Animal Microbiome

SOFTWARE Open Access

The AnimalAssociatedMetagenomeDB reveals a bias towards livestock and developed countries and blind spots in functional-potential studies of animal-associated microbiomes

Anderson Paulo Avila Santos[1,3], Muhammad Kabiru Nata'ala[1,2], Jonas Coelho Kasmanas[1,2,3], Alexander Bartholomäus[4], Tina Keller-Costa[6], Stephanie D. Jurburg[1,10], Tamara Tal[7], Amélia Camarinha-Silva[8,9], João Pedro Saraiva[1], André Carlos Ponce de Leon Ferreira de Carvalho[3], Peter F. Stadler[2,11,12,13,14,15,16], Danilo Sipoli Sanches[5] and Ulisses Rocha[1*]

Abstract

Background Metagenomic data can shed light on animal-microbiome relationships and the functional potential of these communities. Over the past years, the generation of metagenomics data has increased exponentially, and so has the availability and reusability of data present in public repositories. However, identifying which datasets and associated metadata are available is not straightforward. We created the Animal-Associated Metagenome Metadata Database (AnimalAssociatedMetagenomeDB - AAMDB) to facilitate the identification and reuse of publicly available non-human, animal-associated metagenomic data, and metadata. Further, we used the AAMDB to (i) annotate common and scientific names of the species; (ii) determine the fraction of vertebrates and invertebrates; (iii) study their biogeography; and (iv) specify whether the animals were wild, pets, livestock or used for medical research.

Results We manually selected metagenomes associated with non-human animals from SRA and MG-RAST. Next, we standardized and curated 51 metadata attributes (e.g., host, compartment, geographic coordinates, and country). The AAMDB version 1.0 contains 10,885 metagenomes associated with 165 different species from 65 different countries. From the collected metagenomes, 51.1% were recovered from animals associated with medical research or grown for human consumption (i.e., mice, rats, cattle, pigs, and poultry). Further, we observed an over-representation of animals collected in temperate regions (89.2%) and a lower representation of samples from the polar zones, with only 11 samples in total. The most common genus among invertebrate animals was Trichocerca (rotifers).

Conclusion Our work may guide host species selection in novel animal-associated metagenome research, especially in biodiversity and conservation studies. The data available in our database will allow scientists to perform meta-

*Correspondence:
Ulisses Rocha
ulisses.rocha@ufz.de

Full list of author information is available at the end of the article

analyses and test new hypotheses (e.g., host-specificity, strain heterogeneity, and biogeography of animal-associated metagenomes), leveraging existing data. The AAMDB WebApp is a user-friendly interface that is publicly available at https://webapp.ufz.de/aamdb/.

Keywords Metagenome, Animal-Associated Microbiomes, Microbial Ecology, Metadata, Database, FAIR principles

Background

Metagenomics is an expanding field of study, and the number of metagenomes in public databases has grown exponentially [1]. While genomic studies consider a specific organism's genetic material, metagenomic studies consider the genetic material of entire communities of organisms [2, 3]. They have served to inform a wide range of fields, including Earth Sciences, Life Sciences, Biomedical Sciences, Bioenergy, Bioremediation, Biotechnology, Agriculture and Biodefense, and Microbial Forensics [4]. Metagenomic data is generally deposited in public sequence repositories, the largest of which is the Sequence Read Archive (SRA) [5] of the National Center for Biotechnology Institute (NCBI) [6], which is associated with the International Nucleotide Sequence Database Collaboration (INSDC) [7], a collaboration between the DNA Databank of Japan (DDBJ) [8], and the European Nucleotide Archive (ENA) [9]. Smaller repositories for metagenomic data include MG-RAST [10], IMG/M [11], MGnify [12], and gcMeta [13]. However, identifying relevant studies is not straightforward. As the number of publicly available metagenomic studies continues to grow, developing centralized resources and curated metadata is essential to improving the findability and reusability of these data. While several such resources exist for environmental [1, 14, 15] and human-associated metadata [16–19], animal-associated metagenomes have received less attention. Searching and reusing samples can be difficult due to incomplete or poor-quality metadata accompanying the metagenome data [14–16, 20–22].

Databases allow for the re-analysis of samples to test new hypotheses and make novel discoveries [23]. For instance, a study by Stewart and collaborators [24] assembled 4,941 rumen microbial metagenome-assembled genomes (MAGs). Reusing metagenomes in public repositories may lead to new genomic and protein resources, enabling a better understanding of the structure and functions of the (in this case rumen) microbiota. This highlights the importance of creating user-friendly metagenome repositories as the starting point for meta-studies of animal microbiomes. Nevertheless, mining metagenomes of interest from public databases is time-consuming and requires specialist knowledge in bioinformatics and microbiology, as metadata are not easily accessible for those with little experience in data science [25]. Some initiatives try to simplify this process, such as the Genomics Standard Consortium [26], the BioProject, and the BioSample project [27], which defined the minimum necessary information about a metagenomic sample to facilitate and provide better organization of metadata [28].

Nevertheless, tools are still needed to filter samples based on metadata in a user-friendly way. For example, the HumanMetagenomeDB contains standardized metadata of about 70,000 human metagenomes [16]. Further, the TerrestrialMetagenomeDB contains curated metadata of more than 20,000 terrestrial metagenomes [14], and the MarineMetagenomeDB over 11,449 marine metagenomes [15]. Recently, Hu and collaborators [20] developed an animal metagenome database that contains 10,672 publicly available metagenomes and 63,214 amplicon sequencing samples from animal-associated microbiomes and divides the data into domestic and wild animal categories. Users can download the metadata of interest according to filters but cannot download raw data or visualize the distribution of selected data on a world map. Furthermore, an analysis of data distribution across different species and regions is not available. Research into animal-associated metagenomes could be significantly accelerated by developing attributes tailored for non-human, animal-associated data and presenting a user-friendly catalogue of available data.

Improved data accessibility can shed light on current data collection biases. One study demonstrated a higher data collection from temperate zones and vertebrate animals in biodiversity studies [29]. Indeed, most research on biodiversity and microbiomes happens in countries with larger economies [30]. At the same time, up to 50% of all species on Earth may be native to 6–7% of the Earth's land area that is covered by tropical moist forests [31]. Many developing nations are located in the tropics, where the level of biological diversity is the highest, and the threats to its maintenance are the greatest [32]. Further, while it is estimated that 925,000 species of invertebrates [29] and 100,000 species of vertebrate animals [33] inhabit Earth, biodiversity research shows a general bias towards vertebrates [29]. Identifying gaps is essential, especially in the young and rapidly growing field of metagenomic research. Awareness alone may aid in curving these biases in research as they develop [29]. In addition, biodiversity loss and species extinction are two critical environmental problems. Microbiome-targeted interventions have been studied as potential options to reverse the deterioration of biodiversity [34].

To tackle the challenges of the non-standardization and ambiguity of the metadata, we developed the AAMDB, which has standardized and manually curated animal-associated metadata to help researchers quickly identify animal-associated metagenomes of interest through a user-friendly web interface. Further, we used this resource to evaluate bias in the distribution of animal-associated metagenomes, indicating which are the most studied and the under-represented species in public repositories.

Implementation

We constructed the AAMDB in three steps (Fig. 1): (1) metadata retrieval from the source databases (SRA and MG-RAST) and removal of human and non-animal metagenomes (Fig. 1A); (2) selection, parsing, and standardization of available metadata attributes (Fig. 1B); (3) identification of animal-associated metagenomes terms (Fig. 1C); (4) merging datasets (Fig. 1D); and (5) development and implementation of a user-friendly web application (Fig. 1E).

At different points of the implementation, we will use the expression 'manually curated'. Any manual process has risks and setbacks, such as human errors, biases, and limitations, so it cannot be considered inherently superior to an automatic process. In this study, manual curation was done by specialists in the field who identified issues in terminology, standardization and usage of different types of metadata. Once issues were identified, we generated automated or semi-automated scripts to address them, minimizing the risks and setbacks.

Metadata retrieval and removal of non-whole genome sequencing data

Metadata from SRA were retrieved from SRAdb using a list of sample identifiers (SRA run IDs) labeled as whole-genome sequencing from complex microbial communities (hereafter, metagenomic data) or amplicon sequencing was downloaded from PARTIE (Ref=Torres, Edwards, e McNair, "PARTIE") (https://github.com/linsalrob/partie). PARTIE is a Machine Learning model based on supervised and unsupervised classification, which classifies sequence data into metagenomic or amplicon sequence data sets. Sample identifiers labeled as WGS were extracted from the list, and metadata of WGS samples was retrieved using SRAdb R package. This package provides access to metadata of samples available in SRA. We retrieved quality scores and the creation date of the SRA libraries using SRA-Tinder (https://github.com/NCBI-Hackathons/SRA_Tinder) and Entrez Direct (https://www.ncbi.nlm.nih.gov/books/NBK179288) respectively. Finally, we recovered the PubMed and BioProject ID using the rentrez tool (https://github.com/ropensci/rentrez). Concurrently, all metadata from the

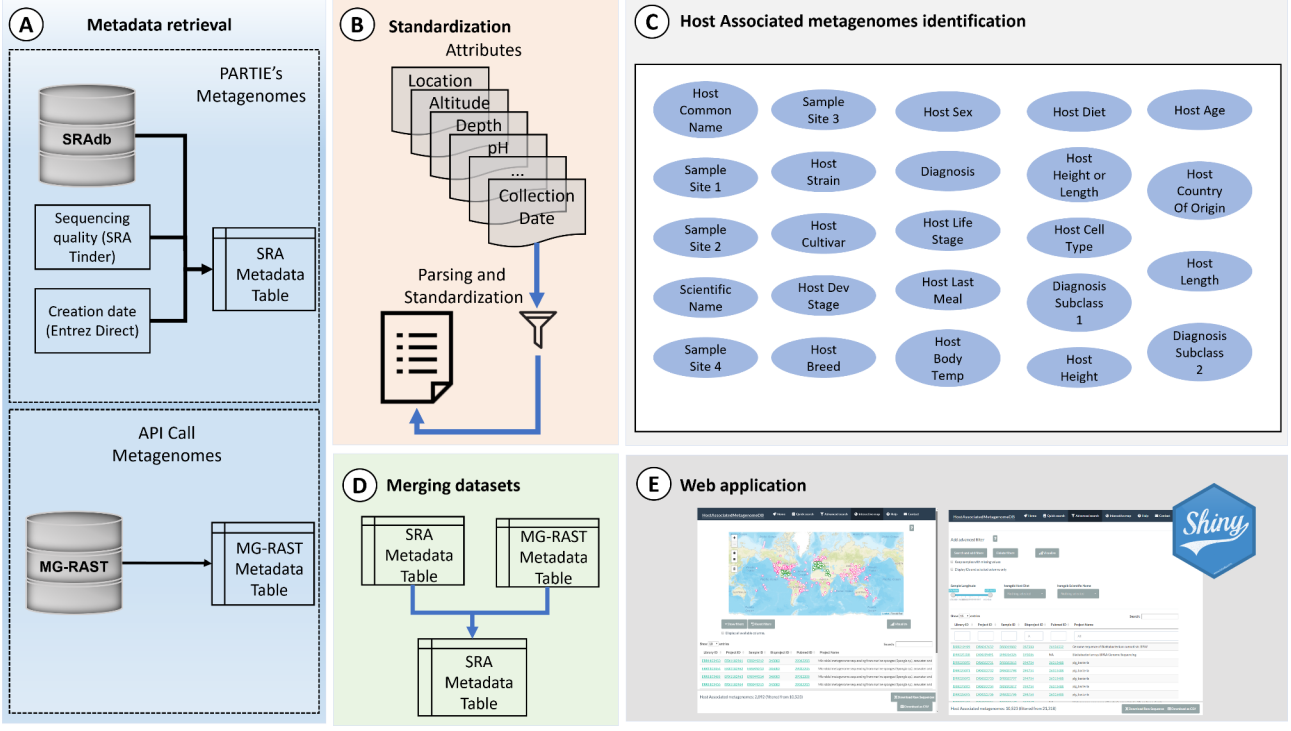

Fig. 1 Overview of the AAMDB construction workflow. (**A**) Metadata retrieval present in SRA and MG_RAST; (**B**) standardization of attributes; (**C**) identification of animal-associated metagenomes terms; (**D**) merging SRA and MG_RAST dataset; (**E**) The AAMDB was made available through a Shiny web application

MG-RAST repository were retrieved using their application program interface (API).

Once the metadata were downloaded, we removed non-whole genome sequencing (non-WGS) from SRA and MG-RAST. For SRA, the non-WGS samples were eliminated by removing samples that contained 'AMPLICON' or '*RNA*' in the 'library_strategy' category. We also removed samples labeled '*PCR*' in the 'library_selection' category. After, we removed all samples labeled with anything other than 'METAGENOMIC' or 'GENOMIC' as their 'library_source' and manually checked for entries belonging to single species genomes. We removed all samples marked with anything other than 'WGS' for MG-RAST as in the categories 'investigation_type' and 'seq_meth'.

Selection of non-human animal-associated metagenomes, parsing, and standardization of attributes

To select animal metagenomes, the column 'sample_attribute' was manually explored, and host names were extracted, and a dictionary of terms containing the host names and human-related terms was created (Supplementary Table S1). The columns 'center_project_name', 'sample_attribute', and 'study_title' were extracted, as they contained information on the source of the sequence data. Then, a vector with the flags 'keep', 'remove', or 'NA' was created for each metadata column (Supplementary Table S2). For each one of the three columns, samples were labeled 'keep' when they contained non-human, animal host names, 'remove' when they contained human terms, or both terms, human and non-human hosts term or 'NA' when they contained none of the terms in the dictionary and samples marked with only 'remove' or 'remove' and 'NA' were eliminated, while samples marked with either only 'keep' or 'keep' and 'NA' were retained. Samples were manually curated when they contained both 'keep' and 'remove', while samples marked with only 'NA' were labeled 'undefined'. We extracted the column 'study_abstract' to determine, case by case, whether samples classified as 'check' and 'undefined' were non-human host-associated. We manually inspected all samples after that to confirm their suitability.

In SRAdb, all the sample features are found in a single field named 'sample_attributes'. Therefore, the feature names and values of the samples in the 'sample_attributes' field are not coherently organized into distinct and well-defined metadata categories. We parsed the field attribute names and determined their frequency of occurrence. We removed sample attribute names with less than ten occurrences across a single dataset as they would make manual curation impossible. Next, we manually grouped synonymous attribute names (Supplementary Table S3).

Further, we extracted and standardized the values of ten attributes: sample altitude, sample elevation, sample collection date, sample temperature, sample pH, sample salinity, sample depth, sample latitude, sample longitude, and sample location (country and ocean/sea). Dates were standardized using International Standard Organization (ISO) 8601 (YYYY-MM-DD) [35]. Sample latitude and longitude were standardized to the format of decimal degrees. Location (country) was manually labeled following the standard of ISO 3166-1 [36]. The hosts' common and scientific names were derived from the metadata. The taxonomy of the hosts was identified using the R package ritis (Integrated Taxonomic Information System Client) [37]. Samples with no scientific names in the metadata were treated manually, and the following actions were taken sequentially: (a) their scientific names were searched in different taxonomy browsers on the internet (i.e., NCBI Taxonomy [38] and ITIS (Integrated Taxonomic Information System) [39]); (b) the taxonomic level of the common name was identified; (c) the complete taxonomy string of the host was deduced; (d) lastly, all the taxonomic levels are marked as NA.

We also identified and standardized four categories of host attributes: host characteristics (e.g., age, sex, height, length), host exposure (diet and last meal), host diagnosis, and sampling site/material. Host ages were standardized to years, length and height were standardized to a uniform SI system unit, and sex was standardized based on the metadata in the respective repositories. The sampling sites of mammals were organized into eight main categories: ear, gut, liver, lung, nose, oral, skin, biofluid, and the whole organism. We used the BRENDA [40] Tissue and Enzyme Source Ontology (http://bioportal.bioontology.org/ontologies/BTO/?p=classes&conceptid=root) to homogenize the 'Sample_Site' attributes. Terms used during the standardization of the attributes can be found in Supplementary Table S3.

For MG-RAST, the selected attributes were mostly already standardized, e.g., 'project_name', 'seq_meth', 'latitude', and 'country', among others. Therefore, the columns were adapted to the standard created during the standardization of the SRA retrieved metadata. The kingdom of the hosts was used to eliminate non-animal hosts. The complete set of standardized attributes can be found in Supplementary Table S4.

Combining SRA and MG-RAST

We identified equivalent and comparable attributes from the curated SRA metadata with those in metadata provided by MG-RAST (Supplementary Table S5), and the two metadata tables were merged. Five attributes (three related to library sequencing quality and two to sample attributes) were specific to SRA and MG-RAST, respectively. They are; 'quality_above_30_SRA',

'mean_quality_SRA', 'sample_pH', 'sample_salinity', for SRA and 'drisee_score_raw_MGRAST' for MG-RAST.

Web app implementation

The AAMDB web application was implemented using Shiny (version 1.5.0), an R package (version 3.6.3). The app was designed with a tab layout. The tabs include. 'Home', 'Quick search', 'Advanced search', 'Interactive map', 'Help', and 'Contact'. The 'Home' tab steers users around the application. The tabs 'Quick search' and 'Advanced search' provide filter options to aid users in selecting samples of interest. The 'Interactive map' tab allows users to select samples based on location. The interactive map functionality was implemented using the leaflet package (version 2.0.3), and we implemented the toolbox for selecting areas on the map with geoshaper (version 0.1.0) and the sp packages (version 1.4-2). The remaining R packages and their respective versions can be found in Supplementary Table S6 (Supplementary Material 6) [41–152]. The web application is available at https://webapp.ufz.de/aamdb/.

Quick search

The 'Quick Search' tab gives users access to all AAMDB content, and it is also possible to filter by main attributes. This feature shows all metagenomes, including those that do not have a valid geographic coordinate. The WebApp allows users to set up input filters using 20 available filters or by typing in the search box at the top of the table. After filtering the metadata, one can download a sample table with the associated metadata as comma-separated values (.csv) file. The metadata of the entire dataset can be downloaded if the user does not apply any filter. The steps to obtain raw sequencing data are described below.

Advanced search

The 'Advanced Search' tab allows the creation of a dynamic filter for the available attributes. We implemented a checkbox to allow users to filter out samples with missing values for the chosen attributes. The user can click the 'Search and add filters' button and a window will open. Searches for attributes can be made by name. Further, they are also organized using the following categories: 'Sample Attributes', 'Host Characteristics', 'Host Identity', 'Sample Site', 'Sample Location', 'Sequencing Features', 'Host Diagnosis', and 'Host Exposure'. At the top of the 'Advanced Search' page, we added a dropdown menu where users may select attributes to be added in the table below. After applying the filters and associated values, the metadata of selected entries can be downloaded as a comma-separated values (.csv) file.

Interactive map

The 'Interactive map' tab allows users to find the samples according to their location on the world map. Only samples with valid coordinates can be selected on the map. We implemented drawing tools (e.g., rectangular or polygon shapes) to help users select samples on the map.

It is important to note that individual points marked on the map may represent more than one sample since multiple samples can come from the same coordinate position. After selecting samples on the map, the selected metagenomes will be shown in the data table below the map. The users can further filter the samples using the 'Show filters' option, which will open the 'Quick Search' tab to filter the samples. After filtering the dataset, the resulting metadata table can be downloaded as a comma-separated values (.csv) file.

Downloading the raw data from selected metagenomes

We developed a simple download procedure to obtain raw sequence data from SRA. Unfortunately, MG-RAST does not allow public downloads anymore. Our Python scripts enable simple installation of the SRA-toolkit (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software) and the download of specific metagenomes or all metagenomes using the table exported by the WebApp, with two user-friendly commands. To support less experienced users, a script with a graphical user interface (GUI) is available (link at the end of the paragraph). Although most users may operate on Linux systems, we provide Windows executables to allow instant execution without installation. The download scripts are additionally [AB1] compatible with the CSV exports from the TerrestrialMetagenomeDB, HumanMetagenomeDB, and the MarineMetagenomeDB and are provided at https://github.com/mdsufz/downloadtool.

Results and discussion

Sample distribution and under-represented species and areas

The current version of the AAMDB version 1.0 includes metadata for 10,885 animal-associated metagenomes. Among them, 7,817 (71.81%) samples were retrieved from the SRA, and 3,040 (28.19%) samples were retrieved from MG-RAST. Hosts represented in our database span 10 phyla, 28 classes, 74 orders, 122 families, 174 genera, and 283 species (Fig. 2A). Samples with information on common host names were concentrated in four host species (mouse, cattle, pig, and chicken), which accounted for 5,560 (51.1%) samples (Fig. 2B). All samples had the attribute 'species name' filled. We split the host species' names into their common and scientific names. Most samples, 10,847 (99.65%), had the common host name, and 6,935 (63.71%) had the scientific host name (Fig. 2C). Illumina-based technologies were the most frequent

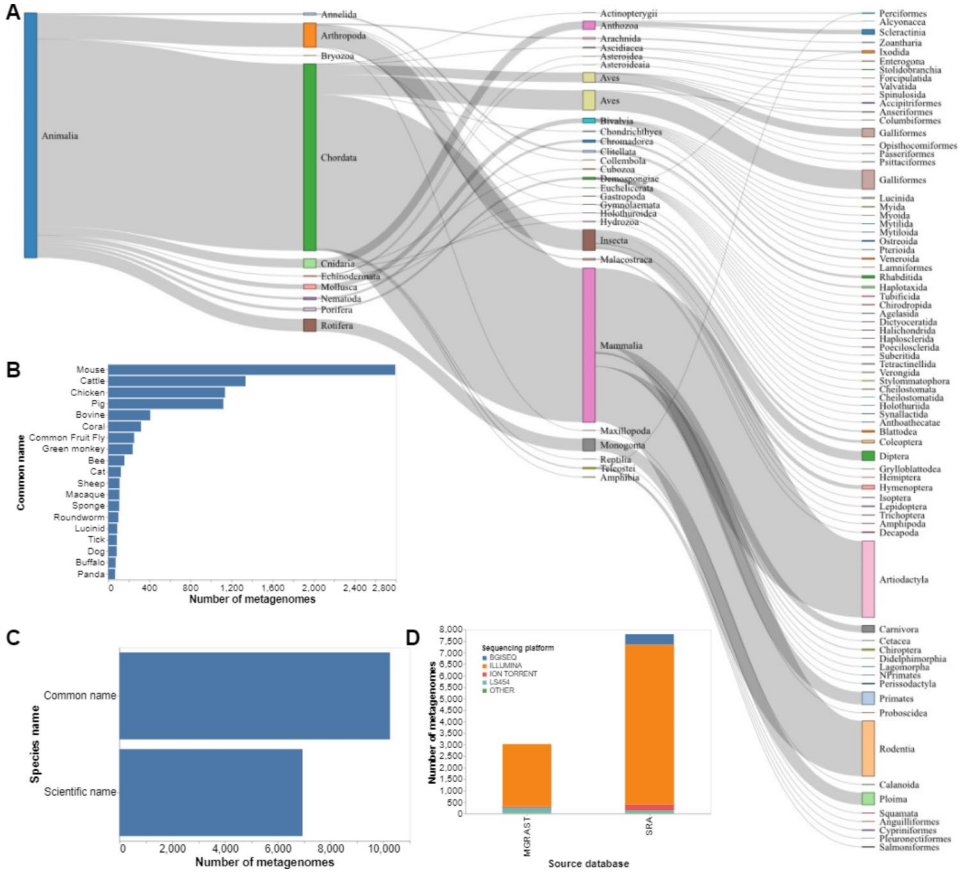

Fig. 2 Descriptive statistics of the AAMDB content. (**A**) Sankey plot containing the taxonomic classification in Kingdom, Phylum, Class, and Order. (**B**) Bar plot of the distribution of the top twenty (20) common host animal names. (**C**) Bar plot of the distribution of the samples with the host animal species name. (**D**) Bar plot of the distribution of sequencing technologies (Sequencing platform) per database of origin(Source database)

sequencing technology, with 9,663 (88.77%) samples. BGISEQ followed Illumina data with 462 (4.24%) samples, followed by Roche LS454 (3.19%), ION TORRENT (2.93%), PacBio (0,3%) and Nanopore (0,1%) (Fig. 2D). All animal attributes' frequency and co-occurrence were examined (Fig. 3). Beyond the differences in technology, we also analyzed the biogeography of the entries in our database regarding the following host distributions: (a) taxonomy; (b) use (domestic, wild animals, food stock and medical research); (c) aquatic vs. terrestrial animals; (d) vertebrate vs. invertebrate; (e) climate and economy of the country generating the data.

The metadata metagenome samples are distributed according to the taxonomy. There were 10,258 (94.24%) samples with a phylum assigned, 10,204 (93.74%) had a class and 9,942 (91.33%) had an order assignment. Family and genus were assigned to 6,635 (60.95%) and 6,585 (60.50%) samples, respectively. Using the taxonomic classification information from our data, we observed a bias towards vertebrate animals. Vertebrates represented 7,831 (76.34%) samples, whereas 2,427 (23.66%) were collected from invertebrates (Fig. 4C). The most common genus among invertebrates was *Trichocerca* (rotifers), with 512 samples (21% of the total invertebrate samples). Most vertebrates were used for livestock, medical research, and pets, representing 6,999 samples (64.30%, Fig. 4A). Of these samples, 2,792 (25.64%) represented mice, usually used in medical research. Livestock animals (bovine, chicken, pig) represented 4,001 (36.75%) samples while 206 (1.90%) samples represented pets. On the other hand, there were samples from wild animals such as gorillas with 43 samples, green monkeys with 238 samples, and dwarf tiger lucines with 7 samples.

Aquatic animals represented 1,017 (9.34%) entries, of which 326 (3%) were coral species and 512 (4.70%) were rotifers. The phylum Arthropoda comprised 1,010 (9.27%) samples. Other relatively frequent invertebrate groups of the dataset were mollusks. A study concerning biomass distribution on Earth [153] highlighted the impact of human civilization on global biomass through livestock. The AAMDB could be explored to examine the anthropogenic activity in animal microbiomes (e.g., following biodiversity and functional-potential profiles of wild animals, pets, and food stock over time).

A total of 6,808 entries had geographical coordinates. We found that temperate biomes were more extensively

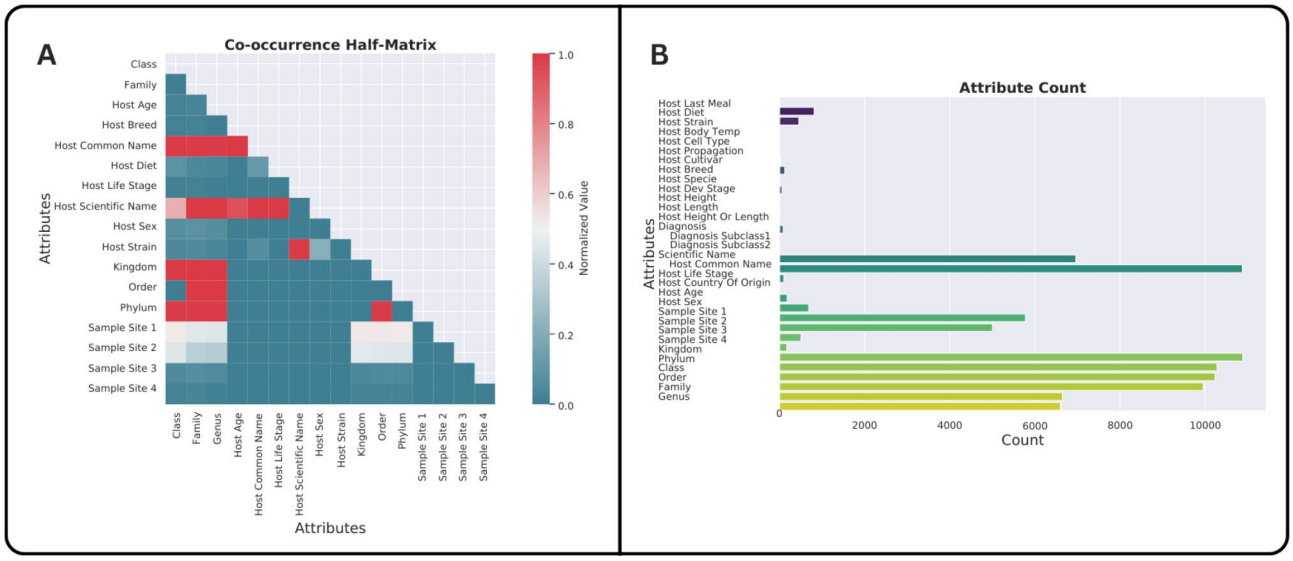

Fig. 3 Co-occurrence relationships and prevalence of attributes. (**A**) Half-matrix visualization providing a detailed representation of the frequencies of animal-associated attributes, capturing the complex co-occurrence relationships among different attributes. (**B**) Bar chart display showcasing the occurrence count of the predominant attributes, broken down per individual sample. This analysis illustrates the distribution and prevalence of these attributes within the collected samples

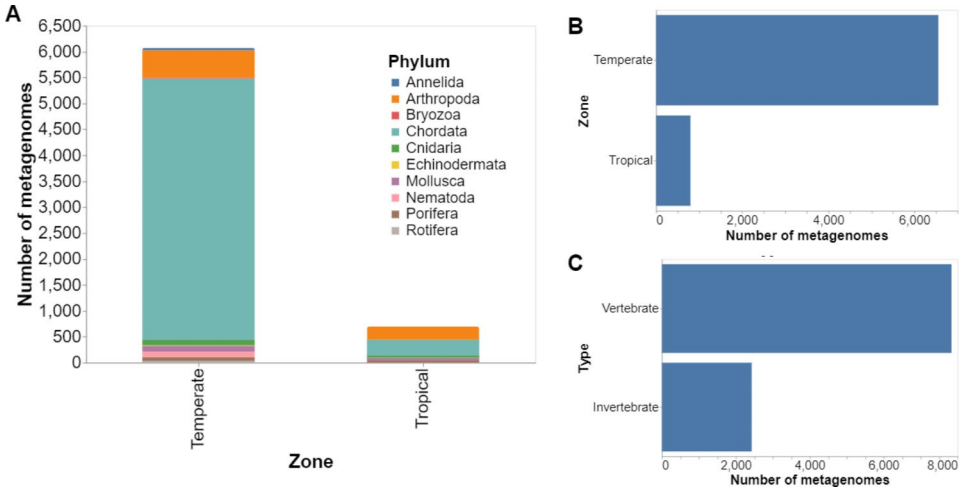

Fig. 4 Bias in sampling animal-associated metagenomes. (**A**) Bar Plot of the number of metagenomes by host phylum divided by tropical and temperate zone. (**B**) Bar plot of the samples' distribution in the tropical and temperate zone. (**C**) Bar Plot of the distribution of microbial metagenome samples from vertebrate and invertebrate hosts

sampled than tundra, flooded grasslands, savannas, mangroves, and most tropical biomes, likely reflecting differences in funding availability, access to modern molecular biology laboratories, and/or expertise in metagenomics analyses across countries [154]. In this work, 6,553 (89.16%) samples were retrieved from temperate climate zones, and only 796 (10.84%) were collected in tropical regions (Fig. 4B). Only 11 of the samples that had coordinates were from polar regions. Figure 5 shows that sampling has been concentrated in areas with higher economic indexes. However, these regions do not have the highest biodiversity [30].

To start addressing these gaps, awareness and support from the international scientific community are urgently needed for nations located in the tropics, where biological diversity is highest and the threats to its maintenance are the greatest [34]. Indeed, animal microbiomes in tropical countries are less studied than in temperate countries [10].

Usage and functionalities

The AAMDB has a simple and user-friendly interface consisting of three main sections, which allow users to select the type of query that suits their needs (Fig. 6). The

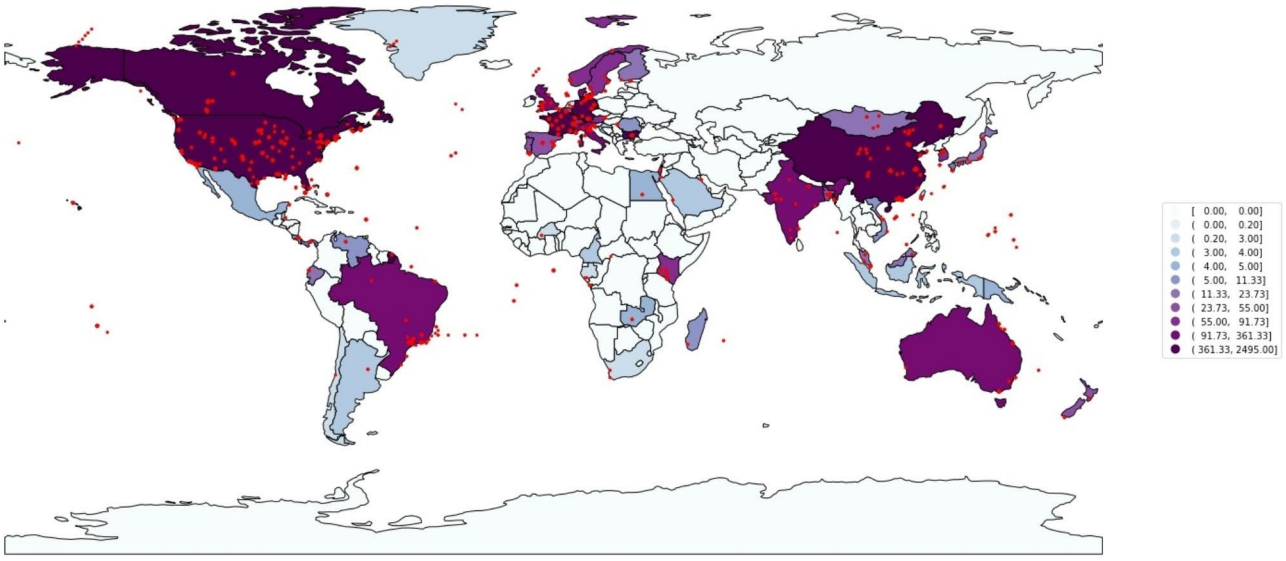

Fig. 5 World map showing the number of metagenome samples by region/country

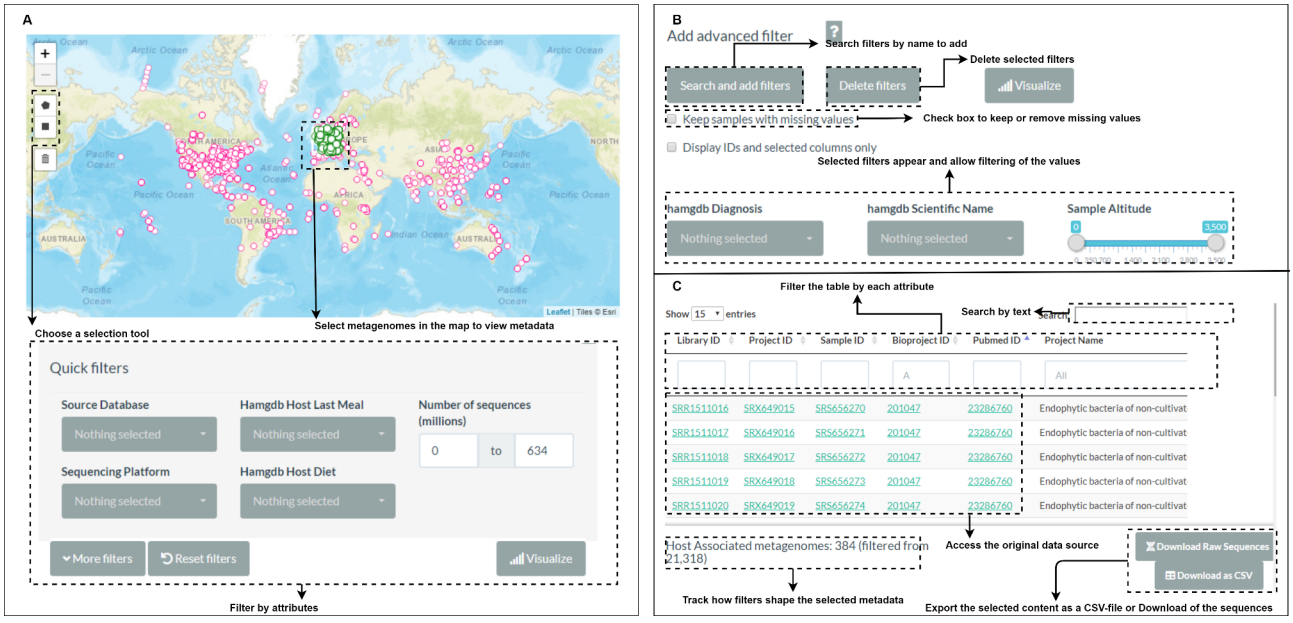

Fig. 6 AAMDB user-interface overview. (**A**) The 'Interactive Map' allows users to select samples according to their geographical location on the map using a selection tool. (**B**) The 'Advanced search' tab allows users to select as many filters as they want, and the metadata is displayed under the filtering options

'Quick Search' section allows users to apply quick filters with some attributes that appear initially. Clicking the "more filters" button allows the selection of further filters according to the user's needs. The 'Advanced Search' section has filters for all dataset attributes, allowing for a dynamic combination of these attributes. Finally, the 'Interactive map' section provides the user with a visual and intuitive search to select metagenomes according to the geographic coordinates of the samples. This functionality is limited to metagenomes with a valid geographic coordinate. Each sample has identification attribute hyperlinks ('sample_id', 'project_id', 'library_id', 'PubMed ID', and 'BioProject ID') to the source database (MG-RAST and SRA). All sections include functionalities to visualize the data distribution as a pie chart with the percentages of selected data or a histogram with the data distribution selected by the user, where the X-axis represents the value for a specific attribute chosen by the user, and the Y-axis represents the number of samples with this value. These features are intended to help users better understand the distribution of the selected data after

filtering and the number of samples with a given attribute value.

While developing the AAMDB, we ensured that users could filter metadata using all attributes and combine filters to optimize data searches. Indeed, our database contains 51 attributes (e.g., Host Body Temperature, Host Characteristics, Host Diagnosis, Host Strain and Host Sex). We also provide a tool for automatically downloading metagenome samples from the SRA database (see item 'Downloading the raw data from selected metagenomes' later in the manuscript).

Usage example

Bees are considered essential for maintaining biodiversity on Earth [155]. Scientists interested in bees may use the AAMDB to find metagenomes recovered from bees. The user can select the 'More filters' tab on the Quick Search tab to search for *bee* under 'AAMDB Host Common Name', resulting in a list of 157 samples. Users can select samples from countries of interest under the 'Sample Location Country' filter. Further, the user may select samples from 'Switzerland', decreasing the number of samples to 24. The user can click 'Visualize' to explore the selection. After the selection, the user can download the selected metadata as a CSV file for further analysis using an icon in the bottom right of the webpage and use our download tool (see the previous item) to retrieve the raw sequence data of the selected samples.

Database update plan

The number of metagenomic experiments submitted to public repositories, like SRA, is growing rapidly. Therefore, due to the manual curation steps that are necessary to construct our database, we will update the AAMDB with newly submitted microbial metagenome samples every February. Updates will serve as opportunities to include novel features or modify existing ones. Users can send questions and suggestions using the information included in the contact tab.

Suggestions for good practices

One of the goals of this work was to facilitate metagenome meta-analyses. To this end the AAMDB provides curated taxonomic classifications of the host animals, and includes a help guide for the community to improve metadata annotation when submitting novel metagenome samples to public repositories. Suggested ontologies can be located under Point 7 in the 'Help' tab of the AAMDB website under the title 'What should I do to include my metagenomes in AAMDB?'. Our database is not a repository for raw data. Still, when users follow our suggestions when submitting new entries to SRA, these new samples will be added to our metadata database during the yearly update.

The study by Wilkinson and collaborators [156] recognizes Findability, Accessibility, Interoperability, and Reusability (FAIR principles) as key pillars of robust research and effective data management. The AAMDB strictly adheres to these four principles. Firstly, the Findability of our data is ensured as each sample is associated with a unique identifier. The metadata has undergone manual curation and is entirely searchable in our WebApp. Secondly, Accessibility is guaranteed as the metadata is recoverable and made open access, promoting a culture of transparency and shared knowledge. The third point, Interoperability, is upheld by employing knowledge representation and standardized vocabularies that are consistent with international data management practices rather than merely adhering to the FAIR principles. Lastly, the Reusability of our data is ensured as the metadata, which has been manually curated in our work, comprises relevant attributes that conform to the standards accepted by the broader community.

Conclusions

Our work facilitates the reuse of metagenomics data, providing a WebApp with tools to search for information in a user-friendly way. Furthermore, it allows downloading metadata and metagenomes through our Download Tool. In different parts of this manuscript, we pointed out potential ways scientists interested in animal-associated metagenomes could reuse the data present in our database in new studies. It is relevant to indicate that, in most cases, further, targeted metadata would be necessary for such studies. The AnimalAssociatedDB may raise awareness of the importance of rich metadata to the reusability of metagenomic data. Besides being a valuable tool for scientists studying animal microbiomes and conducting meta-analyses, we shed light on the current bias in the data in public repositories toward vertebrates, temperate regions, and animals used in livestock, medical research, and pets. This work demonstrates that more studies on animal microbiomes outside these fields are necessary (e.g., studies involving biodiversity-, conservation-, and biotechnology-oriented surveys). Moreover, our study highlights the need for more research in underexplored geographic regions of the globe (e.g., tropical areas; Fig. 5), which contain most of the animal biodiversity on Earth and hold an untapped potential in their associated microbiomes.

Availability and requirements

Project name AnimalAssociatedMetagenomeDB.
Project home page: https://webapp.ufz.de/aamdb/.

Operating system(s) Platform independent.

Programming language R, Python.

Other requirements Python3.

License GNU GPL v3.

Any restrictions to use by non-academics See License.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s42523-023-00267-3.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Acknowledgements
We thank Dr. Sebastian Canzler, Dr. Andreas Schuttler, Dr. Mathias Bernt, and Sven Petruschke for supporting the shiny app deployment.

Authors' contributions
UNR and DSS conceptualized and supervised the study. UNR, DSS, APAS and MKN created the data frame and wrote the manuscript. APAS built the web app and created the figures. TKC assisted in filtering and standardization of the data. JCK assisted in checking and debugging codes during the project. JCK and AB developed the download tool. TKC, SDJ, TT and ACS provided specialist guidance for the standardization of the attributes, beta-tested the web app, and reviewed the manuscript before submission. All authors read and approved the final manuscript.

Funding
This work was funded by the Helmholtz Young Investigator grant VH-NG-1248 Micro' Big Data' and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 460129525. MKN was supported by the Petroleum Trust Development Fund (PTDF), German Academic Exchange Service (DAAD) (#91759074). TKC is the recipient of an investigator contract (CEECIND/00788/2017) conceded by the Fundação para a Ciência e a Tecnologia (FCT). JCK was supported by São Paulo Research Foundation (FAPESP; grant 2019/03396-9 and 2022/03534-5).
Open Access funding enabled and organized by Projekt DEAL.

Data availability
The dataset used during the current study is available in the SRA [http://www.ncbi.nlm.nih.gov/Traces/sra] and MG-RAST [http://metagenomics.anl.gov/] repositories.

Declarations

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Not Applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Environmental Microbiology, Helmholtz Centre for Environmental Research - UFZ GmbH, 04318 Leipzig, Germany
[2]Department of Computer Science and Interdisciplinary Centre of Bioinformatics, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Saxony, Germany
[3]Institute of Mathematics and Computer Sciences, University of Sao Paulo, Sao Carlos, Brazil
[4]GFZ German Research Centre for Geosciences, Section 3.7 Geomicrobiology, 14473 Telegrafenberg, Potsdam, Germany
[5]Federal Univ. of Technology - Paraná (UTFPR), Cornélio Procópio, Brazil
[6]Institute for Bioengineering and Biosciences (iBB) and Institute for Health and Bioeconomy (i4HB), Instituto Superior Tecnico (IST), Universidade de Lisboa, Lisbon 1049-001, Portugal
[7]Department of Bioanalytical Ecotoxicology, Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany
[8]Hohenheim Center for Livestock Microbiome Research (HoLMiR), University of Hohenheim, Stuttgart, Germany
[9]Institute of Animal Science, University of Hohenheim, Stuttgart, Germany
[10]German Centre of Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstraße 4, Leipzig 04103, Germany
[11]Max Planck Institute for Mathematics in the Sciences, Inselstraße, 04103 Leipzig, Germany
[12]Institute for Theoretical Chemistry, Universität Wien, Währingerstraße 17, Vienna A-1090, Austria
[13] Center for Scalable Data Analytics and Artificial Intelligence Dresden-Leipzig, Leipzig University, Leipzig, Germany
[14] Faculdad de Ciencias, Universidad Nacional de Colombia, Sede Bogotá, Bogotá, Colombia
[15] Center for non-coding RNA in Technology and Health, University of Copenhagen, Frederiksberg, Denmark
[16]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

Received: 9 November 2022 / Accepted: 18 September 2023
Published online: 05 October 2023

References

1. Xu Y, Lei B, Zhang Q, Lei Y, Li C, Li X, et al. ADDAGMA: a database for domestic animal gut microbiome atlas. Comput Struct Biotechnol J. 2022;20:891–8.
2. Bharucha T, Oeser C, Balloux F, Brown JR, Carbo EC, Charlett A, et al. STROBE-metagenomics: a STROBE extension statement to guide the reporting of metagenomics studies. Lancet Infect Dis. 2020;20:e251–60.
3. Aguiar-Pulido V, Huang W, Suarez-Ulloa V, Cickovski T, Mathee K, Narasimhan G, Metagenomics. Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis: supplementary issue: Bioinformatics methods and applications for big Metagenomics Data. Evol Bioinforma. 2016;12s1:EBO.S36436.
4. Singh S, Singh H, Rout B, Tripathi RBM, Chopra C, Chopra RS. The new science of metagenomics: revealing the secrets of microbial physiology. Metagenomics Tech Appl Chall Oppor. Springer; 2020. pp. 3–22.
5. Kodama Y, Shumway M, Leinonen R, on behalf of the International Nucleotide Sequence Database Collaboration. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40:D54–6.
6. Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2021;49:D10–7.
7. Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2021;49:D121–4.
8. Okido T, Kodama Y, Mashima J, Kosuge T, Fujisawa T, Ogasawara O. DNA Data Bank of Japan (DDBJ) update report 2021. Nucleic Acids Res. 2022;50:D102–5.
9. Harrison PW, Ahamed A, Aslam R, Alako BTF, Burgin J, Buso N, et al. The european Nucleotide Archive in 2020. Nucleic Acids Res. 2021;49:D82–5.
10. Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, Keegan KP, et al. The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Res. 2016;44:D590–4.
11. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, et al. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res. 2007;36:D534–8.
12. Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, et al. EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Res. 2018;46:D726–35.
13. Shi W, Qi H, Sun Q, Fan G, Liu S, Wang J, et al. gcMeta: a global catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. Nucleic Acids Res. 2019;47:D637–48.

14. Corrêa FB, Saraiva JP, Stadler PF, da Rocha UN. TerrestrialMetagenomeDB: a public repository of curated and standardized metadata for terrestrial metagenomes. Nucleic Acids Res. 2019;gkz994.
15. da Rocha UN, Nata'ala MK, Santos APA, Kasmanas JC, Bartholomäus A, Saraiva JP et al. MarineMetagenomeDB: a public repository for curated and standardized metadata for marine metagenomes [Internet]. In Review; 2022 Apr. Available from: https://www.researchsquare.com/article/rs-1431837/v1.
16. Kasmanas JC, Bartholomäus A, Corrêa FB, Tal T, Jehmlich N, Herberth G, et al. HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. Nucleic Acids Res. 2021;49:D743–50.
17. Agostinetto G, Bozzi D, Porro D, Casiraghi M, Labra M, Bruno A. SKIOME Project: a curated collection of skin microbiome datasets enriched with study-related metadata. Database [Internet]. 2022;2022. https://doi.org/10.1093/database/baac033.
18. Zhang Q, Yu K, Li S, Zhang X, Zhao Q, Zhao X et al. gutMEGA: a database of the human gut MEtaGenome Atlas. Brief Bioinform [Internet]. 2020;22. https://doi.org/10.1093/bib/bbaa082.
19. Forster SC, Browne HP, Kumar N, Hunt M, Denise H, Mitchell A, et al. HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. Nucleic Acids Res. 2015;44:D604–9.
20. Hu R, Yao R, Li L, Xu Y, Lei B, Tang G, et al. A database of animal metagenomes. Sci Data. 2022;9:312.
21. Torres PJ, Edwards RA, McNair KA. PARTIE: a partition engine to separate metagenomic and amplicon projects in the Sequence Read Archive. Valencia A, editor. Bioinformatics. 2017;33:2389–91.
22. Jurburg SD, Konzack M, Eisenhauer N, Heintz-Buschart A. The archives are half-empty: an assessment of the availability of microbial community sequencing data. Commun Biol. 2020;3:474.
23. Metch JW, Burrows ND, Murphy CJ, Pruden A, Vikesland PJ. Metagenomic analysis of microbial communities yields insight into impacts of nanoparticle design. Nat Nanotechnol. 2018;13:253–9.
24. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol. 2019;37:953–61.
25. Zhu F, Ju Y, Wang W, Wang Q, Guo R, Ma Q, et al. Metagenome-wide association of gut microbiome features for schizophrenia. Nat Commun. 2020;11:1612.
26. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008;26:541–7.
27. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. 2012;40:D57–63.
28. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29:415–20.
29. Titley MA, Snaddon JL, Turner EC. Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. Schierwater B, editor. PLOS ONE. 2017;12:e0189577.
30. Raven PH, Gereau RE, Phillipson PB, Chatelain C, Jenkins CN, Ulloa Ulloa C. The distribution of biodiversity richness in the tropics. Sci Adv. 2020;6:eabc6228.
31. Assessment USCO. of T. Technologies to maintain biological diversity. Congress of the US, Office of Technology Assessment; 1987.
32. Brady NC. International development and the protection of biological diversity. Biodiversity. 1988;409:411.
33. Harfoot MBJ, Johnston A, Balmford A, Burgess ND, Butchart SHM, Dias MP, et al. Using the IUCN Red List to map threats to terrestrial vertebrates at global scale. Nat Ecol Evol. 2021;5:1510–9.
34. Peixoto RS, Voolstra CR, Sweet M, Duarte CM, Carvalho S, Villela H et al. Harnessing the microbiome to prevent global biodiversity loss. Nat Microbiol [Internet]. 2022 [cited 2022 Oct 24]; Available from: https://www.nature.com/articles/s41564-022-01173-1.
35. ISO - ISO. 8601 — Date and time format [Internet]. ISO. 2017 [cited 2023 Jul 3]. Available from: https://www.iso.org/iso-8601-date-and-time-format.html.
36. ISO - ISO. 3166 — Country Codes [Internet]. ISO. [cited 2023 Jul 3]. Available from: https://www.iso.org/iso-3166-country-codes.html.
37. ritis. : Integrated Taxonomic Information System Client version 1.0.0 from CRAN [Internet]. [cited 2023 Jul 3]. Available from: https://rdrr.io/cran/ritis/.
38. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012;40:D136–43.
39. ITIS - MEMORANDUM OF UNDERSTANDING [Internet]. [cited 2023 Jul 4]. Available from: https://www.itis.gov/mou.html.
40. Jupp S, Burdett T, Leroy C, Parkinson HE. A new Ontology Lookup Service at EMBL-EBI. Workshop Semantic Web Appl Tools Life Sci. 2015.
41. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing. ; 2021. Available from: https://www.R-project.org/.
42. Xie Y, Cheng J, Tan X. DT: A Wrapper of the JavaScript Library "DataTables" [Internet]. 2021. Available from: https://CRAN.R-project.org/package=DT.
43. Venables WN, Ripley BD. Modern Applied Statistics with S [Internet]. Fourth. New York: Springer; 2002. Available from: https://www.stats.ox.ac.uk/pub/MASS4/.
44. Bates D, Maechler M, Matrix. Sparse and Dense Matrix Classes and Methods [Internet]. 2021. Available from: https://CRAN.R-project.org/package=Matrix.
45. Chang W. R6: Encapsulated Classes with Reference Semantics [Internet]. 2021. Available from: https://CRAN.R-project.org/package=R6.
46. Neuwirth E, RColorBrewer. ColorBrewer Palettes [Internet]. 2014. Available from: https://CRAN.R-project.org/package=RColorBrewer.
47. Eddelbuettel D, Balamuta JJ. Extending extitR with extitC++: a brief introduction to extitRcpp. Am Stat. 2018;72:28–36.
48. Ooms Jaskpass. Safe Password Entry for R, Git, and SSH [Internet]. 2019. Available from: https://CRAN.R-project.org/package=askpass.
49. Urbanek S. base64enc: Tools for base64 encoding [Internet]. 2015. Available from: https://CRAN.R-project.org/package=base64enc.
50. Horner J. brew: Templating Framework for Report Generation [Internet]. 2011. Available from: https://CRAN.R-project.org/package=brew.
51. Csárdi G, Chang W. callr: Call R from R [Internet]. 2021. Available from: https://CRAN.R-project.org/package=callr.
52. Csárdi G, cli. Helpers for Developing Command Line Interfaces [Internet]. 2021. Available from: https://CRAN.R-project.org/package=cli.
53. Lincoln Mclipr. Read and Write from the System Clipboard [Internet]. 2020. Available from: https://CRAN.R-project.org/package=clipr.
54. Zeileis A, Fisher JC, Hornik K, Ihaka R, McWhite CD, Murrell P, et al. Colorspace: a toolbox for manipulating and assessing colors and palettes. J Stat Softw. 2020;96:1–49.
55. Ooms J, commonmark. High Performance CommonMark and Github Markdown Rendering in R [Internet]. 2018. Available from: https://CRAN.R-project.org/package=commonmark.
56. Hester J, François R. cpp11: A C + + 11 Interface for R's C Interface [Internet]. 2021. Available from: https://CRAN.R-project.org/package=cpp11.
57. Csárdi G. crayon: Colored Terminal Output [Internet]. 2021. Available from: https://CRAN.R-project.org/package=crayon.
58. Cheng J, Sievert C. crosstalk: Inter-Widget Interactivity for HTML Widgets [Internet]. 2021. Available from: https://CRAN.R-project.org/package=crosstalk.
59. Ooms J. curl: A Modern and Flexible Web Client for R [Internet]. 2021. Available from: https://CRAN.R-project.org/package=curl.
60. Dowle M, Srinivasan A. data.table: Extension of `data.frame` [Internet]. 2021. Available from: https://CRAN.R-project.org/package=data.table.
61. Csárdi G, Müller K, Hester J. desc: Manipulate DESCRIPTION Files [Internet]. 2021. Available from: https://CRAN.R-project.org/package=desc.
62. Wickham H, Hester J, Chang W. devtools: Tools to Make Developing R Packages Easier [Internet]. 2021. Available from: https://CRAN.R-project.org/package=devtools.
63. Lucas DE with contributions by, Tuszynski A, Bengtsson J, Urbanek H, Frasca S, Lewis M et al. B,. digest: Create Compact Hash Digests of R Objects [Internet]. 2021. Available from: https://CRAN.R-project.org/package=digest.
64. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation [Internet]. 2021. Available from: https://CRAN.R-project.org/package=dplyr.
65. Wickham H. ellipsis: Tools for Working with ... [Internet]. 2021. Available from: https://CRAN.R-project.org/package=ellipsis
66. Wickham H, Xie Y. evaluate: Parsing and Evaluation Tools that Provide More Details than the Default [Internet]. 2019. Available from: https://CRAN.R-project.org/package=evaluate.
67. Gaslam B. fansi: ANSI Control Sequence Aware String Functions [Internet]. 2021. Available from: https://CRAN.R-project.org/package=fansi.
68. Pedersen TL, Nicolae B, François R, farver. High Performance Colour Space Manipulation [Internet]. 2021. Available from: https://CRAN.R-project.org/package=farver.
69. Chang W. fastmap: Fast Data Structures [Internet]. 2021. Available from: https://CRAN.R-project.org/package=fastmap.

70. Hester J, Wickham H, fs. Cross-Platform File System Operations Based on "libuv" [Internet]. 2020. Available from: https://CRAN.R-project.org/package=fs.
71. Wickham H, Kuhn M, Vaughan D. generics: Common S3 Generics not Provided by Base R Methods Related to Model Fitting [Internet]. 2021. Available from: https://CRAN.R-project.org/package=generics.
72. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org.
73. Bryan J, Wickham H, gh. "GitHub" "API" [Internet]. 2021. Available from: https://CRAN.R-project.org/package=gh.
74. Hester J. glue: Interpreted String Literals [Internet]. 2020. Available from: https://CRAN.R-project.org/package=glue.
75. Auguie B, gridExtra. Miscellaneous Functions for "Grid" Graphics [Internet]. 2017. Available from: https://CRAN.R-project.org/package=gridExtra.
76. Wickham H, Pedersen TL. gtable: Arrange "Grobs" in Tables [Internet]. 2019. Available from: https://CRAN.R-project.org/package=gtable.
77. Xie Y, Qiu Y, highr. Syntax Highlighting for R Source Code [Internet]. 2021. Available from: https://CRAN.R-project.org/package=highr.
78. Cheng J, Sievert C, Schloerke B, Chang W, Xie Y, Allen J. htmltools: Tools for HTML [Internet]. 2021. Available from: https://CRAN.R-project.org/package=htmltools.
79. Cheng J, Chang W. httpuv: HTTP and WebSocket Server Library [Internet]. 2021. Available from: https://CRAN.R-project.org/package=httpuv.
80. Wickham H. httr: Tools for Working with URLs and HTTP [Internet]. 2020. Available from: https://CRAN.R-project.org/package=httr.
81. Dias DV. ini: Read and Write ".ini" Files [Internet]. 2018. Available from: https://CRAN.R-project.org/package=ini.
82. Wilke CO, Pedersen TL. isoband: Generate Isolines and Isobands from Regularly Spaced Elevation Grids [Internet]. 2021. Available from: https://CRAN.R-project.org/package=isoband.
83. Ooms J. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. ArXiv14032805 StatCO [Internet]. 2014; Available from: https://arxiv.org/abs/1403.2805.
84. Xie Y, knitr:. A General-Purpose Package for Dynamic Report Generation in R [Internet]. 2021. Available from: https://yihui.org/knitr/.
85. Justin Talbot. labeling: Axis Labeling [Internet]. 2020. Available from: https://CRAN.R-project.org/package=labeling.
86. Chang W, Cheng J. later: Utilities for Scheduling Functions to Execute Later with Event Loops [Internet]. 2021. Available from: https://CRAN.R-project.org/package=later.
87. Sarkar D, Lattice. Multivariate Data Visualization with R [Internet]. New York: Springer; 2008. Available from: http://lmdvr.r-forge.r-project.org.
88. Wickham H. lazyeval: Lazy (Non-Standard) Evaluation [Internet]. 2019. Available from: https://CRAN.R-project.org/package=lazyeval.
89. Cheng J, Karambelkar B, Xie Y. leaflet: Create Interactive Web Maps with the JavaScript "Leaflet" Library [Internet]. 2021. Available from: https://CRAN.R-project.org/package=leaflet.
90. Karambelkar B, Schloerke B. leaflet.extras: Extra Functionality for "leaflet" Package [Internet]. 2018. Available from: https://CRAN.R-project.org/package=leaflet.extras.
91. Huang L. leaflet.providers: Leaflet Providers [Internet]. 2019. Available from: https://CRAN.R-project.org/package=leaflet.providers.
92. Henry L, Wickham H. lifecycle: Manage the Life Cycle of your Package Functions [Internet]. 2021. Available from: https://CRAN.R-project.org/package=lifecycle.
93. Bache SM, Wickham H. magrittr: A Forward-Pipe Operator for R [Internet]. 2020. Available from: https://CRAN.R-project.org/package=magrittr.
94. Bivand R, Lewin-Koh N. maptools: Tools for Handling Spatial Objects [Internet]. 2021. Available from: https://CRAN.R-project.org/package=maptools.
95. Allaire JJ, Horner J, Xie Y, Marti V, Porte N. markdown: Render Markdown with the C Library "Sundown" [Internet]. 2019. Available from: https://CRAN.R-project.org/package=markdown.
96. Wickham H, Hester J, Chang W, Müller K, Cook D. memoise: Memoisation of Functions [Internet]. 2021. Available from: https://CRAN.R-project.org/package=memoise.
97. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc B. 2011;73:3–36.
98. Xie Y. mime: Map Filenames to MIME Types [Internet]. 2021. Available from: https://CRAN.R-project.org/package=mime.
99. Wickham C. munsell: Utilities for Using Munsell Colours [Internet]. 2018. Available from: https://CRAN.R-project.org/package=munsell.
100. Ooms J. openssl: Toolkit for Encryption, Signatures and Certificates Based on OpenSSL [Internet]. 2021. Available from: https://CRAN.R-project.org/package=openssl.
101. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: Linear and Nonlinear Mixed Effects Models [Internet]. 2021. Available from: https://CRAN.R-project.org/package=nlme.
102. Müller K, Wickham H. pillar: Coloured Formatting for Columns [Internet]. 2021. Available from: https://CRAN.R-project.org/package=pillar.
103. Wickham H, Hester J, pkgbuild. Find Tools Needed to Build R Packages [Internet]. 2020. Available from: https://CRAN.R-project.org/package=pkgbuild.
104. Csárdi G, pkgconfig. Private Configuration for "R" Packages [Internet]. 2019. Available from: https://CRAN.R-project.org/package=pkgconfig.
105. Wickham H, Hester J, Chang W, pkgload. Simulate Package Installation and Attach [Internet]. 2021. Available from: https://CRAN.R-project.org/package=pkgload.
106. Sievert C. Interactive Web-Based Data Visualization with R, plotly, and shiny [Internet]. Chapman and Hall/CRC; 2020. Available from: https://plotly-r.com.
107. Urbanek S. png: Read and write PNG images [Internet]. 2013. Available from: https://CRAN.R-project.org/package=png.
108. Csardi G, Sorhus S. praise: Praise Users [Internet]. 2015. Available from: https://CRAN.R-project.org/package=praise.
109. Csardi Gprettyunits. Pretty, Human Readable Formatting of Quantities [Internet]. 2020. Available from: https://CRAN.R-project.org/package=prettyunits.
110. Csárdi G, Chang Wprocessx. Execute and Control System Processes [Internet]. 2021. Available from: https://CRAN.R-project.org/package=processx.
111. Cheng J. promises: Abstractions for Promise-Based Asynchronous Programming [Internet]. 2021. Available from: https://CRAN.R-project.org/package=promises.
112. Loden J, Daeschler D, Rodola' G. ps: List, Query, Manipulate System Processes [Internet]. 2021. Available from: https://CRAN.R-project.org/package=ps.
113. Henry L, Wickham H, purrr. Functional Programming Tools [Internet]. 2020. Available from: https://CRAN.R-project.org/package=purrr.
114. Hijmans RJ. raster: Geographic Data Analysis and Modeling [Internet]. 2021. Available from: https://CRAN.R-project.org/package=raster.
115. Csárdi Grcmdcheck. Run "R CMD check" from "R" and Capture Results [Internet]. 2021. Available from: https://CRAN.R-project.org/package=rcmdcheck.
116. Csárdi G. rematch2: Tidy Output from Regular Expression Matching [Internet]. 2020. Available from: https://CRAN.R-project.org/package=rematch2.
117. Hester J, Csárdi G, Wickham H, Chang W, Morgan M, Tenenbaum D. remotes: R Package Installation from Remote Repositories, Including "GitHub" [Internet]. 2021. Available from: https://CRAN.R-project.org/package=remotes.
118. Henry L, Wickham H. rlang: Functions for Base Types and Core R and "Tidyverse" Features [Internet]. 2021. Available from: https://CRAN.R-project.org/package=rlang.
119. Allaire JJ, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A et al. rmarkdown: Dynamic Documents for R [Internet]. 2022. Available from: https://github.com/rstudio/rmarkdown.
120. Wickham H, Danenberg P, Csárdi G, Eugster M. roxygen2: In-Line Documentation for R [Internet]. 2021. Available from: https://CRAN.R-project.org/package=roxygen2.
121. Müller K, rprojroot. Finding Files in Project Subdirectories [Internet]. 2020. Available from: https://CRAN.R-project.org/package=rprojroot.
122. Müller K, Ushey K, Allaire JJ, Wickham H, Ritchie G. rstudioapi: Safely Access the RStudio API [Internet]. 2020. Available from: https://CRAN.R-project.org/package=rstudioapi.
123. Csárdi Grversions. Query "R" Versions, Including "r-release" and "r-oldrel" [Internet]. 2021. Available from: https://CRAN.R-project.org/package=rversions.
124. Wickham H, Seidel D. scales: Scale Functions for Visualization [Internet]. 2020. Available from: https://CRAN.R-project.org/package=scales.
125. Wickham H, Chang W, Flight R, Müller K, Hester J. sessioninfo: R Session Information [Internet]. 2021. Available from: https://CRAN.R-project.org/package=sessioninfo.
126. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. shiny: Web Application Framework for R [Internet]. 2020. Available from: https://CRAN.R-project.org/package=shiny.
127. Bailey E, shinyBS. Twitter Bootstrap Components for Shiny [Internet]. 2015. Available from: https://CRAN.R-project.org/package=shinyBS.
128. Perrier V, Meyer F, Granjon D, shinyWidgets. Custom Inputs Widgets for Shiny [Internet]. 2021. Available from: https://CRAN.R-project.org/package=shinyWidgets.

129. Attali D, shinyjs. Easily Improve the User Experience of Your Shiny Apps in Seconds [Internet]. 2020. Available from: https://CRAN.R-project.org/package=shinyjs.
130. Chang W. shinythemes: Themes for Shiny [Internet]. 2021. Available from: https://CRAN.R-project.org/package=shinythemes.
131. Ushey K. sourcetools: Tools for Reading, Tokenizing and Parsing R Code [Internet]. 2018. Available from: https://CRAN.R-project.org/package=sourcetools.
132. Bivand RS, Pebesma E, Gomez-Rubio V. Applied spatial data analysis with R, Second edition [Internet]. Springer, NY; 2013. Available from: https://asdar-book.org/.
133. Gagolewski M. Stringi: fast and portable character string processing in R. J Stat Softw. 2021.
134. Wickham Hstringr. Simple, Consistent Wrappers for Common String Operations [Internet]. 2019. Available from: https://CRAN.R-project.org/package=stringr.
135. Ooms J. sys: Powerful and Reliable Tools for Running System Commands in R [Internet]. 2020. Available from: https://CRAN.R-project.org/package=sys.
136. Wickham H. Testthat: get started with testing. R J. 2011;3:5–10.
137. Müller K, Wickham H. tibble: Simple Data Frames [Internet]. 2021. Available from: https://CRAN.R-project.org/package=tibble.
138. Wickham H. tidyr: Tidy Messy Data [Internet]. 2021. Available from: https://CRAN.R-project.org/package=tidyr.
139. Henry L, Wickham H. tidyselect: Select from a Set of Strings [Internet]. 2021. Available from: https://CRAN.R-project.org/package=tidyselect.
140. Xie Y, tinytex. Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents [Internet]. 2022. Available from: https://github.com/yihui/tinytex.
141. Wickham H, Bryan J, Barrett M. usethis: Automate Package and Project Setup [Internet]. 2021. Available from: https://CRAN.R-project.org/package=usethis.
142. Perry PO. utf8: Unicode Text Processing [Internet]. 2021. Available from: https://CRAN.R-project.org/package=utf8.
143. Wickham H, Henry L, Vaughan D. vctrs: Vector Helpers [Internet]. 2021. Available from: https://CRAN.R-project.org/package=vctrs.
144. Garnier S, Ross, Noam, Rudis R et al. viridis - Colorblind-Friendly Color Maps for R [Internet]. 2021. Available from: https://sjmgarnier.github.io/viridis/.
145. Coene J. waiter: Loading Screen for "Shiny" [Internet]. 2021. Available from: https://CRAN.R-project.org/package=waiter.
146. de Jonge E. whisker: mustache for R, Logicless Templating [Internet]. 2019. Available from: https://CRAN.R-project.org/package=whisker.
147. Hester J, Müller K, Ushey K, Wickham H, Chang W. withr: Run Code "With" Temporarily Modified Global State [Internet]. 2021. Available from: https://CRAN.R-project.org/package=withr.
148. Xie Y, xfun. Supporting Functions for Packages Maintained by "Yihui Xie" [Internet]. 2022. Available from: https://CRAN.R-project.org/package=xfun.
149. Wickham H, Hester J, Ooms J. xml2: Parse XML [Internet]. 2020. Available from: https://CRAN.R-project.org/package=xml2.
150. Csárdi G, Boudra F, Dieter R, Krammer K, White J, xopen. Open System Files, "URLs", Anything [Internet]. 2018. Available from: https://CRAN.R-project.org/package=xopen.
151. Dahl DB, Scott D, Roosen C, Magnusson A, Swinton J. xtable: Export Tables to LaTeX or HTML [Internet]. 2019. Available from: https://CRAN.R-project.org/package=xtable.
152. Stephens J, Simonov K, Xie Y, Dong Z, Wickham H, Horner J et al. yaml: Methods to Convert R Data to YAML and Back [Internet]. 2020. Available from: https://CRAN.R-project.org/package=yaml.
153. Bar-On YM, Phillips R, Milo R. The biomass distribution on Earth. Proc Natl Acad Sci. 2018;115:6506–11.
154. Guerra CA, Heintz-Buschart A, Sikorski J, Chatzinotas A, Guerrero-Ramírez N, Cesarz S, et al. Blind spots in global soil biodiversity and ecosystem function research. Nat Commun. 2020;11:3870.
155. Hung K-LJ, Kingston JM, Albrecht M, Holway DA, Kohn JR. The worldwide importance of honey bees as pollinators in natural habitats. Proc R Soc B Biol Sci. 2018;285:20172140.
156. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.

Publisher's Note