nmdc
National Microbiome
Data Collaborative

National Microbiome Data Collaborative

Progress Report 2021
U.S. Department of Energy

Vision

Empower the research community to harness microbiome data exploration and discovery through a collaborative integrative data science ecosystem.

Mission

Work with the community to iteratively develop and pilot an integrated, open-source microbiome science gateway that leverages existing resources and enables comprehensive access to multidisciplinary microbiome data and standardized, reproducible data products.

Cover credit: Marilyn Sargent, Berkeley Lab
This page credit: Roy Kaltschmidt, Berkeley Lab

Forging a Collaborative Path for Microbiome Data Science

Amidst the uncertainty of COVID-19, our team at the National Microbiome Data Collaborative (NMDC) has made remarkable progress this past year. We have pushed forward our strategic priorities — infrastructure and engagement — to launch the **Data Portal**, build bioinformatic workflows, and engage members of the microbiome community. I am proud of the progress the NMDC team has made in building an integrated data science ecosystem for microbiome research using an inclusive and collaborative community-driven approach.

In March 2021, our team launched the Data Portal with valuable input from researchers in the microbiome field. The Data Portal is built on top of a Data Schema that enables scientific discovery through adoption of standards and coordination with the Genomic Standards Consortium (GSC). The NMDC team deployed meaningful, intuitive architecture and bioinformatics workflows, through direct contributions by the research community. These workflows are publicly available as stand-alone containers, offering a unique opportunity for any institute or individual to obtain, install, and run the workflows in their own computing environments.

Emiley Eloe-Fadrosh
National Microbiome Data Collaborative Lead

Important to the workflow development is the close coordination with software developers at two US Department of Energy (DOE) user facilities — the Joint Genome Institute (JGI) and the Environmental Molecular Sciences Laboratory (EMSL) — and at Los Alamos National Laboratory (LANL) to ensure the NMDC leverages these capabilities and integrates updates (**learn more on page 7**). Our collaboration with the JGI's Integrated Microbial Genomes and Microbiomes (IMG/M) team allows us to provide read-based analysis services to augment the assembly-based annotation and comparative analysis tools in IMG/M. Our team's focus on infrastructure development has led to collaborations and key outputs including:

1. Publication of "The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource" in the *Nucleic Acids Research* database issue.
2. Development of a feature in the Data Portal to link to associated biosample records in the JGI's Genomes OnLine Database (GOLD) and annotated multi-omics data in IMG/M.
3. Integration of several NMDC workflows into the Empowering the Development of Genomics Expertise (EDGE) web-based interface for point-and-click access for data processing.

The NMDC team's cross-laboratory structure has enabled partnerships across a variety of DOE projects. While taking advantage of the DOE user facilities' capabilities, we have deepened these partnerships by working with DOE Science Focus Area (SFA) projects that use the JGI and EMSL to generate data from their samples. We have also worked closely with the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) and DOE Systems Biology Knowledgebase (KBase) teams to support data exchange across systems, and coordinate directly with research teams to learn how to better support their science. These collaborations are lowering barriers to data discovery and interoperability.

At Oak Ridge National Laboratory (ORNL), our team is coordinating with the Bio-Scales project, which aims to understand how genes influence ecosystem-level processes in plant-microbiome systems. The Bio-Scales team's first suite of samples have been processed at the JGI, ensuring that data collection, processing, analysis, and management adhere to widely recognized and accessible data standards by following the NMDC team's standard recommendations. These data will be integrated in the Data Portal in early 2022, serving as a model for Findable, Accessible, Interoperable, and Reusable (FAIR) Principles for environmental microbiome science (learn more on page 24).

Beyond our work within the DOE, we continue to develop our partnerships with scientific societies and researchers. This past year, the NMDC collaborated with the American Society for Microbiology (ASM) to launch the ASM Microbiome Data Prize recognizing outstanding contributions at the intersection of microbiome research and computing science. To expand our reach in the microbiome community, we developed and launched the Ambassador Program for 13 early career researchers across the microbiome field who are passionate about FAIR microbiome data creation. Our Ambassadors have hosted or presented in 10 NMDC-related events since their initiation in July, including a collaborative seminar with the Microbiome Centers Consortium.

In the coming year, our focus will be increasing the number of studies and multi-omics data available in the Data Portal and expanding strategic partnerships. Our team will develop and launch a sample submission system that will integrate with the JGI and EMSL to

collect study and sample information from users, making it easier for microbiome researchers to adhere to community-driven standards and submit metadata to the Data Portal. We will continue to build out data exchange services across microbiome resources and work with the KBase team to support advanced analyses from available NMDC data.

The NMDC is excited to begin working across federal agencies. Our team is developing a key strategic partnership with the National Science Foundation's (NSF) National Ecological Observatory Network (NEON). NEON, an environmental infrastructure facility, monitors ecosystems across the US with the mission of serving as a long-term ecological observatory for collection and dissemination of ecological data. The NMDC and NEON have a shared vision of promoting open access to environmental data to enhance scientific discovery.

As 2021 comes to a close, the NMDC team is well on our way to building a collaborative, integrative data science ecosystem for microbiome research. We invite you to read on to learn more about our team's accomplishments this past year, and we hope that you will engage with the NMDC to shape the future of microbiome data discovery.

Emiley Eloe-Fadrosh
National Microbiome
Data Collaborative Lead

Credit: Marilyn Sargent, Berkeley Lab

A Distributed Data Infrastructure for Microbiome Research

The NMDC is tackling infrastructure challenges through iteratively developing an open-source, integrated ecosystem for multi-omics data. Metadata standards and consistently processed multi-omics data are essential to successfully developinga FAIR microbiome data infrastructure. This past year has marked the launch of the Data Portal, broad access to our suite of bioinformatics workflows, and important technical developments in support of evolving metadata standards.

Standards for FAIR Multi-omics Data

The development and adoption of standards is key to the success of the NMDC, and key to being able to integrate and search across data derived from different studies and stored in distributed data resources. The Data Schema has been developed to define how metadata elements (e.g., biosamples, environmental descriptors) are related. This schema has been applied to a translation process that integrates metadata from multiple environmental data repositories (e.g., the JGI's GOLD and EMSL's NEXUS). Coordinating community-driven development of standardized metadata formats is an essential first step for data to be findable and interoperable. In 2021, the NMDC team formalized the Data Schema using the Linked data Modeling Language (LinkML), which leverages existing data exchange standards, enables automatic validation of data, and allows the standard to be machine-readable. Together with the GSC, the NMDC team has made the Minimum Information about any (x) Sequence (MIxS) standardized templates for describing the environmental contexts computationally accessible beyond a set of spreadsheets. This transition marks a new phase for how genomic standards are updated and used by the research community, and paves the way for future omics standards.

ACCOMPLISHMENTS

- **Developed metadata curation processes using community-driven standards and in collaboration with research teams**
- **Developed machine-readable metadata templates in collaboration with the GSC**
- **Expanded the Data Schema to support amplicon data standards**
- **Submitted new terms in the Environment Ontology (EnvO) to meet community needs**

The NMDC team has directly contributed to the enhancement of GOLD metadata and driven new updates in the GOLD system. In collaboration with the GOLD team, we devised a system for mapping GOLD ecosystem classification path descriptors to the EnvO to be fully compliant with the GSC MIxS standards. Our team has worked closely with the GOLD team to curate over 63,000 biosamples. Going forward, the NMDC team is working closely with staff at the JGI and EMSL to develop a streamlined sample submission system. This new work, initiated towards the end of 2021, is in close collaboration with the **DataHarmonizer** team at Simon Fraser University. Adoption of the DataHarmonizer tool will advance our efforts to lower barriers to sample metadata submission and support community-driven standards. As we look forward to the coming year, the NMDC team will release the NMDC sample submission system as part of the Data Portal to broadly support submission of new studies, samples, and associated standardized metadata.

LEARN MORE

Credit: Roy Kaltschmidt, Berkeley Lab

Bioinformatics Workflows for Data Integration

The lack of standardardized bioinformatics methods impedes researcher's ability to compare microbiome data generated by many research groups. To address this issue, the NMDC team has integrated existing open-source bioinformatics workflows, which can be used to process raw multi-omics data and produce interoperable and reusable annotated data from metagenome, metatranscriptome, metaproteome, metabolome, and natural organic matter characterizations. The workflows include production-quality bioinformatics tools developed by the JGI, EMSL, and LANL. To support broad availability of these workflows, the NMDC team has made the metagenomic (including quality control, read-based analyses, assembly, annotation, and binning), metatranscriptomic, metabolomic, metaproteomic, and natural organic matter workflows available as **stand-alone downloadable software containers**.

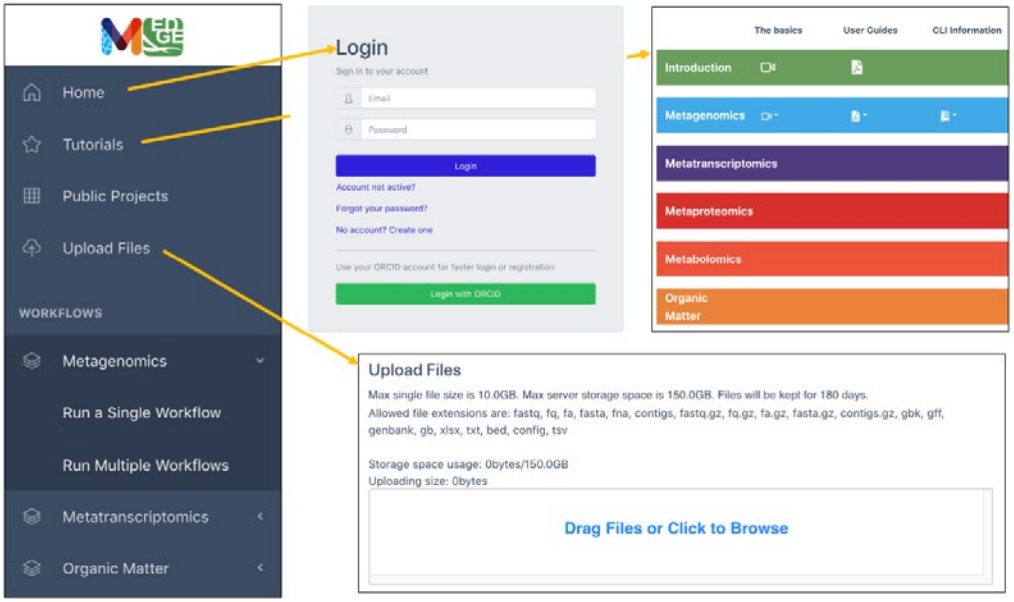
The **NMDC EDGE** platform has been developed to provide an intuitive, user-friendly graphical interface and currently supports the metagenomic and natural organic matter workflows. Detailed **workflow documentation** and **tutorials** enable both novice and experienced users to understand the various components of the data processing steps. The NMDC EDGE platform has

ACCOMPLISHMENTS

- **Made available all NMDC workflows as stand-alone software containers in Docker Hub and codebase in GitHub**
- **Created a beta version of NMDC EDGE, a user-friendly user interface for workflow execution that supports user uploaded files for metagenome and natural organic matter workflows**
- **Expanded documentation for all workflows and created NMDC EDGE tutorials and user guides to increase ease of use**

been deployed and made available to process multi-omics data through a partnership with the San Diego Supercomputer Center (SDSC) and leverages the NSF's Extreme Science and Engineering Discovery Environment (XSEDE) network. In 2021, the research community was invited to beta test the metagenomics and natural organic matter workflows in NMDC EDGE, resulting in 24 recommendations for the user interface and documentation. The NMDC team will continue to gather and incorporate feedback from the community as new workflows are released.

LEARN MORE



The left menu bar in NMDC EDGE allows users to log in, access video and written tutorials, upload data using an intuitive drag and drop feature, and select desired workflows to run.

A New Data Portal for Multi-omics Search and Access

The NMDC infrastructure leverages the DOE's high-performance supercomputing facilities with our core Data Portal infrastructure and centralized metadata store residing within the National Energy Research Scientific Computing Center (NERSC) Spin service. Our approach relies on a data federation with distributed compute resources, which will enable our architecture to scale to increasing volumes of multi-omics data in the coming years. The NMDC infrastructure does not serve as a centralized data repository, and instead supports integration across distributed resources at the JGI and EMSL. The NMDC metadata store implements an application programming interface (API) that allows search of the data and communication with satellite data storage sites at the EMSL and LANL. This past year, we have expanded this architecture to include workflow execution at the SDSC to demonstrate a model for diverse, integrated distributed resources outside of the DOE's supercomputing facilities.

The NMDC team publicly released the NMDC Data Portal in March 2021. This initial release included three multi-omics projects with data generated from the Facilities Integrating Collaborations for User Science (FICUS)

ACCOMPLISHMENTS

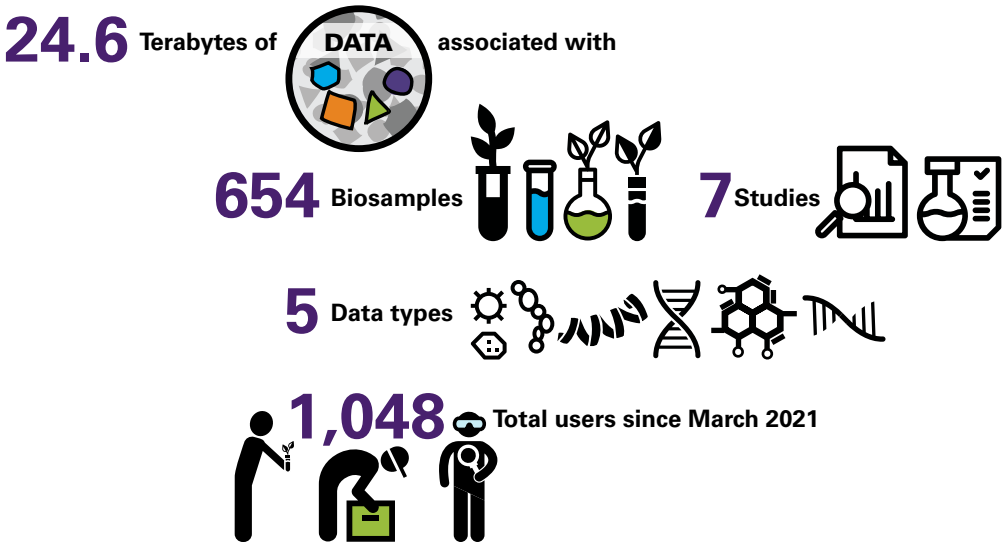
- **Launched the Data Portal in March 2021**
- **Developed new features including:**
  - **Faceted search based on sample environment metadata**
  - **Functional annotation search across multi-omics**
  - **Integrated Contributor Roles Taxonomy (CRediT) to support researcher attribution**
  - **Links to primary data resources i.e. National Center for Biotechnology Information, IMG/M, GOLD, and ESS-DIVE**
- **Published in the *Nucleic Acids Research* database issue**
- **Created an in-depth user guide for the Data Portal**

program from NMDC collaborators. Since the launch of the Data Portal, the NMDC team continues to add studies, samples, and standardized data products from our multi-omics bioinformatics workflows. The Data Portal currently contains 10.2 terabytes of data associated with 638 biosamples, 7 studies, and 5 data types from a breadth of environmental microbiomes, spanning river sediments, subsurface shale carbon reservoirs, plant-microbe associations, and temperate and tropical soils.

A primary feature of the Data Portal is enabling the research community to discover multi-omic data through a variety of search capabilities, such as faceted search and interactive visualizations using both the environmental metadata and functional annotations. Our team has developed a detailed user guide to help users understand the capabilities of the Data Portal.

LEARN MORE

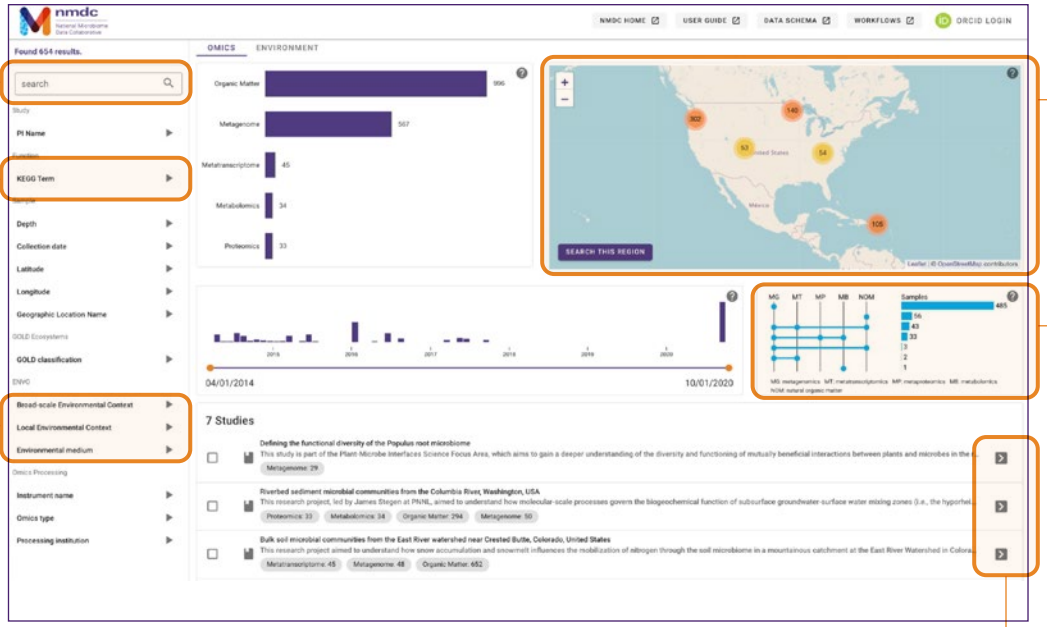
DATA PORTAL STATS



NMDC Data Portal

Search by text with automatic suggestions for available query options

Search by functional annotation



Search by geography using the interactive map

Select and filter by bio-samples that have been characterized with one or more types of omics data

Select and filter by exploring EnvO terms in the hierarchical structure within each facet

Go to study pages to see publications related to the data and learn more about the team who collected the biosamples

User-Centered Design Process

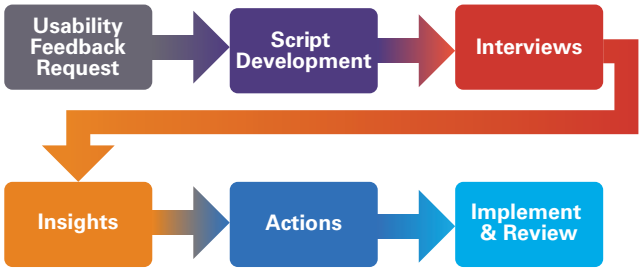
The NMDC is a resource designed with and for the scientific community. We have engaged in extensive user research through interviews and direct collaboration with the scientific community that have informed the design, development, and display of data through the Data Portal. Our user-centered design methodology enables the scientific community to provide feedback leading to iterative and continuous improvement of our systems and ensuring that our systems enable a high level of scientific productivity. Over the past year, we have conducted four rounds of user interviews (31 user interviews total) focused on targeted goals:

1. Understanding the current challenges of searching and accessing microbiome data with available resources.
2. Gathering usability feedback to improve general search and discovery in the Data Portal.
3. Gathering usability feedback to improve targeted search and download in the Data Portal.
4. Gathering usability feedback to improve a proof-of-concept sample submission system.

This user-centered design process, shown to the right, resulted in the generation of 45 user insights that have led to the creation of key features in the Data Portal, such as the ability to bulk download data and use an interactive map visualization to search for samples.

ACCOMPLISHMENTS

- **Developed processes for iterative feedback from the research community for bioinformatics workflows, the Data Portal, and sample submission interface**
- **Generated 45 user insights from 31 user interviews**
- **Created features such as bulk download and interactive map visualizations based on user feedback**



User-centered design process.

Integration Across DOE User Facilities and Computational Resources

Resource	Mission, Scope, and Capabilities Integration with the NMDC	Integration with the NMDC
IMG/M JGI, Lawrence Berkeley National Laboratory (Berkeley Lab)	**Integrated Microbial Genomes and Microbiomes:** A JGI resource for the integration of sequence data to support the annotation, analysis, and distribution of microbial genomes, metagenomes, and metatranscriptomes. Read more	IMG/M's production-quality annotation workflow is used by the NMDC with the JGI's quality control and assembly components. The NMDC team collaborated with the IMG team platform to provide read-based analysis services to augment the assembly-based annotation and comparative analysis tools in IMG/M. In the fall 2021 quarterly release, we provided a new feature in the Data Portal to link to associated biosample records in GOLD and annotated multi-omics data in IMG/M.
GOLD JGI, Berkeley Lab	**Genomes OnLine Database:** A JGI resource for comprehensive access to information regarding genome, metagenome, and metatranscriptome sequencing projects, and their associated metadata. Read more	The NMDC has contributed to the enhancement of GOLD metadata through updated EnvO curation complementing GOLD's five-level ecosystem classification. Over 63,000 biosamples have been curated with at least one EnvO term.
EnvO Open Biological and Biomedical Ontology Foundry, Berkeley Lab	**Environment Ontology:** A community- driven ontology for the concise, controlled description of environments. Read more	The NMDC search functionality uses a faceted approach including EnvO annotation. The use of EnvO terms in the Data Portal facilitates dataset aggregation from different sources, and increases the findability and interoperability of the data.
KBase Berkeley Lab	**DOE Systems Biology Knowledgebase:** A software and data science platform designed to meet the grand challenge of systems biology: predicting and designing biological function. KBase integrates data and tools in a unified graphical interface. Read more	By integrating KBase's platform with the NMDC search APIs, users of KBase will have access to NMDC data and the open-source components of the NMDC workflows for sequence data. These data can be used to create a KBase Narrative, an interactive digital notebook. A prototype model has been generated, and we will continue to build out data exchange to support advanced analyses.

Continued on page 16

Resource	Mission, Scope, and Capabilities Integration with the NMDC	Integration with the NMDC
EDGE LANL	**Empowering the Development of Genomics Expertise:** A bioinformatics platform that can be installed anywhere providing a user-friendly web-based interface to enable novice microbiome and bioinformatics users to process and explore the analysis of raw multi-omics data. Read more	The NMDC workflows have been integrated into the EDGE web-based interface for point-and-click access to data processing. NMDC EDGE is hosted at the SDSC to allow any user to run their own multi-omics data through the available workflows.
NEXUS Pacific Northwest National Laboratory (PNNL)	**NEXUS:** A central data repository for semi-automated capture and storage of data generated by EMSL staff and users, including the capture of basic metadata associated with instruments and user proposals.	The data standards, ontologies, and standardized workflows developed through the NMDC will form the basis of the community standards used by EMSL. The NEXUS repository will be the primary data provider of metaproteomics, natural organic matter, and metabolomics data to the Data Portal for integration.
DAAC ORNL	**Distributed Active Archive Center:** A data center for biogeochemical measurements and other Earth system data as part of the NASA Earth Observing System Data and Information System.	To enhance the available biosample metadata, the NMDC team has integrated API calls to DAAC servers to obtain environmental data based on geolocation information including ground elevation and Zobler soil types.
ESS-DIVE Berkeley Lab	**Environmental Systems Science Data Infrastructure for a Virtual Ecosystem:** A repository of DOE-generated earth science data.	The NMDC team has identified studies that are supported within the ESS-DIVE archive, and this past year implemented links from the Data Portal to ESS-DIVE data to associate microbiome data and complementary environmental measurements. We plan to work closely with the ESS-DIVE team to expand data exchange services and infrastructure links.

Advancing Microbiome Science, Together with the Research Community

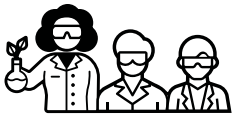
The cross-cutting nature of microbiome research, from environmental sciences, agriculture, and energy, to human, natural, and built environments, creates a diverse community united by a common challenge: the need for seamless data discovery, exploration, and access. In order to develop inclusive solutions to address this challenge, the NMDC team works closely with the research community to learn and better understand specific needs across the diverse array of microbiome research.

Credit: R Dodsworth, Joint Genome Institute

NMDC Community Engagement

The NMDC's tiered engagement model supports anyone working with microbiome research data to be part of the NMDC community at a level of involvement that best suits them. We consider individuals who sign up for our newsletters, follow us on Twitter, or attend our events to be part of our community. Those who want to be more involved with the NMDC are invited to apply to join the NMDC Champions program, where they can be among the first in line to provide input on NMDC activities. Early career researchers who want to be more involved in formally partnering with the NMDC can apply to become NMDC Ambassadors. Our Ambassadors receive training and support to engage with their respective communities to advance the NMDC mission in a one-year cohort-based program.

BY THE NUMBERS



13 AMBASSADORS

24 CHAMPIONS

26 Events hosted, including 10 from NMDC Ambassadors

1000+ People reached at events

ACCOMPLISHMENTS

- **Launched the NMDC Ambassador Program with 13 outstanding early career microbiome researchers**
- **Published the first NMDC community manuscript, "Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative's Workshop and Follow-On Activities," in *mSystems***
- **Launched the monthly NMDC Community Conversations series**
- **Developed the Inclusion, Diversity, Equity and Accountability Strategic Plan to ensure the NMDC supports a diverse research community**

Promoting Community Adoption of Data Standards

The NMDC team hosted a workshop in 2019 with over 50 attendees with expertise in microbiome research, data standards, genome annotation, bioinformatics, and community engagement to discuss shared challenges and solutions to improving community adoption of and compliance with data standards. In early 2021, the NMDC team and workshop participants summarized their workshop discussions with a list of recommendations (Vangay et al. 2021), including the need to reduce barriers to data submission by understanding how communities are currently using data standards and providing training for a variety of learning styles.

Further, the NMDC team conducted a community survey in early 2021 to understand how the microbiome research community currently shares, accesses, and reuses data. The survey received responses from 773 individuals across a variety of scientific disciplines and organizations spanning academia, industry, government, and nonprofit sectors. From this survey, the NMDC team found that microbiome data search criteria are primarily based on environmental context and omic data type of interest, and researchers prefer a graphical user interface (GUI) for both searching and downloading of data. These preferences are supported as primary features of the Data Portal. The survey results also showed that while about a third of respondents have used the GSC MIxS standard, another third of respondents have never used any data standard, irrespective of domain of research.

While metadata standards are key to the integration and search of microbiome data across studies and distributed data resources, the NMDC workshop recommendations and the NMDC community survey underscore that there are significant barriers to the adoption of these standards. The NMDC team is contributing to and linking community-driven metadata standards, and is currently developing an interactive and intuitive sample submission system with built-in use of standards. To further this effort, the NMDC launched the Ambassador Program to provide training and support for early career researchers who are motivated to engage with their respective research communities to lower barriers to adoption of metadata standards.

NMDC Ambassadors

NMDC Ambassadors are early career researchers who are motivated to engage with their research communities to support the generation of FAIR microbiome data. Ambassadors are an important part of the NMDC's goal of creating culture change to encourage use of metadata standards. In July 2021, through a competitive selection process, 13 researchers were selected to form the inaugural cohort, with representation from a variety of institutions, including the National Institutes of Health, Scripps Institution of Oceanography, and the US Food and Drug Administration. Since the kick-off in July, NMDC Ambassadors have grown their technical knowledge about metadata standards by attending workshops with the NMDC team and metadata experts from the GSC, Qiita, and the American Gut Project. They have also learned how to lead engaging virtual events and effective discussions through trainings led by the Center for Scientific Collaboration and Community Engagement. NMDC Ambassadors recently represented the NMDC at a Microbiomes Center Consortium seminar, have presented on the NMDC in a number of events, and have led four metadata workshops for their respective microbiome communities, reaching over 150 microbiome researchers in total. In 2022, these early career researchers will continue to host workshops and events with their communities to lower barriers to adoption of metadata standards.

MEET OUR AMBASSADORS


Arwa Abbas


Mikayla Borton


Emily Davenport


Natalia Erazo


Chloe Herman


Lisa Karstens


Brandon Kocurek


Holly Lutz


Kevin Myers


Jaci Saunders


Michael Shaffer


Emily Vogtmann


Amanda Windsor

NMDC Champions

NMDC Champions understand and appreciate the value of well-curated data, and are willing to advocate for the importance of FAIR microbiome data. NMDC Champions are the first in line to provide feedback on the usability of NMDC platforms, training materials, and many other external-facing programs. In 2021, our champions' contributions resulted in important features and enhancements on the Data Portal and NMDC EDGE platform. Several champions were also invited to participate in a number of opportunities in 2021, such as a World Microbe Forum panel discussion focused on FAIR microbiome data and an ASM congressional briefing on the broader impacts of sharing microbiome data.

MEET OUR CHAMPIONS


Yigal Achmon


Kai Blumberg


John-Marc Chandonia


Sean Cleveland


Sneha Couvillion


Justine Debelius


JP Dundore-Arias


Cassie Ettinger


Alexis Garretson


Sean Gibbons


Judson Hervey


Alex Honeyman


Bonnie Hurwitz


Ulas Karaoz


Marie Kroeger


Kate Lane


Jessica Audrey Lee


Ryan McClure


Nancy Merino


Ahmed Shibl


Venkat Subramanian


Luke Thompson


Alonna Wright


Ying Zhang

Inclusion, Diversity, Equity, and Accountability Plan

Diversity within microbiome research, in all its forms (racial, gender, sexual identity, class, and more), strengthens research teams and practice, and helps advance science. Significant parallel, nontechnical efforts are required to ensure microbiome data science, new technologies, and infrastructure developments work in the best interests of the research community and society at large. We are committed to supporting the diversity of experiences, expertise, backgrounds, needs, and perspectives of the microbiome research community, and to actively work towards an inclusive culture at a programmatic and individual level. We have developed a **strategic plan** to hold ourselves accountable to the goals listed below..

NMDC Team's Goals

Goal 1: Promote transparency and accountability within the NMDC's team and operations.

Goal 2: Promote transparency and accountability within the NMDC's governance structure.

Goal 3: Engage and support diverse stakeholders and users.

NMDC Community Conversations Series

The NMDC team has hosted three webinars in 2021 kicking off a monthly series aiming to connect with and educate the scientific community on FAIR data and its applications in microbiome research. These three Community Conversations featured discussions on data management plans, the use of ontologies in microbiome research, and how to create community champion programs to promote usage of scientific data resources, reaching over 100 community members with our live webinars and recordings. We are looking forward to many more great conversations with our community in the coming year.

LEARN MORE

"Communities provide valuable opportunities to try out new ways of doing things — which enables us to shape together the way that we do modern science."

— **Lou Woodley**
Center for Scientific Collaboration & Community Engagement*

*Lou was a panelist in our Community Conversations entitled **Developing Data Science Resources in Partnership with Scientific Communities**

Partnership Highlights

American Society for Microbiology (ASM)

The NMDC team has a longstanding partnership with the ASM, beginning with the ideation of the NMDC and program development through its program launch in 2019 at the ASM's annual conference. Our multifaceted partnership includes involvement in the ASM's microbiome stakeholder group and participation of the ASM's CEO Stefano Bertuzzi on the NMDC Scientific Advisory Board. This past year, the NMDC partnered with the ASM for three main activities, including the launch of the ASM Microbiome Data Prize, communications and science advocacy training for the NMDC Ambassadors and Champions, and a congressional microbiome briefing focused on the role the NMDC plays in developing a microbiome data ecosystem together with the scientific research community and the federal government.

Center for Scientific Collaboration and Community Engagement (CSCCE)

The CSCCE is a research and training center to support and study the emerging field of scientific community engagement, and has partnered with the NMDC team since the initiation of the NMDC program. The CSCCE has provided valuable trainings for the NMDC team and Ambassadors on best practices for community engagement and has provided a framework for planning, coordinating, and implementing engagement activities. With a shared vision of creating an engaged scientific community, the CSCCE has been instrumental in the development and success of our ambassador and champion programs, and the NMDC's continued work with the microbiome research community.

California Digital Libraries (DMPTool)

In partnership with the University of California Curation Center of the California Digital Library, the NMDC team has created a microbiome-specific DMPTool template. DMPTool is an open-source application that assists researchers in the creation of data management plans compliant with federal funding requirements. The NMDC DMPTool template supports microbiome data management best practices with specifications unique to microbiome standards and data processing. This template will help researchers create data management plans that lead to data being FAIR.

Genomic Standards Consortium (GSC)

As standards underpin all NMDC work, our team has a longstanding partnership with the GSC, a community-based organization focused on developing and maintaining internationally recognized standards for genomic data. The GSC's president, Lynn Schriml, serves on the NMDC Scientific Advisory Board, and several members of the NMDC team are heavily involved in the GSC's Compliance and Interoperability Group to support development and integration, along with NMDC Lead Eloe-Fadrosh serving on the GSC board. This past year, the NMDC team worked closely with the GSC to support adoption of the LinkML implementation of the MIxS standards.

International Microbiome and Multi-Omics Standards Alliance (IMMSA)

IMMSA, led by the National Institute of Standards and Technology (NIST), was launched this past year to support coordinating cross-cutting efforts that address microbiome measurement challenges and broadly engage microbiome-focused researchers from industry, academia, and government. NMDC Lead Eloe-Fadrosh serves on the IMMSA Steering Committee, and several NMDC team members are actively engaged in the current working groups on bioinformatics tools, workshop planning, and the newly formed metabolomics group. The NMDC was featured as part of the first quarterly IMMSA webinars in April 2021.

Microbiome Centers Consortium (MCC)

The MCC is a cooperative and collaborative network of microbiome centers led by Dr. Jennifer Martiny at UC Irvine. MCC Lead Jennifer Martiny serves on the NMDC Scientific Advisory Board, and members of the NMDC team served on the MCC seminar series committee this past year. The NMDC team actively participates in the MCC seminar series, which provides a much-needed platform for early career microbiome researchers across broad disciplines to share and discuss their research during a time when most in-person events have been canceled. The MCC partnered with the NMDC to showcase the NMDC Ambassadors in the final seminar of 2021. The Ambassadors delivered flash talks on their research and discussed how they implemented best practices in making microbiome data more accessible.

Science Highlights

The NMDC is driven by the needs of the research community, and aims to serve as a catalyst for innovative research leveraging multi-omics data across diverse environmental microbiomes. The NMDC team is collaborating with research teams doing cutting-edge environmental microbiome research and bioinformaticians building up computational capabilities for cross-study multi-omics comparative analyses. On the following pages are highlights from our research engagements this past year.

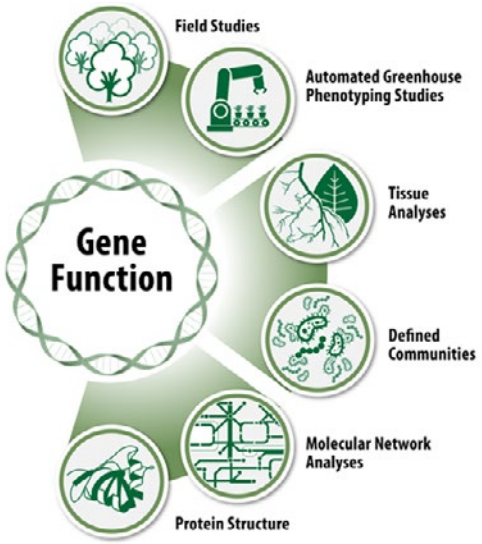
Credit: Pacific Northwest National Laboratory

Bio-Scales

Identifying gene functions across the plant-soil-microbial system

The **Bio-Scales pilot project**, led by Mitchel J. Doktycz at ORNL, is focused on the concept of rapidly identifying gene function in order to understand how genes influence ecosystem-level processes. These pilot-scale efforts are driven by the hypothesis that combinations of host and microbial traits influence nitrogen transformation patterns and fluxes across the coupled plant-soil-microbial system. The findings of this study will have important implications for ecosystem-level carbon and nitrogen cycling. The pilot project uses *Populus*, as a model plant host system and genotypes representing trait extremes (e.g., those involved in plant control of bacterial nitrogen cycling processes) across a genome-wide association mapping study population. Bio-Scales is in the process of partnering with the JGI to generate hundreds of new metagenomes and metabolomes that will be made available through the Data Portal. The collaboration of the NMDC with the Bio-Scales pilot project serves as a model for integrated and FAIR environmental microbiome science by ensuring that data collection, analysis, and management are tightly coupled and adhere to widely recognized and accessible metadata standards.

The **Bio-Scales study**, sample metadata information, and processed data are available for exploration in the Data Portal.



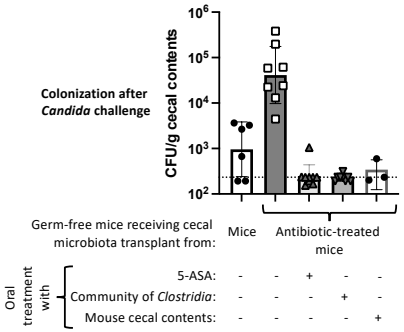
The Bio-Scales team applies a range of expertise and capabilities to translate DNA sequences into new knowledge about gene functions.

Tri-Institutional Partnership in Microbiome Research

In early 2020, the University of California, San Francisco (UCSF), UC Davis, and Berkeley Lab formed a Tri-Institutional Partnership in Microbiome Research (TrIP Microbiome) to inspire synergies in this area. This partnership aims to catalyze and fund novel, bold, and potentially transformative collaborative microbiome research projects proposed jointly by researchers at these three institutions. The unique aspect of TrIP Microbiome is its data-driven focus and data infrastructure brought through the participation of the NMDC. The NMDC is working with TrIP Microbiome researchers to catalyze experimental co-design between biologists and computational scientists, adoption of data management best practices, and open science to promote cross-study comparison and machine learning.

Colonization resistance against Candida

TEAM: Baumler, Andreas* (UC Davis); Reagan, Krystle (UC Davis); Savage, Hannah (UC Davis); Noble, Suzanne (UCSF); Tritt, Andrew (LBNL)

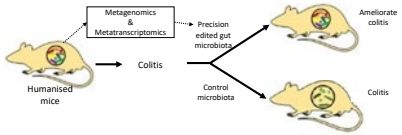


Colonization after *Candida* challenge.

This project aims to develop gut microbiome therapeutics to prevent antibiotic-induced blooms of the opportunistic pathogen *C. albicans* by identifying metabolic pathways that are linked to colonization resistance and differentially expressed during expansion of the pathogen using canine and gnotobiotic mouse models. In early 2021, the research team began enrolling patients, and has started bioinformatics analyses of the canine fecal microbiota in parallel with starting the germ-free mouse experiments this winter.

Precision editing of gut dysbiosis in inflammatory bowel disease for ameliorating inflammation

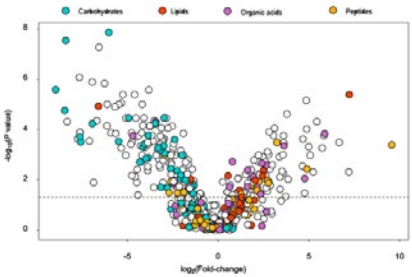
TEAM: Dave, Maneesh* (UC Davis); Arkin, Adam (LBNL)



Approach for precision editing of gut microbiota in IBD mouse model.

This project utilizes a targeted, personalized approach to develop phage and probiotic interventions to address inflammatory bowel diseases (IBD). Using an immunodeficient mouse model with fecal microbiota transplantation from IBD human patients, together with the research team's isolate and phage reference collections, the team will develop phage/probiotic interventions to target the growth of "healthy" microbiome members after inducing colitis in the humanized mice. This past year, the research team has started the mouse experiments.

FODMAP utilization by the microbiota



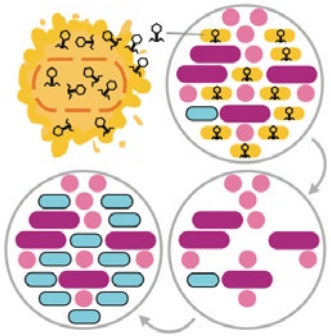
Metabolites: mice engrafted with Clostridia versus Germ-free mice.

Team: Baumler, Andreas* (UC Davis); Turnbaugh, Peter (UCSF); Eloe-Fadrosh, Emiley (LBNL)

This project evaluates the effect of dietary FODMAPs (an acronym for poorly absorbed fermentable oligosaccharides, disaccharides, monosaccharides, and polyols) on the microbiome, and their impact on gut physiology in a mouse model of irritable bowel syndrome. In 2021, the research team completed the mouse experiments and generated initial microbiome profiles using 16S rRNA sequencing. They have also generated metagenome data and submitted samples for RNAseq analysis, and are currently preparing for bioinformatics analysis.

Unraveling strain-level virus-host dynamics in diverse ecosystems

TEAM: Emerson, Joanne* (UC Davis); Brown, C. Titus (UC Davis); Turnbaugh, Peter (UCSF); Roux, Simon (LBNL)

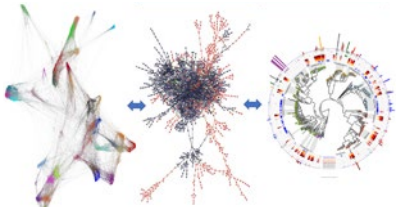


Virus-driven selective sweeps in host populations or strains.

Using existing data from agricultural soils, and new data from soil and the human gut, the research team will integrate both long- and short-read metagenomic data to assess paired virus-host strain heterogeneity and virus-host dynamics in response to disturbance, and evaluate the generalizability of observed virus-host dynamics across ecosystems. This past year, the research team has started bioinformatic analyses to test different tools for unraveling strain heterogeneity, started to develop and test pipelines for merging long-read and short-read metagenomic and viromic sequencing data, and coordinated applying soil viromic laboratory approaches to human gut samples.

MicroMetabolome: A biological and analytical framework of the human microbiome-metabolome axis

TEAM: Spitzer, Matthew* (UCSF); Fiehn, Oliver (UC Davis); Brown, James Bentley (LBNL)



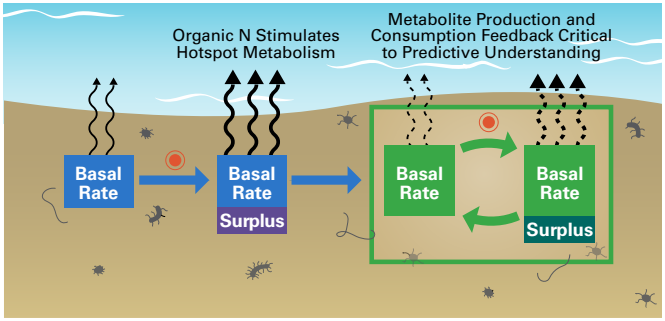
Multi-modal/Multi-OMIC high dimensional profiling approach.

This project aims to perform whole metagenome shotgun sequencing and mass spectrometry on human stool samples to generate a multi-omic dataset of the fecal metagenome and metabolome of 100 healthy individuals over multiple intervals of time. In 2021, the research team has started to process the fecal samples for metabolomics analysis.

*Contact Principal Investigator

Microbes in Mixing Zones

An integrative multi-omics approach to explore impacts of groundwater-surface water mixing on microbial communities

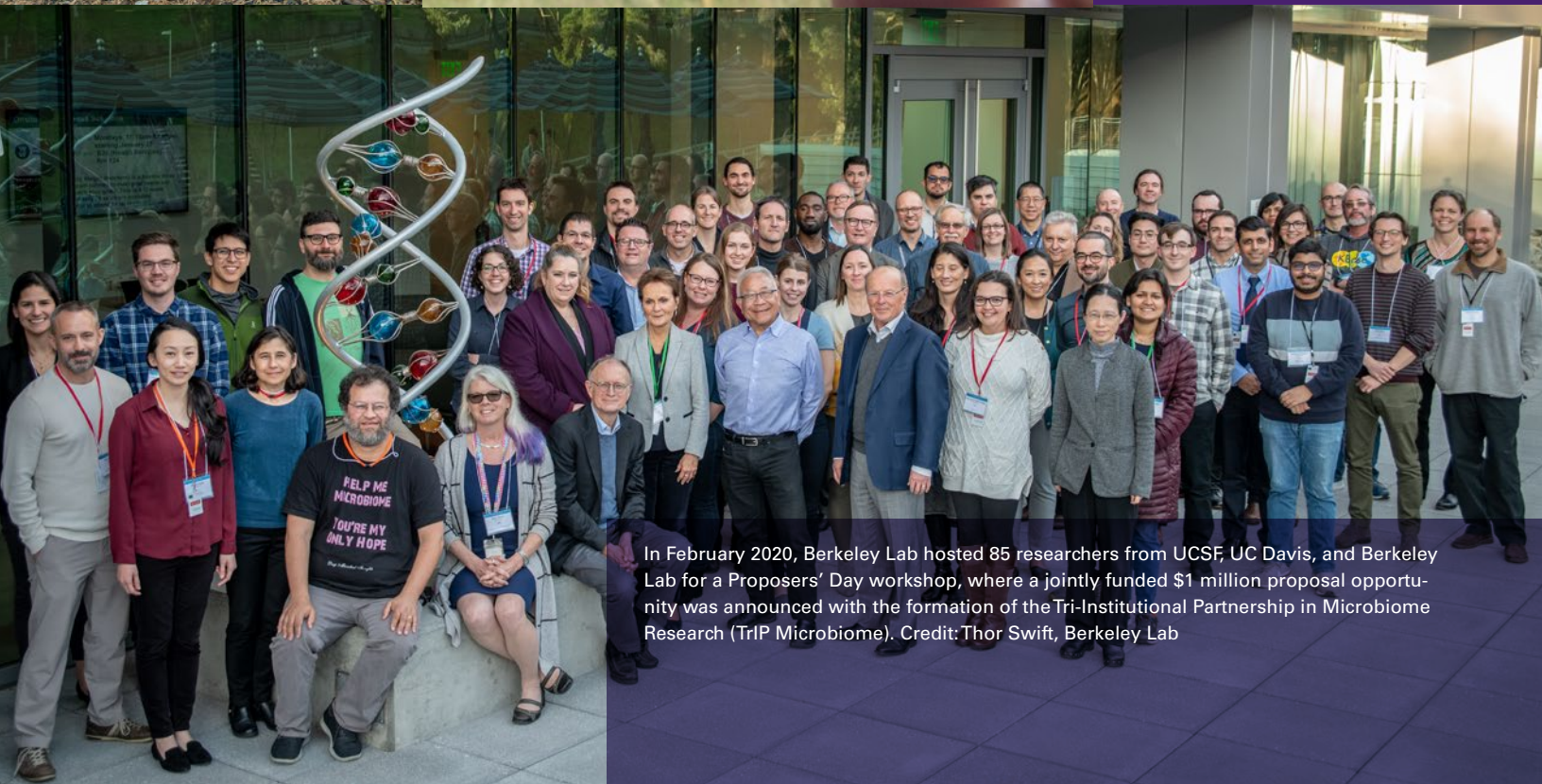


Conceptualization of a role for metabolic pathways in CO_2 flux predictions.

Led by James Stegen at PNNL, this study aimed to understand how molecular-scale processes govern the biogeochemical function of subsurface groundwater-surface water mixing zones (e.g., the hyporheic zone). This project, part of PNNL's River Corridor Hydrobiogeochemistry SFA, aims to develop models that can simulate impacts of disturbance on river corridor hydro-biogeochemistry by understanding fundamental molecular processes that lead to emergent function. This project was conducted along the Columbia River in Eastern Washington State, which exhibits variation in microbiome composition, biogeochemical activity, and substrate biogeochemistry, making it an ideal environment for studying biogeochemical hotspots. To capture a range of biogeochemical activities, samples were collected from areas with dense vegetation and virtually no vegetation. This project found that microbiome composition (metagenomics) and expression (metaproteomics) was similar between biogeochemical hotspots and low-activity sediments, diverse nitrogenous metabolites and biochemical transformations characterized hotspots, metabolite chemistry explained most of the variation in aerobic metabolism, and bulk carbon and nitrogen independently were insufficient predictors of aerobic respiration (Graham 2017b, Graham 2018). One of the most important outcomes of this study is the discovery that thermodynamic properties of organic matter chemistry are key to explaining variation in respiration rates. Learn more about this study and explore its 85 sediment samples, which generated metagenomes, proteomics, metabolomics, and organic matter characterizations, on the NMDC portal.

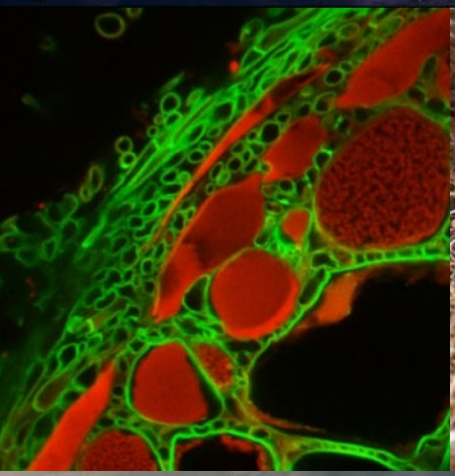
For more information, contact:
Marisa Rudolph, Editor
mrudolph@lbl.gov

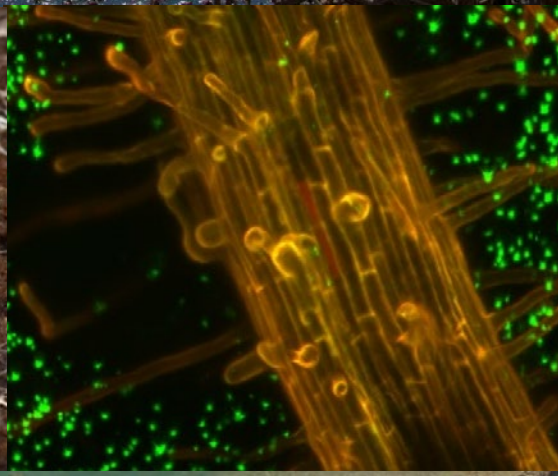
Design by Creative Services, Berkeley Lab



In February 2020, Berkeley Lab hosted 85 researchers from UCSF, UC Davis, and Berkeley Lab for a Proposers' Day workshop, where a jointly funded $1 million proposal opportunity was announced with the formation of the Tri-Institutional Partnership in Microbiome Research (TrIP Microbiome). Credit: Thor Swift, Berkeley Lab

A series of photos from the Plant Microbe Interfaces Scientific Focus Area. From left to right: Transverse section of an ectomycorrhizal root tip. Credit: Claire Veneault-Fourrey; A plant root system. Credit: Gregory Bonito; Bacterial colonization Pantoea YR343. Credit: Amber Webb





nmdc
National Microbiome Data Collaborative

U.S. DEPARTMENT OF ENERGY

microbiomedata.org @microbiomedata #NMDC @microbiomedata

22-BAO-3058