The National Microbiome Data Collaborative: Empowering the Research Community to More Effectively Harness Microbiome Data

Pajau Vangay[1]* (pvangay@lbl.gov), Lee Ann McCue[2], Chris Mungall[1], Stanton Martin[3], Shane Canon[1], Patrick Chain[4], Kjiersten Fagnan[1], Elisha Wood-Charlson[1], Nigel Mouncey[1], and **Emiley Eloe-Fadrosh**[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Pacific Northwest National Laboratory, Richland, WA; [3]Oak Ridge National Laboratory, Oak Ridge, TN; and [4]Los Alamos National Laboratory, Los Alamos, NM

https://microbiomedata.org/

Project Goals: The vision of the National Microbiome Data Collaborative (NMDC) is to empower the research community to harness microbiome data exploration and discovery through a collaborative and integrative data science ecosystem. The NMDC will address fundamental roadblocks in microbiome data science through implementation of guiding principles to make data findable, accessible, interoperable, and reusable (FAIR). To realize this vision, our multi-Lab collaborative partnership will pilot an integrated, community-centric framework to fully leverage existing microbiome data science resources and high-performance computing systems available within the DOE complex for data access, integration, and advanced analyses.

Abstract

The cross-cutting nature of microbiome research in environmental sciences, health, agriculture, energy, and natural and built environments requires the development of new solutions and community coordination to tackle grand challenges that will accelerate basic discovery and lead to transformative advances. The velocity at which microbiome data are generated has far outpaced current infrastructure resources for collection, processing, and distribution of these data in an effective, uniform, and reproducible manner, even at the largest data centers. The Interagency Strategic Plan for Microbiome Research outlined three areas of focus for strategic investments over the next five years, importantly highlighting the development of platform technologies and specifically support for open and transparent data through the development of a user-friendly, robust, integrated system with expert curation. The NMDC will tackle these infrastructure challenges in microbiome data science through developing a community-centric framework based on large-scale, collaborative partnerships leveraging unique capabilities, expertise, and resources available across four DOE National Laboratories (Lawrence Berkeley National Laboratory (LBNL), Los Alamos National Laboratory (LANL), Pacific Northwest National Laboratory (PNNL) and Oak Ridge National Laboratory (ORNL)).

During Phase I (the first 27 months), we aim to deliver a set of unique microbiome data science capabilities through leveraging existing microbiome resources hosted across the DOE complex,

as well as taking advantage of DOE’s HPC systems at LBNL. The activities that will enable a fully functional NMDC are organized into four aims, which include: leveraging existing ontology mapping software and curation resources to enable automated annotation of standardized metadata; developing microbiome workflows for metagenome, metatranscriptome, metaproteome, and metabolomics data processing leveraging HPC systems, and integrating the execution of these pipelines to produce NMDC-compliant data products; developing data registration, indexing, and access services to link data through a suite of publicly available APIs; and, developing communication and sustainability strategies to assess current and future needs and capabilities to empower users, collaborate on web-based interfaces for search functionality, and promote the NMDC to the larger scientific community.

Funding statement

The work conducted by the National Microbiome Data Collaborative is supported by the Office of Science of the U.S. Department of Energy.