Statistical methods for microbiome análisis

Practical exercises

Data set Hansen et al. (2018) Gluten Diet study:

https://pubmed.ncbi.nlm.nih.gov/30425247/

1. Upload the data as a phyloseq object as follows:

data<-readRDS("hansen.rds")

1. Use phyloseq library to preprocess and describe the dataset:
   1. Number of taxa, number of samples, number of variables in metadata
   2. Focus on the metadata: number of subjects and number of time points for each subject.
   3. Summarize and plot with a histogram the variable *days\_from\_first\_collection*. Check which is the minimum and the maximum time point in the dataset
   4. Agglomerate at the species level and normalize the data to relative abundances
   5. Summarize the data at the species level: number of taxa and summary and histogram of log10(mean taxa abundances)
   6. ~~Normalize the data at species level to relative abundances~~
   7. Filter out those taxa at the species level with a mean relative abundance larger than 0.0001. Check how many species are left.
   8. Normalize the data at species level to relative abundances
   9. ~~Compute alpha diversity Shannon index and the effective number of taxa as follows:~~

~~rich<-estimate\_richness(data, measures =c("Shannon"))~~

~~effnum <- exp(rich$Shannon)~~

* 1. ~~Test for differences in the effective number of species between the two types of gluten diet (boxplot and Wilcoxon test)~~
  2. ~~Perform ordination plots using both MDS and NMDS and the BC, UF and wUF distances. Display the samples according to disease.~~
  3. ~~Test for global differences between cases and controls with PERMANOVA~~

1. Use coda4microbiome R package to analyze hansen dataset at genus level
   1. Perform an exploratory analysis of log-ratios
   2. Implement variable selection to obtain the trajectory of a microbial signature over time that is predictive of the kind of gluten diet.