Gut microbiome signatures of Colorectal Cancer

Background

There have been several studies investigating the link between the gut microbiome and colorectal cancer (CRC). We recently published a meta-analyses and have shown a link between the gut microbiome and CRC across a variety of cohorts from around the world:

> Wirbel *et al* (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine* (25):679-689.
>
> https://www.nature.com/articles/s41591-019-0406-6

We showed that CRC patient can be classified based on the species composition in their gut microbiome. We also showed that models trained on one cohort can classify patients from a different cohort even from a different continent, suggesting that gut microbiome signatures of CRC are global. See Figure 3a from the article.

Data description

You are given a zip file `CRC_meta.zip` with 8 tab-delimited tsv files. They are:

```
AT-CRC_species.tsv
CN-CRC_species.tsv
DE-CRC_species.tsv
FR-CRC_species.tsv
IT-CRC-2_species.tsv
IT-CRC_species.tsv
JP-CRC_species.tsv
US-CRC_species.tsv
```

Each file represents gut microbiome species compositions of several individuals in one cohort. The first column specifies the condition (control or CRC) and the rest of the columns represent relative abundances of many species. The first row is a header where the names of the species can be obtained starting from 2nd column. All cohorts have the same number of columns, so the features match across cohorts.

Problem statement

In the publication above, we used a LASSO model for this classification. Based on what you have learnt in the course, you will use the species composition data from this study and build a classifier (ideally different from LASSO) to identify a microbiome signature that distinguishes a cohort of patients between those with colorectal cancer (CRC) and healthy controls.

The specific tasks for you are:

1. Can you build a classifier that robustly classifies CRC patients for each of the cohorts? How does it compare with the classifier performance in the publication (Figure 3a, diagonal values)?
2. Can the classifier trained on one cohort robustly classify CRC patients from another cohort? How does it compare with the classier performance in the publication (Figure 3a, non-diagonal values)?
3. From your classifier, can you identify biomarker species that can distinguish CRC patients from control individuals?

Some useful pointers

1. Always be mindful of information leaking from training data to the test data. E.g., if you are optimizing a hyperparameter during cross validation, use nested cross validation.
2. Relative abundances normally sum to 1. In the given dataset, most rows will sum to a value near 1 but not necessarily 1. This is because there are species that we do not know of.
3. It is common to transform the relative abundances before using them for classification or biomarker discovery. Feel free to try different data transformations.