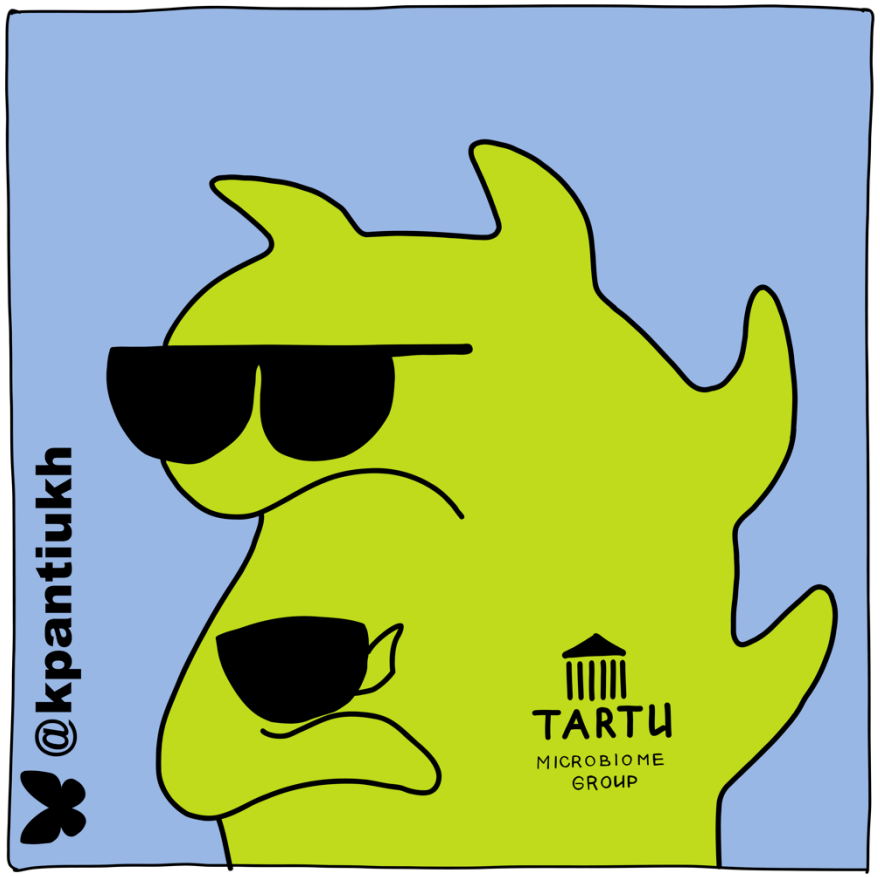


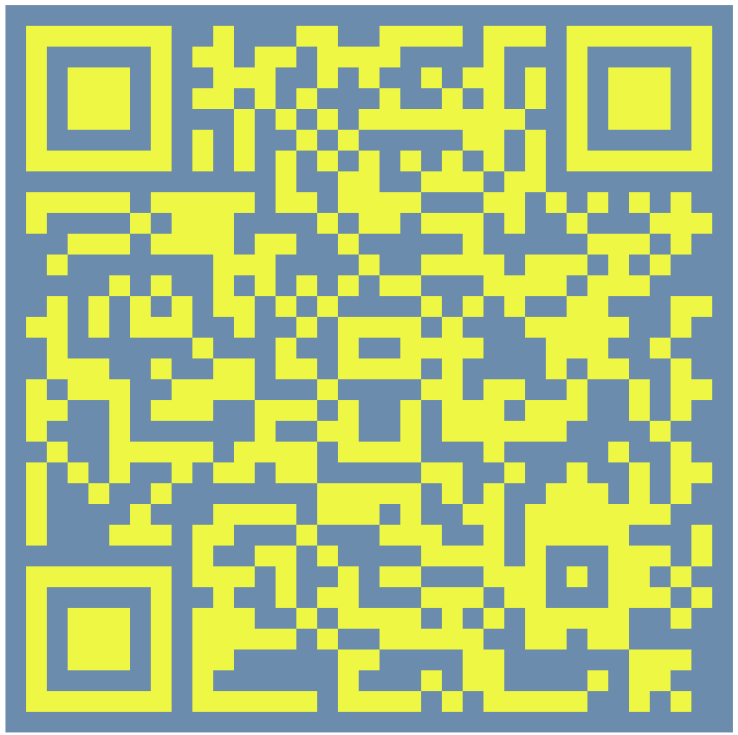
Data analysis
Lecture 2

Metalog data

Kateryna Pantiukh

pantiukh@ut.ee

GitHub



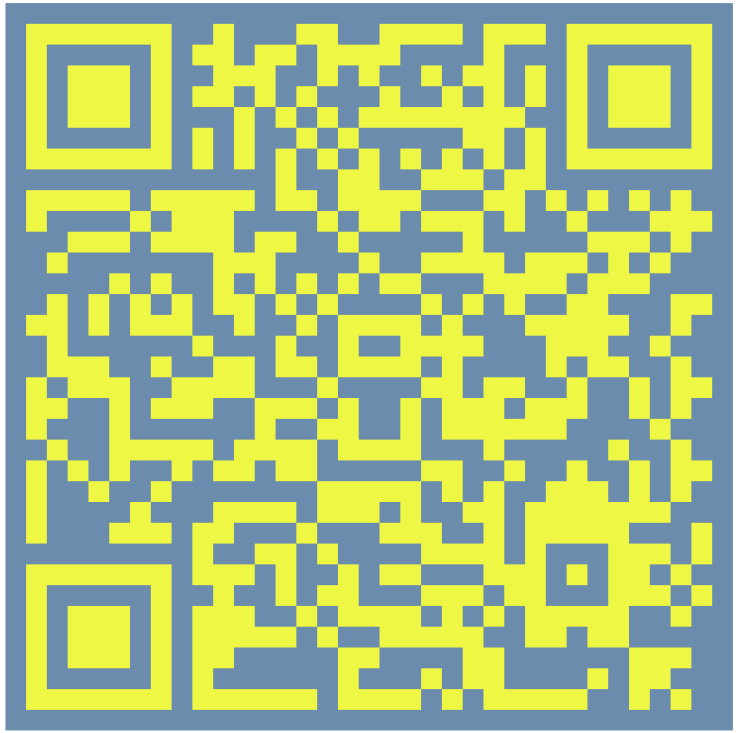
Please set up your GitHub repository and code editor (Visual Studio Code)

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

Discussion

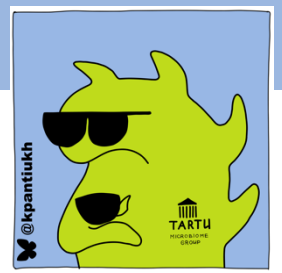
What was your research question from previous HW?

Metalog. Open-source data

https://metalog.embl.de

Metalog is a repository of manually curated metadata for metagenomic samples

EMBL European Molecular Biology Laboratory



MWAS

Task 1. Prepare the abundance table

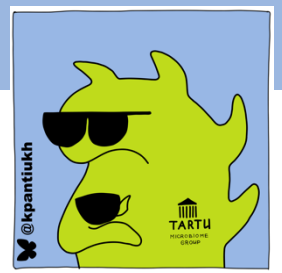
Decide on the design of your association study, including the taxonomic level to analyse and the criteria for selecting taxa (e.g., minimum prevalence or abundance thresholds).

Analysis:

- Filter out species with very low prevalence or extremely low abundance across samples.
- Handle zeros appropriately (e.g., add a small pseudocount for log transformation).
- Apply **centered log-ratio (CLR) transformation** to account for compositionality.
- Ensure the table aligns with the metadata for downstream regression analyses.

Conclusions

- Discuss the following points:
- Why you select particular taxanomic level?
- Why you select particular prevalence and abundance treshhold?



MWAS

Task 2. Perform MWAS using linear regression

Decide on the design of your association study, including the taxonomic level to analyse and the criteria for selecting taxa (e.g., minimum prevalence or abundance thresholds).

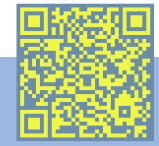
Analysis:

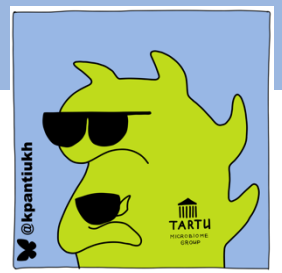
- Use CLR-transformed abundance data as the outcome or predictor (depending on modeling choice).
- Fit linear regression models for each species to test for associations with group membership.
- Adjust for potential confounders (e.g., age, sex, batch effects).
- Correct for multiple testing (FDR or Bonferroni).

Visualisation - Manhatten plot

Conclusions:

Highlight species positively or negatively associated with groups, formulate statement about this species, like for example – "*Alistipes putredinis* is positively associated with female infertility status (p-value=X), indicating higher relative abundance in diagnosed individuals compared with controls, however the estimated effect size is small (beta=X)"





MWAS

Task 3. Convert abundances to presence/absence and perform logistic regression MWAS
Analyze species associations using binary presence/absence data to account for sparse taxa.

Analysis:
- Convert the filtered abundance table to binary (1 = species present, 0 = absent).
- Fit logistic regression models for each species to test for associations with group membership.
- Adjust for potential confounders (e.g., age, sex, batch effects).
- Correct for multiple testing (FDR or Bonferroni).

Visualization - volcano plot

Conclusions
- Highlight species positively or negatively associated with groups, formulate statement about this species.
- Compare linear regression results and logistic regression results.

Metalog. Open-source data

Task 1. Prepare the abundance table

Task 2. Perform MWAS using linear regression

Task 3. Convert abundances to presence/absence and perform logistic regression MWAS

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: 23-02-26