHub.

LA-UR-21-21661

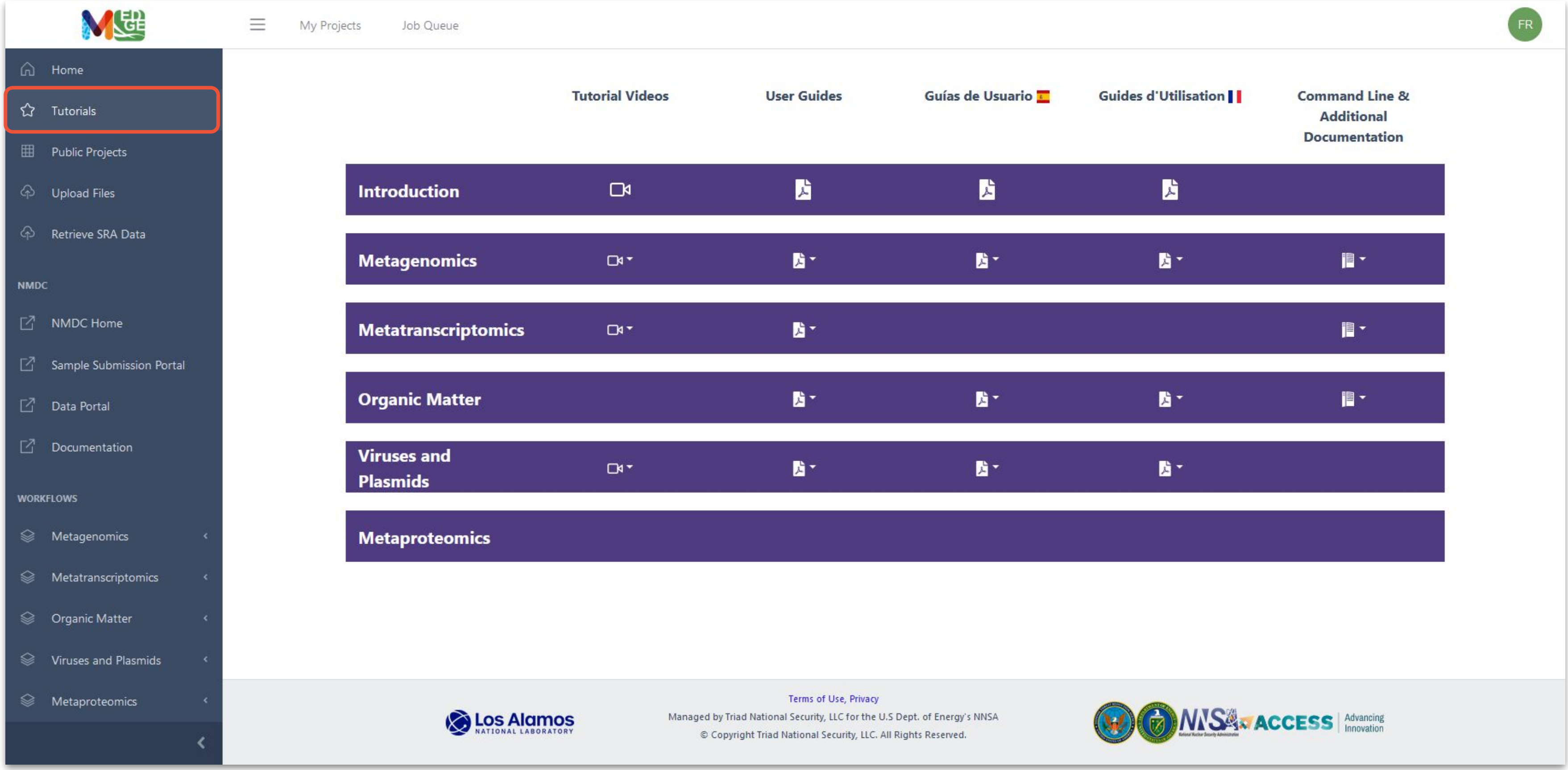
Terms of Use, Privacy

Managed by Triad National Security, LLC for the U.S Dept. of Energy's NNSA

© Copyright Triad National Security, LLC. All Rights Reserved.

Tutorials, User guides, Information

nmdc
National Microbiome Data Collaborative



	Tutorial Videos	User Guides	Guías de Usuario 🇪🇸	Guides d'Utilisation 🇫🇷	Command Line & Additional Documentation
Introduction	🎥	📄	📄	📄	
Metagenomics	🎥▾	📄▾	📄▾	📄▾	📑▾
Metatranscriptomics	🎥▾	📄▾			📑▾
Organic Matter		📄▾	📄▾	📄▾	📑▾
Viruses and Plasmids	🎥▾	📄▾	📄▾	📄▾	
Metaproteomics					

Los Alamos NATIONAL LABORATORY

Terms of Use, Privacy

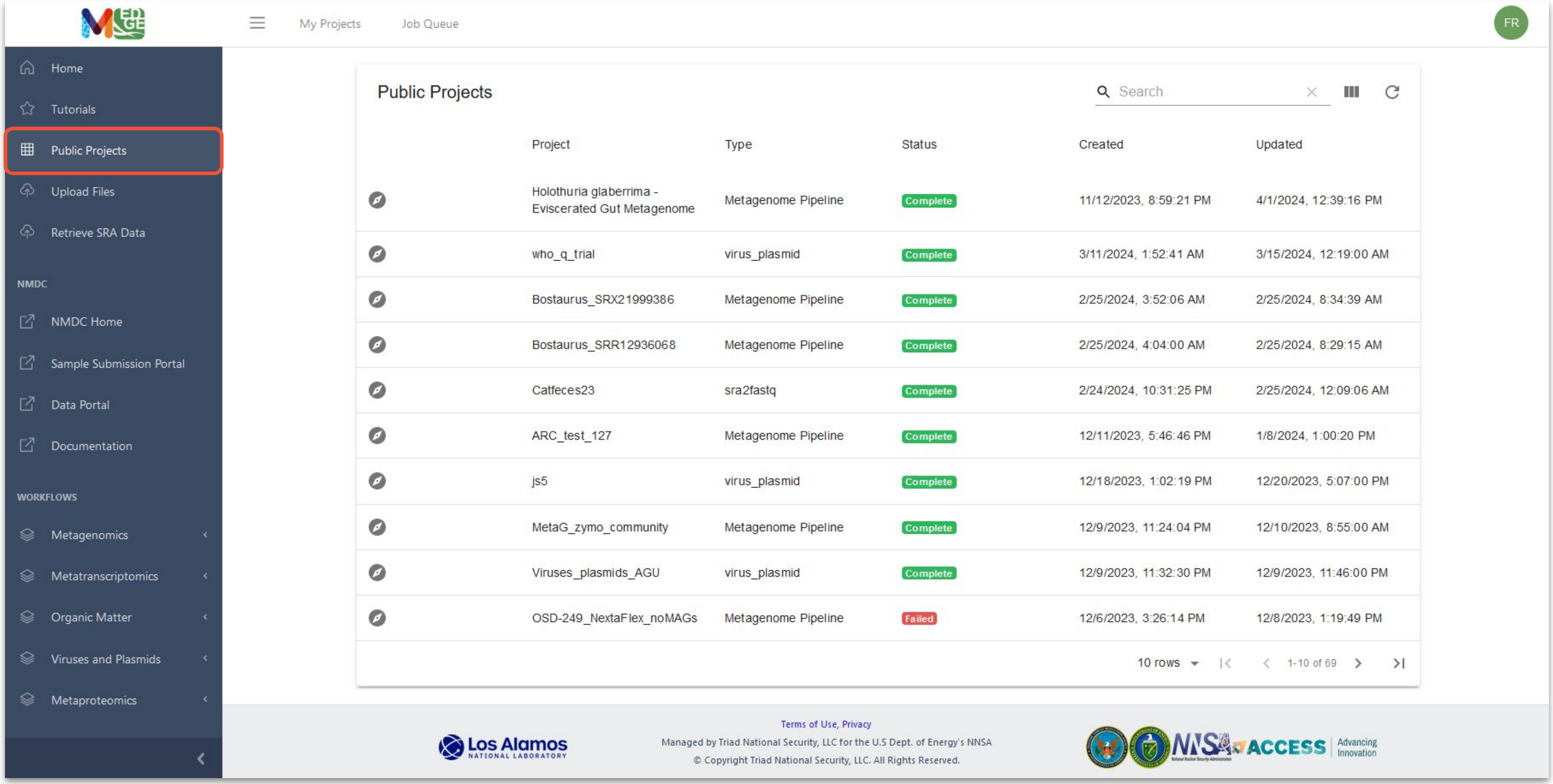
Managed by Triad National Security, LLC for the U.S Dept. of Energy's NNSA

© Copyright Triad National Security, LLC. All Rights Reserved.

NNSA ACCESS Advancing Innovation

Public Projects

nmdc
National Microbiome Data Collaborative



Public Projects

Project	Type	Status	Created	Updated
Holothuria glaberrima - Eviscerated Gut Metagenome	Metagenome Pipeline	Complete	11/12/2023, 8:59:21 PM	4/1/2024, 12:39:16 PM
who_q_trial	virus_plasmid	Complete	3/11/2024, 1:52:41 AM	3/15/2024, 12:19:00 AM
Bostaurus_SRX21999386	Metagenome Pipeline	Complete	2/25/2024, 3:52:06 AM	2/25/2024, 8:34:39 AM
Bostaurus_SRR12936068	Metagenome Pipeline	Complete	2/25/2024, 4:04:00 AM	2/25/2024, 8:29:15 AM
Catfeces23	sra2fastq	Complete	2/24/2024, 10:31:25 PM	2/25/2024, 12:09:06 AM
ARC_test_127	Metagenome Pipeline	Complete	12/11/2023, 5:46:46 PM	1/8/2024, 1:00:20 PM
js5	virus_plasmid	Complete	12/18/2023, 1:02:19 PM	12/20/2023, 5:07:00 PM
MetaG_zymo_community	Metagenome Pipeline	Complete	12/9/2023, 11:24:04 PM	12/10/2023, 8:55:00 AM
Viruses_plasmids_AGU	virus_plasmid	Complete	12/9/2023, 11:32:30 PM	12/9/2023, 11:46:00 PM
OSD-249_NextaFlex_noMAGs	Metagenome Pipeline	Failed	12/6/2023, 3:26:14 PM	12/8/2023, 1:19:49 PM

10 rows 1-10 of 69

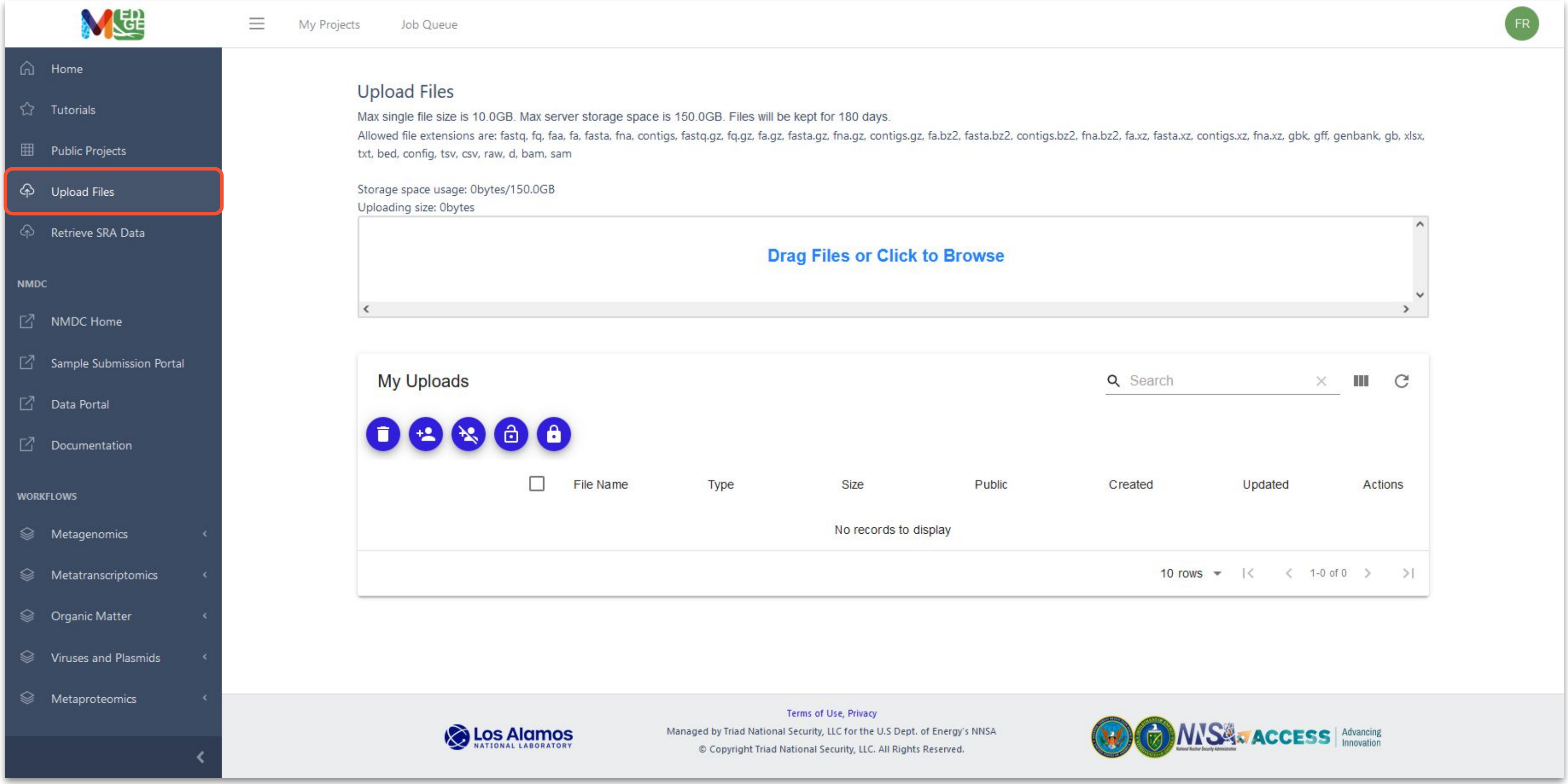
Los Alamos NATIONAL LABORATORY

Terms of Use, Privacy
Managed by Triad National Security, LLC for the U.S Dept. of Energy's NNSA
© Copyright Triad National Security, LLC. All Rights Reserved.

NNSA ACCESS Advancing Innovation

Upload Files

nmdc
National Microbiome Data Collaborative



Upload Files

Max single file size is 10.0GB. Max server storage space is 150.0GB. Files will be kept for 180 days.

Allowed file extensions are: fastq, fq, faa, fa, fasta, fna, contigs, fastq.gz, fq.gz, fa.gz, fasta.gz, fna.gz, contigs.gz, fa.bz2, fasta.bz2, contigs.bz2, fna.bz2, fa.xz, fasta.xz, contigs.xz, fna.xz, gbk, gff, genbank, gb, xlsx, txt, bed, config, tsv, csv, raw, d, bam, sam

Storage space usage: 0bytes/150.0GB
Uploading size: 0bytes

Drag Files or Click to Browse

My Uploads

Search

□	File Name	Type	Size	Public	Created	Updated	Actions

No records to display

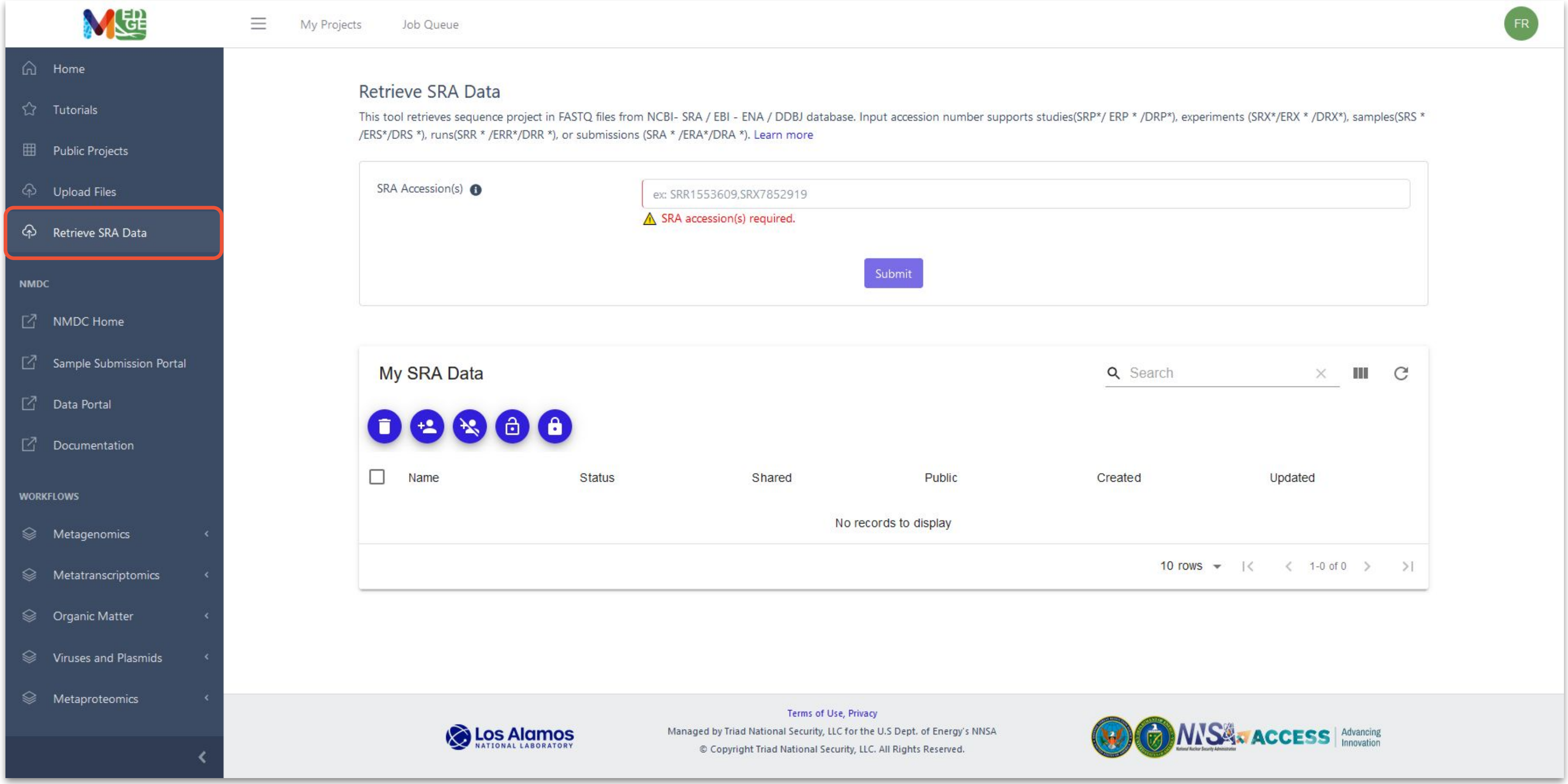
10 rows 1-0 of 0

Los Alamos NATIONAL LABORATORY

Terms of Use, Privacy
Managed by Triad National Security, LLC for the U.S Dept. of Energy's NNSA
© Copyright Triad National Security, LLC. All Rights Reserved.

NNSA ACCESS Advancing Innovation

Retrieve SRA Data



Retrieve SRA Data

This tool retrieves sequence project in FASTQ files from NCBI- SRA / EBI - ENA / DDBJ database. Input accession number supports studies(SRP*/ ERP * /DRP*), experiments (SRX*/ERX * /DRX*), samples(SRS * /ERS*/DRS *), runs(SRR * /ERR*/DRR *), or submissions (SRA * /ERA*/DRA *). Learn more

SRA Accession(s): ex: SRR1553609,SRX7852919

⚠ SRA accession(s) required.

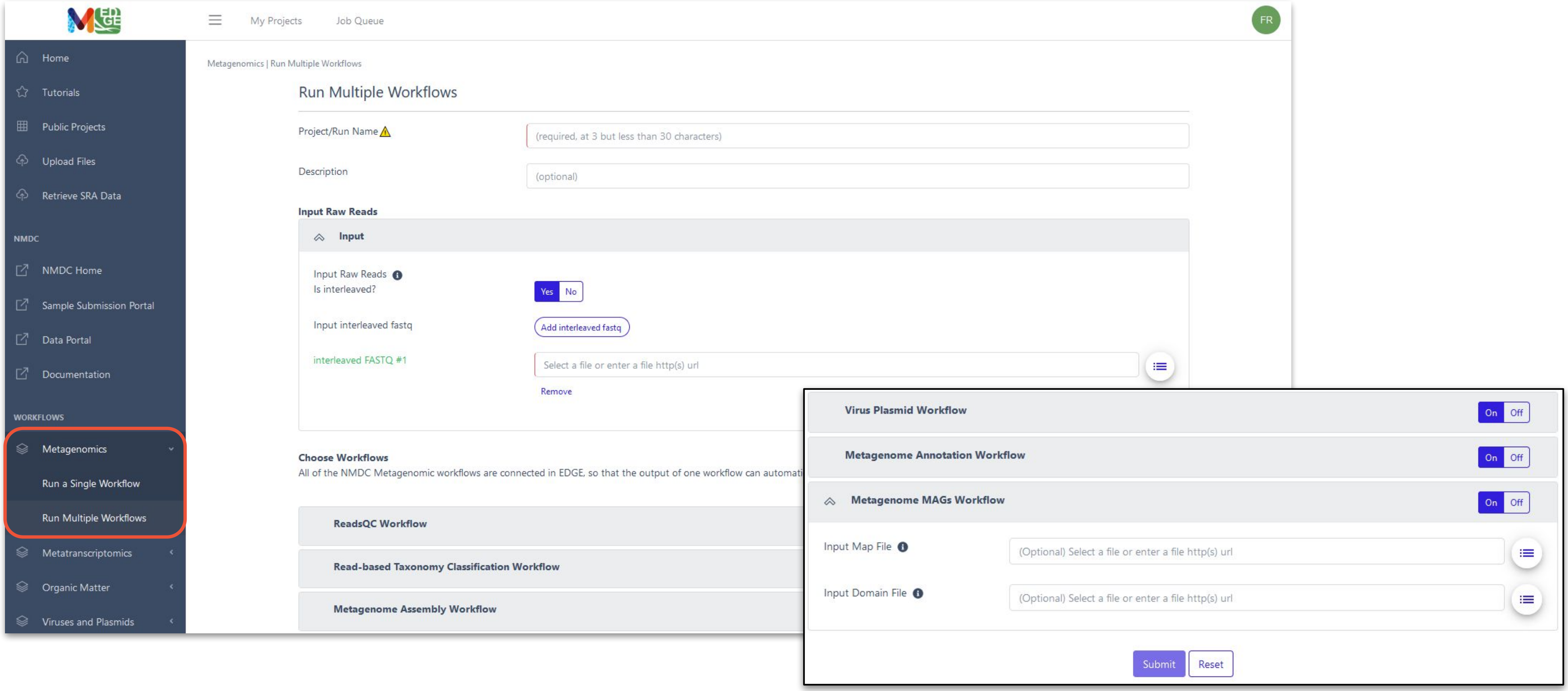
My SRA Data

□	Name	Status	Shared	Public	Created	Updated

No records to display

Terms of Use, Privacy
Managed by Triad National Security, LLC for the U.S Dept. of Energy's NNSA
© Copyright Triad National Security, LLC. All Rights Reserved.

Run Metagenome Pipeline



Metagenomics | Run Multiple Workflows

Run Multiple Workflows

Project/Run Name ⚠ (required, at 3 but less than 30 characters)

Description (optional)

Input Raw Reads

Input

Input Raw Reads
Is interleaved? [✓ Yes] No

Input interleaved fastq: Add interleaved fastq

interleaved FASTQ #1: Select a file or enter a file http(s) url

Remove

Choose Workflows

All of the NMDC Metagenomic workflows are connected in EDGE, so that the output of one workflow can automati

ReadsQC Workflow

Read-based Taxonomy Classification Workflow

Metagenome Assembly Workflow

Virus Plasmid Workflow [✓ On] Off

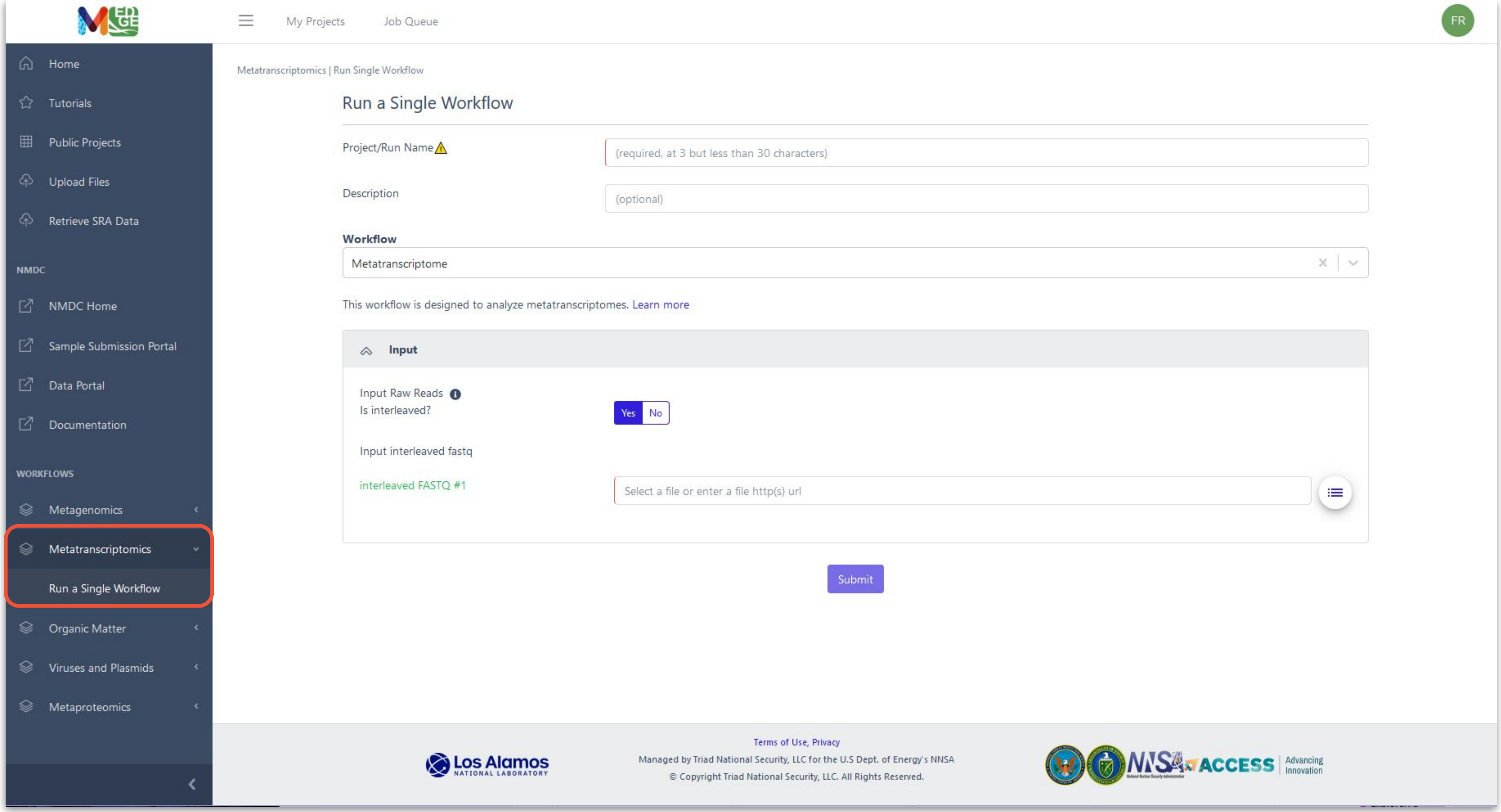
Metagenome Annotation Workflow [✓ On] Off

Metagenome MAGs Workflow [✓ On] Off

Input Map File: (Optional) Select a file or enter a file http(s) url

Input Domain File: (Optional) Select a file or enter a file http(s) url

Metatranscriptomics Workflow



Metatranscriptomics | Run Single Workflow

Run a Single Workflow

Project/Run Name: (required, at 3 but less than 30 characters)

Description: (optional)

Workflow: Metatranscriptome

This workflow is designed to analyze metatranscriptomes. Learn more

Input

Input Raw Reads
Is interleaved? [✓ Yes] No

Input interleaved fastq

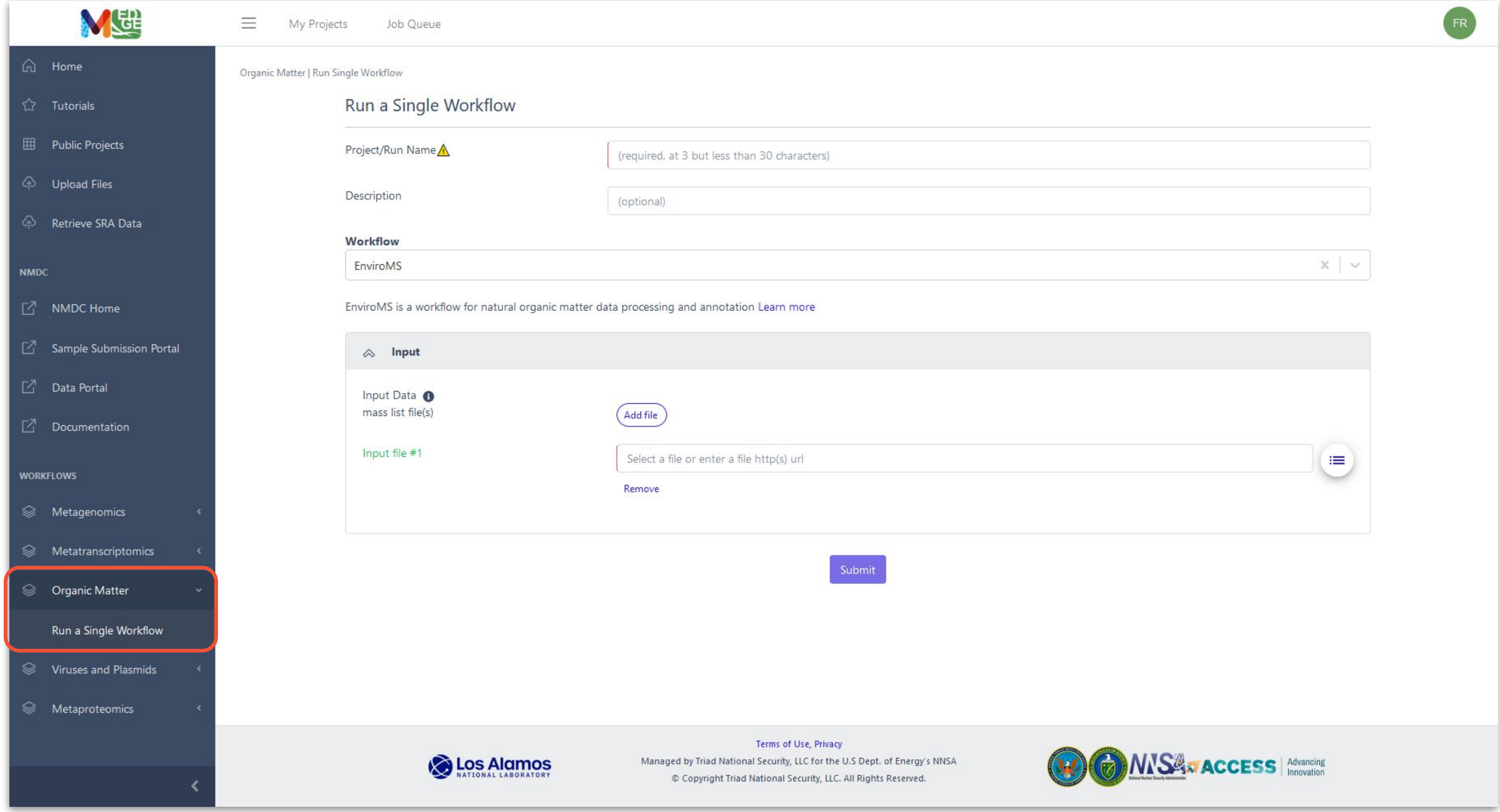
interleaved FASTQ #1: Select a file or enter a file http(s) url

Submit

Terms of Use, Privacy
Managed by Triad National Security, LLC for the U.S Dept. of Energy's NNSA
© Copyright Triad National Security, LLC. All Rights Reserved.

Natural Organic Matter Workflow

nmdc
National Microbiome Data Collaborative



Organic Matter | Run Single Workflow

Run a Single Workflow

Project/Run Name: (required, at 3 but less than 30 characters)

Description: (optional)

Workflow: EnviroMS

EnviroMS is a workflow for natural organic matter data processing and annotation Learn more

Input

Input Data mass list file(s)

Input file #1: Select a file or enter a file http(s) url

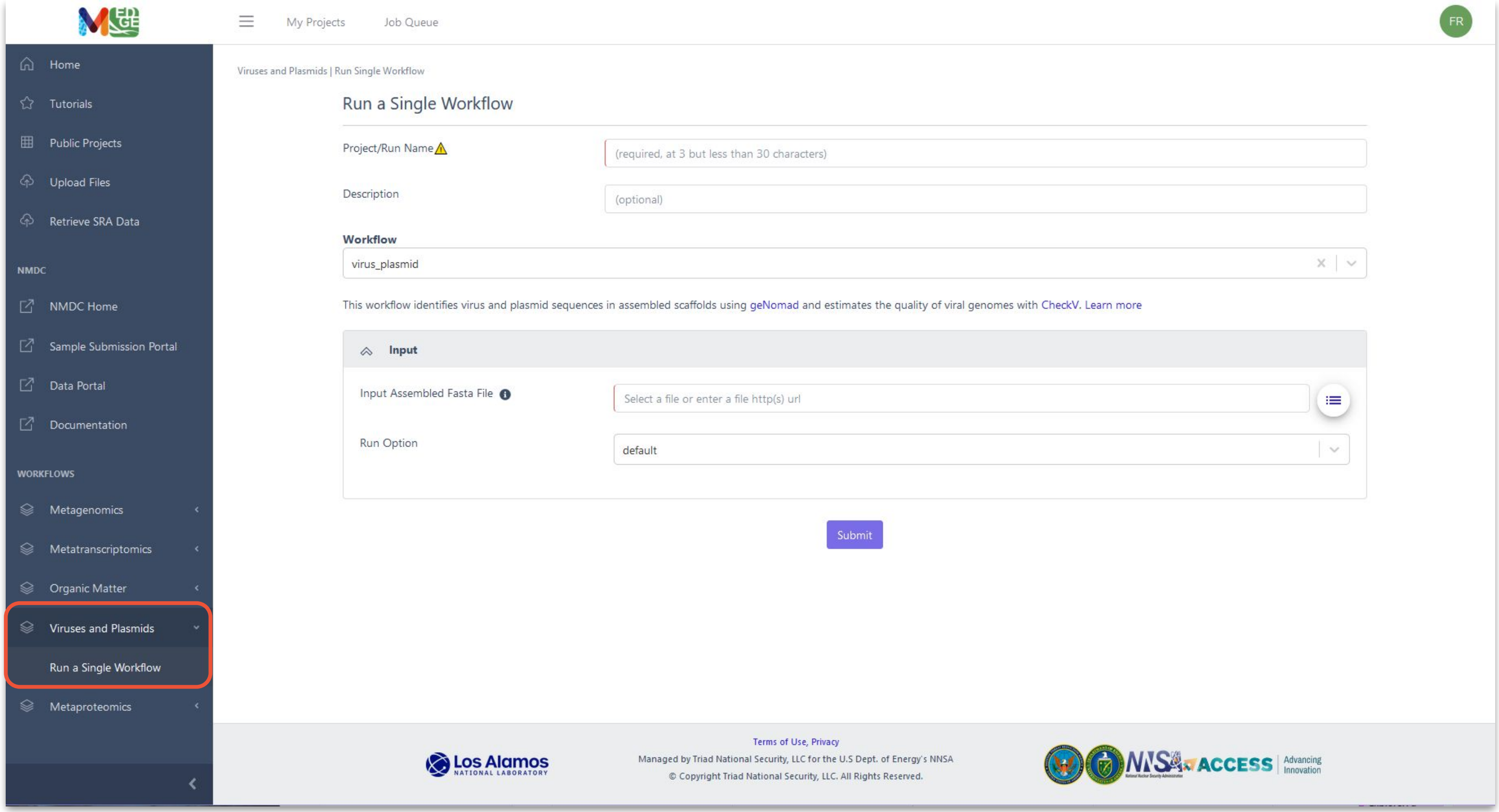
Terms of Use, Privacy

Managed by Triad National Security, LLC for the U.S Dept. of Energy's NNSA

Viruses & Plasmids Workflow

nmdc

National Microbiome Data Collaborative



Viruses and Plasmids | Run Single Workflow

Run a Single Workflow

Project/Run Name ⚠ (required, at 3 but less than 30 characters)

Description (optional)

Workflow

virus_plasmid

This workflow identifies virus and plasmid sequences in assembled scaffolds using geNomad and estimates the quality of viral genomes with CheckV. Learn more

Input

Input Assembled Fasta File: Select a file or enter a file http(s) url

Run Option: default

Submit

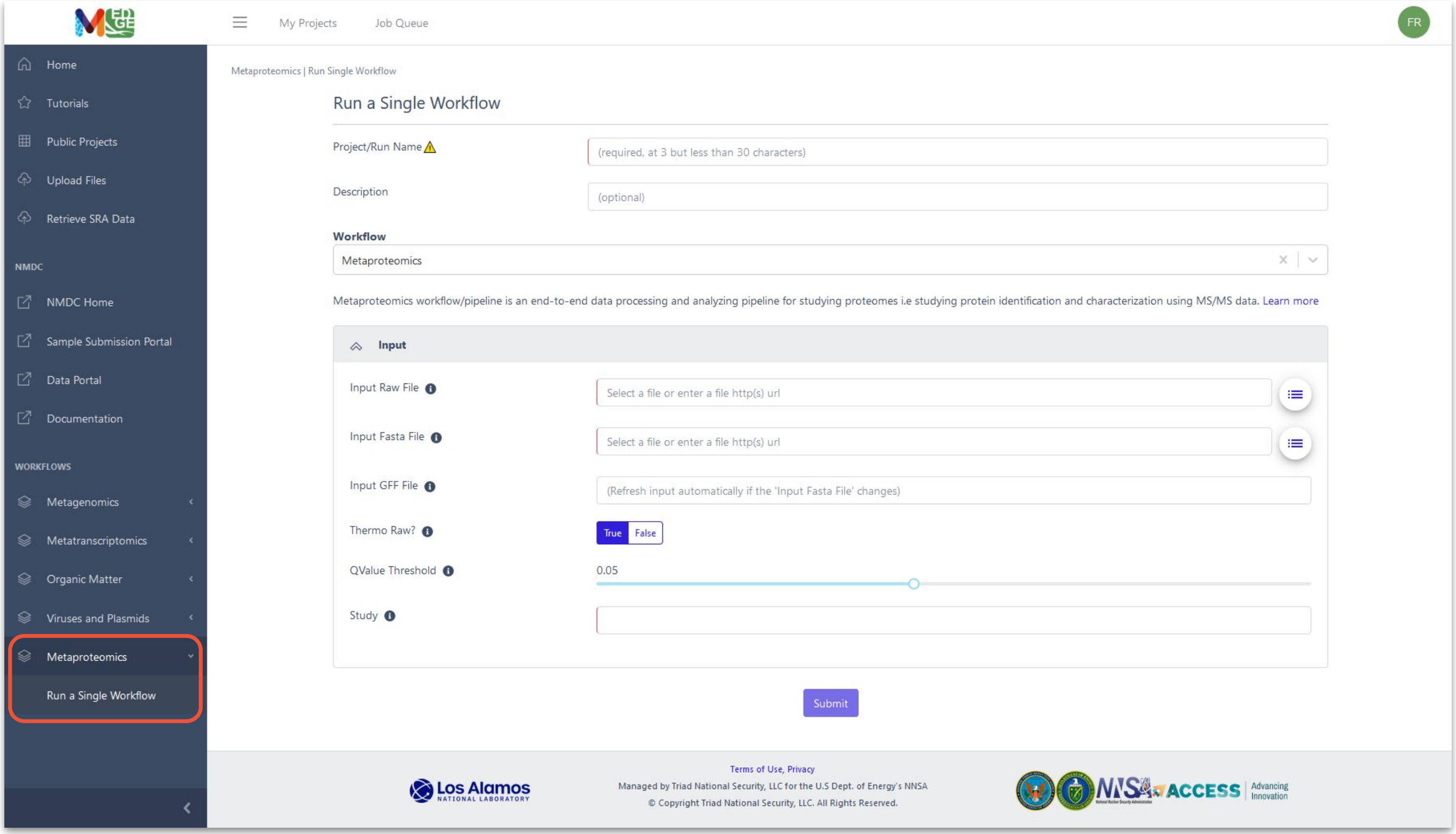
Terms of Use, Privacy

Managed by Triad National Security, LLC for the U.S Dept. of Energy's NNSA

© Copyright Triad National Security, LLC. All Rights Reserved.

Metaproteomics Workflow

nmdc
National Microbiome Data Collaborative



Metaproteomics | Run Single Workflow

Run a Single Workflow

Project/Run Name ⚠ (required, at 3 but less than 30 characters)

Description (optional)

Workflow

Metaproteomics

Metaproteomics workflow/pipeline is an end-to-end data processing and analyzing pipeline for studying proteomes i.e studying protein identification and characterization using MS/MS data. Learn more

Input

Input Raw File Select a file or enter a file http(s) url

Input Fasta File Select a file or enter a file http(s) url

Input GFF File (Refresh input automatically if the 'Input Fasta File' changes)

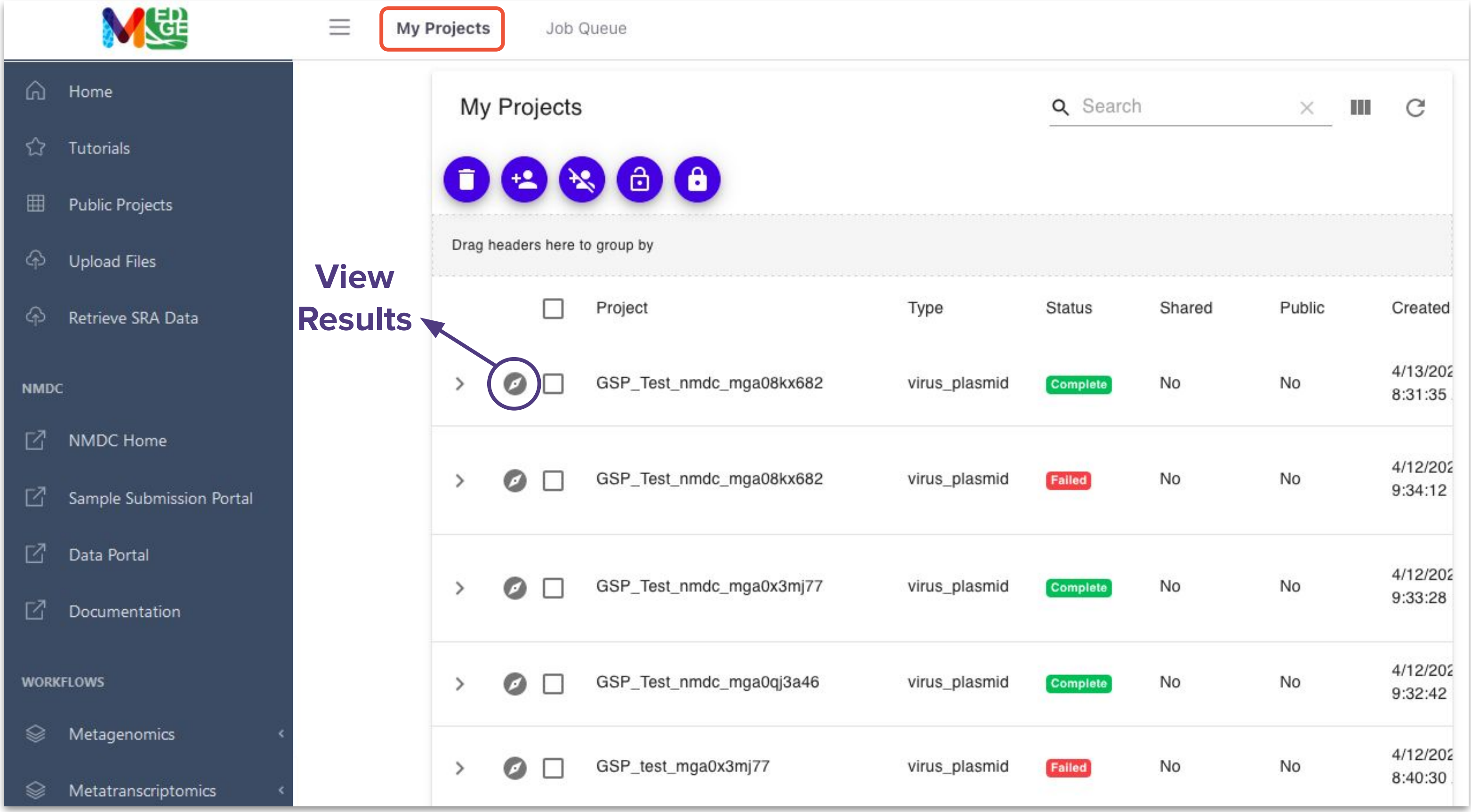
Thermo Raw? [✓ True] False

QValue Threshold 0.05

Study

Terms of Use, Privacy
Managed by Triad National Security, LLC for the U.S Dept. of Energy's NNSA
© Copyright Triad National Security, LLC. All Rights Reserved.

My Projects

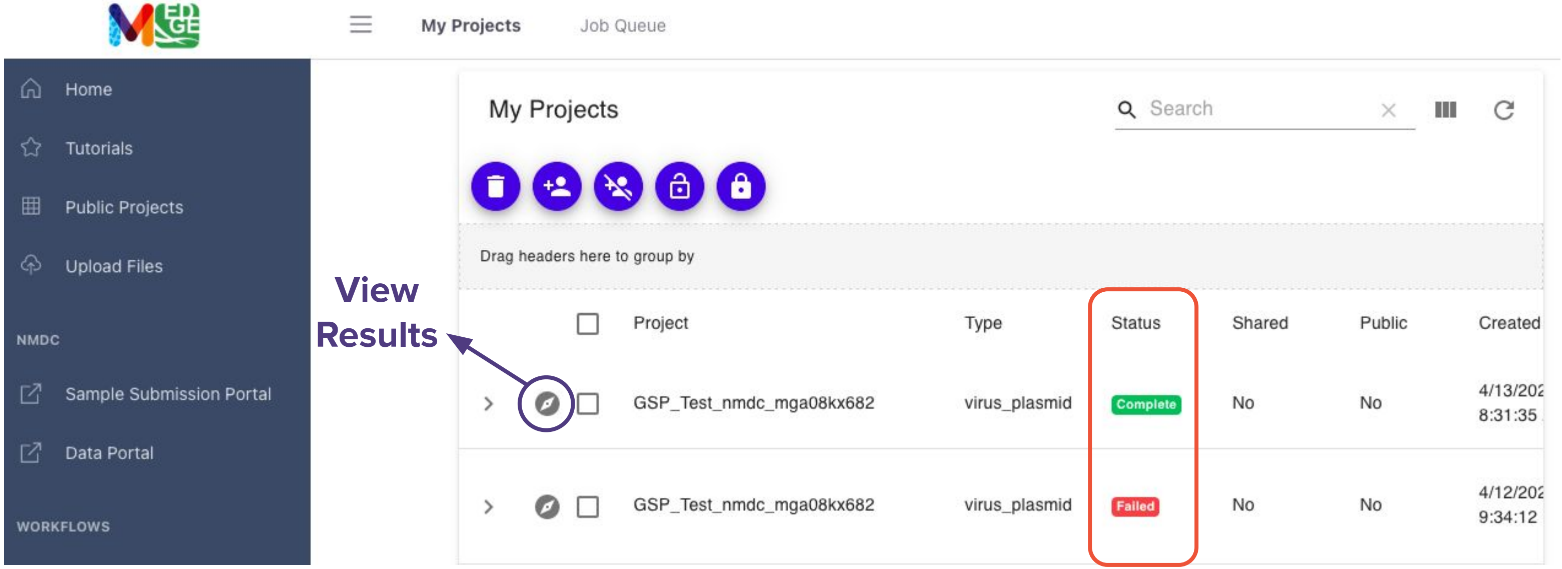


View Results

My Projects

			Project	Type	Status	Shared	Public	Created
›	◎	□	GSP_Test_nmdc_mga08kx682	virus_plasmid	Complete	No	No	4/13/202 8:31:35
›	◎	□	GSP_Test_nmdc_mga08kx682	virus_plasmid	Failed	No	No	4/12/202 9:34:12
›	◎	□	GSP_Test_nmdc_mga0x3mj77	virus_plasmid	Complete	No	No	4/12/202 9:33:28
›	◎	□	GSP_Test_nmdc_mga0qj3a46	virus_plasmid	Complete	No	No	4/12/202 9:32:42
›	◎	□	GSP_test_mga0x3mj77	virus_plasmid	Failed	No	No	4/12/202 8:40:30

My Projects: Possible Statuses



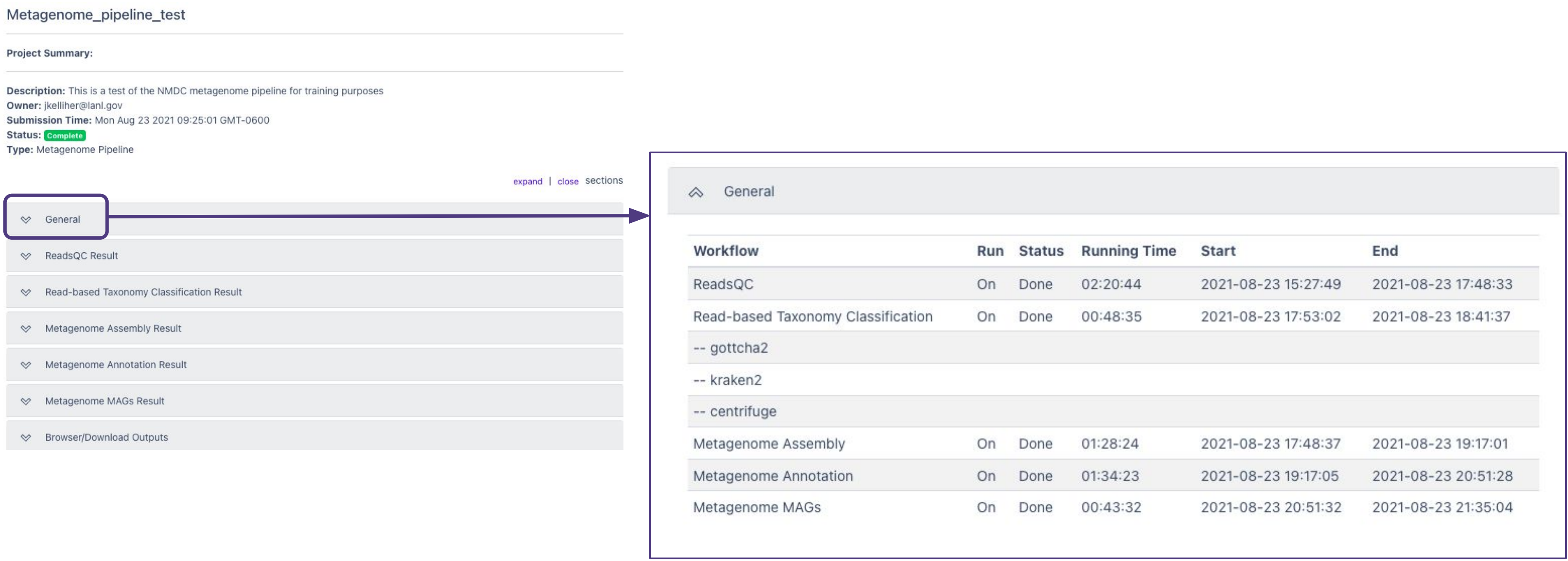
My Projects

View Results

	Project	Type	Status	Shared	Public	Created
□	GSP_Test_nmdc_mga08kx682	virus_plasmid	Complete	No	No	4/13/202 8:31:35
□	GSP_Test_nmdc_mga08kx682	virus_plasmid	Failed	No	No	4/12/202 9:34:12

Submitted In queue Running Complete Failed

Example Results: Metagenome



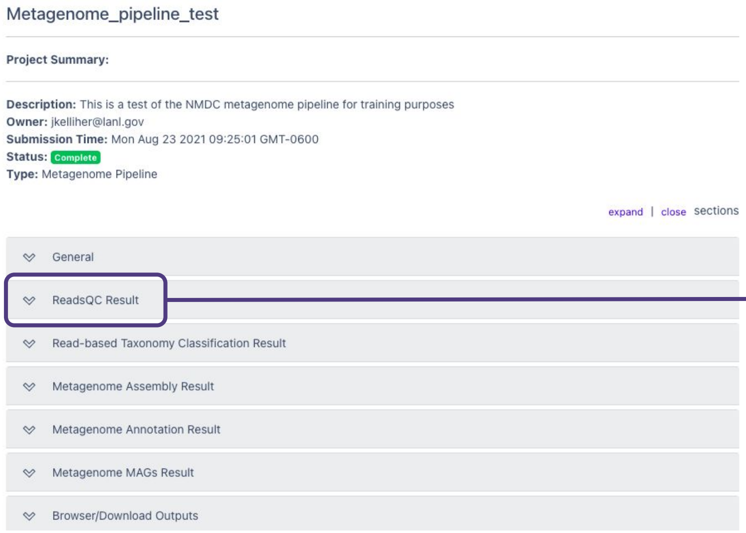
Metagenome_pipeline_test

Project Summary:

Description: This is a test of the NMDC metagenome pipeline for training purposes
Owner: jkelliher@lanl.gov
Submission Time: Mon Aug 23 2021 09:25:01 GMT-0600
Status: Complete
Type: Metagenome Pipeline

Workflow	Run	Status	Running Time	Start	End
ReadsQC	On	Done	02:20:44	2021-08-23 15:27:49	2021-08-23 17:48:33
Read-based Taxonomy Classification	On	Done	00:48:35	2021-08-23 17:53:02	2021-08-23 18:41:37
-- gottcha2					
-- kraken2					
-- centrifuge					
Metagenome Assembly	On	Done	01:28:24	2021-08-23 17:48:37	2021-08-23 19:17:01
Metagenome Annotation	On	Done	01:34:23	2021-08-23 19:17:05	2021-08-23 20:51:28
Metagenome MAGs	On	Done	00:43:32	2021-08-23 20:51:32	2021-08-23 21:35:04

Example Results: Metagenome

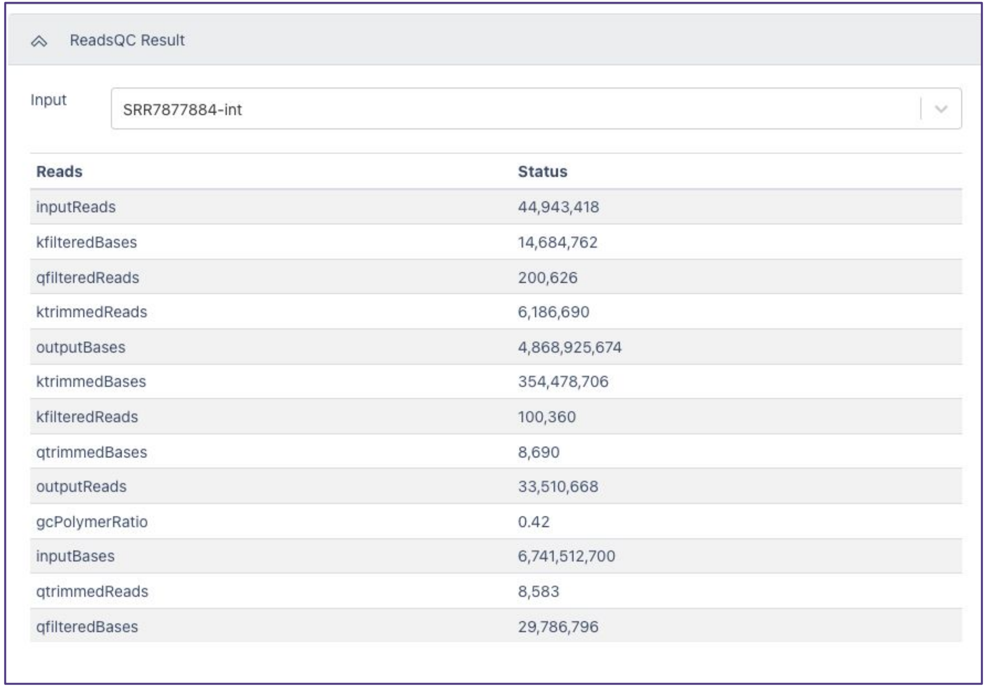


Metagenome_pipeline_test

Project Summary:

Description: This is a test of the NMDC metagenome pipeline for training purposes
Owner: jkelliher@lanl.gov
Submission Time: Mon Aug 23 2021 09:25:01 GMT-0600
Status: Complete
Type: Metagenome Pipeline

expand | close sections
General
ReadsQC Result
Read-based Taxonomy Classification Result
Metagenome Assembly Result
Metagenome Annotation Result
Metagenome MAGs Result
Browser/Download Outputs



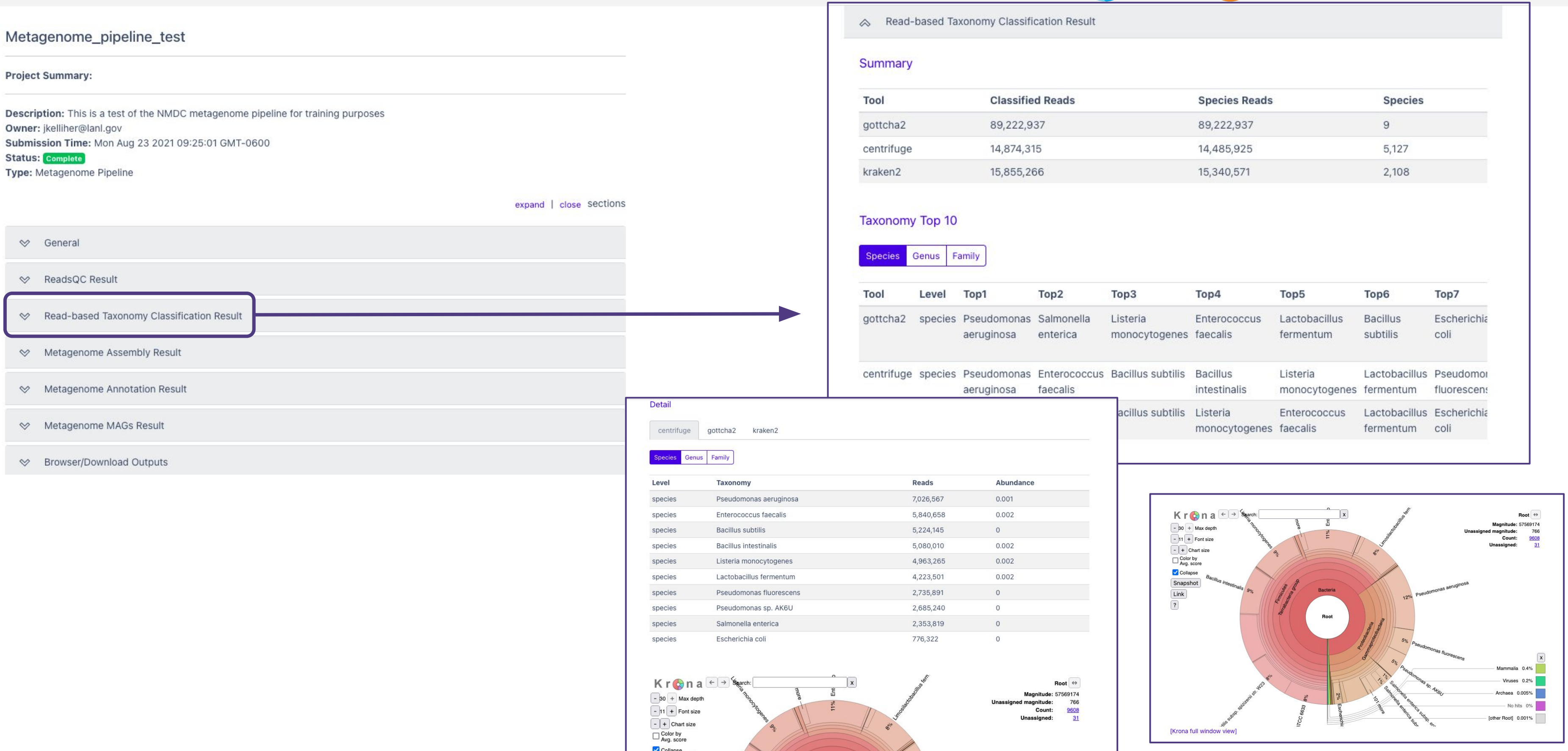
Input: SRR7877884-int

Reads	Status
inputReads	44,943,418
kfilteredBases	14,684,762
qfilteredReads	200,626
ktrimmedReads	6,186,690
outputBases	4,868,925,674
ktrimmedBases	354,478,706
kfilteredReads	100,360
qtrimmedBases	8,690
outputReads	33,510,668
gcPolymerRatio	0.42
inputBases	6,741,512,700
qtrimmedReads	8,583
qfilteredBases	29,786,796

Example Results: Metagenome

nmdc

National Microbiome Data Collaborative



Metagenome_pipeline_test

Project Summary:

Description: This is a test of the NMDC metagenome pipeline for training purposes
Owner: jkelliher@lanl.gov
Submission Time: Mon Aug 23 2021 09:25:01 GMT-0600
Status: Complete
Type: Metagenome Pipeline

expand | close sections

- General
- ReadsQC Result
- Read-based Taxonomy Classification Result
- Metagenome Assembly Result
- Metagenome Annotation Result
- Metagenome MAGs Result
- Browser/Download Outputs

Read-based Taxonomy Classification Result

Summary

Tool	Classified Reads	Species Reads	Species
gottcha2	89,222,937	89,222,937	9
centrifuge	14,874,315	14,485,925	5,127
kraken2	15,855,266	15,340,571	2,108

Taxonomy Top 10

Species | Genus | Family

Tool	Level	Top1	Top2	Top3	Top4	Top5	Top6	Top7
gottcha2	species	Pseudomonas aeruginosa	Salmonella enterica	Listeria monocytogenes	Enterococcus faecalis	Lactobacillus fermentum	Bacillus subtilis	Escherichia coli
centrifuge	species	Pseudomonas aeruginosa	Enterococcus faecalis	Bacillus subtilis	Bacillus intestinalis	Listeria monocytogenes	Lactobacillus fermentum	Pseudomonas fluorescens
				acillus subtilis	Listeria monocytogenes	Enterococcus faecalis	Lactobacillus fermentum	Escherichia coli

Detail

centrifuge | gottcha2 | kraken2

Species | Genus | Family

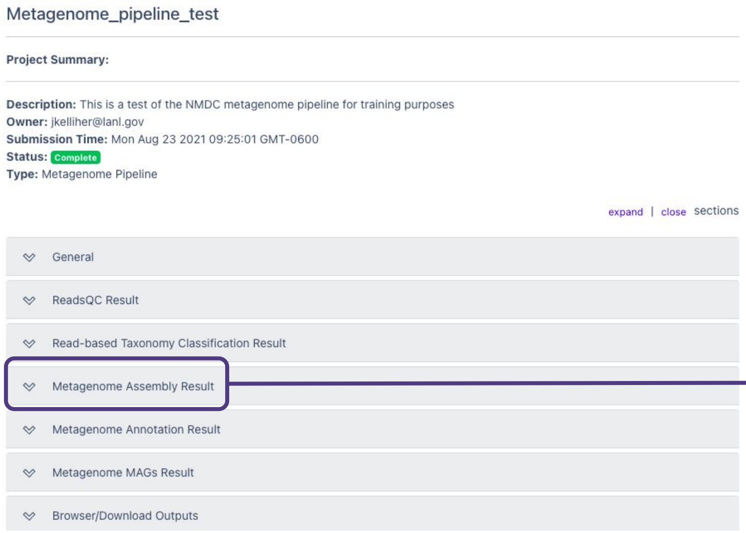
Level	Taxonomy	Reads	Abundance
species	Pseudomonas aeruginosa	7,026,567	0.001
species	Enterococcus faecalis	5,840,658	0.002
species	Bacillus subtilis	5,224,145	0
species	Bacillus intestinalis	5,080,010	0.002
species	Listeria monocytogenes	4,963,265	0.002
species	Lactobacillus fermentum	4,223,501	0.002
species	Pseudomonas fluorescens	2,735,891	0
species	Pseudomonas sp. AK6U	2,685,240	0
species	Salmonella enterica	2,353,819	0
species	Escherichia coli	776,322	0

Root

Magnitude: 57569174
Unassigned magnitude: 766
Count: 9608
Unassigned: 31

Example Results: Metagenome

nmdc
National Microbiome Data Collaborative

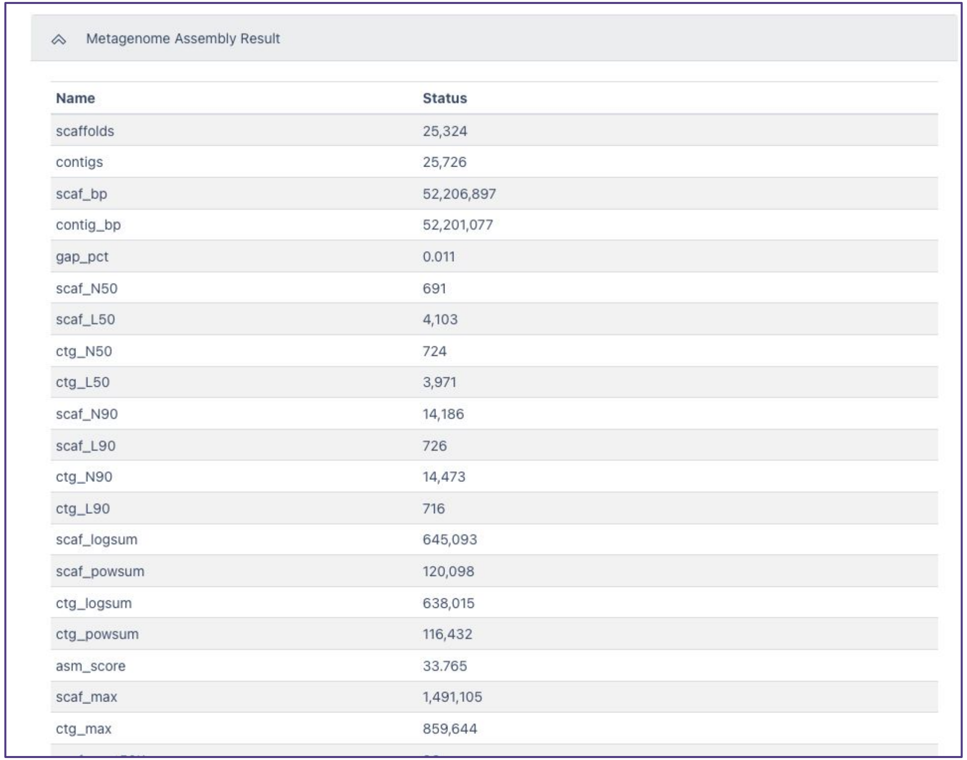


Metagenome_pipeline_test

Project Summary:

Description: This is a test of the NMDC metagenome pipeline for training purposes
Owner: jkelliher@lanl.gov
Submission Time: Mon Aug 23 2021 09:25:01 GMT-0600
Status: Complete
Type: Metagenome Pipeline

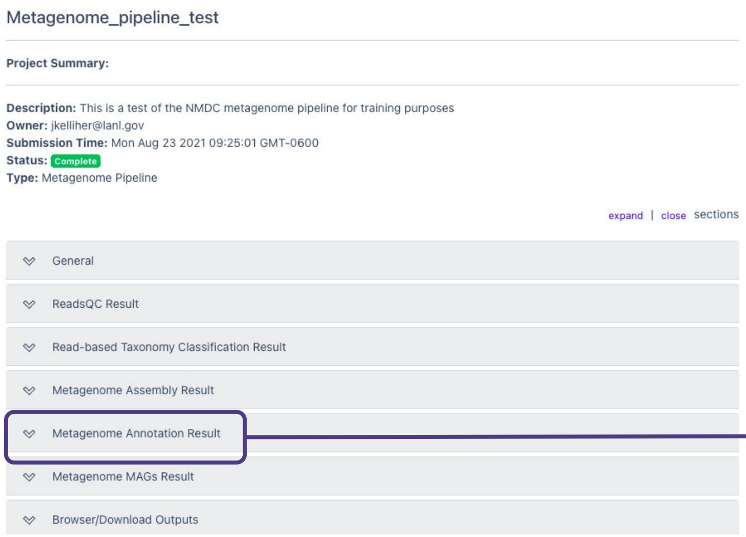
expand | close sections
General
ReadsQC Result
Read-based Taxonomy Classification Result
Metagenome Assembly Result
Metagenome Annotation Result
Metagenome MAGs Result
Browser/Download Outputs



Metagenome Assembly Result

Name	Status
scaffolds	25,324
contigs	25,726
scaf_bp	52,206,897
contig_bp	52,201,077
gap_pct	0.011
scaf_N50	691
scaf_L50	4,103
ctg_N50	724
ctg_L50	3,971
scaf_N90	14,186
scaf_L90	726
ctg_N90	14,473
ctg_L90	716
scaf_logsum	645,093
scaf_powsum	120,098
ctg_logsum	638,015
ctg_powsum	116,432
asm_score	33.765
scaf_max	1,491,105
ctg_max	859,644

Example Results: Metagenome



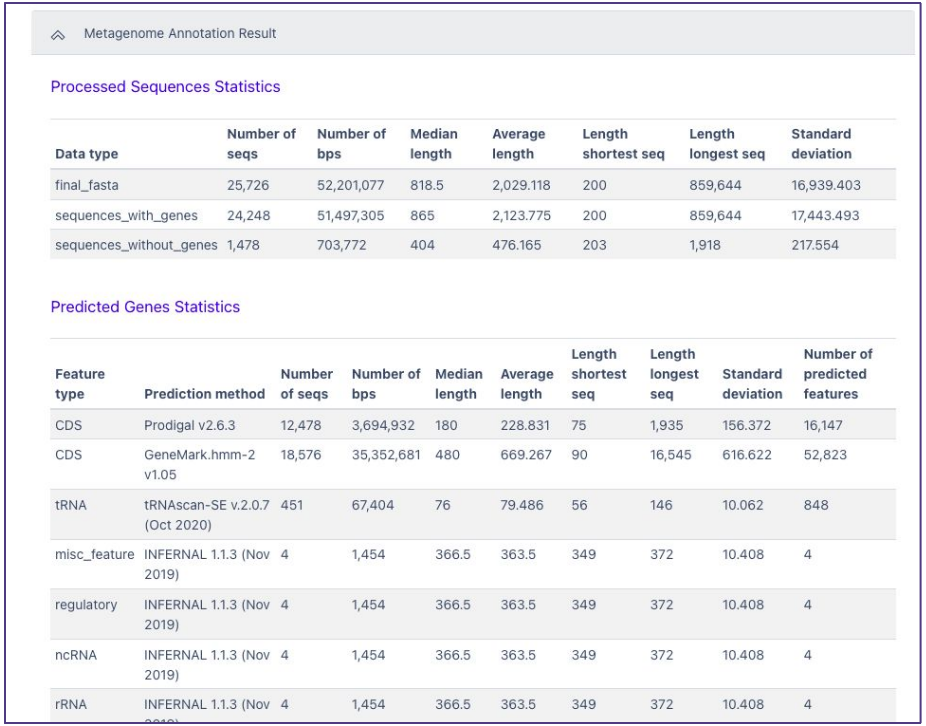
Metagenome_pipeline_test

Project Summary:

Description: This is a test of the NMDC metagenome pipeline for training purposes
Owner: jkelliher@lanl.gov
Submission Time: Mon Aug 23 2021 09:25:01 GMT-0600
Status: Complete
Type: Metagenome Pipeline

expand | close sections

- General
- ReadsQC Result
- Read-based Taxonomy Classification Result
- Metagenome Assembly Result
- Metagenome Annotation Result
- Metagenome MAGs Result
- Browser/Download Outputs



Metagenome Annotation Result

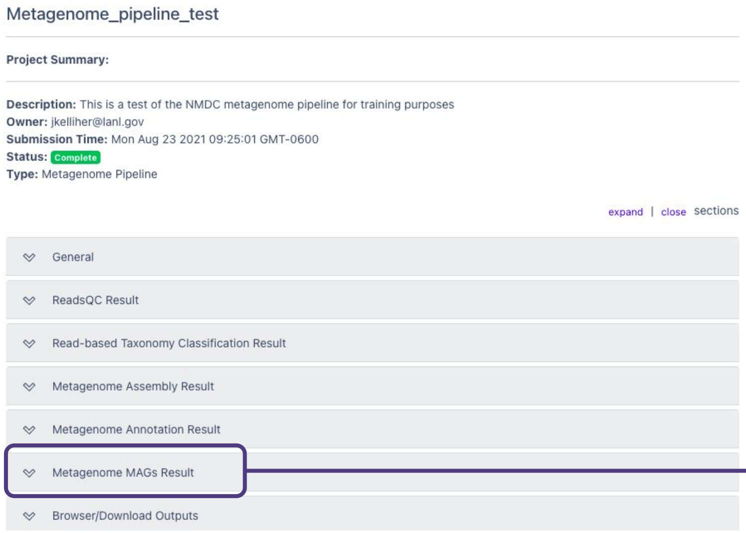
Processed Sequences Statistics

Data type	Number of seqs	Number of bps	Median length	Average length	Length shortest seq	Length longest seq	Standard deviation
final_fasta	25,726	52,201,077	818.5	2,029.118	200	859,644	16,939.403
sequences_with_genes	24,248	51,497,305	865	2,123.775	200	859,644	17,443.493
sequences_without_genes	1,478	703,772	404	476.165	203	1,918	217.554

Predicted Genes Statistics

Feature type	Prediction method	Number of seqs	Number of bps	Median length	Average length	Length shortest seq	Length longest seq	Standard deviation	Number of predicted features
CDS	Prodigal v2.6.3	12,478	3,694,932	180	228.831	75	1,935	156.372	16,147
CDS	GeneMark.hmm-2 v1.05	18,576	35,352,681	480	669.267	90	16,545	616.622	52,823
tRNA	tRNAscan-SE v.2.0.7 (Oct 2020)	451	67,404	76	79.486	56	146	10.062	848
misc_feature	INFERNAL 1.1.3 (Nov 2019)	4	1,454	366.5	363.5	349	372	10.408	4
regulatory	INFERNAL 1.1.3 (Nov 2019)	4	1,454	366.5	363.5	349	372	10.408	4
ncRNA	INFERNAL 1.1.3 (Nov 2019)	4	1,454	366.5	363.5	349	372	10.408	4
rRNA	INFERNAL 1.1.3 (Nov	4	1,454	366.5	363.5	349	372	10.408	4

Example Results: Metagenome

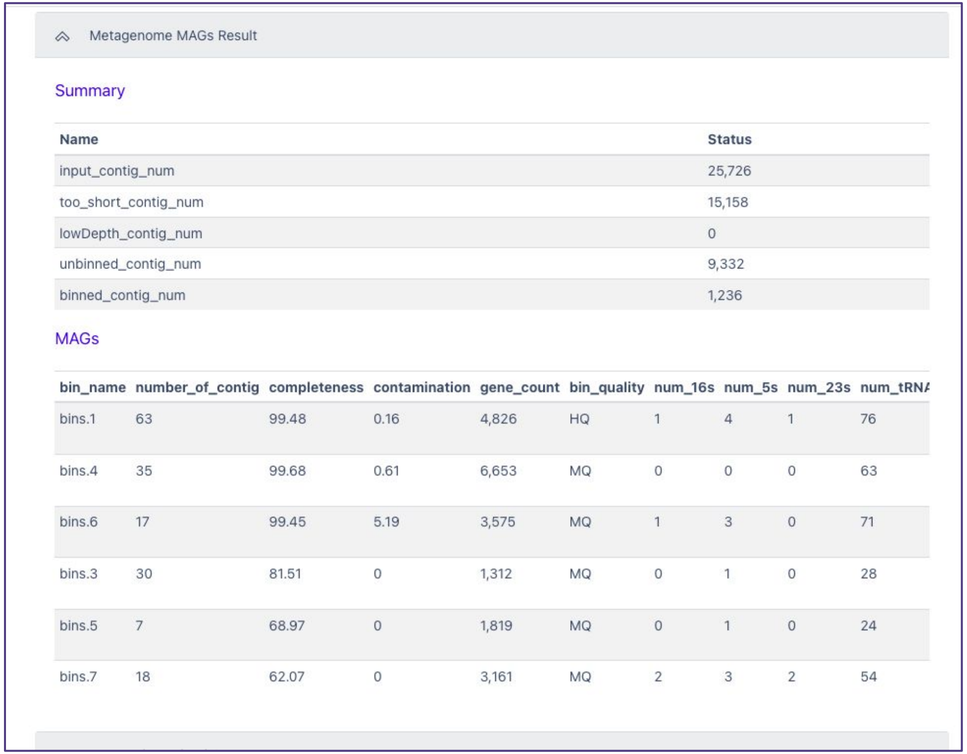


Metagenome_pipeline_test

Project Summary:

Description: This is a test of the NMDC metagenome pipeline for training purposes
Owner: jkelliher@lanl.gov
Submission Time: Mon Aug 23 2021 09:25:01 GMT-0600
Status: Complete
Type: Metagenome Pipeline

expand | close sections
General
ReadsQC Result
Read-based Taxonomy Classification Result
Metagenome Assembly Result
Metagenome Annotation Result
Metagenome MAGs Result
Browser/Download Outputs



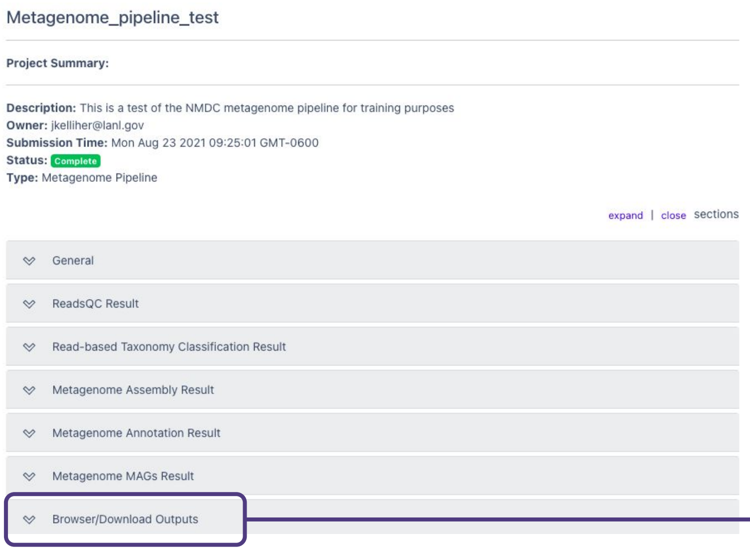
Summary

Name	Status
input_contig_num	25,726
too_short_contig_num	15,158
lowDepth_contig_num	0
unbinned_contig_num	9,332
binned_contig_num	1,236

MAGs

bin_name	number_of_contig	completeness	contamination	gene_count	bin_quality	num_16s	num_5s	num_23s	num_tRNA
bins.1	63	99.48	0.16	4,826	HQ	1	4	1	76
bins.4	35	99.68	0.61	6,653	MQ	0	0	0	63
bins.6	17	99.45	5.19	3,575	MQ	1	3	0	71
bins.3	30	81.51	0	1,312	MQ	0	1	0	28
bins.5	7	68.97	0	1,819	MQ	0	1	0	24
bins.7	18	62.07	0	3,161	MQ	2	3	2	54

Example Results: Metagenome

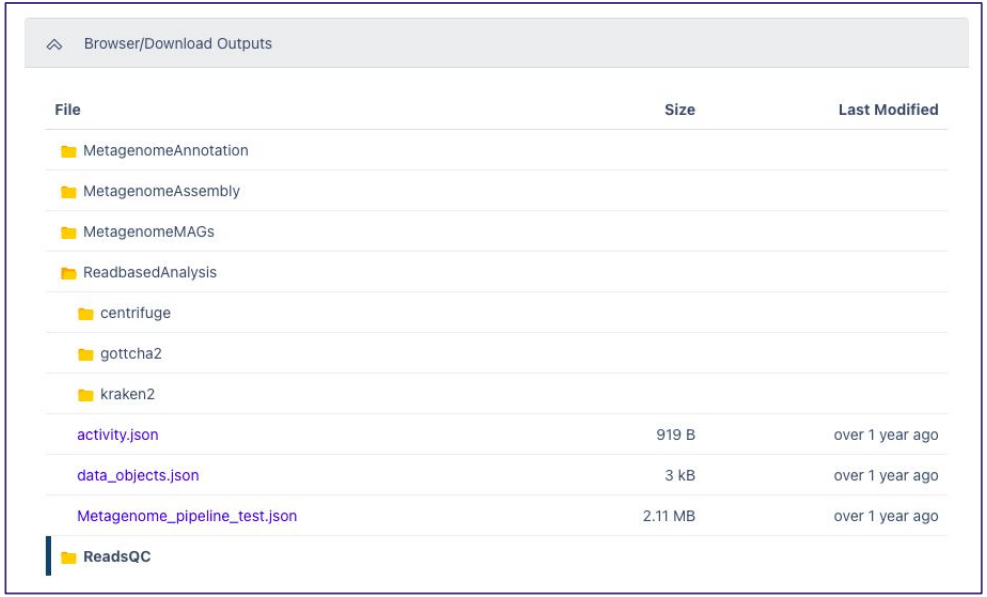


Metagenome_pipeline_test

Project Summary:

Description: This is a test of the NMDC metagenome pipeline for training purposes
Owner: jkelliher@lanl.gov
Submission Time: Mon Aug 23 2021 09:25:01 GMT-0600
Status: Complete
Type: Metagenome Pipeline

expand | close sections
General
ReadsQC Result
Read-based Taxonomy Classification Result
Metagenome Assembly Result
Metagenome Annotation Result
Metagenome MAGs Result
Browser/Download Outputs



Browser/Download Outputs

File	Size	Last Modified
MetagenomeAnnotation		
MetagenomeAssembly		
MetagenomeMAGs		
ReadbasedAnalysis		
centrifuge		
gottcha2		
kraken2		
activity.json	919 B	over 1 year ago
data_objects.json	3 kB	over 1 year ago
Metagenome_pipeline_test.json	2.11 MB	over 1 year ago
ReadsQC		

nmdc

National Microbiome Data Collaborative

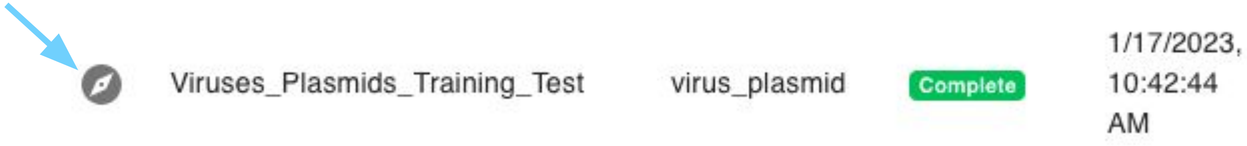
Questions?

Hands-on Activity

- Navigate to https://nmdc-edge.org/home
- Log in using your ORCiD
- Select Viruses & Plasmids workflow
    - Select ‘Run a single workflow’
- Enter a Project/Run name
    - Example: LastName_geNomad_Test
- Optional: enter a project description
- Select button to the right of ‘Select a File’ for the Input Assembled Fasta File
- Go to public data → virus_plasmid → you can choose any of the test files here (we recommend one of the ones that says _over5k)
- Run Option: Default
- Select Submit

Results

- Go to Public Projects; Select button on corresponding

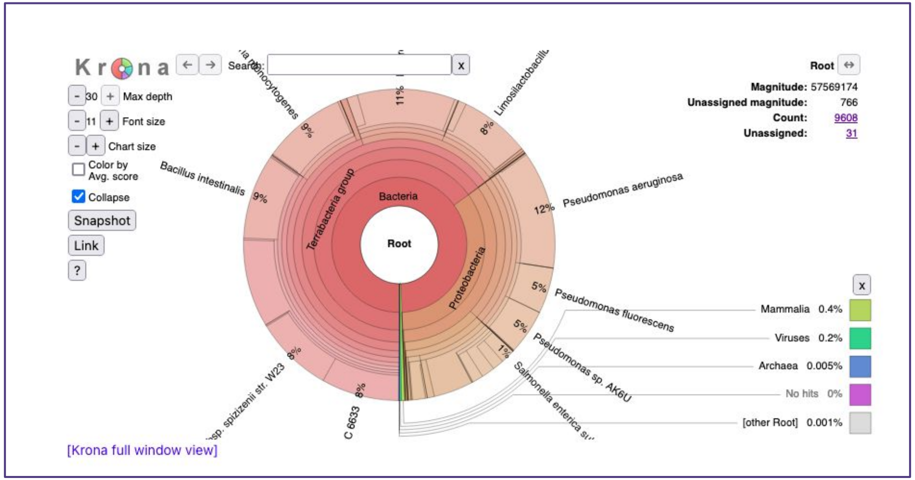


Viruses_Plasmids_Training_Test virus_plasmid Complete 1/17/2023, 10:42:44 AM

- General: provides run information
- Explore the virus_plasmid results!
- Browser/Download outputs: Provides downloadable files

Other Runs

- You can run any of the workflows at any time!
  - Can upload and run your own data
- Feel free to explore other public project results
  - Metagenome_pipeline_test
- Send comments or issues to: nmdc-edge@lanl.gov



30 Max depth

11 Font size

☐ Color by Avg. score

☑ Collapse

Root

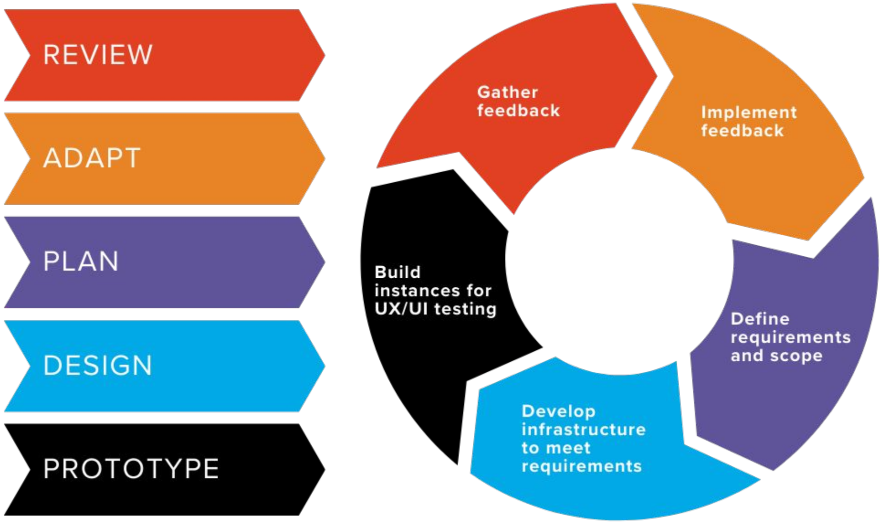
Magnitude:	57569174
Unassigned magnitude:	766
Count:	9608
Unassigned:	31

nmdc

National Microbiome
Data Collaborative

Questions?

Community-centered design process



REVIEW

ADAPT

PLAN

DESIGN

PROTOTYPE

Gather feedback

Implement feedback

Define requirements and scope

Develop infrastructure to meet requirements

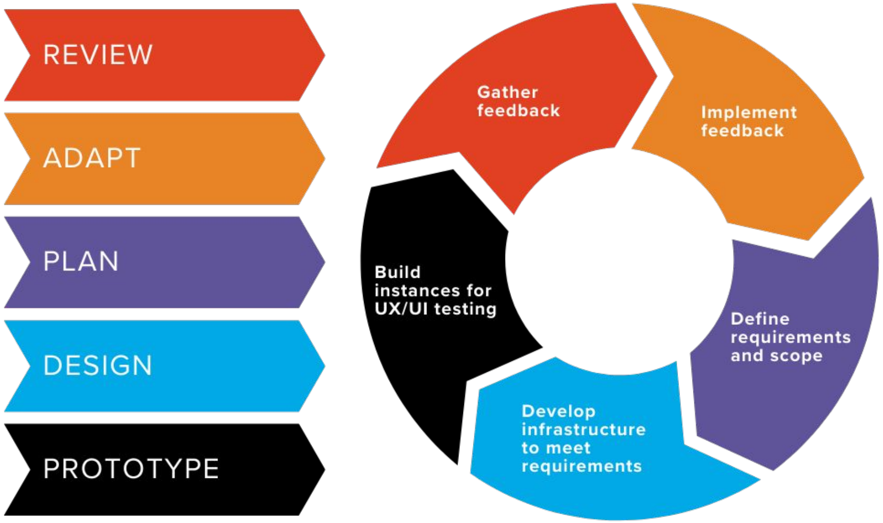
Build instances for UX/UI testing

2023-2024 User Research

34	321	120
User research participants	Insights from user interviews	Action items from insights

- The NMDC program utilizes a community-centered design approach to its product development process
    - User interviews, usability testing, beta-testing
    - Understand the needs of the community and how researchers use the products
- Workshops and event feedback is critical too
- Want to develop products that are as useful to the community as possible

Community-centered design process



REVIEW

ADAPT

PLAN

DESIGN

PROTOTYPE

Gather feedback

Implement feedback

Define requirements and scope

Develop infrastructure to meet requirements

Build instances for UX/UI testing

2023-2024 User Research

34
User research participants

321
Insights from user interviews

120
Action items from insights

NMDC EDGE Feedback examples

- Data visualization updates
- Support for batch processing
- Updates to how results are displayed and accessed
- Import SRA data easily with SRA accession numbers

NMDC EDGE Beta Testing

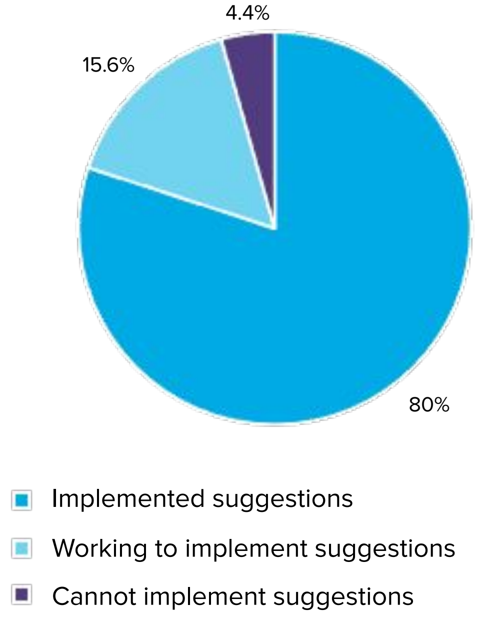
NMDC EDGE feedback areas:

- Running the workflows in NMDC EDGE
- NMDC EDGE user interface
- NMDC workflow training materials

Interested in becoming a beta-tester? We will give you credit on the website and through ORCiD!

- Volunteer at https://microbiomedata.org/user-research/
- Select “NMDC EDGE” for product interest

Implementation of NMDC EDGE beta-tester feedback



Reporting Mechanisms

- Beta tester form (NMDC EDGE homepage)
    - Specific workflows that you ran
    - Information about file sizes, issues with jobs (killed or errors), etc
- Email: nmdc-edge@lanl.gov
    - Can get troubleshooting help with NMDC EDGE workflows
- General reporting form
    - Did you come across any issues when using an NMDC product?
    - Features feedback
    - https://forms.gle/yxu9gkbufPigtbrB8

nmdc

National Microbiome Data Collaborative

Questions?

NMDC Resources

nmdc
National Microbiome Data Collaborative

Website: https://microbiomedata.org/
Data Portal: https://data.microbiomedata.org/
Submission Portal: https://data.microbiomedata.org/submission/home
NMDC EDGE: https://nmdc-edge.org/home
Github: https://github.com/microbiomedata GitHub
Docker Hub: https://hub.docker.com/u/microbiomedata docker
Documentation:
https://nmdc-documentation.readthedocs.io/en/latest/overview/nmdc_overview.html
YouTube: https://www.youtube.com/channel/UCyBqKc46NQZ_YgZIKGYeglw/featured

Get involved!

Sign up for our newsletter
microbiomedata.org

Become a NMDC Champion
bit.ly/champions-program

Find us on X/Twitter
@microbiomedata

Find us on LinkedIn
https://bit.ly/NMDC_LinkedIn

Find us on Instagram
@microbiomedata

Read more about the NMDC

nature microbiology

Kelliher JM *et al.* Cohort-based learning for microbiome research community standards. *Nat Microbiol* (2023). doi.org/10.1038/s41564-023-01361-7.

frontiers

Hu B, Canon S, Eloe-Fadrosh EA, et al.. Challenges in Bioinformatics Workflows for Processing Microbiome Omics Data at Scale. *Front Bioinform*. 1:826370. (2022) doi: 10.3389/fbinf.2021.826370.

Nucleic Acids Research

Eloe-Fadrosh EA *et al*. The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res.* 7;60(D1):D828–D836. (2022) doi: 10.1093/nar/gkab990.

nature REVIEWS MICROBIOLOGY

Wood-Charlson, E.M., Anubhav, Auberry, D. *et al.* The National Microbiome Data Collaborative: enabling microbiome science. *Nat Rev Microbiol* **18,** 313–314 (2020). doi.org/10.1038/s41579-020-0377-0

mSystems

Vangay, P *et al.* Microbiome metadata standards: Report of the National Microbiome Data Collaborative's workshop and follow-on activities. *mSystems* **6**, e01194-20 (2021). doi.org/10.1128/mSystems.01194-20

nmdc

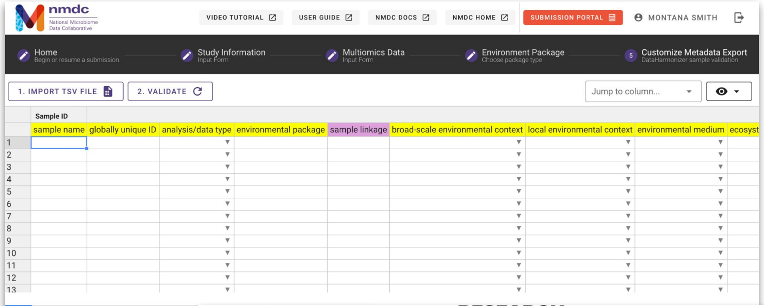
National Microbiome Data Collaborative

Metadata Standards and Submission Portal

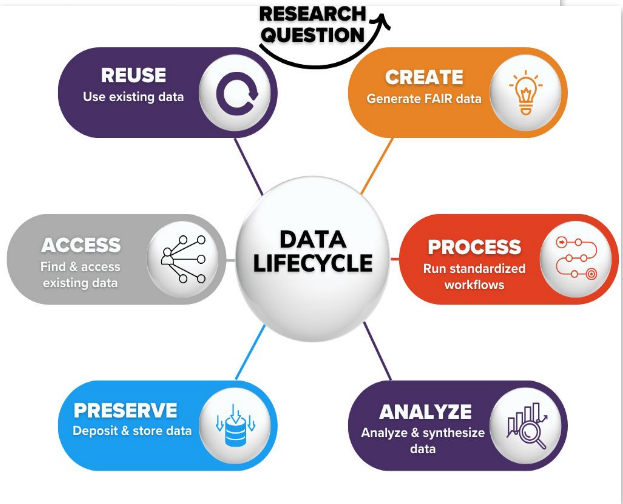
2024 Ambassador Cohort

Purpose of this training

- Provide an overview of the benefits of metadata standardization and how this promotes interoperability
- Discuss the importance of metadata standards to enable data reuse
- Provide template slides to use for events
- Introduce audience to the NMDC Submission Portal
  - The hands-on activity can be used (or modified and used) for use in your events too!



	Sample ID								
	sample name	globally unique ID	analysis/data type	environmental package	sample linkage	broad-scale environmental context	local environmental context	environmental medium	ecosyst
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									



nmdc

National Microbiome Data Collaborative

What are metadata?

What are Metadata?

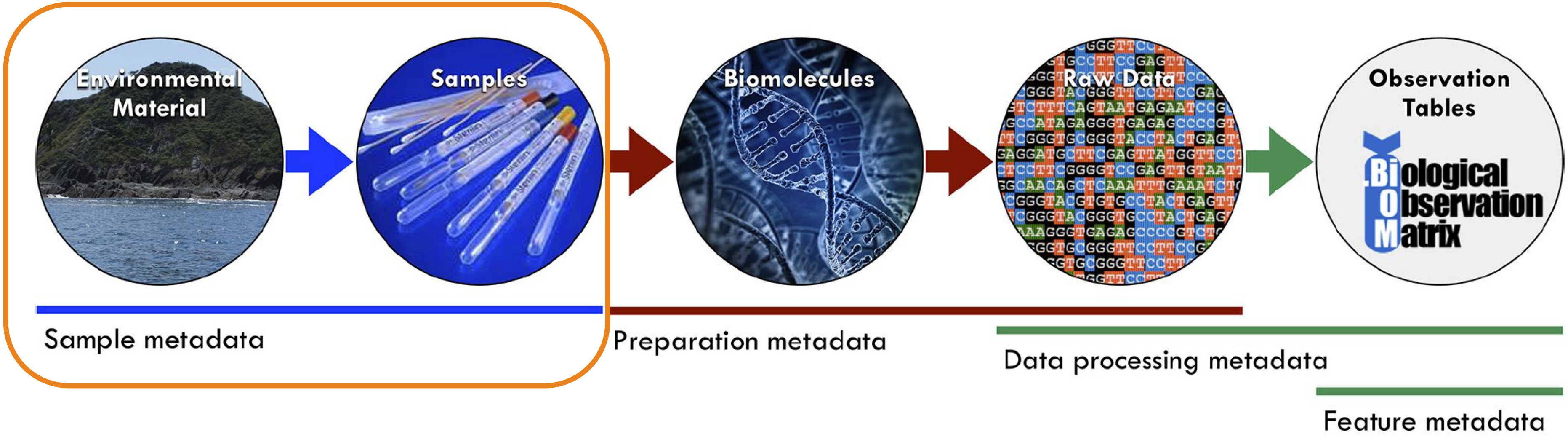
Metadata are …

- Contextual data about your data
- Vital for data
  - Publication & deposition
  - Preservation
  - Discovery
  - Access
  - Reuse

Sample Metadata

nmdc
National Microbiome Data Collaborative

Microbiome Environmental & Sample Metadata



From: **Introduction to Metadata and Ontologies:** Everything You Always Wanted to Know About Metadata and Ontologies (But Were Afraid to Ask) DOI: 10.25979/1607365

What are Metadata?

nmdc
National Microbiome
Data Collaborative

Sample metadata includes information about:

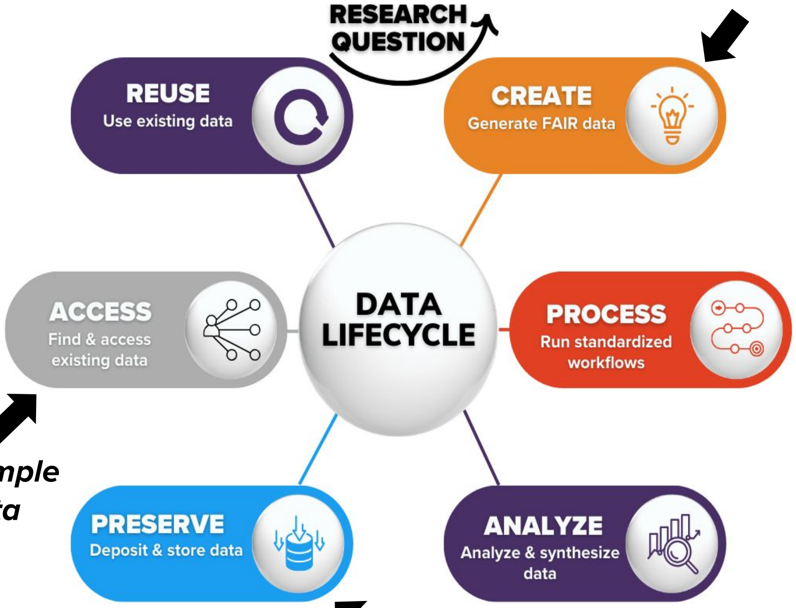
- ***When*** it was collected
- ***Where*** it was collected
- ***How*** it was collected
- ***What*** kind of sample is it
- ***Treatment*** applied during experimentation
- ***Environmental Properties*** from which the sample was taken

Data Lifecycle and Metadata

nmdc
National Microbiome Data Collaborative

During experimental design:

- Plan for every stage: be intentional about the process of collecting, storing, processing, and protecting (meta)data
- Detail a data management plan that includes metadata for your samples



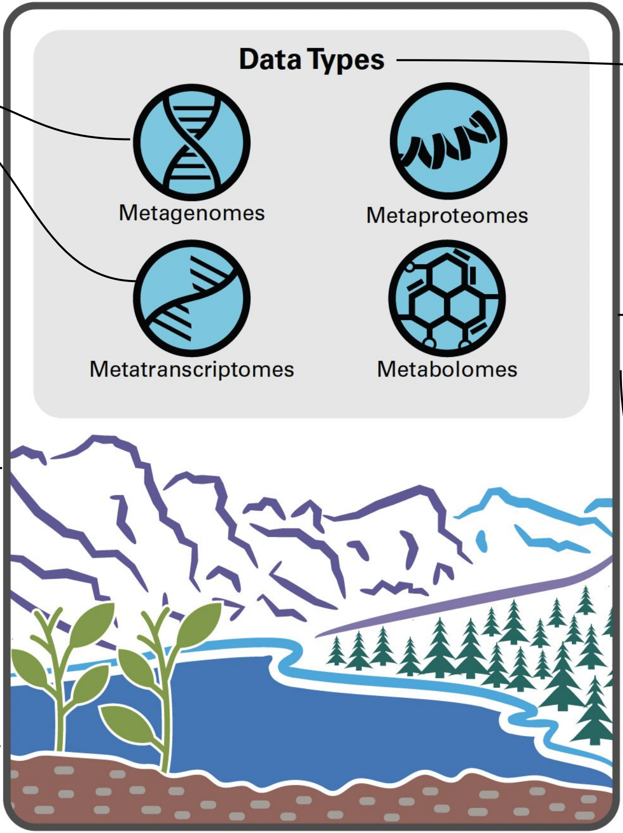
Collect Sample metadata

Access Sample metadata

Submit Sample metadata

Breadth of information in microbiome science

nmdc
National Microbiome Data Collaborative



Sequencing method

Nucleic acid extraction
Sequencing method

Multiple different biogeochemical and environmental parameters

climate
elevation

veg type
plant part

lat/lon
biome
type
material
...

pH
temperature
carbon
nitrogen
...

depth
soil type
land use

Study information

PI & Contributors
experimental design

Different sample Treatments and preparations

Sample collection device
Sample processing

Analysis outputs

assembly statistics
gene function
metabolite and peptide counts
taxon abundance

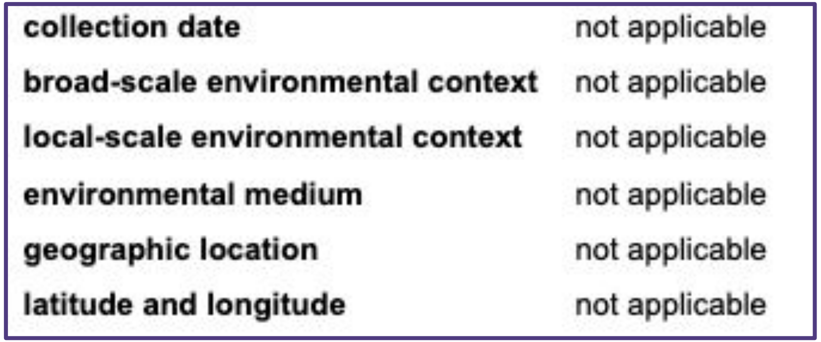
Valid and complete metadata

nmdc
National Microbiome Data Collaborative

Sampling date	04.06.2023
pH	100
Geographic location	n/a
Host	Plant
Elevation	10,000 m
Sequencing technology	

- Lots of samples and studies exist with little to no metadata
- Metadata can be incorrect (e.g. pH of 100) or not usable
- For your own studies, important that you collect valid and complete metadata for data to be FAIR

Real example from NCBI SRA



collection date	not applicable
broad-scale environmental context	not applicable
local-scale environmental context	not applicable
environmental medium	not applicable
geographic location	not applicable
latitude and longitude	not applicable

Valid and complete metadata

Sampling date	04.06.2023	Format not standardized, not ISO compliant
pH	100	Invalid entry, not in range
Geographic location	n/a	Field not filled out
Host	Plant	Entry not very specific
Elevation	10,000 m	Likely incorrect units
Sequencing technology		Missing fields

Why metadata standards?

nmdc
National Microbiome
Data Collaborative

Datasets can be difficult to compare and reuse if they lack consistent language and formatting

Why aren't these datasets directly comparable? What isn't standardized?

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

sampleNum	substance	sample depth	temp
8725	dirt	45 cm	21.1
2312	ground liquid	105 cm	7

Why metadata standards?

Metadata fields and titles can be inconsistent

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

sampleNum	substance	sample depth	temp
8725	dirt	45 cm	21.1
2312	ground liquid	105 cm	7

Why metadata standards?

The terms used to describe metadata can be different

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

sampleNum	substance	sample depth	temp
8725	dirt	45 cm	21.1
2312	ground liquid	105 cm	7

Why metadata standards?

The units used can be inconsistent and not directly comparable

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

sampleNum	substance	sample depth	temp
8725	dirt	45 cm	21.1
2312	ground liquid	105 cm	7

Why metadata standards?

Some metadata may lack units or descriptive information

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

sampleNum	substance	sample depth	temp
8725	dirt	45 cm	21.1
2312	ground liquid	105 cm	7

Why metadata standards?

Adopting standards for reporting makes data human and machine readable.

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

idNumber	material	sample depth	temperature
8725	soil	.45 m	21.1 °C
2312	groundwater	1.05 m	7 °C

Why metadata standards?

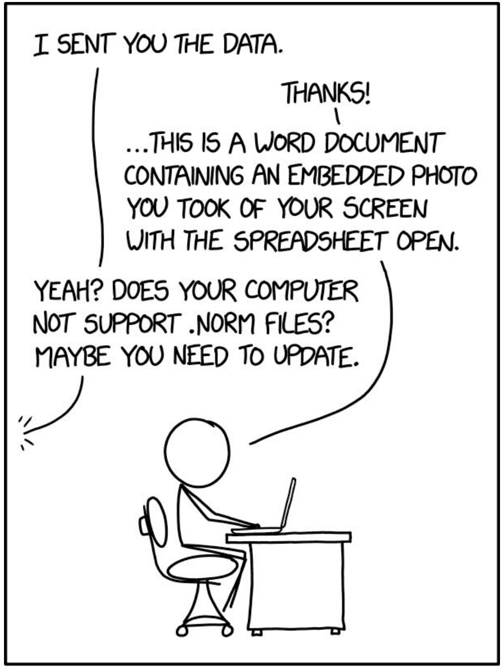
nmdc
National Microbiome Data Collaborative

Adopting standards for reporting makes data human and machine readable.

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

idNumber	material	sample depth	temperature
8725	soil	.45 m	21.1 °C
2312	groundwater	1.05 m	7 °C

Consequences of not standardizing metadata


I SENT YOU THE DATA.
THANKS!
...THIS IS A WORD DOCUMENT CONTAINING AN EMBEDDED PHOTO YOU TOOK OF YOUR SCREEN WITH THE SPREADSHEET OPEN.
YEAH? DOES YOUR COMPUTER NOT SUPPORT .NORM FILES? MAYBE YOU NEED TO UPDATE.

SINCE EVERYONE SENDS STUFF THIS WAY ANYWAY, WE SHOULD JUST FORMALIZE IT AS A STANDARD.

https://xkcd.com/2116/

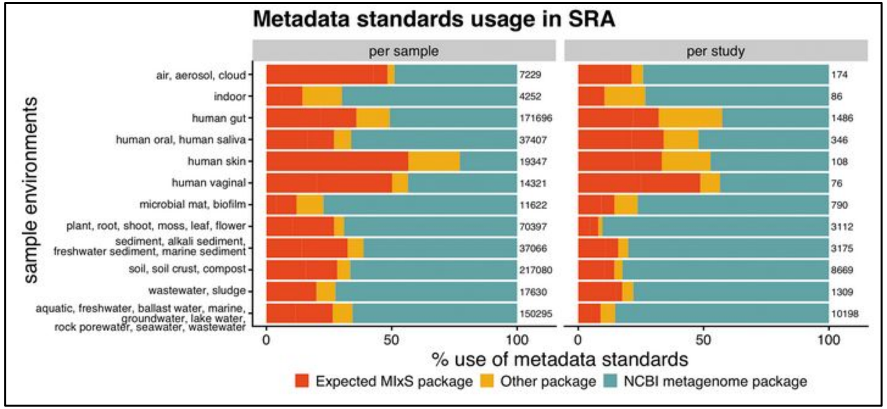
- Can miss critical contextual information when performing analyses
    - May miss other confounding variables
- Data may not be able to be published or deposited into certain repositories
- Limits data comparisons and reuse within a group and beyond
- Prevents reproducibility
- Difficult to search for, difficult to compare with other datasets

nmdc

National Microbiome Data Collaborative

Metadata Standards

The standards gap



What we found:

- Lack of awareness of metadata standards, inconsistent usage, lack of training
- Metadata managed as Excel spreadsheets
- Not machine-actionable
- Not following FAIR

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

PERSPECTIVE
January/February 2021 Volume 6 Issue 1 e01194-20
https://doi.org/10.1128/mSystems.01194-20

Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative's Workshop and Follow-On Activities

Pajau Vangay [a], Josephine Burgin [b], Anjanette Johnston[c], Kristen L. Beck [d], Daniel C. Berrios [e], Kai Blumberg [f], Shane Canon[a], Patrick Chain[g], John-Marc Chandonia [a], Danielle Christianson[a], Sylvain V. Costes[e], Joan Damerow[a], William D. Duncan[a], Jose Pablo Dundore-Arias [h], Kjiersten Fagnan[a], Jonathan M. Galazka [e], Sean M. Gibbons [i,j], David Hays[a], Judson Hervey [k], Bin Hu [g], Bonnie L. Hurwitz [f], Pankaj Jaiswal [l], Marcin P. Joachimiak[a], Linda Kinkel[m], Joshua Ladau[a], Stanton L. Martin[n], Lee Ann McCue [o], Kayd Miller [a], Nigel Mouncey[a], Chris Mungall[a], Evangelos Pafilis [p], T. B. K. Reddy [a], Lorna Richardson [b], Simon Roux [q], Lynn M. Schriml[w], Justin P. Shaffer [r], Jagadish Chandrabose Sundaramurthi [a], Luke R. Thompson [s,t], Ruth E. Timme [u], Jie Zheng [v], Elisha M. Wood-Charlson [a], Emiley A. Eloe-Fadrosh [a]

[a]Lawrence Berkeley National Laboratory, Berkeley, California, USA
[b]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
[c]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
[d]IBM Almaden Research Center, San Jose, California, USA
[e]NASA Ames Research Center, Moffett Field, California, USA

Community Metadata Standards

The NMDC metadata standards utilize and enhance *existing* community-driven standards

1. MIxS: Minimum Information about any (x) Sequence

   Genomic Standards Consortium (GSC)

2. GOLD: Genomes OnLine Database

   Joint Genome Institute (JGI)

4. EnvO: Environment Ontology

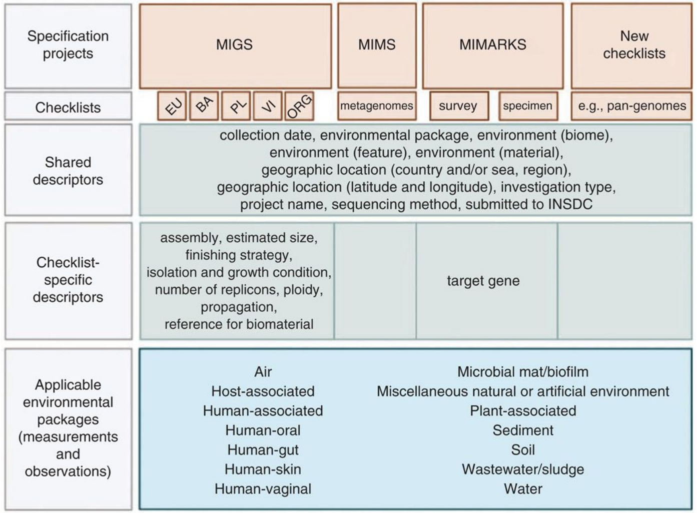
   Open Biological and Biomedical Ontology (OBO) Foundry

Standards for describing sample context

GSC MIxS

- Minimal Information about any (x) Sequence (MIxS)
- ~600 standardized fields for reporting
    - Different environmental packages
        - Soil
        - Water
        - Sediment
        - Plant-associated
        - ...

genomic STANDARDS consortium

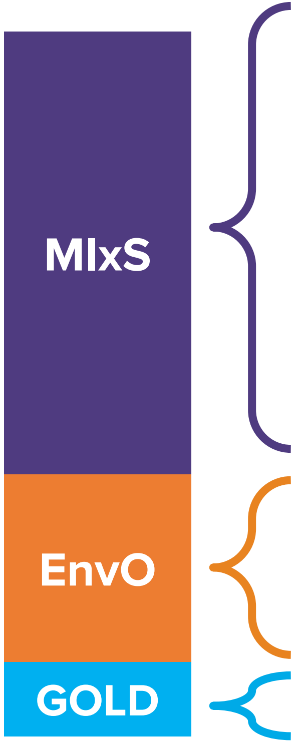


Specification projects	MIGS	MIMS	MIMARKS	New checklists
Checklists	EU BA PL VI ORG	metagenomes	survey specimen	e.g., pan-genomes
Shared descriptors	collection date, environmental package, environment (biome), environment (feature), environment (material), geographic location (country and/or sea, region), geographic location (latitude and longitude), investigation type, project name, sequencing method, submitted to INSDC			
Checklist-specific descriptors	assembly, estimated size, finishing strategy, isolation and growth condition, number of replicons, ploidy, propagation, reference for biomaterial		target gene	
Applicable environmental packages (measurements and observations)	Air Host-associated Human-associated Human-oral Human-gut Human-skin Human-vaginal Microbial mat/biofilm Miscellaneous natural or artificial environment Plant-associated Sediment Soil Wastewater/sludge Water			

The NMDC + Community Standards

nmdc
National Microbiome
Data Collaborative

NMDC derives ***required metadata*** from community standards



Examples for MIxS:

Sample Identifiers

Growth Facility

Geographic Location (latitude longitude)

Geographic Region (country and/or sea, region)

Collection Date

Sample Material Processing

Storage Conditions & Temperature

Broad-scale Environmental Context

Local Environmental Context

Environmental Medium

GOLD Environment Path

MIxS - Sample Metadata Standards

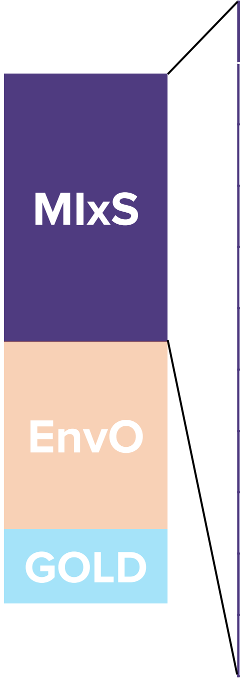
MIxS
EnvO
GOLD

Genomic Standards Consortium

- **MIxS: Environmental Extensions**
- System of unified standards for describing microbiome samples from diverse environments (e.g., plant, soil, sediment, water)
- Each package includes a list of ***metadata descriptors*** for each parameter/field
- NMDC leverages MIxS ***Mandatory Metadata Fields***

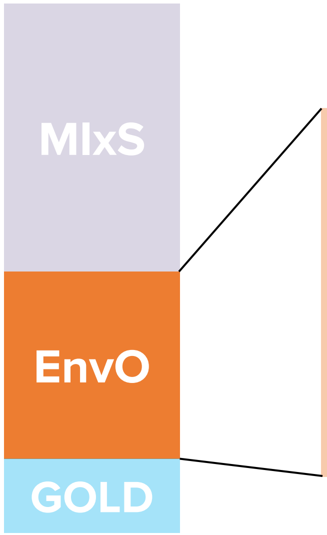
MIxS Environmental Extensions

17 extensions (*currently available in NMDC)
New extensions added with community input



MIxS Environmental Extensions	
*air	*built environment
*host-associated	human-associated
human-gut	human-oral
human-skin	human-vaginal
*hydrocarbon resources-cores	*hydrocarbon resources-fluids/swabs
*microbial mat/biofilm	*miscellaneous natural or artificial environment
*plant associated	*sediment
*soil	wastewater/sludge
*water	agriculture
Pending: food, parasite, etc.	

EnvO - Sample Metadata Standards



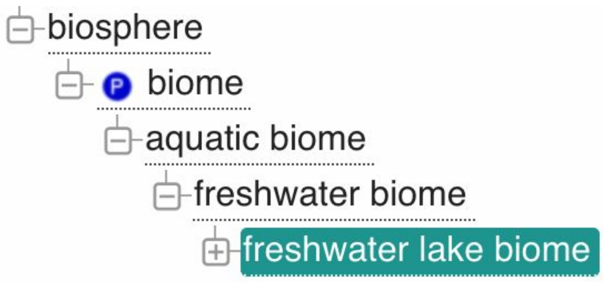
EnvO (Environment Ontology)

- Dynamic, community resource
- Hierarchical classification of samples by environment
- Mandated by MIxS for environment fields

MIxS requires environment terms

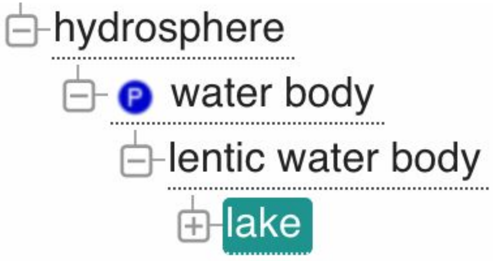
MIxS descriptors specify the ***sample environment*** with the Environment Ontology (EnvO)

broad-scale environmental context



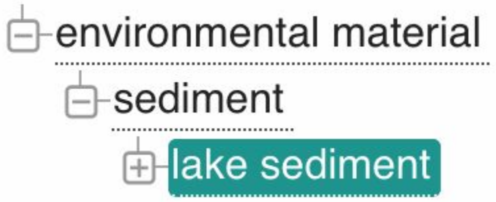
- biosphere
  - P biome
    - aquatic biome
      - freshwater biome
        - freshwater lake biome

local-scale environmental context



- hydrosphere
  - P water body
    - lentic water body
      - lake

environmental medium



- environmental material
  - sediment
    - lake sediment

http://obofoundry.org/ontology/envo

The NMDC + Community Standards

nmdc
National Microbiome
Data Collaborative

MIxS

EnvO

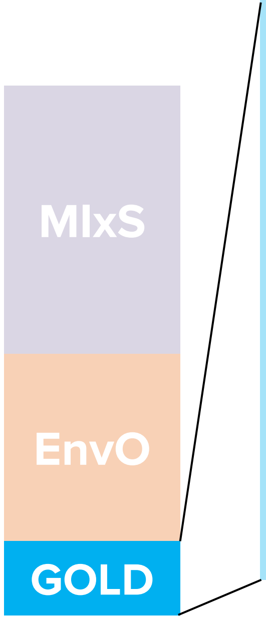
GOLD

MIxS / EnvO		
Broad-scale environment	Local-scale environment	Environmental Medium
Freshwater lake biome	Lake Shore	Sediment
Freshwater lake biome	Lake	Algal bloom

EnvO & GOLD terms together gives us improved environmental context to the microbiome biosamples!

GOLD Ecosystem classification				
Ecosystem	Ecosystem Category	Ecosystem Type	Specific Ecosystem	Ecosystem Tree
Environment	Aquatic	Freshwater	Lake	Sediment
Environment	Aquatic	Freshwater	Lake	Algal bloom

GOLD - Sample Metadata Standards



The Genomes OnLine Database (GOLD)

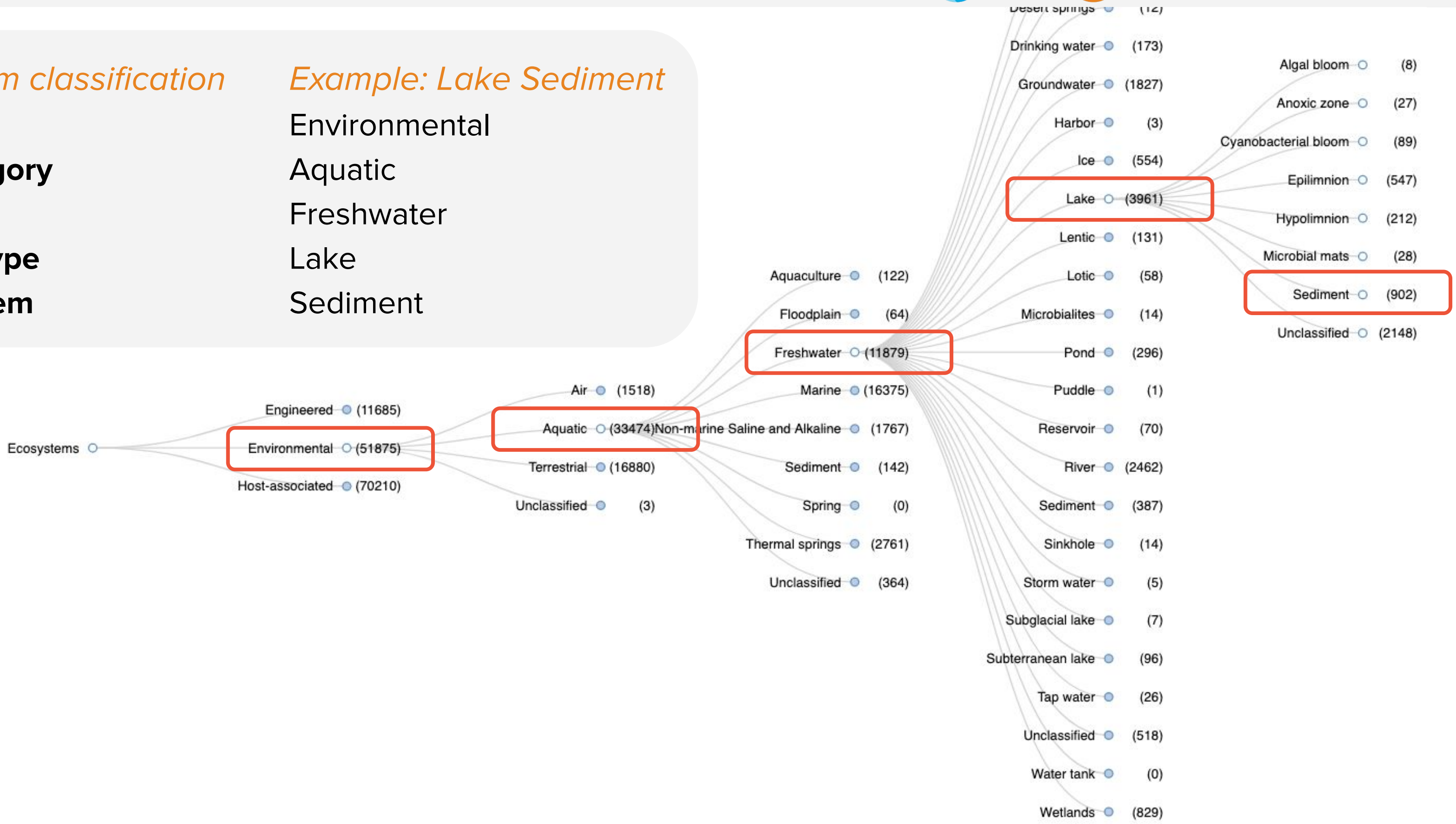
- Manually curated collection of genome projects and their metadata
- Metadata fields: ~600
- Controlled Vocabulary fields: 76 (3,873 terms)
- Currently contains hundreds of thousands of microbiome biosamples

GOLD - Five-level ecosystem path

nmdc

National Microbiome Data Collaborative

GOLD Ecosystem classification	*Example: Lake Sediment*
Ecosystem	Environmental
Ecosystem Category	Aquatic
Ecosystem Type	Freshwater
Ecosystem Subtype	Lake
Specific Ecosystem	Sediment



GOLD - Five-level ecosystem path

Activity: Explore the GOLD ecosystem tree viewer to identify your sample metadata path

https://gold.jgi.doe.gov/ecosystemtree

Product Initiatives

Submission Portal

Lower barriers to collect study and biosample data

NMDC EDGE

Streamline multi-omics data processing

Data Portal & API

Access and discovery of microbiome information

Engagement

User Facilities

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

EMSL

Individuals

The NMDC Champions Program

The NMDC Ambassador Program

Strategic Partners

NSF

neon Operated by Battelle

NASA

GeneLab

Product Initiatives

Submission Portal

Lower barriers to collect study and biosample data

NMDC EDGE

Streamline multi-omics data processing

Data Portal & API

Access and discovery of microbiome information

Engagement

User Facilities

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

EMSL

Individuals

The NMDC Champions Program

The NMDC Ambassador Program

Strategic Partners

NSF

neon Operated by Battelle

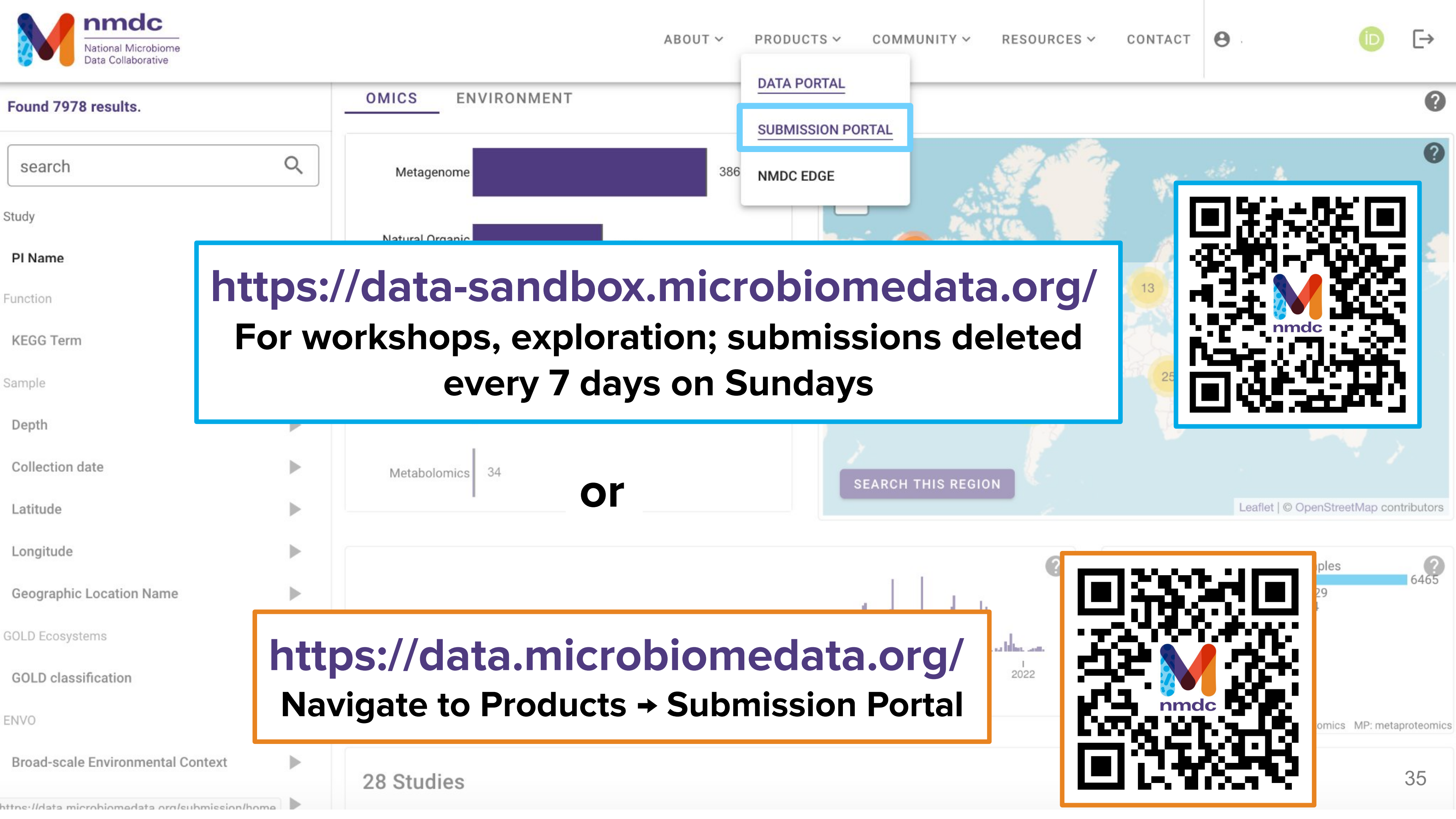
NASA

GeneLab

nmdc

National Microbiome Data Collaborative

NMDC Submission Portal



Found 7978 results.

search

Study

PI Name

Function

KEGG Term

Sample

Depth

Collection date

Latitude

Longitude

Geographic Location Name

GOLD Ecosystems

GOLD classification

ENVO

Broad-scale Environmental Context

28 Studies

https://data.microbiomedata.org/submission/home

https://data-sandbox.microbiomedata.org/

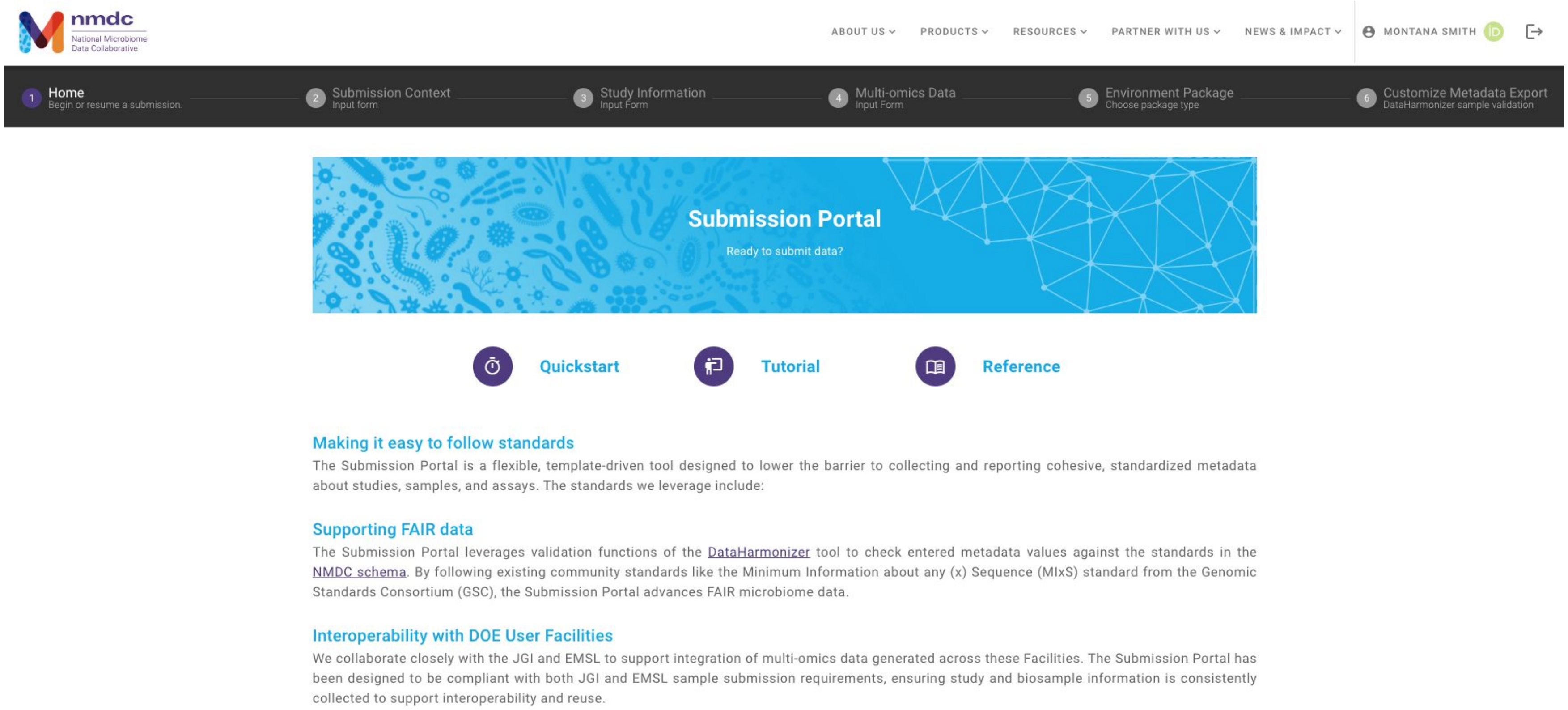
For workshops, exploration; submissions deleted every 7 days on Sundays

or

https://data.microbiomedata.org/

Navigate to Products → Submission Portal

The NMDC Submission Portal



Making it easy to follow standards

The Submission Portal is a flexible, template-driven tool designed to lower the barrier to collecting and reporting cohesive, standardized metadata about studies, samples, and assays. The standards we leverage include:

Supporting FAIR data

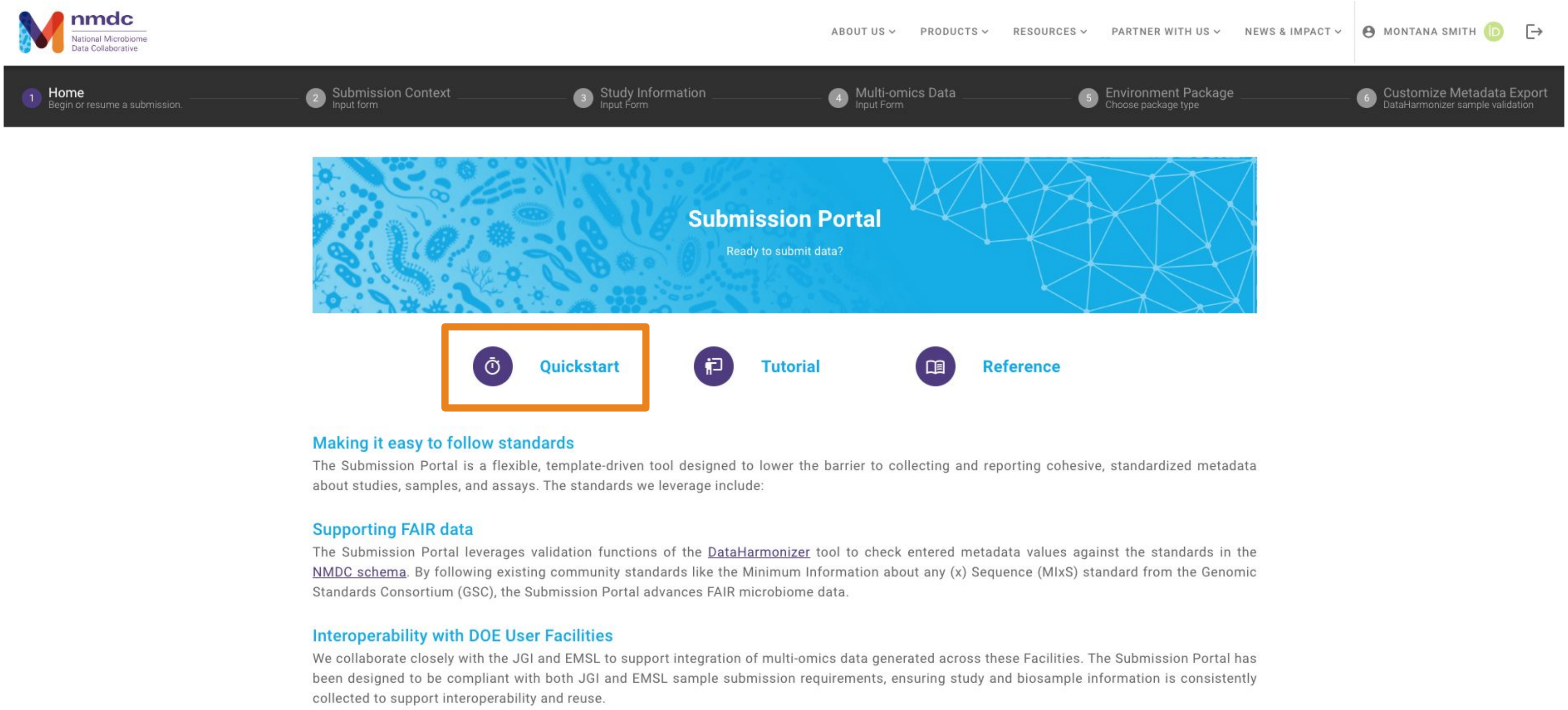
The Submission Portal leverages validation functions of the DataHarmonizer tool to check entered metadata values against the standards in the NMDC schema. By following existing community standards like the Minimum Information about any (x) Sequence (MIxS) standard from the Genomic Standards Consortium (GSC), the Submission Portal advances FAIR microbiome data.

Interoperability with DOE User Facilities

We collaborate closely with the JGI and EMSL to support integration of multi-omics data generated across these Facilities. The Submission Portal has been designed to be compliant with both JGI and EMSL sample submission requirements, ensuring study and biosample information is consistently collected to support interoperability and reuse.

The NMDC Submission Portal

nmdc
National Microbiome Data Collaborative



1 Home
Begin or resume a submission.

2 Submission Context
Input form

3 Study Information
Input Form

4 Multi-omics Data
Input Form

5 Environment Package
Choose package type

6 Customize Metadata Export
DataHarmonizer sample validation

Submission Portal

Ready to submit data?

Making it easy to follow standards

The Submission Portal is a flexible, template-driven tool designed to lower the barrier to collecting and reporting cohesive, standardized metadata about studies, samples, and assays. The standards we leverage include:

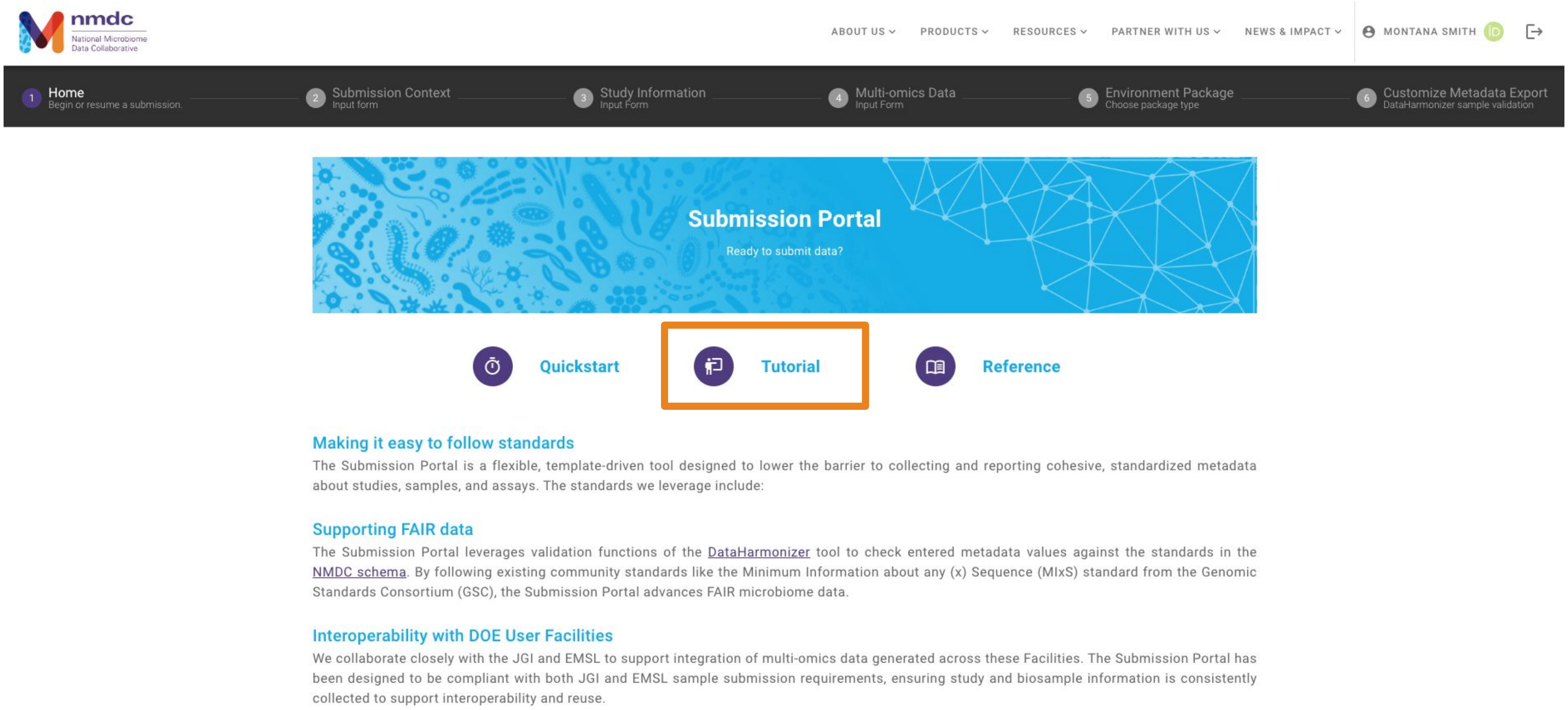
Supporting FAIR data

The Submission Portal leverages validation functions of the DataHarmonizer tool to check entered metadata values against the standards in the NMDC schema. By following existing community standards like the Minimum Information about any (x) Sequence (MIxS) standard from the Genomic Standards Consortium (GSC), the Submission Portal advances FAIR microbiome data.

Interoperability with DOE User Facilities

We collaborate closely with the JGI and EMSL to support integration of multi-omics data generated across these Facilities. The Submission Portal has been designed to be compliant with both JGI and EMSL sample submission requirements, ensuring study and biosample information is consistently collected to support interoperability and reuse.

The NMDC Submission Portal



Submission Portal

Ready to submit data?

Making it easy to follow standards

The Submission Portal is a flexible, template-driven tool designed to lower the barrier to collecting and reporting cohesive, standardized metadata about studies, samples, and assays. The standards we leverage include:

Supporting FAIR data

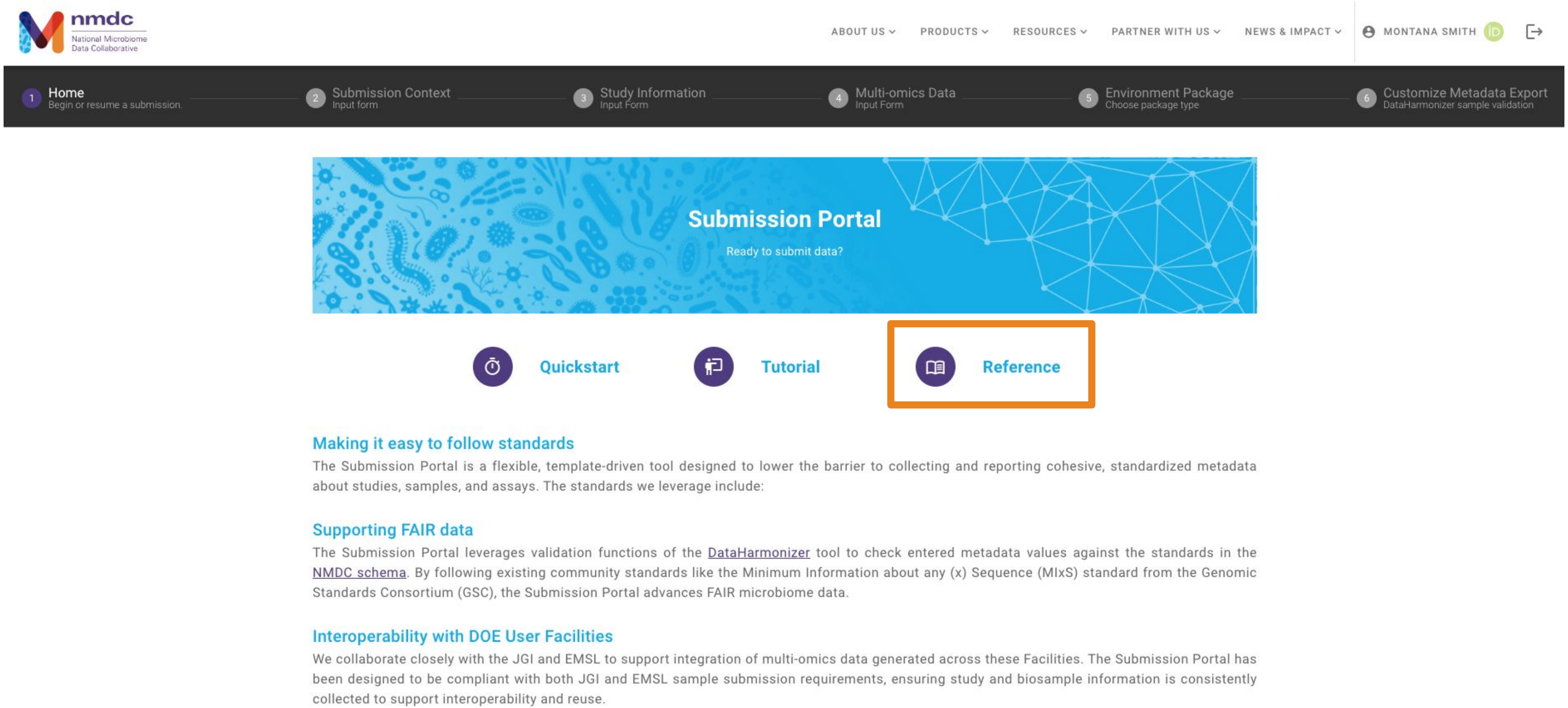
The Submission Portal leverages validation functions of the DataHarmonizer tool to check entered metadata values against the standards in the NMDC schema. By following existing community standards like the Minimum Information about any (x) Sequence (MIxS) standard from the Genomic Standards Consortium (GSC), the Submission Portal advances FAIR microbiome data.

Interoperability with DOE User Facilities

We collaborate closely with the JGI and EMSL to support integration of multi-omics data generated across these Facilities. The Submission Portal has been designed to be compliant with both JGI and EMSL sample submission requirements, ensuring study and biosample information is consistently collected to support interoperability and reuse.

The NMDC Submission Portal

nmdc
National Microbiome Data Collaborative



Making it easy to follow standards

The Submission Portal is a flexible, template-driven tool designed to lower the barrier to collecting and reporting cohesive, standardized metadata about studies, samples, and assays. The standards we leverage include:

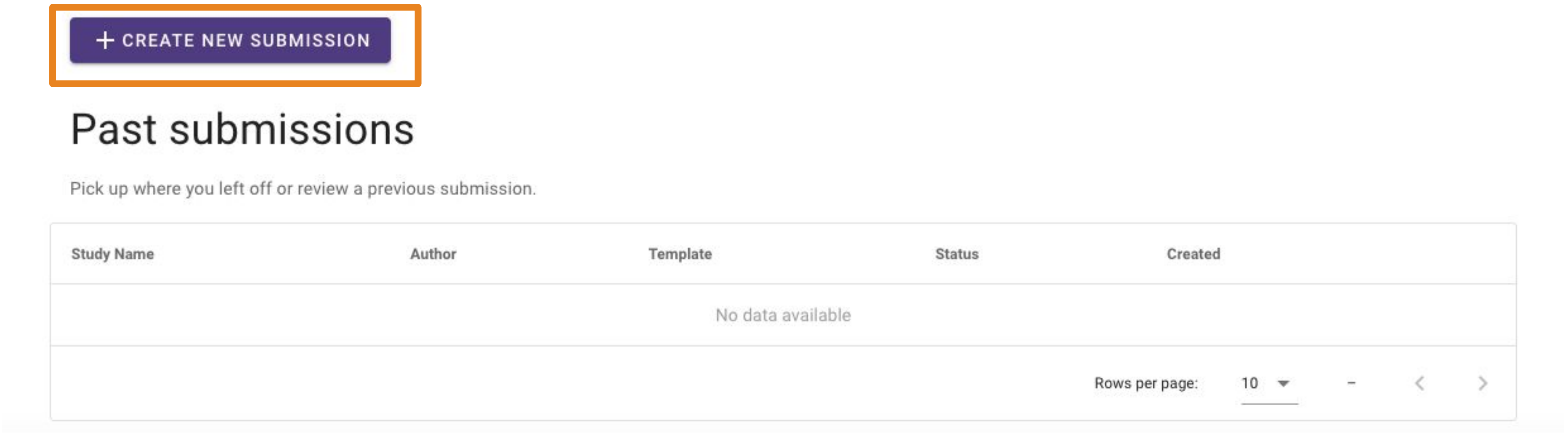
Supporting FAIR data

The Submission Portal leverages validation functions of the DataHarmonizer tool to check entered metadata values against the standards in the NMDC schema. By following existing community standards like the Minimum Information about any (x) Sequence (MIxS) standard from the Genomic Standards Consortium (GSC), the Submission Portal advances FAIR microbiome data.

Interoperability with DOE User Facilities

We collaborate closely with the JGI and EMSL to support integration of multi-omics data generated across these Facilities. The Submission Portal has been designed to be compliant with both JGI and EMSL sample submission requirements, ensuring study and biosample information is consistently collected to support interoperability and reuse.

The NMDC Submission Portal



Past submissions

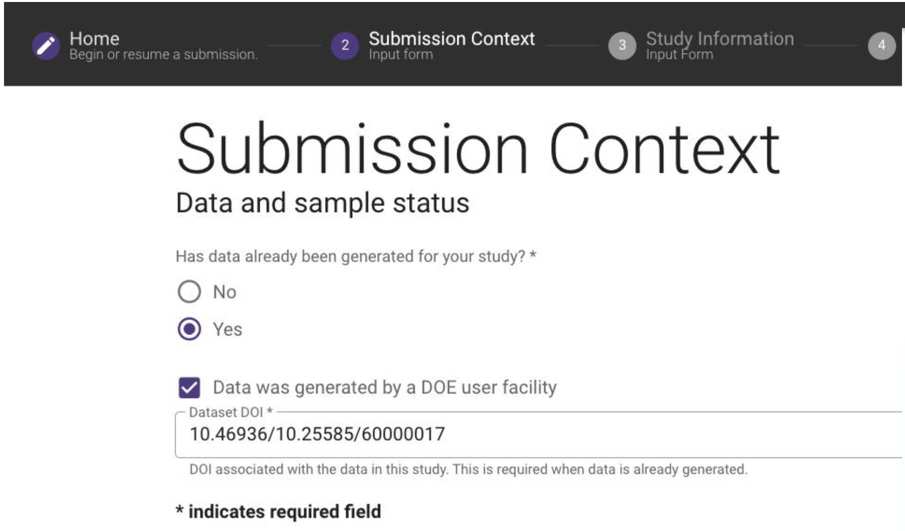
Pick up where you left off or review a previous submission.

Study Name	Author	Template	Status	Created
No data available				

Rows per page: 10 –

The NMDC Submission Portal

nmdc
National Microbiome Data Collaborative



Submission Context

Data and sample status

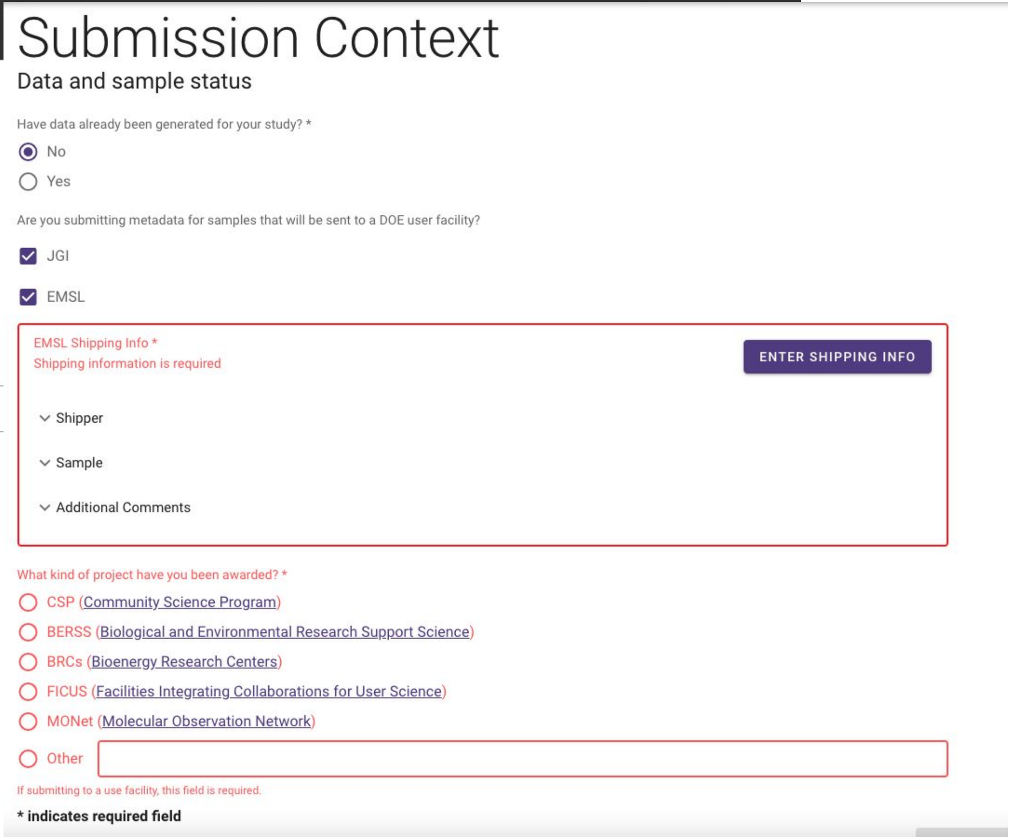
Has data already been generated for your study? *

○ No
● Yes

☑ Data was generated by a DOE user facility

Dataset DOI *
10.46936/10.25585/60000017
DOI associated with the data in this study. This is required when data is already generated.

*** indicates required field**



Submission Context

Data and sample status

Have data already been generated for your study? *

● No
○ Yes

Are you submitting metadata for samples that will be sent to a DOE user facility?

☑ JGI

☑ EMSL

EMSL Shipping Info *
Shipping information is required

What kind of project have you been awarded? *

○ CSP (Community Science Program)
○ BERSS (Biological and Environmental Research Support Science)
○ BRCs (Bioenergy Research Centers)
○ FICUS (Facilities Integrating Collaborations for User Science)
○ MONet (Molecular Observation Network)
○ Other

If submitting to a use facility, this field is required.

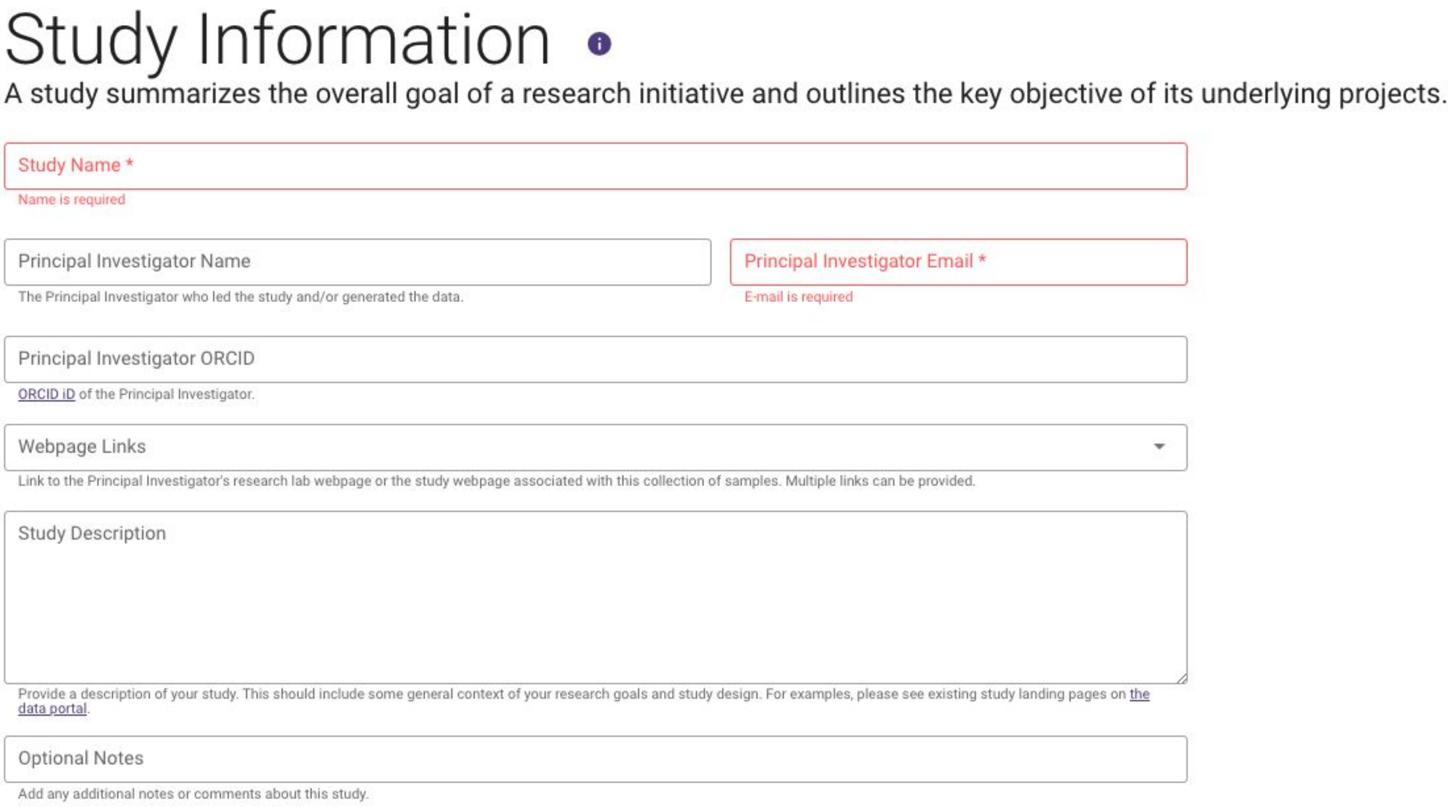
*** indicates required field**

Submission Context

- Indicate if data already exists
- Are you using NMDC to complete metadata for samples going to a DOE user facility?

The NMDC Submission Portal

nmdc
National Microbiome Data Collaborative



Study Information

A study summarizes the overall goal of a research initiative and outlines the key objective of its underlying projects.

Study Name *

Name is required

Principal Investigator Name

The Principal Investigator who led the study and/or generated the data.

Principal Investigator Email *

E-mail is required

Principal Investigator ORCID

ORCID iD of the Principal Investigator.

Webpage Links

Link to the Principal Investigator's research lab webpage or the study webpage associated with this collection of samples. Multiple links can be provided.

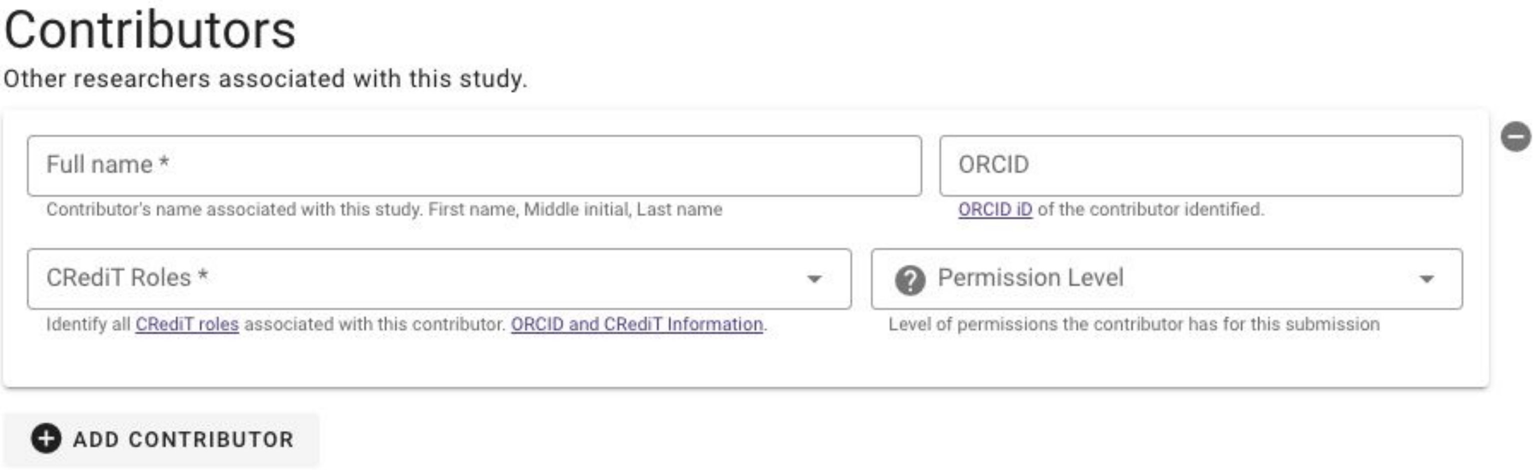
Study Description

Provide a description of your study. This should include some general context of your research goals and study design. For examples, please see existing study landing pages on the data portal.

Optional Notes

Add any additional notes or comments about this study.

The NMDC Submission Portal



Contributors

Other researchers associated with this study.

Full name *

Contributor's name associated with this study. First name, Middle initial, Last name

ORCID

ORCID iD of the contributor identified.

CRediT Roles *

Identify all CRediT roles associated with this contributor. ORCID and CRediT Information.

Permission Level

Level of permissions the contributor has for this submission

Contributor Roles & Access Permissions

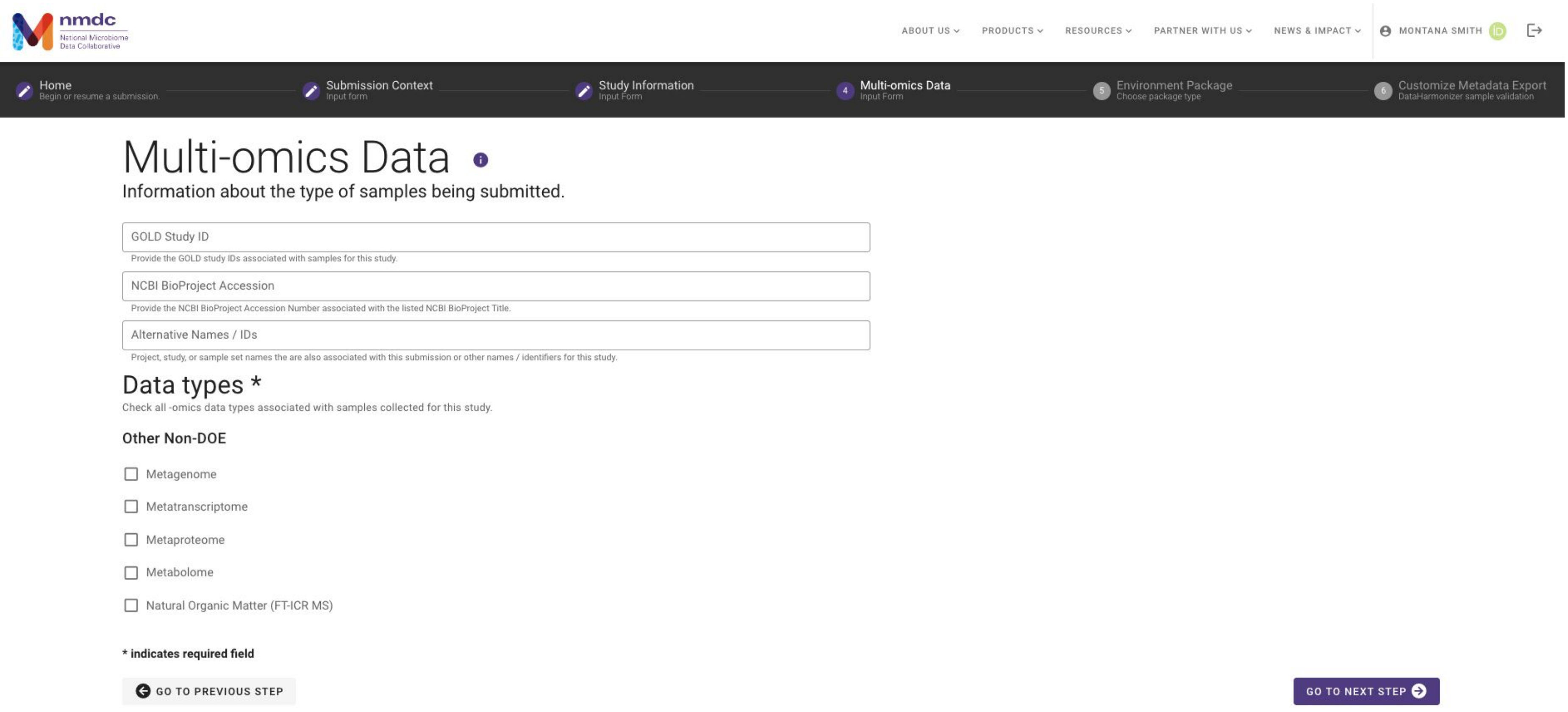
CRediT Roles : https://credit.niso.org/

Permission Levels

- Author/Owner, Editor, Metadata Contributor, Viewer

The NMDC Submission Portal

nmdc National Microbiome Data Collaborative



nmdc National Microbiome Data Collaborative
ABOUT US
PRODUCTS
RESOURCES
PARTNER WITH US
NEWS & IMPACT
MONTANA SMITH
Home Begin or resume a submission.
Submission Context Input form
Study Information Input Form
4 Multi-omics Data Input Form
5 Environment Package Choose package type
6 Customize Metadata Export DataHarmonizer sample validation

Multi-omics Data

Information about the type of samples being submitted.

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.

Alternative Names / IDs

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

Data types *

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

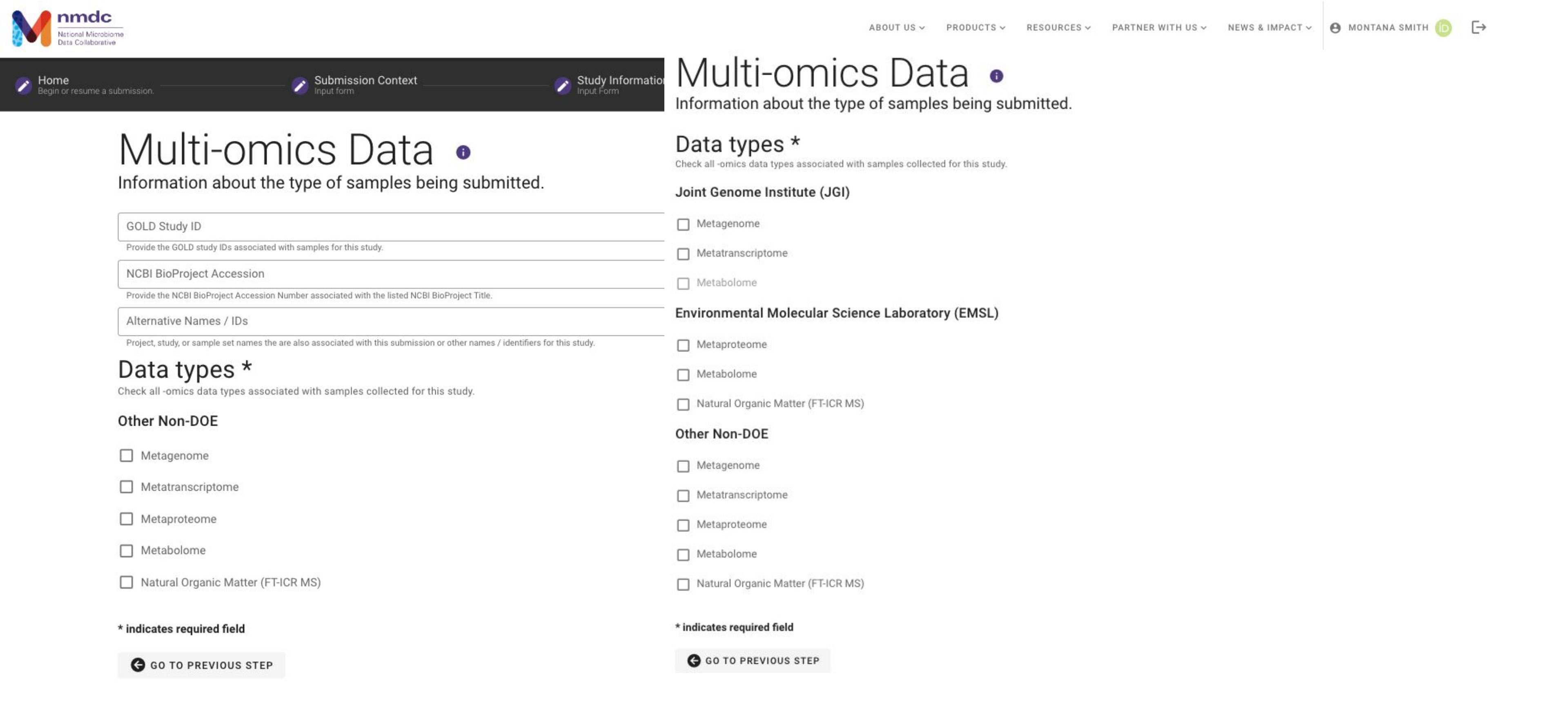
Other Non-DOE

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

*** indicates required field**

The NMDC Submission Portal

nmdc
National Microbiome
Data Collaborative



Multi-omics Data

Information about the type of samples being submitted.

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.

Alternative Names / IDs

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

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

Multi-omics Data

Information about the type of samples being submitted.

Data types *

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

Joint Genome Institute (JGI)

- □ Metagenome
- □ Metatranscriptome
- □ Metabolome

Environmental Molecular Science Laboratory (EMSL)

- □ Metaproteome
- □ Metabolome
- □ Natural Organic Matter (FT-ICR MS)

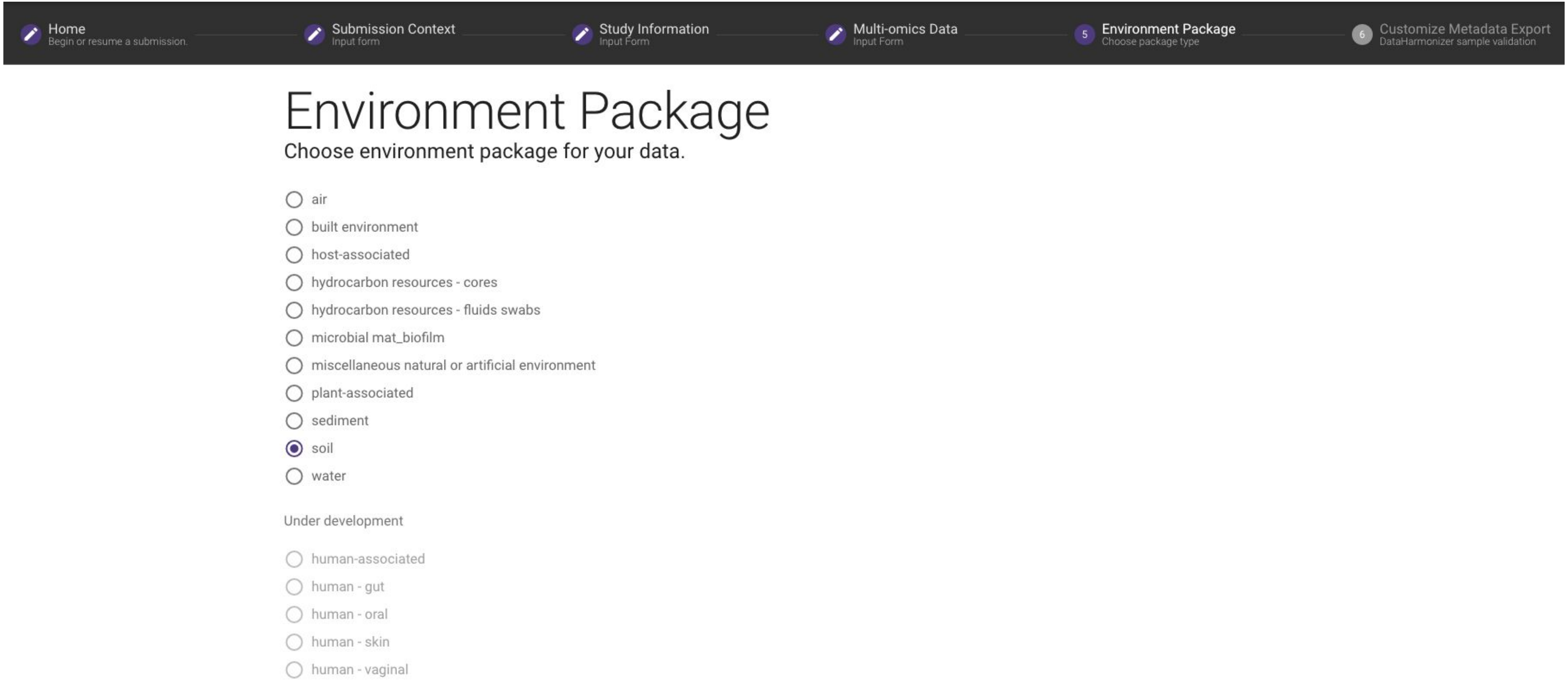
Other Non-DOE

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

*** indicates required field**

GO TO PREVIOUS STEP

The NMDC Submission Portal



Environment Package

Choose environment package for your data.

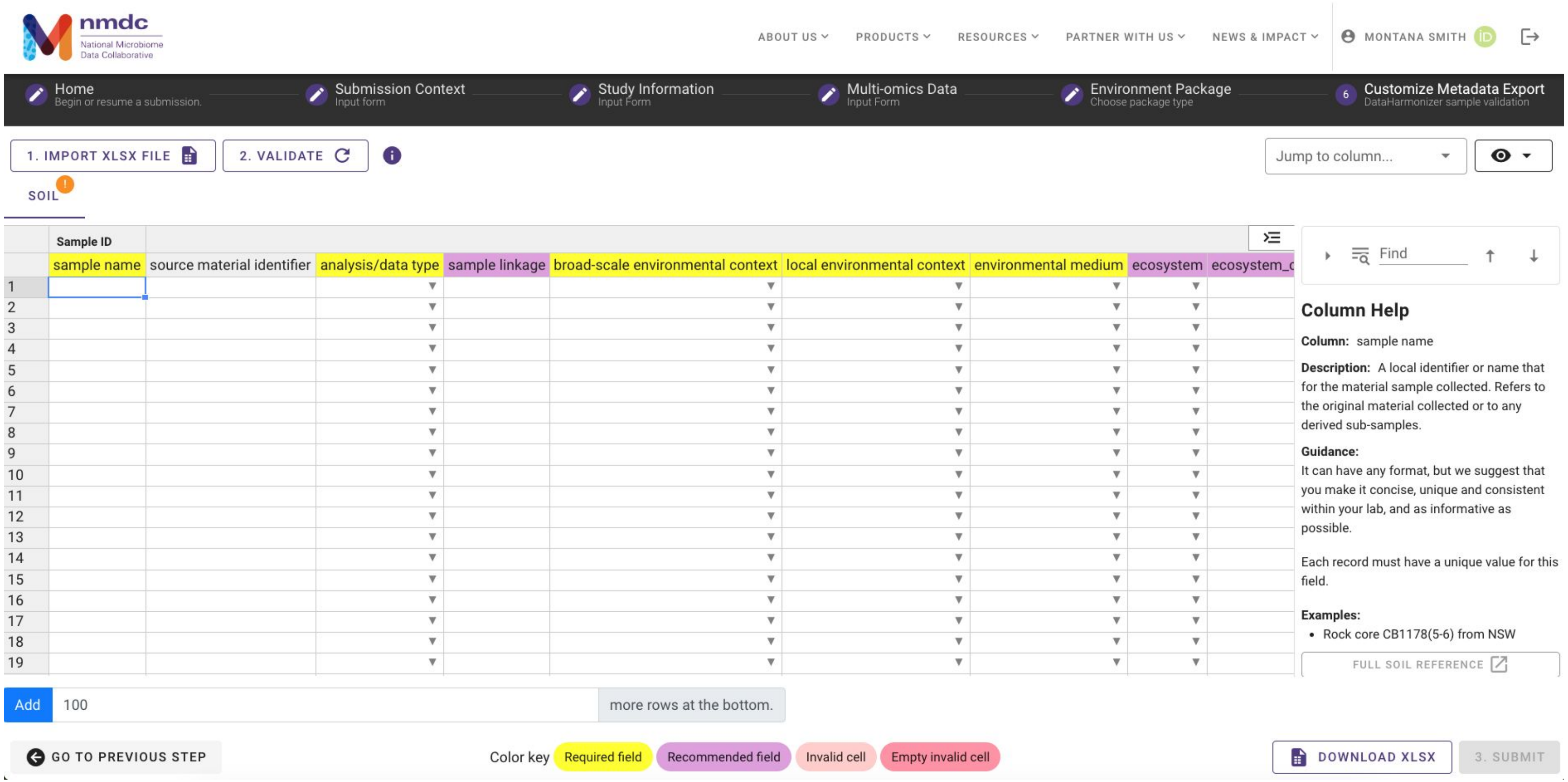
- ○ air
- ○ built environment
- ○ host-associated
- ○ hydrocarbon resources - cores
- ○ hydrocarbon resources - fluids swabs
- ○ microbial mat_biofilm
- ○ miscellaneous natural or artificial environment
- ○ plant-associated
- ○ sediment
- ● soil
- ○ water

Under development

- ○ human-associated
- ○ human - gut
- ○ human - oral
- ○ human - skin
- ○ human - vaginal

The NMDC Submission Portal

nmdc
National Microbiome
Data Collaborative



	Sample ID								
	sample name	source material identifier	analysis/data type	sample linkage	broad-scale environmental context	local environmental context	environmental medium	ecosystem	ecosystem_c
1			▼		▼	▼	▼	▼	
2			▼		▼	▼	▼	▼	
3			▼		▼	▼	▼	▼	
4			▼		▼	▼	▼	▼	
5			▼		▼	▼	▼	▼	
6			▼		▼	▼	▼	▼	
7			▼		▼	▼	▼	▼	
8			▼		▼	▼	▼	▼	
9			▼		▼	▼	▼	▼	
10			▼		▼	▼	▼	▼	
11			▼		▼	▼	▼	▼	
12			▼		▼	▼	▼	▼	
13			▼		▼	▼	▼	▼	
14			▼		▼	▼	▼	▼	
15			▼		▼	▼	▼	▼	
16			▼		▼	▼	▼	▼	
17			▼		▼	▼	▼	▼	
18			▼		▼	▼	▼	▼	
19			▼		▼	▼	▼	▼	

Add 100 more rows at the bottom.

Column Help

Column: sample name

Description: A local identifier or name that for the material sample collected. Refers to the original material collected or to any derived sub-samples.

Guidance:
It can have any format, but we suggest that you make it concise, unique and consistent within your lab, and as informative as possible.

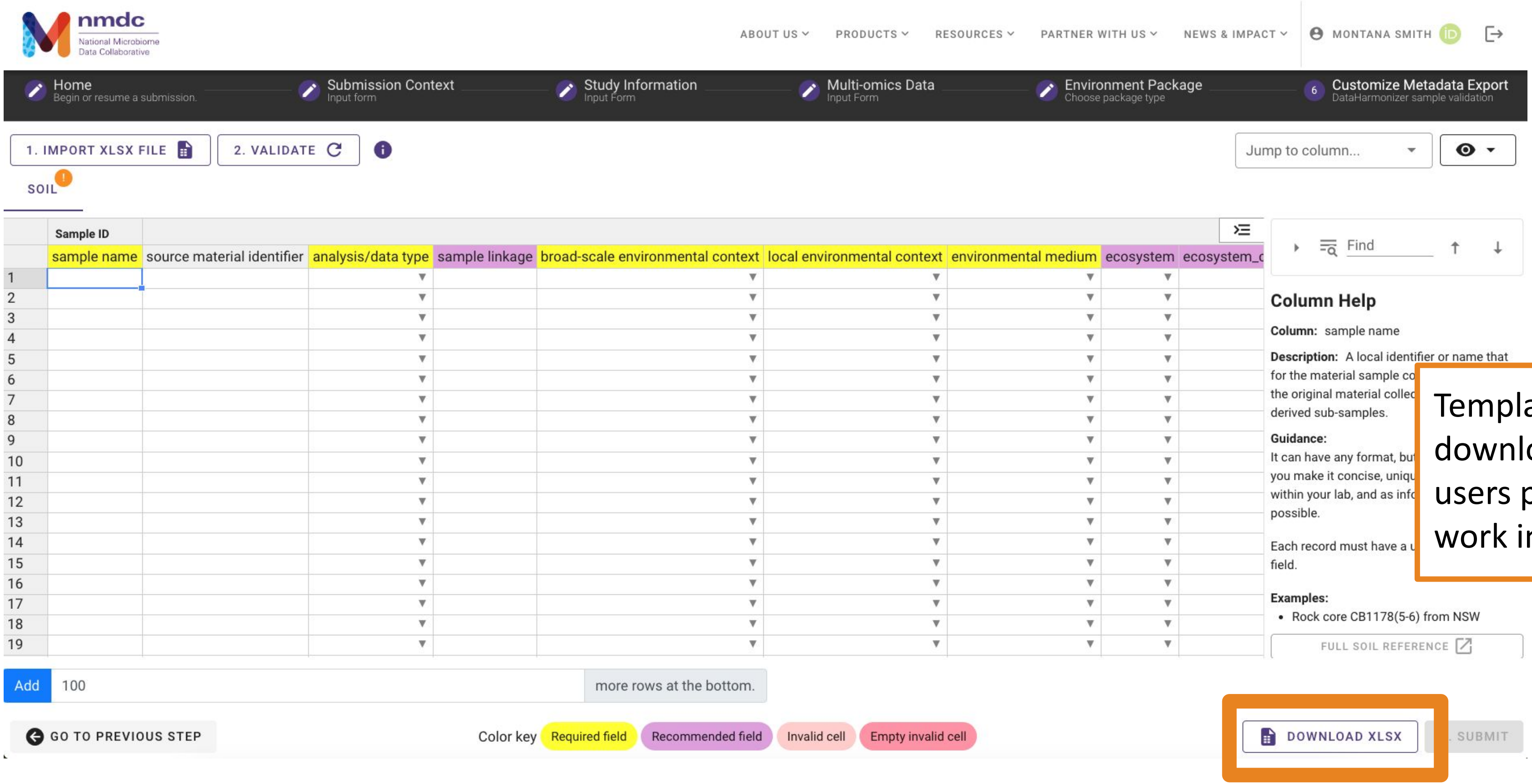
Each record must have a unique value for this field.

Examples:

- Rock core CB1178(5-6) from NSW

Color key: Required field | Recommended field | Invalid cell | Empty invalid cell

The NMDC Submission Portal



Home – Begin or resume a submission.
Submission Context – Input form
Study Information – Input Form
Multi-omics Data – Input Form
Environment Package – Choose package type
6 Customize Metadata Export – DataHarmonizer sample validation

	Sample ID								
	sample name	source material identifier	analysis/data type	sample linkage	broad-scale environmental context	local environmental context	environmental medium	ecosystem	ecosystem_c
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									

Column Help

Column: sample name

Description: A local identifier or name that for the material sample co the original material collec derived sub-samples.

Guidance:
It can have any format, but you make it concise, uniqu within your lab, and as info possible.

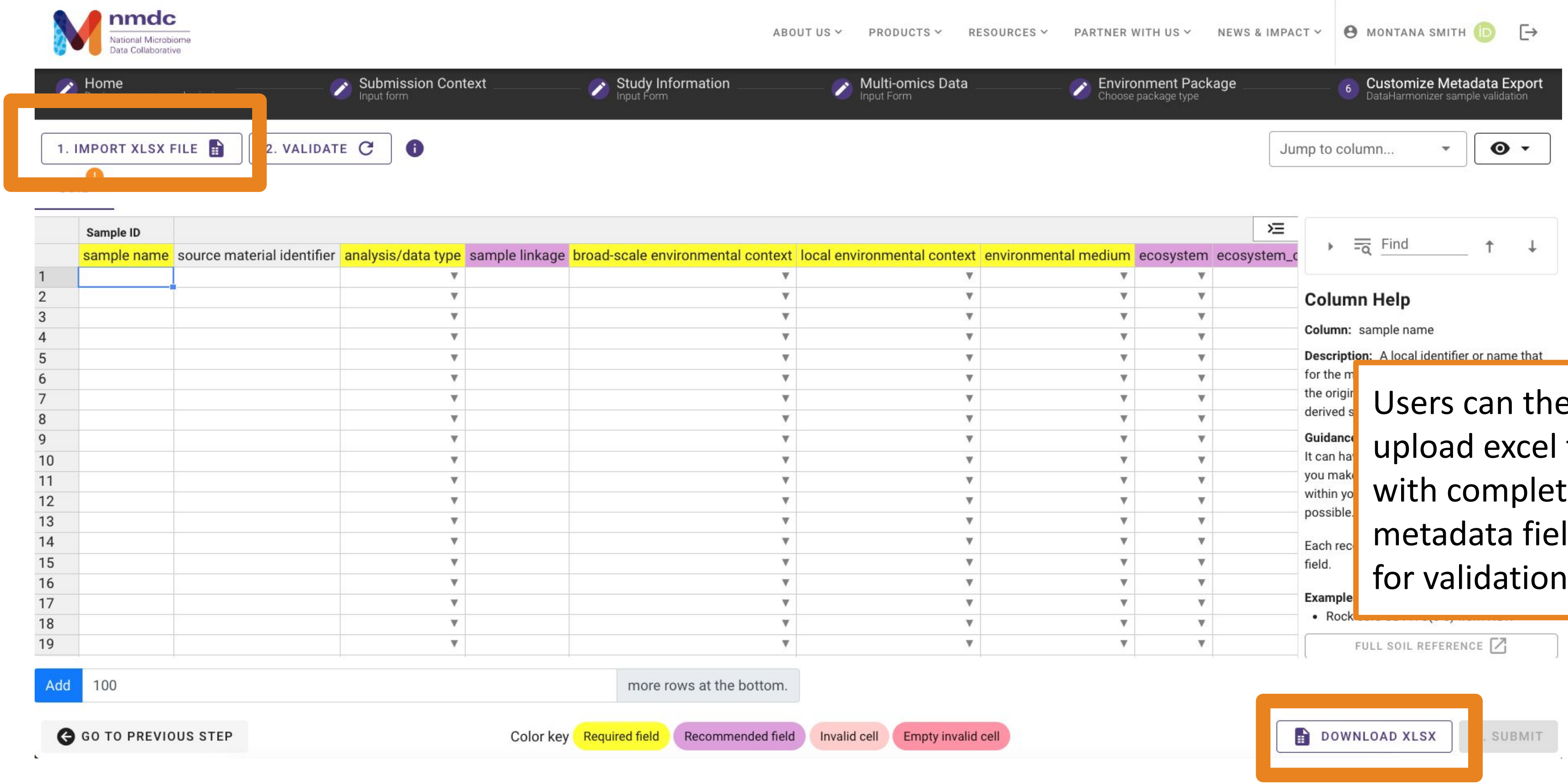
Each record must have a u field.

Examples:
- Rock core CB1178(5-6) from NSW

Color key: Required field | Recommended field | Invalid cell | Empty invalid cell

Templates are downloadable if users prefer to work in Excel

The NMDC Submission Portal



Column Help

Column: sample name

Description: A local identifier or name that for the m the origi derived s

Guidance It can ha you mak within yo possible.

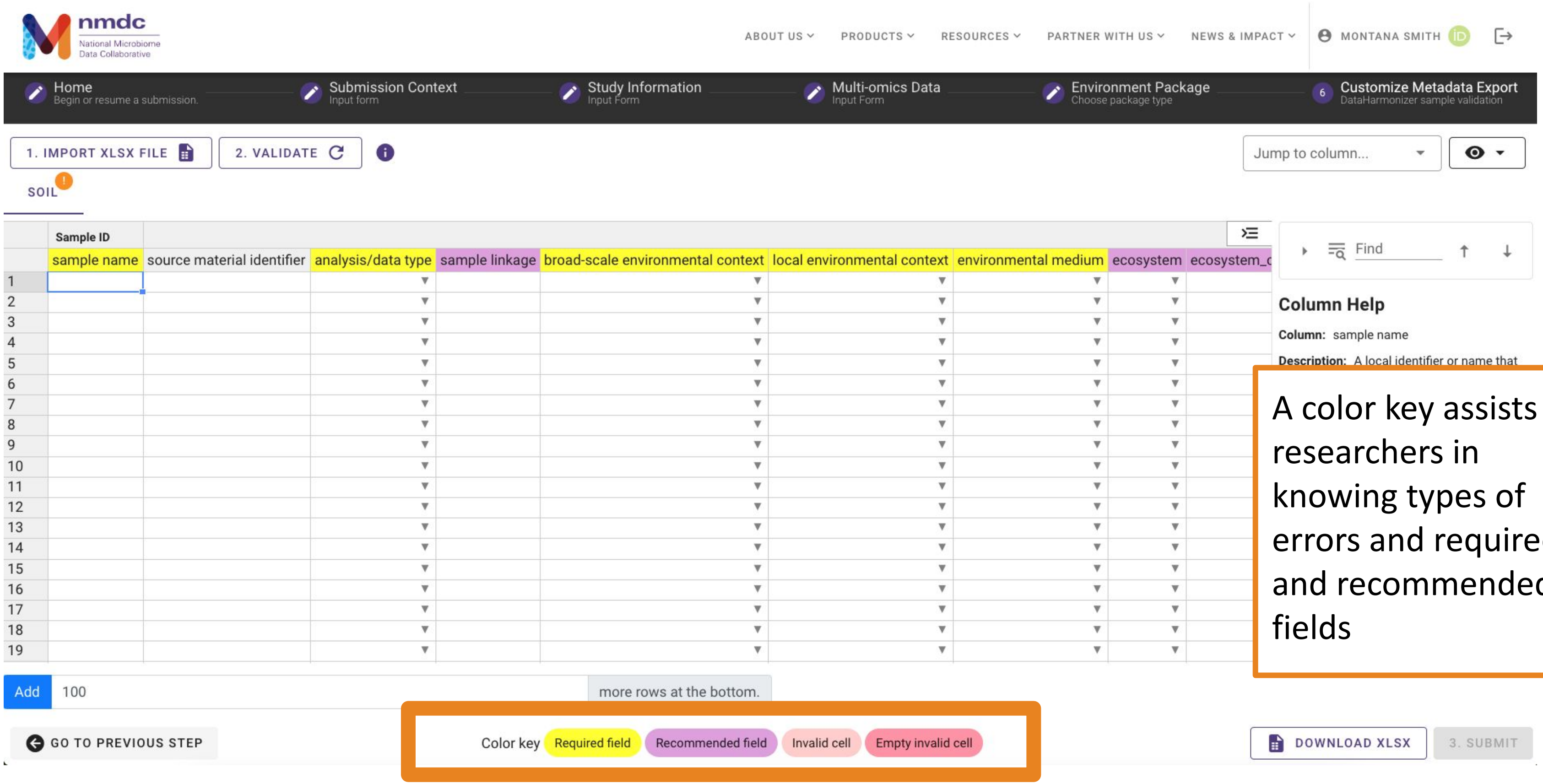
Each rec field.

Example

- Rock

Users can then upload excel files with completed metadata fields for validation

The NMDC Submission Portal



	Sample ID								
	sample name	source material identifier	analysis/data type	sample linkage	broad-scale environmental context	local environmental context	environmental medium	ecosystem	ecosystem_c
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									

Column Help

Column: sample name

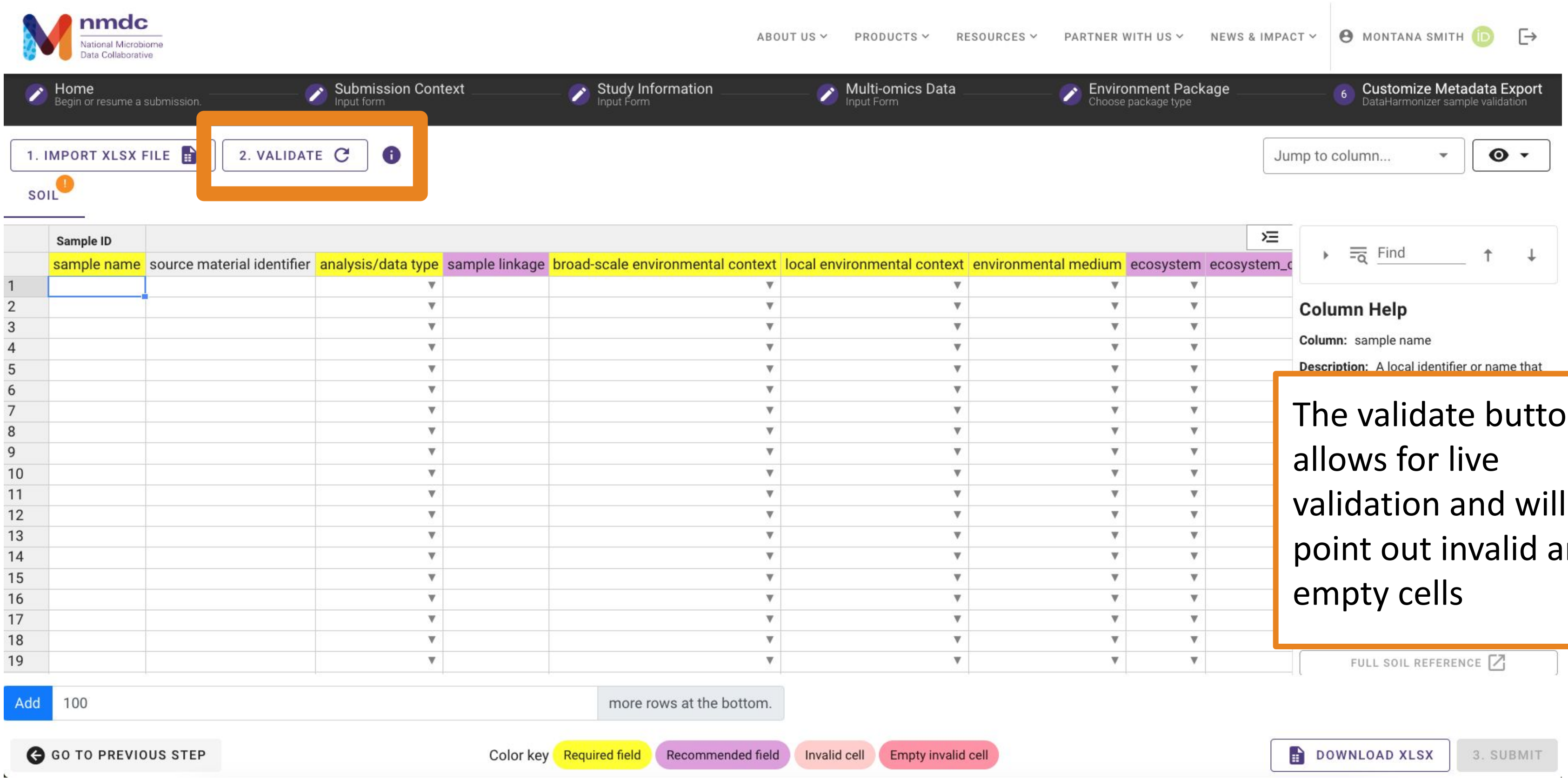
Description: A local identifier or name that

Color key Required field Recommended field Invalid cell Empty invalid cell

A color key assists researchers in knowing types of errors and required and recommended fields

The NMDC Submission Portal

nmdc
National Microbiome Data Collaborative



	Sample ID									
	sample name	source material identifier	analysis/data type	sample linkage	broad-scale environmental context	local environmental context	environmental medium	ecosystem	ecosystem_c
1			▼		▼	▼	▼	▼	
2			▼		▼	▼	▼	▼	
3			▼		▼	▼	▼	▼	
4			▼		▼	▼	▼	▼	
5			▼		▼	▼	▼	▼	
6			▼		▼	▼	▼	▼	
7			▼		▼	▼	▼	▼	
8			▼		▼	▼	▼	▼	
9			▼		▼	▼	▼	▼	
10			▼		▼	▼	▼	▼	
11			▼		▼	▼	▼	▼	
12			▼		▼	▼	▼	▼	
13			▼		▼	▼	▼	▼	
14			▼		▼	▼	▼	▼	
15			▼		▼	▼	▼	▼	
16			▼		▼	▼	▼	▼	
17			▼		▼	▼	▼	▼	
18			▼		▼	▼	▼	▼	
19			▼		▼	▼	▼	▼	

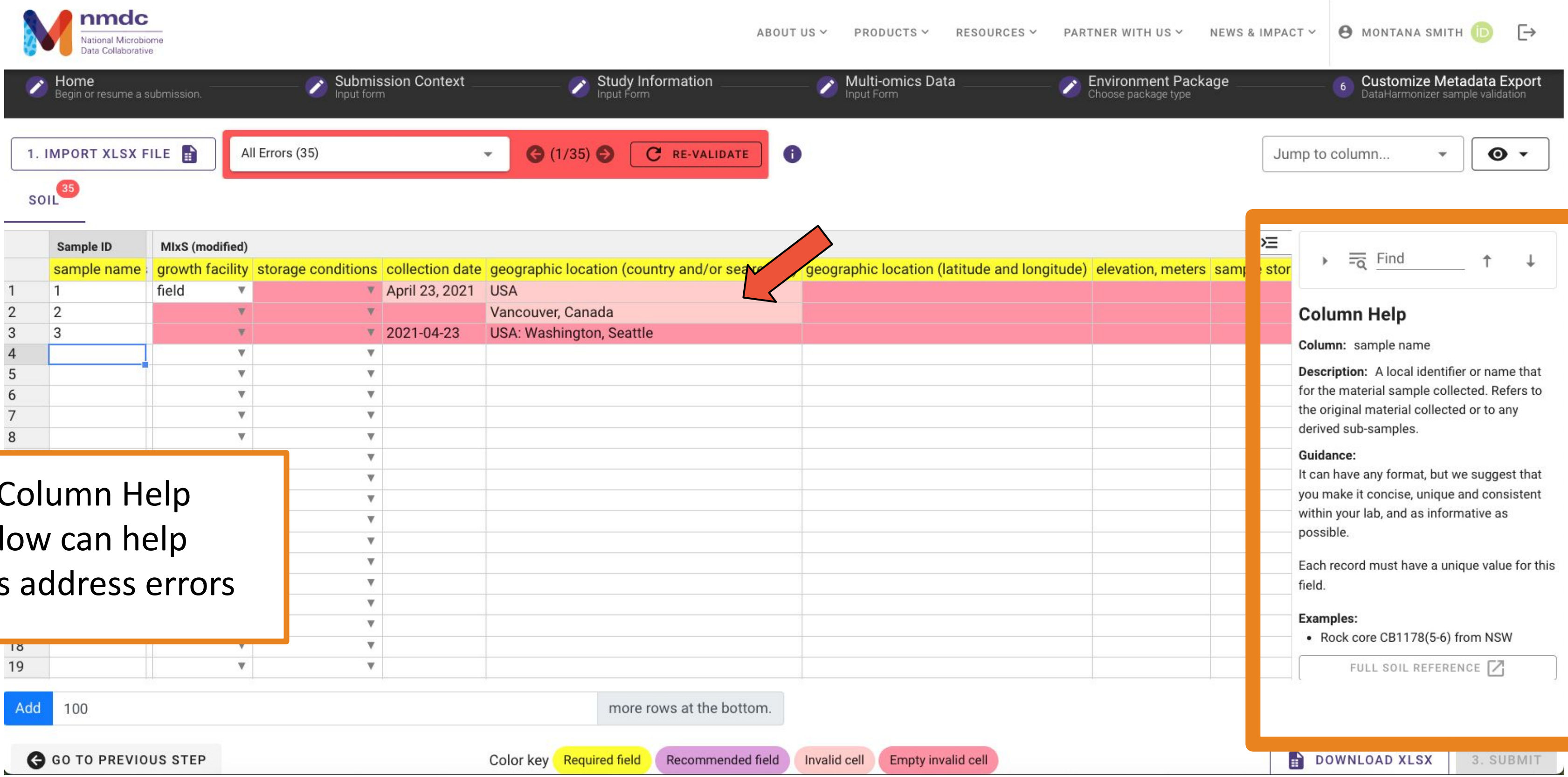
Column Help

Column: sample name

Description: A local identifier or name that

The validate button allows for live validation and will point out invalid and empty cells

The NMDC Submission Portal



	Sample ID	MIxS (modified)							
	sample name	growth facility	storage conditions	collection date	geographic location (country and/or sea	geographic location (latitude and longitude)	elevation, meters	samp	stor
1	1	field		April 23, 2021	USA				
2	2				Vancouver, Canada				
3	3			2021-04-23	USA: Washington, Seattle				
4									
5									
6									
7									
8									
18									
19									

Column Help

Column: sample name

Description: A local identifier or name that for the material sample collected. Refers to the original material collected or to any derived sub-samples.

Guidance:
It can have any format, but we suggest that you make it concise, unique and consistent within your lab, and as informative as possible.

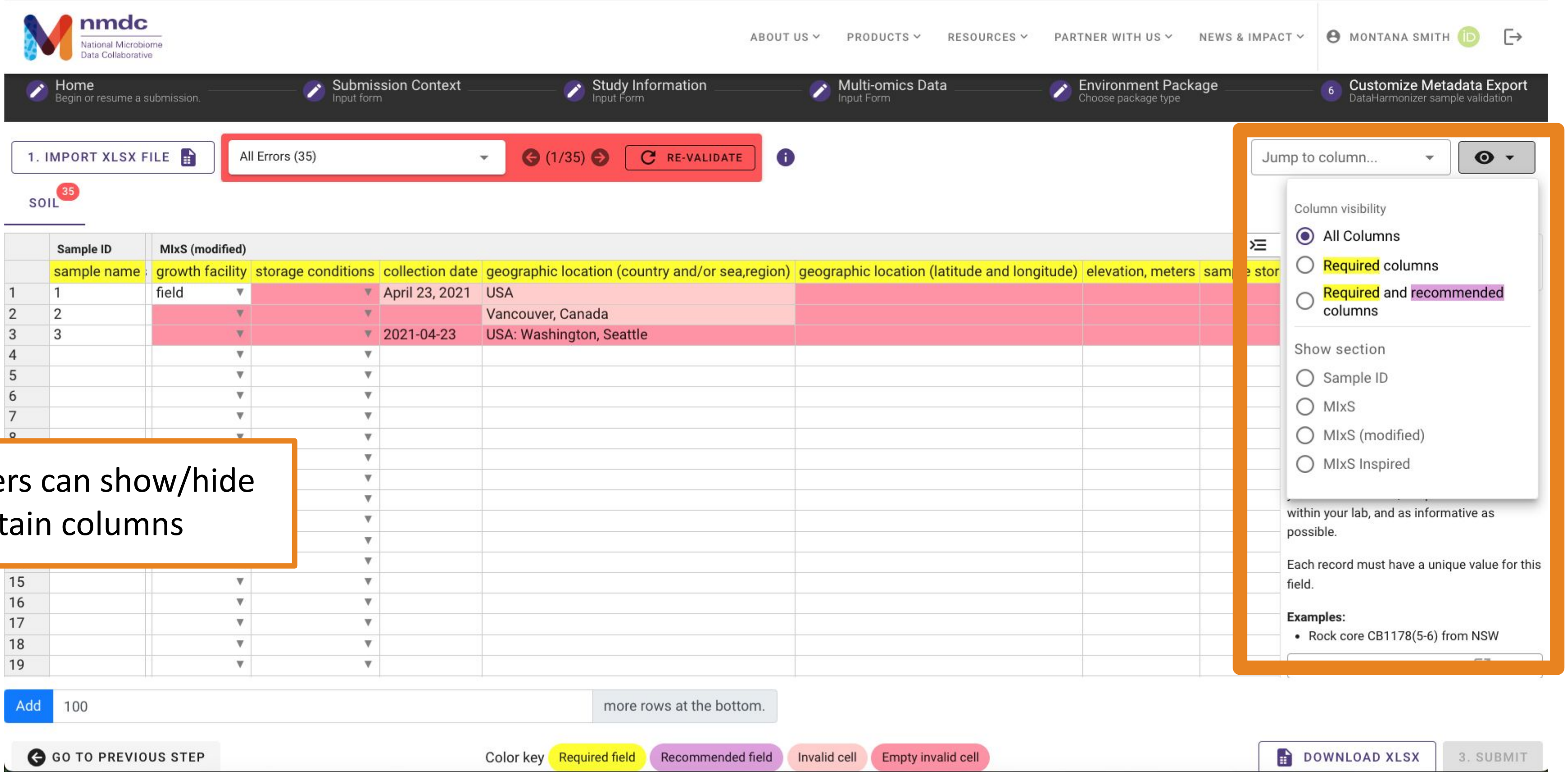
Each record must have a unique value for this field.

Examples:
- Rock core CB1178(5-6) from NSW

Color key: Required field | Recommended field | Invalid cell | Empty invalid cell

The Column Help window can help users address errors

The NMDC Submission Portal



	Sample ID	MIxS (modified)							
	sample name	growth facility	storage conditions	collection date	geographic location (country and/or sea,region)	geographic location (latitude and longitude)	elevation, meters	sam	stor
1	1	field ▼	▼	April 23, 2021	USA				
2	2	▼	▼		Vancouver, Canada				
3	3	▼	▼	2021-04-23	USA: Washington, Seattle				
4		▼	▼						
5		▼	▼						
6		▼	▼						
7		▼	▼						
8		▼	▼						
15		▼	▼						
16		▼	▼						
17		▼	▼						
18		▼	▼						
19		▼	▼						

Column visibility

- ● All Columns
- ○ Required columns
- ○ Required and recommended columns

Show section

- ○ Sample ID
- ○ MIxS
- ○ MIxS (modified)
- ○ MIxS Inspired

within your lab, and as informative as possible.

Each record must have a unique value for this field.

Examples:

- Rock core CB1178(5-6) from NSW

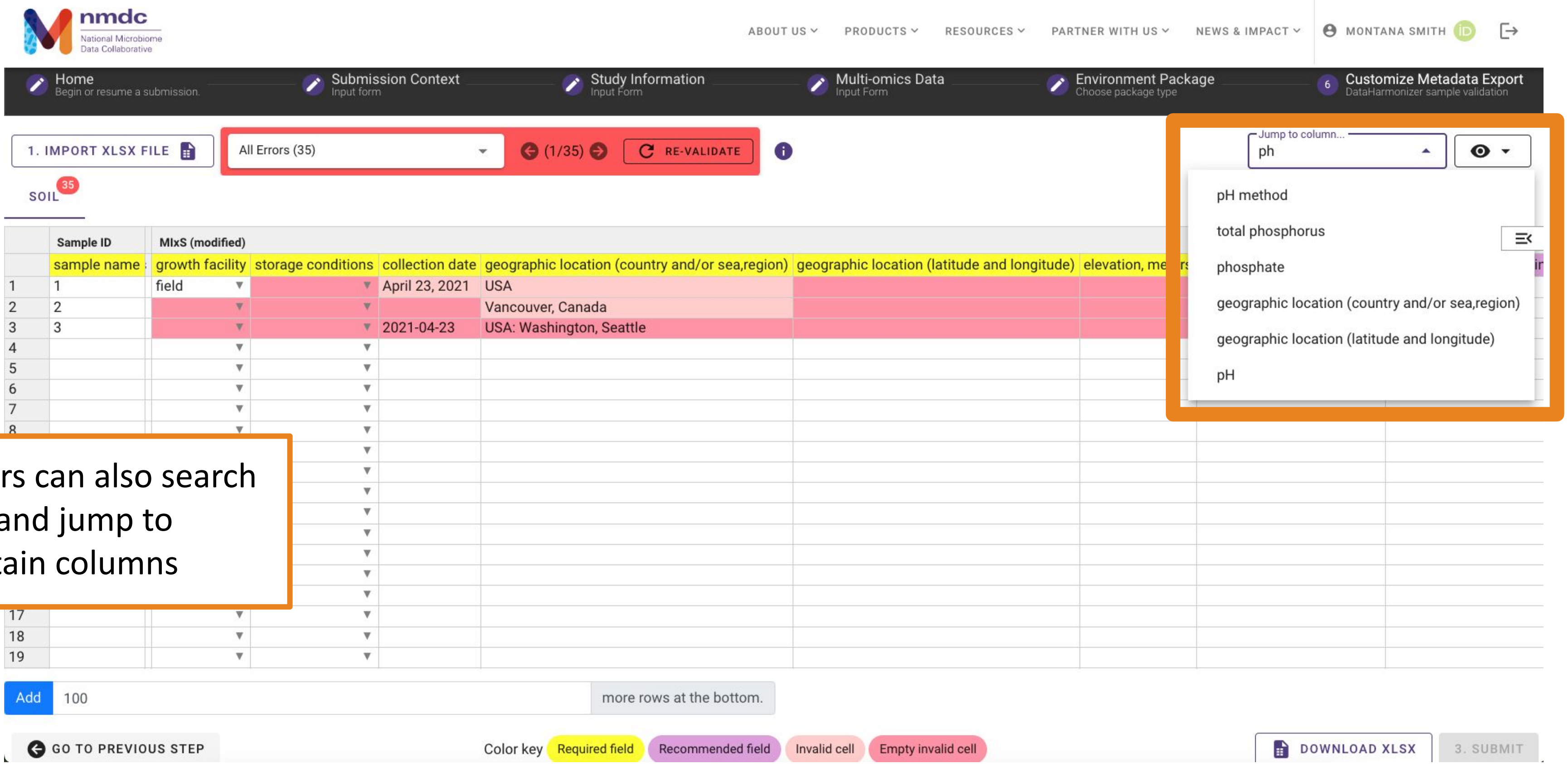
Add 100 more rows at the bottom.

Color key: Required field | Recommended field | Invalid cell | Empty invalid cell

Users can show/hide certain columns

The NMDC Submission Portal

nmdc
National Microbiome Data Collaborative



nmdc National Microbiome Data Collaborative

ABOUT US ⌄ PRODUCTS ⌄ RESOURCES ⌄ PARTNER WITH US ⌄ NEWS & IMPACT ⌄ MONTANA SMITH

Home — Begin or resume a submission.
Submission Context — Input form
Study Information — Input Form
Multi-omics Data — Input Form
Environment Package — Choose package type
6 Customize Metadata Export — DataHarmonizer sample validation

1. IMPORT XLSX FILE | All Errors (35) | (1/35) | RE-VALIDATE

SOIL 35

Jump to column...: ph

- pH method
- total phosphorus
- phosphate
- geographic location (country and/or sea,region)
- geographic location (latitude and longitude)
- pH

	Sample ID	MIxS (modified)					
	sample name	growth facility	storage conditions	collection date	geographic location (country and/or sea,region)	geographic location (latitude and longitude)	elevation, me
1	1	field		April 23, 2021	USA		
2	2				Vancouver, Canada		
3	3			2021-04-23	USA: Washington, Seattle		
4							
5							
6							
7							
8							
17							
18							
19							

Add 100 more rows at the bottom.

GO TO PREVIOUS STEP

Color key: Required field | Recommended field | Invalid cell | Empty invalid cell

DOWNLOAD XLSX | 3. SUBMIT

Users can also search for and jump to certain columns

nmdc

National Microbiome Data Collaborative

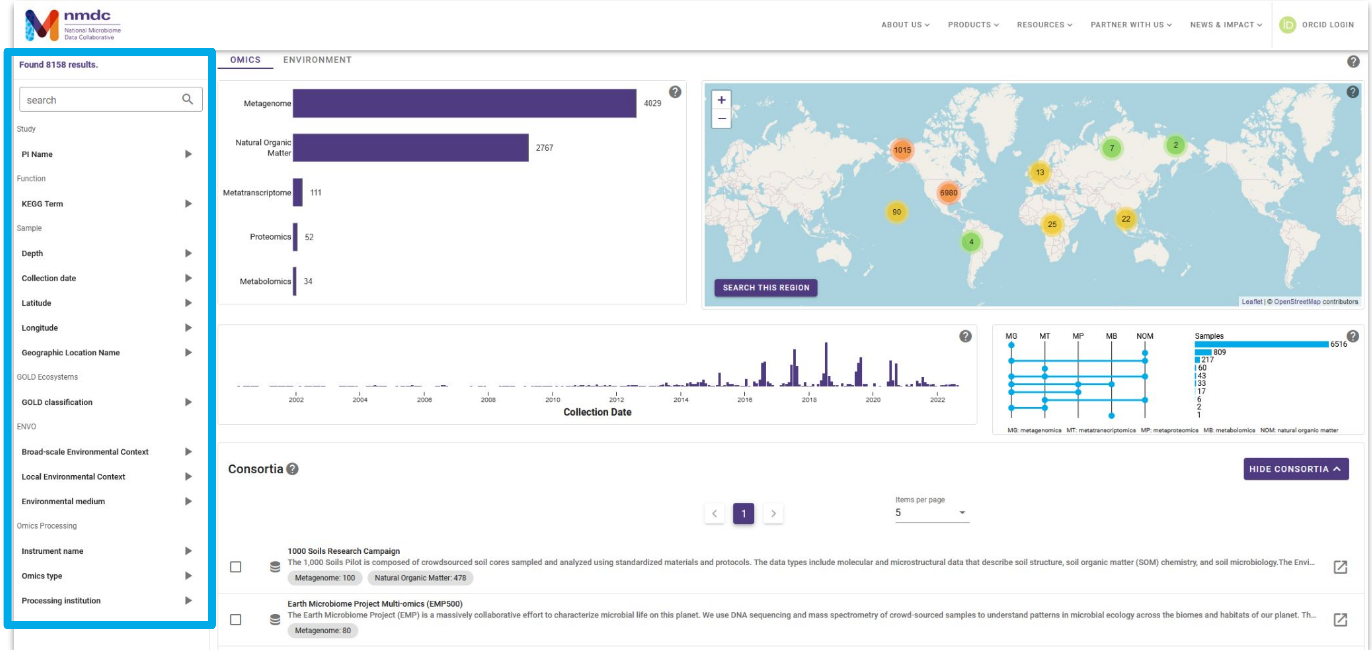
NMDC Data Portal

NMDC Data Portal

nmdc
National Microbiome Data Collaborative

Data Portal & API

Access and discovery of microbiome information



□ **1000 Soils Research Campaign**
The 1,000 Soils Pilot is composed of crowdsourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology.The Envi...
Metagenome: 100 | Natural Organic Matter: 478

□ **Earth Microbiome Project Multi-omics (EMP500)**
The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ecology across the biomes and habitats of our planet. Th...
Metagenome: 80

Filter datasets by standardized metadata terms

nmdc

National Microbiome Data Collaborative

Questions?

nmdc

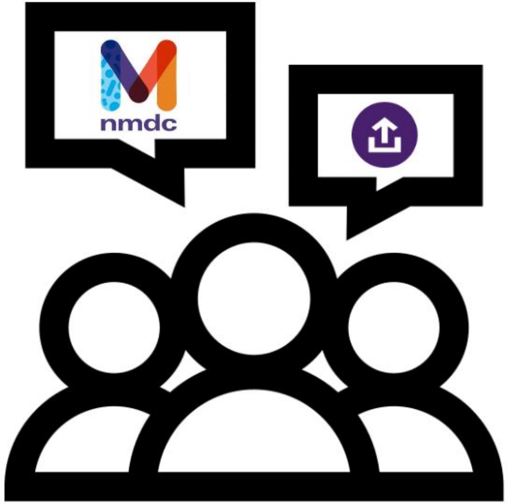
National Microbiome
Data Collaborative

Activities

Submission Portal Activity

- Select ‘**Create New Submission**’
- For Submission Context, select ‘No’ for if the data has been generated for this study
- Select ‘JGI’ then select ‘Other’ and type something along the lines of ‘test submission’
- Select Go to Next Step
- Name the study: “Ambassador Training: Validation activity”
- Put a fake email address or your own email address into the Principal Investigator Email field
- No need to fill out non-required information on this page
- Select Go to Next Step
- Select metagenome under JGI and type a fake study ID (123456)
- Select Go to Next Step
- Select the Soil Environment Package
- Select Go to Next Step
- <u>**Pause**</u> for discussion
- Upload the xlsx file, select the ‘Validate’ button and try to fix all of the invalid fields!

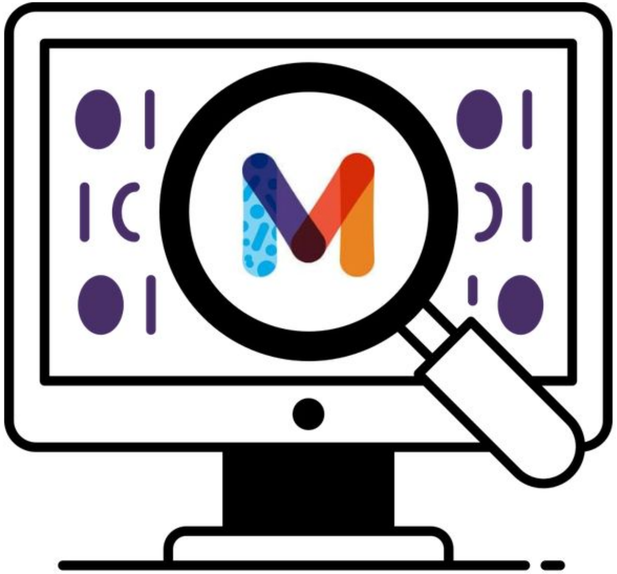
Example Break Out Activity



A more challenging example activity

- Complete study and data information using a written study design
  - Ambassadors can write out a paragraph of study design (we have examples you can use)
  - Attendees read the paragraph and identify relevant metadata fields
  - Attendees complete the study and data information sections of the Submission Portal

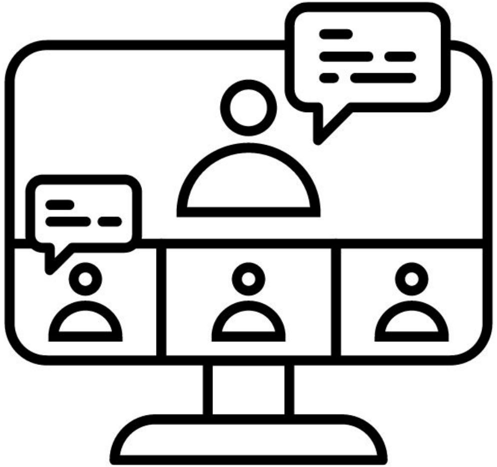
Discussion



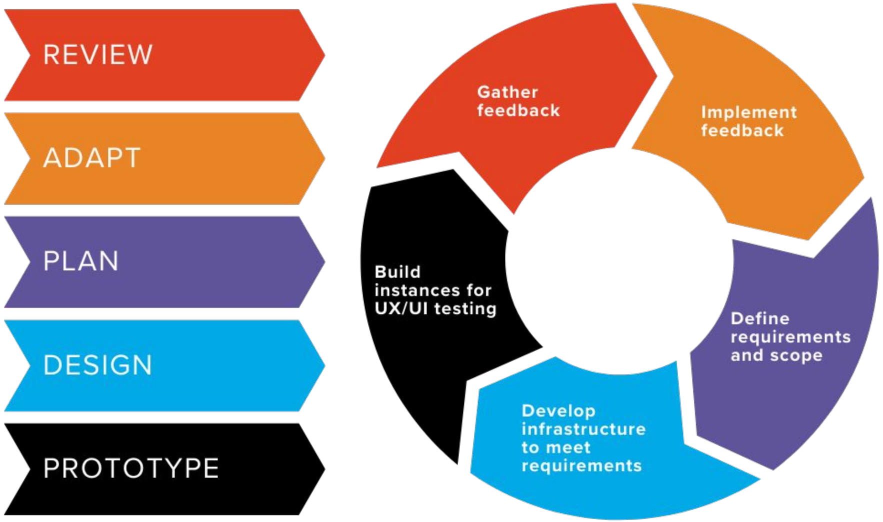
- What do you currently use to track metadata and sample information?
- Which metadata standards or templates (if any) are your group currently using?
- What did you like about the NMDC Submission Portal?
- What could be improved about the Submission Portal?
- Did you encounter any bugs in the Submission Portal?

Submission Portal Experience

- Do you have projects you would like to track in the Submission Portal?
  - Walk through your Submission Portal experience with our team!
  - Please add your name to shared notes to participate
- Live feedback from your workshops
  - Features people like/don't like
  - Things that aren't clear during the workshop
  - Bugs that come up while people are doing activities
- General reporting google form for issues: https://forms.gle/yxu9gkbufPigtbrB8



Community-centered design process


REVIEW
ADAPT
PLAN
DESIGN
PROTOTYPE
Gather feedback
Implement feedback
Define requirements and scope
Develop infrastructure to meet requirements
Build instances for UX/UI testing

34
User research participants

321
Insights from user interviews

120
Action items from insights

Submission Portal Feedback

- Updates to how the templates are accessed and downloaded
- Feature and UI updates to improve experience with interface
- Help and tutorial guidance updates

nmdc

National Microbiome Data Collaborative

Questions?

NMDC Resources

nmdc
National Microbiome Data Collaborative

Website: https://microbiomedata.org/
Data Portal: https://data.microbiomedata.org/
Submission Portal: https://data.microbiomedata.org/submission/home
NMDC EDGE: https://nmdc-edge.org/home
Github: https://github.com/microbiomedata
Docker Hub: https://hub.docker.com/u/microbiomedata
Documentation: https://nmdc-documentation.readthedocs.io/en/latest/overview/nmdc_overview.html
YouTube: https://www.youtube.com/channel/UCyBqKc46NQZ_YgZlKGYegIw/featured

Get involved!

Sign up for our newsletter
microbiomedata.org

Become a NMDC Champion
bit.ly/champions-program

Find us on X/Twitter
@microbiomedata

Find us on LinkedIn
https://bit.ly/NMDC_LinkedIn

Find us on Instagram
@microbiomedata

Read more about the NMDC

Kelliher JM *et al.* Cohort-based learning for microbiome research community standards. *Nat Microbiol* (2023). doi.org/10.1038/s41564-023-01361-7.

Hu B, Canon S, Eloe-Fadrosh EA, et al.. Challenges in Bioinformatics Workflows for Processing Microbiome Omics Data at Scale. *Front Bioinform*. 1:826370. (2022) doi: 10.3389/fbinf.2021.826370.

Eloe-Fadrosh EA *et al*. The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res.* 7;60(D1):D828–D836. (2022) doi: 10.1093/nar/gkab990.

Wood-Charlson, E.M., Anubhav, Auberry, D. *et al.* The National Microbiome Data Collaborative: enabling microbiome science. *Nat Rev Microbiol* **18,** 313–314 (2020). doi.org/10.1038/s41579-020-0377-0

Vangay, P *et al.* Microbiome metadata standards: Report of the National Microbiome Data Collaborative's workshop and follow-on activities. *mSystems* **6**, e01194-20 (2021). doi.org/10.1128/mSystems.01194-20

nmdc

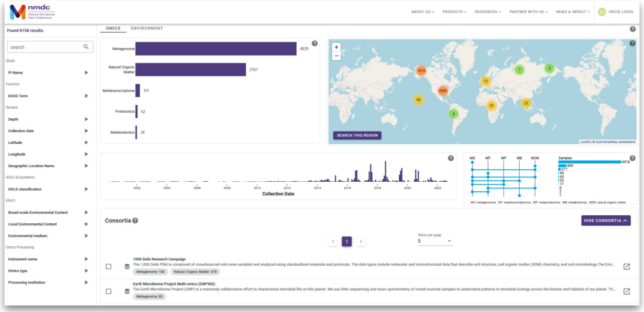
National Microbiome
Data Collaborative

Data Stewardship and NMDC Data Portal

2024 Ambassador Cohort

Purpose of this training

- Provide an overview of data stewardship and relevant guiding principles
- Discuss the importance of proper data stewardship for the future of microbiome data
- Provide template slides for events
- Introduce audience to the NMDC Data Portal
  - The hands-on scavenger hunt can be used (or modified and used) for use in your events too!



Found 8158 results.

1000 Soils Research Campaign
The 1,000 Soils Pilot is composed of crowdsourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology.The Envi...
Metagenome: 100 Natural Organic Matter: 478

Earth Microbiome Project Multi-omics (EMP500)
The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ecology across the biomes and habitats of our planet. Th...
Metagenome: 80

Findable
Accessible
Interoperable
Reusable

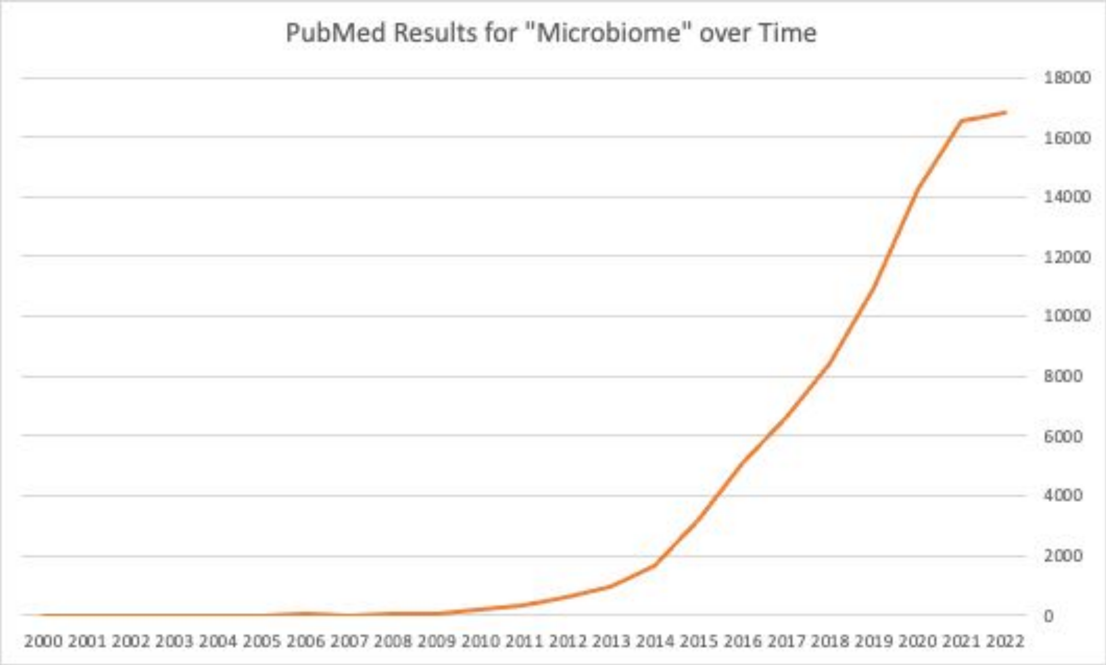
nmdc

National Microbiome Data Collaborative

Data Stewardship & FAIR Data

Microbiome Data

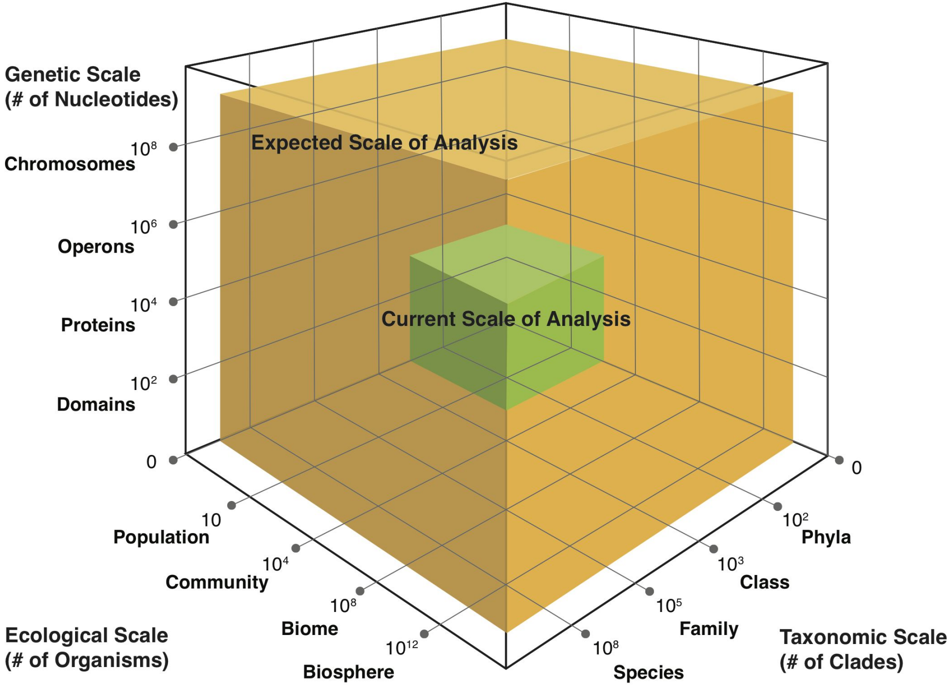
Cover credit: Photo courtesy of Jeremy Snyder
This page credit: Roy Kaltschmidt, Berkeley Lab

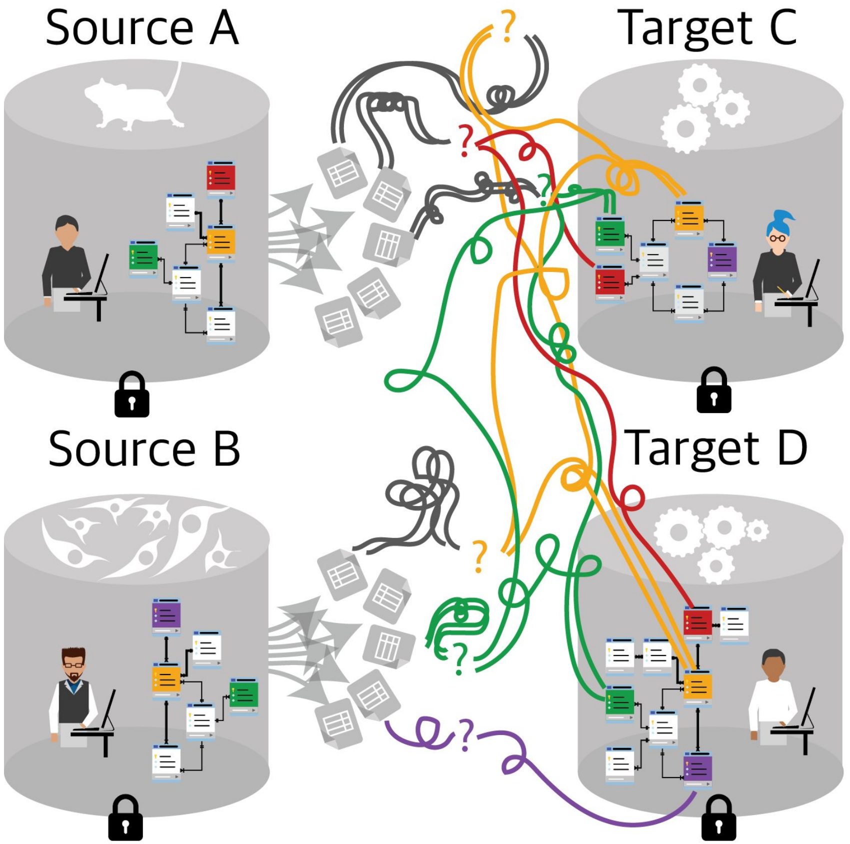


The immense scale of omics data

nmdc
National Microbiome Data Collaborative

Advances in sequencing and omics technologies have **far outpaced** data infrastructure

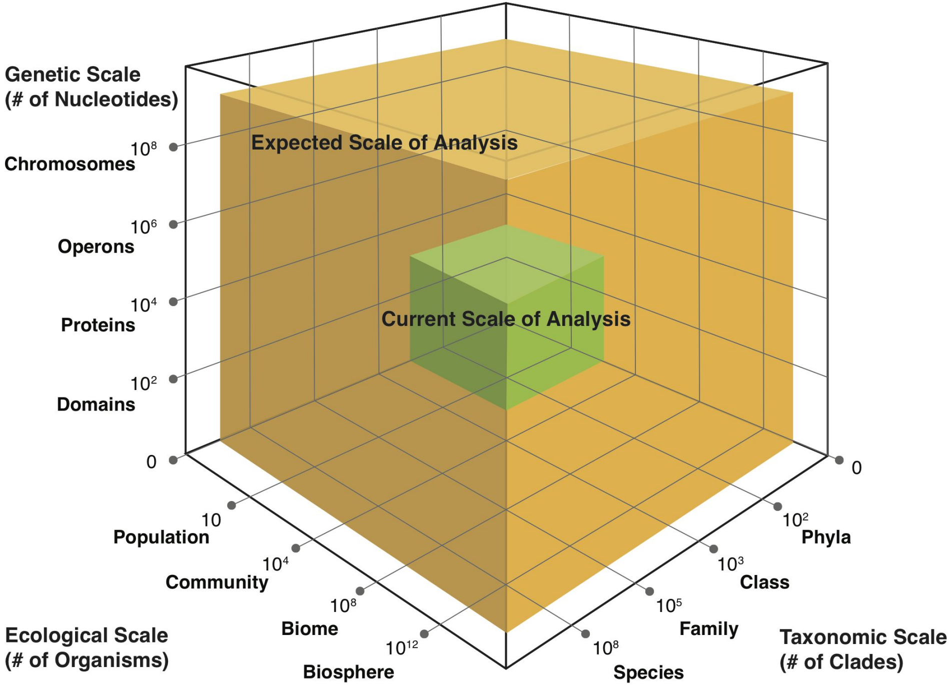


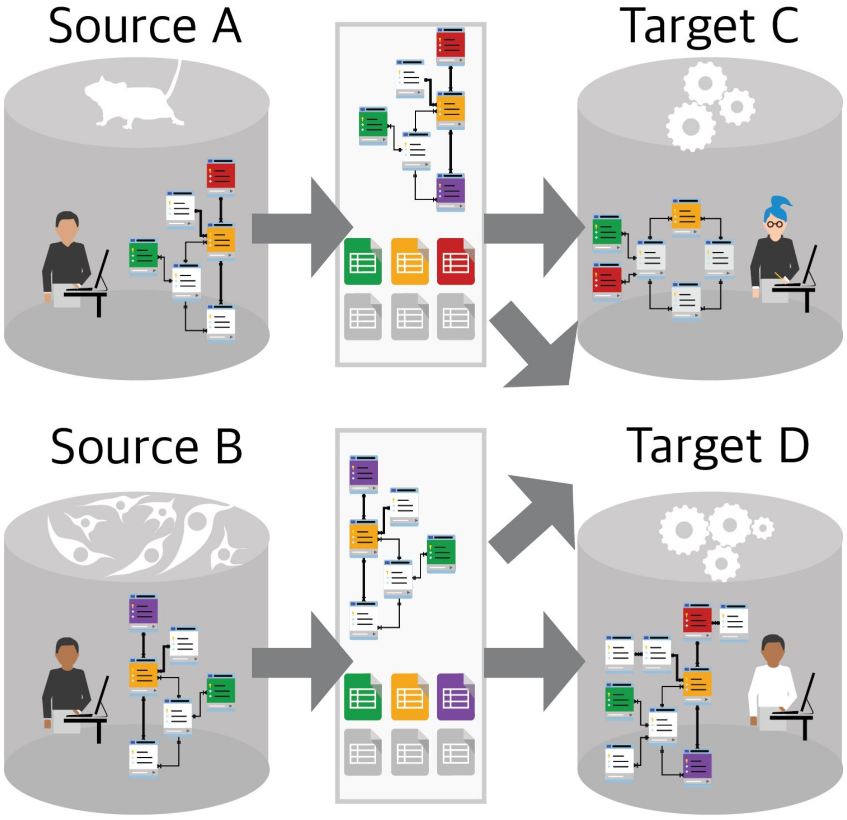


The immense scale of omics data

nmdc
National Microbiome Data Collaborative

Advances in sequencing and omics technologies have **far outpaced** data infrastructure





Data Stewardship

nmdc
National Microbiome
Data Collaborative

Aspects of data stewardship:

- Data generation and acquisition
- Data storage
- Data processing
- Data publication, release
- Data reuse

Everyone’s responsibility to ensure data is properly managed, trustworthy, and FAIR

Accountability in data stewardship

- All researchers
- Funders
- User facilities
- Publishers
- Societies
- Institutions
- Data storage facilities

Important to implement data stewardship best practices as early in the research process as possible

Why should you care about data management and data stewardship?

nmdc
National Microbiome Data Collaborative

Streamlines your research process

- Less data loss and waste, easier to find information
- Can answer new scientific questions, easily share data
- Publication of data
- Data preservation in the future

Makes data accessible for others

- Collecting and providing data and metadata allows other researchers to understand your full study context for data reuse
- Increased scientific exposure

Ensures you receive credit for your work

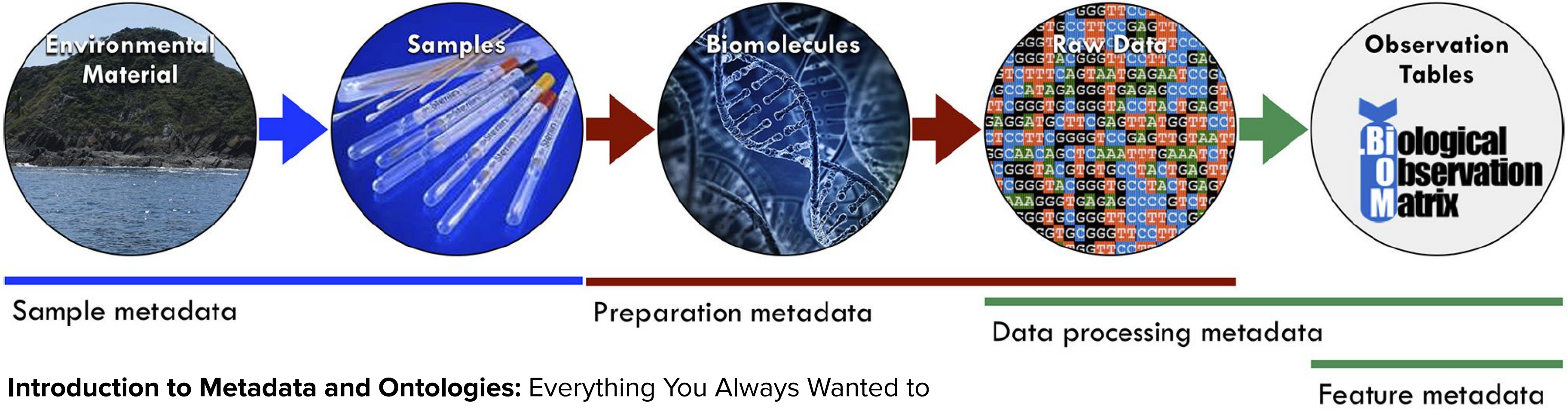
- Datasets can get DOIs, included in publications
- Poorly managed data may not retain provenance or may not be able to be reused, limiting your contributions to your field
- Impact is quantifiable, measurable

Metadata standards

A pillar for data stewardship best practices

nmdc

National Microbiome Data Collaborative



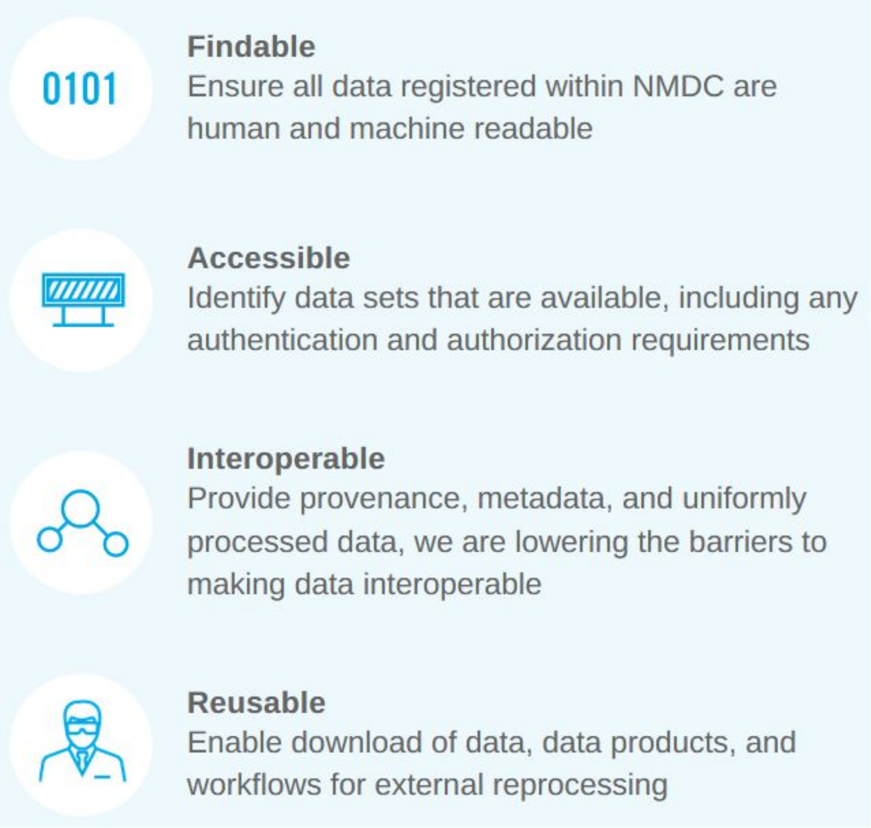
Introduction to Metadata and Ontologies: Everything You Always Wanted to Know About Metadata and Ontologies (But Were Afraid to Ask)

DOI: 10.25979/1607365

FAIR Data

nmdc National Microbiome Data Collaborative

- FAIR is about:
  - Data and metadata
    - Metadata and metadata standards should be articulated and made publicly available to the greatest extent possible
  - Machine-actionability
    - Relevant on all levels of data aggregation
    - Human and machine readable considerations
  - Controlled data access
    - Explicit, well-defined and readily available terms and conditions under which data will be shared or made accessible



Findable
Ensure all data registered within NMDC are human and machine readable

Accessible
Identify data sets that are available, including any authentication and authorization requirements

Interoperable
Provide provenance, metadata, and uniformly processed data, we are lowering the barriers to making data interoperable

Reusable
Enable download of data, data products, and workflows for external reprocessing

FAIR guiding principles: https://doi.org/10.1038/sdata.2016.18

Findable

The first step in (re)using data is to find datasets. Metadata and data should be machine-readable and easy to find by the community.

- (Meta)data are assigned a globally unique and persistent identifier
- Data are described with rich metadata
- Metadata clearly and explicitly include the identifier of the data they describe
- (Meta)data are registered or indexed in a searchable resource

Accessible

Once the user finds the required data, they need to know how the data can be accessed, possibly including authentication and authorization.

- (Meta)data are retrievable by their identifier using a standardized communications protocol
- The protocol is open, free, and universally implementable
- The protocol allows for an authentication and authorization procedure, where necessary
- Metadata are accessible, even when the data are no longer available

Interoperable

Data usually needs to be integrated with other data. In addition, data needs to interoperate with applications or workflows for analysis, storage, and processing.

- (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- (Meta)data use vocabularies that follow FAIR principles
- (Meta)data include qualified references to other (meta)data

Reusable

The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

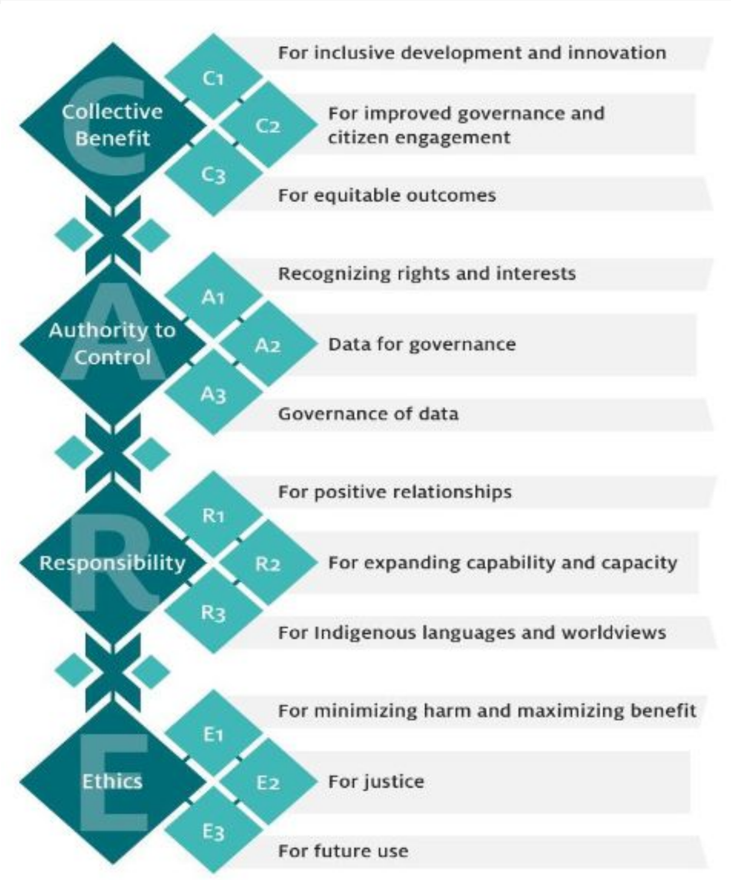
- (Meta)data are richly described with a plurality of accurate and relevant attributes
- (Meta)data are released with a clear and accessible data usage license
- (Meta)data are associated with detailed provenance
- (Meta)data meet domain-relevant community standards

Open Science & Open Data

Why is openly accessible data important?

- Part of **FAIR** data practices
- Scientific advances can be made by synthesizing new and old studies
- Open data can create new research avenues
- Open data increases accessibility to all researchers

CARE principles



Collective Benefit
C1 For inclusive development and innovation
C2 For improved governance and citizen engagement
C3 For equitable outcomes

Authority to Control
A1 Recognizing rights and interests
A2 Data for governance
A3 Governance of data

Responsibility
R1 For positive relationships
R2 For expanding capability and capacity
R3 For Indigenous languages and worldviews

Ethics
E1 For minimizing harm and maximizing benefit
E2 For justice
E3 For future use

- Indigenous data: “Data generated by Indigenous Peoples or by other governments and institutions on or about Indigenous Peoples and territories”
- CARE principles aim for data stewardship practices that align with Indigenous interests and governance needs
  - Making data FAIR while acknowledging power differentials and historical contexts
  - People- and person-oriented, reflecting the crucial role of data in advancing Indigenous innovation and self-determination

CARE principles for Indigenous Data Governance

CARE Principles

Be
FAIR
Findable
Accessible
Interoperable
Reusable
and
CARE
Collective Benefit
Authority to Control
Responsibility
Ethics

The CARE Principles are available in Spanish, Vietnamese, Māori, German, and Khmer

CARE principles for Indigenous Data Governance

IDEA and Data Stewardship

Inclusion means creating environments, large and small, that foster welcoming and belonging.

Diversity refers to the variety of backgrounds, cultures, disciplines, approaches, perspectives, and ways in which we solve problems.

Equity means achieving the aspirational state of 'opportunity parity' for all. To achieve equity, we must increase access to and remove barriers to opportunity, taking into consideration individual needs wherever feasible.

Accountability means taking individual and collective responsibility for our actions, behavior, and impact on others.

How can **data stewardship** incorporate IDEA principles?

- Openly accessible data makes resources more equitable
- Training and educational resources make research more accessible
- From CARE principles - how does your data impact groups involved, who benefits from your data, is anyone excluded who should not be?

https://diversity.lbl.gov/ideaberkeleylab/

nmdc

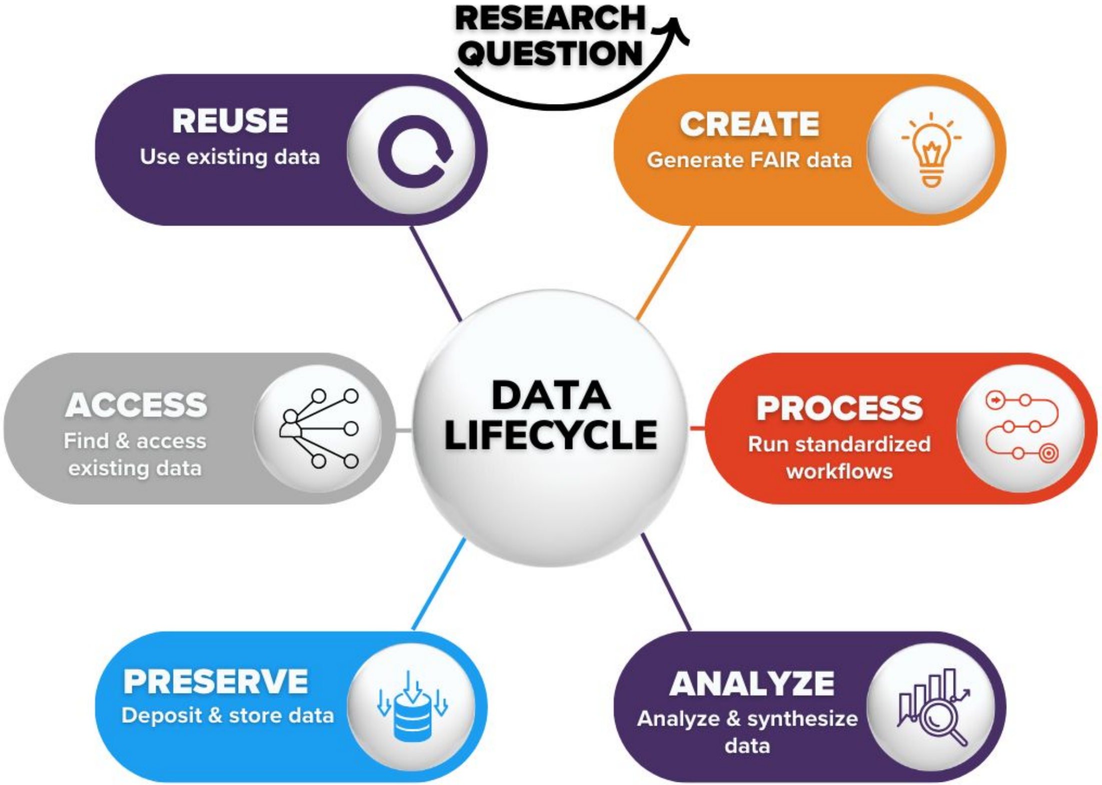
National Microbiome
Data Collaborative

Data Management

Data Lifecycle and Metadata

nmdc

National Microbiome Data Collaborative



Best Practices in Data Management support the entire Data Lifecycle

Data Management

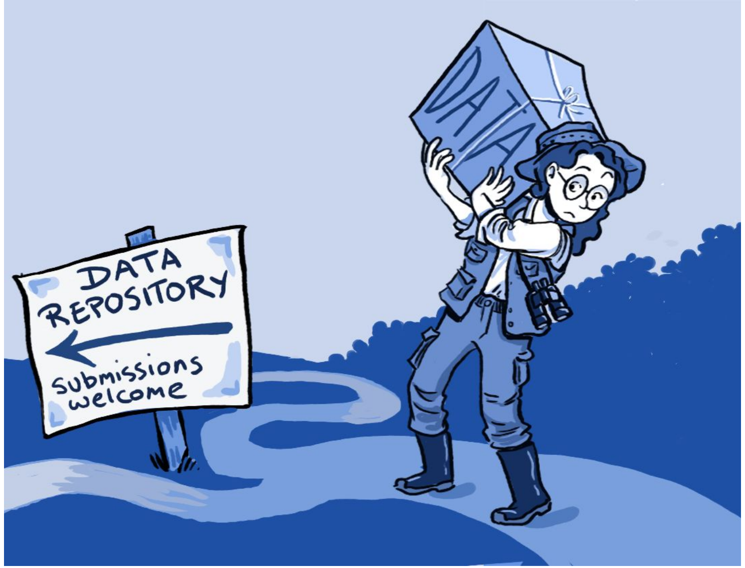
nmdc National Microbiome Data Collaborative

For research: the *intentional* process of collecting, storing, processing, and protecting data

For data preservation: data are **FAIR** now and into the future

Benefits of good data management:

- Decreased data loss
- Streamlined data deposition and publication
- Work is appropriately credited
- Impact is measurable / quantifiable
- Increased scientific exposure



From: Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, et al. (2014) Troubleshooting Public Data Archiving: Suggestions to Increase Participation. PLoS Biol 12(1): e1001779. https://doi.org/10.1371/journal.pbio.1001779

What is a Data Management Plan?

Required part of any federally funded grant proposal

- How you will collect, categorize, store, and share any data produced during the duration of a grant
- How that data will be preserved and made accessible after the completion of a project

Impact:

- Well-managed data can be published as a product
  - DOIs assigned to datasets
- Data management plans extend beyond the researchers
- Leads to greater re-use of data (internally and externally)

What to include in your DMP?

What goes into a data management plan?

SAMPLE AND DATA TYPES AND SOURCES

Outlines what kinds of data will be produced throughout the project.

DATA STANDARDS AND FORMATS

Defines all variables of interest and communicates that you are aware of and will abide by community best practices whenever possible.

DATA DISSEMINATION & ARCHIVING

Describes what the final data products will be and how you will protect data, if applicable.

POLICIES FOR DATA SHARING, PUBLIC ACCESS, AND RE-USE

Communicates that you understand your funders data sharing policies and that you have a plan to ensure public availability.

DATA AND SAMPLE PRESERVATION

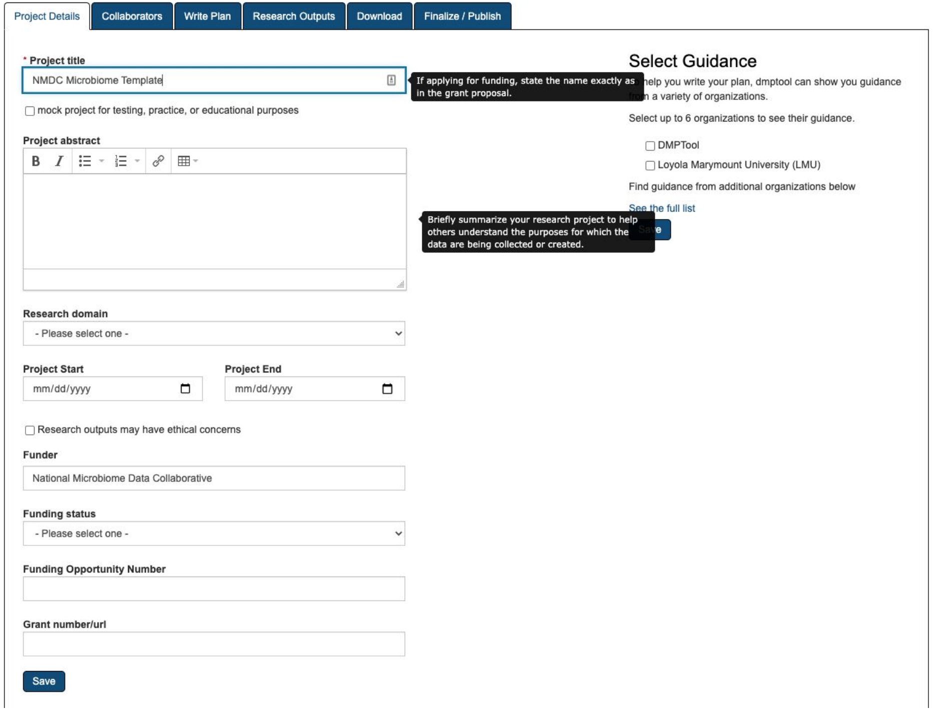
Communicates the sustainability plan for your data, showing your funder that the data products will last after the completion of the project.

ROLES AND RESPONSIBILITIES

Shows how your data management plan will be executed and ensures that your team's data management responsibilities are clearly defined.

NMDC DMP Resources

DMPTool Template



Project Details | Collaborators | Write Plan | Research Outputs | Download | Finalize / Publish

* Project title

NMDC Microbiome Template

If applying for funding, state the name exactly as in the grant proposal.

□ mock project for testing, practice, or educational purposes

Project abstract

Briefly summarize your research project to help others understand the purposes for which the data are being collected or created.

Research domain

- Please select one -

Project Start

mm/dd/yyyy

Project End

mm/dd/yyyy

□ Research outputs may have ethical concerns

Funder

National Microbiome Data Collaborative

Funding status

- Please select one -

Funding Opportunity Number

Grant number/url

Save

Select Guidance

help you write your plan, dmptool can show you guidance from a variety of organizations.

Select up to 6 organizations to see their guidance.

□ DMPTool

□ Loyola Marymount University (LMU)

Find guidance from additional organizations below

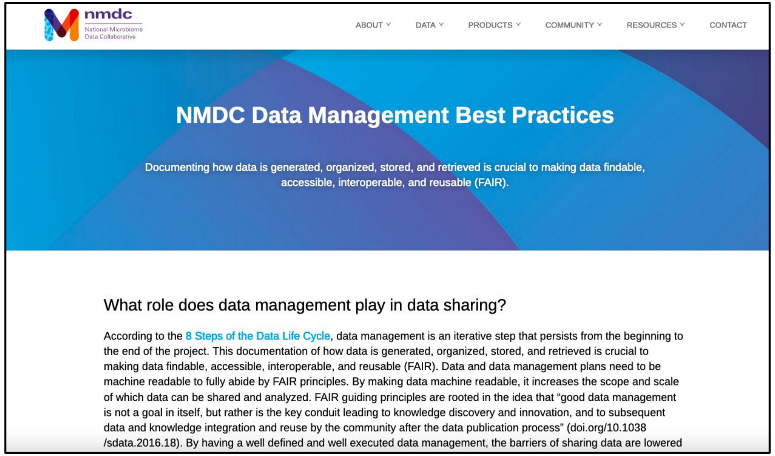
See the full list

Save

DMP Consultancies

One-on-one guidance on how to make a FAIR Data Management Plan

NMDC team has the expertise to offer guidance on the creation of DMPs for the **Department of Energy Office of Science** proposals



NMDC Data Management Best Practices

Documenting how data is generated, organized, stored, and retrieved is crucial to making data findable, accessible, interoperable, and reusable (FAIR).

What role does data management play in data sharing?

According to the 8 Steps of the Data Life Cycle, data management is an iterative step that persists from the beginning to the end of the project. This documentation of how data is generated, organized, stored, and retrieved is crucial to making data findable, accessible, interoperable, and reusable (FAIR). Data and data management plans need to be machine readable to fully abide by FAIR principles. By making data machine readable, it increases the scope and scale of which data can be shared and analyzed. FAIR guiding principles are rooted in the idea that “good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process” (doi.org/10.1038 /sdata.2016.18). By having a well defined and well executed data management, the barriers of sharing data are lowered

https://microbiomedata.org/data-management/

Tools for Data Management

Raise your hand if you use the following for managing your data:

- Lab notebooks
- Field notes, pieces of paper
- Remembering the information
- Google sheets
- Excel
- Tablet, iPad
- Laboratory Information Management Systems (LIMS) systems
- Online tools, data management plan software
- Post-it notes, paper towels, the back of a glove in the lab

Data Management Reflection

nmdc
National Microbiome
Data Collaborative

- How do you implement data management best practices in your organization?
    - Is there anything you would add/change to make your data more FAIR?
    - What tools work for you for data management? Are there tools that could improve this process for you?
    - How can we implement the principles we discussed and ensure accountability?
- Are the FAIR principles enough?

nmdc

National Microbiome Data Collaborative

Questions?

nmdc

National Microbiome Data Collaborative

NMDC Data Portal

Product Initiatives

nmdc
National Microbiome Data Collaborative

Submission Portal

Lower barriers to collect study and biosample data

NMDC EDGE

Streamline multi-omics data processing

Data Portal & API

Access and discovery of microbiome information

Engagement

User Facilities

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

EMSL

Individuals

The NMDC Champions Program

The NMDC Ambassador Program

Strategic Partners

NSF

neon Operated by Battelle

NASA

GeneLab

Product Initiatives

Submission Portal

Lower barriers to collect study and biosample data

NMDC EDGE

Streamline multi-omics data processing

Data Portal & API

Access and discovery of microbiome information

Engagement

User Facilities

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

EMSL

Individuals

The NMDC Champions Program

The NMDC Ambassador Program

Strategic Partners

NSF

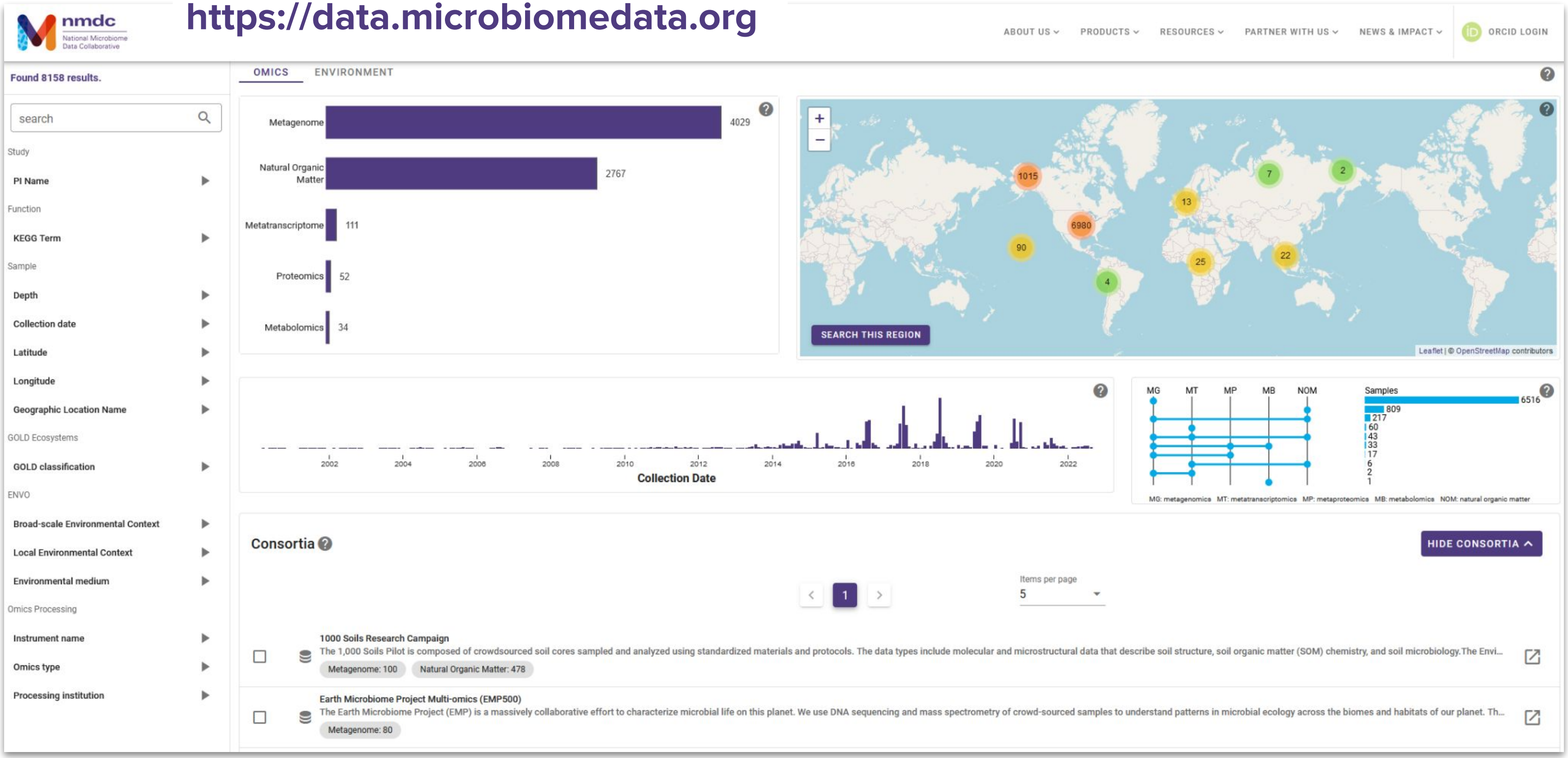
neon Operated by Battelle

NASA GeneLab

NMDC Data Portal

nmdc
National Microbiome Data Collaborative

https://data.microbiomedata.org



Found 8158 results.

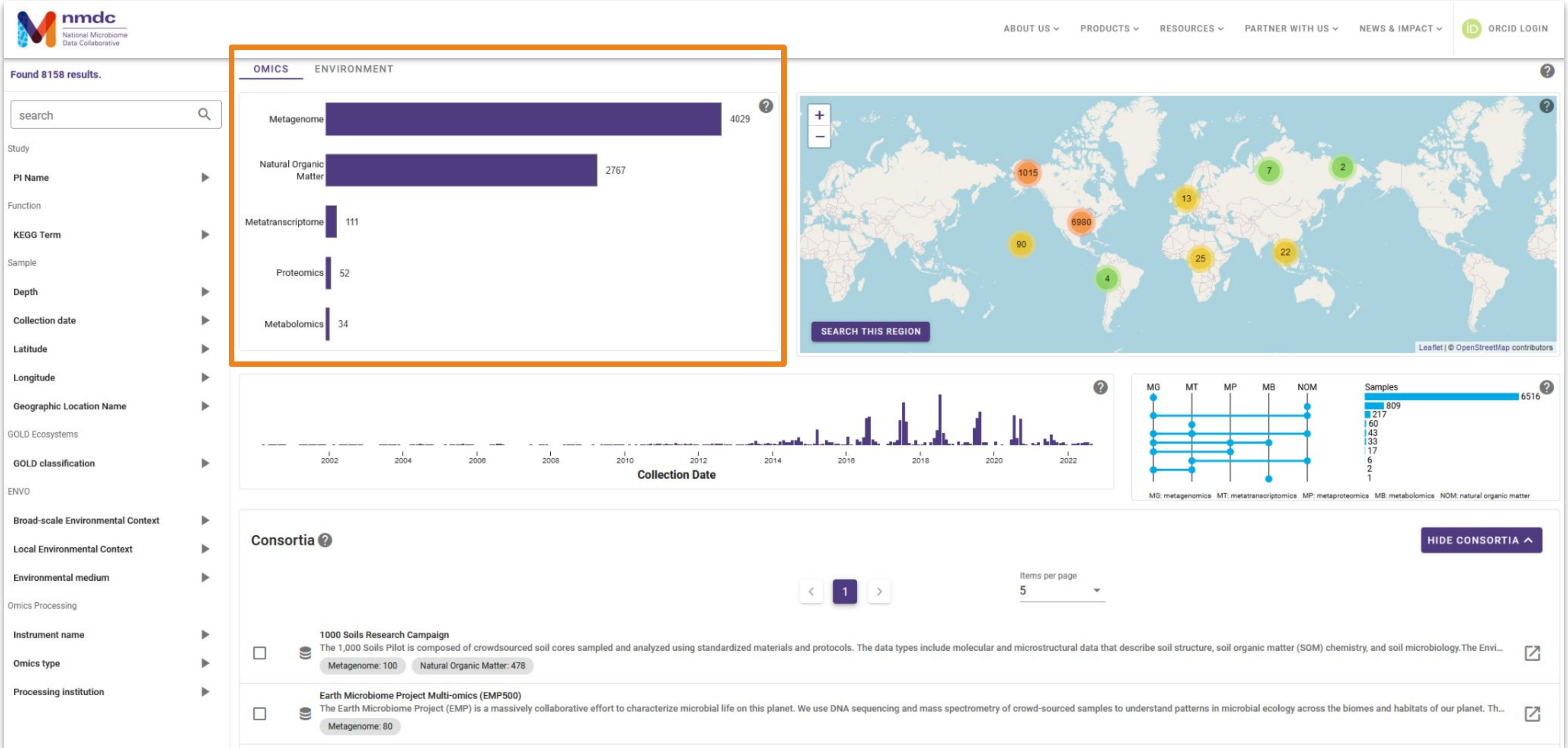
Consortia

□ **1000 Soils Research Campaign**
The 1,000 Soils Pilot is composed of crowdsourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology.The Envi...
Metagenome: 100 Natural Organic Matter: 478

□ **Earth Microbiome Project Multi-omics (EMP500)**
The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ecology across the biomes and habitats of our planet. Th...
Metagenome: 80

‘Omics Type

nmdc
National Microbiome
Data Collaborative



Found 8158 results.

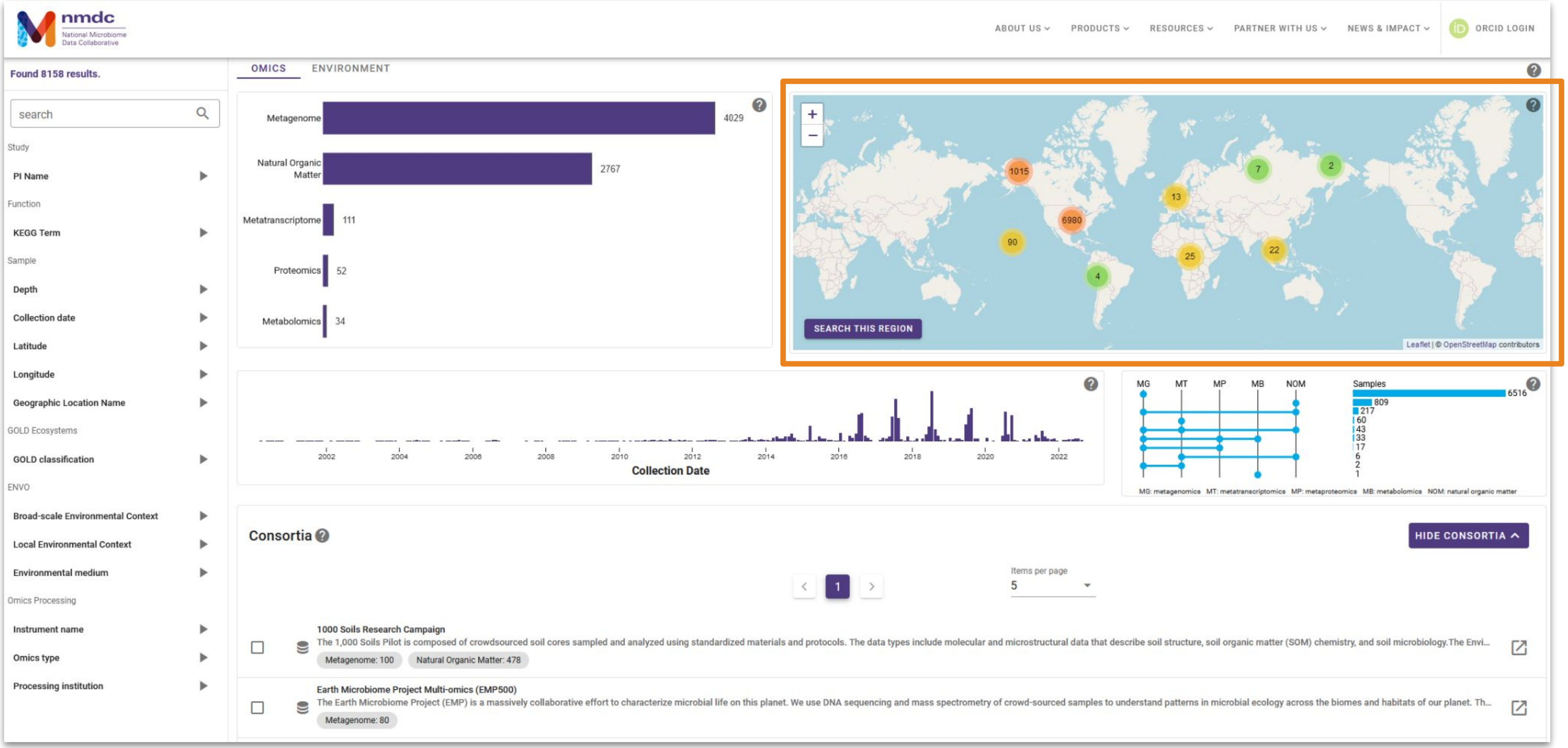
Consortia

□ **1000 Soils Research Campaign**
The 1,000 Soils Pilot is composed of crowdsourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology.The Envi...
Metagenome: 100 | Natural Organic Matter: 478

□ **Earth Microbiome Project Multi-omics (EMP500)**
The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ecology across the biomes and habitats of our planet. Th...
Metagenome: 80

Geographic Location

nmdc
National Microbiome Data Collaborative



Found 8158 results.

search

Study

- PI Name ▶

Function

- KEGG Term ▶

Sample

- Depth ▶
- Collection date ▶
- Latitude ▶
- Longitude ▶
- Geographic Location Name ▶

GOLD Ecosystems

- GOLD classification ▶

ENVO

- Broad-scale Environmental Context ▶
- Local Environmental Context ▶
- Environmental medium ▶

Omics Processing

- Instrument name ▶
- Omics type ▶
- Processing institution ▶

Consortia

HIDE CONSORTIA

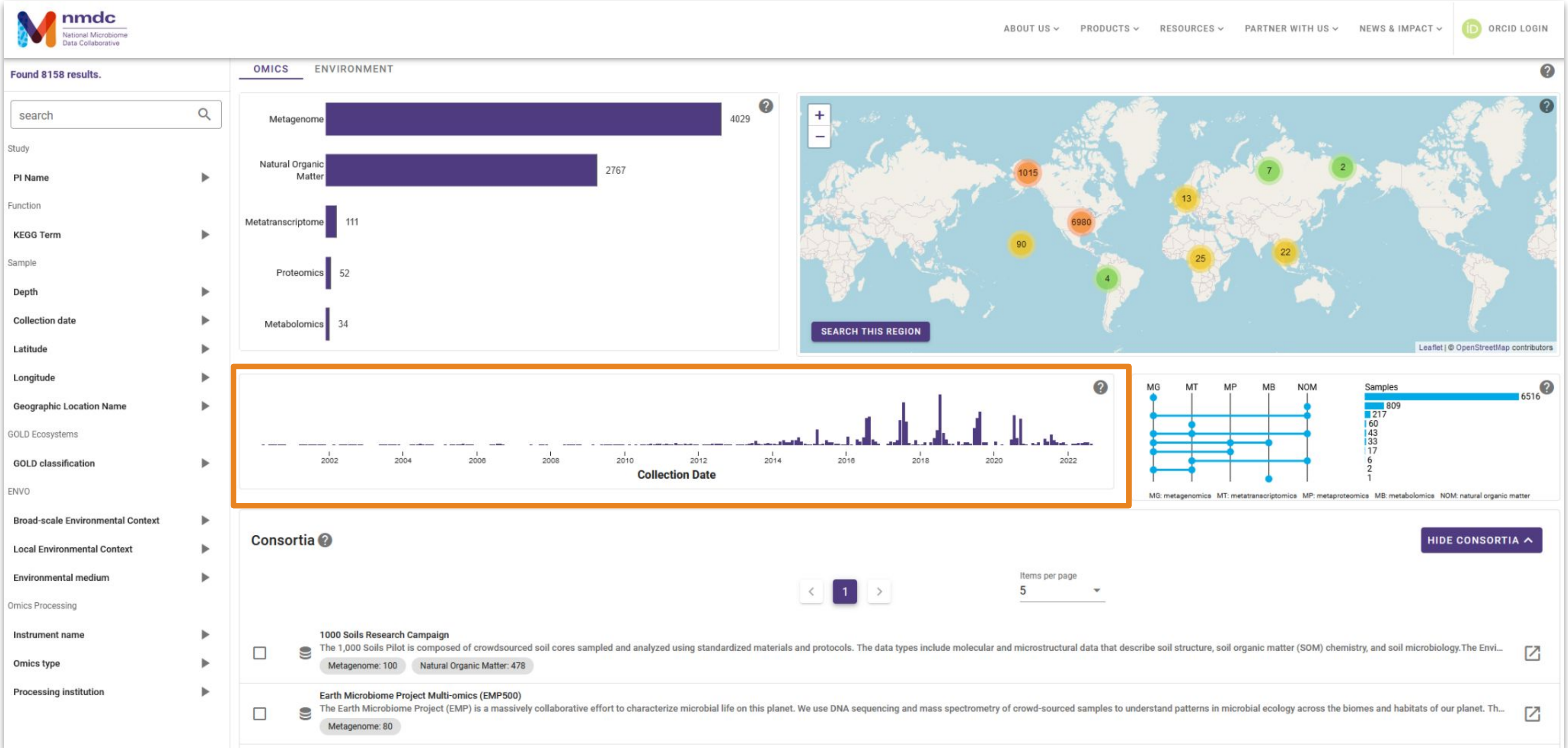
‹ 1 ›

Items per page
5

□ **1000 Soils Research Campaign**
The 1,000 Soils Pilot is composed of crowdsourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology.The Envi...
Metagenome: 100 Natural Organic Matter: 478

□ **Earth Microbiome Project Multi-omics (EMP500)**
The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ecology across the biomes and habitats of our planet. Th...
Metagenome: 80

Collection Date



Found 8158 results.

Consortia

□ 1000 Soils Research Campaign

The 1,000 Soils Pilot is composed of crowdsourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology.The Envi...

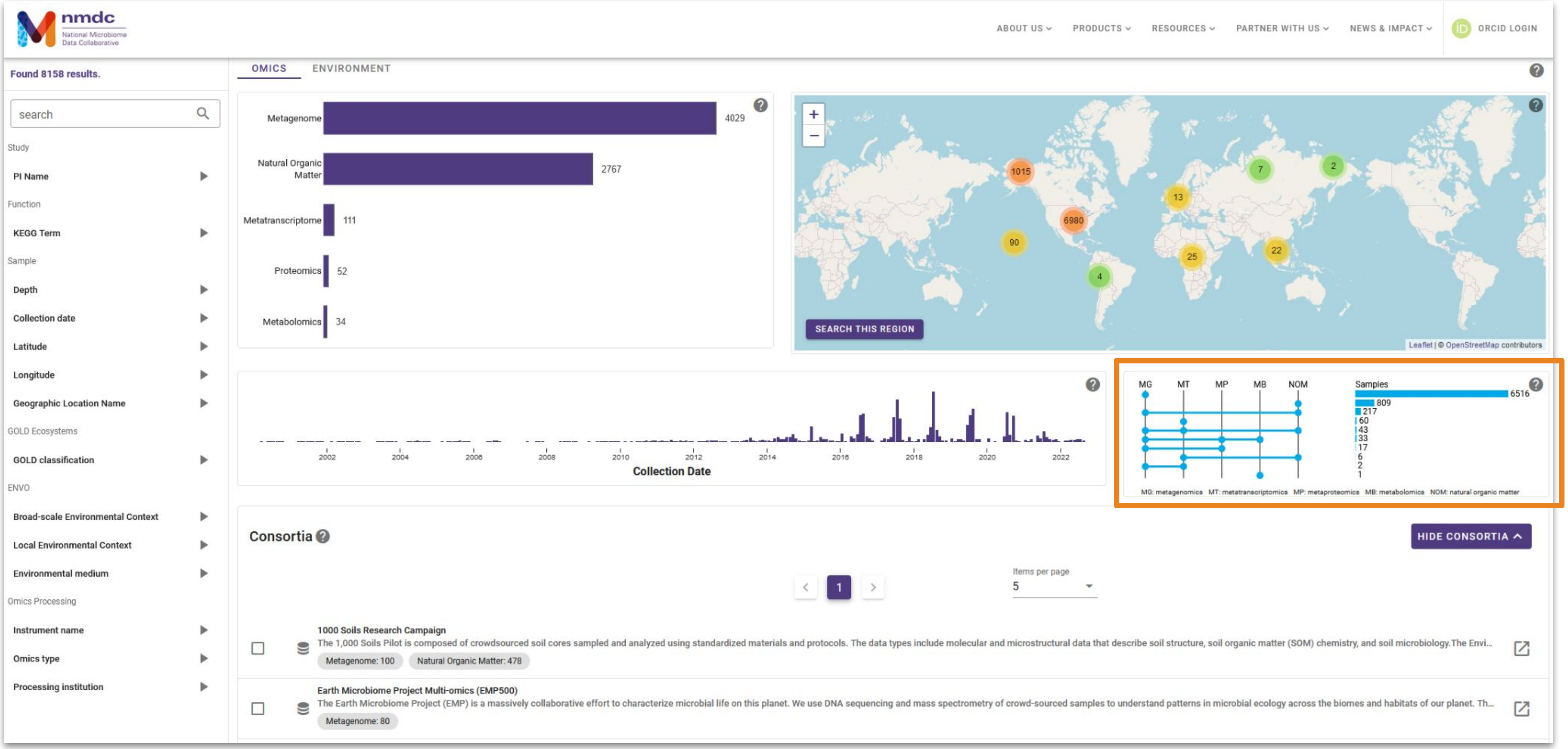
Metagenome: 100 | Natural Organic Matter: 478

□ Earth Microbiome Project Multi-omics (EMP500)

The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ecology across the biomes and habitats of our planet. Th...

Metagenome: 80

Combination of ‘omics types



nmdc National Microbiome Data Collaborative

ABOUT US PRODUCTS RESOURCES PARTNER WITH US NEWS & IMPACT ORCID LOGIN

Found 8158 results.

search

Study
- PI Name

Function
- KEGG Term

Sample
- Depth
- Collection date
- Latitude
- Longitude
- Geographic Location Name

GOLD Ecosystems
- GOLD classification

ENVO
- Broad-scale Environmental Context
- Local Environmental Context
- Environmental medium

Omics Processing
- Instrument name
- Omics type
- Processing institution

OMICS ENVIRONMENT

Consortia

HIDE CONSORTIA

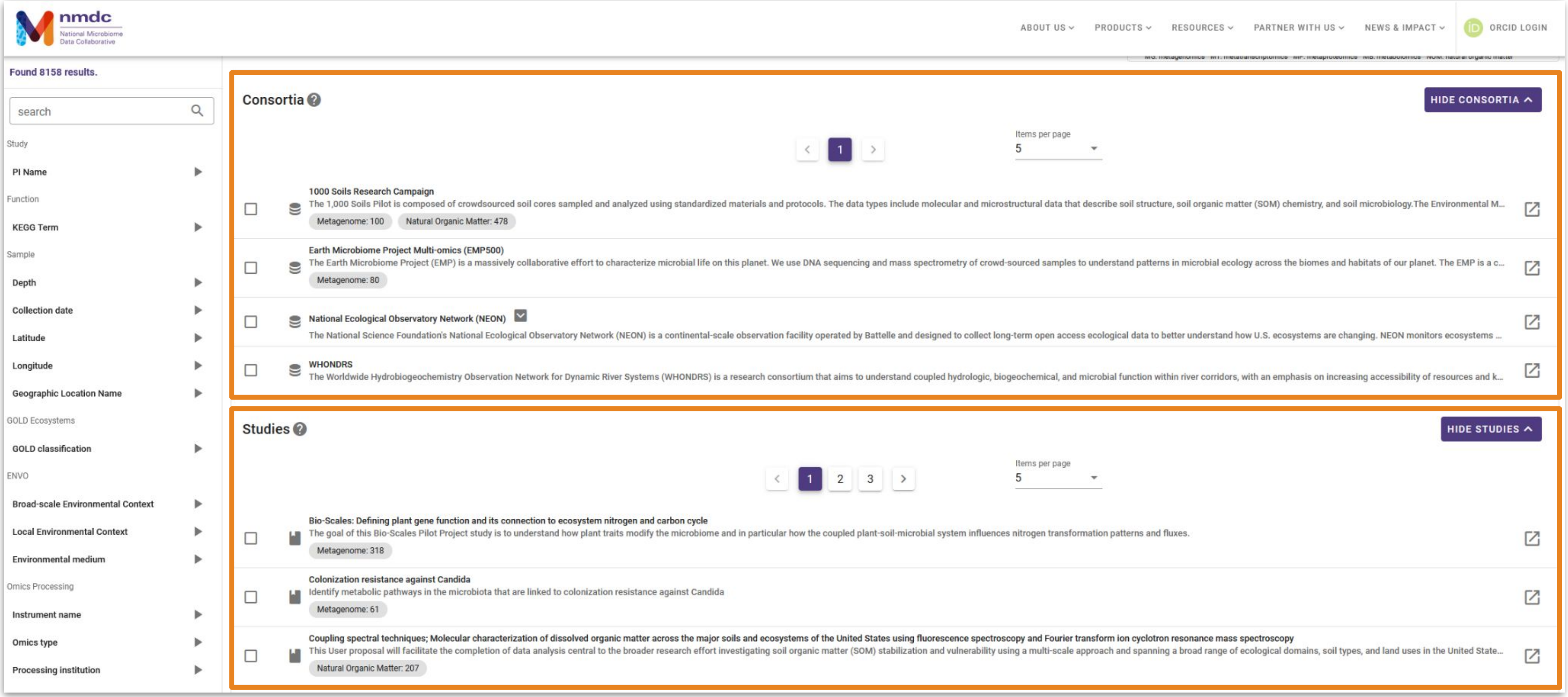
Items per page 5

□ **1000 Soils Research Campaign**
The 1,000 Soils Pilot is composed of crowdsourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology.The Envi...
Metagenome: 100 Natural Organic Matter: 478

□ **Earth Microbiome Project Multi-omics (EMP500)**
The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ecology across the biomes and habitats of our planet. Th...
Metagenome: 80

Consortia and Studies

nmdc
National Microbiome Data Collaborative



Found 8158 results.

Consortia

Items per page: 5

□ **1000 Soils Research Campaign**
The 1,000 Soils Pilot is composed of crowdsourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology.The Environmental M...
Metagenome: 100 | Natural Organic Matter: 478

□ **Earth Microbiome Project Multi-omics (EMP500)**
The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ecology across the biomes and habitats of our planet. The EMP is a c...
Metagenome: 80

□ **National Ecological Observatory Network (NEON)**
The National Science Foundation's National Ecological Observatory Network (NEON) is a continental-scale observation facility operated by Battelle and designed to collect long-term open access ecological data to better understand how U.S. ecosystems are changing. NEON monitors ecosystems ...

□ **WHONDRS**
The Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS) is a research consortium that aims to understand coupled hydrologic, biogeochemical, and microbial function within river corridors, with an emphasis on increasing accessibility of resources and k...

Studies

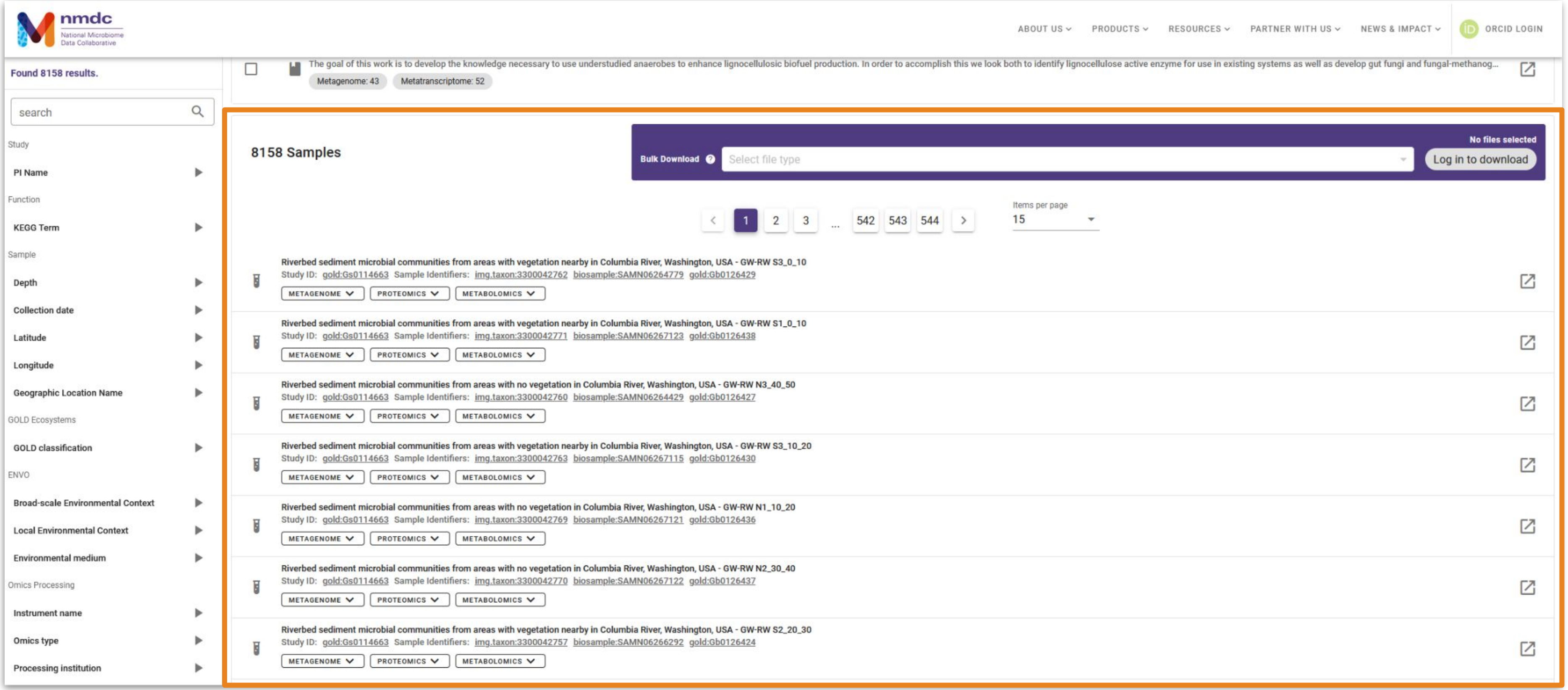
Items per page: 5

□ **Bio-Scales: Defining plant gene function and its connection to ecosystem nitrogen and carbon cycle**
The goal of this Bio-Scales Pilot Project study is to understand how plant traits modify the microbiome and in particular how the coupled plant-soil-microbial system influences nitrogen transformation patterns and fluxes.
Metagenome: 318

□ **Colonization resistance against Candida**
Identify metabolic pathways in the microbiota that are linked to colonization resistance against Candida
Metagenome: 61

□ **Coupling spectral techniques; Molecular characterization of dissolved organic matter across the major soils and ecosystems of the United States using fluorescence spectroscopy and Fourier transform ion cyclotron resonance mass spectroscopy**
This User proposal will facilitate the completion of data analysis central to the broader research effort investigating soil organic matter (SOM) stabilization and vulnerability using a multi-scale approach and spanning a broad range of ecological domains, soil types, and land uses in the United State...
Natural Organic Matter: 207

Samples



Found 8158 results.

The goal of this work is to develop the knowledge necessary to use understudied anaerobes to enhance lignocellulosic biofuel production. In order to accomplish this we look both to identify lignocellulose active enzyme for use in existing systems as well as develop gut fungi and fungal-methanog...

Metagenome: 43 Metatranscriptome: 52

8158 Samples

Riverbed sediment microbial communities from areas with vegetation nearby in Columbia River, Washington, USA - GW-RW S3_0_10
Study ID: gold:Gs0114663 Sample Identifiers: img.taxon:3300042762 biosample:SAMN06264779 gold:Gb0126429
METAGENOME PROTEOMICS METABOLOMICS

Riverbed sediment microbial communities from areas with vegetation nearby in Columbia River, Washington, USA - GW-RW S1_0_10
Study ID: gold:Gs0114663 Sample Identifiers: img.taxon:3300042771 biosample:SAMN06267123 gold:Gb0126438
METAGENOME PROTEOMICS METABOLOMICS

Riverbed sediment microbial communities from areas with no vegetation in Columbia River, Washington, USA - GW-RW N3_40_50
Study ID: gold:Gs0114663 Sample Identifiers: img.taxon:3300042760 biosample:SAMN06264429 gold:Gb0126427
METAGENOME PROTEOMICS METABOLOMICS

Riverbed sediment microbial communities from areas with vegetation nearby in Columbia River, Washington, USA - GW-RW S3_10_20
Study ID: gold:Gs0114663 Sample Identifiers: img.taxon:3300042763 biosample:SAMN06267115 gold:Gb0126430
METAGENOME PROTEOMICS METABOLOMICS

Riverbed sediment microbial communities from areas with no vegetation in Columbia River, Washington, USA - GW-RW N1_10_20
Study ID: gold:Gs0114663 Sample Identifiers: img.taxon:3300042769 biosample:SAMN06267121 gold:Gb0126436
METAGENOME PROTEOMICS METABOLOMICS

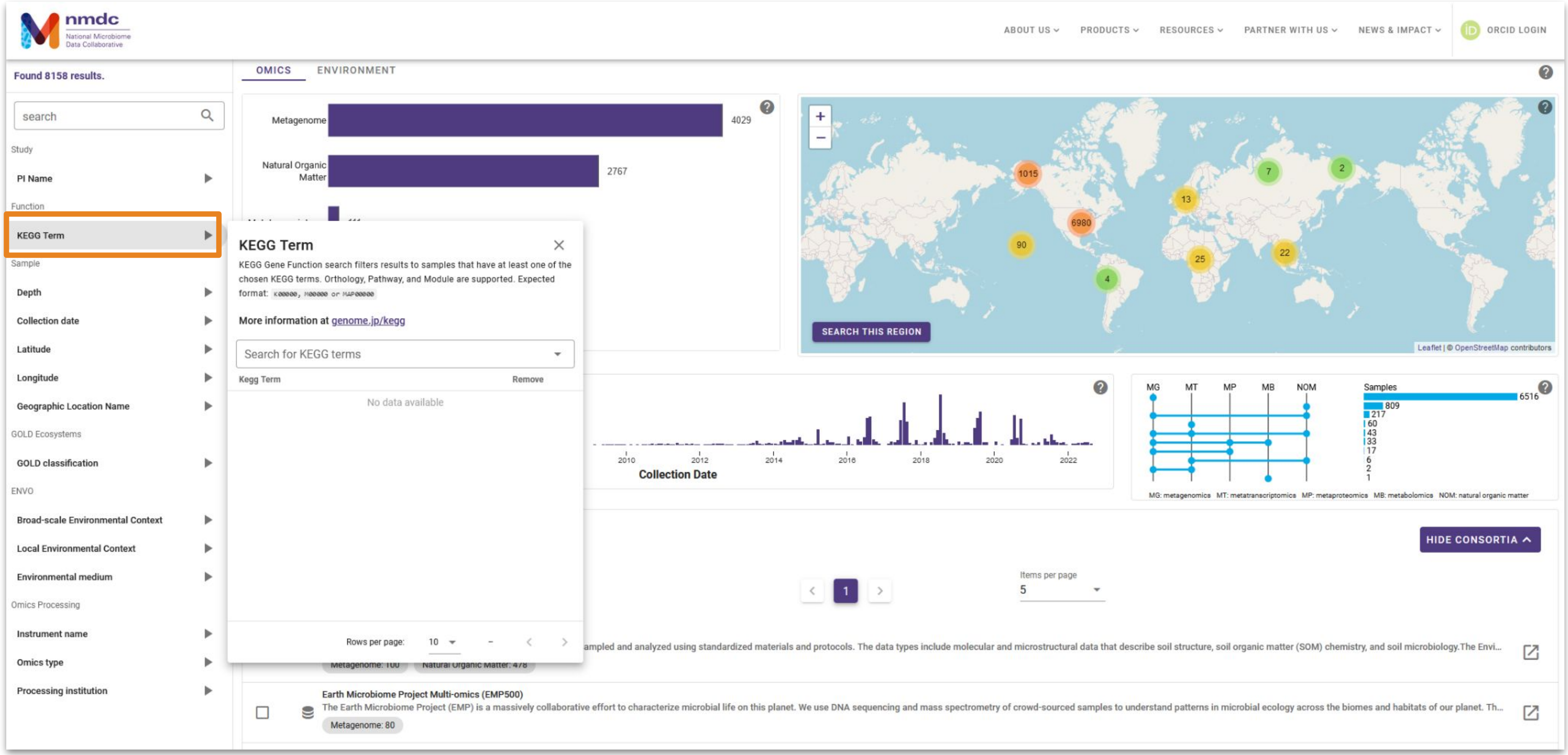
Riverbed sediment microbial communities from areas with no vegetation in Columbia River, Washington, USA - GW-RW N2_30_40
Study ID: gold:Gs0114663 Sample Identifiers: img.taxon:3300042770 biosample:SAMN06267122 gold:Gb0126437
METAGENOME PROTEOMICS METABOLOMICS

Riverbed sediment microbial communities from areas with vegetation nearby in Columbia River, Washington, USA - GW-RW S2_20_30
Study ID: gold:Gs0114663 Sample Identifiers: img.taxon:3300042757 biosample:SAMN06266292 gold:Gb0126424
METAGENOME PROTEOMICS METABOLOMICS

KEGG Term Search

nmdc

National Microbiome Data Collaborative



Found 8158 results.

KEGG Term

KEGG Gene Function search filters results to samples that have at least one of the chosen KEGG terms. Orthology, Pathway, and Module are supported. Expected format: `K00000, M00000 or MAP00000`

More information at genome.jp/kegg

Search for KEGG terms

Kegg Term	Remove
No data available	

Rows per page: 10

...ampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology.The Envi...

Metagenome: 100 Natural Organic Matter: 478

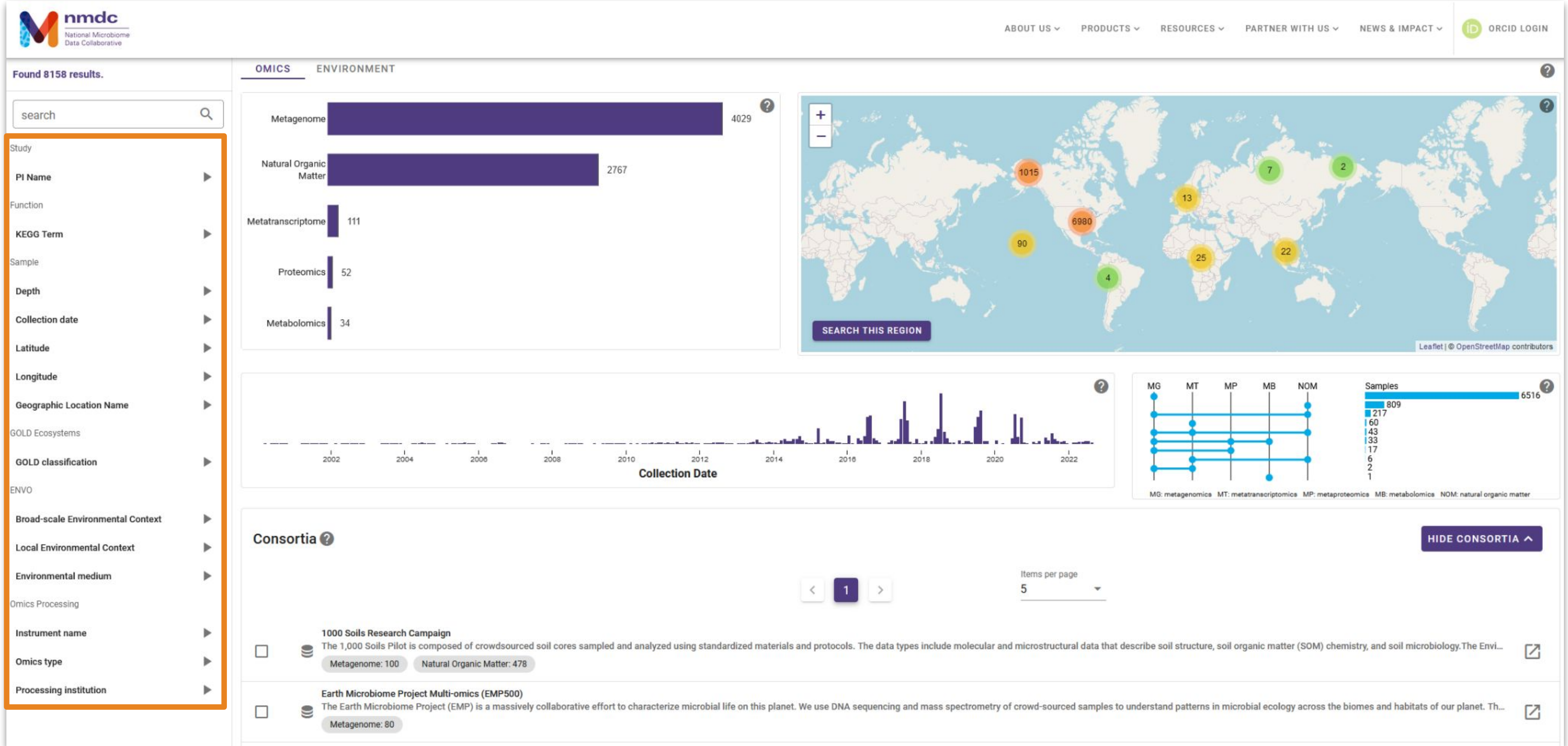
□ Earth Microbiome Project Multi-omics (EMP500)

The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ecology across the biomes and habitats of our planet. Th...

Metagenome: 80

Other searchable metadata fields

nmdc
National Microbiome Data Collaborative



Found 8158 results.

Study
- PI Name ▶

Function
- KEGG Term ▶

Sample
- Depth ▶
- Collection date ▶
- Latitude ▶
- Longitude ▶
- Geographic Location Name ▶

GOLD Ecosystems
- GOLD classification ▶

ENVO
- Broad-scale Environmental Context ▶
- Local Environmental Context ▶
- Environmental medium ▶

Omics Processing
- Instrument name ▶
- Omics type ▶
- Processing institution ▶

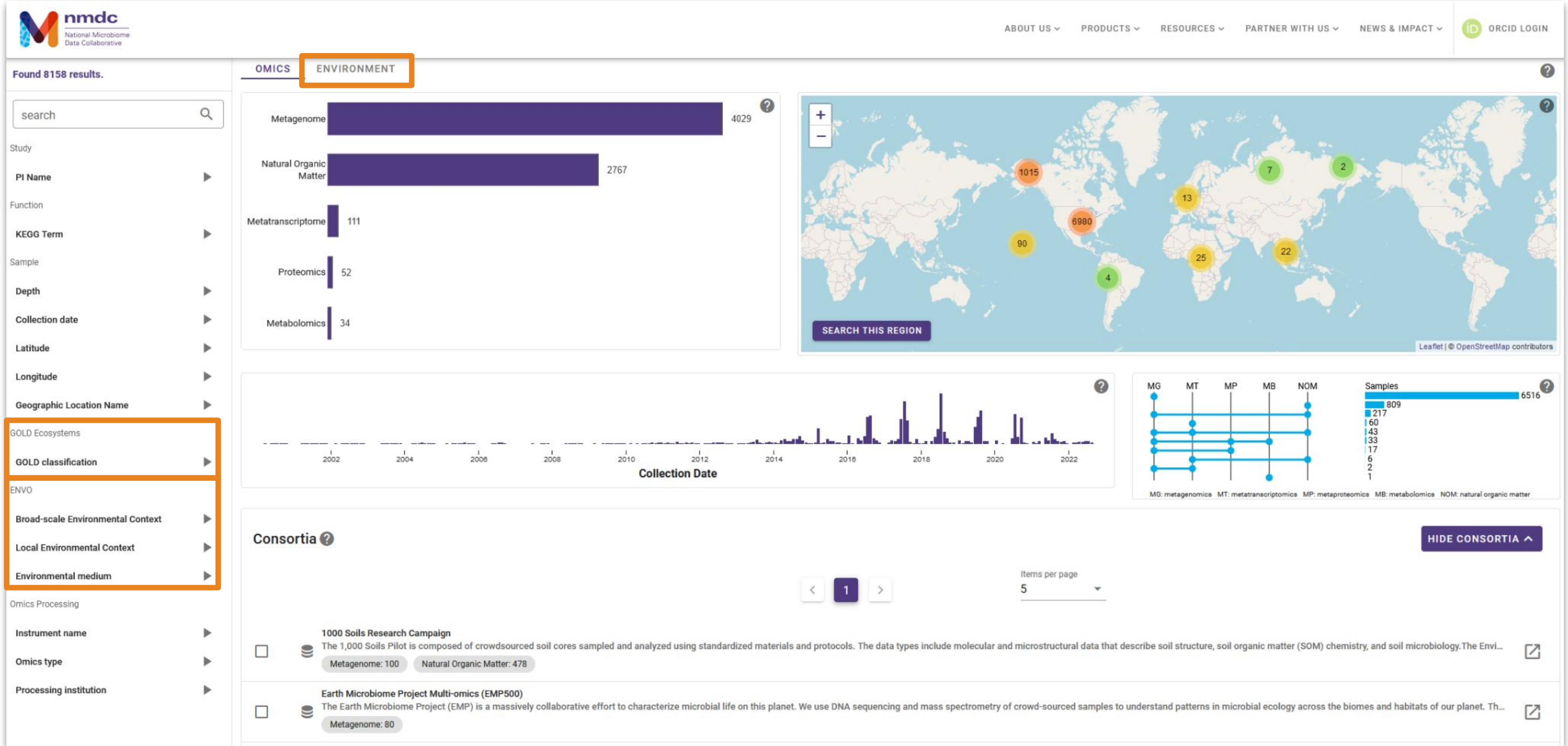
Consortia

☐ **1000 Soils Research Campaign**
The 1,000 Soils Pilot is composed of crowdsourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology.The Envi...
Metagenome: 100 | Natural Organic Matter: 478

☐ **Earth Microbiome Project Multi-omics (EMP500)**
The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ecology across the biomes and habitats of our planet. Th...
Metagenome: 80

Environmental Metadata

nmdc
National Microbiome Data Collaborative



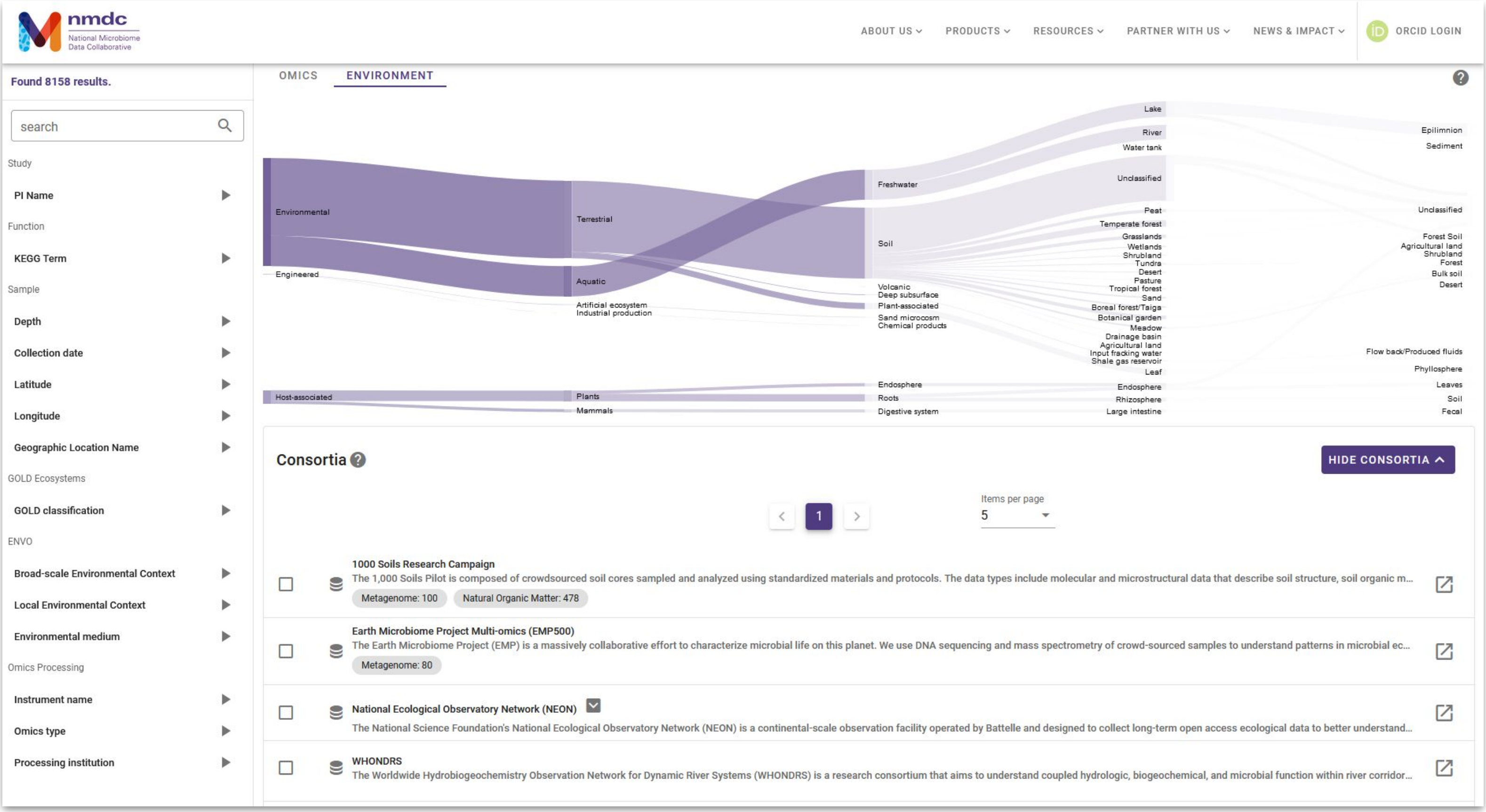
Found 8158 results.

Consortia

□ 1000 Soils Research Campaign
The 1,000 Soils Pilot is composed of crowdsourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology.The Envi...
Metagenome: 100 Natural Organic Matter: 478

□ Earth Microbiome Project Multi-omics (EMP500)
The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ecology across the biomes and habitats of our planet. Th...
Metagenome: 80

Environment tab



Found 8158 results.

Consortia

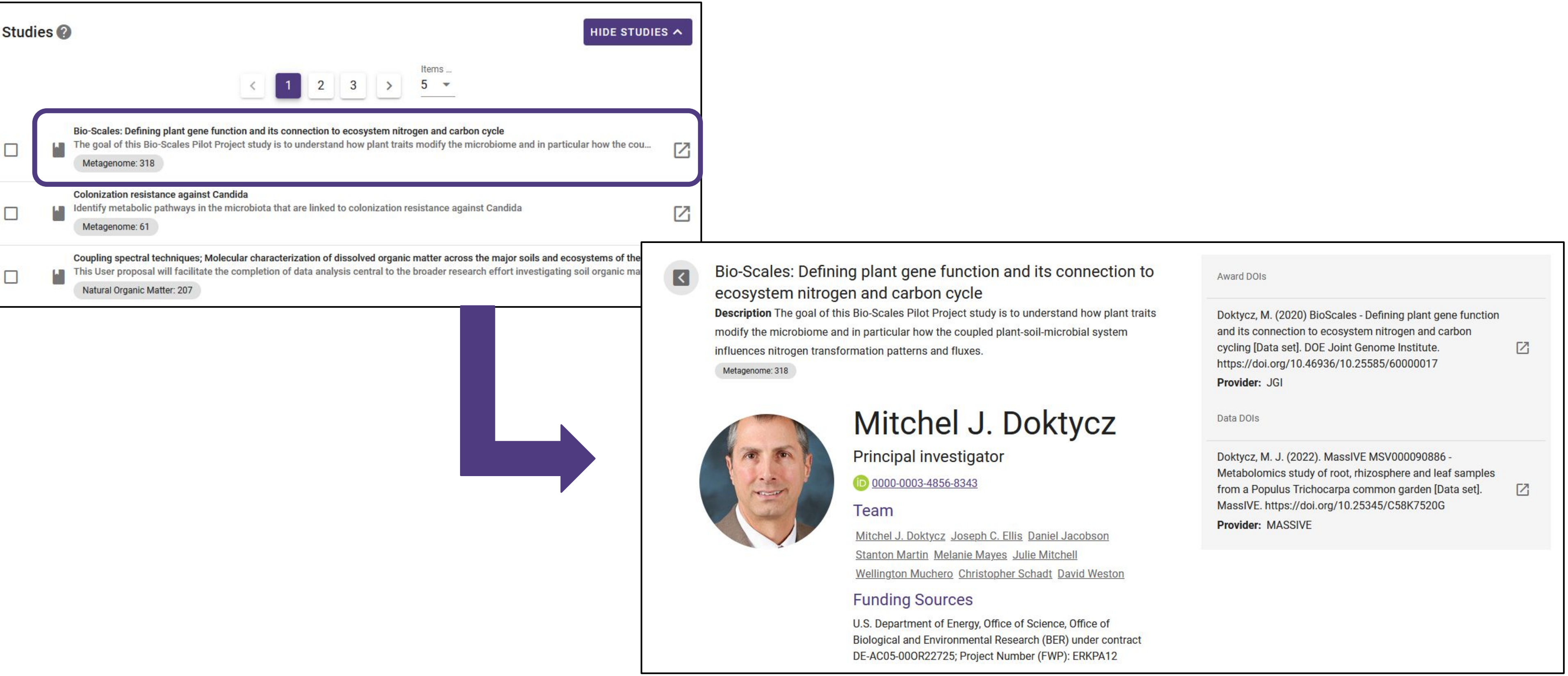
□ 1000 Soils Research Campaign
The 1,000 Soils Pilot is composed of crowdsourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic m...
Metagenome: 100 Natural Organic Matter: 478

□ Earth Microbiome Project Multi-omics (EMP500)
The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. We use DNA sequencing and mass spectrometry of crowd-sourced samples to understand patterns in microbial ec...
Metagenome: 80

□ National Ecological Observatory Network (NEON)
The National Science Foundation's National Ecological Observatory Network (NEON) is a continental-scale observation facility operated by Battelle and designed to collect long-term open access ecological data to better understand...

□ WHONDRS
The Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS) is a research consortium that aims to understand coupled hydrologic, biogeochemical, and microbial function within river corridor...

Study Pages



Studies

□ Bio-Scales: Defining plant gene function and its connection to ecosystem nitrogen and carbon cycle
The goal of this Bio-Scales Pilot Project study is to understand how plant traits modify the microbiome and in particular how the cou...
Metagenome: 318

□ Colonization resistance against Candida
Identify metabolic pathways in the microbiota that are linked to colonization resistance against Candida
Metagenome: 61

□ Coupling spectral techniques; Molecular characterization of dissolved organic matter across the major soils and ecosystems of the
This User proposal will facilitate the completion of data analysis central to the broader research effort investigating soil organic ma
Natural Organic Matter: 207

Bio-Scales: Defining plant gene function and its connection to ecosystem nitrogen and carbon cycle

Description The goal of this Bio-Scales Pilot Project study is to understand how plant traits modify the microbiome and in particular how the coupled plant-soil-microbial system influences nitrogen transformation patterns and fluxes.

Metagenome: 318

Mitchel J. Doktycz

Principal investigator

0000-0003-4856-8343

Team

Mitchel J. Doktycz Joseph C. Ellis Daniel Jacobson Stanton Martin Melanie Mayes Julie Mitchell Wellington Muchero Christopher Schadt David Weston

Funding Sources

U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER) under contract DE-AC05-00OR22725; Project Number (FWP): ERKPA12

Award DOIs

Doktycz, M. (2020) BioScales - Defining plant gene function and its connection to ecosystem nitrogen and carbon cycling [Data set]. DOE Joint Genome Institute. https://doi.org/10.46936/10.25585/60000017

Provider: JGI

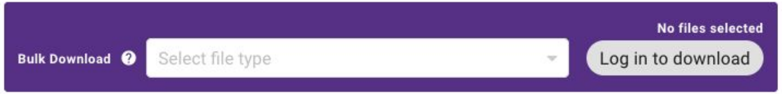
Data DOIs

Doktycz, M. J. (2022). MassIVE MSV000090886 - Metabolomics study of root, rhizosphere and leaf samples from a Populus Trichocarpa common garden [Data set]. MassIVE. https://doi.org/10.25345/C58K7520G

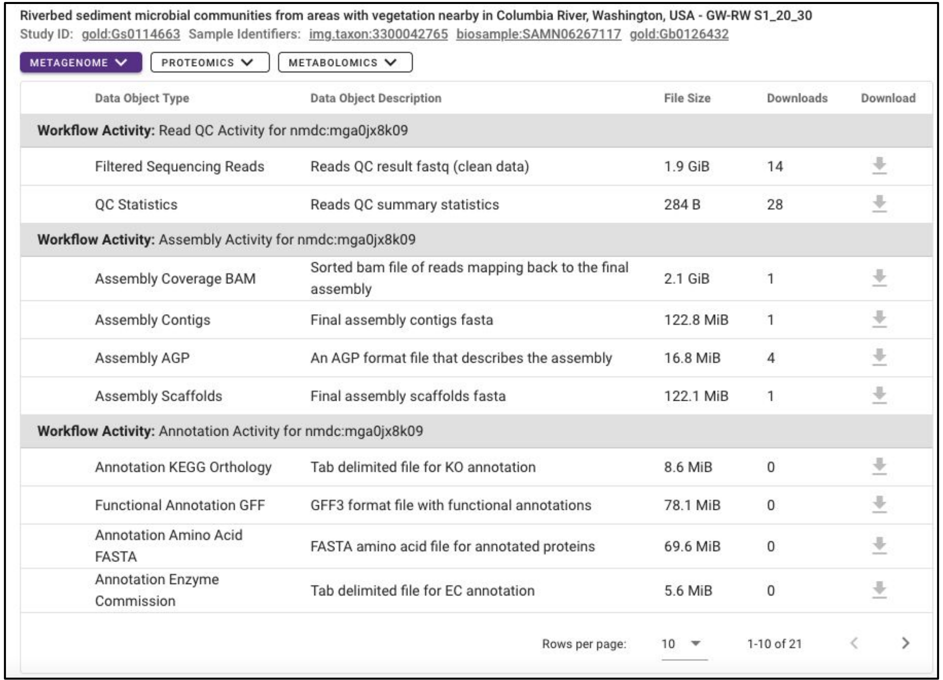
Provider: MASSIVE

Data Download

Need to log in with ORCiD to download data



Can directly download all files from the processed multi-omics data



Riverbed sediment microbial communities from areas with vegetation nearby in Columbia River, Washington, USA - GW-RW S1_20_30

Study ID: gold:Gs0114663 Sample Identifiers: img.taxon:3300042765 biosample:SAMN06267117 gold:Gb0126432

Data Object Type	Data Object Description	File Size	Downloads	Download
Workflow Activity: Read QC Activity for nmdc:mga0jx8k09				
Filtered Sequencing Reads	Reads QC result fastq (clean data)	1.9 GiB	14	⬇
QC Statistics	Reads QC summary statistics	284 B	28	⬇
Workflow Activity: Assembly Activity for nmdc:mga0jx8k09				
Assembly Coverage BAM	Sorted bam file of reads mapping back to the final assembly	2.1 GiB	1	⬇
Assembly Contigs	Final assembly contigs fasta	122.8 MiB	1	⬇
Assembly AGP	An AGP format file that describes the assembly	16.8 MiB	4	⬇
Assembly Scaffolds	Final assembly scaffolds fasta	122.1 MiB	1	⬇
Workflow Activity: Annotation Activity for nmdc:mga0jx8k09				
Annotation KEGG Orthology	Tab delimited file for KO annotation	8.6 MiB	0	⬇
Functional Annotation GFF	GFF3 format file with functional annotations	78.1 MiB	0	⬇
Annotation Amino Acid FASTA	FASTA amino acid file for annotated proteins	69.6 MiB	0	⬇
Annotation Enzyme Commission	Tab delimited file for EC annotation	5.6 MiB	0	⬇

Rows per page: 10 1-10 of 21

Instructional Content

Data Portal User Guide:

https://nmdc-documentation.readthedocs.io/en/latest/howto_guides/portal_guide.html

Data Portal Tutorial Video:

https://nmdc-documentation.readthedocs.io/en/latest/tutorials/nav_data_portal.html

Data Portal Documentation:

https://nmdc-documentation.readthedocs.io/en/latest/reference/data_portal.html

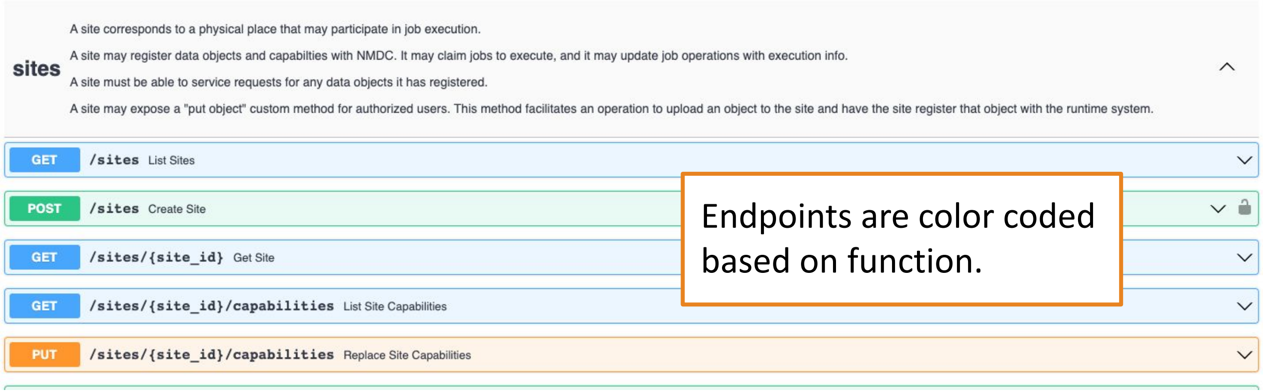
More information found by clicking the “?”s

NMDC API

nmdc
National Microbiome Data Collaborative

The NMDC API

- Auto-generated documentation at and UI at https://api.microbiomedata.org/docs#
- User friendly detailed documentation



sites

A site corresponds to a physical place that may participate in job execution.

A site may register data objects and capabilties with NMDC. It may claim jobs to execute, and it may update job operations with execution info.

A site must be able to service requests for any data objects it has registered.

A site may expose a "put object" custom method for authorized users. This method facilitates an operation to upload an object to the site and have the site register that object with the runtime system.

GET `/sites` List Sites

POST `/sites` Create Site

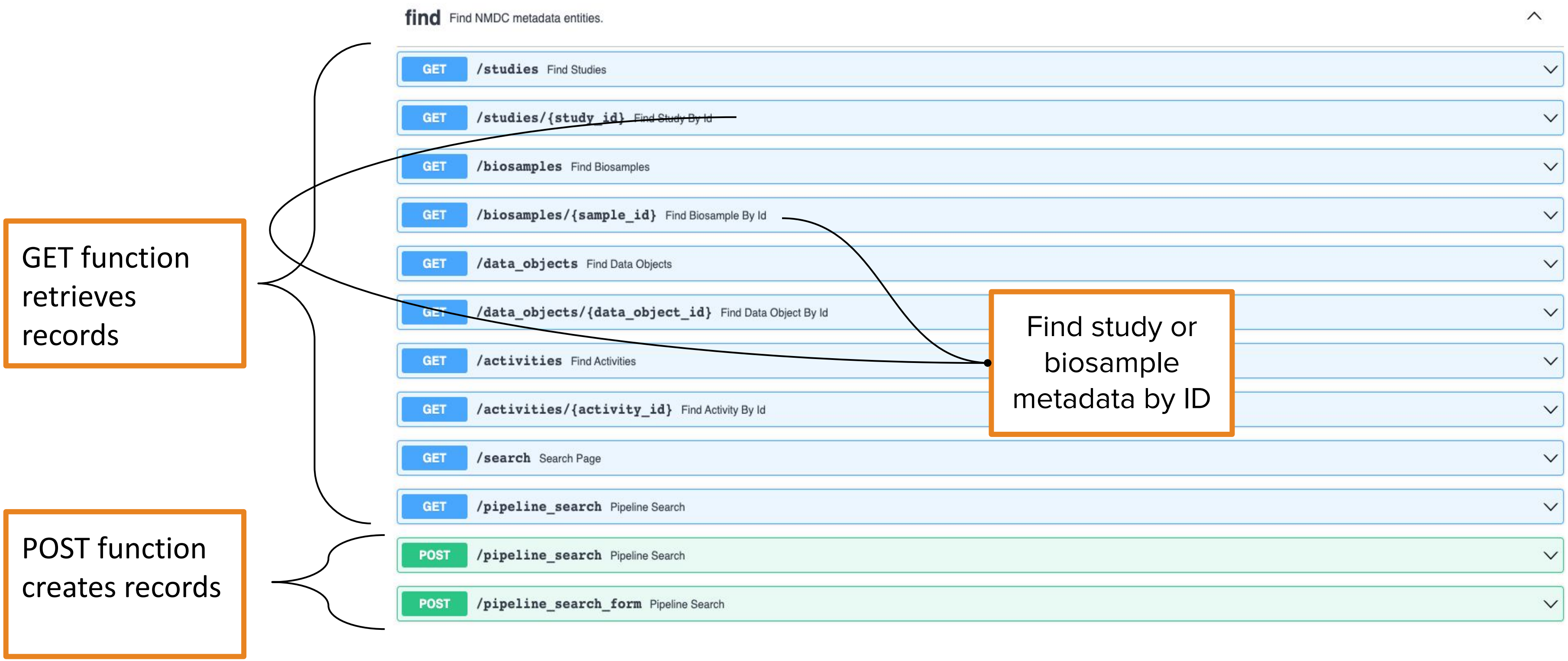
GET `/sites/{site_id}` Get Site

GET `/sites/{site_id}/capabilities` List Site Capabilities

PUT `/sites/{site_id}/capabilities` Replace Site Capabilities

Endpoints are color coded based on function.

The NMDC API



find Find NMDC metadata entities.

GET /studies Find Studies

GET /studies/{study_id} Find Study By Id

GET /biosamples Find Biosamples

GET /biosamples/{sample_id} Find Biosample By Id

GET /data_objects Find Data Objects

GET /data_objects/{data_object_id} Find Data Object By Id

GET /activities Find Activities

GET /activities/{activity_id} Find Activity By Id

GET /search Search Page

GET /pipeline_search Pipeline Search

POST /pipeline_search Pipeline Search

POST /pipeline_search_form Pipeline Search

GET function retrieves records

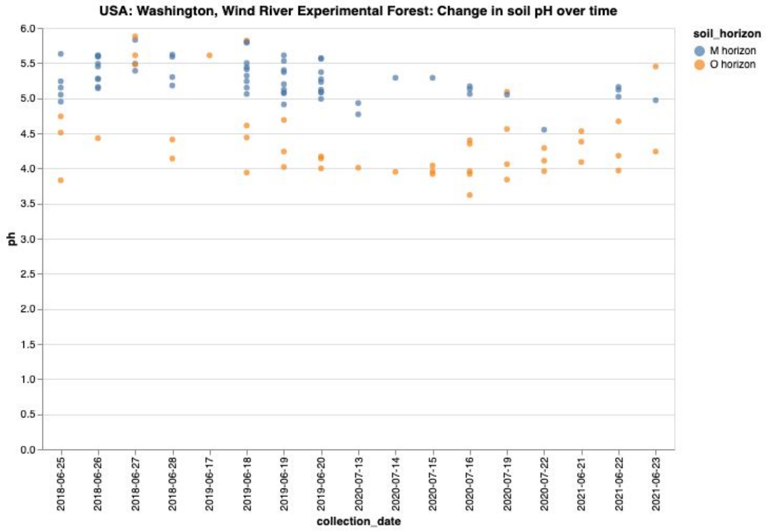
POST function creates records

Find study or biosample metadata by ID

Available Jupyter notebooks

The NMDC has several publicly available Jupyter notebooks that researchers can explore, use, and modify

https://github.com/microbiomedata/notebook_hackathons/tree/main



nmdc

National Microbiome Data Collaborative

Questions?

nmdc

National Microbiome
Data Collaborative

Data Portal Walkthrough

Adding Data

- Do you have data you’d like to add to the data portal?
- We are accepting environmental microbiome omics data with sufficient metadata
- NMDC team can process for you and upload to the Portal
- Why add your data?
    - Increased citations, increased visibility, inclusion in meta-analyses, comparisons with other datasets in the portal, ethical benefits

nmdc

National Microbiome Data Collaborative

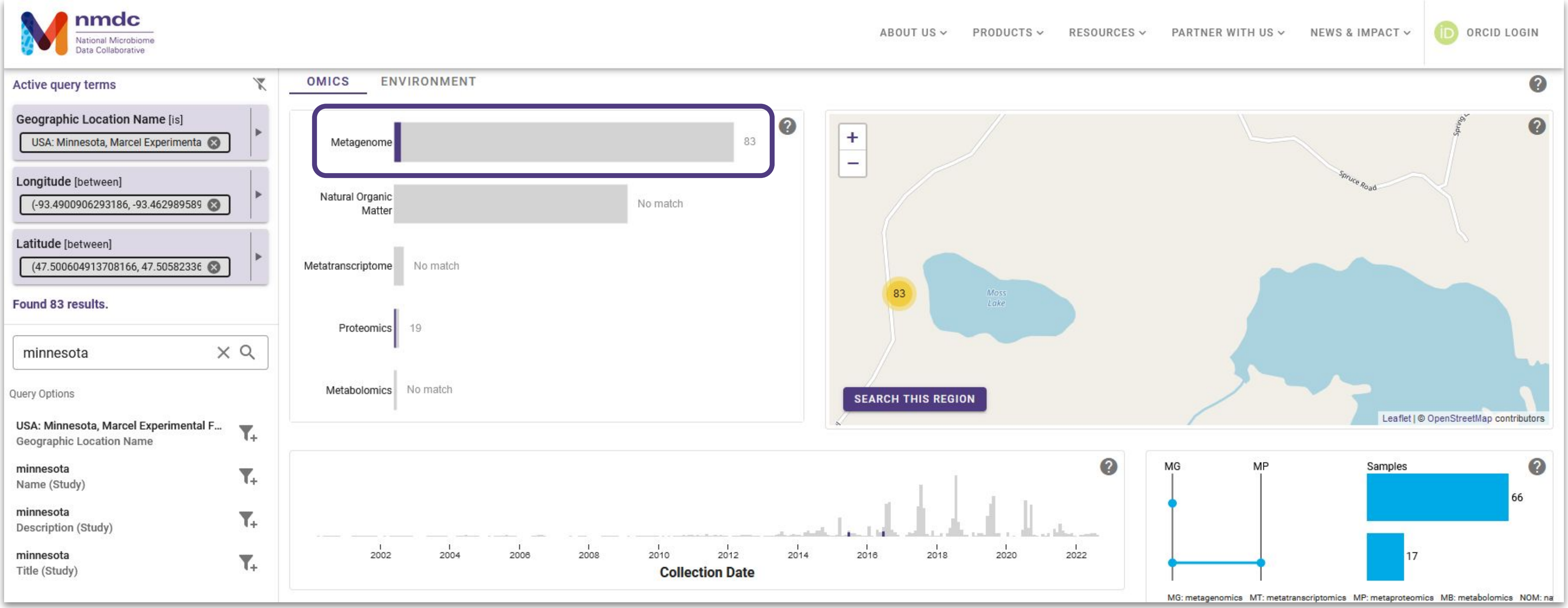
Scavenger Hunt

Scavenger Hunt

How many metagenome samples were collected next to Moss Lake, Minnesota?

Scavenger Hunt

How many metagenome samples were collected next to Moss Lake, Minnesota?



Active query terms

Geographic Location Name [is]
USA: Minnesota, Marcel Experimenta

Longitude [between]
(-93.4900906293186, -93.462989589

Latitude [between]
(47.500604913708166, 47.50582336

Found 83 results.

minnesota

Query Options

USA: Minnesota, Marcel Experimental F...
Geographic Location Name

minnesota
Name (Study)

minnesota
Description (Study)

minnesota
Title (Study)

Scavenger Hunt

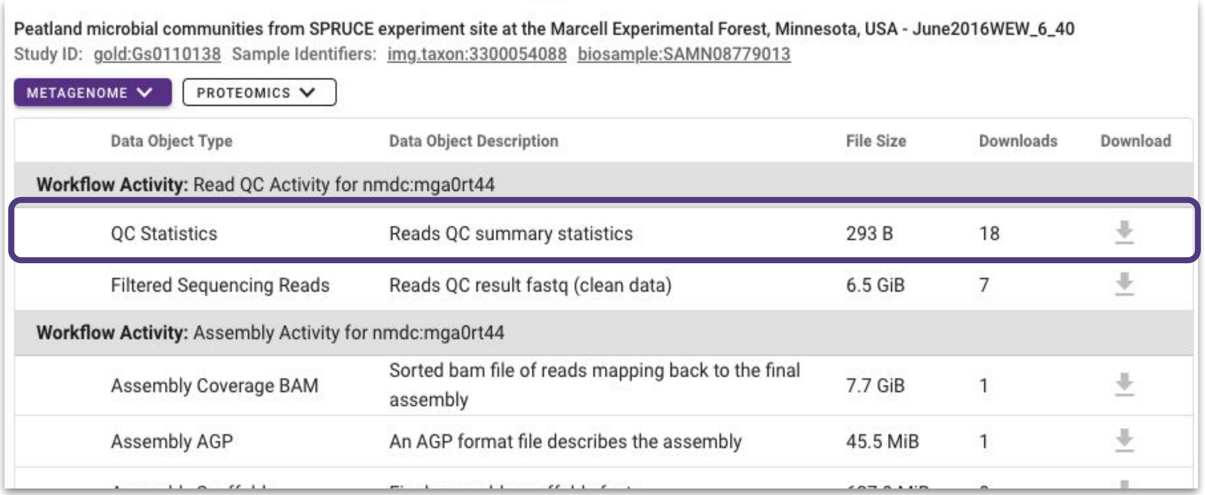
How many metagenome samples were collected next to Moss Lake, Minnesota?

- Find where to download the metagenome Reads QC statistics file from the first resulting sample
- Find the link to the data in IMG

Scavenger Hunt

How many metagenome samples were collected next to Moss Lake, Minnesota?

- Find where to download the metagenome Reads QC statistics file from the first resulting sample



Peatland microbial communities from SPRUCE experiment site at the Marcell Experimental Forest, Minnesota, USA - June2016WEW_6_40

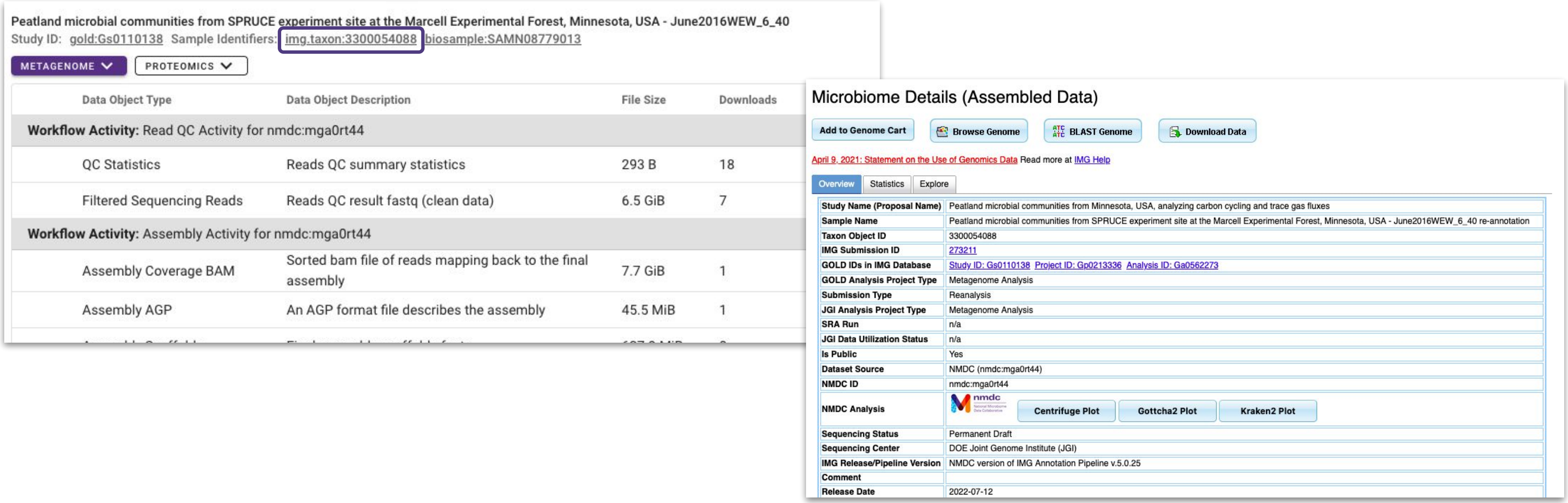
Study ID: gold:Gs0110138 Sample Identifiers: img.taxon:3300054088 biosample:SAMN08779013

Data Object Type	Data Object Description	File Size	Downloads	Download
Workflow Activity: Read QC Activity for nmdc:mga0rt44				
QC Statistics	Reads QC summary statistics	293 B	18	⬇
Filtered Sequencing Reads	Reads QC result fastq (clean data)	6.5 GiB	7	⬇
Workflow Activity: Assembly Activity for nmdc:mga0rt44				
Assembly Coverage BAM	Sorted bam file of reads mapping back to the final assembly	7.7 GiB	1	⬇
Assembly AGP	An AGP format file describes the assembly	45.5 MiB	1	⬇

Scavenger Hunt

How many metagenome samples were collected from peatland soil in 2016?

- Find the link to the data in IMG



Peatland microbial communities from SPRUCE experiment site at the Marcell Experimental Forest, Minnesota, USA - June2016WEW_6_40
Study ID: gold:Gs0110138 Sample Identifiers: img.taxon:3300054088 biosample:SAMN08779013

Data Object Type	Data Object Description	File Size	Downloads
Workflow Activity: Read QC Activity for nmdc:mga0rt44			
QC Statistics	Reads QC summary statistics	293 B	18
Filtered Sequencing Reads	Reads QC result fastq (clean data)	6.5 GiB	7
Workflow Activity: Assembly Activity for nmdc:mga0rt44			
Assembly Coverage BAM	Sorted bam file of reads mapping back to the final assembly	7.7 GiB	1
Assembly AGP	An AGP format file describes the assembly	45.5 MiB	1

Microbiome Details (Assembled Data)

April 9, 2021: Statement on the Use of Genomics Data Read more at IMG Help

Study Name (Proposal Name)	Peatland microbial communities from Minnesota, USA, analyzing carbon cycling and trace gas fluxes
Sample Name	Peatland microbial communities from SPRUCE experiment site at the Marcell Experimental Forest, Minnesota, USA - June2016WEW_6_40 re-annotation
Taxon Object ID	3300054088
IMG Submission ID	273211
GOLD IDs in IMG Database	Study ID: Gs0110138 Project ID: Gp0213336 Analysis ID: Ga0562273
GOLD Analysis Project Type	Metagenome Analysis
Submission Type	Reanalysis
JGI Analysis Project Type	Metagenome Analysis
SRA Run	n/a
JGI Data Utilization Status	n/a
Is Public	Yes
Dataset Source	NMDC (nmdc:mga0rt44)
NMDC ID	nmdc:mga0rt44
NMDC Analysis	Centrifuge Plot Gottcha2 Plot Kraken2 Plot
Sequencing Status	Permanent Draft
Sequencing Center	DOE Joint Genome Institute (JGI)
IMG Release/Pipeline Version	NMDC version of IMG Annotation Pipeline v.5.0.25
Comment	
Release Date	2022-07-12

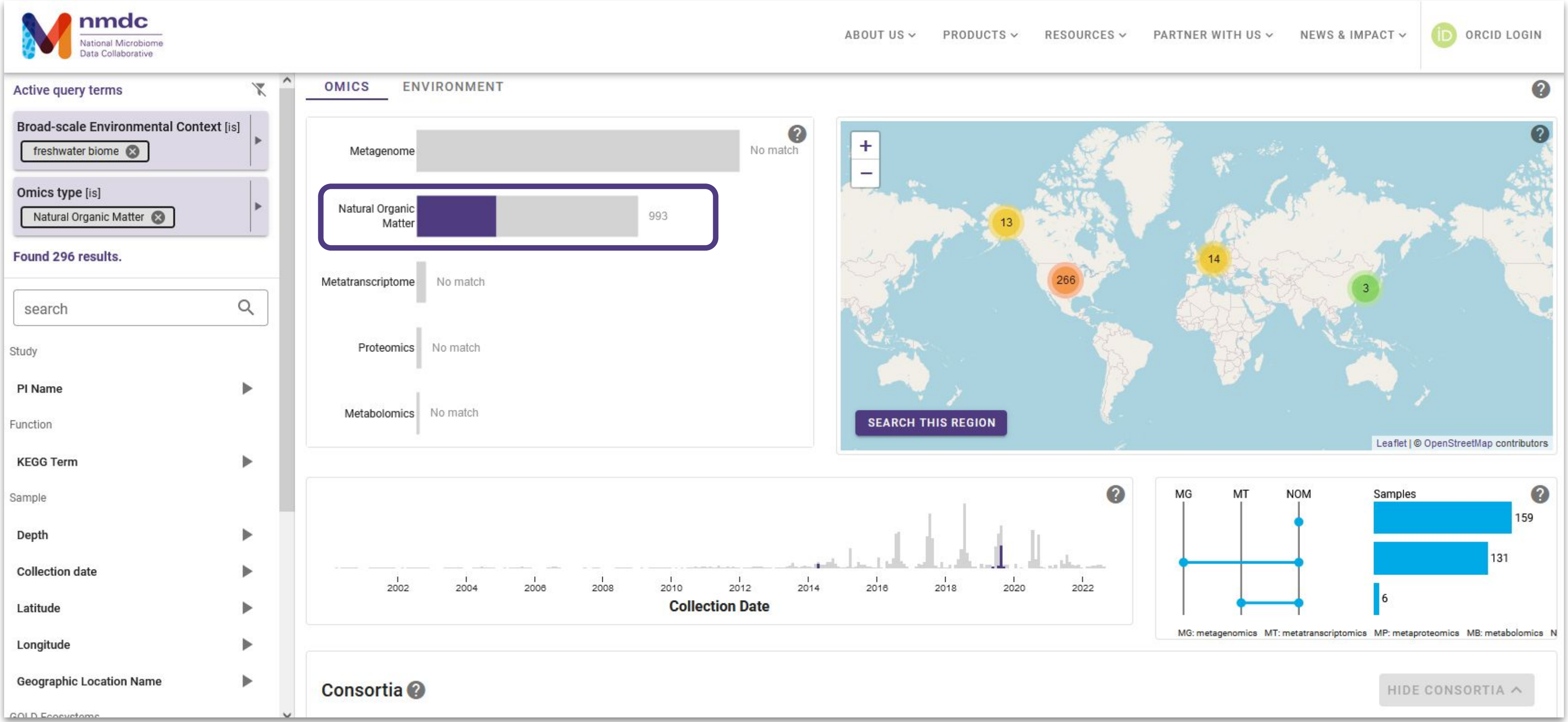
Scavenger Hunt

How many Natural Organic Matter samples are from Freshwater biomes?

Scavenger Hunt

nmdc
National Microbiome Data Collaborative

How many Natural Organic Matter samples are from Freshwater Biomes?



Active query terms

Broad-scale Environmental Context [is]
freshwater biome

Omics type [is]
Natural Organic Matter

Found 296 results.

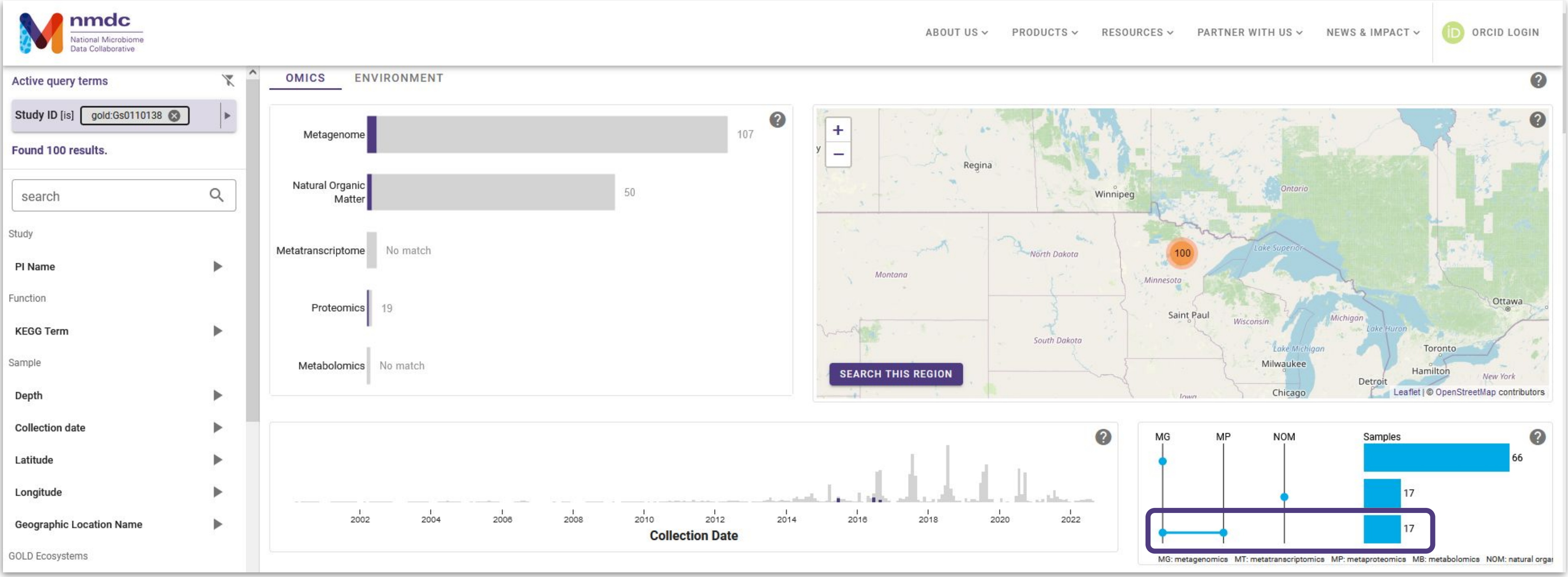
Scavenger Hunt

How many SPRUCE biosamples have been characterized by both metagenomics *and* metaproteomics?

Scavenger Hunt

nmdc National Microbiome Data Collaborative

How many SPRUCE biosamples have been characterized by both metagenomics *and* metaproteomics?



Active query terms

Study ID [is] gold:Gs0110138

Found 100 results.

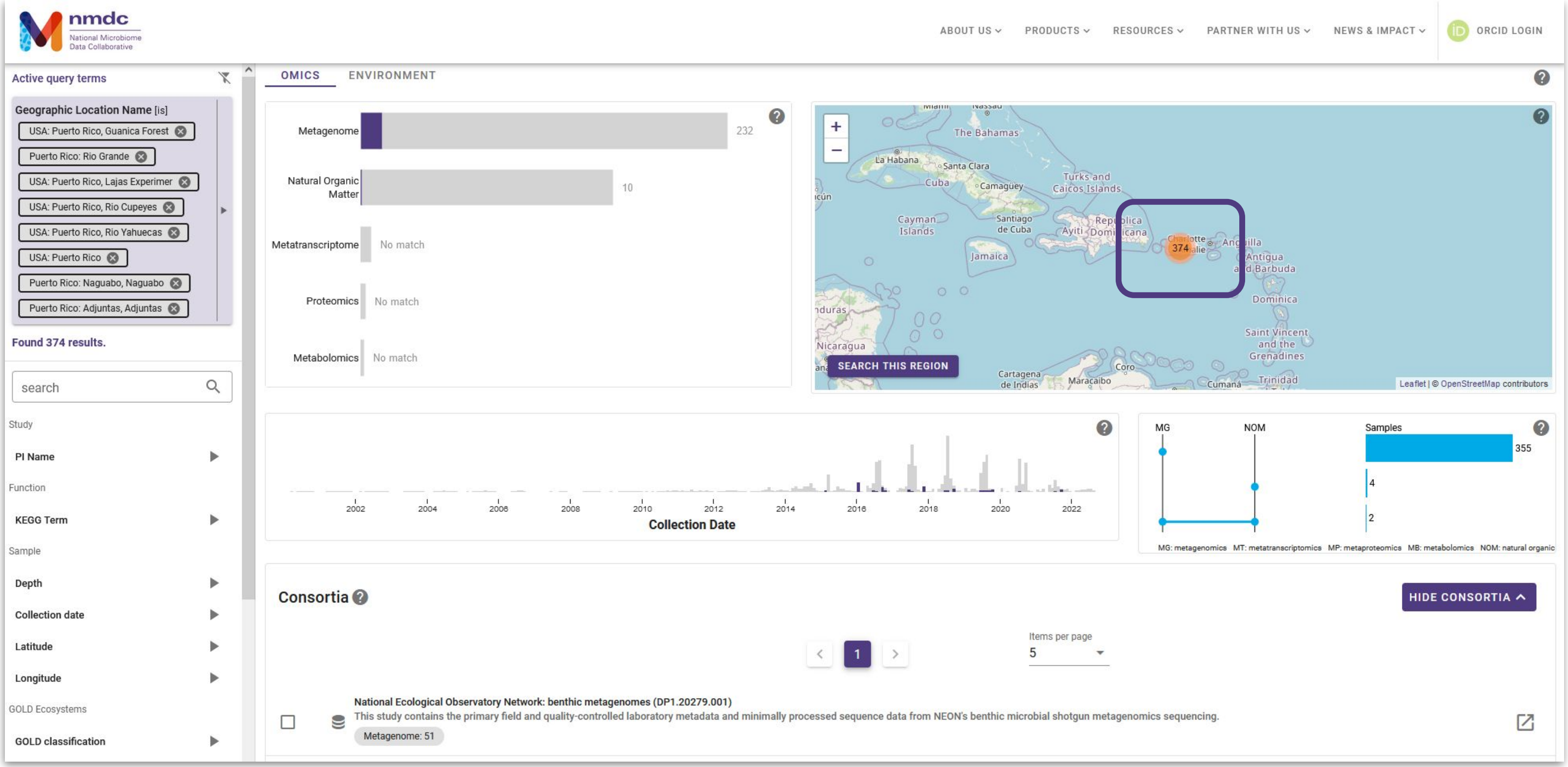
Scavenger Hunt

How many samples are from Puerto Rico?

Scavenger Hunt

nmdc
National Microbiome
Data Collaborative

How many samples are from Puerto Rico?



Active query terms

Geographic Location Name [is]

- USA: Puerto Rico, Guanica Forest
- Puerto Rico: Rio Grande
- USA: Puerto Rico, Lajas Experimer
- USA: Puerto Rico, Rio Cupeyes
- USA: Puerto Rico, Rio Yahuecas
- USA: Puerto Rico
- Puerto Rico: Naguabo, Naguabo
- Puerto Rico: Adjuntas, Adjuntas

Found 374 results.

search

Study
- PI Name

Function
- KEGG Term

Sample
- Depth
- Collection date
- Latitude
- Longitude

GOLD Ecosystems
- GOLD classification

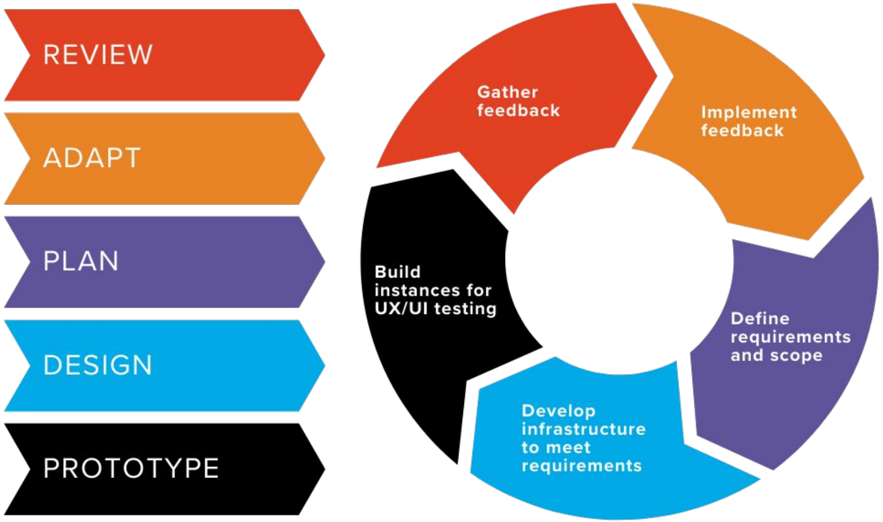
Consortia

□ **National Ecological Observatory Network: benthic metagenomes (DP1.20279.001)**
This study contains the primary field and quality-controlled laboratory metadata and minimally processed sequence data from NEON's benthic microbial shotgun metagenomics sequencing.
Metagenome: 51

Data Portal Feedback

- Were the metadata search fields comprehensive? Are there any other metadata terms you would want to search or sort by?
- Favorite feature(s)? Least favorite?
- Was the Data Portal intuitive to use?
- What did you look for? What were you hoping to find? Could you always find what you were looking for?
- Any sticking points?
- Other datasets or data types you'd like to see?
- Other questions?

Community-centered design process



REVIEW

ADAPT

PLAN

DESIGN

PROTOTYPE

Gather feedback

Implement feedback

Define requirements and scope

Develop infrastructure to meet requirements

Build instances for UX/UI testing

34
User research participants

321
Insights from user interviews

120
Action items from insights

Data Portal Feedback

- Bulk data download updates
- Visualization updates - map, upset plot
- Implementing taxonomy-based search
- Improved search and filtering

nmdc

National Microbiome Data Collaborative

Questions?

NMDC Resources

Website: https://microbiomedata.org/
Data Portal: https://data.microbiomedata.org/
Submission Portal: https://data.microbiomedata.org/submission/home
NMDC EDGE: https://nmdc-edge.org/home
Github: https://github.com/microbiomedata
Docker Hub: https://hub.docker.com/u/microbiomedata
Documentation:
https://nmdc-documentation.readthedocs.io/en/latest/overview/nmdc_overview.html
YouTube: https://www.youtube.com/channel/UCyBqKc46NQZ_YgZIKGYeglw/featured

Get involved!

Sign up for our newsletter
microbiomedata.org

Become a NMDC Champion
bit.ly/champions-program

Find us on X/Twitter
@microbiomedata

Find us on LinkedIn
https://bit.ly/NMDC_LinkedIn

Find us on Instagram
@microbiomedata

Read more about the NMDC

Kelliher JM *et al.* Cohort-based learning for microbiome research community standards. *Nat Microbiol* (2023). doi.org/10.1038/s41564-023-01361-7.

Hu B, Canon S, Eloe-Fadrosh EA, et al.. Challenges in Bioinformatics Workflows for Processing Microbiome Omics Data at Scale. *Front Bioinform*. 1:826370. (2022) doi: 10.3389/fbinf.2021.826370.

Eloe-Fadrosh EA *et al*. The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res.* 7;60(D1):D828–D836. (2022) doi: 10.1093/nar/gkab990.

Wood-Charlson, E.M., Anubhav, Auberry, D. *et al.* The National Microbiome Data Collaborative: enabling microbiome science. *Nat Rev Microbiol* **18,** 313–314 (2020). doi.org/10.1038/s41579-020-0377-0

Vangay, P *et al.* Microbiome metadata standards: Report of the National Microbiome Data Collaborative's workshop and follow-on activities. *mSystems* **6**, e01194-20 (2021). doi.org/10.1128/mSystems.01194-20