nature microbiology

Consensus Statement

https://doi.org/10.1038/s41564-025-02116-2

A roadmap for equitable reuse of public microbiome data

Received: 27 March 2023

Accepted: 13 August 2025

Published online: 26 September 2025

Check for updates

Laura A. Hug[1,217], Roland Hatzenpichler[2,3,4,217], Cristina Moraru[5,217], André R. Soares[5,6,217], Folker Meyer[7,8,217], Anke Heyder[9], The Data Reuse Consortium* & Alexander J. Probst[5,6,217]

Science benefits from rapid open data sharing, but current guidelines for data reuse were established two decades ago, when databases were several million times smaller than they are today. These guidelines are largely unfamiliar to the scientific community, and, owing to the rapid increase in biological data generated in the past decade, they are also outdated. As a result, there is a lack of community standards suited to the current landscape and inconsistent implementation of data sharing policies across institutions. Here we discuss current sequence data sharing policies and their benefits and drawbacks, and present a roadmap to establish guidelines for equitable sequence data reuse, developed in consultation with a data consortium of 167 microbiome scientists. We propose the use of a Data Reuse Information (DRI) tag for public sequence data, which will be associated with at least one Open Researcher and Contributor ID (ORCID) account. The machine-readable DRI tag indicates that the data creators prefer to be contacted before data reuse, and simultaneously provides data consumers with a mechanism to get in touch with the data creators. The DRI aims to facilitate and foster collaborations, and serve as a guideline that can be expanded to other data types.

Sequence data reuse has been an evolving topic over the past two decades. The Fort Lauderdale Agreement (FLA), a public declaration by biomedicine scientists supporting the free and unrestricted use of genome sequencing data, was coined in 2003, before the advent of metagenomics and during a time when sequencing was still too costly to be performed by individual laboratories[1]. The FLA concluded that large genome projects should be released before publication to allow unrestricted and immediate reuse, which would accelerate the advancement of science. The FLA strengthened the Bermuda Principles defined in 1996[2], which advocated for the release of sequence data 24 h after generation and before publication of research papers. In 2009, after the Human Genome Project highlighted the advantages of sharing data early and widely, the Toronto Statement (TOR)[3] advocated for the prepublication release of other biological data types beyond genomics data. Finally, in 2014, 141 United Nations member states and the European Union entered into the Nagoya Protocol[4] (Regulation (EU) No 511/2014), which calls on data creators and data users to develop, update and use voluntary codes of conduct, guidelines and

[1]Department of Biology, University of Waterloo, Waterloo, Ontario, Canada. [2]Department of Microbiology and Cell Biology, Thermal Biology Institute, Montana State University, Bozeman, MT, USA. [3]Department of Chemistry and Biochemistry, Thermal Biology Institute, Montana State University, Bozeman, MO, USA. [4]Center for Biofilm Engineering, Montana State University, Bozeman, MT, USA. [5]Environmental Metagenomics, Research Center One Health Ruhr, University Alliance Ruhr, Faculty of Chemistry, University of Duisburg-Essen, Essen, Germany. [6]Centre for Water and Environmental Research (ZWU), University of Duisburg-Essen, Essen, Germany. [7]Institute for AI in Medicine, University Hospital Essen, University of Duisburg-Essen, Essen, Germany. [8]Department of Computer Science, University of Duisburg-Essen, Essen, Germany. [9]Department of Psychology, Ruhr University Bochum, Bochum, Germany. [217]These authors contributed equally: Laura A. Hug, Roland Hatzenpichler, Cristina Moraru, André R. Soares, Folker Meyer, Alexander J. Probst. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: alexander.probst@uni-due.de

BOX 1

Definitions of roles of scientists and legal entities related to microbiome and sequence data

Data consumer: any legal entity interested in using public data

Data creator: entity or entities, that is, individual or multiple researchers, who designed the study, obtained the samples and intended to publish an analysis; typically assumed to have priority on analysis

Data distributor: public databases that provide access to the digital data (for example, GenBank, ENA, DDBJ; Box 2)

Data generator: entity that renders the sample into a digital object (for example, a sequencing facility that processes biological material and produces sequence-related files to transmit to another entity)

Data owner: legal entity who owns the data by rights; this can be different from 'data creator' (for example, an institute at which the data creator is employed or a nation or state)

best practices in relation to access and benefit-sharing of genetic data (see Box 1 for descriptions of the different roles of researchers working with sequence datasets).

Large-scale sequence data analysis has become mainstream with a wide array of tools available, making data mining accessible to many labs. Now, ~20 years after the FLA, GenBank holds an estimated ~5.09 terabase pairs (Tbp)[5] of biological sequence data, and the Sequence Read Archive (SRA) holds 90.89 petabase pairs (Pbp) as of February 2024 (Supplementary Fig. 1 and Box 2). These databases are several million times larger than the available sequence data at the time that the FLA or TOR was formulated. With the rapid, continuous increase in public sequence data (projected to reach ~500 Pbp in 2030, Supplementary Fig. 1), data mining projects (or those requiring large public datasets for artificial intelligence training) have increased in both frequency and scope, necessitating a revisit and potential overhaul of the 20-year-old guidelines depicted in the FLA and TOR[1,3].

In 2016, the FAIR principles for data management were defined, which place an emphasis on submitted data being machine actionable (that is, computational systems should be able to Find, Access, Interoperate and Reuse data with minimal human intervention)[6]. These principles were designed to promote good scientific practice and to serve as a guideline for those wishing to enhance the reusability of their data. The FAIR principles have since been adopted as recommendations or requirements by major funding bodies, including the US National Institute of Health and the European Commission[7]. The FAIR data principles prioritize data reuse and computer-driven data mining, and include a specific requirement for (meta)data to be released with a clear and accessible data reuse licence (principle R.1). To date, this aspect of the FAIR principles has not been implemented in a straightforward or machine-readable way, lacking a coordinated implementation between databases and the community.

Biological sciences, and particularly their subdisciplines associated with generating sequence data, have been at the forefront of data availability compared with the fields of earth sciences, mathematics, physics and chemistry[8]. For instance, astrophysics is a subdiscipline that traditionally relies on data sharing and reuse owing to the exorbitant costs of research data. A study investigating the motivational factors behind data sharing and reuse within this field identified several demotivating factors[9]. Among them were the lack of data standards, the lack of facilitating platforms, inconsistency between datasets, limited documentation, difficulties finding and reusing data, and last but not least, competition and fear of being 'scooped' (accidentally or purposefully)[9]. The latter point is considered in the FLA. While the FLA recommends swift prepublication of data generated by large sequencing consortia, it also states that "[...] the contributions and interests of the large-scale data producers should be recognized and respected by the users of the data, and the ability of the production centers to analyse and publish their own data should be supported by their funding agencies [...]"[1]. This highlights one of the significant and enduring tensions between data creators and data consumers. Both data creators and data consumers are indispensable to advancing biological sciences, particularly in the realm of sequence data analysis, in which many data creators also act as data consumers. Unrestricted public use of microbiome data, on which data creators have not yet published, does not always align with the interests of data creators.

How to achieve unrestricted data reuse and, at the same time, give due credit to data creators has been discussed by the scientific community[10]. Data reuse in this work refers specifically to those cases in which the sequence data will be featured in a publication prepared by a data consumer, whether in figures, tables or text, or as an important aspect of the workflow that leads to new insights or conclusions (see Supplementary Table 2 for some Data Reuse Information (DRI) usage scenarios). In this spirit, a recent study thoroughly analysed the pros and cons of early data release and considered both the needs of data creators and consumers. The authors proposed immediate, unrestricted release of sequence data before publication, in parallel with the adoption of a reward system (for example, separate promotion and tenure tracks) for acknowledgement of data creators by universities and research institutions[11]. In addition, the authors proposed making the datasets and the protocols used for their generation citable through Digital Object Identifiers (DOIs). If implemented, these measures would create a safer environment for data sharing, benefiting all parties involved and, most importantly, supporting the advancement of science.

Mechanisms for crediting data creators beyond citing associated publications are not yet widespread in the scientific community. Creating separate tenure tracks or other incentives for data creators and data consumers requires sizable changes in evaluation criteria, and would require substantial time to propagate through institutions. DOIs for datasets, on the other hand, seem relatively easy to implement and would provide data creators with a reportable impact metric. However, their use has not yet been widely adopted, possibly owing to associated costs with purchasing and maintaining DOIs, which can be prohibitive for many publicly funded research institutions. Potential measures to lower DOI costs could include large-scale agreements between research institutions and DOI providers. Other mechanisms of data citation have been discussed in the community but have also not been widely adopted[12,13]. Currently, data creators do not have any incentive or reward for releasing sequence data before an associated publication.

It is crucial to implement methodological and ethical guidelines that are based on the principles of good scientific practice and which are driven by the scientific community to facilitate appropriate use of public data. This need has been highlighted by recent conflicts between data creators and data consumers that played out over social media. Implementing and following guidelines for unpublished data usage by all scientists would create 'safe spaces' for data creators to publish their first analyses of data—particularly if they are delayed by resource, time or personnel constraints. The research topic also affects the expectations for open data. Research related to public health necessitates swift data release to counteract pandemics or identify zoonotic diseases. For example, in the event of a pandemic, there should be no data restriction

BOX 2

Abbreviations extensively used in microbiome data research

COGs (clusters of orthologous groups): represents a collection of proteins from complete bacterial and archaeal genomes, grouped into clusters of orthologues, and associated with functional annotations

DDBJ (DNA Data Bank of Japan): a database of nucleotide sequence data maintained by the National Institute of Genetics (NIG) in Japan

EMBL (European Molecular Biology Laboratory): a research organization that conducts basic research in molecular biology and offers a range of scientific resources

EMBL-EBI (European Bioinformatics Institute): a bioinformatics research centre belonging to the EMBL, which maintains and provides access to several sequence-related databases (for example, ENA, Interpro, PDBe, UniProt)

ENA (European Nucleotide Archive): a database of raw sequence data and annotated sequence data, from a wide range of organisms, maintained by EMBL-EBI

GenBank (Genetic Sequence Database): a database of DNA and RNA sequences from a wide range of organisms, along with associated annotations and metadata

GSC (Genomic Standards Consortium): an international organization dedicated to the development and implementation of standards and best practices in genomics and related fields

IMG/M (Integrated Microbial Genomes and Metagenomes): a data management and analysis system for microbial genomes and metagenomes maintained by the Department of Energy's Joint Genome Institute (JGI) in the USA

IMG/VR (Integrated Microbial Genomes with Virus-related Datasets): a specialized database storing virus genomic and metagenomic sequences, annotations and metadata

InterPro (integrated resource of protein domains and functional sites): a database that integrates information on protein domains, motifs and functional sites from a variety of sources

INSDC (International Nucleotide Sequence Database Collaboration): a data-sharing initiative between DDBJ, EMBL-EBI and NCBI

KEGG (Kyoto Encyclopedia of Genes and Genomes): a comprehensive database and knowledge base of biological systems, including genes, proteins and biochemical pathways; maintained by Kanehisa Laboratories in Japan

KOG (Eukaryotic Orthologous Groups): a collection of proteins from eukaryotic genomes, grouped into clusters of orthologues, and associated with functional annotations

NCBI (National Center for Biotechnology Information): part of NIH; a central repository for molecular sequence data, including several databases (for example, GenBank, RefSeq, SRA, COG, KOG and so on)

NIH (National Institutes of Health): a biomedical research agency of the federal government of the USA

PDBe (Protein Data Bank in Europe): a comprehensive collection of 3D structures of proteins and other macromolecules

PFAM (protein family database): a database of protein families, domains and functional sites

RefSeq (Reference Sequence Database): a comprehensive, non-redundant database of reference genomic, transcriptomic and proteomic sequences, for a wide range of organisms

SRA (Sequence Read Archive): a public repository for raw sequence data generated by platforms such as Sanger, Illumina, Ion Torrent and Pacific Biosciences

UniProt (Universal Protein Resource): a comprehensive protein sequence database, including additional field-specific contextual information (for example, protein domain structure and known interactions)

InterPro (Integrated resource of protein domains and functional sites): a database that integrates information on protein domains, motifs and functional sites from a variety of sources

INSDC (International Nucleotide Sequence Database Collaboration): a data sharing initiative between DDBJ, EMBL-EBI and NCBI

KEGG (Kyoto Encyclopedia of Genes and Genomes): a comprehensive database and knowledge base of biological systems, including genes, proteins and biochemical pathways; maintained by Kanehisa Laboratories in Japan

KOG (euKaryotic Orthologous Groups): a collection of proteins from eukaryotic genomes, grouped into clusters of orthologues, and associated with functional annotations

NCBI (National Center for Biotechnology Information): part of NIH; a central repository for molecular sequence data, including several databases (for example, GenBank, RefSeq, SRA, COG, KOG and so on)

NIH (National Institutes of Health): a biomedical research agency of the federal government of the USA

PDBe (Protein Data Bank in Europe): a comprehensive collection of 3D structures of proteins and other macromolecules

PFAM (Protein family database): a database of protein families, domains and functional sites

RefSeq (Reference Sequence Database): a comprehensive, non-redundant database of reference genomic, transcriptomic and proteomic sequences, for a wide range of organisms

SRA (Sequence Read Archive): a public repository for raw sequence data generated by platforms such as Sanger, Illumina, Ion Torrent and Pacific Biosciences

UniProt (Universal Protein Resource): a comprehensive protein sequence database, including additional field-specific contextual information (as, for example, protein domain structure and known interactions) on research related to the pandemic[14]. The general goal should be to promote open sharing of complete datasets as early and as widely as possible, across all institutions and individuals. This necessitates a technical framework that enhances the communication between data creators and data consumers regarding data reuse.

Here we propose a roadmap to enable equitable reuse of public microbiome data. This roadmap (1) addresses the lack of consensus in the field of microbiome research regarding public microbiome data use and reuse, (2) promotes communication between data consumers and data creators and (3) facilitates the rapid advancement of the microbiome field, including supporting the continued increases in data mining. To achieve the goal of this roadmap, we propose the introduction of a new machine-readable metadata tag, named DRI, containing Open Researcher and Contributor IDs (ORCIDs) of the data creators associated with data in public databases. The DRI will clearly indicate the point of contact for communication and if communication is desired by data creators. The ability to provide a point of contact for data reuse will lead to more rapid and complete data deposition. Following adoption by databases, authors and scientific journals would ideally integrate statements confirming that the best practices governed by DRI use were used in manuscripts and submission processes.

The roadmap is directly in line with the FAIR data principles, specifically contributing to FAIR principle R.1 in providing a machine-readable licence for data usage. This roadmap and its adoption by the scientific community (222 scientists as part of the Data Reuse Consortium—Supplementary Table 1—totalling 229 supporters, including the co-authors of this paper) will provide a citable resource regarding guidelines for public data reuse, will enable appropriate data reuse by data consumers and will reduce tension for data creators when submitting data. Ultimately, this roadmap outlines the expected practices for open data use for sequence data and represents a model for other biological data such as metabolomics or proteomics data.

Survey on data reuse

We created an anonymous survey with Google Forms, which was distributed to the international scientific community on 15 January 2024, to accumulate opinion data on a number of key topics related to this manuscript (Supplementary Box 1). Questions included in the survey were formulated in a neutral fashion with the intent of not biasing responses, which were anonymized to ensure openness and transparency. This survey was online and open for responses over a total of 21 days. Efforts to ensure widespread awareness of this survey included actively advertising it across multiple social media platforms (X.com, LinkedIn, Bluesky.app) over the duration of the survey. To achieve this, 39 authors made use of their accounts across these platforms while also leveraging their working group and other institutional accounts. A blog advertising this survey was additionally hosted at the Springer Nature Research Communities blog[15]. Finally, a total of 78 microbiology institutions across the world were contacted to increase participation from the Global South and underrepresented segments of the global scientific community related to this survey. Raw data pertaining to the anonymous responses to this survey in TSV format are available at the Open Science Foundation[16].

This resulted in responses from 306 scientists representing all continents (except Antarctica) with feedback on community interest in and likelihood of adopting the proposed roadmap (Supplementary Figs. 2–10 and Supplementary Tables 3–7). Survey questions were designed to enable quantitative analysis, namely, to appreciate the fraction of the community that agreed or disagreed with proposed aspects of the roadmap delineated in this manuscript. Raw data for this survey were imported into R 4.3.1 running in RStudio 2023.12.0 (Build 369) using the tidyverse ecosystem of data analysis packages to read TSV inputs (readr, version 2.1.5), filter (dplyr, version 1.1.4) and generate plots (ggplot2, version 3.5.1)[17,18]. For Fig. 1, survey data were processed using R, and visualizations in Fig. 1a–c were generated using ggplot2. Figure 1d was generated with a copyright-free image, coloured by normalizing height (in pixels) to percentage categories. Colours and formatting were manually edited. Sankey diagrams were generated via the ggsankey package (version 0.0.99999). Tables were generated with the kableExtra package (version 1.3.4)[19]. R code as well as system and package versions used for all data analysis in this manuscript are publicly available at a GitHub repository: https://github.com/GeoMicroSoares/DataUsage_Data_Analysis.

The initial DRI tag and the strategy for its implementation were defined by the Data Reuse Core Team before the survey. After the survey, the Data Reuse Core Team refined the roadmap and DRI strategy through conversations with members of the Data Reuse Consortium, of the Joint Genome Institute, of the European Nucleotide Archive and of the Genomic Standards Consortium, and the reviewers of this manuscript. Throughout the refinement process, the Data Reuse Core Team identified compromises between differing priorities and ensured that the DRI strategy was actionable within current database structures.

Microbiome data types most frequently reused include amplicon datasets of single genes, such as 16S rRNA or internal transcribed spacer (ITS) genes, as well as reads and assemblies of genomes, metagenomes and metatranscriptomes. These sequence data, while typically structured as machine-readable files, can come in various formats depending on the hosting repository.

Organizations hosting data usually have their own legal framework governing data download. In the case of the EMBL-EBI ENA database, for example, while records are typically available with no restrictions on reuse, the Terms of Use[20] recognize that third parties may assert restrictions on reuse for a variety of reasons. The database is working towards more systematic implementation of categorizing reusability, probably under Creative Commons (CC0) licensing[21]. This puts the responsibility on the data consumer to ensure their action is indeed legal and covered by the copyrights that may be associated with the data they are accessing, even if the licensing information is not machine readable.

The current state of data policies across online repositories of biological data, including EMBL-EBI, is largely reflective of a now outdated statement in the TOR. Here data users are advised to contact data creators to 'discuss publication plans', a demanding and often unfeasible approach given today's rate of data production and the availability of new data analysis pipelines. The TOR, from 2009, and the FLA, from 2003, are the agreements guiding the field to date, although their contents are often not well known by microbiome scientists across all career stages (Supplementary Figs. 4–6). The FLA specifies, among other things, that "sequence assemblies of 2 kb or greater by large-scale sequencing efforts" must be rapidly released. The language around the scale of data reflects the outdated guidance, and the emphasis on large-scale data creators is no longer in line with the prevalence of independent labs as data creators today. Interestingly, the conflict between data sharing and publishing a first analysis was acknowledged in the FLA, but it does not provide guidelines for navigating this concern. The TOR elaborated further on the conflict between "data [creators] and data users", stating that, in the author's experience, conflicts have rarely arisen.

The TOR lists a set of conditions (scale, utility, reference data, community acceptance) to consider for prepublication data sharing that are of limited relevance for today's data landscape[3]. Currently, individual research groups can contribute substantial datasets, both in terms of size and scientific value, and the idea of individualized, private agreements on data sharing, as suggested by the TOR, is no longer viable. In response to this need, repositories have developed their own policies (for example, the EMBL-EBI licence information). Data distributors recognize that some public datasets, although hosted by their respective services, carry additional restrictions that are currently neither easily visible nor machine readable. Given the ease of accessibility and sheer volume of sequence data, it has become impractical and, in some cases, impossible for a data consumer to verify and to comply with recommendations and restrictions.

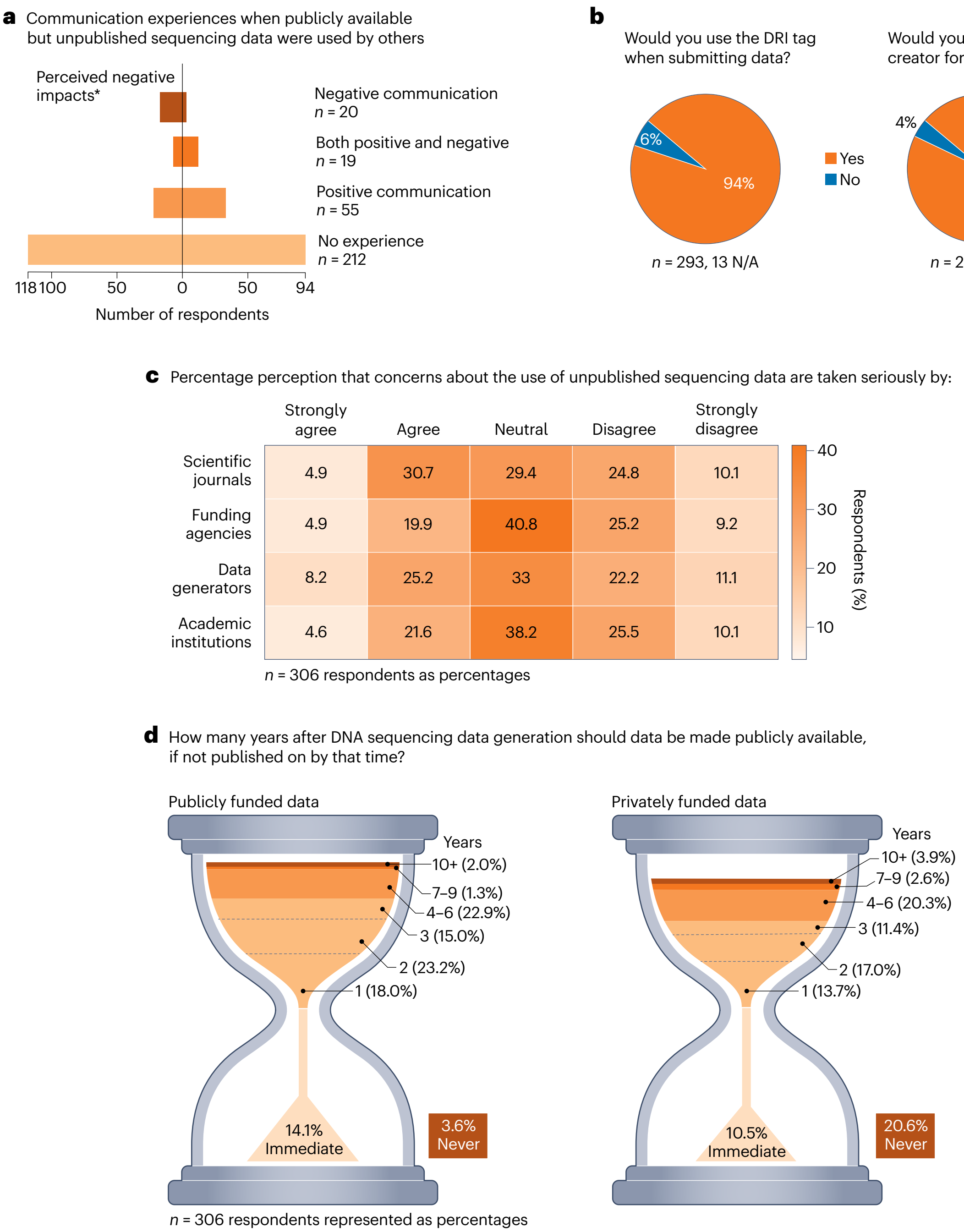

Fig. 1 | Summary results from a survey of 306 scientists on data reuse. A 21-question survey examining the scientific community's perceptions of data reuse was conducted in January 2024, with the survey distributed through social media, a blog and email to hundreds of scientists and two dozen scientific societies. A total of 306 respondents contributed to the survey. **a**–**d**, Responses were summarized for key questions and visualized using Adobe Illustrator, in which the first panel corresponds to questions 15–17 (**a**), the second panel to questions 20–21 (**b**), the third panel to questions 9–12 (**c**) and the fourth panel to questions 7–8 (**d**) (a descriptive analysis of responses to this survey and anonymous raw data to all questions are available in Supplementary Figs. 2–10 and Supplementary Tables 3–7). N/A indicates the question was left blank. Positive communication was defined as being contacted before data analysis or publication and asked for collaboration and opinion, or a positive answer when you requested data removal from a manuscript. Negative communication was defined as no contact before publication, or refusal to remove data from a manuscript upon request. The asterisk in **a** indicates that respondents agree or strongly agree that 'Unauthorized use of my sequence data by other authors has had (or will have) negative impacts on my research programme and/or my mentees'. Respondents selected single-year intervals for the data presented in the hourglass image; years 4–6 and 7–9 were combined given very low proportions for all years except year 5. Notably, respondents did not agree on a time interval after which data should be made available in the absence of an available publication (Supplementary Fig. 7). As a result, our roadmap does not include a recommendation as to when publicly available data with a DRI but no publication can be reused without contacting the data creators (Fig. 2).

Identifying conflicts of interest between data consumers and data creators

The current scale of open sequence data, including massive open datasets such as the *Tara* Oceans (7.2 Tbp of metagenomic data) and Integrative Human Microbiome (1.3 Tbp of metagenomic data as of 2019) projects, has made meta-analyses drawing on public data a powerful avenue to explore microbial systems[22,23]. Use of public data is now routine; close to 80% of respondents to our poll on data usage identified themselves as both data creators and data consumers (Box 1 and Supplementary Table 7). Access to public data has unequivocally improved the depth of the science conducted. However, it is often difficult to assess whether data reuse follows the expectations of communication and collaboration outlined in the TOR. There are many widely used software tools that include use of secondarily accessed data for which no primary publication exists (for example, the Genome Taxonomy Database, GToTree)[24,25]. Identification of the publication status associated with specific data in many repositories is not straightforward. As more governments begin to require data deposition on short or immediate timelines, there is a growing tension between data creators and data consumers around public data use. Clarifying and facilitating data reuse is therefore in the best interest of the community.

The first step is to identify roots of the potential conflict of interest between data creators and data consumers, as established following discussions between the authors of this manuscript as well as within their academic networks. These are discussed in detail below.

Disconnect between efforts of data creation and ease of reuse. One source of conflict is a disconnect between the efforts expended by data creators in generating sequence data and associated metadata, and the ease of reuse and limited or inconsistent acknowledgement of data origins by consumers. Creators sometimes feel that their monetary, time and intellectual investments to design and conduct sample collection and experiments; obtain permission for, plan, fund and carry out research expeditions; process samples; and deposit data and metadata are not adequately acknowledged or are potentially ignored by data consumers. Data creators must obtain legal documents (for example, sampling permits, visas) and follow international agreements (for example, the Convention on Biological Diversity, Nagoya Protocol), and manage the risks that come with certain fieldwork (for example, treacherous terrain, wilderness areas, areas with high criminal activity). In addition, they must secure funding for custom-design vehicles and instrumentation needed for sample collection (for example, drill ships, research vessels, submarines, buoys, remote samplers) and maintain research sites in hard-to-access areas (for example, polar regions or the Amazon). Unbeknownst to data consumers, the original data creators may be bound by restrictive agreements on appropriate or ethical data use if research was conducted in a national park, on private land or land owned by Indigenous nations, and/or for samples obtained from human specimens or biobanks[26]. There are currently limited rewards for data creators when their data are reused, and data creators have little incentive to make detailed metadata available. Systems for reporting and incentivizing data deposition have been proposed but are not yet the norm[11,27].

Timely deposition contrasts with lengthy multi-omics analyses. Both data creators and consumers generally share ideals of open science and rapid advancement of science. However, conflicts arise from disagreements in the timing of sharing data and prioritization of access. Data creators must balance long trainee timelines with publication of datasets intended for multiple research questions. Publishing a first paper on a large dataset and depositing the full data may make additional research projects associated with that dataset vulnerable to scooping. Results from our poll suggest that 53% of researchers are concerned about negative impacts on their research programme and/or mentees from unauthorized data reuse (Fig. 1 and Supplementary Figs. 7–9). As a result, partial or raw datasets or datasets lacking key metadata are deposited in place of more polished, complete datasets with full metadata to guide interpretation of genome data. In the absence of open data, data consumers are frequently unable to access contextual information (for example, physicochemical parameters, geolocations) of the field site that are essential to accurately interpret the data, to the detriment of downstream analyses. These issues are exacerbated by public sequence databases lacking (links to) the associated metadata and a general lack of familiarity of many data consumers with the literature on, or the environmental context of, a specific system.

Perceived threats to research and career goals. Duplication of effort and the potential for lowered impact or difficulties publishing replicated results are a loss for both creators and consumers. For data creators, raw data underlying published scientific results must be made public to meet expectations for reproducibility. However, unrestricted access to public data can compromise permits, site access agreements and research ethics board approvals, all of which can negatively impact the data creators' and their mentees' ongoing research. For data consumers, even unintentional reuse of restricted data can slow research progress while appropriate permissions are sought, delay publications while data are removed and, in extreme cases, lead to paper retractions. The perceived threat is that unauthorized data reuse can also negatively impact planned research directions, funded research goals, acquisition of new funds or career perspectives of early career researchers for both creators and consumers. A lack of formal structure for data reuse causes tension for data creators and consumers alike.

A roadmap to reduce tension between data creators and data consumers

There are multiple potential avenues for mitigating the three conflicts of interest discussed above, yet not every approach is suitable or can be realized. For instance, funding agencies have the power to set rules for data release and data reuse in principle. However, besides the differentiation between private and taxpayer-funded agencies, funders usually have diverging agendas that not only differ across political borders but are also heterogeneous within a single country. To address the current tension(s) between data creators and data consumers and to update the existing agreements from more than 15 years ago, we propose a comprehensive roadmap for data reuse (Fig. 2). We recommend following this roadmap except in cases in which institutions or funding agencies have a different policy for data reuse in place or there is a restricting licence associated with the dataset itself. This roadmap was developed with the aim of minimizing friction between data creators and data consumers, while promoting open science, and involves the introduction of a new machine-readable DRI metadata tag for facilitating communication between data creators, generators and consumers.

Transparent, equitable and ethical use of public data necessitates clear labelling of its usability by data distributors. Although we initially considered a system that places limitations on free data use, following consultations with the Genomic Standards Consortium and polling 306 microbiome scientists, we have converged on an approach that focuses on both simplicity and openness while achieving nearly all the desired effects. The DRI tag would attach ORCIDs of the data creators to deposited data, signalling that data creators wish to be contacted (for example, via email) before data use[28]. This way, the DRI tag also provides stable contact information to allow data consumers to easily reach data creators. ORCIDs are both free and ubiquitous and, most importantly, are already used internally by the INSDC community. The absence of a DRI tag would signal that the authors of the dataset agree to its reuse without the need to be further involved or contacted.

The DRI will consist of a tag with one or more associated ORCIDs identifying the data creators. In computer science notation, the DRI tag will have the following structure:

DRI = {ORCID1, ORCID2, ...}

We note that, implicitly, data consumers are expected to acknowledge or cite any data they use for their scientific work. For the new ethical use of data with DRI, we expect data consumers to follow the approach summarized in Fig. 2.

Sequencing datasets published in public databases (for example, in Genbank) have tags, attributes and fields that indicate which publications are connected to the respective datasets. Examples of such tags

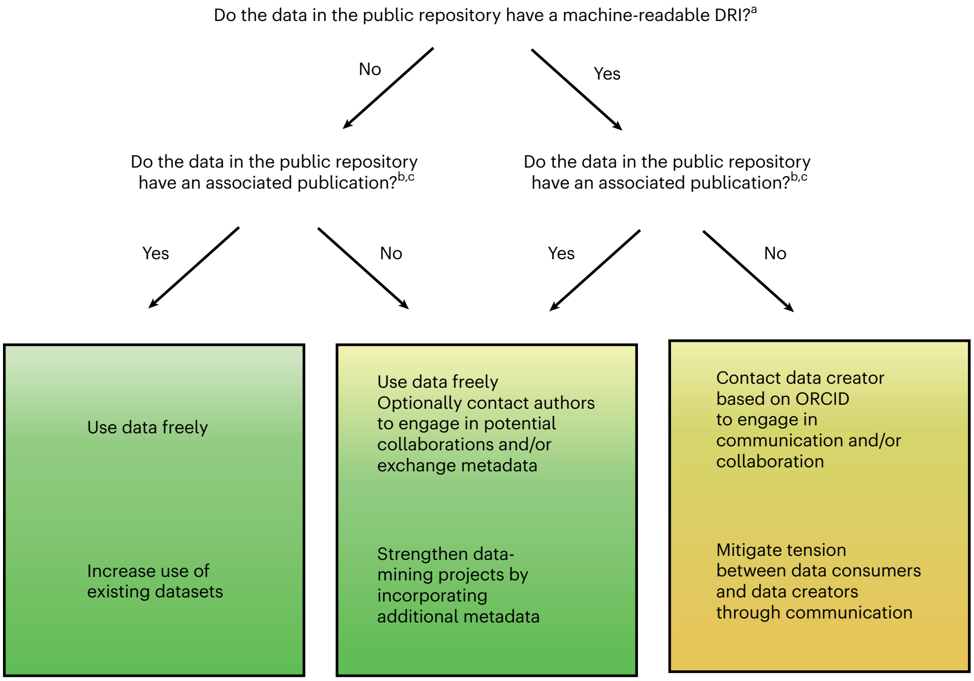
Do the data in the public repository have a machine-readable DRI?[a]
No
Yes
Do the data in the public repository have an associated publication?[b,c]
Do the data in the public repository have an associated publication?[b,c]
Yes
No
Yes
No
Use data freely
Increase use of existing datasets
Use data freely
Optionally contact authors to engage in potential collaborations and/or exchange metadata
Strengthen data-mining projects by incorporating additional metadata
Contact data creator based on ORCID to engage in communication and/or collaboration
Mitigate tension between data consumers and data creators through communication

Fig. 2 | Recommendations for equitable reuse of public microbiome data. This flow diagram is applicable unless there is a differing policy in place by the institution using the data consumer or there is a restricting licence associated with the dataset itself. [a]DRI can be updated or added. [b]Manuscripts on a preprint server or in a peer-reviewed journal are included in these recommendations. [c]Repositories could more easily update publication status.

are the following: (1) for GenBank entries, 'REFERENCE', 'REFERENCE/AUTHORS', 'REFERENCE/TITLE' and 'REFERENCE/JOURNAL'; (2) for BioSample, 'reference for biomaterial'; and (3) for BioProject, 'Publications'. The content of these tags is input initially and/or updated by the data submitter. In theory, they can be updated later either manually by the submitter or automatically by the system. Large scientific literature databases such as PubMed (https://pubmed.ncbi.nlm.nih.gov/) and Europe PMC (https://europepmc.org/) actively monitor published scientific articles and index listings of sequencing accession numbers, generating crucial linkage information that can help address updating such information in public sequence databases. However, as of 6 November 2024, only 147,632 and 78,430 publications for PubMed and Europe PMC, respectively, have had 196,632 and 91,440 sequence accession numbers assigned. Linkage information on literature and sequence data stored in these databases contrasts poorly with the scale of exponentially growing data hosted in the NCBI SRA, amounting to 9 million accession numbers as of 28 March 2023 (12 petabytes of sequence data). Coordination between publication and sequencing databases could in theory be improved if both database types use ORCID associated with their entries. In our roadmap, the DRI, which will contain the ORCID of at least one of the data creators (typically the corresponding creators, that is, the project leader), could address these issues. The content of this tag would be input or updated solely by the data submitter, preferably during the initial data deposition.

The presence of a DRI tag indicates that data creators prefer to be contacted if a data consumer reuses their data, especially if the respective data have no associated publication. The intentions behind this preference can be manifold, including a willingness to share additional metadata or datasets, or the preference to collaborate to help protect early career researchers' (for example, PhD students) ability to finish their studies and graduate. Including one or several ORCID(s) with the DRI will provide a stable point of contact, bolster transparency in science and adherence to the FAIR principles (findable, accessible, interoperable and reusable)[6] and also facilitate science through the exchange of metadata and increased collaboration. There have been instances in which such collaboration with data creators, for example, provision of additional metadata that are not publicly available, has strengthened the content of research studies[29,30]. This should be the norm rather than the exception. In other cases, communication with data creators has allowed proper citation of datasets used, and acknowledgement of funding that supported critical datasets, thus providing some benefit to the data creator for their data reuse[29,31].

In the long run, the DRI tag would help facilitate automatic updates of the publication-associated tags. For example, a GenBank sequence entry could be updated as follows: if the GenBank entry or its corresponding BioProject and BioSample entries are mentioned in a PubMed-indexed publication together with the ORCID(s) found in the DRI, then the new publication will be automatically added to the respective publication tags. The proposed metadata tag (that is, the DRI) will substantially reduce the amount of time needed to clarify the status of any public dataset, enabling automated rules for dataset screening and reducing tension between data creators and data consumers in the future. The DRI thus bridges a gap generated by the FAIR principles, for those who are invested in making their data open access, but where it is equitable for the creator to maintain some control over its reuse.

Ideally, DRIs would propagate automatically within databases (for example, from datasets of sequencing reads to the assembled genomes) and across other online databases of -omics sequence data for a given data creator. In the absence of automatic connections, data consumers can manually add DRIs to downstream data depositions (for example, metagenome assembled genomes) via a custom metadata field. These can be the original DRIs from the original data creator or, following conversation with the data creator, new DRIs connected to the data consumer.

Outlook

In this Consensus Statement, we propose a roadmap to facilitate equitable data reuse in the microbiome field. The implementation of machine-readable labels reflecting contact information of data creators will permit efficient reuse of data, accelerate scientific discoveries

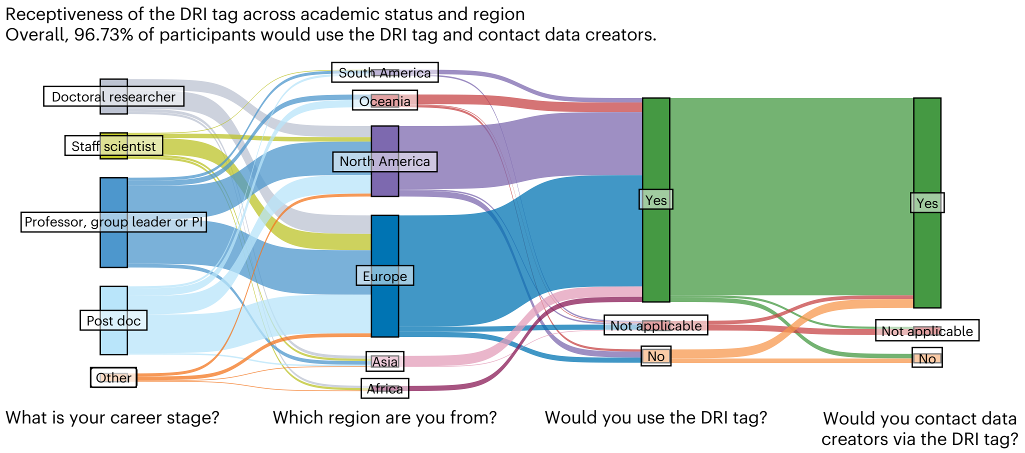

Fig. 3 | The receptiveness of survey participants towards implementing a DRI tag. Flow height (response category, in colour) per stage (survey question, *x*-axis) is proportional to number of responses. See Supplementary Table 2 for respondent numbers in each category. PI, principal investigator.

and hopefully lead data creators to more readily share their valuable datasets with the scientific public at an earlier stage. Using the standards devised by the GSC[32], complete metadata are expected to be made available by data creators. Here we propose the addition of a DRI tag to GSC metadata. Enabling equitable use of public microbiome data will rely on close collaborative work between the data creators, data distributors and data consumers. We envision that the GSC will play a pivotal role in helping all parties (that is, data distributors) implement a DRI tag within submission systems for microbiome data to public repositories, as well as in socializing the new approach to data mining (that is, the flowchart and algorithm given above). We note that even tangible incentives (for example, MG-RAST granting priority access to computing for data made public[33]) have not alleviated data creators' hesitancy to make their data available. However, while we do not anticipate the DRI will achieve a complete resolution to this challenge, we do think it is an essential first step in the right direction.

The full adoption of a DRI tag for microbiome data in public repositories will ultimately require broad support from the scientific community, data distributors, and journals and publishing houses. As an encouraging first step, the ENA has independently implemented an ORCID metadata category, allowing data creators to attach identifying information to their submissions. Propagation of this practice to other major databases will set the stage for the DRI to be used to screen for data availability. In the meantime, data creators are encouraged to apply DRI metadata tags to their datasets. This will allow data consumers to connect with data creators to discuss data reuse. We, the Data Reuse Core Team and Data Reuse Consortium, propose that scientific manuscripts partially or exclusively making use of public data should include a written statement by the authors confirming that they have complied with these guidelines for public data use. This statement would include protocols for data download and use of analysis tools, or reproducible workflows describing how the tag was incorporated in the workflow or, in case of a missing DRI, how authors adhered to the roadmap outlined in Fig. 2.

We are aware that the implementation of the DRI could substantially impact the timeline from analysis to publication by increasing the workload (for example, needing to identify email addresses of data creators). This is especially true for projects accessing many datasets (for example, mining single genes for phylogenetic trees). We are confident that, with time, automated and standardized informatic tools will become available that will lower the administrative burden of following the DRI guidelines.

At the same time, the high proportion of participants who stated that they would respect the DRI when reusing data (266 participants (96.73%); Fig. 3) suggests that data creators will be able to more freely and more frequently share their data in the future. This would facilitate adoption of the FAIR principles in science and hugely benefit the scientific community in the long run. Moreover, by fostering collaborations, the scientific best practices, and the roadmap for data sharing in microbiome research as introduced here, will enable research by lower-resourced laboratories, reducing financial bias in scientific progress. However, we do not expect that the recommendations in this roadmap will be applied retroactively to datasets already deposited in public databases before the implementation of the DRI.

The scale of nucleic acid sequence data required early pioneers to establish public databases across political borders and to confront data sharing considerations early. Other omics technologies are maturing (for example, proteomics), and scientists are recognizing the need to establish data mining approaches[34]. We propose that this roadmap for equitable reuse of public sequence data should be expanded to other fields including but not limited to proteomics, lipidomics, metabolomics, phenomics, microscopy and spectroscopy as data mining becomes routine with these types of data.

References

1. The Wellcome Trust *Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility* (National Human Genome Research Institute, 2003).
2. *Report of the International Strategy Meeting on Human Genome Sequencing held at the Princess Hotel, Southampton, Bermuda, on 25th–28th February 1996* (unpublished manuscript, 1996); http://hdl.handle.net/10161/7715
3. Toronto International Data Release Workshop Authors. Prepublication data sharing. *Nature* **461**, 168–170 (2009).
4. Parties to the Convention on Biological Diversity. *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity* 234–249 (Official Journal of the European Union, 2014); https://eur-lex.europa.eu/eli/agree_prot/2014/283/oj
5. *GenBank and WGS Statistics* (National Center for Biotechnology Information, accessed February 2024); https://www.ncbi.nlm.nih.gov/genbank/statistics/
6. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
7. Data Management. *NIH Grants & Funding* https://sharing.nih.gov/data-management-and-sharing-policy/data-management (2025).
8. Womack, R. P. Research data in core journals in biology, chemistry, mathematics, and physics. *PLoS ONE* **10**, e0143460 (2015).

9. Zuiderwijk, A. & Spiers, H. Sharing and re-using open data: a case study of motivations in astrophysics. *Int. J. Inf. Manag.* **49**, 228–241 (2019).
10. Uhlir, P. *Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop* (National Academies, 2012).
11. Amann, R. I. et al. Toward unrestricted use of public genomic data. *Science* **363**, 350–352 (2019).
12. Borgman, C. L. in *Theories of Informetrics and Scholarly Communication* (ed. Sugimoto, C. R.) 93–116 (De Gruyter, 2016).
13. Cousijn, H., Feeney, P., Lowenberg, D., Presani, E. & Simons, N. *Bringing citations and usage metrics together make data count* **18**, 9 (2019).
14. Rourke, M., Eccleston-Turner, M., Phelan, A. & Gostin, L. Policy opportunities to enhance sharing for pandemic research. *Science* **368**, 716–718 (2020).
15. Hug, L. Contribution needed for developing a new community standard for reusing sequencing data. *Springer Nature Research Communities* https://communities.springernature.com/posts/contribution-needed-for-developing-a-new-community-standard-for-reusing-sequencing-data (2024).
16. Soares, A. R. *Data Usage Manuscript - Data Deposition* (OSFHOME, 2025); https://osf.io/skw4a/
17. Wickham, H. et al. Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
18. Wickham, H. et al. *dplyr: a grammar of data manipulation*. R package version 3.1 (2023).
19. Zhu, H. et al. kableExtra: construct complex table with 'kable' and pipe syntax. R package version 3.1 (2024).
20. EMBL-EBI Terms of Use. *EMBL-EBI* https://www.ebi.ac.uk/about/terms-of-use/ (2025).
21. Licensing of EMBL-EBI data resources. *EMBL-EBI* https://www.ebi.ac.uk/licencing/ (2025).
22. Sunagawa, S. et al. Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
23. Proctor, L. M. et al. The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).
24. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* **35**, 4162–4164 (2019).
25. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
26. Jennings, L. et al. Applying the 'CARE Principles for Indigenous Data Governance' to ecology and biodiversity research. *Nat. Ecol. Evol.* **7**, 1547–1551 (2023).
27. Westoby, M., Falster, D. S. & Schrader, J. Motivating data contributions via a distinct career currency. *Proc. R. Soc. B* **288**, 20202830 (2021).
28. Credit where credit is due. *Nature* **462**, 825–825 (2009).
29. Buessecker, S. et al. An essential role for tungsten in the ecology and evolution of a previously uncultivated lineage of anaerobic, thermophilic Archaea. *Nat. Commun.* **13**, 3773 (2022).
30. McKay, L. J. et al. Co-occurring genomic capacity for anaerobic methane and dissimilatory sulfur metabolisms discovered in the Korarchaeota. *Nat. Microbiol.* **4**, 614–622 (2019).
31. Viljakainen, V. R. & Hug, L. A. The phylogenetic and global distribution of bacterial polyhydroxyalkanoate bioplastic-degrading genes. *Environ. Microbiol.* **23**, 1717–1731 (2021).
32. Field, D. et al. The Genomic Standards Consortium. *PLoS Biol.* **9**, e1001088 (2011).
33. Meyer, F. et al. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* **9**, 386 (2008).
34. Vaudel, M. et al. Exploring the potential of public proteomics data. *Proteomics* **16**, 214–225 (2016).

Acknowledgements

The Data Reuse Core Team thanks the numerous participants of our survey who chose not to be listed as authors on this manuscript for their valuable perspectives and the many people who discussed this endeavour with us on multiple occasions. We thank L. Rothe for consultations on data visualizations. A.J.P. acknowledges funding by the German Research Foundation (DFG), CRC 1439/1 and CRC 1439/2, project number 426547801 (project INF). L.A.H. acknowledges support from the Canada Research Chairs. R.H. acknowledges support from the US National Science Foundation (OCE-2049445). C. Moraru acknowledges funding by the German Research Foundation (DFG), Priority Program SPP 2330, project number MO 3498/2-1. F.M. acknowledges support from the German Federal Ministry of Education and Research (BMBF project number 01ZZ2013).

Author contributions

A.J.P. was invited to write this manuscript and formed the 'Data Reuse Core Team' via invitation. All members of this team (A.J.P., L.A.H., A.R.S., C. Moraru, F.M. and R.H.) contributed equally to this manuscript. A.J.P., F.M. and A.R.S. led discussions with JGI, ENA and GSC. L.A.H. and A.R.S. performed the analysis and visualization of survey response data. A.H. supervised the construction of the scientific survey, ensuring the quality and impartiality of questions. The 'Data Reuse Consortium' provided support and feedback to this manuscript.

Competing interests

R.H. was a member (2021–2024) of the User Executive Committee of the US Department of Energy's (DOE) Joint Genome Institute. A.P. is an affiliate scientist at JGI and sits on the Prokaryotic Advisory Committee. All opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of the DOE. R.K. is a scientific advisory board member and consultant for BiomeSense, Inc., has equity and receives income. He is a scientific advisory board member and has equity in GenCirq. He has equity in and acts as a consultant for Cybele. He is a co-founder of Biota, Inc. and has equity. He is a co-founder of Micronoma and has equity and is a scientific advisory board member. He is a board member of Microbiota Vault, Inc. He is a board member of N=1 IBS advisory board and receives income. He is a Senior Visiting Fellow of HKUST Jockey Club Institute for Advanced Study. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-025-02116-2.

Correspondence and requests for materials should be addressed to Alexander J. Probst.

Peer review information *Nature Microbiology* thanks Marnix Medema and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

The Data Reuse Consortium

R. Z. Abdallah[10], A. Abdalrahem[11], N. Abdulkadir[12], I. M. Adesiyan[13], L. Alteio[14], K. Anantharaman[15], R. Anderson[16], A-S. Andrei[17], J. A. Baeza[18,19], F. Bak[20], B. Baker[21], A. Bartholomäus[22], N. Bejerman[23], J. Biddle[24], A. Bissett[25], J. A. Blakeley-Ruiz[26], K. Block[27], J. Boldt[28], G. Bonilla-Rosso[29], T. L. Bornemann[5,6], V. S. Brauer[30], W. Brazelton[31], A. Bremges[32], E. Buelow[33], Z. M. Burcham[34], A. Cansdale[35], J. G. Caporaso[36], T. Cernava[37], I. Chatzigiannidou[38], R. Costa[39], C. R. Currie[40], A. Daebeler[41], V. De Anda[42], A. De Santiago[43], L. M. Arake de Tacca[44], J. Debelius[45], S. M. Dittami[46], X. Dong[47], M. Džunková[48], A. Edwards[49], R. Edwards[50], S. Egbert[51], J. C. Engelmann[52], S. P. Esser[5], T. J. G. Ettema[53], C. L. Ettinger[54], A. Petrovic Fabijan[55,56], R. M. W. Ferguson[57], P. Ferretti[58], P. Foucault[59], J. A. Fuhrman[60], A. M. Gada[61], P. Geesink[53], I. R. Gerhardt[62], M. O. Gessner[12,63], D. Giovannelli[64], D. Gittins[65], G. B. Gloor[66], R. A. González-Pech[67], C. Gopalakrishnappa[68], C. Greening[69], R. Gregor[68], A. C. Gregory[70], H.-P. Grossart[12,71], M. Groussin[72], B. Valenzuela Guerrero[73], M. Guzel[74], N. Hamamura[75], T. L. Hamilton[76], J. N. Hamm[77], L. Hart[78], C. Hassenrück[79], M. Hay[80], R. M. Hechler[81], P. Hellwig[82], M. Henson[83], M. Herold[84], P. J. Hesketh-Best[76], M. Hess[85], L. Hillary[85], T. C. Hitch[86], S. S. Hivarkar[87], K. J. Hoff[88], E. F. Hom[89], S. Hou[90], L. W. Hugerth[91], Y. Hwang[92], N. Ilott[93], Z. J. Jay[3], S. P. Jungbluth[94], E. Karimi[95], Y. M. Kaspareit[96], C. Keating[97], M. Kellom[98], E. A. Kiledal[99], I. Klarenberg[100], R. Knight[101,102,103,104], A. K. Koech[105], E. V. Koonin[106], K. Kormas[107], K. Kujala[108], N. C. Kyrpides[98], S. L. La Rosa[109], C. C. Laczny[110], K. Lahmers[111], X. Lan[112], A. A. Lateef[113,114], S. H. Lau[115], F. Leese[6,116], M. Á. Lezcano[117], S. S. Li[118], R. N. Lima[119], S. Lücker[120], A. Mahnert[121], S. Majidian[122], L. Malfertheiner[123], A. Marshall[124], S. Meaden[35], C. J. Meehan[125], D. V. Meier[126], C. Melkonian[127,128], D. R. Mende[129], J. L. Meyer[130], G. Michoud[131], V. Mikryukov[132], S. Miravet-Verde[133], J. Muschiol[134], M. K. Nata'ala[135,136], J. D. Neufeld[1], S. Neuhauser[137], O. Osuolale[138], J. Osvatic[139,140], K. M. Pappas[141], D. H. Parks[142], R. H. Parry[143], P. V. Pascoal[44], C. Pavloudi[144], B. Peyton[145], J. Plewka[5], M. Poyet[146], T. Priest[133], E. K. Quaye[146], S. Ramganesh[147], T. Rattei[148], P. Rausch[72], E. Rech[149], C. Rinke[150], C. Robinson[151], A. Rodríguez-Gijón[152], L. M. Rodriguez-R[153,216], R. R. Rohwer[154], T. Roloff[155], J. A. Rothman[156], S. Rückert[6,157], S. E. Ruff[158], J. S. Saini[159], M. G. Santiago-Martínez[160], L. Santoferrara[161], M. S. Sarhan[162], J. H. Saw[163], T. Sbaffi[164], R. B. Schäfer[165], G. Schaible[145], M. Schloter[166], R. A. Schmitz[167], C. Schubert[168], O. Schwengers[169], L. Sehnal[170], A. Sekar[171], J. Sekar[172], M. M. Seyoum[173], M. B. Shah[5], I. Sharon[174], B. Siebers[6,175], E. T. Sieradzki[176], D. Skliros[177], O. L. Snoeyenbos-West[178], A. Sorbie[179], D. R. Speth[148], C. G. Sprehn[98], P. Srivastava[180], T. L. Stach[5,6], J. E. Stajich[156], J. Starke[5], A. D. Steen[181], R. Stöckl[182], T. Stoikidou[183], N. Stopnisek[184], R. Sukumaran[185], B. Sures[6,186], S. Suzuki[187], D. Tamarit[127], P. Thieringer[188], R. Y. Tito[189,190], C. B. Trivedi[191], G. Trubl[192], J. Truu[193], M. Tsiknia[194], J. Ugalde[195], L. E. Valentin-Alvarado[196], X. Vázquez-Campos[197], J. Vierheilig[198,199], F. A. B. von Meijenfeldt[200], M. Wagner[148], C. J. Walsh[201], S. Wang[202], Y. Wang[203], C.-E. Wegner[204], T. Weir[205], L. C. Weiss[206], J. L. Weissman[207], A. Wichels[208], C. L. Williams[49], T. A. Williams[209], A. Z. Worden[158], T. Woyke[98], M. Wu[210], W. Xiu[211], Y. Zhang[212], J. Zhu[213], R. M. Ziels[214] & B. Zwirzitz[215]

[10]Department of Bioinformatics and Genomics, College of Biotechnology, Misr University for Science and Technology, Giza, Egypt. [11]Université de Lorraine, INRAE, IAM, Nancy, France. [12]Department of Plankton and Microbial Ecology, Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany. [13]Department of Environmental and Occupational Health, School of Public Health, University of Medical Sciences, Ondo, Nigeria. [14]Austrian Competence Centre for Feed and Food Quality, Safety and Innovation, FFoQSI GmbH, Tulln an der Donau, Austria. [15]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA. [16]Carleton College, Northfield, MN, USA. [17]Department of Plant and Microbial Biology, University of Zurich, Zurich, Switzerland. [18]Department of Biological Sciences, Clemson University, Clemson, SC, USA. [19]Departamento de Biologia Marina, Universidad Catolica del Norte, Coquimbo, Chile. [20]Austrian Institute of Technology, Wien, Austria. [21]The University of Texas at Austin, Austin, TX, USA. [22]GFZ Helmholtz Centre for Geosciences, Potsdam, Germany. [23]CONICET, Buenos Aires, Argentina. [24]University of Delaware, Newark, DE, USA. [25]CSIRO, Canberra, Australian Capital Territory, Australia. [26]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, USA. [27]Universitätsklinikum Essen—IKIM, Essen, Germany. [28]Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany. [29]Agroscope, Zurich, Switzerland. [30]Aquatic Microbiology, Environmental Microbiology and Biotechnology, Faculty of Chemistry, University of Duisburg-Essen, Essen, Germany. [31]School of Biological Sciences, University of Utah, Salt Lake City, UT, USA. [32]Evonik Operations GmbH, RD&I Biotechnology, Halle, Germany. [33]CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, Univ. Grenoble Alpes, Grenoble, France. [34]University of Tennessee, Knoxville, TN, USA. [35]Department of Biology, University of York, York, UK. [36]Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA. [37]School of Biological Sciences, Faculty of Environmental and Life Sciences, University of Southampton, Southampton, UK. [38]Department of Biotechnology and Biomedicine, Technical University of Denmark, Kongens Lyngby, Denmark. [39]Department of Bioengineering, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. [40]Department of Biochemistry and Biomedical Sciences, M.G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada. [41]Institute of Soil Biology and Biogeochemistry, Biology Centre CAS, České Budějovice, Czechia. [42]Department of Microbiology and Cell Sciences, Fort Lauderdale Research and Education Center, University of Florida, Fort Lauderdale, FL, USA. [43]Institute of Bioinformatics, University of Georgia, Athens, GA, USA. [44]Embrapa Cenargen, Brasília, Brazil. [45]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [46]UMR8228, Station Biologique de Roscoff, Roscoff, France. [47]Third Institute of Oceanography, Ministry of Natural Resources, Xiamen, China. [48]Institute for Integrative Systems Biology (I2SysBio), University of Valencia and Spanish National Research Council, Valencia, Spain. [49]Aberystwyth University, Aberystwyth, UK. [50]Flinders University, Adelaide, South Australia, Australia. [51]Washington University at St. Louis, St. Louis, MO, USA. [52]NIOZ—Royal Netherlands Institute for Sea Research, Den Burg, The Netherlands. [53]The Laboratory of Microbiology, Wageningen University and Research, Wageningen, The Netherlands. [54]University of California, Riverside, CA, USA. [55]Westmead Institute for Medical Research, Sydney, New South Wales, Australia. [56]Sydney Medical School, The University of Sydney, Sydney, New South Wales, Australia. [57]The University of Essex, Colchester, UK. [58]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA. [59]Biology of Intracellular Bacteria Unit, Pasteur Institute,

Paris, France. [60]University of Southern California, Los Angeles, CA, USA. [61]Sokoto State University, Sokoto, Nigeria. [62]Embrapa Agricultura Digital, Campinas, Brazil. [63]Institute of Ecology, Berlin Institute of Technology (TUB), Berlin, Germany. [64]Department of Biology, University of Naples Federico II, Naples, Italy. [65]University of California, Berkeley, Berkeley, CA, USA. [66]Western University, London, Ontario, Canada. [67]Department of Biology, The Pennsylvania State University, University Park, PA, USA. [68]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [69]Monash University, Melbourne, Victoria, Australia. [70]University of Calgary, Calgary, Alberta, Canada. [71]Institute of Biochemistry and Biology, Potsdam University, Potsdam, Germany. [72]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany. [73]Departamento de Educación, Facultad de Educación, Universidad de Antofagasta, Antofagasta, Chile. [74]Department of Microbiology, University of Tennessee, Knoxville, TN, USA. [75]Department of Biology, Faculty of Science, Kyushu University, Fukuoka, Japan. [76]Plant and Microbial Biology, University of Minnesota, Minneapolis, MN, USA. [77]Department of Marine Microbiology and Biogeochemistry, Royal Netherlands Institute for Sea Research (NIOZ), Texel, The Netherlands. [78]Life Sciences Institute, University of Michigan, Ann Arbor, MI, USA. [79]Leibniz Institute for Baltic Sea Research Warnemünde (IOW), Rostock, Germany. [80]Royal Veterinary College, London, UK. [81]Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada. [82]Otto-von-Guericke-University Magdeburg, Magdeburg, Germany. [83]Northern Illinois University, DeKalb, IL, USA. [84]Luxembourg Institute of Science and Technology, Esch-sur-Alzette, Luxembourg. [85]University of California, Davis, CA, USA. [86]University Hospital of RWTH Aachen, Aachen, Germany. [87]Agharkar Research Institute, Pune, India. [88]University of Greifswald, Greifswald, Germany. [89]University of Mississippi, Oxford, MS, USA. [90]Southern University of Science and Technology, Shenzhen, China. [91]Uppsala University, Uppsala, Sweden. [92]Harvard University, Cambridge, MA, USA. [93]University of Oxford, Oxford, UK. [94]Lawrence Berkeley National Laboratory, United States Department of Energy, Berkeley, CA, USA. [95]Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, Jouy-en-Josas, France. [96]Wageningen University & Research, Wageningen, The Netherlands. [97]Durham University, Durham, UK. [98]US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [99]University of Michigan, Ann Arbor, MI, USA. [100]Systems Ecology, Amsterdam Institute for Life and Environment (A-LIFE), Faculty of Science, Vrije Universiteit, Amsterdam, The Netherlands. [101]Department of Pediatrics, University of California, San Diego, San Diego, CA, USA. [102]Department of Computer Science & Engineering, University of California, San Diego, San Diego, CA, USA. [103]Shu Chien-Gene Lay Department of Bioengineering, University of California, San Diego, San Diego, CA, USA. [104]Halıcıoğlu Data Science Institute, University of California, San Diego, San Diego, CA, USA. [105]Masinde Muliro University of Science and Technology, Kakamega, Kenya. [106]Computational Biology Branch, Division of Intramural Research, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [107]Department of Ichthyology and Aquatic Environment, University of Thessaly, Volos, Greece. [108]Natural Resources Institute Finland, Helsinki, Finland. [109]Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Aas, Norway. [110]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. [111]Virginia Tech, Blacksburg, VA, USA. [112]Nankai University, Tianjin, China. [113]University of Ilorin, Ilorin, Nigeria. [114]Department of Forest Sciences, University of Helsinki, Helsinki, Finland. [115]Department of Microbiology, The University of Hong Kong, Hong Kong, China. [116]Faculty of Biology, University of Duisburg-Essen, Essen, Germany. [117]IMDEA Water Institute, Alcalá de Henares, Madrid, Spain. [118]Microbiome Systems Laboratory, Biomedicine Discovery Institute, Monash University, Melbourne, Victoria, Australia. [119]EMBRAPA/INCT_BioSyn, Brasília, Brazil. [120]Department of Microbiology, RIBES, Radboud University, Nijmegen, The Netherlands. [121]Medical University of Graz, Graz, Austria. [122]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [123]Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [124]Department of Biology, Royal Melbourne Institute of Technology, Melbourne, Victoria, Australia. [125]Nottingham Trent University, Nottingham, UK. [126]University of Bayreuth, Bayreuth Center of Ecology and Environmental Research, Bayreuth, Germany. [127]Utrecht University, Theoretical Biology and Bioinformatics, Utrecht, The Netherlands. [128]Wageningen University & Research, Bioinformatics Group, Wageningen, The Netherlands. [129]Amsterdam University Medical Center, Amsterdam, The Netherlands. [130]Department of Soil, Water, and Ecosystem Sciences, University of Florida, Gainesville, FL, USA. [131]River Ecosystems Laboratory, Alpine and Polar Environmental Research Center, Ecole Polytechnique Fédérale de Lausanne EPFL, Lausanne, Switzerland. [132]Institute of Ecology & Earth Sciences, University of Tartu, Tartu, Estonia. [133]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, Zurich, Switzerland. [134]Ocean EcoSystems Biology Unit, Marine Ecology Research Division, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany. [135]Department of Computer Science and Interdisciplinary Centre of Bioinformatics, University of Leipzig, Leipzig, Germany. [136]Department of Data Science in Bioeconomy, Leibniz Institute for Agricultural Engineering and Bioeconomy (ATB), Potsdam, Germany. [137]Institute of Microbiology, University of Innsbruck, Innsbruck, Austria. [138]Elizade University, Ilara-Mokin, Nigeria. [139]Division of Clinical Microbiology, Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria. [140]Joint Microbiome Facility of the Medical University of Vienna and the University of Vienna, Vienna, Austria. [141]Department of Biology, National and Kapodistrian University of Athens, Athens, Greece. [142]The University of Queensland, Brisbane, Queensland, Australia. [143]School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, Queensland, Australia. [144]European Marine Biological Resource Centre-European Research Infrastructure Consortium (EMBRC-ERIC), Paris, France. [145]Montana State University, Bozeman, MT, USA. [146]Institute of Experimental Medicine, Kiel University, Kiel, Germany. [147]Department of Biochemistry, J.J College of Arts and Science (Autonomous), Pudukkottai, India. [148]Centre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria. [149]National Institute of S&T in Synthetic Biology/EMBRAPA, Brasília, Brazil. [150]University of Queensland, Brisbane, Queensland, Australia. [151]Indiana University, Bloomington, IN, USA. [152]Department of Ecology, Environment, and Plant Sciences, Science for Life Laboratory, Stockholm University, Stockholm, Sweden. [153]Department of Microbiology and Digital Science Center (DiSC), University of Innsbruck, Innsbruck, Austria. [154]UT Austin, Austin, TX, USA. [155]Institute for Medical Microbiology, University of Zurich, Zurich, Switzerland. [156]Department of Microbiology and Plant Pathology, University of California-Riverside, Riverside, CA, USA. [157]Department of Eukaryotic Microbiology, Faculty of Biology, University of Duisburg-Essen, Essen, Germany. [158]Marine Biological Laboratory, Woods Hole, MA, USA. [159]EPFL, Lausanne, Switzerland. [160]Department of Molecular and Cell Biology, The University of Connecticut (UConn), Storrs, CT, USA. [161]Department of Biology, Hofstra University, Hempstead, NY, USA. [162]Institute for Biomedicine, Eurac Research, Bolzano, Italy. [163]The George Washington University, Washington DC, USA. [164]Water Research Institute (IRSA), National Research Council of Italy (CNR), Verbania, Italy. [165]RC One Health Ruhr, Research Alliance Ruhr and Faculty of Biology, University of Duisburg-Essen, Essen, Germany. [166]Helmholtz Munich, Research Unit for Comparative Microbiome Analysis, Munich, Germany. [167]Kiel University, Kiel, Germany. [168]ETH Zürich, Institute of Microbiology, Zurich, Switzerland. [169]Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen, Germany. [170]RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic. [171]Institute of Medical Genetics and Applied Genomics, Universitätsklinikum Tübingen, Tübingen, Germany. [172]M. S. Swaminathan Research Foundation, Chennai, India. [173]Department of Poultry Science, University of Arkansas, Fayetteville, AR, USA. [174]Tel Hai Academic College, Qiryat Shemona, Israel. [175]Molecular Enzyme Technology and Biochemistry (MEB), Environmental Microbiology and Biotechnology (EMB), Department of Chemistry, University of Duisburg-Essen, Essen, Germany. [176]Department of Agroecology, Aarhus University, Denmark. [177]Laboratory of Environmental Biotechnology, Department of Biotechnology, School of Applied Biology and Biotechnology, Agricultural University of Athens, Athens, Greece. [178]Department of Environmental Science, University of Arizona, Tucson, AZ, USA. [179]Institute for Stroke and Dementia Research, LMU Klinikum, Munich, Germany. [180]School of Engineering, Cardiff University, Cardiff, UK. [181]University of Tennessee - Knoxville,

Knoxville, TN, USA. [182]Institute of Microbiology and Archaea Centre, University of Regensburg, Regensburg, Germany. [183]School of Biological Sciences, Institute for Global Food Security, Queen's University Belfast, Belfast, UK. [184]National Laboratory of Health Environment and Food, Maribor, Slovenia. [185]University of Kerala, Kerala, India. [186]Department of Aquatic Ecology, University of Duisburg-Essen, Essen, Germany and Research Center One Health Ruhr, Research Alliance Ruhr, University of Duisburg-Essen, Essen, Germany. [187]RIKEN, Wakō, Japan. [188]University of California Santa Barbara, Santa Barbara, CA, USA. [189]Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium. [190]Center for Microbiology, VIB, Leuven, Belgium. [191]LGC Biosearch Technology, Petaluma, CA, USA. [192]Lawrence Livermore National Laboratory, Livermore, CA, USA. [193]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia. [194]Soil Science and Agricultural Chemistry Lab, Dept. of Natural Resources and Agricultural Engineering, Agricultural University of Athens, Athens, Greece. [195]Universidad Andres Bello, Santiago, Chile. [196]Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA, USA. [197]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia. [198]Institute of Water Quality and Resource Management, TU Wien, Vienna, Austria. [199]Interuniversity Cooperation Centre Water & Health, Vienna, Austria. [200]NIOZ Royal Netherlands Institute for Sea Research, Yerseke, The Netherlands. [201]Department of Microbiology and Immunology, University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia. [202]Department of Microbiology, The Chinese University of Hong Kong, Hong Kong, China. [203]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. [204]Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [205]Colorado State University, Fort Collins, CO, USA. [206]Department of Animal Ecology, Evolution and Biodiversity, Ruhr University Bochum, Bochum, Germany. [207]Department of Biology, The City College of New York, New York, NY, USA. [208]Alfred-Wegener-Institut Helmholtz Zentrum für Polar-und Meeresforschung, Helgoland, Germany. [209]University of Bath, Bath, UK. [210]Australian Centre for Water and Environmental Biotechnology, University of Queensland, Brisbane, Queensland, Australia. [211]State Key Laboratory of Geomicrobiology and Environmental Changes (GMEC), China University of Geosciences, Beijing, China. [212]Department of Cell and Molecular Biology, College of the Environment and Life Sciences, University of Rhode Island, Kingston, RI, USA. [213]Li Ka Shing Institute of Health Sciences, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China. [214]Department of Civil Engineering, University of British Columbia, Vancouver, British Columbia, Canada. [215]Institute of Food Science, BOKU University, Vienna, Austria. [216]Present address: Department of Chemistry and Biosciences, Aalborg University, Aalborg, Denmark. A full list of members and their affiliations appears in the Supplementary Information.