frontiers | Frontiers in Bioinformatics

TYPE Perspective
PUBLISHED 09 April 2025
DOI 10.3389/fbinf.2025.1585717

OPEN ACCESS

EDITED BY
Alberto Paccanaro,
FGV EMAp -- School of Applied Mathematics, Brazil

REVIEWED BY
Suzana De Siqueira Santos,
Federal University of ABC, Brazil

*CORRESPONDENCE
Julia M. Kelliher,
jkelliher@lanl.gov
Leah Y. D. Johnson,
leahjohnson@lanl.gov
Emiley A. Eloe-Fadrosh,
eaeloefadrosh@lbl.gov

†PRESENT ADDRESS
Mark McCauley,
U.S. Geological Survey, Wetland and Aquatic Research Centre, Gainesville, FL, United States

‡These authors have contributed equally to this work and share first authorship

RECEIVED 01 March 2025
ACCEPTED 17 March 2025
PUBLISHED 09 April 2025

CITATION
Kelliher JM, Johnson LYD, Rodriguez FE, Saunders JK, Kroeger ME, Hanson B, Robinson AJ, Anthony WE, Van Goethem MW, Kiledal A, Shibl AA, Andrade AAS, Ettinger CL, Gupta CL, Robinson CRP, Zuniga C, Sprockett D, Machado DT, Skoog EJ, Oduwole I, Rothman JA, Prime K, Lane KR, Lemos LN, Karstens L, McCauley M, Seyoum MM, Elmassry MM, Guzel M, Longley R, Roux S, Pitot TM and Eloe-Fadrosh EA (2025) A cost and community perspective on the barriers to microbiome data reuse.
Front. Bioinform. 5:1585717.
doi: 10.3389/fbinf.2025.1585717

A cost and community perspective on the barriers to microbiome data reuse

Julia M. Kelliher[1,2*‡], Leah Y. D. Johnson[1*‡], Francisca E. Rodriguez[1], Jaclyn K. Saunders[3], Marie E. Kroeger[4], Buck Hanson[1], Aaron J. Robinson[1], Winston E. Anthony[5], Marc W. Van Goethem[6], Anders Kiledal[7], Ahmed A. Shibl[8], Amanda Araujo Serrao de Andrade[9], Cassandra L. Ettinger[10], Chhedi Lal Gupta[11,12], Chris R. P. Robinson[13], Cristal Zuniga[14,15], Daniel Sprockett[16], Douglas Terra Machado[17], Emilie J. Skoog[18], Iyanu Oduwole[19], Jason A. Rothman[20], Kaelan Prime[1], Katherine R. Lane[21], Leandro Nascimento Lemos[22], Lisa Karstens[23], Mark McCauley[24†], Mitiku Mihiret Seyoum[25], Moamen M. Elmassry[26], Mustafa Guzel[27], Reid Longley[1], Simon Roux[28], Thomas M. Pitot[29] and Emiley A. Eloe-Fadrosh[28*]

[1]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, United States, [2]Department of Microbiology, Genetics, and Immunology, Michigan State University, East Lansing, MI, United States, [3]Department of Marine Sciences, University of Georgia, Athens, GA, United States, [4]In-Pipe Technology, Wood Dale, IL, United States, [5]Pacific Northwest National Laboratory, Richland, WA, United States, [6]Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, [7]Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, MI, United States, [8]Public Health Research Center, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates, [9]Department of Biological Sciences, University of Calgary, Calgary, AB, Canada, [10]Department of Microbiology and Plant Pathology, University of California, Riverside, Riverside, CA, United States, [11]ICMR-CRMCH, National Institute of Immunohaematology, Chandrapur Unit, Chandrapur, Maharashtra, India, [12]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, Uttar Pradesh, India, [13]Department of Biology, Indiana University, Bloomington, IN, United States, [14]Department of Biology, Cell, and Molecular Biology, San Diego State University, San Diego, CA, United States, [15]DOE Great Lakes Bioenergy Research Center, San Diego State University, San Diego, CA, United States, [16]Department of Microbiology and Immunology, Wake Forest University School of Medicine, Winston-Salem, NC, United States, [17]Bioinformatics Laboratory, National Laboratory for Scientific Computing, Quitandinha, Rio de Janeiro, Brazil, [18]Scripps Institution of Oceanography, UC San Diego, La Jolla, CA, United States, [19]Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, Knoxville, TN, United States, [20]Department of Microbiology and Plant Pathology, University of California: Riverside, Riverside, CA, United States, [21]Massachusetts Institute of Technology, Cambridge, MA, United States, [22]Ilum School of Science, Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, São Paulo, Brazil, [23]Division of Oncological Sciences, Department of Obstetrics and Gynecology, Knight Cancer Institute, Oregon Health and Science University, Portland, OR, United States, [24]The Whitney Laboratory for Marine Bioscience and Sea Turtle Hospital, University of Florida, St. Augustine, FL, United States, [25]Department of Poultry Science, University of Arkansas, Fayetteville, AR, United States, [26]Department of Molecular Biology, Princeton University, Princeton, NJ, United States, [27]Department of Food Engineering, Hitit University, Corum, Türkiye, [28]Lawrence Berkeley National Laboratory, Berkeley, CA, United States, [29]Department of Biochemistry, Microbiology, and Bioinformatics, Université Laval, Québec, QC, Canada

Microbiome research is becoming a mature field with a wealth of data amassed from diverse ecosystems, yet the ability to fully leverage multi-omics data for reuse remains challenging. To provide a view into researchers' behavior and attitudes towards data reuse, we surveyed over 700 microbiome researchers to

evaluate data sharing and reuse challenges. We found that many researchers are impeded by difficulties with metadata records, challenges with processing and bioinformatics, and problems with data repository submissions. We also explored the cost constraints of data reuse at each step of the data reuse process to better understand "pain points" and to provide a more quantitative perspective from sixteen active researchers. The bioinformatics and data processing step was estimated to be the most time consuming, which aligns with some of the most frequently reported challenges from the community survey. From these two approaches, we present evidence-based recommendations for how to address data sharing and reuse challenges with concrete actions for future work.

KEYWORDS

microbiome, multi-omics, data reuse, FAIR data, survey, metadata, data standards

1 Introduction

Researchers investigating microbiomes, whether from host, plant, water, or soil ecosystems, collectively generate large amounts of data increasingly from multi-omics experiments. The current paradigm in the field is to generate data for a given scientific question, yet the nature of multi-omics data often lends itself to reuse for data exploration and discovery purposes outside of the original study. More recent studies have demonstrated the tremendous value of a data reuse approach, including meta-analyses and comparative (meta)genomics, modeling efforts, and machine learning training (Duvallet et al., 2017; Kiledal et al., 2021; Madrigal et al., 2022; Saucedo et al., 2024; Li et al., 2025; Abdill et al., 2025; Elmassry et al., 2025). Microbiome data reuse has also enabled researchers to address their scientific questions at broader scales with data that they would not normally be able to generate themselves, such as continental and global-scale data (Zhang and Ning, 2015; McCauley et al., 2023; Lang et al., 2023; Osburn et al., 2024; Graham et al., 2024; Abdill et al., 2025), and difficult to acquire samples such as those from remote terrestrial and marine locations and even the International Space Station (Saunders and Rocap, 2016; Alexander et al., 2023; Pitot et al., 2024; Gonzalez et al., 2024; Nastasi et al., 2024). Reuse of published microbiome data has enabled the discovery of novel organisms and relationships, and informed our collective understanding of the biogeography of microorganisms and genetically encoded traits such as secondary metabolite production (Parks et al., 2017; Nayfach et al., 2021; Robinson et al., 2021; Edgar et al., 2022; Lima et al., 2022; Sanders et al., 2023; Machado et al., 2024; Elmassry et al., 2025).

To facilitate microbiome data reuse, several calls have been made to promote standardization, open data, and to increase data sharing (Gomez-Cabrero et al., 2014; Bhandary et al., 2018; Huttenhower et al., 2023; Sielemann et al., 2020; Eckert et al., 2020). More research teams, primary repositories, and institutions are promoting data reuse to facilitate enhanced analyses across samples, geographic locations, data types, and time scales (Jurburg et al., 2024). However, a more nuanced view into researchers' behavior and attitudes towards data reuse, along with the associated costs, have not been explored in depth.

Here, we conducted a community survey of over 700 microbiome researchers to evaluate data sharing and reuse challenges. Based on the survey results, we next explored the cost and personnel time constraints of data reuse at each step of the data reuse process to better understand the "pain points" that could be improved upon. Together, we present recommendations for how to address these challenges and concrete actions to improve how the research community can further leverage data reuse for microbiome science.

2 Community analysis of barriers to microbiome data reuse

To better understand the barriers to microbiome data reuse that researchers face, we conducted a survey in 2020 to assess various aspects of microbiome science. The survey was designed by the National Microbiome Data Collaborative (NMDC) team and reviewed and approved by the Human Subjects Committee at Lawrence Berkeley National Laboratory as an exempt IRB protocol under #394NR001. A total of 783 participants participated in the survey spanning 60 countries with approximately 50% of participants indicating they resided in the United States (415 out of 783 participants). Survey participants were asked a series of questions about their data sharing and data reuse practices, with the survey questions and anonymized results publicly available [https://doi.org/10.5281/zenodo.14948343] (Kelliher et al., 2025). Beyond the multiple choice questions, we specifically were interested in gathering feedback to understand the biggest challenges for (a) searching for microbiome data in available resources and (b) sharing microbiome data (Figure 1). For data search, we received responses from participants that outlined 637 challenges, with a plurality of responses (22%, 140/637) describing missing or incorrect metadata (Figure 1A). Other related issues with metadata were also reported, specifically a lack of standardized metadata (e.g., different ontologies and requirements across repositories) making it challenging to find data (7%, 44/637), along with issues linking primary data to the metadata (6%, 38/637). The next two categories with the most responses included challenges with processing data (16%, 105/637) and the user-friendliness of data repositories (11%, 71/637). Data processing challenges included issues with a lack of data interoperability (e.g., a lack of standardized formatting hindering data processing and different workflows leading to different outputs) and bioinformatics limitations (e.g., compute power, quality checks after data retrieval,

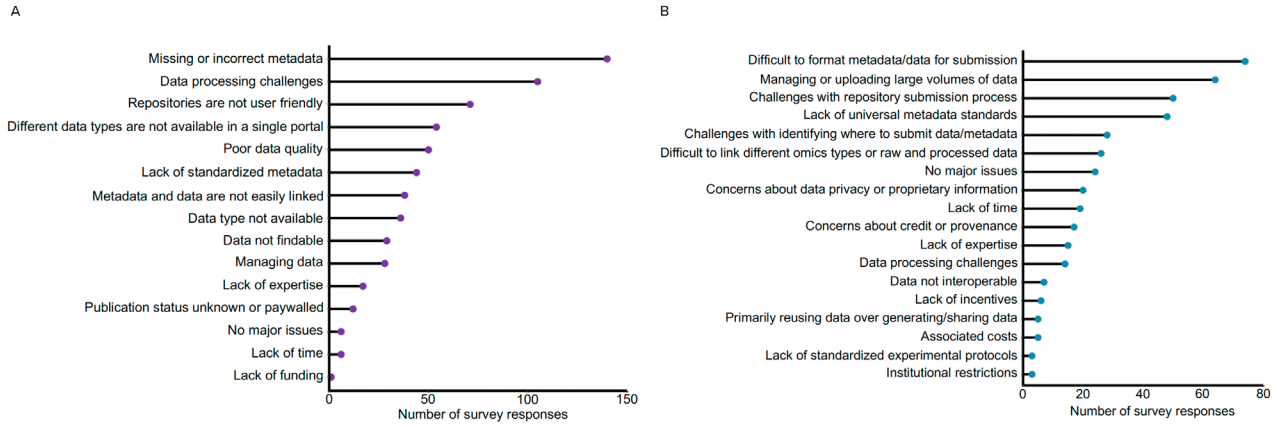

FIGURE 1
Synthesis of results from the community survey about challenges for data discovery and sharing. The answers to both free-form questions were classified into the broader categories for **(A)** data search: respondents were asked, "What is currently your biggest challenge when searching for microbiome data in available resources?" and **(B)** data sharing: respondents were asked, "What is currently your biggest challenge with respect to sharing microbiome data?". All respondent answers and bins are available here: https://doi.org/10.5281/zenodo.14948343 (Kelliher et al., 2025).

downloaded data is not in a useful format for programmatic usage).

Survey feedback regarding "user-friendliness of data repositories" focused on issues with filtering or searching for data, a lack of interoperability between user interfaces and platforms, lack of programmatic access for downloading files, and issues with repositories or databases not being maintained. Related to challenges with data, many respondents specifically noted poor quality data (8%, 50/637) and difficulties managing data (4%, 28/637). Other response categories encompassed data accessibility, including the inability to find specific data of interest (Data type not available - 6%, 36/637; e.g., research area too niche, researchers not sharing data), concerns regarding data findability (5%, 29/637; e.g., challenges identifying relevant datasets, sorting through vast amounts of datasets, concerns about missing relevant datasets), or limited publication access (e.g., paywalls or datasets and publications not linked) to identify the study context (2%, 12/637). Lack of expertise was reported as one of the least limiting factors (3%, 17/637), while the lack of time or funding together only garnered 1% (n = 7) of 637 responses.

For data sharing, we received 428 separate issues. The top responses to this question related to issues submitting data, including difficulties in formatting metadata/data for submission (17%, 74/428), managing or uploading large volumes of data (15%, 64/428), and general challenges with repository submission processes (12%, 50/428) (Figure 1B). Related to difficulties formatting metadata, many responses specified that a lack of universal metadata standards (11%, 48/428) hindered the sharing process (e.g., uncertainty about which standards or ontologies to use). Other issues related to data management included difficulty linking raw and processed data or linking different omics types (6%, 26/428) (e.g., repositories not accepting different omics data types) and challenges navigating and choosing where to submit their data from the vast amount of repositories (7%, 28/428). Participants also reported concerns about data privacy (5%, 20/428) and concerns about credit or provenance (4%, 17/428), indicating that issues surrounding data reuse ethics may contribute to reduced data sharing. Lack of time (4%, 19/428), expertise (4%, 15/428), and incentives (1%, 6/428) were additional limiting factors, with researchers reporting that the data deposition process is time-consuming and tedious, and that there are insufficient resources for navigating proper data management and repository submissions. Similar to the responses regarding data reuse, the associated costs were not reported as a major issue (1%, 5/428).

Taken together, the responses to both data search and sharing indicate that challenges with metadata, data repositories, data management, and data processing represent major issues that limit effective data reuse across the microbiome research field.

3 A case study to assess the costs of reusing microbiome data

To expand upon the community survey results and assess the costs associated with each step of the typical microbiome data reuse process, we collated information from active microbiome researchers part of the NMDC Champions program (https://microbiomedata.org/community/championsprogram/). Sixteen researchers provided estimates of personnel time and other resources associated with each step of the data reuse process from their own experiences. Figure 2 outlines the estimated personnel hours (excluding salaries or other associated costs) for each step. Personnel time was emphasized because it allowed for more direct comparisons between microbiome studies, regardless of institution or salary level, and could be used as a proxy for cost and burden estimations. Sixteen Champions assessed their "level of expertise" for large-scale data reuse (7 Intermediate and 9 Expert), and estimated the personnel time investment required at each research step, as well as the required computational resources. These estimates widely varied for each step, but overall the bioinformatics step was reported as the largest time burden (average: 160.5 hours (h); median: 100 h), followed by the downstream statistics, analyses, and figure generation step (average: 91.5 h; median: 72 h) and the

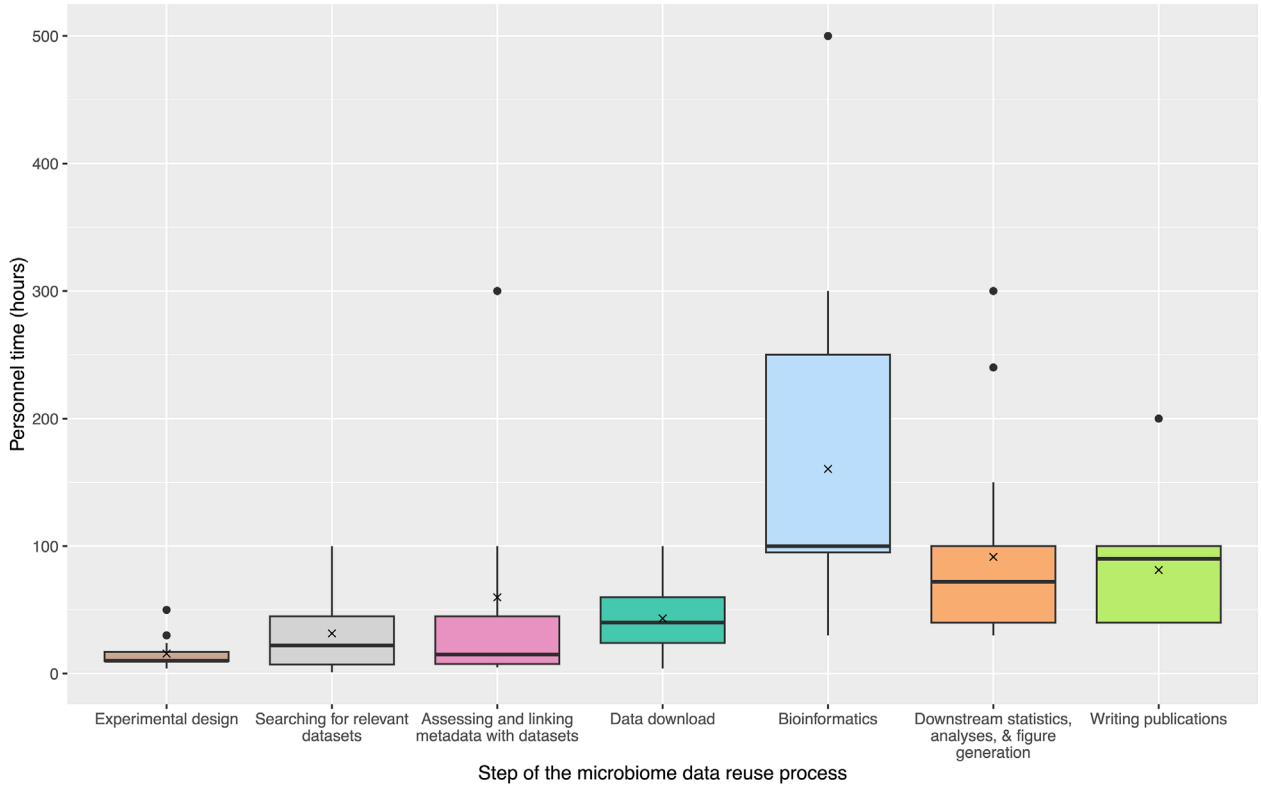

FIGURE 2
Estimates of personnel time for microbiome data reuse steps. These steps include: Experimental design, Searching for relevant datasets, Assessing and linking metadata with datasets, Data download, Bioinformatics, Downstream statistics, analyses, and figure generation, and Writing publications. Estimates were gathered from 16 NMDC Champions. Several Champions estimated "greater than" the hours that were used in the final average and median calculations (e.g., >100 h was treated as 100 h). Figure was generated using the ggplot2 R package and the *microshades* color palette (Wickham, 2011; Dahl et al., 2022).

publication writing step (average: 81.25 h; median: 90 h). In the community survey, many researchers reported difficulties in managing and uploading large amounts of data. To further quantify the amounts of data involved in typical reuse studies, Champions estimated the amount of computational resources required for data storage as well as the computational resources required for data analysis and bioinformatics. A range from 1 TB to 10 TB was estimated for data storage, and up to 10,000 core hours were reported for data analysis and bioinformatics, although this metric was not able to be estimated by all Champions. Together, this case study to estimate time constraints and costs illustrates practical data reuse steps in a more quantitative way. Based on these data, we offer evidence-based recommendations to improve the process with an eye towards streamlining future data reuse.

4 Discussion

This view into researchers' attitudes towards data reuse and cost estimates is instructive and allows for an enhanced understanding of how microbiome researchers can leverage existing data investments. By clarifying how challenges are perceived and the associated costs of data reuse, we are able to establish evidence-based recommendations for future work. Using a community survey approach, we found that there are several issues that disincentivize researchers from reusing data. One major theme of reported challenges in both sharing and reusing data involved metadata quality, standardization, and availability. This echoes other reports surrounding lagging adoption of metadata standards and best practices (Vangay et al., 2021; Cernava et al., 2022; Fraga-Gonzalez et al., 2025), further emphasizing that this is a barrier that needs to be universally addressed. Challenges with processing large volumes of data was reported in the community survey, and this burden was also reflected by the NMDC Champion estimates that the bioinformatics steps are the most time consuming and require large amounts of computational resources. Difficulties with repositories for data sharing as well as for finding and accessing reusable microbiome data were also reported.

The case study estimation analysis, while limited to sixteen individuals, provides more quantitative information on data reuse that, to our knowledge, has not been reported in other studies discussing barriers to data reuse. This analysis is meant as a preliminary assessment of current practices to elaborate upon discussion points that have been reported in other perspectives (Tenopir et al., 2011; Tenopir et al., 2015; Huttenhower et al., 2023). We recognize that other costs such as those associated with computational resources widely vary across institutions and that it can be difficult or impossible to obtain quotes or financial

information to quantify and standardize cost assessments across the entire microbiome research community. We assessed personnel time requirements as a more comparable metric across microbiome research questions, researchers, and institutions, however this also has its limitations when used as a measure of burden.

4.1 Recommendations

Moving forward, it is increasingly clear that data reuse challenges must be addressed from the perspectives of both depositors and reusers to make it more broadly feasible. Below, we provide specific recommendations based on our synthesis of both the community survey and Champions feedback on data sharing, reuse, and costs.

4.2 Metadata

From our survey, researchers reported issues with metadata collection, standardization, reporting, deposition, and access as significant barriers to both sharing and reuse. The Genomic Standards Consortium (GSC), the Environment Ontology (EnvO), and the Open Biomedical Ontologies (OBO) Foundry are all examples of community-driven efforts that have significantly advanced how microbiome metadata can be collected and standardized (Field et al., 2011; Buttigieg et al., 2013; Smith et al., 2007). More recently, two large community-driven efforts have emerged to assist researchers with consistently reporting and publishing on microbiome data: the STORMS and STREAMS guidelines for human microbiome and environmental microbiome data, respectively (Mirzayi et al., 2021; Kelliher et al., 2024a). Despite these efforts, there is generally a lack of awareness of existing metadata standards and an even more pronounced lack of adoption (Vangay et al., 2021; Cernava et al., 2022). We suspect that this lack of awareness and utilization of metadata standards significantly contributes to the challenges the community faces with sharing and reusing data. Additional training, tutorials, and awareness of metadata and data standards would provide significant benefits at the individual and community levels to increase adoption and implementation of these efforts.

4.3 Finding and accessing data through data repositories

Searching for, finding, and accessing relevant datasets was emphasized as a pain point for researchers in both the community survey and as a time burden in the Champions' estimates. Enhanced interoperability between datasets and repositories can assist researchers in these steps, and it is often the responsibility of the data submitter to ensure that there are decipherable connections between the data and metadata. Survey participants reported issues navigating repositories for both sharing and searching for data, and a lack of training or tutorials hindering this process. Educational resources with a focus on data repositories could enhance researchers' ability to effectively share their data and adhere to FAIR [Findable, Accessible, Interoperable, and Reusable] principles, thus making data more accessible overall (Wilkinson et al., 2016). Several data repositories offer links to other related repositories or datasets which can help in the search process (Gebre et al., 2025). It can also be important to note whether repositories have been curated and in what manner to minimize the reported issues with insufficient data and metadata quality (Eloe-Fadrosh et al., 2022; Muller et al., 2022). Additionally, when publishing on data reuse studies, it is important to note that many repositories accept processed data, which can improve reproducibility of the published work (Baker et al., 2000; McWilliam et al., 2013). Researching and adhering to the data use policies and citing data from repositories properly can save time during revisions and can foster trust and incentives for those that report hesitancy with data sharing. Lastly, we anticipate newer tools like incorporating machine learning or artificial intelligence within repositories will also help to address challenges in data quality control for both raw and processed data (reviewed in Hernández Medina et al., 2022; Kumar et al., 2024).

4.4 Data processing and bioinformatics

Bioinformatics was reported as the most time-consuming step for the NMDC Champions. Data processing steps would be significantly less time consuming for researchers if free, publicly available software and tools were more readily available. Several web-based cyberinfrastructures exist that increase the accessibility of bioinformatics workflows (Swetnam et al., 2024; Lo et al., 2022; Arkin et al., 2018; Li et al., 2017; Kelliher et al., 2024b). For data download, tools such as the Sequence Read Archive (SRA) toolkit and Globus can help researchers to improve their download procedures (Foster and Kesselman, 1997; Chard et al., 2016; Heldenbrand et al., 2017; Sayers et al., 2022). More transfer services at the institutional and individual levels (such as the services provided by IMG/M) as well as more publicly available application programming interfaces (APIs) would also minimize data download burdens for researchers (Chen et al., 2023). Additional training, webinars, workshops, tutorials, and documentation would all lower the barriers to these steps, especially for researchers that are not as familiar with these processes. Publishing and sharing code used for data reuse can also facilitate collective improvements across the field.

5 Conclusion

Taken together, many of the responsibilities for promoting microbiome data reuse have been discussed from the perspective of the data generators (Huttenhower et al., 2023). Other new tools, resources, and recommendations can also improve the data reuse process and decrease researcher burden. Increased collaboration and discussions between data generators and reusers can also lead to the sharing and adoption of data and metadata best practices. We encourage the continuation of calls to action for increased reuse of microbiome data, ideally from the perspective of all organizational partners including research teams, data repository representatives, funding agencies, and publishers. Grant funding calls for research projects specifically reusing data and providing

compute time may incentive more meta-analyses. While community awareness and adoption of FAIR and open data management practices is increasing, addressing the reported challenges will facilitate further implementation across the field. The two lines of investigation reported herein provide insight into the behaviors and practices of microbiome researchers, and the barriers they encounter with microbiome data reuse. This perspective contributes to our collective understanding of researcher attitudes towards data reuse, and provides recommendations for how the community can work together to address the most pressing challenges in microbiome data reuse.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: The data analyzed for this study can be found in Zenodo [https://doi.org/10.5281/zenodo.14948343].

Ethics statement

The studies involving humans were approved by the Human Subjects Committee at Lawrence Berkeley National Laboratory as an exempt IRB protocol under #394NR001. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

JK: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Supervision, Visualization, Writing–original draft, Writing–review and editing. LJ: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Visualization, Writing–original draft, Writing–review and editing. FR: Conceptualization, Data curation, Investigation, Methodology, Project administration, Supervision, Writing–original draft, Writing–review and editing. JS: Conceptualization, Data curation, Investigation, Methodology, Writing–original draft, Writing–review and editing. MK: Conceptualization, Data curation, Investigation, Methodology, Writing–original draft, Writing–review and editing. BH Conceptualization, Data curation, Investigation, Methodology, Writing–original draft, Writing–review and editing. AR: Conceptualization, Data curation, Investigation, Methodology, Visualization, Writing–original draft, Writing–review and editing. WA: Conceptualization, Data curation, Writing–original draft, Writing–review and editing. MV: Data curation, Writing–original draft, Writing–review and editing. AK: Conceptualization, Data curation, Writing–original draft, Writing–review and editing. AS: Writing–original draft, Writing–review and editing. AA: Writing–original draft, Writing–review and editing. CE: Writing–original draft, Writing–review and editing. CG: Writing–original draft, Writing–review and editing. CR: Writing–original draft, Writing–review and editing. CZ: Writing–original draft, Writing–review and editing. DS: Writing–original draft, Writing–review and editing. DM: Writing–original draft, Writing–review and editing. ES: Writing–original draft, Writing–review and editing. IO: Writing–original draft, Writing–review and editing. JR: Writing–original draft, Writing–review and editing. KP: Conceptualization, Investigation, Writing–original draft, Writing–review and editing. KL: Writing–original draft, Writing–review and editing. LL: Writing–original draft, Writing–review and editing. LK: Writing–original draft, Writing–review and editing. MM: Writing–original draft, Writing–review and editing. MS: Writing–original draft, Writing–review and editing. ME: Writing–original draft, Writing–review and editing. MG: Writing–original draft, Writing–review and editing. RL: Data curation, Writing–original draft, Writing–review and editing. SR: Conceptualization, Investigation, Writing–original draft, Writing–review and editing. TP: Writing–original draft, Writing–review and editing. EE-F: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The work conducted by the National Microbiome Data Collaborative (https://ror.org/05cwx3318) is supported by the Genomic Science Program in the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER) under contract numbers DE-AC02-05CH11231 (LBNL), 89233218CNA000001 (LANL), and DE-AC05-76RL01830 (PNNL).

Conflict of interest

Author MK is employed by In-Pipe Technology.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Abdill, R. J., Graham, S. P., Rubinetti, V., Ahmadian, M., Hicks, P., Chetty, A., et al. (2025). Integration of 168,000 samples reveals global patterns of the human gut microbiome. *Cell* 188, 1100–1118.e17. doi:10.1016/j.cell.2024.12.017

Alexander, H., Hu, S. K., Krinos, A. I., Pachiadaki, M., Tully, B. J., Neely, C. J., et al. (2023). Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton. *mBio* 14, e0167623. doi:10.1128/mbio.01676-23

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., et al. (2018). KBase: the United States department of Energy systems biology knowledgebase. *Nat. Biotechnol.* 36, 566–569. doi:10.1038/nbt.4163

Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G., et al. (2000). The EMBL nucleotide sequence database. *Nucleic Acids Res.* 28, 19–23. doi:10.1093/nar/28.1.19

Bhandary, P., Seetharam, A. S., Arendsee, Z. W., Hur, M., and Wurtele, E. S. (2018). Raising orphans from a metadata morass: a researcher's guide to re-use of public 'omics data. *Plant Sci.* 267, 32–47. doi:10.1016/j.plantsci.2017.10.014

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., and the ENVO Consortium (2013). The environment ontology: contextualising biological and biomedical entities. *J. Biomed. Semant.* 4, 43. doi:10.1186/2041-1480-4-43

Cernava, T., Rybakova, D., Buscot, F., Clavel, T., McHardy, A. C., Meyer, F., et al. (2022). Metadata harmonization–Standards are the key for a better usage of omics data for integrative microbiome analysis. *Environ. Microbiome* 17, 33. doi:10.1186/s40793-022-00425-1

Chard, K., Tuecke, S., and Foster, I. (2016). "Globus: recent enhancements and future plans," in *Proceedings of the XSEDE16 conference on diversity, big data, and science at scale* (New York, NY, USA: Association for Computing Machinery), 1–8. doi:10.1145/2949550.2949554

Chen, I.-M. A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., et al. (2023). The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.* 51, D723–D732. doi:10.1093/nar/gkac976

Dahl, E. M., Neer, E., Bowie, K. R., Leung, E. T., and Karstens, L. (2022). Microshades: an R package for improving color accessibility and organization of microbiome data. *Microbiol. Resour. Announc.* 11, e0079522–22. doi:10.1128/mra.00795-22

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* 8, 1784. doi:10.1038/s41467-017-01973-8

Eckert, E. M., Cesare, A. D., Fontaneto, D., Berendonk, T. U., Bürgmann, H., Cytryn, E., et al. (2020). Every fifth published metagenome is not available to science. *PLOS Biol.* 18, e3000698. doi:10.1371/journal.pbio.3000698

Edgar, R. C., Taylor, B., Lin, V., Altman, T., Barbera, P., Meleshko, D., et al. (2022). Petabase-scale sequence alignment catalyses viral discovery. *Nature* 602, 142–147. doi:10.1038/s41586-021-04332-2

Elmassry, M. M., Sugihara, K., Chankhamjon, P., Kim, Y., Camacho, F. R., Wang, S., et al. (2025). A meta-analysis of the gut microbiome in inflammatory bowel disease patients identifies disease-associated small molecules. *Cell Host Microbe* 33, 218–234.e12. doi:10.1016/j.chom.2025.01.002

Eloe-Fadrosh, E. A., Ahmed, F., Babinski, M., Maumes, J., Borkum, M., Bramer, L., et al. (2022). The national microbiome data collaborative data portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res.* 50, D828–D836. doi:10.1093/nar/gkab990

Field, D., Amaral-Zettler, L., Cochrane, G., Cole, J. R., Dawyndt, P., Garrity, G. M., et al. (2011). The genomic standards Consortium. *PLoS Biol.* 9, e1001088. doi:10.1371/journal.pbio.1001088

Foster, I., and Kesselman, C. (1997). Globus: a metacomputing infrastructure toolkit. *Int. J. Supercomput. Appl. High Perform. Comput.* 11, 115–128. doi:10.1177/109434209701100205

Fraga-González, G., van de Wiel, H., Garassino, F., Kuo, W., de Zélicourt, D., Kurtcuoglu, V., et al. (2025). Affording reusable data: recommendations for researchers from a data-intensive project. *Sci. Data* 12, 258. doi:10.1038/s41597-025-04565-0

Gebre, S. G., Scott, R. T., Saravia-Butler, A. M., Lopez, D. K., Sanders, L. M., and Costes, S. V. (2025). NASA open science data repository: open science for life in space. *Nucleic Acids Res.* 53, D1697–D1710. doi:10.1093/nar/gkae1116

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8, I1. doi:10.1186/1752-0509-8-S2-I1

Gonzalez, E., Lee, M. D., Tierney, B. T., Lipieta, N., Flores, P., Mishra, M., et al. (2024). Spaceflight alters host-gut microbiota interactions. *npj Biofilms Microbiomes* 10, 71–19. doi:10.1038/s41522-024-00545-1

Graham, E. B., Camargo, A. P., Wu, R., Neches, R. Y., Nolan, M., Paez-Espino, D., et al. (2024). A global atlas of soil viruses reveals unexplored biodiversity and potential biogeochemical impacts. *Nat. Microbiol.* 9, 1873–1883. doi:10.1038/s41564-024-01686-x

Heldenbrand, J., Ren, Y., Asmann, Y., and Mainzer, L. S. (2017). Step-by-Step guide for downloading very large datasets to a supercomputer using the SRA Toolkit. Available online at: https://protocols.io/view/step-by-step-guide-for-downloading-very-large-data-kb6csre.

Hernández Medina, R., Kutuzova, S., Nielsen, K. N., Johansen, J., Hansen, L. H., Nielsen, M., et al. (2022). Machine learning and deep learning applications in microbiome research. *ISME Commun.* 2 (1), 98. doi:10.1038/s43705-022-00182-9

Huttenhower, C., Finn, R. D., and McHardy, A. C. (2023). Challenges and opportunities in sharing microbiome data and analyses. *Nat. Microbiol.* 8, 1960–1970. doi:10.1038/s41564-023-01484-x

Jurburg, S. D., Álvarez Blanco, M. J., Chatzinotas, A., Kazem, A., König-Ries, B., Babin, D., et al. (2024). Datathons: fostering equitability in data reuse in ecology. *Trends Microbiol.* 32, 415–418. doi:10.1016/j.tim.2024.02.010

Kelliher, J., Aljumaah, M., Bordenstein, S., Brister, J. R., Chain, P., Dundore-Arias, J. P., et al. (2024a). *Microbiome data management in action workshop: Atlanta, ga, USA, june 12-13, 2024*. doi:10.5281/zenodo.13829669

Kelliher, J., Johnson, L., and Eloe-Fadrosh, E. (2025). *NMDC community survey questions and binned responses*. doi:10.5281/zenodo.14948343

Kelliher, J. M., Xu, Y., Flynn, M. C., Babinski, M., Canon, S., Cavanna, E., et al. (2024b). Standardized and accessible multi-omics bioinformatics workflows through the NMDC EDGE resource. *Comput. Struct. Biotechnol. J.* 23, 3575–3583. doi:10.1016/j.csbj.2024.09.018

Kiledal, E. A., Keffer, J. L., and Maresca, J. A. (2021). Bacterial communities in concrete reflect its composite nature and change with weathering. *mSystems* 6, e01153. doi:10.1128/msystems.01153-20

Kumar, B., Lorusso, E., Fosso, B., and Pesole, G. (2024). A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions. *Front. Microbiol.* 15, 1343572. doi:10.3389/fmicb.2024.1343572

Lang, A. K., Pett-Ridge, J., McFarlane, K. J., and Phillips, R. P. (2023). Climate, soil mineralogy and mycorrhizal fungi influence soil organic matter fractions in eastern US temperate forests. *J. Ecol.* 111, 1254–1269. doi:10.1111/1365-2745.14094

Li, P.-E., Lo, C.-C., Anderson, J. J., Davenport, K. W., Bishop-Lilly, K. A., Xu, Y., et al. (2017). Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Res.* 45, 67–80. doi:10.1093/nar/gkw1027

Li, Z., Chen, Y., Sun, Y., McArthur, K., Carnes, M., Liu, T., et al. (2025). *In harmony? A scoping review of methods to combine multiple 16S amplicon data sets*. doi:10.1101/2025.02.11.637740

Lima, S. T., Fallon, T. R., Cordoza, J. L., Chekan, J. R., Delbaje, E., Hopiavuori, A. R., et al. (2022). Biosynthesis of guanitoxin enables global environmental detection in freshwater cyanobacteria. *J. Am. Chem. Soc.* 144, 9372–9379. doi:10.1021/jacs.2c01424

Lo, C.-C., Shakya, M., Connor, R., Davenport, K., Flynn, M., Gutiérrez, A. M. y., et al. (2022). EDGE COVID-19: a web platform to generate submission-ready genomes from SARS-CoV-2 sequencing efforts. *Bioinformatics* 38, 2700–2704. doi:10.1093/bioinformatics/btac176

Machado, D. T., Dias, B. do C., Cayô, R., Gales, A. C., Carvalho, F. M. de, and Vasconcelos, A. T. R. (2024). Uncovering new Firmicutes species in vertebrate hosts through metagenome-assembled genomes with potential for sporulation. *Microbiol. Spectr.* 12, e0211324. doi:10.1128/spectrum.02113-24

Madrigal, P., Singh, N. K., Wood, J. M., Gaudioso, E., Hernández-Del-Olmo, F., Mason, C. E., et al. (2022). Machine learning algorithm to characterize antimicrobial resistance associated with the International Space Station surface microbiome. *Microbiome* 10, 134. doi:10.1186/s40168-022-01332-w

McCauley, M., Goulet, T. L., Jackson, C. R., and Loesgen, S. (2023). Systematic review of cnidarian microbiomes reveals insights into the structure, specificity, and fidelity of marine associations. *Nat. Commun.* 14, 4899. doi:10.1038/s41467-023-39876-6

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., et al. (2013). Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.* 41, W597–W600. doi:10.1093/nar/gkt376

Mirzayi, C., Renson, A., Zohra, F., Elsafoury, S., Geistlinger, L., Kasselman, L. J., et al. (2021). Reporting guidelines for human microbiome research: the STORMS checklist. *Nat. Med.* 27, 1885–1892. doi:10.1038/s41591-021-01552-x

Muller, E., Algavi, Y. M., and Borenstein, E. (2022). The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *npj Biofilms Microbiomes* 8, 79–87. doi:10.1038/s41522-022-00345-5

Nastasi, N., Haines, S. R., Bope, A., Meyer, M. E., Horack, J. M., and Dannemiller, K. C. (2024). Fungal diversity differences in the indoor dust microbiome from built environments on earth and in space. *Sci. Rep.* 14, 11858. doi:10.1038/s41598-024-62191-z

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., et al. (2021). A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* 39, 499–509. doi:10.1038/s41587-020-0718-6

Osburn, E. D., McBride, S. G., Bahram, M., and Strickland, M. S. (2024). Global patterns in the growth potential of soil bacterial communities. *Nat. Commun.* 15, 6881. doi:10.1038/s41467-024-50382-1

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2, 1533–1542. doi:10.1038/s41564-017-0012-7

Pitot, T. M., Rapp, J. Z., Schulz, F., Girard, C., Roux, S., and Culley, A. I. (2024). Distinct and rich assemblages of giant viruses in Arctic and Antarctic lakes. *ISME Commun.* 4, ycae048. doi:10.1093/ismeco/ycae048

Robinson, A. J., House, G. L., Morales, D. P., Kelliher, J. M., Gallegos-Graves, L. V., LeBrun, E. S., et al. (2021). Widespread bacterial diversity within the bacteriome of fungi. *Commun. Biol.* 4, 1168–1213. doi:10.1038/s42003-021-02693-y

Sanders, J. G., Sprockett, D. D., Li, Y., Mjungu, D., Lonsdorf, E. V., Ndjango, J.-B. N., et al. (2023). Widespread extinctions of co-diversified primate gut bacterial symbionts from humans. *Nat. Microbiol.* 8, 1039–1050. doi:10.1038/s41564-023-01388-w

Saucedo, B., Saldivar, A., Martinez, D., Canto-Encalada, G., Norena-Caro, D., Peeler, I., et al. (2024). *Mathematical modeling is unraveling the metabolism of photosynthetic organisms to drive novel culturing*. London, United Kingdom: IntechOpen. doi:10.5772/intechopen.1007463

Saunders, J. K., and Rocap, G. (2016). Genomic potential for arsenic efflux and methylation varies among global Prochlorococcus populations. *ISME J.* 10, 197–209. doi:10.1038/ismej.2015.85

Sayers, E. W., O'Sullivan, C., and Karsch-Mizrachi, I. (2022). "Using GenBank and SRA," in *Plant bioinformatics: methods and protocols*. Editor D. Edwards (New York, NY: Springer US), 1–25. doi:10.1007/978-1-0716-2067-0_1

Sielemann, K., Hafner, A., and Pucker, B. (2020). The reuse of public datasets in the life sciences: potential risks and rewards. *PeerJ* 8, e9954. doi:10.7717/peerj.9954

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255. doi:10.1038/nbt1346

Swetnam, T. L., Antin, P. B., Bartelme, R., Bucksch, A., Camhy, D., Chism, G., et al. (2024). CyVerse: cyberinfrastructure for open science. *PLOS Comput. Biol.* 20, e1011270. doi:10.1371/journal.pcbi.1011270

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., et al. (2011). Data sharing by scientists: practices and perceptions. *PLOS ONE* 6, e21101. doi:10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., et al. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS ONE* 10, e0134826. doi:10.1371/journal.pone.0134826

Vangay, P., Burgin, J., Johnston, A., Beck, K. L., Berrios, D. C., Blumberg, K., et al. (2021). Microbiome metadata standards: report of the national microbiome data collaborative's workshop and follow-on activities. *mSystems* 6, e01194. doi:10.1128/msystems.01194-20

Wickham, H. (2011). ggplot2. *WIREs Comput. Stat.* 3, 180–185. doi:10.1002/wics.147

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi:10.1038/sdata.2016.18

Zhang, H., and Ning, K. (2015). The tara oceans project: new opportunities and greater challenges ahead. *Genomics, Proteomics and Bioinforma.* 13, 275–277. doi:10.1016/j.gpb.2015.08.003