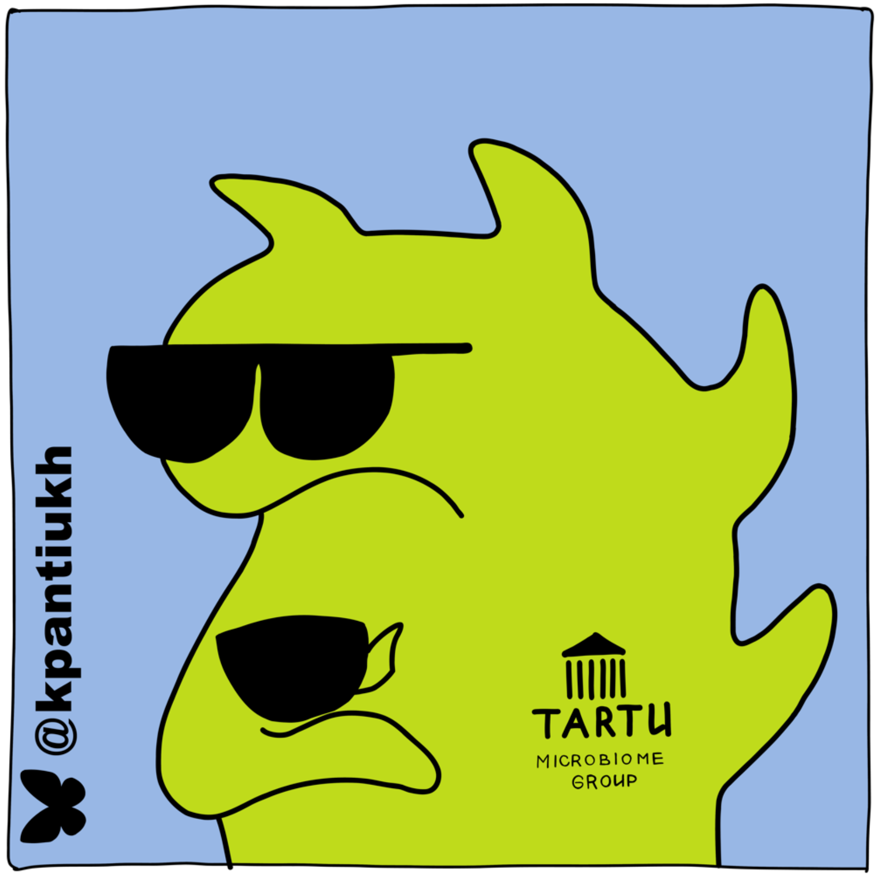


# Data analysis
## Lecture 2

Metalog data

Kateryna Pantiukh
pantiukh@ut.ee

GitHub

Please set up your GitHub repository and code editor (Visual Studio Code)

## GitHub

https://github.com

## Visual Studio Code

https://code.visualstudio.com

Please clone our working directory from GitHub

GitHub
https://github.com/Chartiza/KSE_microbiome

# Feedback

Please write in the chat how much time you spent on your homework.

# Metalog. Open-source data

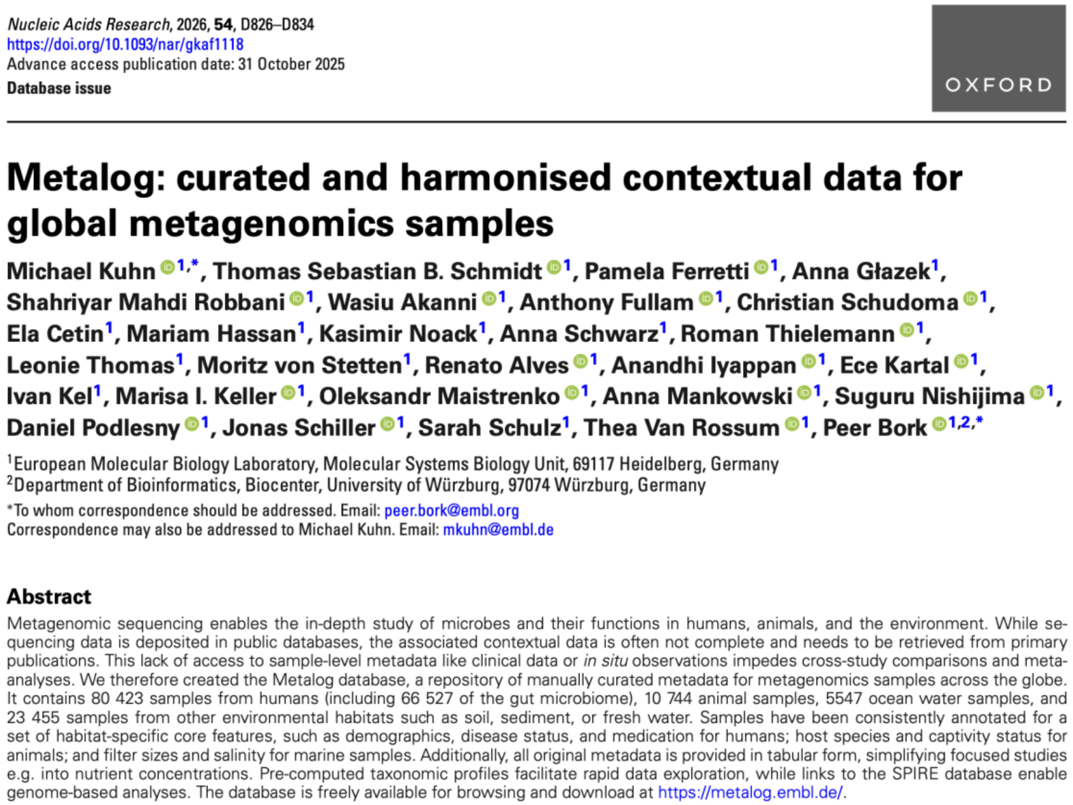
https://metalog.embl.de

Metalog is a repository of manually curated metadata for metagenomic samples

It is being developed in the Bork Group at EMBL Heidelberg

With support from more than 30 countries, the European Molecular Biology Laboratory (EMBL) has more than 110 independent research groups and service teams covering the spectrum of molecular biology at six sites in Barcelona, Grenoble, Hamburg, Heidelberg, EMBL-EBI Hinxton, and Rome.

# Metalog. Open-source data



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

OXFORD

## 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/.

Database:

https://metalog.embl.de

Paper:

https://academic.oup.com/nar/article/54/D1/D826/8307355

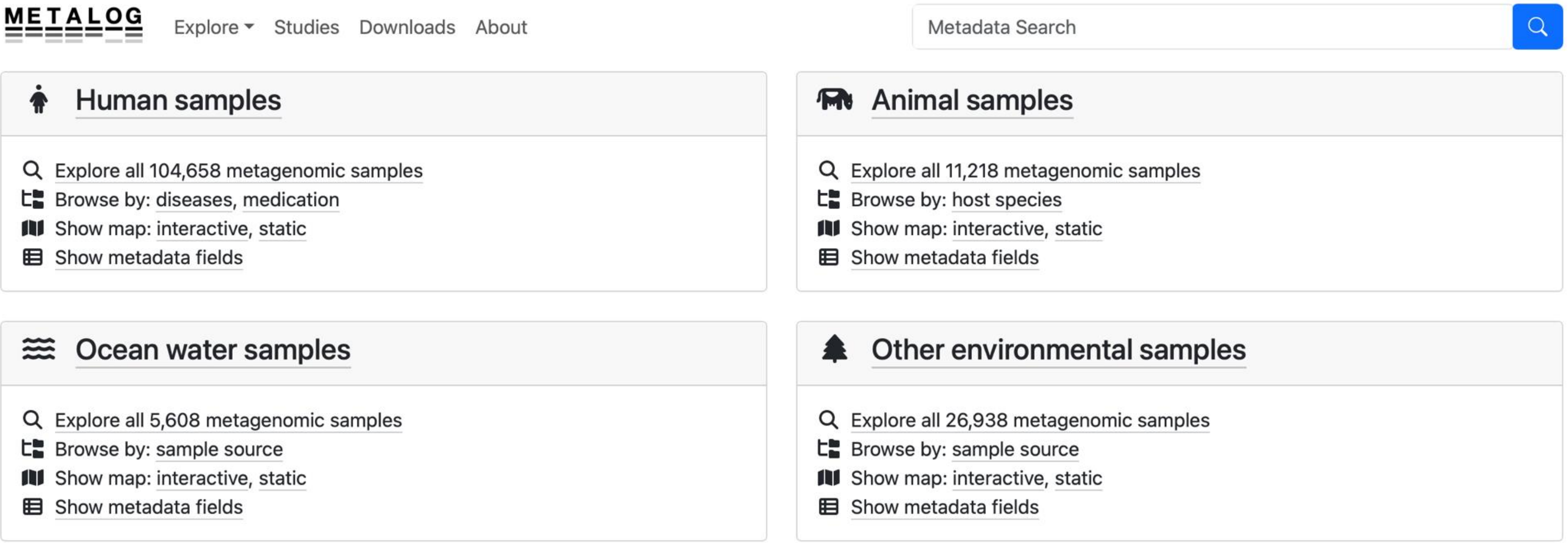
# Metalog. Open-source data

Metadata + Taxonomic profile

*(MetaPlAn 4.0)*

https://metalog.embl.de

# Metalog. Open-source data



**Human samples**

- Explore all 104,658 metagenomic samples
- Browse by: diseases, medication
- Show map: interactive, static
- Show metadata fields

**Animal samples**

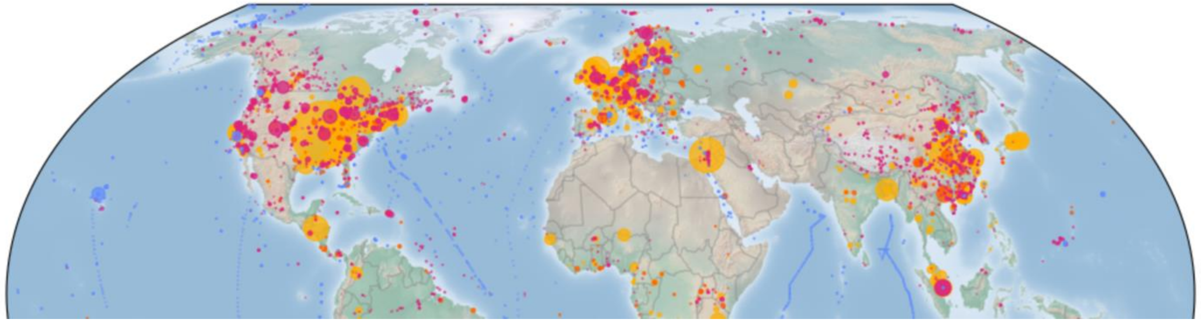
- Explore all 11,218 metagenomic samples
- Browse by: host species
- Show map: interactive, static
- Show metadata fields

**Ocean water samples**

- Explore all 5,608 metagenomic samples
- Browse by: sample source
- Show map: interactive, static
- Show metadata fields

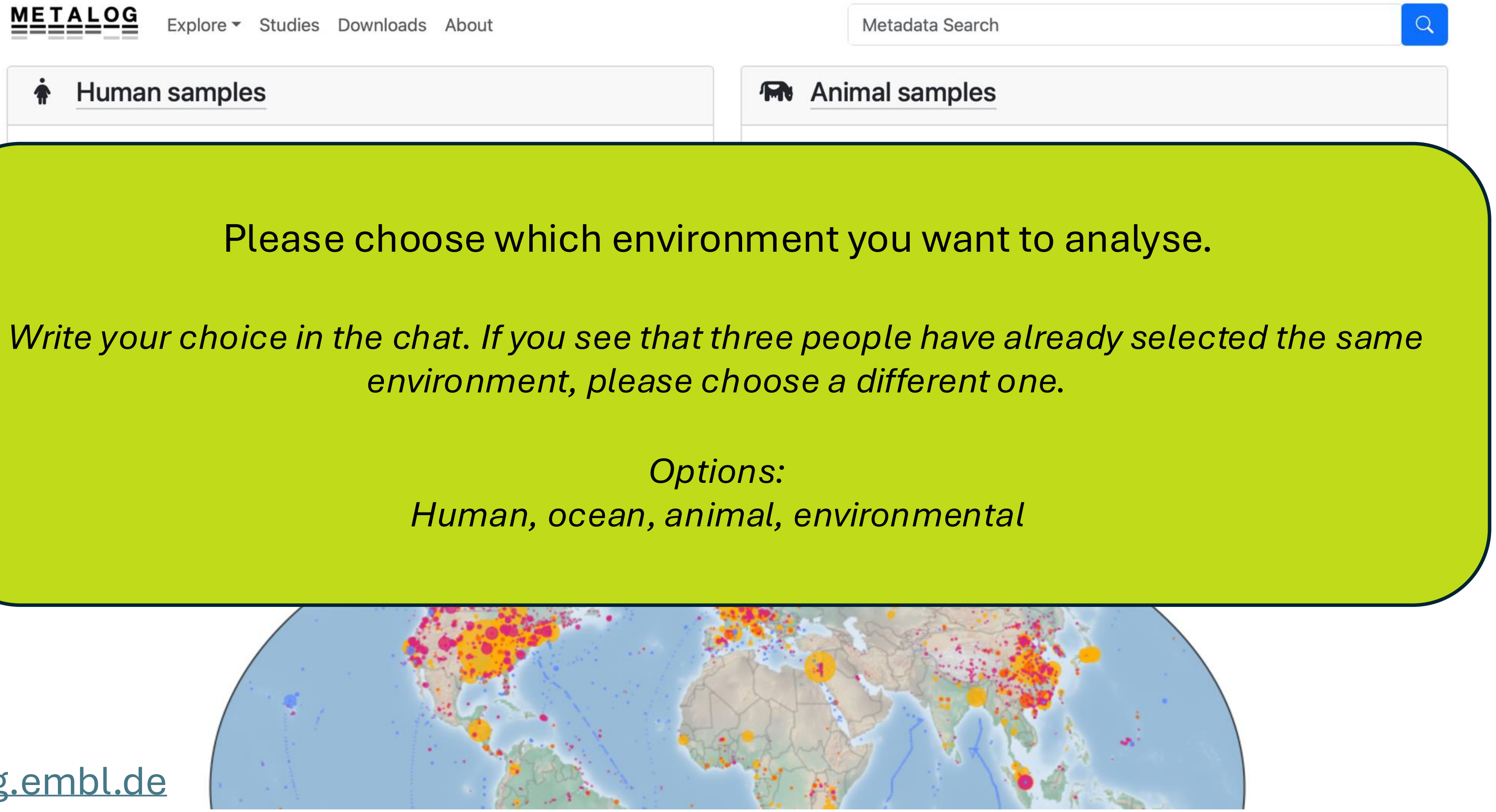
**Other environmental samples**

- Explore all 26,938 metagenomic samples
- Browse by: sample source
- Show map: interactive, static
- Show metadata fields



https://metalog.embl.de

# Metalog. Open-source data



Please choose which environment you want to analyse.

*Write your choice in the chat. If you see that three people have already selected the same environment, please choose a different one.*

*Options:*
*Human, ocean, animal, environmental*

https://metalog.embl.de

# Metalog. Open-source data

**Task 1. Exploratory analysis of the selected dataset**

Please write your ideas in the chat about what you can check to understand your dataset better ...

https://metalog.embl.de

# Metalog. Open-source data

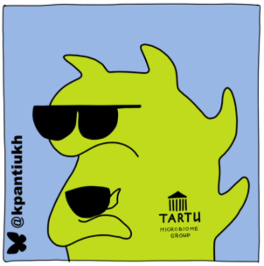
**Task 1. Exploratory analysis of the selected dataset**

**General.** How many samples are included in the dataset? How many studies does the dataset contain? What are the minimum, maximum, and average numbers of samples per study?

**Dataset balance and bias & missing data structure.** Which metadata fields are available, and are they consistently present across all samples and studies? …
Are any metadata variables strongly correlated or effectively duplicating the same information?

**Temporal coverage.** Do we have studies with multiple time points for the same samples?
**Technical heterogeneity.** Assess whether sequencing platform, library preparation, or read length vary across studies.
**Redundancy at the sample level.** Assess whether the dataset contains duplicate samples.

https://metalog.embl.de



# Metalog. Open-source data

**Task 2. Formulate a question that can be answered with the data**

Please write your ideas in the chat ...

*For each selected question, clearly specify the **response variables**, the explanatory **variables of interest**, and any **covariates** or confounders that need to be controlled for.*

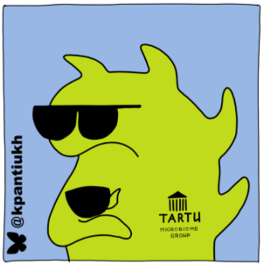
https://metalog.embl.de

# Metalog. Open-source data

**Task 2. Formulate a question that can be answered with the data**

1. Does microbiome community composition differ between sampling locations, different geographic regions, different body sites?
2. Is there a relationship between environmental conditions such as temperature, salinity, or oxygen concentration and the observed microbiome structure?
3. Do samples collected in different seasons or months show systematic differences, suggesting temporal or seasonal effects on the microbiome?
4. Does microbiome community composition differ between sampling protocols?

https://metalog.embl.de

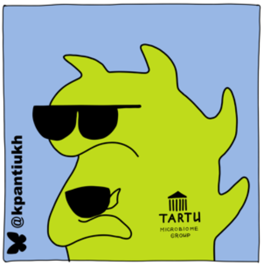


# Metalog. Open-source data

**Task 3. Select a subset of data for your analysis**

1. Select one study or a small set of studies that contain the metadata required to answer your question.
2. Aim for a balanced dataset to avoid confounding signals. For example, if you include multiple studies, they should have comparable structure (such as age, gender distribution, or other relevant variables).
3. Define case and control groups. Verify that these groups have similar structures, or construct matched case–control pairs where appropriate.

[https://metalog.embl.de](https://metalog.embl.de)



# Metalog. Open-source data

**Task 4. Answer your question**

1. Compare samples from the case and control groups.
2. Analyse and compare alpha-diversity distributions between the groups.
3. Calculate beta diversity and perform a PCA to determine whether the groups differ.

https://metalog.embl.de

# Metalog. Open-source data

**Task 1.** Exploratory analysis of the selected dataset

**Task 2.** Formulate a question that can be answered with the data

**Task 3.** Select a subset of data for your analysis

**Task 4.** Answer your question

Dataset:

https://metalog.embl.de

## HW

- Well-documented and reproducible code, including clear variable names and informative comments
- A concise written report summarizing your analyses, key results, biological interpretations, and limitations of the dataset

**Deadline:** 12-02-26