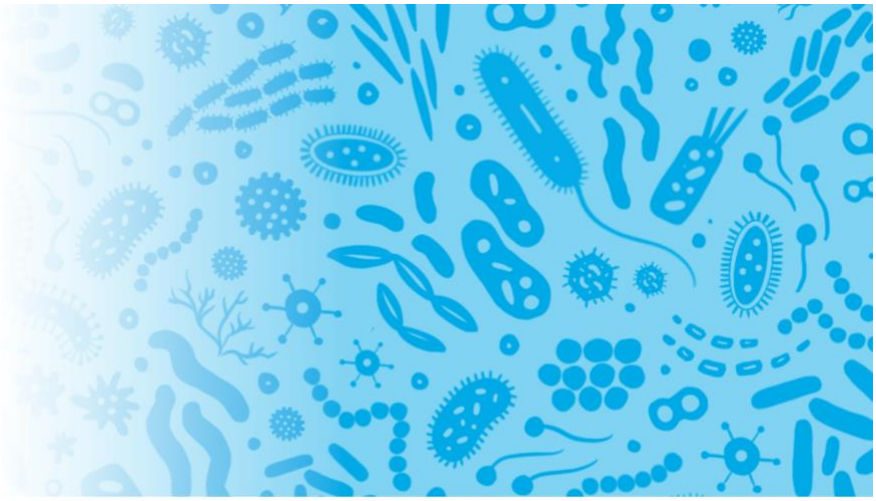

DOE BSSD Performance Management Metrics Report Q4

September 20, 2024

Investigators: Emiley A. Eloe-Fadrosh[1] (Lead, eaeloefadrosh@lbl.gov), Patrick S. G. Chain[2], Shreyas Cholia[1], Kjiersten Fagnan[1], Douglas Mans[3], Lee Ann McCue[3], Christopher J. Mungall[1], Nigel J. Mouncey[1]

Participating Institutions: [1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [2]Los Alamos National Laboratory, Los Alamos, NM 87545; [3]Pacific Northwest National Laboratory, Richland, WA 99354.

Deliverable Q4: Provide an update on the latest capabilities developed within the NMDC for providing a unique data resource to the larger microbiome research community.

Executive Summary

The National Microbiome Data Collaborative (NMDC) is building a unique and valuable multi-omics data resource to transform how researchers create, use, and reuse data. Our efforts in data standardization and innovative software solutions aim to facilitate microbiome research and discovery, addressing challenges posed by the increasing volume and complexity of microbiome data [1]. At the core of the NMDC is our data schema, developed using the Linked Data Modeling Language (LinkML), which ensures data is represented consistently and adheres to the FAIR (findable, accessible, interoperable, and reusable) data principles [2]. The NMDC schema is a unique resource in itself, providing a structured and standardized way to represent the entire lifecycle of multi-omics microbiome data analysis. Further, our use of persistent identifiers ensures long-term data accessibility and facilitates the linking of data across different studies and platforms.

Our suite of user-friendly software products serve as enabling capabilities for researchers, from making metadata capture easy to accessing publicly available multi-omics data. The NMDC Field Notes App, designed for use on mobile devices, simplifies metadata collection in the field using automated features and syncs with the web-based Submission Portal. For researchers seeking to process their data, NMDC EDGE offers easy access to standardized bioinformatics workflows to generate interoperable data outputs. The NMDC Data Portal acts as a central hub for discovering and accessing multi-omics data. The Data Portal is unique in its focus on environmental samples and connecting samples to their multi-omics data, which may otherwise be stored across different data repositories that do not always contain information on whether the data originated from the same sample. Researchers can utilize various search functionalities, including faceted search, interactive visualizations, and a map-based search to explore data from thousands of unique samples. Recognizing the importance of programmatic data access, the NMDC also offers a public Application Programming Interface (API) that allows researchers to access NMDC data directly.

We also collaborate with other facilities and data resources, including the Joint Genome Institute (JGI), the Environmental Molecular Sciences Laboratory (EMSL), the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), KBase, and the National Ecological Observatory Network (NEON). These collaborations aim to address critical gaps in data harmonization. By fostering collaborations with key organizations and engaging directly with the research community, we are establishing the NMDC as a central and invaluable microbiome data integration “hub” for microbiome research, poised to facilitate future discoveries. Herein, we describe how the NMDC’s data standards and infrastructure are designed to provide unique capabilities for researchers to create, use, and reuse microbiome data.

Background

Platform technologies in support of open and transparent data sharing in a user-friendly, robust, and integrated approach are needed for the volume of microbiome data that is now being generated. The available infrastructure resources for collection, processing, and distribution of metagenome, metatranscriptome, metabolome, metaproteome, and lipidome microbiome data in an effective, uniform, and reproducible manner have become inadequate [1]. Technical advances in data and metadata standards, interoperable database systems for open data sharing, and advanced analytical technologies that scale with the volumes of data are needed [3].

NMDC's role within the growing field of multi-omics research is to serve the scientific community in a way that enables microbiome innovation and discovery. To do this, we have built an ambitious framework for data capture, standardization, and exploration to support integrative science. The NMDC distributed data infrastructure and linked data technologies provide new ways for researchers to ask questions about how genes, proteins, metabolites, individual microbes, and communities are associated with environments and ecosystem processes. Further, by fostering strong community partnerships and developing a set of robust community outreach and training programs, we have fostered a community with diverse perspectives that is deeply involved in setting the vision and defining requirements for NMDC's software and data resources.

To date, the NMDC products have benefited a wide swath of environmental research programs which have previously lacked a coordinated effort in standardized microbiome methods, environmental metadata, and data streams, creating unnecessary barriers to data integration across studies and ecosystems. Here, we outline our accomplishments towards developing new capabilities and unique data resources for the broader research community to advance microbiome science.

Data Standards

The NMDC production platform leverages a federated data model with multi-omics microbiome data and metadata. The foundation of this platform is the NMDC schema, developed using the Linked Data Modeling Language (LinkML), to support handling and modeling of data aligned with the findable, accessible, interoperable, and reusable (FAIR) data principles [2]. This unified data model leverages vocabulary selected from external resources, including the Environment Ontology (EnvO) [4], the Genomic Standards Consortium's (GSC) Minimum Information about any (x) Sequence standard (MIxS) [5], and the Ontology of Biomedical Investigations (OBI) framework [6] for modeling sample processing and data generation. The NMDC schema provides a consistent, modular representation of the whole lifecycle of multi-omics analysis of environmental samples.

With the implementation of our data schema in LinkML, we have accelerated the transition of microbiome metadata standards from error-prone spreadsheets to a flexible, machine readable format that can incorporate metadata validation. This effort has also been closely coordinated with the development of the primary GSC MIxS standard, with our team directly involved in the release of MIxS version 6.2. Further, our most recent

enhancements of the schema to model laboratory and data analysis processes will capture the full lifecycle of omics data generation from microbiome samples. This is a major first step **to fully model contextual metadata for multi-omics microbiome data and provide the research community a path towards data integration across studies and across computational platforms**. The effort introduced higher-level organization with the introduction of "Material Processing" and "Data Generation" classes to separate laboratory processes and instrumentation for data generation. Similarly, lower-level organization has focused on granular subtypes (e.g., Mass Spectrometry, Library Preparation, Chromatographic Separation) to specify the processes behind how data is generated and what instrumentation is utilized. This new schema structure will make it easier to maintain moving forward, "future-proofing" new technologies that could be applied to microbiome samples and the instruments used to generate data. Further, the schema is also structured in a consistent and coherent manner for programmatic queries of the data. The new NMDC schema will be in place at the end of this calendar year and serves as a unique resource for the research community.

We are also leading the way towards FAIR data through the NMDC persistent identifier service that was launched in January 2023. Persistent identifiers support links across studies, samples, and workflow runs in stable, unique, and long-term ways to reference digital objects. Persistent identifiers enable interoperability, attribution, and linking across resources. Our service generates persistent identifiers for all data hosted within the NMDC and this past year, we successfully completed a large update across all legacy data to include NMDC persistent identifiers as the primary identifier. Persistent identifiers are now available for all NMDC studies, biosamples, and multi-omics workflow outputs. For future studies and biosamples, NMDC persistent identifiers will be minted across all NMDC workflows and can be referenced and linked in other systems.

The NMDC Products: unique resources for microbiome data

The NMDC's user-facing software products: the Field Notes app (a mobile application for metadata collection in the field), the Submission Portal (a user friendly spreadsheet-like web interface to enter and validate multi-omics metadata), NMDC EDGE (a web interface for external users to run NMDC standardized bioinformatics workflows), and the Data Portal (a portal for data exploration and access through an integrated, distributed data framework aligned with the FAIR data principles) are built on the foundation of the NMDC schema and provide information to NMDC's central metadata store. These products support data, information, knowledge sharing, and access, and are driven by community needs. Outlined below is further information on how we have developed these unique resources.

Field Notes and Submission Portal: making metadata capture easy

The NMDC Field Notes App and Submission Portal are flexible, template-driven tools designed to lower the barrier to collecting and reporting cohesive, standardized metadata about studies, samples, and assays (**Figures 1 & 2**). These tools were designed to make the capture and adherence to community standards easy for sample contextual information by leveraging the MIxS environmental extensions

(https://w3id.org/mixs) and the validation functions of the DataHarmonizer tool to check entered metadata values against the standards in the NMDC schema [7,8]. Together, they provide powerful and unique resources for researchers to streamline metadata capture.

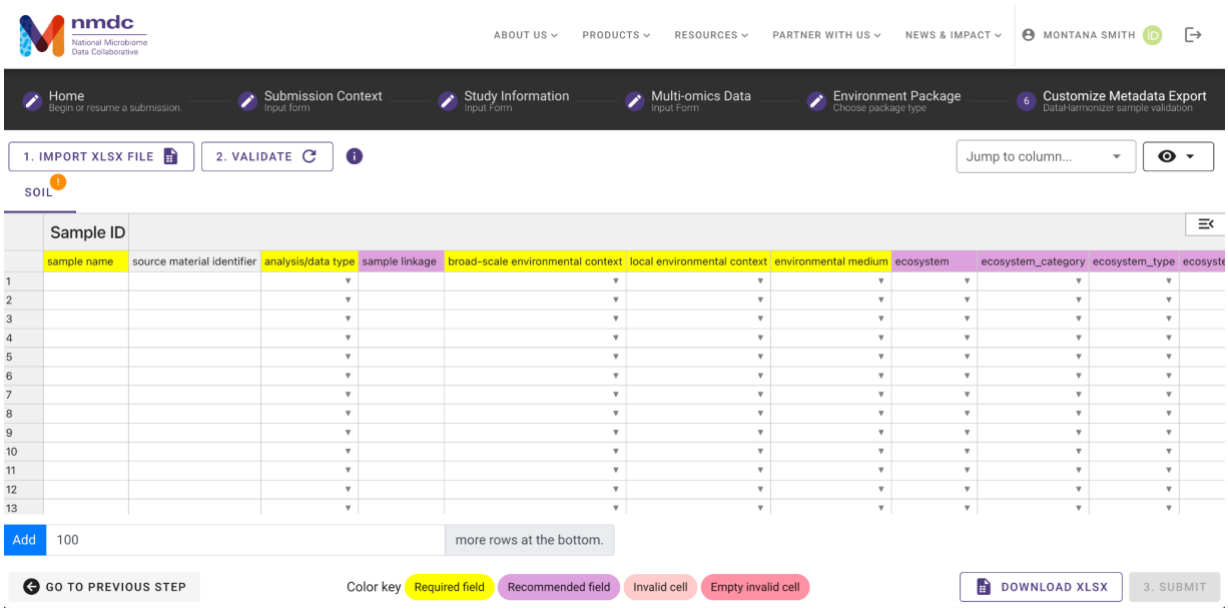

Figure 1. The NMDC Submission Portal provides a web-based interface designed to lower the barrier to capturing metadata that conforms to the community standards. Metadata are validated against these standards in real time during entry.

The mobile app streamlines collection of sample information and uses several of the mobile automated features like geolocation and date and time on either an Android device or an iOS device. The app was designed for both online and offline mode to accommodate access in the field regardless of internet connectivity, and automatically syncs with the NMDC Submission Portal through ORCiD credentials in online mode. We created a detailed document for how to access the app through Apple's TestFlight app or Android's APK files, and are currently in active beta testing mode with nearly 50 individuals. We plan to gather this feedback, fix bugs, and make updates prior to a full public release of the mobile app in 2025.

To support data submission across the DOE user facilities, the JGI and EMSL, the Submission Portal validates compliant metadata consistent with each facility sample submission requirements. These requirements include EMSL's sample management information and the JGI's DNA and RNA quality metrics. This past year, we made many updates based on user research and feedback that included providing easy access links to documentation and how-to guides, providing confirmation of a saved or submitted submission and functionality to allow multiple editors access to a submission.

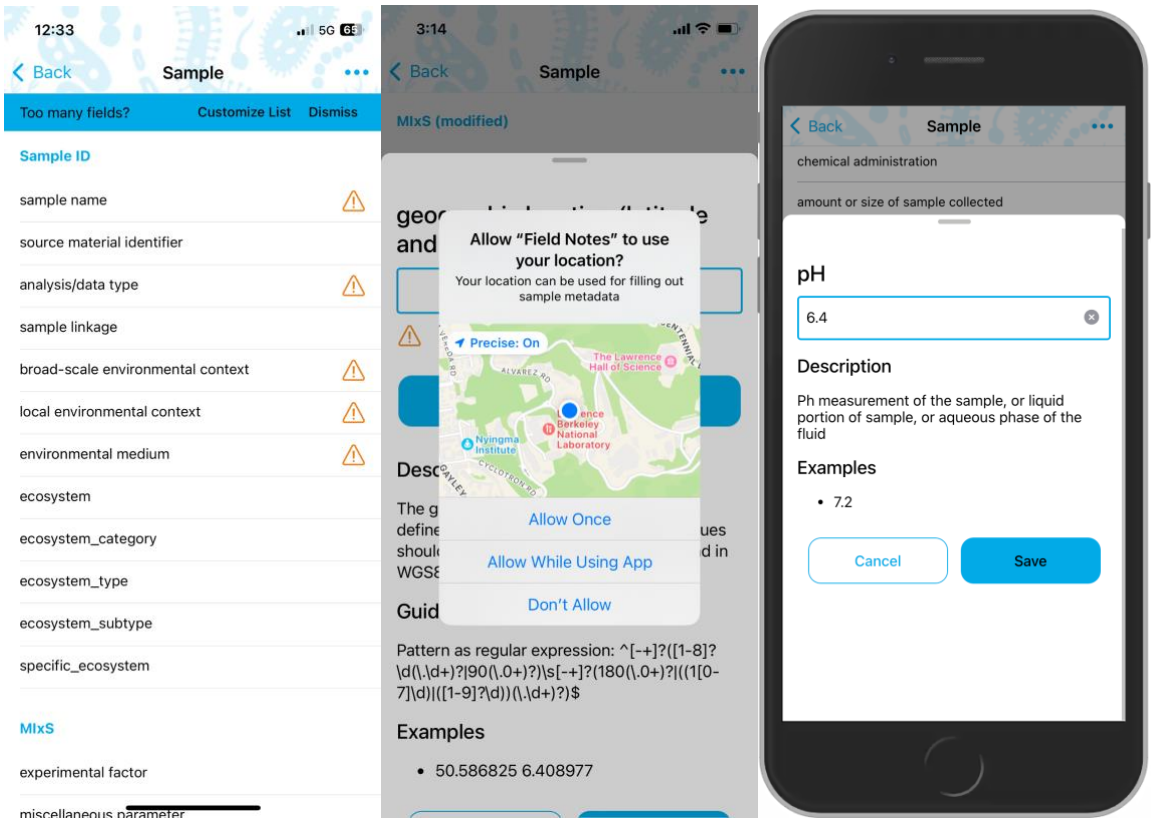

Figure 2. The NMDC Field Notes App is designed to capture metadata using a mobile device. By capturing metadata while in the field, this application will help researchers reliably collect metadata in real time and reduce errors related to delayed transcription.

NMDC EDGE: providing user-friendly bioinformatics tools

NMDC EDGE is a web application that provides broad access to the standardized production-quality workflows developed by the JGI and EMSL that have largely been limited to the facilities for which they were developed. The workflows process raw multi-omics data and produce interoperable annotated data from metagenomes, metatranscriptomes, metaproteomes, and natural organic matter characterizations. This unique resource lowers barriers by (i) making standardized bioinformatics workflows more accessible particularly for bioinformatics novices and (ii) supports powerful compute that may not be available for processing microbiome data.

To make it easy to run NMDC workflows, the NMDC EDGE home page includes quick-start guides and tutorials and has also translated most of the workflow user guides into Spanish and French. In addition to the ability to download all processed data outputs, there are also basic visualizations of data for researchers to easily view and interpret the outputs. Most of the updates and feature requests are a result of our ongoing user research process to improve NMDC EDGE. For example, the 'Retrieve SRA Data' workflow enables users to access publicly available data from NCBI's Sequence Read

Archive (SRA) and directly import into NMDC EDGE [9]. This feature streamlines the use of public data for processing with NMDC's workflows.

The NMDC Data Portal & Public API: accessing multi-omics data

The Data Portal democratizes access to studies integrating different data types, supporting persistent identifiers and links to external repositories. The Data Portal relies on a distributed framework and enables the research community to discover multi-omics data through a variety of search functionalities, such as faceted search and interactive visualizations using both the environmental sample information and functional annotations through KEGG orthology, module, and pathway terms. Work is ongoing this quarter to expand search for functional annotations to Pfam, COG and GO. Pfam and COG were chosen based on user research conducted this year and expanding to GO is part of our roadmap to specifically increase FAIR-ness for multi-omics. Users may also search through two systems for environmental ecosystem classifications, including the GOLD ecosystem classification paths [10] and the Environment Ontology (EnvO) classification terms [4].

The Data Portal hosts metadata and data from thousands of unique samples representing a large and diverse geographic distribution and environments, including river water and sediments, plant-microbe associations, and a range of diverse soils. Geographic diversity (**Figure 3**) has been a major focus in our efforts to support continental-scale biology; the map feature, geographic name or latitude/longitude coordinates can be provided to search based on location in the Data Portal. The Data Portal is a unique resource in that it contains connections of samples to their multi-omics data and in its focus on environmental samples. Sample information is captured either via translation code or Submission Portal entries and offers a single source to find multi-omics data which previously could only be discovered by searching in technology specific data repositories such as NCBI's SRA [9] for sequencing records, UCSD's MassIVE for metabolomics data or EBI's PRIDE [11] for proteomics data without necessarily containing information on which data were generated from the same original sample. **Figure 3B** is an upset plot showing counts of samples with multi-omics data.

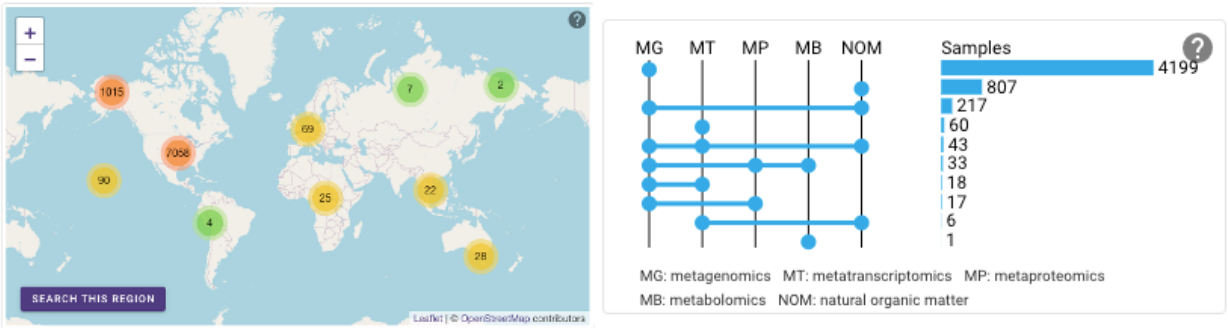

Figure 3. Visualizations of biosamples within the NMDC Data Portal. a) Geographic distribution of samples. b) The upset plot shows counts of samples which have multiple data types. Clicking on the bar chart or count will apply a search filter.

For programmatic access, we have developed a web application programming interface (API). The public NMDC API enables query and access for all study, sample and processing information, and workflow outputs. This allows NMDC metadata and data to be directly pulled into programmatic scripts or analysis tools. The NMDC API can be used broadly by the research community to access and share data. The ability to return such comprehensive and standardized records is unique. This couples nicely with the fact that our LinkML modeled schema provides automatically generated documentation, enabling users to access any metadata in our database. This year we performed user research on our API and the feedback was positive and we are working on implementing action items. We have also made improvements to the API along the way, including support for advanced custom queries, aggregations, and new endpoints based on user needs.

A spotlight on working with the community

Our multi-pronged community building approach spans individual researchers, research teams, consortia and scientific societies, and institutions and federal agencies. Through our user research program, we are able to gather feedback and work directly with the microbiome community to ensure the NMDC's resources are serving their research needs. Here, we outline how the NMDC is being used and what capabilities are helping researchers to create, use, and reuse data.

Ambassadors & Champions

NMDC Ambassadors and Champions frequently contribute to user research, the development of training materials, and advocate for data stewardship within their networks. These engaged researchers are on the front lines of working with the NMDC software and data resources, and provide unique perspectives to our team. For example, an Ambassador from the 2023 cohort hosted two NMDC workshops at the University of Puerto Rico in conjunction with the Microbial Bioprospecting and Biotechnology Laboratory. For these workshops, undergraduate researchers leveraged NMDC EDGE for standardized metagenome data processing (read QC, assembly, annotation, and binning) and published their work in a Data in Brief report [12].

Data Resource Partnerships

Our team collaborates across DOE's Biological and Environmental Research (BER) program to support access to multi-omics data. The recent workshop report "A Unified Data Infrastructure for Biological and Environmental Research: Report from the BER Advisory Committee" provides a summary of existing capabilities for data management and recommendations for future data infrastructure. To improve interoperability across existing resources and advance community standards, the NMDC serves as a data integration "hub" for standardized microbiome data to support how researchers create, use, and reuse data.

Through our close collaboration with the JGI and EMSL, users of these flagship User Facilities, individually or through the Facilities Integrating Collaborations for User Science (FICUS) program, can submit metadata compliant with established community standards and consistent with the sample submission requirements of both User Facilities. This

collaboration fills a critical gap in metadata harmonization across the JGI and EMSL, and allows researchers to readily associate multiple data types generated at the User Facilities from the same original sample, and discover data associated with specific environmental factors or data generation methods. Further, compliance with the community metadata standards implemented in the Submission Portal ensures that the data are FAIR, from collection of metadata about samples, to how they were processed, and the data analysis protocols used.

We continue to build on these successful metadata collaborations, working with the ESS-DIVE team to develop methods for linking samples and metadata across data infrastructures and with the KBase team to support data sharing of NMDC's multi-omics data within their narrative infrastructure.

While it is essential to partner across BER user facilities and resources to advance microbiome data sharing, there is a much broader sphere of microbiome research spanning federal agencies, industry, and internationally. By partnering with the National Ecological Observatory Network (NEON), the NMDC has been able to map NEON sample metadata to the community metadata standard terms in NMDC's schema. This makes a volume of metagenome data for soil, benthic, and surface water samples combined with carbon flux measurements that are valuable for researchers interested in continental-scale environmental science available through the Data Portal.

References

1. Kyrpides NC, Eloe-Fadrosh EA, Ivanova NN. Microbiome data science: Understanding our microbial planet. Trends Microbiol. 2016;24: 425–427.

2. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018.

3. Huttenhower C, Finn RD, McHardy AC. Challenges and opportunities in sharing microbiome data and analyses. Nat Microbiol. 2023;8: 1960–1970.

4. Buttigieg P, Morrison N, Smith B, Mungall CJ, Lewis SE. The environment ontology: contextualising biological and biomedical entities. J Biomed SemJournal of Biomedical Semantics. 2013;4: 43.

5. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29: 415–420.

6. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The Ontology for Biomedical Investigations. Xue Y, editor. PLoS One. 2016;11: e0154556.

7. Gill IS, Griffiths EJ, Dooley D, Cameron R, Savić Kallesøe S, John NS, et al. The DataHarmonizer: a tool for faster data harmonization, validation, aggregation and analysis of pathogen genomics contextual information. Microb Genom. 2023;9.

doi:10.1099/mgen.0.000908

8. Eloe-Fadrosh EA, Mungall CJ, Miller MA, Smith M, Patil SS, Kelliher JM, et al. A practical approach to using the Genomic Standards Consortium MIxS reporting standard for comparative genomics and metagenomics. Methods Mol Biol. 2024;2802: 587–609.

9. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res. 2022;50: D387–D390.

10. Mukherjee S, Stamatis D, Li CT, Ovchinnikova G, Bertsch J, Sundaramurthi JC, et al. Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v.9. Nucleic Acids Res. 2023;51: D957–D963.

11. Jones P, Côté RG, Martens L, Quinn AF, Taylor CF, Derache W, et al. PRIDE: a public repository of protein and peptide identifications for the proteomics community. Nucleic Acids Res. 2006;34: D659–63.

12. Rivera-Lopez EO, Nieves-Morales R, Melendez-Martinez G, Paez-Diaz JA, Rodriguez-Carrio SM, Rodriguez-Ramos J, et al. Sea cucumber (Holothuria glaberrima) intestinal microbiome dataset from Puerto Rico, generated by shotgun sequencing. Data Brief. 2024;54: 110421.