Title

Tier-based standards for FAIR sequence data and metadata sharing in microbiome research

Authors

Lina Kim[1]*, Anton Lavrinienko[1]*, Zuzana Sebechlebska[1], Sven Stoltenberg[1], Nicholas A. Bokulich[1#]

** These authors contributed equally to this publication*

Affiliations

[1] Department of Health Sciences and Technology, ETH Zurich, Switzerland

[#] Corresponding author: Nicholas A. Bokulich, nicholas.bokulich@hest.ethz.ch

Abstract

Microbiome research is a growing, data-driven field within the life sciences. While policies exist for sharing microbiome sequence data and using standardized metadata schemes, compliance among researchers varies. To promote open research data best practices in the microbiome research community, we (1) propose two tiered badge systems to evaluate data/metadata sharing compliance, and (2) showcase an automated evaluation tool to determine adherence to data reporting standards in publications with amplicon and metagenome sequence data. In a systematic evaluation of publications (n ~ 3000) spanning human gut microbiome research, we found that nearly half of publications do not meet minimum standards for sequence data availability. Moreover, poor standardization of metadata creates a high barrier to harmonization and cross-study comparison. Using this badge system and evaluation tool, our proof-of-concept work exposes (i) the ineffectiveness of sequence data availability statements, and (ii) the lack of consistent metadata standards for annotating microbial data. In this Perspective, we highlight the need for improved practices and infrastructure that reduce barriers to data submission and maximize reproducibility in microbiome research. We anticipate our tiered badge framework will promote dialogue regarding data sharing practices and facilitate microbiome data reuse, supporting best practices that make microbiome data FAIR.

Keywords: microbiome, data availability statement, data sharing, meta-research, metadata

Main

Reproducibility and data reuse are fundamental to advancing scientific discovery and innovation, yet both remain persistent challenges across the life sciences [1]. Open research data (ORD) are publicly accessible, usable, and redistributable. Ideally, ORD should also adhere to Findable, Accessible, Interoperable, and Reusable (FAIR) [2] principles and best practices that facilitate data reuse in research. Such openness is crucial in the life sciences that are increasingly reliant on large-scale, high-dimensional datasets. Nearly half of all researchers use data generated by other scientists [3], and this number is likely to grow due to the increasing adoption of computational methods in all areas of research [4,5]. However, data reuse across research disciplines is hindered by subpar sharing practices including poor accessibility and interoperability, and particularly insufficient data and metadata reporting [6].

A growing and data-driven field within the life sciences involves the study of microbiomes: microbial communities, their components, products, interactions, and functional activities [7,8]. Rapidly advancing technology provides the means to catalog microbiomes using culture-independent methods, including marker-gene amplicon sequencing [9–11] and genome-resolved metagenomics [12]. Microbiome research has already demonstrated practical potential in diverse areas of biomedicine [13], and direct applications in environmental sustainability [14], and agriculture and food research [15]. Addressing the complexity of microbial communities in diverse heterogeneous ecosystems requires enormous quantities of sequence data, the interpretation of which relies on the associated sample and study metadata. Access to the raw sequence data and metadata is key to scientific transparency and innovative reuse, yet data accessibility [16] and metadata standardization issues [17] are salient in microbiome research.

Applying FAIR principles to both the raw sequence data and metadata is necessary for reproducibility and efficient data reuse. Microbiome research is an actively developing field and several data sharing standards already exist for depositing the raw microbiome sequence data (e.g. INSDC repositories), including microbiome-specific databases with integrated analytical workflows (e.g. Qiita [18], MGnify [19], and NMDC [20]). Similarly, various metadata standards were developed to establish a controlled vocabulary for sample collection, preparation, and data processing methods [21,22]. The Genomic Standards Consortium established the Minimum Information for any (x) Sequence standards (MIxS), with specific checklists for marker-gene [17] and metagenome data [23]; these are currently supported by major sequence data repositories.

Despite such institutional efforts, compliance with these data sharing standards among researchers is still limited [22].

Consequently, an increasing number of microbiome studies fail to meet standards and hamper the scientific community's efforts toward open data sharing [24]. Numerous microbiome studies either provide no accessible sequence data, or deposit the raw sequence data with insufficient or mismatched metadata that can not effectively be linked to the provided sequence data [25–27]. This is a wider problem across numerous scientific disciplines, with publication data often remaining incomplete [28,29]. According to estimates, only a quarter of papers in *Science* meet computational reproducibility [30] despite its policy requiring sharing of "all data necessary to understand, assess, and extend the conclusions of the manuscript". Particularly frustrating is the refusal of authors to share published data despite reporting those as "available upon request" [31,32]. The vast majority of researchers are not compliant with their own published data availability statements (DAS): of nearly 1,800 reviewed publications with DAS suggesting data availability upon "reasonable request", a mere 7% of authors provided usable data [33].

Microbiome sequence data and metadata often accumulate in ways unsuitable for future research. And while depositing raw data prior to publication is mandated by most journals and funding agencies, mechanisms for assessing and validating data reporting practices are largely missing [34]. To address these gaps, we propose here a two-pronged solution to promote reproducibility in the field of microbiome science: (1) tier-based standards to assess reproducibility of publications by evaluating data accessibility statements and metadata, and (2) software to programmatically determine adherence to these tiered standards, enabling evaluation pre-publication (e.g., during review) or post-publication (e.g., for selection during meta-analysis).

Tier-based standards for microbiome sequence data and metadata sharing

We propose two tiered badge systems to evaluate the openness of research data within the microbiome research community: one for sequence data reporting, and one for metadata reporting. These parallel standards are linked to provide effective data availability reporting for both nucleotide sequence data and associated metadata (Figure 1, Tables 1-2). We consider in this framework the availability of next-generation sequencing (NGS)-based data, representing

sequences comprising microbial communities; in the microbiome field, these are microbial amplicon and metagenome sequences.

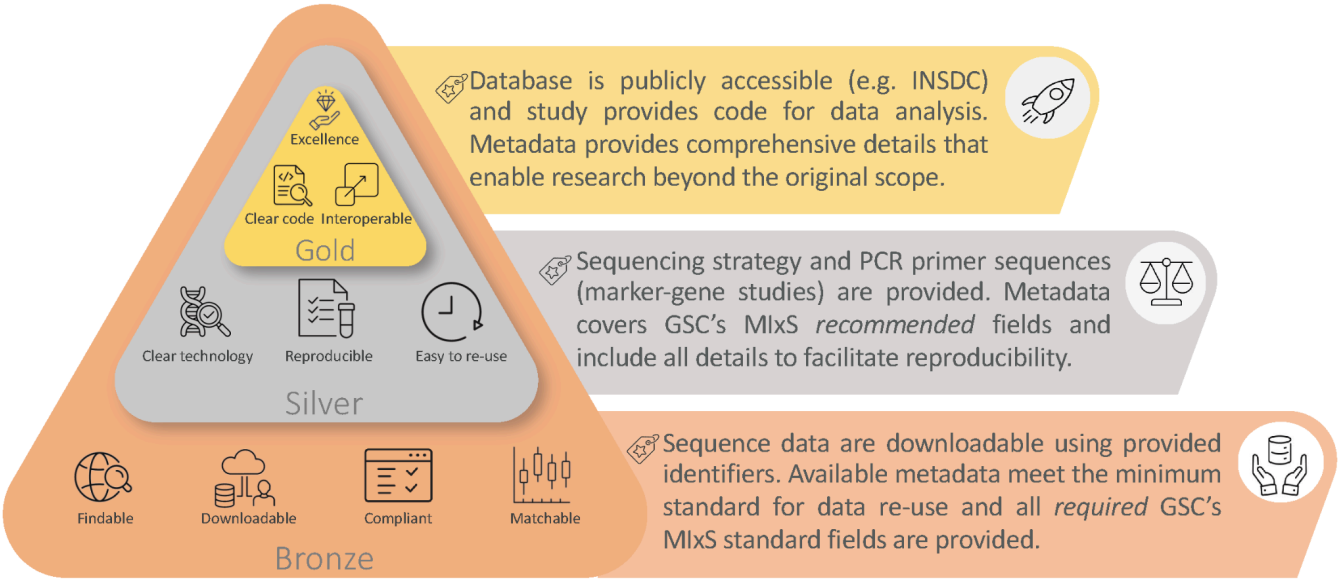

Figure 1. An integrated overview of the tiered badge system.

Table 1. Requirements for nucleotide sequence data reporting. Bronze represents a minimal set of requirements for releasing open data, while the Gold tier represents both open and FAIR research data for easy reuse.

Requirement Type	Bronze	Silver	Gold
Data accessibility	Sequence data are downloadable and not paywalled	Sequence data are downloadable and not paywalled	Sequence data are downloadable and not paywalled
Data accessibility	Accession numbers provided	Accession numbers provided	Accession numbers provided
Data accessibility	Raw sequences given as BCL or FASTQ	Raw sequences given as BCL or FASTQ	Raw sequences given as BCL or FASTQ
Sample processing metadata		Sequencing method provided: 16S or metagenomics	Sequencing method provided: 16S or metagenomics
Sample processing metadata		PCR primer sequences	PCR primer sequences
Data accessibility			If data are not public, DAS clarifies access requirements
Data accessibility			Database is publicly accessible, with no login and/or institutional affiliation e.g. SRA or Figshare
Data processing metadata			Code provided

The badge system for sequence data and metadata availability standards ranks publications by integrating well accepted—but not yet universally adopted—standards:

1. **Bronze**: Available data meet the minimum standard for data reuse. To meet this standard, the publication includes basic details such as the accession ID for the associated sequence data, and the metadata that provides environmental context for given samples. With these, researchers are able to find and, when suitable for their purpose, use data from the original publication.

2. **Silver**: Facilitated reproducibility; this is a balance between comprehensiveness and feasibility for most researchers. This standard integrates more comprehensive metadata with additional fields such as recommended fields in the MixS checklists, and other domain-specific standards. Sequence data in this tier are not necessarily openly accessible, but the publication clarifies access requirements for reuse.

3. **Gold**: These data provide valuable information to address research questions beyond the original scope. With freely and publicly accessible sequence data, any researcher is able to use these data without barriers. And with additional optional metadata details from specific MIxS environmental packages, researchers are able to effectively characterize this dataset for deeper investigation.

Table 2. Example of requirements for metadata reporting. Bronze represents a minimal set of requirements for releasing open data, while the Gold tier represents both open and FAIR research data for easy reuse. Note that while in the example below the tiers do include the fields of a study that can be used to describe an experiment, or set of experiments, in an increasingly detailed manner (from Bronze to Gold), thus enabling data reuse, reproducibility, and eventually, research beyond the original scope, Silver and Gold are somewhat study-specific, i.e., these tiers do not assume the ever-increasing number of reported MIxS fields, but rather their quality and value for reproducibility and future data reuse. The MIxS-specific fields are based on the MIxS v6.0 release (https://www.ebi.ac.uk/ena/browser/view/ERC000015); [1] mandatory field in the NCBI metagenome or environmental package (https://submit.ncbi.nlm.nih.gov/biosample/template/), [2] mandatory in the ENA default sample checklist (https://www.ebi.ac.uk/ena/browser/view/ERC000011).

MIxS v.6.0 Source	**Bronze**	**Silver**	**Gold**	**MIxS requirement**
MIxS core	Sample name	Sample name	Sample name	Mandatory [1]
MIxS core	Project name	Project name	Project name	Mandatory

MIxS core	Taxonomy ID of DNA sample	Taxonomy ID of DNA sample	Taxonomy ID of DNA sample	Mandatory [1]
MIxS core	Collection date	Collection date	Collection date	Mandatory [1,2]
MIxS core	Geographic location (latitude/longitude)	Geographic location (latitude/longitude)	Geographic location (latitude/longitude)	Mandatory [1,2]
MIxS core	Geographic location (country/region)	Geographic location (country/region)	Geographic location (country/region)	Mandatory [1]
MIxS core	Broad-scale environmental context	Broad-scale environmental context	Broad-scale environmental context	Mandatory
MIxS core	Local-scale environmental context	Local-scale environmental context	Local-scale environmental context	Mandatory
MIxS core	Environmental medium	Environmental medium	Environmental medium	Mandatory
MIxS core		Sequencing method	Sequencing method	Mandatory
MIxS core		Negative control type	Negative control type	Optional
MIxS core / MIMARKS checklist		Target gene	Target gene	Mandatory
MIxS core / MIMARKS checklist		PCR primers	PCR primers	Optional
MIxS human-gut package		Host subject ID	Host subject ID	Optional
MIxS human-gut package		Host disease status	Host disease status	Optional
MIxS human-gut package		Host total mass	Host total mass	Optional
MIxS human-gut package		Host sex	Host sex	Optional
MIxS core			Sample material processing	Optional
MIxS core			Nucleic acid extraction	Optional

MIxS human-gut package			Host height	Optional
MIxS human-gut package			Host diet	Optional
MIxS human-gut package			Host occupation	Optional

To support reporting and evaluation of these tier-based standards, we developed software that programmatically validates adherence to these tiered standards in published literature. The Python package MISHMASH (MIcrobiome Sequence and Metadata Availability Standards) is available at https://github.com/bokulich-lab/mishmash. To assess the effectiveness of our standards and software, we validated the MISHMASH program on two Case Studies (Box 1).

Box 1. Case studies reveal poor levels of sequence data reporting and temporal trends

The sequence data availability tiers were evaluated using two case studies: (1) a cross-section of human gut microbiome publications released between 2024 January 01 and 2024 June 30 (n = 85), and (2) a list of human gut mycobiota publications from 2013 to 2023 (n = 183). Publications were screened on the NCBI PubMed database for peer-reviewed, original research papers with licenses allowing for full-text automated retrieval.

Of the studies in Case Study 1, we found that nearly half of all publications did not meet the minimum standards for open sequence data (Figure 2). Nearly half (44.7%) of articles note some variation of “Data are available upon reasonable request”, report BioProject numbers with no associated data, or simply share no data. According to our tiered requirements, such articles do not qualify for a badge assignment at all. The vast majority of the remaining publications (40.0%) were assigned the Silver badge, a higher standard of openness that balances availability with feasibility. About 7% and 8% of publications were assigned Bronze and Gold badges, respectively. The Gold standard publications serve as role models for facilitated reproducibility: with not only required details on sequence data but also associated code, these publications facilitate reproducibility of the original study with ease.

The set of human gut mycobiome literature in Case Study 2 has similar proportions of badge assignments to the 2024 H1 set in Case Study 1 (Figure 2). A higher frequency of publications report minimal data accessibility in Case Study 2 (20.8% rather than 7%), but patterns remain similar:

Publications which fail to qualify for a badge are still abundant, and publications with the Silver badge again outnumber those with Bronze and Gold.

As the human gut mycobiome literature set spans ten years of publications, we next explored temporal patterns in data availability. The primary trend is of total frequency (Supplementary Figure 4): The number of annual publications steadily increases from one article in 2013 to 37 by the end of 2023 quarter 3. This is possibly explained by the decreasing cost of sequencing [35] and growing interest in gut fungi in relation to human health [36]. The proportion of publications without a badge assignment (Nones) slowly decreases over time (Supplementary Figure 4). Though there were few human mycobiome articles published in 2013-2016, they consistently hovered at 50% with unreleased sequence data. From 2017 onward, the proportion of Nones remained steadily below 50% and approached 30% in recent years. The number of publications with accompanying code is greater now than a decade ago; while only 3 (of 42) publications from 2013-2019 released their code, 16 papers (of 141) from 2020-2023 made their code repositories available.

We were also interested in the potential influence of the journal's publisher on data openness. Publishing groups have their own data reporting guidelines, and we investigated their effectiveness in motivating good reporting practices. Of the 183 publications analyzed in Case Study 2, 146 (79.8%) were published in journals whose publishing groups included 10 or more publications from our list. Among these, we found data accessibility varies widely by publisher (Supplementary Figure 5). The proportion of published articles with no publicly available sequence data ranges from 18% to 50%. The Public Library of Science (PLOS) tops our list for publisher with both the lowest proportion and the lowest total count of None-badged publications; on the other hand, PLOS has published no recent Gold-standard publications under our criteria. BioMed Central and MDPI both have high None-badged publications, exceeding 40%, though BioMed Central also has the second-highest proportion of Gold-badged publications.

Data accessibility statements: a ‘window’ into reality of data sharing trends

Quantitative output metrics from screening data accessibility statements (DAS) in 20 years worth of human gut microbiome publications (*n* = 2,947 publications between 2003-2023) provide a domain-specific perspective to the state and needs of the microbiome research field. Large-scale survey of the DAS using our automated badge prediction algorithm MISHMASH

(Supplementary Methods) revealed three consistent trends: (i) nearly half of all publications (45.3%) did not meet the minimum standards for open sequence data (Nones), (ii) on the other extreme, only a few publications (8.7%) qualified for the Gold badge, and (iii) publication with the Silver badge outnumbered those with Bronze and Gold - consistent with our expectation that the Silver standard represents a balance between comprehensiveness and feasibility for many researchers. Our survey (Figure 3) yielded remarkably similar estimates to manual badge assignments for Case Studies 1 and 2 (Box 1), further reinforcing these findings and suggesting that such data reporting practices (i.e., high rates of missing sequence data) are consistent at the repository scale. Interestingly, these trends likely reflect distinct user communities: researchers who simply share no data or report that data are “available upon request”, and researchers who exceed the minimum requirements, aiming for a higher standard of openness that balances availability with feasibility, as indicated by higher inclusion rates for Silver vs. Bronze badges in both case studies and our large-scale literature survey.

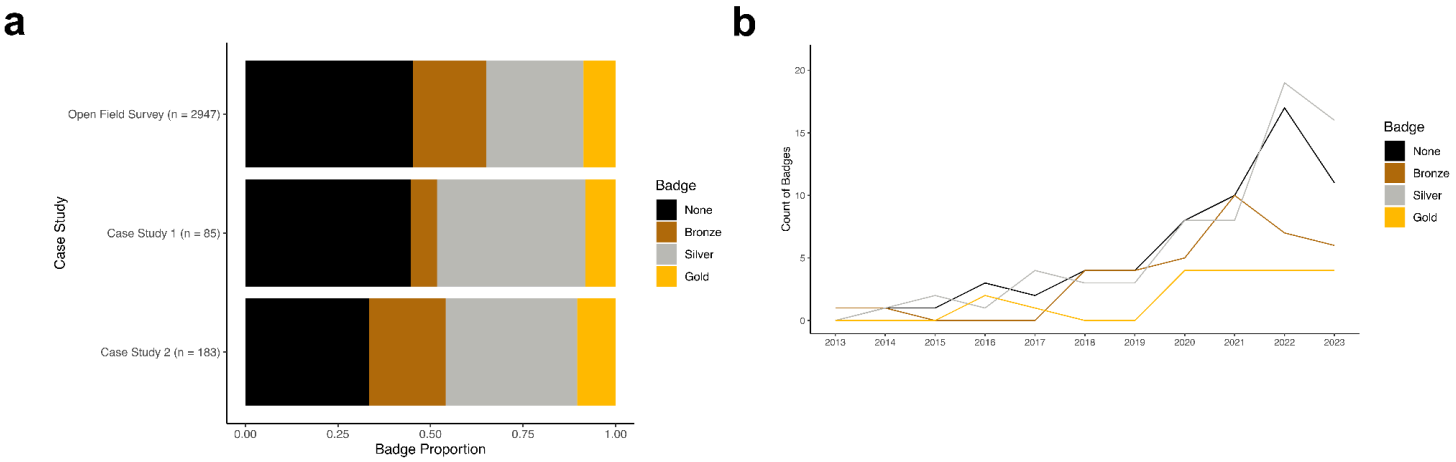

Figure 2. Patterns in badge assignments for various cross-sections of microbiome literature. **a)** Predicted assignments for an open field survey of human gut-associated microbiome publications published from 2003 through 2023 (*n* = 2947), manual badge assignments for human gut-associated microbiome publications published from January through June 2024 (*n* = 85), and manual badge assignments for human mycobiome publications published from 2013 through 2023 (*n* = 183). **b)** A temporal look at badge assignments for Case Study 2, by total frequency per year.

We note that the high proportion (>45%) of None-badged publications with statements of “data available upon reasonable request” is concerning, as such data accessibility statements can include ostensible promises of data which often never materialize [33]. Overall, these findings are consistent with proportions of data availability previously reported by Tedersoo et al. (2021) [16], in

which an average of 54.2% articles had accessible data prior to contact. While the statistics listed above focus on total counts, temporal patterns in data availability revealed encouraging trends (Supplementary Figure 4). In human mycobiome studies, researchers have been adopting better data reporting practices at higher rates than in the mid-2010s, likely due to stricter peer-review monitoring of data submission over the last decade and the growing number of initiatives that encourage data sharing by adding visibility to accessible data [37]. Differences in data reporting guidelines [27] and the adoption rate of ORD standards (including variation in the peer-review process across journals) are also likely responsible for the apparent influence of a journal's publisher on the distribution of assigned badges and data openness (Supplementary Figure 5).

Systematic issues in current metadata sharing practices

Despite the critical role of metadata in microbiome research, standardized metadata is not commonly available, leaving microbiome datasets incomplete and limiting their utility [34]. Indeed, most articles in our Case Study 1 did not qualify for a badge assignment when considering the reported metadata. This contrasts with the data availability statements for the same publications, where about half of the screened papers met eligibility for Bronze, Silver, or even Gold badges (Box 1). Metadata records are generally more complex and can be difficult to evaluate automatically using scraper tools. However, our implementation relies on a set of requirements for metadata reporting defined by the Genomics Standards Consortium's MIxS checklists that should be standardized and machine-readable in a compliant report [17]. Although adoption of the MIxS metadata standards is widespread across the microbiome community [22], metadata associated with most samples in INSDC databases remain poorly standardized, limiting the potential for microbiome data reuse.

To demonstrate some of the issues with current metadata sharing practices, we examined 42 publications with INSDC-based datasets from Case Study 1 (Box 1). We extracted metadata associated with each article's INSDC entry to evaluate badges for metadata reporting. Only 21 of these datasets used a MIxS metadata scheme, while those remaining submitted INSDC-required minimal metadata (e.g., SRA/ENA specific checklists); all datasets also included additional attributes relevant to each study. However, we found that the flexibility of adding non-standard metadata leads to considerable variation, making it challenging to harmonize metadata across studies.

While some metadata attributes such as Organism are complete for every INSDC-submitted sample, others are less commonly utilized. Sample names are submitted to the INSDC through seven different terms depending on the study and data submitter (Supplementary Table 3), such as “Sample Name” and “User_sample_ID”. Sample environmental contexts and microbiome host identifiers are likewise spread across variously named metadata attributes. In all, the union of metadata attributes from 42 studies totaled 220. Only 16 of these attributes were consistently filled across every study, and nearly 80% of metadata attributes had entries from fewer than eight publications (Supplementary Figure 3). The median attribute included metadata from only a single study (2.4%), showcasing the immense sparsity of poorly labeled metadata. Due to the startlingly small intersection of metadata attributes submitted to the INSDC among studies, further standardization is necessary before implementing an automated badge assessment tool for metadata. This highlights two related problems: on the one hand, metadata reporting compliance appears very low in the field if we adhere strictly to the MIxS reporting standards; on the other hand, greater flexibility is clearly necessary to evaluate metadata reporting standards, as different standards are needed in different sub-fields.

More broadly, our findings highlight several systematic issues in current metadata collection and sharing practices. For example, several specific problems hinder efforts to evaluate existing metadata using the badge system alongside the DAS: (i) metadata often lack standardization and are reported without controlled vocabulary or standardized terminology, (ii) similar information is recorded in different fields across studies, (iii) information is reported, but the MIxS package requirements are applied incorrectly. In our Case Studies, the diversity of terms used to describe similar variables increased the dimensions of reported metadata, making it challenging to assess reporting practices automatically (Supplementary Table 3, Supplementary Figure 3). More importantly, these same issues create barriers to data reuse, as it can be difficult or even impossible to harmonize such metadata records across studies without extensive manual inspection and curation. Similarly, a recent survey of public microbiome data identified over 3,000 different descriptors used to report common elements of sample metadata, such as those matching host age or sex [38]. Such basic metadata associated with microbiome studies is also characterized by high levels of sparsity or missingness (e.g., >81.6% missing values for host age [38]). Although a number of resources and tools aimed at simplifying the annotation and formatting of sequence data and metadata exist (e.g., METAGENOTE [39]), their adoption is slow and these trends likely reflect the lack of centralized validation of standards or metadata quality during the submission process. As a result, even microbiome studies that

share data in the INSDC databases (Bronze and above for DAS) often report only the minimally required metadata, which frequently falls short of minimum requirements for study-level reproducibility, lacks standardization, and contains free-text entries prone to inconsistencies and errors.

The misalignment between DAS and metadata reporting in terms of badge assignments in our Case Studies is concerning, as insufficient descriptive metadata directly limits the utility and future reuse of sequence data as they are inter-connected [40]. Recognizing this issue, some initiatives are working on enriching microbiome data using text mining to extract relevant metadata terms from associated publications [41]. Such metadata rescue efforts can help when basic structured metadata is missing (e.g., geographic annotation), but retrospective mapping of more complex metadata terms remains challenging. To this purpose, our proposed metadata badge system may serve as a qualitative guide for metadata reporting and harmonization in the microbiome field.

Ensure quality by (self-)evaluating data sharing practices

In parallel with other fields in the life sciences, microbiome research is an increasingly data-driven domain that relies on large-scale, high-dimensional datasets. Unrestricted access to sequence data and metadata is essential for scientific transparency and innovative reuse, and is therefore increasingly encouraged by the microbiome research community and required by many publishers and funding agencies. Despite the active development of data sharing standards and best practices in microbiome research, here we show that many studies still suffer from poor data accessibility and low-quality metadata, limiting efficient data reuse and constraining the potential benefits of meta-analyses and synthesis of published results. At least partly, this issue is due to the lack of mechanisms for assessing and validating data reporting practices and compliance with ORD reporting requirements, and lack of awareness of best practices among authors.

As a single binary standard is insufficient to cover a wide range of cases, we propose a sliding scale of specific and measurable tiers that reflects usability and ease to other researchers [38]. Our tiered “badge” system is designed to evaluate data reporting based on the level of information provided by authors (Figure 1): Bronze-level contains the minimum required details for data accessibility and reuse, while Gold-ranked contributions allow for data access with little to no effort by users. Such a system provides coherent metrics to (self-)evaluate the quality of a

dataset prior to investing time in reuse. In tandem with a user-friendly software for automated assessment, these unified standards and tier system will enable researchers, journals, and funding agencies to automatically assess compliance and quality of submitted datasets or publications, providing users with immediate, informative feedback [22,42].

Box 2. Future directions for automating data and metadata sharing compliance assessment

Incomplete semantic processing necessitates limited manual checks and validation. That said, accuracy rates for our automated badge prediction algorithm indicate that MISHMASH can correctly assign an appropriate badge to most microbiome studies (overall accuracy 84%; see Supplementary Methods). Most assignment errors can be attributed to the difficulty in discriminating between badge tiers as opposed to between None vs. Badged publications, showing that our prediction algorithm performs exceptionally well (overall accuracy >97%) in establishing sequence data availability. In this initial proof of concept work we were limited to databases that have a robust API and thus were only able to automatically evaluate open-access PMC publications which are available for web scraping; similarly, sequence data were only programmatically assessed if found in INSDC databases and certain common databases such as the Metagenomic Rapid Annotations using Subsystems Technology (MG-RAST) [43] or the Genome Sequence Archive of the Chinese National Genomics Data Center (GSA) [44]. Although the PMC and INSDC do not encompass the entirety of scientific research, we expect our findings to be representative, as PMC and INSDC are among the largest and the most extensive of research databases. However, we hope that future iterations of our code will be able to automatically evaluate sequence data from less commonly used databases as well. We encourage the storage of sequence data in such databases rather than personal web pages due to issues with maintenance and “link rot” in the former [45].

An ethical issue arises in evaluating the accessibility of data locked under confidentiality agreements. Particularly with personally identifiable clinical data, patient confidentiality prevents completely public availability. Initiatives such as the European Genome-Phenome Archive (EGA) [46,47] provide a workaround, as a secure archive released to authorized researchers. While data within the EGA is currently outside the scope of this proof of concept work, we would support a future integration to extend badging assessments to publications associated with EGA data.

Poor standardization of metadata remains a critical issue across the life sciences [1], including microbiome research. Our findings highlight the need for continuing development of standards and best

practices that reduce the barriers to data and metadata submission. The metadata badge system that we propose here, unlike the sequence data badge system, is not yet functional for automated assessments due to the drastic variance in metadata terminology across studies (Supplementary Table 3). In the meantime, it is meant as a qualitative guide for metadata reporting standards in the microbiome field, to support efforts for improving FAIR literacy in the microbiome research domain.

Conclusions and outlook

Preparing published sequence data and metadata for reuse is time-consuming and often inefficient due to the lack of validation methods. Here we propose tier-based standards and contribute an open-source software package intended for assessment of sequence data and metadata sharing practices in microbiome research by a diverse range of downstream users. We hope that these standards and software will help to promote better practices in microbiome research by serving multiple purposes: (i) enabling authors to self-assess compliance with FAIR principles, thereby increasing awareness and adoption of these standards in the field, and (ii) providing data repositories, publishers, funding agencies, and other stakeholders with a prototype model to automatically assess compliance of submitted datasets or publications. Moreover, this functionality will also facilitate meta-analysis as automated predictions can (iii) help users with rapid and accurate identification of studies with minimally available sequence data, and thus datasets potentially suitable for reuse. Further user engagement will be needed to develop and extend this resource to serve the specific needs of user communities across the microbiome research field. While here we focus on microbiome research, our tier-based standards and software approach are extensible and these principles can be easily adopted by life sciences fields relying on similar data types, such as genomics and transcriptomics.

Additional training and coordinated educational campaigns will be needed to overcome perceived barriers to sharing data, which should start by building trust and increasing awareness regarding the value of open research data for individual scientists, wider scientific community, and society at large[48]. We anticipate that evaluating data openness across publications will both promote data-sharing practices in the field and encourage greater community support and efforts toward FAIR ORD sharing and reporting.

Acknowledgments

This work was supported by the ETH Domain's Open Research Data (ORD) Program (2020-2024) (Project MISHMASH and BOSTORD). This project was additionally supported by the grant #2021-362 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology). This work was supported by an ETH Zurich Postdoctoral Fellowship award to AL.

References

1. Price, E. *et al.* What is the *real* value of omics data? Enhancing research outcomes and securing long-term data excellence. *Nucleic Acids Res.* **52**, 12130–12140 (2024).

2. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

3. Science Staff. Challenges and Opportunities. *Science* **331**, 692–693 (2011).

4. Blei, D. M. & Smyth, P. Science and data science. *Proc. Natl. Acad. Sci.* **114**, 8689–8692 (2017).

5. Brus, L. The Rise of Computation. *Nano Lett.* **20**, 801–802 (2020).

6. Wilson, S. L. *et al.* Sharing biological data: why, when, and how. *FEBS Lett.* **595**, 847–863 (2021).

7. Blaser, M., Bork, P., Fraser, C., Knight, R. & Wang, J. The microbiome explored: recent insights and future challenges. *Nat. Rev. Microbiol.* **11**, 213–217 (2013).

8. Eren, A. M. & Banfield, J. F. Modern microbiology: Embracing complexity through integration across scales. *Cell* **187**, 5151–5170 (2024).

9. Song, S. J. *et al.* Comparative Analyses of Vertebrate Gut Microbiomes Reveal Convergence between Birds and Bats. *mBio* **11**, e02901-19 (2020).

10. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

11. Agustinho, D. P. *et al.* Unveiling microbial diversity: harnessing long-read sequencing

technology. *Nat. Methods* **21**, 954–966 (2024).

12. Kim, N. *et al.* Genome-resolved metagenomics: a game changer for microbiome medicine. *Exp. Mol. Med.* **56**, 1501–1512 (2024).

13. Cullin, N., Azevedo Antunes, C., Straussman, R., Stein-Thoeringer, C. K. & Elinav, E. Microbiome and cancer. *Cancer Cell* S1535610821004463 (2021) doi:10.1016/j.ccell.2021.08.006.

14. Averill, C. *et al.* Defending Earth's terrestrial microbiome. *Nat. Microbiol.* **7**, 1717–1725 (2022).

15. Bokulich, N. A., Lewis, Z. T., Boundy-Mills, K. & Mills, D. A. A new perspective on microbial landscapes within food production. *Curr. Opin. Biotechnol.* **37**, 182–189 (2016).

16. Tedersoo, L. *et al.* Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**, 192 (2021).

17. Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29**, 415–420 (2011).

18. Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**, 796–798 (2018).

19. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).

20. The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkab990.

21. Cernava, T. *et al.* Metadata harmonization–Standards are the key for a better usage of omics data for integrative microbiome analysis. *Environ. Microbiome* **17**, 33 (2022).

22. Vangay, P. *et al.* Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative's Workshop and Follow-On Activities. *mSystems* **6**, e01194-20 (2021).

23. Consortium, T. G. S. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

24. Langille, M. G. I., Ravel, J. & Fricke, W. F. “Available upon request”: not good enough for microbiome data! *Microbiome* **6**, 8 (2018).

25. Broderick, D. *et al.* Realising respiratory microbiomic meta-analyses: time for a standardised framework. *Microbiome* **11**, 57 (2023).

26. Butcher, M. C. *et al.* Meta‐analysis of caries microbiome studies can improve upon disease prediction outcomes. *APMIS* **130**, 763–777 (2022).

27. Data availability is the golden rule in research. *Nat. Microbiol.* **9**, 879–879 (2024).

28. Tenopir, C. *et al.* Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLOS ONE* **10**, e0134826 (2015).

29. Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biol.* **13**, e1002295 (2015).

30. Stodden, V., Seiler, J. & Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci.* **115**, 2584–2589 (2018).

31. Borghi, J. A. & Gulick, A. E. Data management and sharing in neuroimaging: Practices and perceptions of MRI researchers. *PLOS ONE* **13**, 0200562 (2018).

32. Ventresca, M. *et al.* Obtaining and managing data sets for individual participant data meta-analysis: scoping review and practical guide. *BMC Med. Res. Methodol.* **20**, 113 (2020).

33. Gabelica, M., Bojčić, R. & Puljak, L. Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *J. Clin. Epidemiol.* **150**, 33–41 (2022).

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

35. Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequence Program (GSP) Available. https://www.genome.gov/sequencingcostsdata.

36. Iliev, I. D. *et al.* Focus on fungi. *Cell* **187**, 5121–5127 (2024).

37. Hrynaszkiewicz, I. Extending Accessible Data to more articles, repositories, and outputs. *The Official PLOS Blog* https://theplosblog.plos.org/2023/07/accessible-data/ (2023).

38. Rosenfeld, G., Angelova, A., Shin, C., Quinones, M. & Hurt, D. Current challenges in microbiome metadata collection. Preprint at https://doi.org/10.1101/2021.05.05.442781 (2021).

39. Quiñones, M. *et al.* "METAGENOTE: a simplified web platform for metadata annotation of genomic samples and streamlined submission to NCBI's sequence read archive". *BMC Bioinformatics* **21**, 378 (2020).

40. Huttenhower, C., Finn, R. D. & McHardy, A. C. Challenges and opportunities in sharing microbiome data and analyses. *Nat. Microbiol.* **8**, 1960–1970 (2023).

41. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* gkz1035 (2019) doi:10.1093/nar/gkz1035.

42. Rideout, J. R. *et al.* Keemei: cloud-based validation of tabular bioinformatics file formats in Google Sheets. *GigaScience* **5**, 27 (2016).

43. Keegan, K. P., Glass, E. M. & Meyer, F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. in *Microbial Environmental Genomics (MEG)* (eds. Martin, F. & Uroz, S.) vol. 1399 207–233 (Springer New York, New York, NY, 2016).

44. CNCB-NGDC Members and Partners *et al.* Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res.* **52**, D18–D32 (2024).

45. Schloss, P. D. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *mBio* **9**, (2018).

46. Freeberg, M. A. *et al.* The European Genome-phenome Archive in 2021. *Nucleic Acids*

Res. **50**, D980–D987 (2022).

47. Lappalainen, I. *et al.* The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* **47**, 692–695 (2015).

48. Gomes, D. G. E. *et al.* Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proc. R. Soc. B Biol. Sci.* **289**, 20221113 (2022).

49. Schaan, A. P. *et al.* Temporal dynamics of gut microbiomes in non-industrialized urban Amazonia. *mSystems* **9**, e00707-23 (2024).