Group Project NUTRITION - hEalTHy Food!

Description of challenge:

You are part of a new unit in the ETH Mensa called "hEalTHy Food". This unit's main task is to optimize student wellbeing via analysis of their gut microbiota to derive eating and drinking habits, and to evaluate their impact on gastrointestinal health.

Your team was asked to analyse which biomarkers and host characteristics are most associated with different eating and drinking habits. To this goal you are given a dataset containing fecal microbiome sequences ("Sequences") and an additional metadata table with additional information on the students' samples ("Sample metadata"). To identify predictive markers for selected eating/drinking habits you should examine the relationship between the fecal microbiome, host covariates and diet.

The Mensa chefs are already very interested in learning about your approach and the positive and negative results of the analysis so that they can decide what to cook tomorrow...

Some specific questions*:

1. Is there a bias in the data cohort you have received? If yes, describe the bias and potential implications to the obtained insights of this analysis.
   - Hint: Explore the metadata to find out and create some visualisations demonstrating potential data biases.
2. What are the most prevalent food and drinking habits in your cohort? Is there scientific evidence that these have an influence on the human gut microbiome?
   - Hint: This is a good moment to practice some pandas summary statistics on your metadata.
3. Explore the microbial communities of patients with different alcohol consumption habits. What characteristics do you observe? What about other dietary habits?

Information on provided datasets:

1. **Sequences**: The sequences originate from the V4 region of the 16S rRNA gene and were sequenced on an Illumina MiSeq machine. The sequence files contain demultiplexed paired-end sequences and are provided as a QIIME 2 Artifact (Semantic type: `SampleData[PairedEndSequencesWithQuality]`).
2. **Sample metadata**: The metadata table contains additional information on each collected sample. Each row represents a unique sample that can be identified with its unique `sampleid`. The remaining columns depict general participant features (e.g. Age, Gender) marked with a prefix "`GEN_`" and nutrition related features (e.g. Frequency of consuming alcohol) marked with a prefix "`NUT_`".

Where to get the data:

1. Sequences: https://polybox.ethz.ch/index.php/s/AsLORIvUbwgBWTq
2. Sample metadata: https://polybox.ethz.ch/index.php/s/QqbHeUIpIR0okB8

Note: use wget to download these links in your analysis notebooks for fully executable analyses.

Questions? Your course instructors are here to help. Schedule a meeting with your instructors during in-class group work sessions.

***Note on questions**: *this is not a list of the only questions that you can/should answer about this dataset. All group semester projects are expected to incorporate most analysis topics covered in the course when possible/appropriate (see Group Semester Projects assignment document for more details). This is merely a list of specific questions that should be answered or attempted in addition to the "standard" analyses covered in the*

course. Not everything will work or give a clear result! But that does not mean that you should not try. Your goal as scientists is to figure out why.