Coding Assignment 03
Title: MWAS -Microbiome Wide Association Study.

Dataset:
https://metalog.embl.de
Metalog is a repository of manually curated, open-source metadata for metagenomic samples. It includes rich contextual information such as clinical and demographic data for human subjects, as well as environmental parameters for non-human samples. Released in 2025, Metalog is actively used by researchers to gain new insights into microbiome research.

Overall Goal:
The objective of this assignment is to extend the previous analysis of microbiome diversity across the defined groups to a microbiome-wide association study (MWAS). Use the same research questions and group definitions as in the previous task. Investigate whether any bacterial species are significantly associated with one of the groups.

Task 1. Prepare the abundance table

Objectives
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.

Visualizations
Propose and generate appropriate visualizations to support your analysis. These may include, for example:
- Boxplots of transformed abundances to check distributions
- Pie chart of filter out and selected species

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

Task 2. Perform MWAS using linear regression

Objective
Identify associations between continuous abundance measures and predefined groups.

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).

Visualizations:

Propose and generate appropriate visualizations to support your analysis. These may include, for example:

- Manhatten plot
- Boxplots showing abundance distributions for top associated species

Conclusions:
Formulate the following points:

- 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)"

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

Objective
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).

Visualizations:
Propose and generate appropriate visualizations to support your analysis. These may include, for example:

- Volcano plot

Conclusions:
Formulate the following points:

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

Final Deliverables:

- Well-documented code (comments and clear variable names required)
- A short written report summarizing results, interpretations, and limitations

IMPORTANT NOTE: The questions and visualizations proposed in this assignment are recommendations and are not compulsory.