Statistical methods for microbiome analysis

Evaluation exercise

Submit the results of the analysis by July 10 to [malu.calle@uvic.cat](mailto:malu.calle@uvic.cat)

Send me a pdf or html document with the code, results and comments

Data set Shao et al. (2019): Microbial samples from children born by different methods (vaginal and c-section) at different time points.

https://pubmed.ncbi.nlm.nih.gov/31534227/

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

data<-readRDS("shao.rds")

1. Use phyloseq library to preprocess and describe the dataset:
   1. Number of taxa, number of samples, number of variables in metadata
   2. Agglomerate at the genus level and normalize the data to relative abundances
   3. Summarize the data at the species level: number of taxa and summary and histogram of log10(mean taxa abundances)
2. Cross-sectional analysis:
   1. Filter data so you work only with children at time point 4 days (infant\_age == 4). You can use subset\_samples() from phyloseq package
   2. Check how many samples from vaginal birth and c-section birth are there before performing the analysis. You can use function table().
   3. 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 among both born methods (vaginal vs c-section). Plot boxplot and perform wilcoxon test
  2. Perform ordination plots using both MDS and NMDS and the BC, UF and wUF distances. Display the samples according to born method.
  3. Test for global differences between vaginal born and s-section with PERMANOVA
  4. Use coda4microbiome R package: implement variable selection to obtain a microbial signature that is predictive of born method.

1. Longitudinal analysis:
   1. Filter data so you work only with children at time points 4, 7 and 21 days (infant\_age == c(4, 7, 21)). You can use subset\_samples() from phyloseq package
   2. Check how many samples from vaginal birth and c-section birth at each time point are there before performing the analysis. You can use function table().
   3. Use coda4microbiome R package: Implement variable selection for longitudinal data to obtain a microbial signature that is predictive of born method over time
      * Create a new data.frame called “metadata.longit” with the metadata of the phyloseq object in step 4a. Then, order both the metadata and the abundance table by ID\_num:

metadata <- metadata.longit[order(metadata.longit$ID\_num),]

x <- x[order(metadata.longit$ID\_num),]