Statistical analysis of read based metagenomic data

Multivariate analysis, diversity and modelling overdispersed count data

Katariina Pärnänen

Metagenomics course 2019

What’s read based metagenomic data like?

- Usually tables, including gene/taxa count tables, taxonomy tables and metadata
- ’Big data’, needs special statistical methods and can’t be analyzed in Excel or SPSS
- Count tables can have many zeros
- Has a lot of variance and is not normally distributed

Microbial ecology statistics can be used for metagenomics too

- Diversity and ordination methods are similar in 16S and metagenomics
- Diversity indexes should take both evenness and the number of species into account
- Multivariate methods are used to relate samples to each other by calculating distances between the samples and ordinating them in a two dimensional space
  - Sharing 0 values should not result in higher similarity, since the data will be sparse and dissimilarity indexes should be selected accordingly

What you should do?

- DO try and analyze your data using statistics so you can go beyond beign just descriptive
- DO spend time selecting the right statistical method and distribution (usually for modeling count data you will go for GLMs with either negative binomial or quasipoisson distributions)
- DO adjust your p-values for multiple testing
- DO use DESEQ, metagenomeseq or EdgeR ,which are designed for metagenome/transcriptome/gene expression data
- DO try machnine learning like randomforests for your data
- DO bear in mind that metagenomic data can be compositional and keep up with new ways to analyze compositional data
  - Check Understanding sequencing data as compositions: an outlook and review https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6084572/

What you shouldn’t do?

- DON’T use normal distribution assumptions or use log transformations
  - Mixed effects models with over dispersion are still not very established so if you need a mixed effect model, you might still need to log transform your data to get a normal distribution
  - Check Do not logtransform count data, https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/j.2041-210X.2010.00021.x)
- DON’T rarefy your data
  - Check: Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003531