Course Data Description

In this course, you will develop machine learning approaches based on the MetaCardis cohort, a large-scale multi-omic European study. This cohort aims to identify novel microbial and metabolomic biomarkers for ischemic heart disease at early, middle, and late clinical stages and includes healthy individuals, individuals with dysmetabolic morbidities, and individuals with heart disease from three states (Denmark, Germany, and France). For more information on the cohort and its main finding, it is advisable to read its two main papers:

1. Fromentin, S., Forslund, S.K., Chechi, K. et al. Microbiome and metabolome features of the cardiometabolic disease spectrum. Nat Med 28, 303–314 (2022). https://doi.org/10.1038/s41591-022-01688-4
2. Forslund, S.K., Chakaroun, R., Zimmermann-Kogadeeva, M. et al. Combinatorial, additive and dose-dependent drug–microbiome associations. Nature 600, 500–505 (2021). https://doi.org/10.1038/s41586-021-04177-9

In this study all subjects have donated a single sample which is identifiable using the SampleID column in the cohort tables. Please be aware that while stool metagenomic samples are available from all cohort participants, metadata or metabolomic information may be unavailable for some subjects due to Danish regulatory and privacy laws as well as missing information.

All data is available in the course website on moodle.

Metadata file

Metadata file (**metadata.csv**) contains general info and lifestyle data on each participant. Note that some patients may have missing values, however, all patients have PatientGroup and CENTER information. The data structure:

- SampleID – unique identifier for each subject
- PatientGroup – patient stratification according to health conditions. **Note that this column will not be available in the test data**. The groups are coded according to these definitions:
  - 1 - Subjects with MS (Metabolic syndrome) without T2D (Type 2 Diabetes) or CAD (Coronary Artery Disease).
  - 2a - Subjects with severe obesity, not eligible for bariatric surgery, without diabetes or CAD
  - 2b - Subjects with severe obesity, eligible for bariatric surgery, without diabetes or CAD
  - 3 - Subjects with T2D without CAD
  - 4 - Subjects with a first acute CAD event

    - 5 - Subjects suffering from chronic CAD without heart failure
    - 6 - Subjects suffering from chronic CAD with heart failure (left ventricular ejection fraction <45)
    - 7 - Subjects suffering from chronic congestive heart failure without CAD
    - 8 - Control group, subjects exempt from CAD, T2D or MS.
- Gender – 0: Males, 1: Females
- AGE – Age
- CENTER – Country of recruitment (France, Denmark, or Germany)
- SMOKE – Current smoker (0 – no, 1 - yes)
- pa_work_2cl – Does the patient work out.
- DDS - Dietary diversity score

Omic data

In this cohort we have both stool metagenomic data produced by shotgun sequencing and metabolomic data. All these feature tables have similar structure:

- SampleID – unique identifier for each subject.
- Microbime/metabolome feature – describe the abundance of each taxa/metabolite.

Metagenomic data

The file (microbiome.csv) contains a taxonomic table with count data representing the abundance of microbial species for each subject. The data is at the species level. If you need to group bacteria at higher taxonomic or phylogenetic levels, contact us for the necessary trees. Keep in mind that, as discussed in class, raw count data is not directly informative and some form of transformation (e.g., relative abundance, CLR) is required before analysis.

Metabolomic data

Here we have metabolomic data based on serum lipoproteins (serum_lipo.csv).