Nucleic Acids Research, 2026, **54**, D826–D834
https://doi.org/10.1093/nar/gkaf1118
Advance access publication date: 31 October 2025
Database issue

Metalog: curated and harmonised contextual data for global metagenomics samples

Michael Kuhn [1,*], Thomas Sebastian B. Schmidt [1], Pamela Ferretti [1], Anna Głazek [1], Shahriyar Mahdi Robbani [1], Wasiu Akanni [1], Anthony Fullam [1], Christian Schudoma [1], Ela Cetin [1], Mariam Hassan [1], Kasimir Noack [1], Anna Schwarz [1], Roman Thielemann [1], Leonie Thomas [1], Moritz von Stetten [1], Renato Alves [1], Anandhi Iyappan [1], Ece Kartal [1], Ivan Kel [1], Marisa I. Keller [1], Oleksandr Maistrenko [1], Anna Mankowski [1], Suguru Nishijima [1], Daniel Podlesny [1], Jonas Schiller [1], Sarah Schulz [1], Thea Van Rossum [1], Peer Bork [1,2,*]

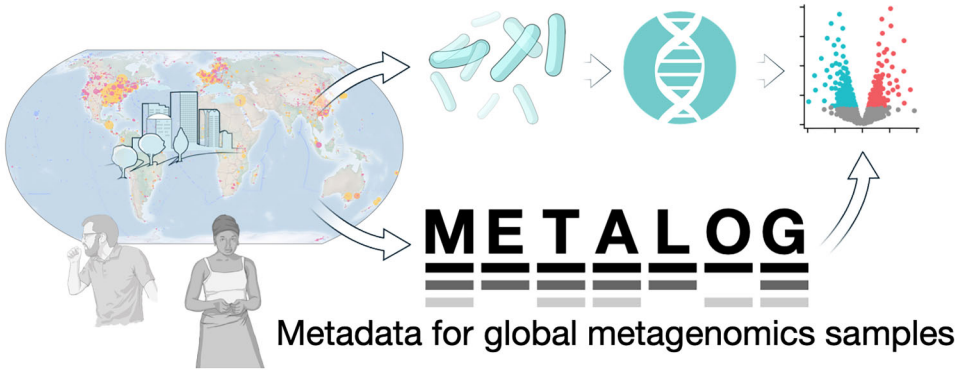
[1]European Molecular Biology Laboratory, Molecular Systems Biology Unit, 69117 Heidelberg, Germany
[2]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany

*To whom correspondence should be addressed. Email: peer.bork@embl.org
Correspondence may also be addressed to Michael Kuhn. Email: mkuhn@embl.de

Abstract

Metagenomic sequencing enables the in-depth study of microbes and their functions in humans, animals, and the environment. While sequencing data is deposited in public databases, the associated contextual data is often not complete and needs to be retrieved from primary publications. This lack of access to sample-level metadata like clinical data or *in situ* observations impedes cross-study comparisons and meta-analyses. We therefore created the Metalog database, a repository of manually curated metadata for metagenomics samples across the globe. It contains 80 423 samples from humans (including 66 527 of the gut microbiome), 10 744 animal samples, 5547 ocean water samples, and 23 455 samples from other environmental habitats such as soil, sediment, or fresh water. Samples have been consistently annotated for a set of habitat-specific core features, such as demographics, disease status, and medication for humans; host species and captivity status for animals; and filter sizes and salinity for marine samples. Additionally, all original metadata is provided in tabular form, simplifying focused studies e.g. into nutrient concentrations. Pre-computed taxonomic profiles facilitate rapid data exploration, while links to the SPIRE database enable genome-based analyses. The database is freely available for browsing and download at https://metalog.embl.de/.

Graphical abstract



Introduction

Metagenomic sequencing has transformed research in microbiology [1–3]. It has made it possible to elucidate the taxonomic profiles and functional potential of microbial communities across the globe [4, 5] and has driven numerous discoveries, e.g. in tracing the microbiome changes upon disease development and drug treatment in type 2 diabetes [6, 7]. Metagenomic sequencing data is useful beyond the scope of the original studies for which it is generated: it can be repurposed in the context of meta-studies, to extract metagenome-assembled genomes [8–10], to create gene catalogs [11], to identify disease-associated species across cohorts [12], and for countless other applications. Combining data across studies increases statistical power and yields more robust findings by covering more diverse populations. However, there are complex interactions between the microbiome and its environment

Received: August 14, 2025. **Revised:** October 2, 2025. **Accepted:** October 8, 2025

Table 1. Non-exhaustive overview of metadata databases

Name	Last update	Number of samples	Manual annotation	Focus
Amplicon data only				
Murine Microbiome Database [26]	2021	762	yes	Mice
Animal Microbiome Database [27]	2021	2530	yes	Animals
Human Microbiome Compendium [28]	2025	168 000	no	Human
Mixed				
mBodyMap [29]	2021	63 148	yes	Human
GMrepo [30]	2021	71 642	yes	Human
MicrobeAtlas [31]	2025	2 056 412	no	All habitats
Metagenomic data only				
AncientMetagenomeDir [32]	2025	2385	yes	Ancient samples
MarineMetagenomeDB [33]	2021	11 449	no	Marine
Meta2DB [34]	2024	13 897	yes	Human
TerrestrialMetagenomeDB [35]	2021	20 206	no	Terrestrial
curatedMetagenomicData [36]	2022	22 588	yes	Human
HumanMetagenomeDB [37]	2020	69 822	no	Human
Metalog (this study)	2025	120 169	yes	All habitats

that can confound associations [13]. It is therefore crucial for analyses to take these factors into account. Unfortunately, the necessary contextual data is often not readily available, is distributed over a variety of sources, and often follows heterogeneous annotation standards that need to be harmonized across studies [14, 15].

Most metagenomic sequencing data is deposited in databases that are part of the International Nucleotide Sequence Database Collaboration (INSDC) between the European Bioinformatics Institute (EMBL-EBI), the National Center for Biotechnology Information (NCBI), and the National Institute of Genetics (NGI): the European Nucleotide Archive (ENA), the Sequence Read Archive (SRA), and the DNA Data Bank of Japan (DDBJ) [16–19]. In the past, authors also used analysis services like MG-RAST to share data [20]. National repositories like the Chinese Genome Sequence Archive (GSA) are also increasing in size [21]. These repositories organize data into a hierarchy of data types: projects, biological samples, experiments, and runs. A biosample accession should uniquely identify a biological sample, like an aliquot of a fecal sample or material collected from a certain size fraction of seawater. From such a sample, multiple readouts may be experimentally prepared e.g. by DNA or RNA extraction. These are then sequenced in one or more runs of a sequencer. Sequencing databases require basic metadata on the experimental process, for example, by specifying the kind of library selection strategy (e.g. whole-genome versus amplicon), but not other important details such as the DNA extraction kits used.

At the sample level, further annotation standards are available, starting with the Minimum Information About a Metagenome or Environmental Sequence (MIMS) [22]. Ideally, complete metadata is directly available and linked to the metagenomic samples via the EMBL-EBI BioSamples [23] or the NCBI BioSample database [24]. More often, metadata needs to be extracted from a paper's text, figures, supplementary tables, or data repositories such as FigShare or Zenodo, and then linked to the deposited sequencing data. While a minimum set of metadata such as MIMS is required, sequencing repositories cannot perform further quality control on the uploaded data—for example, checking whether there is a match between the stated geographic location and the given coordinates. This results in the retention of erroneous annotations, such as switches between latitude and longitude and between longitudes east and west of the meridian (e.g. a sample from Italy may be shown as being located in the Atlantic Ocean). Other errors only become apparent when the metadata is carefully cross-checked, such as mismatches between the location reported in the paper and the individual samples. Both in coordinates and in other metadata columns, we observed the consequences of an inadvertent application of Microsoft Excel's convenience feature to automatically increment values when filling columns from a starting value, where the authors likely intended to replicate the same value but unintentionally introduced "drag-down" errors. Lastly, even if metadata is correctly deposited for one study, there still is a need for harmonization of variable labels and their contents to allow for integration across studies. For example, the type of birth has been given in fields such as "delivery_mode" or "delivery," with a number of synonyms for Caesarean section like "C-section," "Caesarean," and various misspellings of the word.

For the purpose of this paper, "metadata" refers to the contextual data concerning a biological sample, that is, the host-associated or environmental data. For example, this includes age, biological sex, health status, or body mass index (BMI) for human subjects; captivity status and species identifiers for animals; or water depth and filter sizes for ocean samples. In other contexts, metadata can also be regarded as technical data about an experiment—for example, on sample storage, DNA extraction protocols, and sequencers. The contextual data gathered here is considered as being metadata from a microbiological perspective, but is actually the primary data e.g. from the clinical perspective.

Several databases that link metagenomic sequencing data and metadata have been developed in the past, with different focuses for the covered sample types, sequencing approaches, and levels of manual annotation (Table 1). Amplicon sequencing data is more prevalent and allows for more large-scale analyses, but is limited in the taxonomic and functional resolution. Databases with a specialized focus may profit from more in-depth annotation. However, the data can then not be used for global analyses, e.g. to trace a species implicated in disease development like *Fusobacterium nucleatum* in colorectal cancer [25] to other habitats like animal hosts. Unfortunately, many of the databases have not been updated in several years, and there is no repository of manually annotated metadata that covers both host-associated and environmental data. To overcome these limitations, we introduce Metalog, a database of metadata for metagenomes across the globe with

120 169 samples (Fig. 1). We describe the principles in constructing the database and the curation and annotation work, the content of the database, and usage considerations along with a usage example.

Database construction

Data in Metalog is organized by study, each of which can contain any number of samples, potentially from different habitats. Each sample in Metalog corresponds to one defined biological sample. For human and animal samples, the subject or individual is also usually known, and multiple samples may be available for the same subject: different sample types (such as feces and saliva) for the same time point or the same sample type through longitudinal sampling. Whenever possible, we annotated both the subject identity and the time point. A given sample may have been processed in multiple ways (e.g. to extract DNA or RNA), and different sequencing experiments can be performed (e.g. amplicon or WGS sequencing). Metalog only contains samples with available metagenomic sequencing data; amplicon sequencing and metatranscriptomic data are excluded. If multiple metagenomic sequencing runs were available for a sample, they were pooled at the read level for the computation of derived data such as taxonomic profiles.

Samples can be broadly split into host-associated and environmental data. Within the host-associated samples, most are from human subjects, while among the environmental samples, marine samples are the most prevalent category. We therefore defined four templates that contain required and desired metadata items, namely for human samples, animal samples, ocean water samples, and other environmental samples. In the following, we describe the annotation process for Metalog.

Identification of relevant studies and publications

As metadata annotation and quality control are labor-intensive, Metalog cannot cover all published metagenomic datasets. We identified relevant studies from the literature and from the global set of microbiomes within the SPIRE database [10], focusing on data from Illumina and compatible platforms, which represent the bulk of the available sequencing data. We expect that long-read sequencing data will become more prevalent and will be added to Metalog in the future. If necessary, we manually associated publications and project accessions. Studies were assigned a human-readable code consisting of the last name of the first author, the publication year of the preprint or paper, and a short, tag-like description. For large datasets and consortia such as the Human Microbiome Project (HMP) or TARA, the project name is used instead of any individual publication. If no associated paper could be found, the INSDC database project accession number with a human-readable descriptive tag is used instead. While most studies correspond to one INSDC project accession, some uploaders—such as the Joint Genome Institute—create a new project accession for each sample. In this case, a Metalog study encompasses many INSDC project accessions.

Annotation of samples

We matched samples between the uploaded sequencing data and all available metadata using the provided sample identifiers or other indications such as descriptions or names of sequencing libraries. We manually checked for and resolved possible errors in INSDC submissions, such as amplicon data erroneously labeled as shotgun metagenome, individual samples erroneously submitted as distinct runs under a common biosample accession (i.e. the runs need to be treated as separate biosamples), or biosamples erroneously split into multiple sample accessions (i.e. runs from these samples need to be combined at the read level). In this way, each sample in Metalog corresponds to exactly one biological sample, but may not necessarily correspond to exactly one INSDC biosample accession.

Whenever possible, we mapped metadata to standardized vocabularies, such as the Environment Ontology (ENVO) and the Uber-anatomy Ontology (UBERON) for habitats and sample materials [38, 39]. However, ontologies are constantly evolving and have varying coverage of the full breadth of habitats around the world, and can therefore not always provide exactly matching terms [39]. As ontologies change, they may set widely used terms as obsolete without always offering clear alternatives. For example, the ENVO term ENVO:00009003 ("human-associated habitat") has been made obsolete, even though it has been used to denote the habitat for the majority of human metagenomic samples in the INSDC.

Extraction and annotation of tabular metadata

We extracted tabular metadata from the biosamples databases, from supplementary tables or relevant data repositories, and if necessary also added information contained in the papers' text, tables, or figures. We manually checked for and resolved inconsistencies and harmonized selected attributes as detailed below. Each of the four templates (human, animal, ocean water, and other environmental data) has a set of defined attributes that are relevant for the vast majority of studies, and the relevant field names used by data uploaders are mapped onto these attributes. In addition, we noticed that certain attributes occur across papers, such as physicochemical parameters in marine samples or the gestational age in weeks for studies on infants. We harmonized these field names across studies whenever feasible.

In-depth annotation of human samples

In the case of human samples, each deposited sample should be linked to a single subject who participated in the study. Thus, each sample is associated with a unique subject identifier; a few exceptional studies with pooled samples from multiple individuals were excluded from Metalog. This is important in cases of longitudinal sampling, where one subject may have multiple deposited samples. Time points (in days) were therefore calculated for each subject based on the collection dates of the samples or additional information provided in the publication. The first available sample was always set to time point 0. If a study does not specify exact sampling intervals but rather a range, we set the time point to a reasonable approximation (e.g. if there is a follow-up sample taken after 3–5 weeks, the time point would be at 28 days). Some studies contain follow-up samples, but do not provide the necessary information to link subjects or define the time point. In these cases, the time point field was left empty to indicate that there is no reliable information.

Demographic variables such as age, sex, and BMI were also captured in the template after standardization. If the age of infants is given in days or weeks, we converted those values

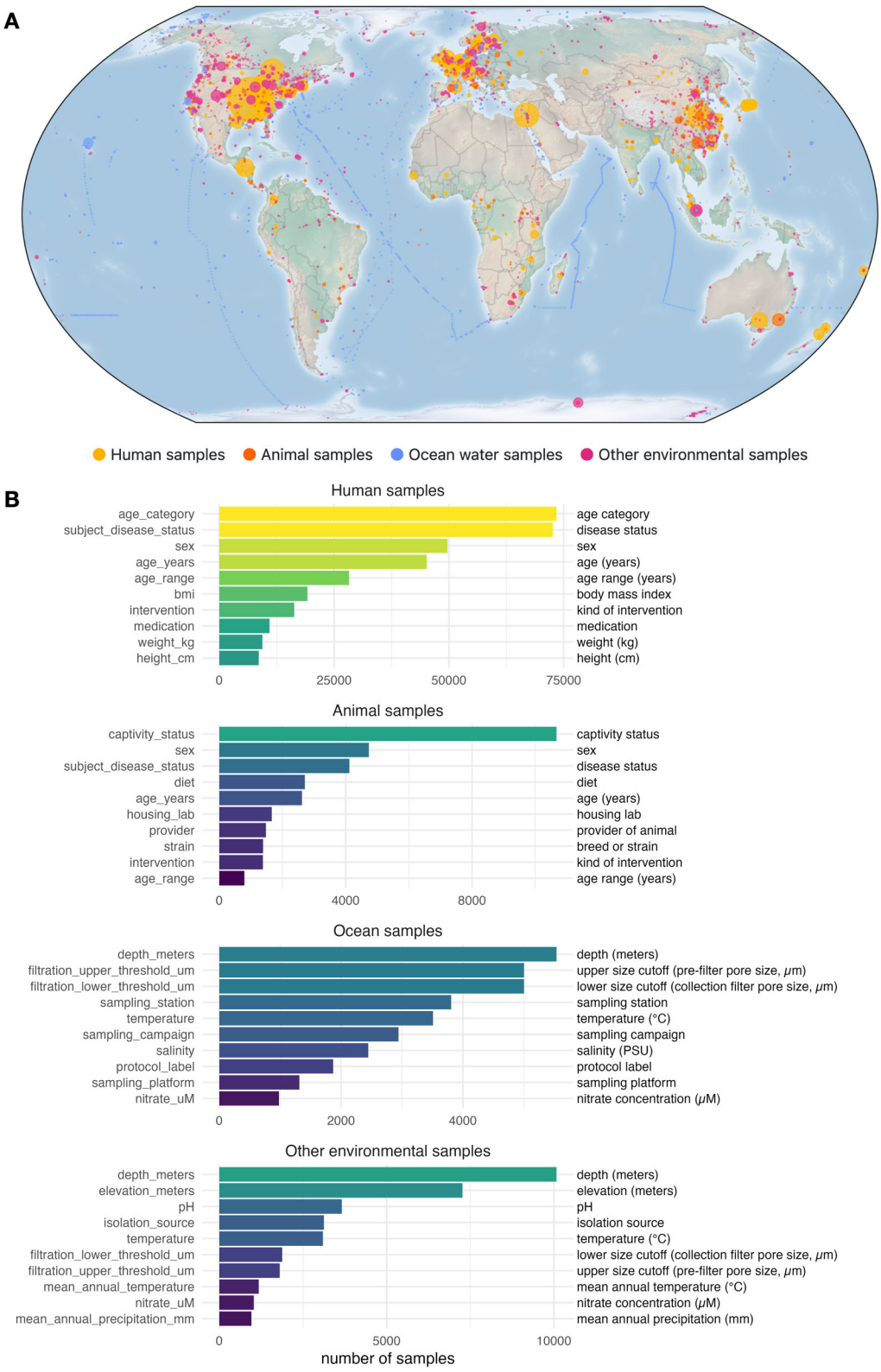


Figure 1. Overview of the content on Metalog. (**A**) A map of the samples contained in Metalog, colored by sample type. Circles are scaled by the number of available samples with the same location. The large yellow circles mostly correspond to country-level associations for human samples, while the chains of blue circles illustrate the course of sampling expeditions. (**B**) Number of available metadata annotations. For each of the four sample categories, the count of the top 10 fields is shown (removing trivial fields such as sample accession, location, or publication identifiers that are known for almost all samples). Left side: column names used in the download files. Right side: human-readable field names shown on the website. Bars are colored according to the number of samples with the respective annotation.

into years. When the exact age is not given, an age range was inferred from the methods section of a publication whenever possible, e.g. labeling subjects as "infants" or within a given range of years.

As several factors impact the human gut microbiome, such as antibiotics or other medication taken, administered FMT, or diet change, capturing them is important. We included these interventions whenever possible. On the broadest level, Metalog offers an "intervention" field that captures such broad categories. Medication metadata in studies may use a variety of synonyms and abbreviations or refer to groups of drugs. To address this, we harmonized the given medication information. Individual drugs are mapped to the ChEMBL database [40], and groups of drugs to the Anatomical Therapeutic Chemical (ATC) classification system (field name: "medication"). The download files also contain an additional automatically generated field, "medication_with_parents," that contains all matching drug classes for the individual medication annotations. For antibiotics, we also annotated the time since the last course of antibiotic treatment, if this information was given by the paper. If diet information was given, we mapped it to four broad categories each for babies (breastfed; breastfed and solids; formula- and breastfed; formula-fed) and everyone else (omnivore; pescetarian; vegan; vegetarian) to enable comparisons across studies. If detailed diet information was available, it is either kept as separate study-specific columns or in the "diet_full" field.

A similar challenge is the annotation of diseases. While some publications provide only a broad classification of patients, for others it can be very detailed, including comorbidities. We annotated a field "subject_disease_status" whenever possible. When a very detailed status was given, we summarized this to a more general term to allow for comparisons across studies. The full disease status was kept in an extra field ("subject_disease_status_full"). In addition, three further categories were identified: cohort members, healthy controls, and control patients. Cohort studies contain individuals drawn from a general population, most of whom will be healthy, but some of them can also have diseases that are not diagnosed or reported in the study. Case–control studies usually feature healthy subjects as controls. However, a few studies also contain control patients, whose actual underlying disease has not been annotated. For data analyses that focus on healthy subjects, we suggest combining data from cohort studies and controls. To investigate the effect of diseases, case–control studies should be used, taking either the controls or control patients into account.

For a small number of studies, inconsistencies in the sample naming and metadata indicated that there might be a wrong assignment of subject identifiers. If these could not be corrected from the given contextual data alone, we clustered samples based on Mash distances between the metagenomic reads [41] to manually identify samples belonging to the same individual based on the clustering pattern, relying on the fact that samples from the same individual generally have lower distances than samples from different individuals [42]. If this was not possible (e.g. in the case of fecal microbiota transplants), we excluded the samples in question.

In-depth annotation of other sample types

For animal-associated data, we also inferred the subjects and time points wherever possible. There are also studies where material from multiple animals is pooled into one sample, and this information is available as an extra field. The host organism is identified with the NCBI Taxonomy identifier [43], usually at species level but in some cases at higher taxonomic levels as appropriate. Environmental datasets often contain measurements of pH, salinity, and other parameters, which we mapped to common column names. The deposited metadata also often contains placeholder values such as "9999" for missing columns, which we removed to indicate the missing measurement.

Generation of associated microbial data

Sequencing data was collected and processed as described previously [10]. Taxonomic profiles were generated based on mOTUs version 3.0 [44], SPIRE-mOTUs [10], and MetaPhlAn 4 [45]. In addition, predicted enterotypes [46] and fecal microbial loads [47] were computed for fecal samples from adults.

Database content

The Metalog database can be accessed at https://metalog.embl.de/. The website allows users to browse the samples by selecting metadata fields of interest, by searching for field names or metadata values, by zooming into interactive maps, and by selecting samples with certain medication, diseases, or habitat from hierarchical trees. Users can download the whole database or individual studies in different formats. While the website makes it possible to select a number of combinations of metadata (e.g. to select gut microbiome data for women from the USA who suffer from colorectal cancer), more complex analysis and filtering tasks should be done by downloading the metadata and analyzing it in statistical analysis software or a programming language. Studies and samples contain links to the original publications, sequencing databases, and the SPIRE database. Downloadable metadata files contain the identifier used by the SPIRE database. This makes it possible to link the annotated metadata from Metalog to functional data in SPIRE, such as antibiotic resistance and protein function annotations [10]. Metalog is continually updated, and the internal development database is synchronized with the public website every weekend. All download files contain the date of the last website update in the filename. Additionally, all samples contain a timestamp of their last update in the database and the downloadable metadata files.

Metalog currently contains metadata for 73 082 human samples, 10 703 animal samples, 5146 ocean water samples, and 21 802 samples from other environmental habitats such as soil, sediment, or fresh water. These samples are distributed all over the world (Fig. 1A), although the bias in the availability of samples toward Western and East Asian countries is apparent. Some metadata attributes are widely available, e.g. age category for humans (91%), captivity status for animals (99%), and water depth for ocean water samples (100%, Fig. 1B). Combinations of metadata items are available with lower counts, e.g. a study investigating the microbiome influenced by both exact age and BMI could make use of 17 324 samples.

In addition to the metadata, taxonomic profiles with matching sample identifiers are available for download based on mOTUs version 3.0 [44], SPIRE-mOTUs [10], and MetaPhlAn 4 [45]. While the mOTUs profiler provides data for bacteria and archaea, MetaPhlAn also yields profiles for fungi and

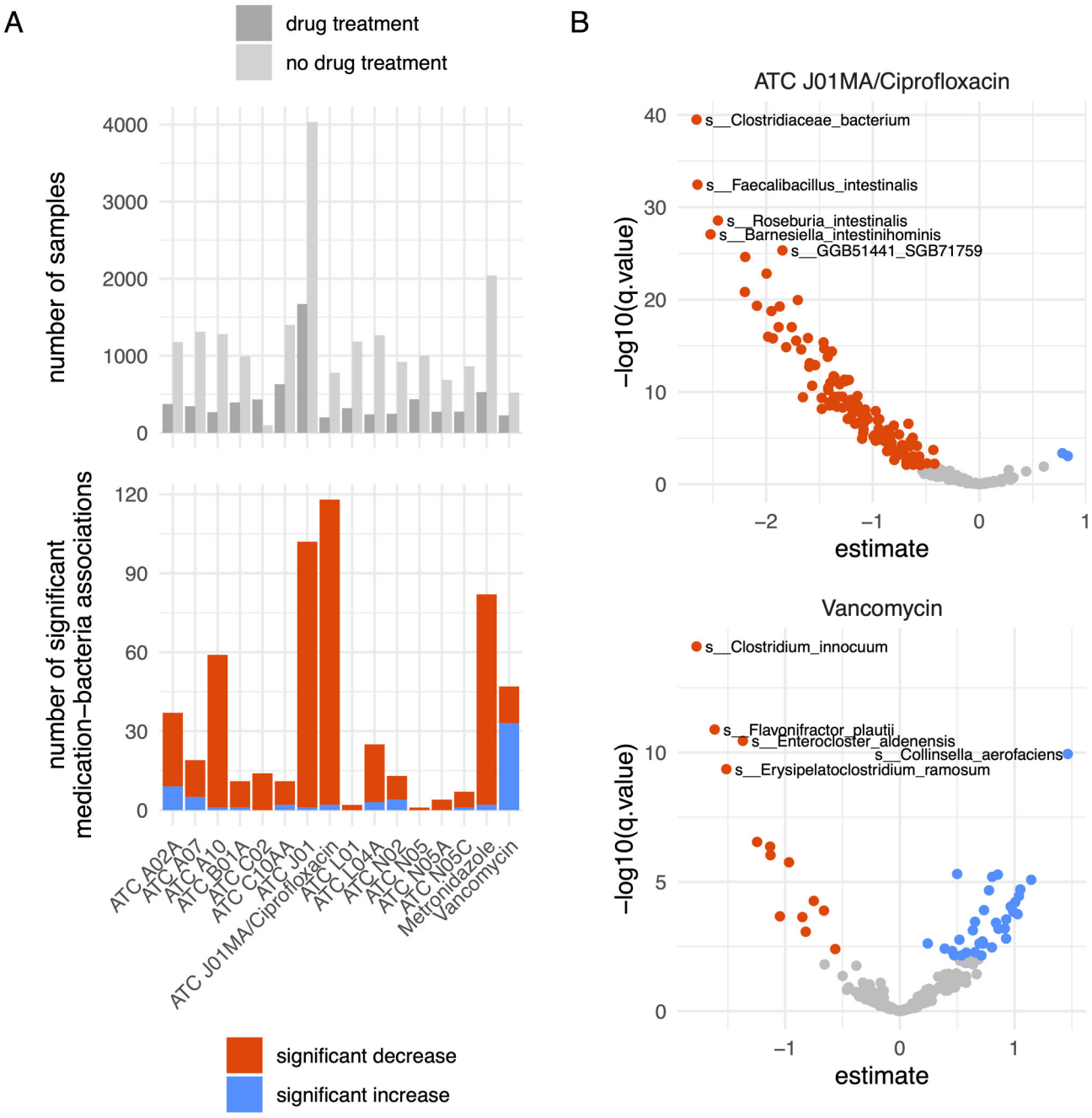


Figure 2. A usage example for Metalog, showing associations between gut bacteria and medication. (**A**) Top: Number of samples per medication for which the drug was given and the number of samples from the same studies without drug treatment. Bottom: Number of significant associations (*q*-value cutoff: 0.01). (**B**) Volcano plots for the two drug classes with the highest number of positive and negative associations.

other eukaryotes. Predicted enterotypes [46] and fecal microbial loads [47] are available for fecal samples from adults.

Usage considerations

Samples with artificial perturbations

A number of studies include negative controls or mock communities for quality control. These are included in the downloadable data and will need to be filtered out for most applications. Studies may also employ cell sorting, *in vitro* cultivation, macrocosm experiments, or spike-in of certain species for specific questions. While these samples are still valuable for meta-analyses that focus on the genomic content, they should be removed for analyses that focus on relative abundances as these experimental treatments may perturb species abundances (field name: "artificial"). We also flagged post-mortem and paleo-samples in this field.

Unreported metadata

The metadata that is available varies greatly between studies. Many papers on clinical findings report only the disease status of the subjects, but not, for example, medication usage. This is often the case even when the authors show in their analysis that drug treatment or other factors play an important role in

shaping the microbiome and the reported biomarkers. A meta-analysis of drug treatment could therefore be restricted to include studies that consist of only healthy (and mostly unmedicated) subjects and of clinical studies that report at least some medication information, so that one can assume that subjects without a reported medication are indeed treatment-free.

Geographic locations

Environmental samples usually have exact geographical coordinates associated with them. For human samples, location information is often only available at the country level, but sometimes also a city, region, or clinical center where the samples have been taken is known. We have annotated this information whenever possible (field "location_resolution"). For country-level data, we set the location to a standard set of coordinates that are close to the center of the country, weighted by regional population data [48]. In this way, all country-level datasets are combined in the map display, and the coordinates are close to the population centers (e.g. being in the south of Canada instead of close to the Arctic Circle). When taking the geographic location of a sample into account, it is important to consider the resolution of the location that has been annotated. A calculation of the distance between two samples will necessarily involve some uncertainty unless the exact coordinates are known for both samples.

Usage example

To illustrate the ease of use and combined strength of metadata and taxonomic profiles, we provide an example use case on the website that also serves as a tutorial in accessing the data. For all human samples, we selected fecal samples from adults with available disease status. We extracted the information on medication taken by the subjects and clustered the medication records to collapse redundant data. In some cases, medication and disease status were completely confounded and therefore excluded (e.g. all subjects in the current Metalog version with helminthiasis had been treated with albendazole). For each medication, we identified all studies in which the medication was administered. We then selected samples for which either the medication under consideration or no medication was given (Fig. 2A, top). We then computed linear models between log-transformed bacterial abundance and medication, taking the study and the disease status into account. Filtering at a false discovery rate threshold of 0.01, we found 16 medications that showed significant associations between bacteria and drug treatment. Fluoroquinolones (ATC code J01MA, like ciprofloxacin) had the highest number of associations (Fig. 2). Interestingly, the antibiotic vancomycin had the highest number of significant positive associations. The analysis presented here only takes study effect and disease status into account. More focused studies would also focus on demographic factors and take other confounding variables into account, like co-treatment with multiple drugs [49].

Discussion

Metalog makes it possible for researchers to integrate metagenomics data and metadata from 919 studies across the globe (and even the International Space Station). It increases the discoverability of studies and provides links between the deposited data and the underlying research papers, encouraging proper citation of the primary data sources. Parts of the metadata collected for Metalog have already been used in previous publications—for example, to ascertain the disease specificity for a biomarker panel for pancreatic duct carcinoma [50], to investigate the associations between fecal microbial load and various host-associated features [47], to trace strain dynamics after FMT [51], and to investigate the prevalence of *C. difficile* across different age groups and environments [52]. In addition to our initial intention to establish Metalog as a resource for microbiome research, we anticipate that it may also be useful for the development of Artificial Intelligence (AI)-assisted annotation systems for the extraction of sample-level metadata. A first such approach extracts information about the ecological environment [53], which could also be cross-checked with sequence-based habitat predictions [54]. Such an AI-based system would need to pull together tabular data from a variety of sources, be able to infer common identifiers, have access to a target vocabulary for given fields, and be able to flag conflicts between different sources of information.

We encourage researchers across all areas of microbiology to make the gathered contextual data as readily available as possible, ideally by adding it as metadata along with the biosamples when uploading sequence data to a repository. This facilitates direct reuse by other researchers and makes it easier for the dataset to be added to integrated databases like Metalog. Crucial information like disease status, medication, time interval between samples, and basic demographics should be reported whenever possible, even if not all available metadata can be shared due to privacy concerns.

Acknowledgements

Author contributions: Michael Kuhn (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Methodology [equal], Software [equal], Supervision [equal], Visualization [equal], Writing—original draft [lead], Writing—review & editing [equal]), Thomas Sebastian B. Schmidt (Conceptualization [equal], Data curation [equal], Methodology [equal], Writing—review & editing [equal]), Pamela Ferretti (Conceptualization [equal], Data curation [equal], Writing—review & editing [equal]), Anna Głazek (Data curation [equal], Methodology [equal], Software [equal]), Shahriyar Mahdi Robbani (Software [equal]), Wasiu Akanni (Software [equal]), Anthony Fullam (Software [equal]), Christian Schudoma (Software [equal]), Ela Cetin (Data curation [equal]), Mariam Hassan (Data curation [equal], Writing—original draft [supporting]), Kasimir Noack (Data curation [equal]), Anna Schwarz (Data curation [equal]), Roman Thielemann (Data curation [equal]), Leonie Thomas (Data curation [equal]), Moritz von Stetten (Data curation [equal]), Renato Alves (Conceptualization [equal], Methodology [equal], Writing—review & editing [equal]), Anandhi Iyappan (Methodology [supporting]), Ece Kartal (Data curation [equal], Writing—review & editing [equal]), Ivan Kel (Data curation [equal]), Marisa I. Keller (Data curation [equal]), Oleksandr Maistrenko (Data curation [equal]), Anna Mankowski (Data curation [equal]), Suguru Nishijima (Data curation [equal]), Daniel Podlesny (Data curation [equal]), Jonas Schiller (Data curation [equal], Writing—review & editing [equal]), Sarah Schulz (Methodology [supporting]), Thea Van Rossum (Data curation [equal]), and Peer Bork (Conceptualization [equal], Funding acquisition [equal], Supervision [equal], Writing—review & editing [equal]).

Conflict of interest

None declared.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement numbers 668031 (GALAXY); the European Research Council under grant agreement number ERC-AdG-669830 (MicrobioS); the European Union's Horizon Europe research and innovation programme under grant agreement number 101059915 (BIOcean5D); the Challenge Grant "MicrobLiver" grant number NNF15OC0016692 from the Novo Nordisk Foundation; the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 460129525 (NFDI4Microbiota); the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI; the German Federal Ministry of Research, Technology and Space in the frame of de.NBI & ELIXIR-DE (W-de.NBI-014); the European Molecular Biology Laboratory (EMBL) and in particular through EMBL Planetary Biology Transversal Theme's seed grant awarded to Peer Bork. Funding to pay the Open Access publication charges for this article was provided by European Molecular Biology Laboratory (EMBL).

Data availability

All data are freely accessible at https://metalog.embl.de/ under the Open Database License.

References

1. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol* 2010;6:e1000667. https://doi.org/10.1371/journal.pcbi.1000667
2. Kim N, Ma J, Kim W *et al.* Genome-resolved metagenomics: a game changer for microbiome medicine. *Exp Mol Med* 2024;56:1501–12. https://doi.org/10.1038/s12276-024-01262-7
3. Pinto Y, Bhatt AS. Sequencing-based analysis of microbiomes. *Nat Rev Genet* 2024;25:829–45. https://doi.org/10.1038/s41576-024-00746-6
4. Schloissnig S, Arumugam M, Sunagawa S *et al.* Genomic variation landscape of the human gut microbiome. *Nature* 2013;493:45–50. https://doi.org/10.1038/nature11711
5. Sunagawa S, Coelho LP, Chaffron S *et al.* Structure and function of the global ocean microbiome. *Science* 2015;348:1261359. https://doi.org/10.1126/science.1261359
6. Wu H, Tremaroli V, Schmidt C *et al.* The gut microbiota in prediabetes and diabetes: a population-based cross-sectional study. *Cell Metabolism* 2020;32:379–90. https://doi.org/10.1016/j.cmet.2020.06.011
7. MetaHIT consortium, Forslund K, Hildebrand F, Nielsen T *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* 2015;528:262–6. https://doi.org/10.1038/nature15766
8. Almeida A, Nayfach S, Boland M *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* 2021;39:105–14. https://doi.org/10.1038/s41587-020-0603-3
9. Pavlopoulos GA, Baltoumas FA, Liu S *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* 2023;622:594–602. https://doi.org/10.1038/s41586-023-06583-7
10. Schmidt TSB, Fullam A, Ferretti P *et al.* SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. *Nucleic Acids Res* 2024;52:D777–83. https://doi.org/10.1093/nar/gkad943
11. Coelho LP, Alves R, del Río ÁR *et al.* Towards the biogeography of prokaryotic genes. *Nature* 2022;601:252–6. https://doi.org/10.1038/s41586-021-04233-4
12. Piccinno G, Thompson KN, Manghi P *et al.* Pooled analysis of 3,741 stool metagenomes from 18 cohorts for cross-stage and strain-level reproducible microbial biomarkers of colorectal cancer. *Nat Med* 2025;31:2416–29. https://doi.org/10.1038/s41591-025-03693-9
13. Schmidt TSB, Raes J, Bork P. The human gut microbiome: from association to modulation. *Cell* 2018;172:1198–215. https://doi.org/10.1016/j.cell.2018.02.044
14. Hassenrück C, Poprick T, Helfer V *et al.* FAIR enough? A perspective on the status of nucleotide sequence data and metadata on public archives. bioRxiv, https://doi.org/10.1101/2021.09.23.461561, 24 September 2021, preprint: not peer reviewed.
15. Kim L, Lavrinienko A, Sebechlebska Z *et al.* Tier-based standards for FAIR sequence data and metadata sharing in microbiome research. bioRxiv, https://doi.org/10.1101/2025.02.06.636914, 8 February 2025, preprint: not peer reviewed.
16. Karsch-Mizrachi I, Arita M, Burdett T *et al.* The international nucleotide sequence database collaboration (INSDC): enhancing global participation. *Nucleic Acids Res* 2025;53:D62–6.
17. O'Cathail C, Ahamed A, Burgin J *et al.* The European Nucleotide Archive in 2024. *Nucleic Acids Res* 2025;53:D49–55. https://doi.org/10.1093/nar/gkae975
18. Sayers EW, Beck J, Bolton EE *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res* 2025;53:D20–9. https://doi.org/10.1093/nar/gkae979
19. Tanizawa Y, Fujisawa T, Kodama Y *et al.* DNA Data Bank of Japan (DDBJ) update report 2022. *Nucleic Acids Res* 2023;51:D101–5. https://doi.org/10.1093/nar/gkac1083
20. Meyer F, Paarmann D, D'Souza M *et al.* The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 2008;9:386. https://doi.org/10.1186/1471-2105-9-386
21. CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2025. *Nucleic Acids Res* 2025;53:D30–44. https://doi.org/10.1093/nar/gkae978
22. Kottmann R, Gray T, Murphy S *et al.* and Genomic Standards Consortium. A standard MIGS/MIMS compliant XML schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* 2008;12:115–21. https://doi.org/10.1089/omi.2008.0A10
23. Courtot M, Gupta D, Liyanage I *et al.* BioSamples database: FAIRer samples metadata to accelerate research data management. *Nucleic Acids Res* 2022;50:D1500–7. https://doi.org/10.1093/nar/gkab1046
24. Barrett T, Clark K, Gevorgyan R *et al.* BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research* 2012;40:D57–63. https://doi.org/10.1093/nar/gkr1163
25. Zepeda-Rivera M, Minot SS, Bouzek H *et al.* A distinct *Fusobacterium nucleatum* clade dominates the colorectal cancer niche. *Nature* 2024;628:424–32. https://doi.org/10.1038/s41586-024-07182-w
26. Yang J, Park J, Park S *et al.* Introducing murine microbiome database (MMDB): a curated database with taxonomic profiling of the healthy mouse gastrointestinal microbiome. *Microorganisms* 2019;7:480. https://doi.org/10.3390/microorganisms7110480
27. Yang J, Park J, Jung Y *et al.* AMDB: a database of animal gut microbial communities with manually curated metadata. *Nucleic Acids Res* 2022;50:D729–35. https://doi.org/10.1093/nar/gkab1009

28. Abdill RJ, Graham SP, Rubinetti V *et al.* Integration of 168,000 samples reveals global patterns of the human gut microbiome. *Cell* 2025;188:1100–18.e17. https://doi.org/10.1016/j.cell.2024.12.017
29. Jin H, Hu G, Sun C *et al.* mBodyMap: a curated database for microbes across human body and their associations with health and diseases. *Nucleic Acids Res* 2022;50:D808–16. https://doi.org/10.1093/nar/gkab973
30. Dai D, Zhu J, Sun C *et al.* GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Res* 2022;50:D777–84. https://doi.org/10.1093/nar/gkab1019
31. Rodrigues JFM, Tackmann J, Malfertheiner L *et al.* The MicrobeAtlas database: global trends and insights into Earth's microbial ecosystems, bioRxiv, https://doi.org/10.1101/2025.07.18.665519, 18 July 2025, preprint: not peer reviewed.
32. Fellows Yates JA, Andrades Valtueña A, Vågene ÅJ *et al.* Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir. *Sci Data* 2021;8:31. https://doi.org/10.1038/s41597-021-00816-y
33. Nata'ala MK, Avila Santos AP, Coelho Kasmanas J *et al.* MarineMetagenomeDB: a public repository for curated and standardized metadata for marine metagenomes. *Environ Microbiome* 2022;17:57.
34. Kok CR, Mulakken NJ, Thissen JB *et al.* Meta2DB: curated shotgun metagenomic feature sets and metadata for health State prediction, bioRxiv, https://doi.org/10.1101/2024.10.03.616398, 12 December 2024, preprint: not peer reviewed.
35. Corrêa FB, Saraiva JP, Stadler PF *et al.* TerrestrialMetagenomeDB: a public repository of curated and standardized metadata for terrestrial metagenomes. *Nucleic Acids Res* 2019;48:D626–D32. https://doi.org/10.1093/nar/gkz994
36. Pasolli E, Schiffer L, Manghi P *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* 2017;14:1023–4. https://doi.org/10.1038/nmeth.4468
37. Kasmanas JC, Bartholomäus A, Corrêa FB *et al.* HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. *Nucleic Acids Res* 2021;49:D743–50. https://doi.org/10.1093/nar/gkaa1031
38. Mungall CJ, Torniai C, Gkoutos GV *et al.* Uberon, an integrative multi-species anatomy ontology. *Genome Biol* 2012;13:R5. https://doi.org/10.1186/gb-2012-13-1-r5
39. Buttigieg PL, Morrison N, Smith B *et al.* The environment ontology: contextualising biological and biomedical entities. *J Biomed Sem* 2013;4:43. https://doi.org/10.1186/2041-1480-4-43
40. Zdrazil B, Felix E, Hunter F *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res* 2024;52:D1180–92. https://doi.org/10.1093/nar/gkad1004
41. Ondov BD, Treangen TJ, Melsted P *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132. https://doi.org/10.1186/s13059-016-0997-x
42. Becker A, Schmartz GP, Gröger L *et al.* Effects of resistant starch on symptoms, fecal markers, and gut microbiota in Parkinson's Disease—The RESISTA-PD trial. *Genom Proteom Bioinform* 2022;20:274–87. https://doi.org/10.1016/j.gpb.2021.08.009
43. Cox E, Tsuchiya MTN, Ciufo S *et al.* NCBI Taxonomy: enhanced access via NCBI datasets. *Nucleic Acids Res* 2025;53:D1711–5. https://doi.org/10.1093/nar/gkae967
44. Ruscheweyh H-J, Milanese A, Paoli L *et al.* Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome* 2022;10:212. https://doi.org/10.1186/s40168-022-01410-z
45. Blanco-Míguez A, Beghini F, Cumbo F *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* 2023;41:1633–44. https://doi.org/10.1038/s41587-023-01688-w
46. Keller MI, Nishijima S, Podlesny D *et al.* Refined enterotyping reveals dysbiosis in global fecal metagenomes, bioRxiv, https://doi.org/10.1101/2024.08.13.607711, 13 August 2024, preprint: not peer reviewed.
47. Nishijima S, Stankevic E, Aasmets O *et al.* Fecal microbial load is a major determinant of gut microbiome variation and a confounder for disease associations. *Cell* 2025;188:222–36.e15. https://doi.org/10.1016/j.cell.2024.10.022
48. Center For International Earth Science Information Network-CIESIN-Columbia University, United Nations Food And Agriculture Programme-FAO, and Centro Internacional De Agricultura Tropical-CIAT. Gridded Population of the World, Version 3 (GPWv3): population count grid. https://data.nasa.gov/dataset/gridded-population-of-the-world-version-3-gpwv3-population-count-grid 2005. (17 May 2024, date last accessed).
49. Forslund SK, Chakaroun R, Zimmermann-Kogadeeva M *et al.* Combinatorial, additive and dose-dependent drug-microbiome associations. *Nature* 2021;600:500–5. https://doi.org/10.1038/s41586-021-04177-9
50. Kartal E, Schmidt TSB, Molina-Montes E *et al.* A faecal microbiota signature with high specificity for pancreatic cancer. *Gut* 2022;71:1359–72. https://doi.org/10.1136/gutjnl-2021-324755
51. Schmidt TSB, Li SS, Maistrenko OM *et al.* Drivers and determinants of strain dynamics following fecal microbiota transplantation. *Nat Med* 2022;28:1902–12. https://doi.org/10.1038/s41591-022-01913-0
52. Ferretti P, Wirbel J, Maistrenko OM *et al. C. difficile* may be overdiagnosed in adults and is a prevalent commensal in infants. 2023;12:RP90111. https://doi.org/10.7554/eLife.90111.1
53. Gaio D, Tackmann J, Perez-Molphe-Montoya E *et al.* Enhanced semantic classification of microbiome sample origins using Large Language Models (LLMs), bioRxiv, https://doi.org/10.1101/2025.04.24.650461, 27 April 2025, preprint: not peer reviewed.
54. Kim CY, Podlesny D, Schiller J *et al.* Planetary microbiome structure and generalist-driven gene flow across disparate habitats, bioRxiv, https://doi.org/10.1101/2025.07.18.664989, 18 July 2025, preprint: not peer reviewed.

Received: August 14, 2025. **Revised:** October 2, 2025. **Accepted:** October 8, 2025