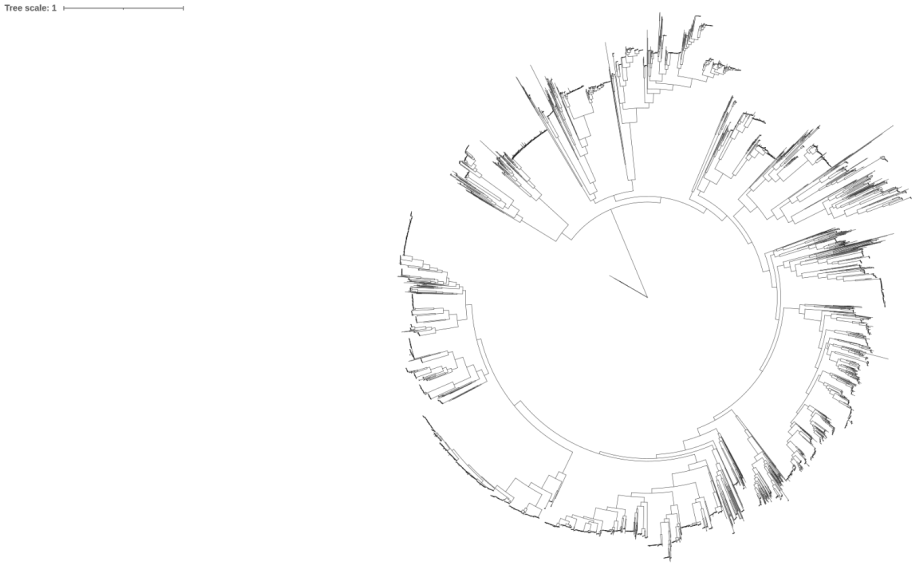
Small Data Analysis for Integrative Microbiome and Metabolomics

Data analysis:

- iTOL phylogenetic tree
- PCAs with Abundance Pie charts
- Sample data correlation matrices
- OTU vs Sample data correlation matrices

iTOL phylogenetic tree

- Too many clusters for data analysis



Pre-processing of the data

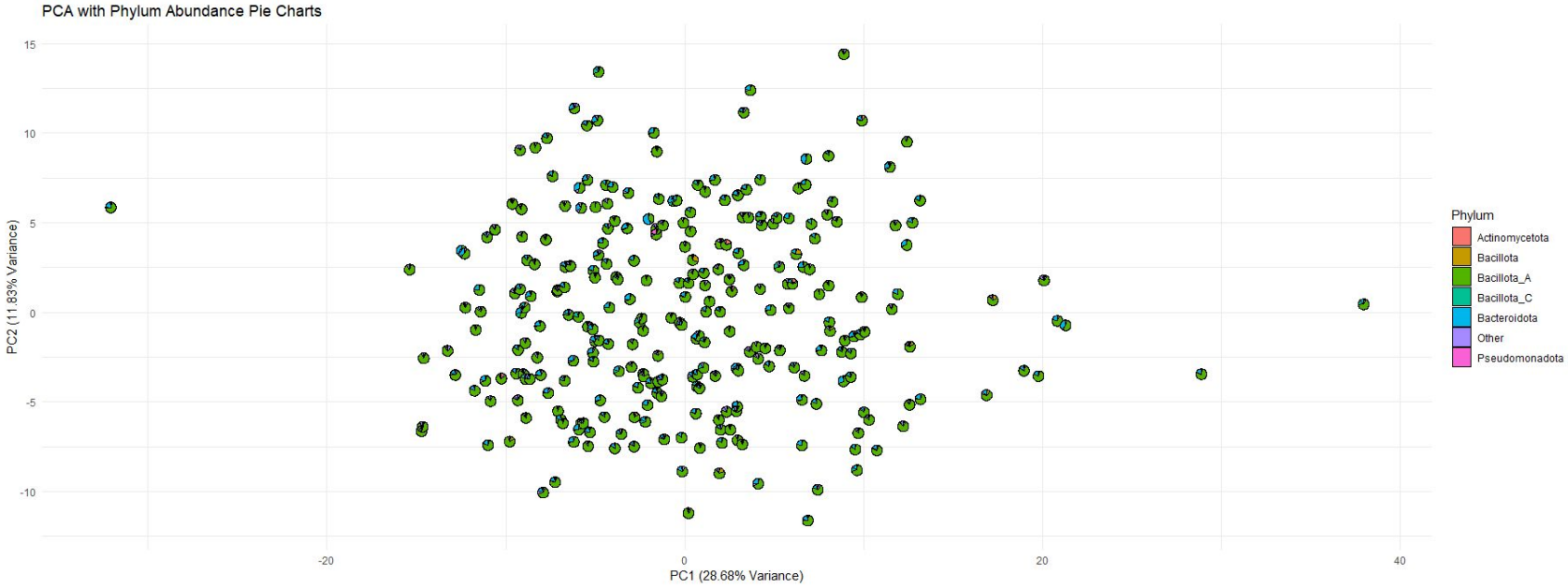
Sample data:

- The samples with NA variables were removed
- Take out the sample variable with constant values across the samples (no information)
- Separate the sample data into 4 dataframes (overall, food, nutrients and metabolomics)

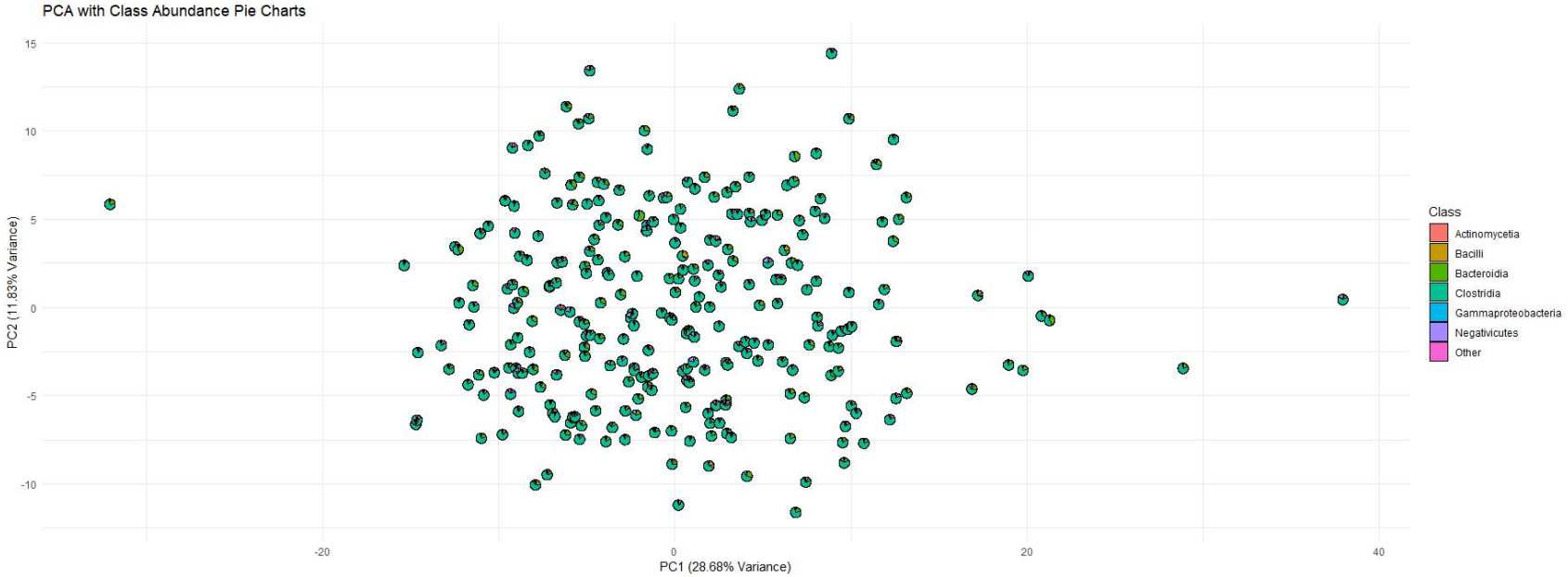
Otu table:

- Normalized the otu table to make sure every sample sums to 1
- Making a taxa table using the relative abundances per cluster of the otu table
- Remove the samples that have been dropped from sample data
- Creating 4 dataframes with abundance per phylum, class, order and family

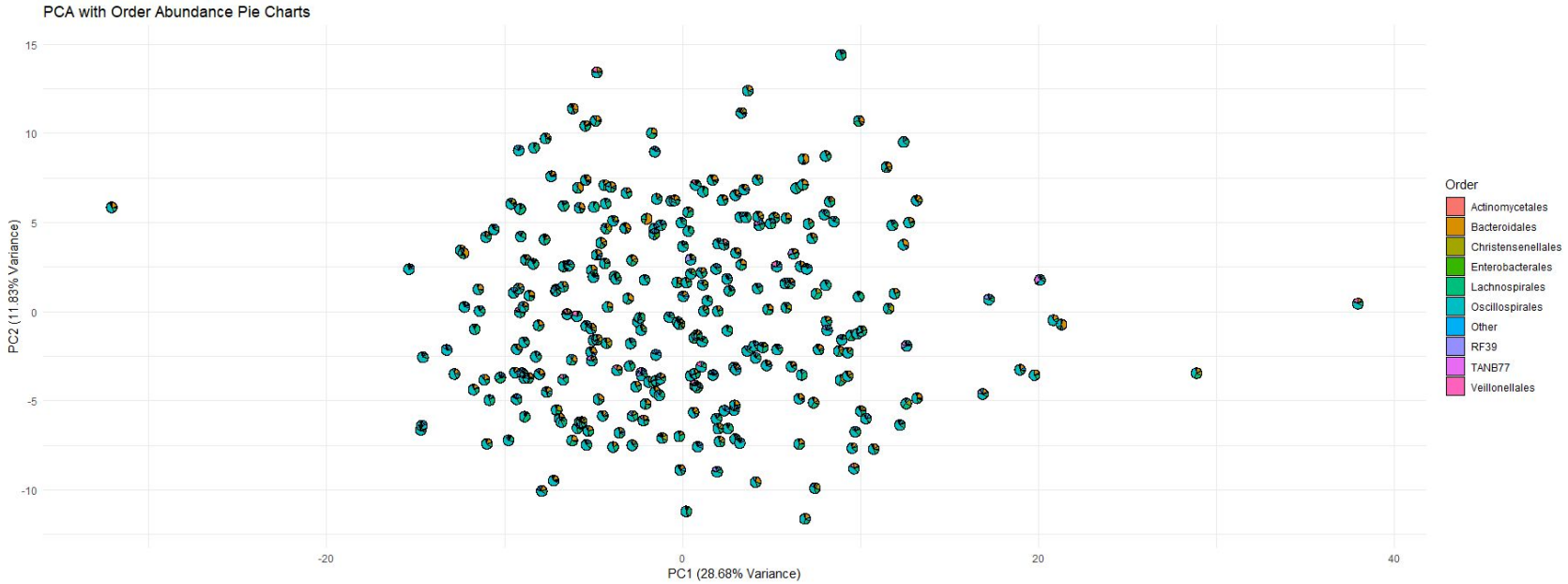
PCAs with Abundance Pie charts



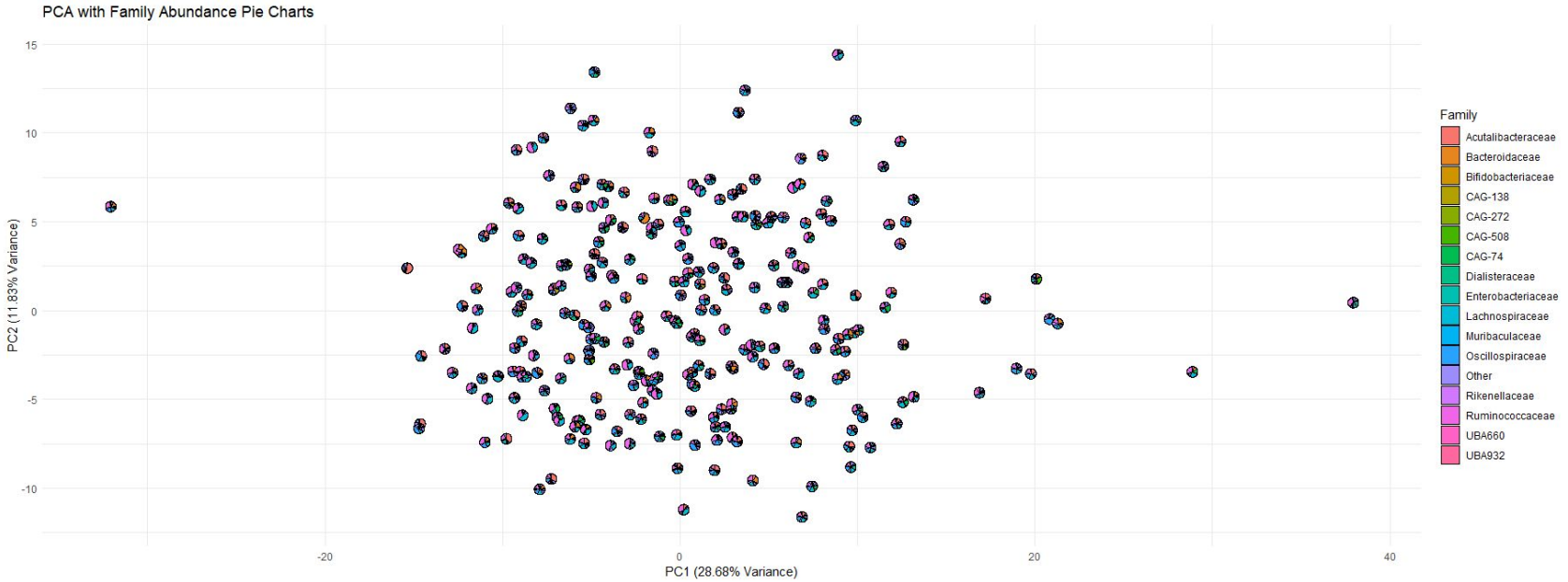
PCAs with Abundance Pie charts



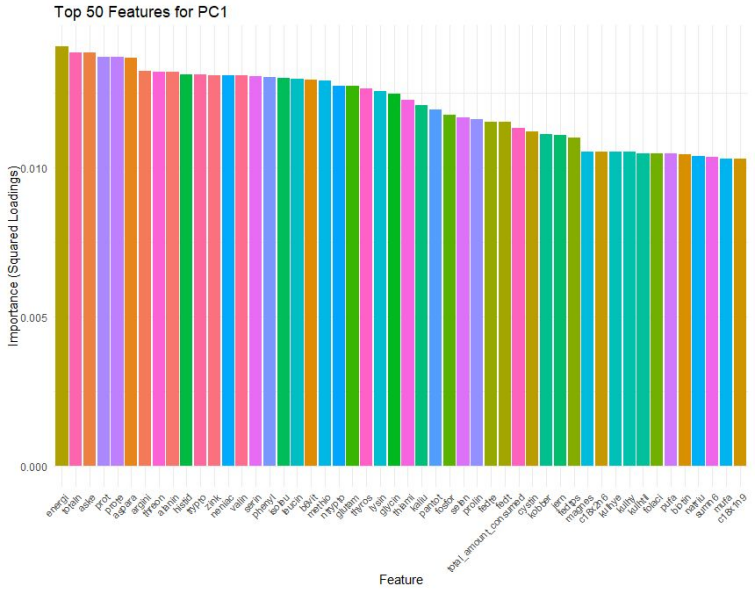
PCAs with Abundance Pie charts

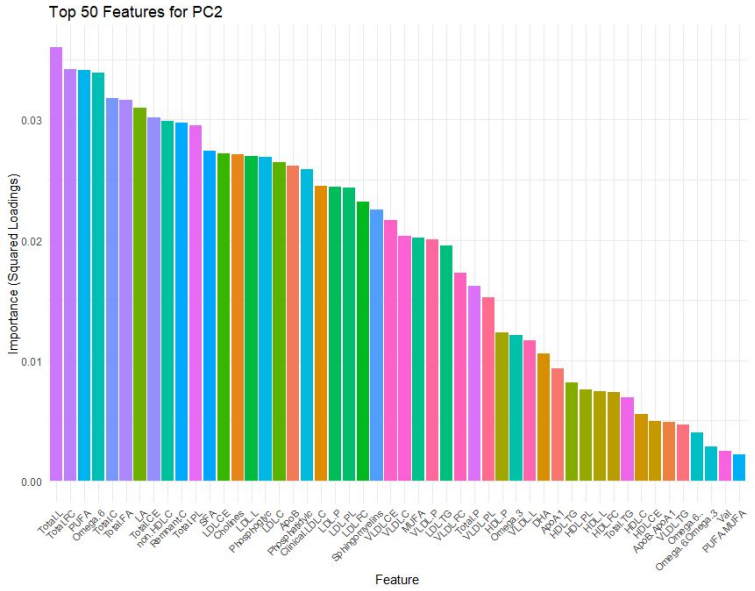


PCAs with Abundance Pie charts



PCAs with Abundance Pie charts





PCAs with Abundance Pie charts

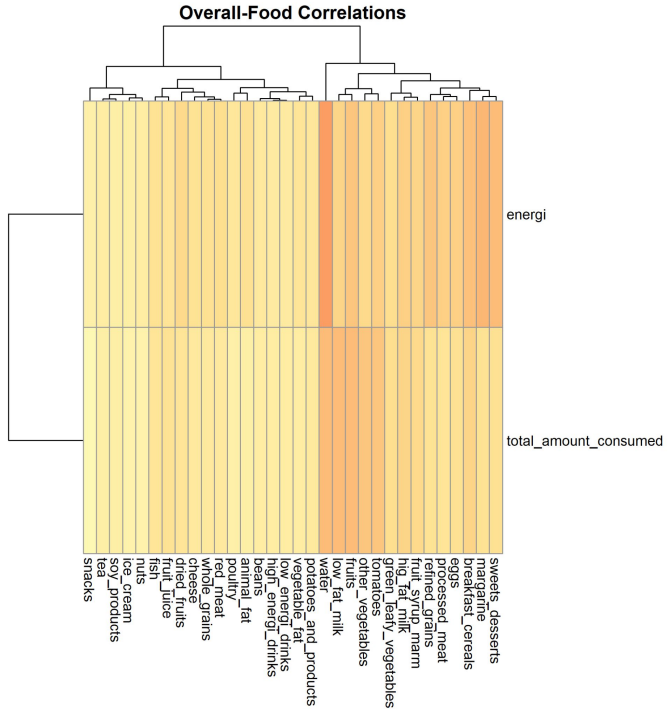
The PCAs using all the sample data have a high explained variance (PC1 11,83%, PC2 28,68%)

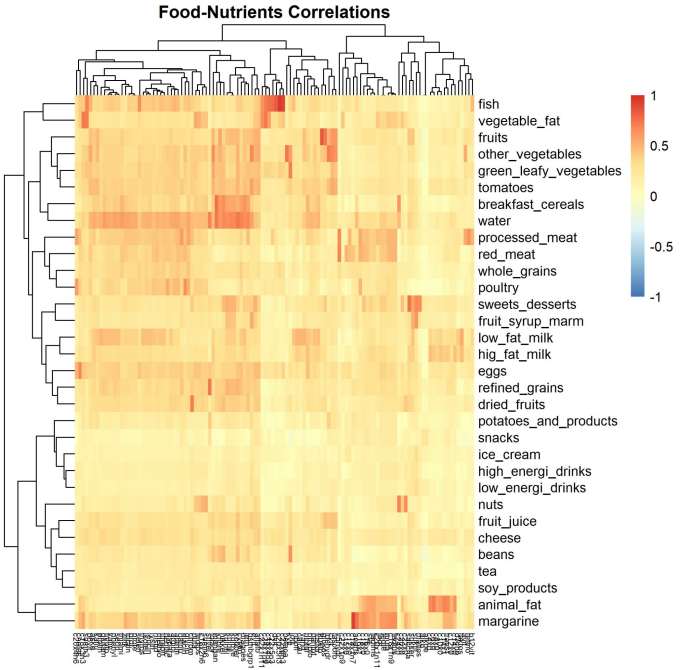
However the features importance shows that the weights of each sample variable is really low (<0.015) for PC1 no features are more important than the others

For the PC2, the weights are a bit higher (around 0.02-0.03 for the highest ones) but explain less variance than PC1

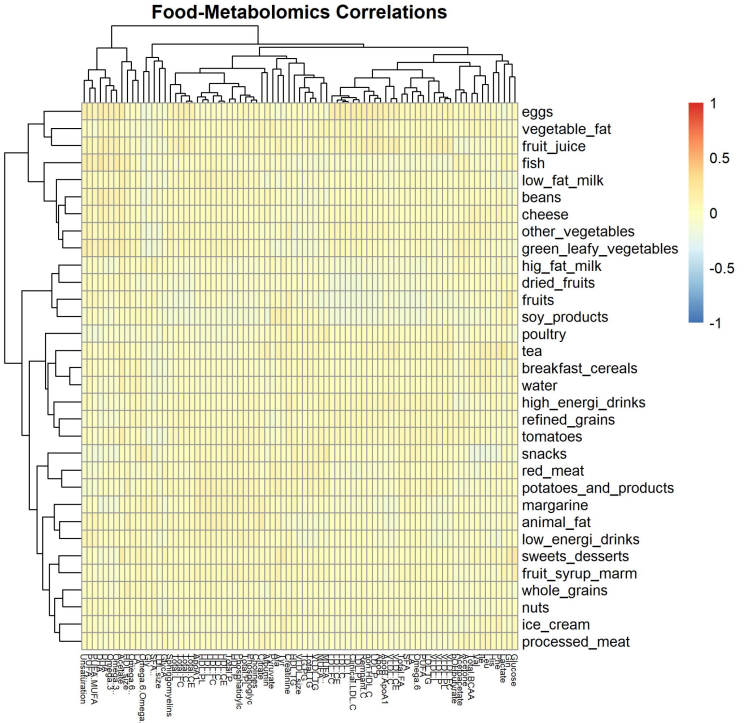
Overall the PCA needs refining (choosing specific sample variables as some might introduce noise that prevent the important features to be detected. (using features selection)

Sample data correlation matrices

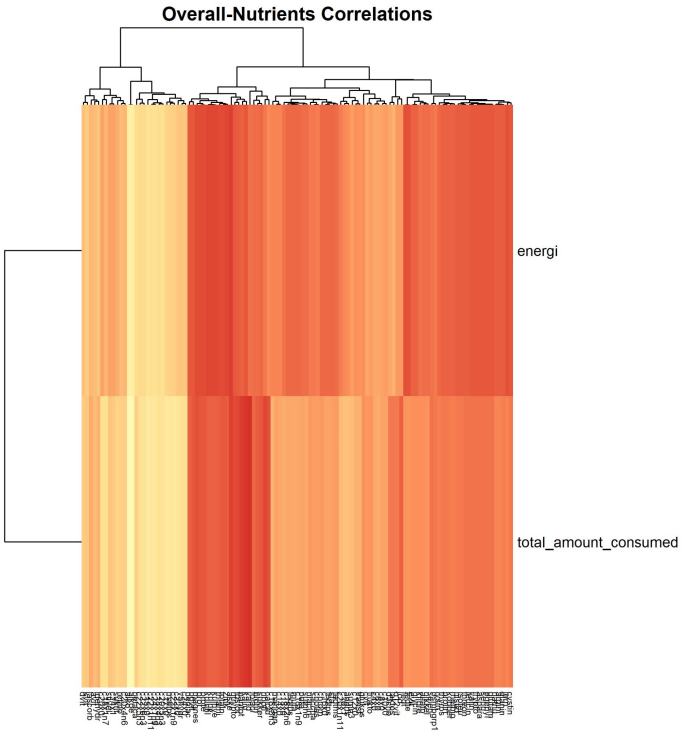


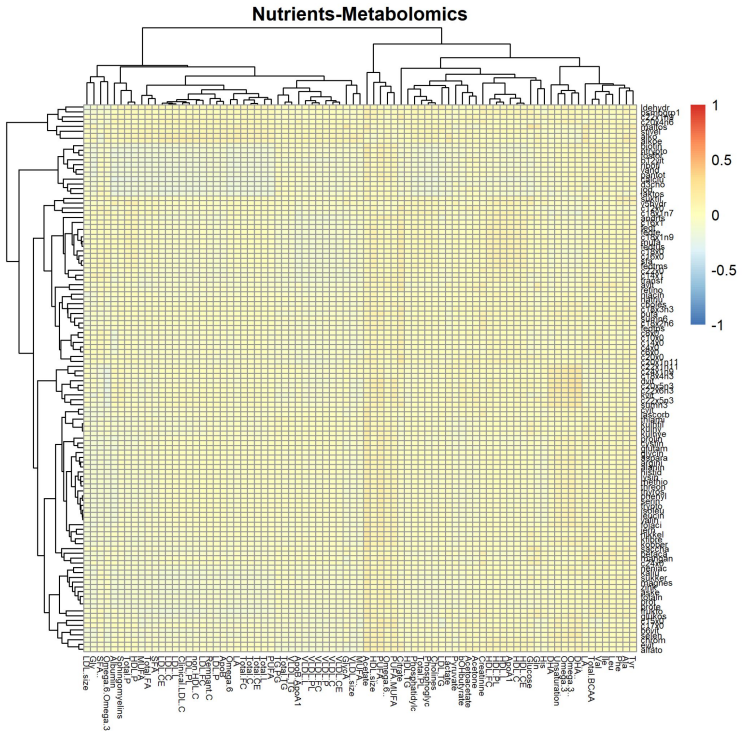


Sample data correlation matrices



Sample data correlation matrices





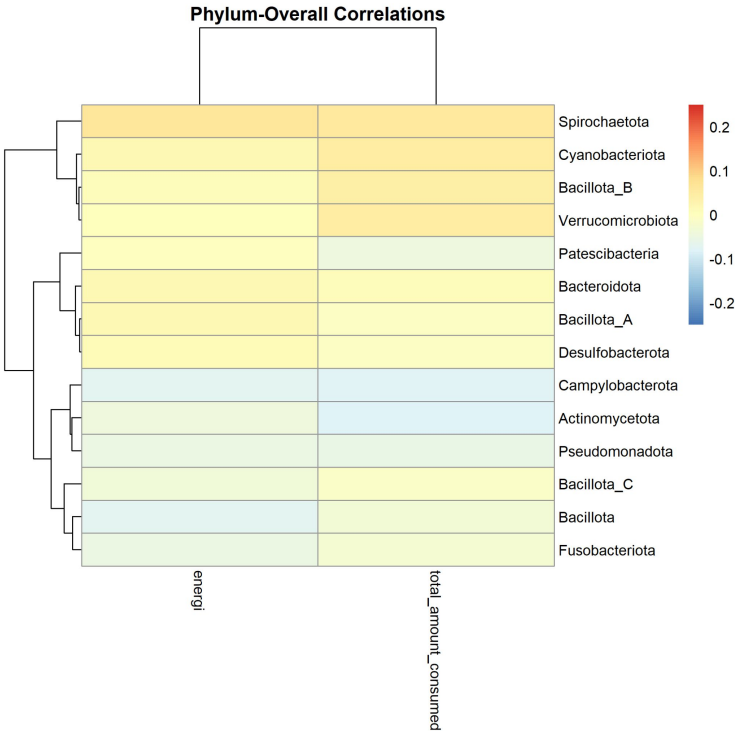
Sample data correlation matrices

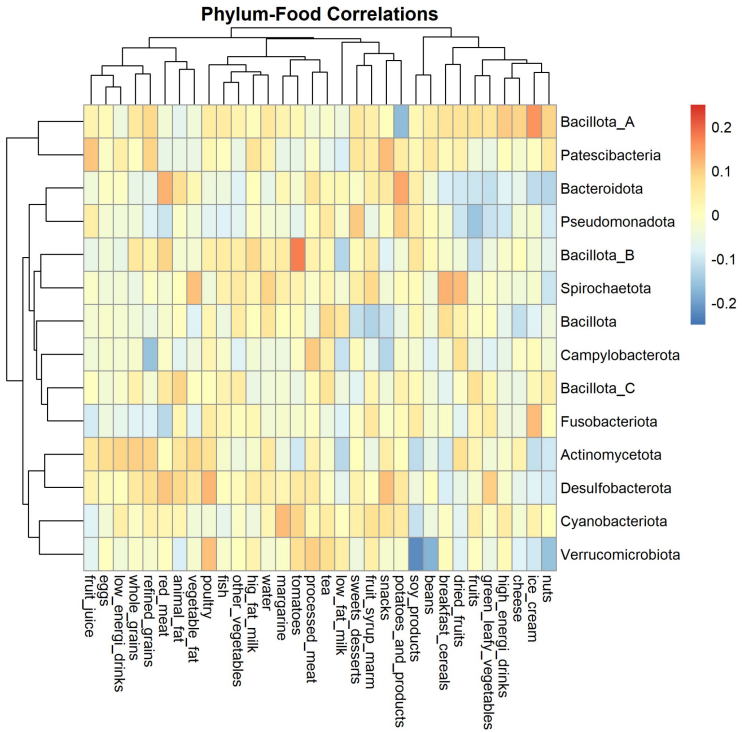
There is high correlation between the overall food and nutrients

This should be investigated more to have a better understanding of the variables that may add the same information and could be removed from the sample data

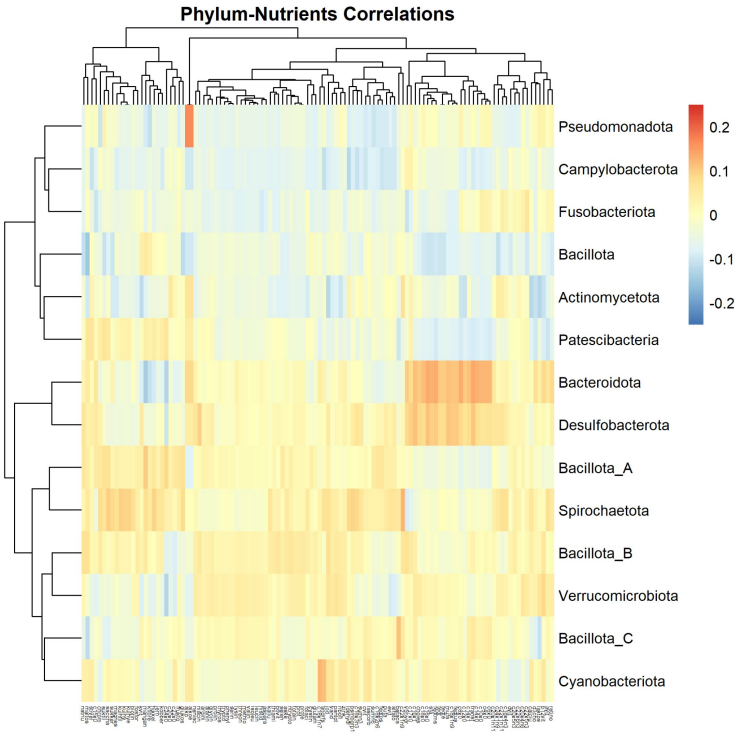
The metabolomics on the other end are not correlated at all with the other sample data, and might need more investigation

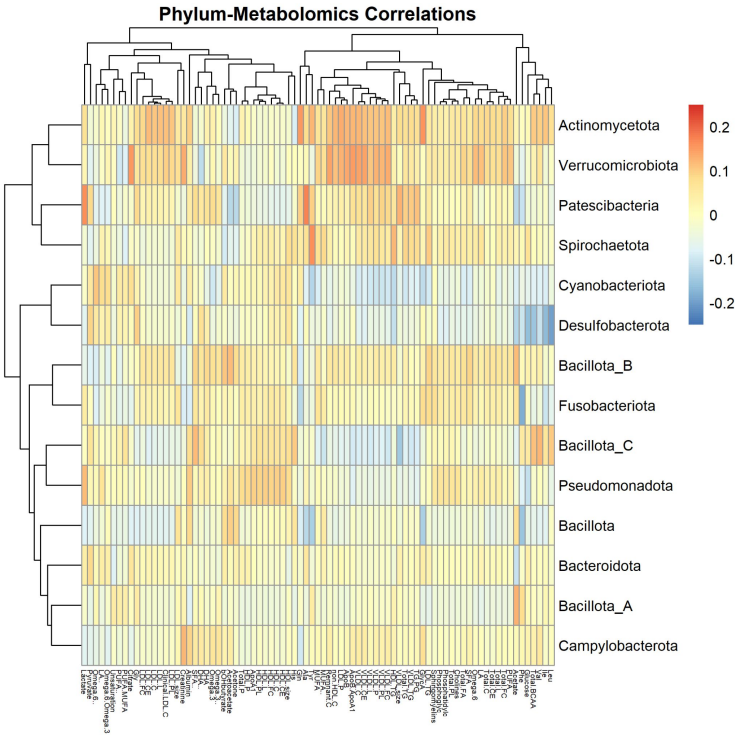
OTU vs Sample data correlation matrices



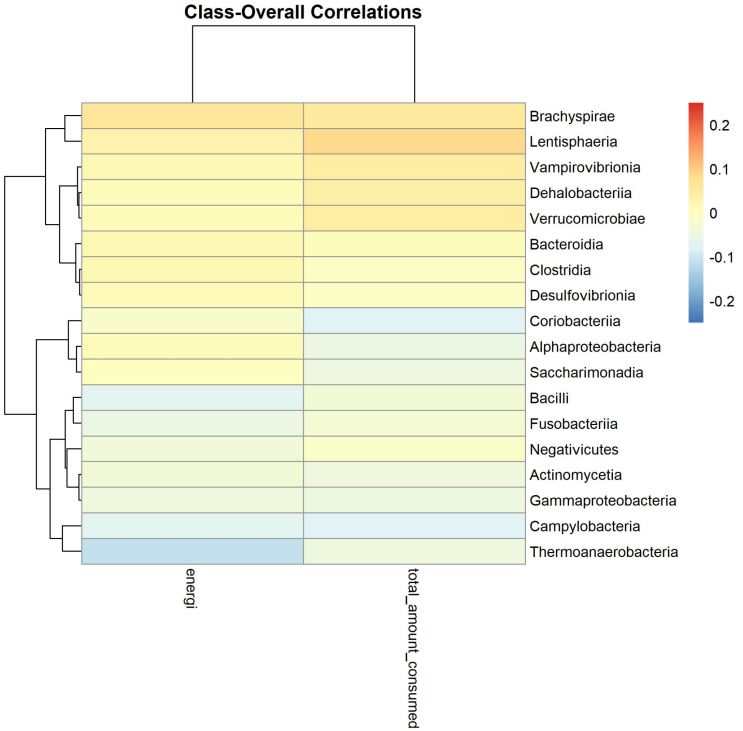


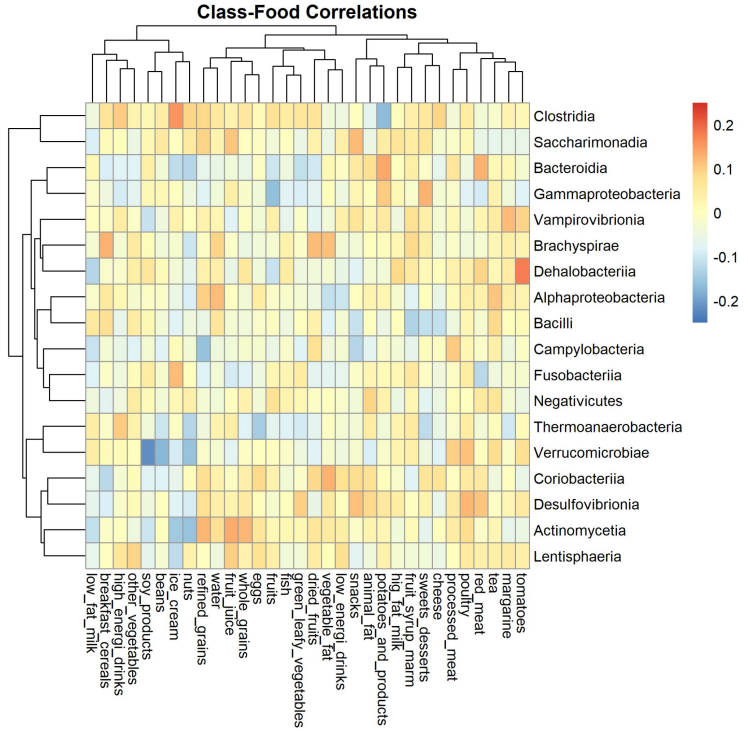
OTU vs Sample data correlation matrices



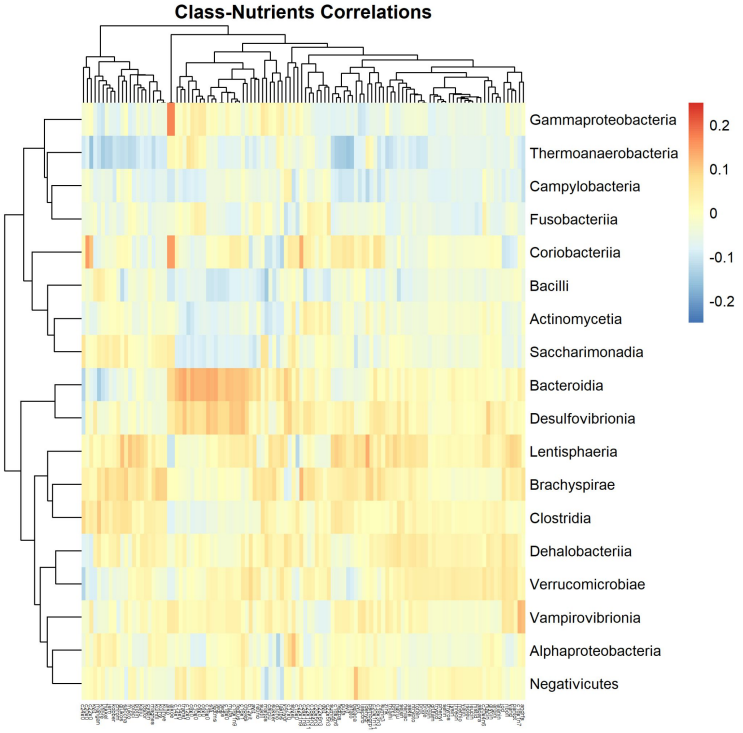


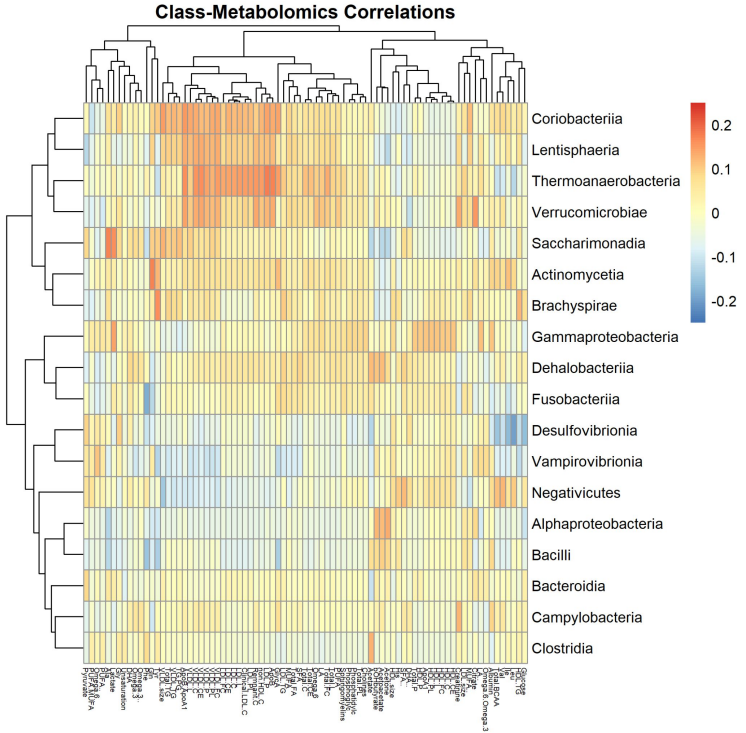
OTU vs Sample data correlation matrices



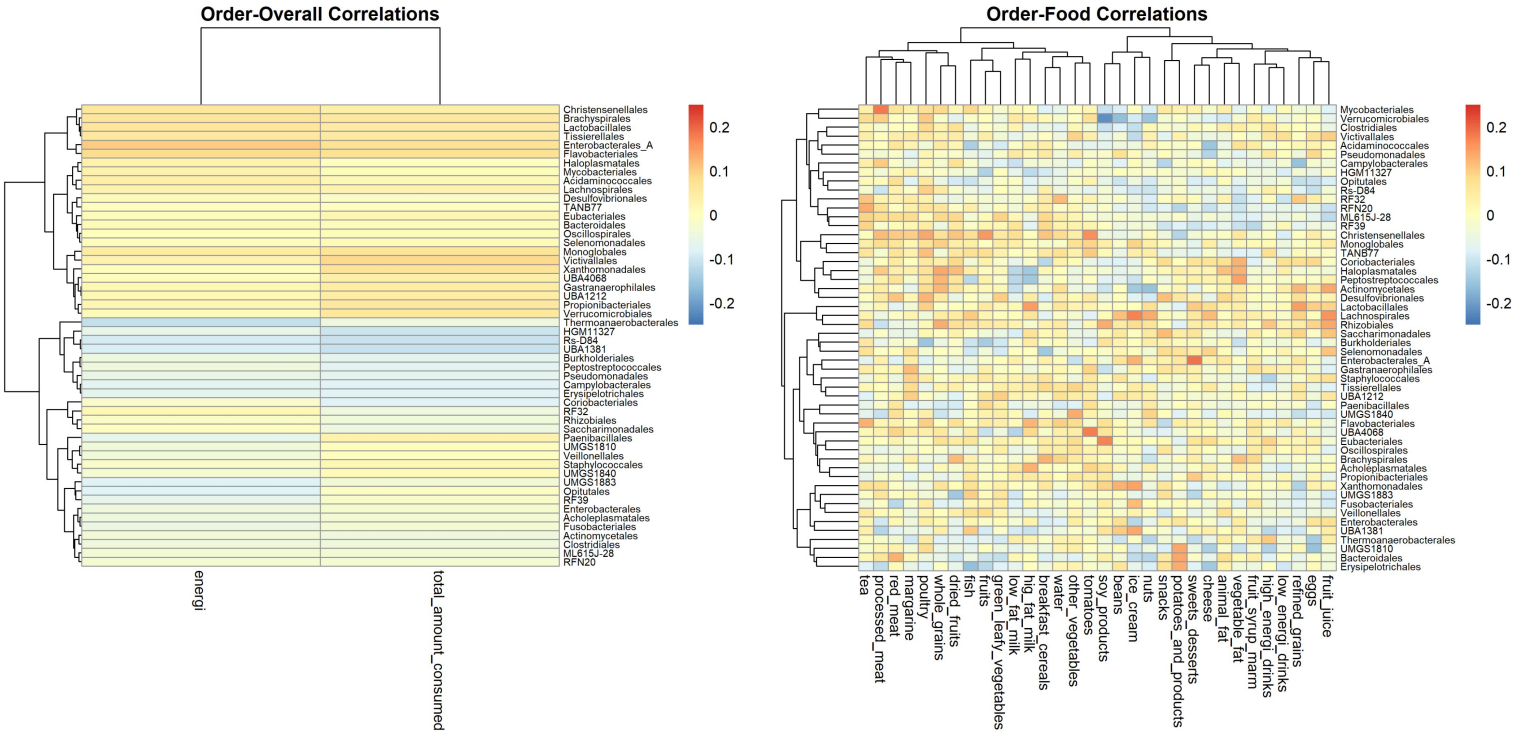


OTU vs Sample data correlation matrices

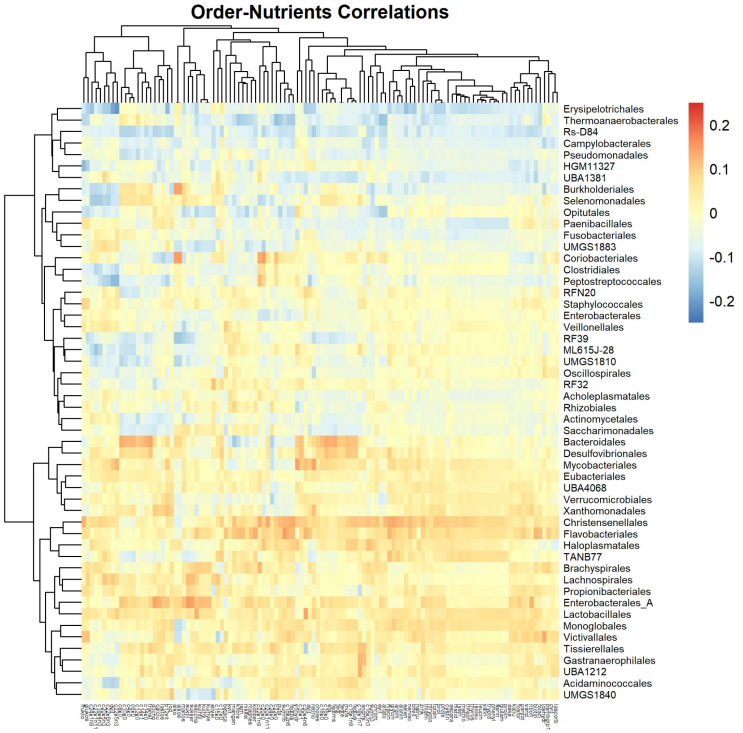


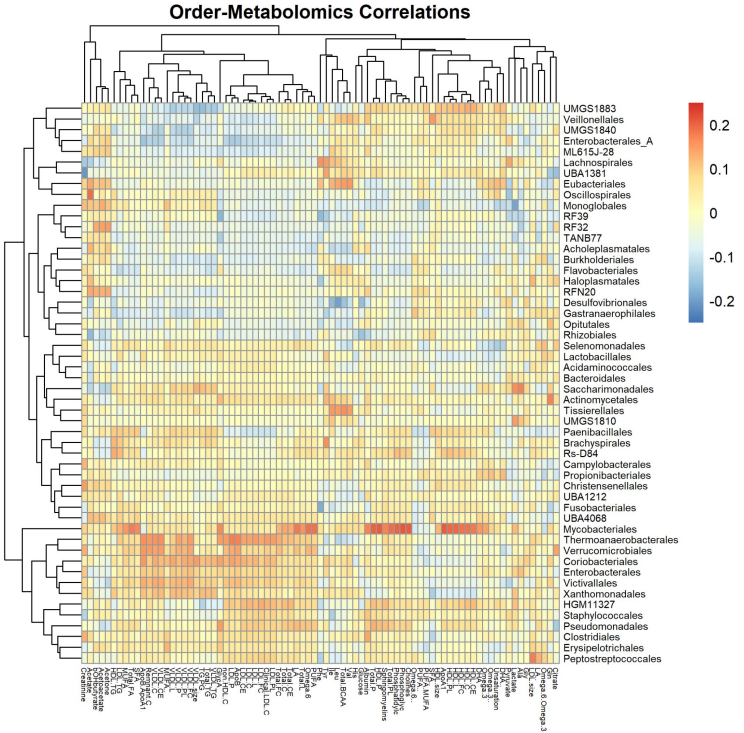


OTU vs Sample data correlation matrices

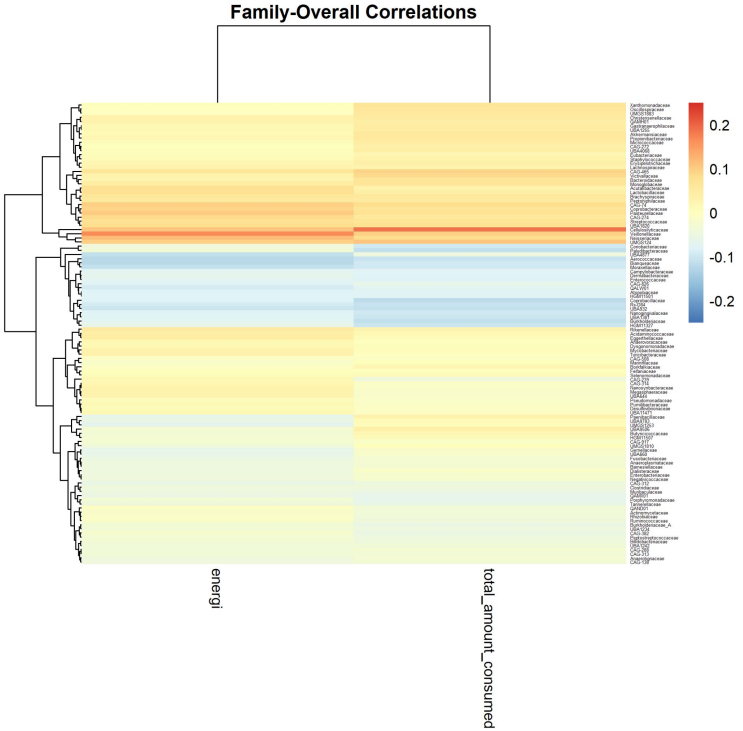


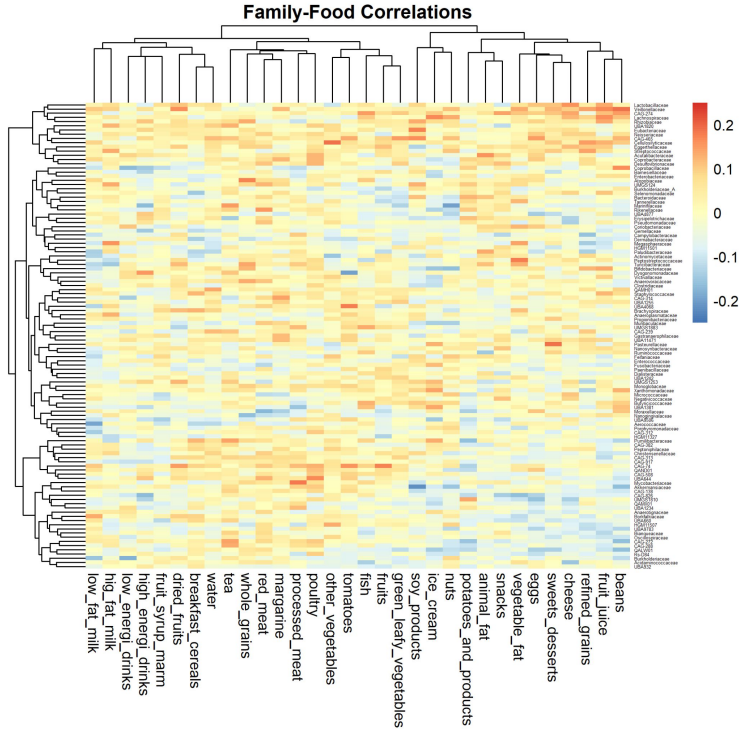
OTU vs Sample data correlation matrices



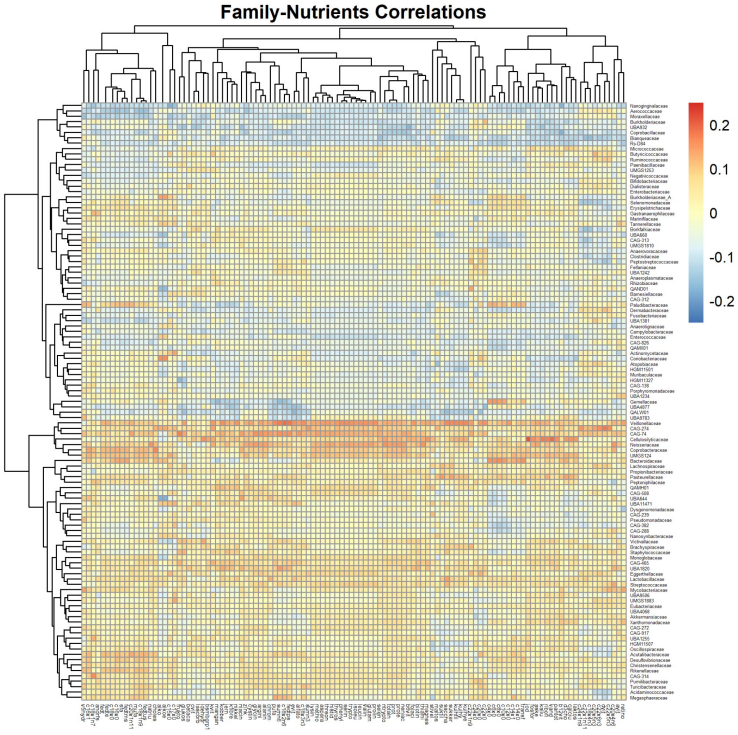


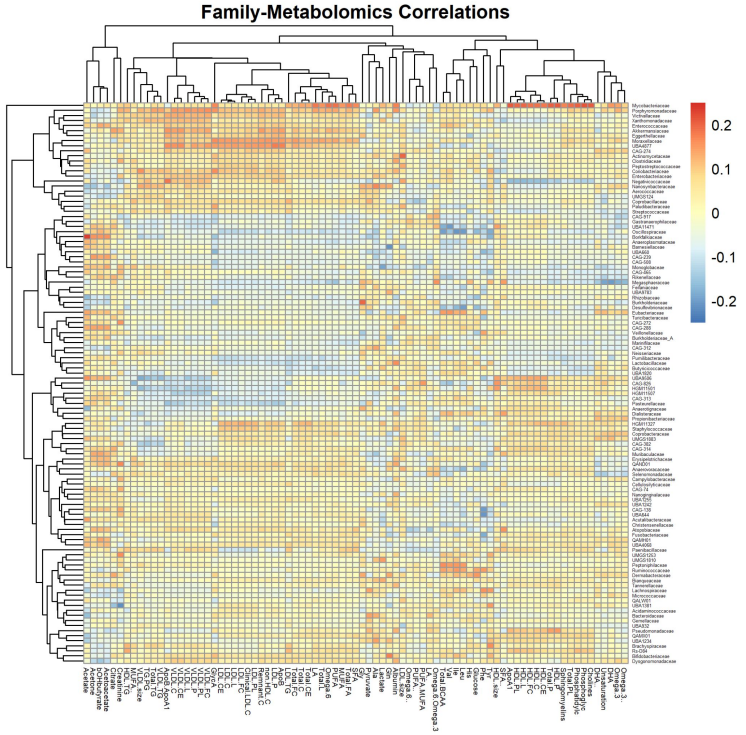
OTU vs Sample data correlation matrices





OTU vs Sample data correlation matrices





OTU vs Sample data correlation matrices

The correlations are pretty low (highest lower than 0.2)

There seems a lot of weak signals that needs to be more investigated mostly on the nutrients and metabolomics as well as on food