Project

January 11, 2025

1 Population dynamics of bacteria in the gut microbiome: a data analysis

1.1 Short description

The project amis at analyzing data from a longitudinal (temporal) study of the bacterial populations in the gut of mice during their whole life duration. The data was collected for the aim of characterizing the microbioma, modeling the ecological interactions among the species and gaining insight into the methabolic pathways. The data was collected from fecal samples through the metagenomic teqnique known as 16smRNA sequencing. The project aims at providing basic analysis of the results with some statistics.

1.2 Tasks

1. **Preprocessing and sample composition analysis**

   - **Data preprocessing**:

     group OTUs corresponding to same species, fix naming issues, double measures etc, retrieve total number of OTUs, total number of species, set up pandas DataFrames and OTU-species dictionary.

   - **Analysis of the cross-subject sample composition variability at different taxonomic-unit levels:**

     Biological organisms are classified in groups on the basis of similarity. Biologists refer to these groups as "taxonomic units" or taxa. Different taxa are also grouped together in a higher-level group ("taxon") based on commonly shared features, resulting in a nested hierarcical classification. The basic scheme of modern classification makes use of 8 levels of classification, which are, from top to bottom: *Domain, Kindom, Phylum, Class, Order, Family, Genus, Species.*

     The task is to assess the cross-subject variability of the sample composition at the various levels of the taxonomic classification. The output will be given in the form of stacked-bar plots. (AND POSSIBLY SOME QUANTITATIVE ANALYSIS?)

   - **Rank Abundance Distribution (RAD):**

     The frequency of each species is computed (= #counts_species/ #counts sample) . Species are ranked with an integer index from the most frequent (i= 1) to the least frequent (i= N). The curve displaying the frequency versus the rank (so-called rank-abundance distribution, or RAD) is experimentally known to assume a lognormal or power-law -like shape, with the first few species occupying the majority of the sample, and a long tail of rare species contributing to the remaining part. The task is to plot the RAD (AND FIT WITH SOME DISTRIBUTION?)

- OTHER THINGS HERE?

2. **Time-series analysis**

   Say $\vec{X}(t) = [x_1(t), x_2(t), \cdots x_N(t)]$ is the abudance of the first N most abudant species at time t. We model $\vec{X}(t)$ as a stochastic process, for which each subject $\vec{X}(t)$ represents a different sample path.

   - non stationary (i.e. to have time dependent PDF and, consequently,expectation values $\mathbb{E}[\vec{X}(t)]$, $\mathbb{E}[\vec{X}^2(t)]$);
   - noisy, due to both statistical measurement erros and systematic noise (mismatch of OTU-species identification, sequencing errors);

   The basic analysis to perform is:

   - Compute the time dependent mean and variance $\mathbb{E}[\vec{X}(t)]$, $\mathbb{E}[\vec{X}^2(t)]$. Plot as a summary the time evolution for the mean +- std of each species
   - Correlations: filter the most abundant species upon a certain threshold and compute their covariance matrix
   - Autocorrelation: compute the autocorrelation function $\mathbb{E}[x(t) \cdot x(s)]$

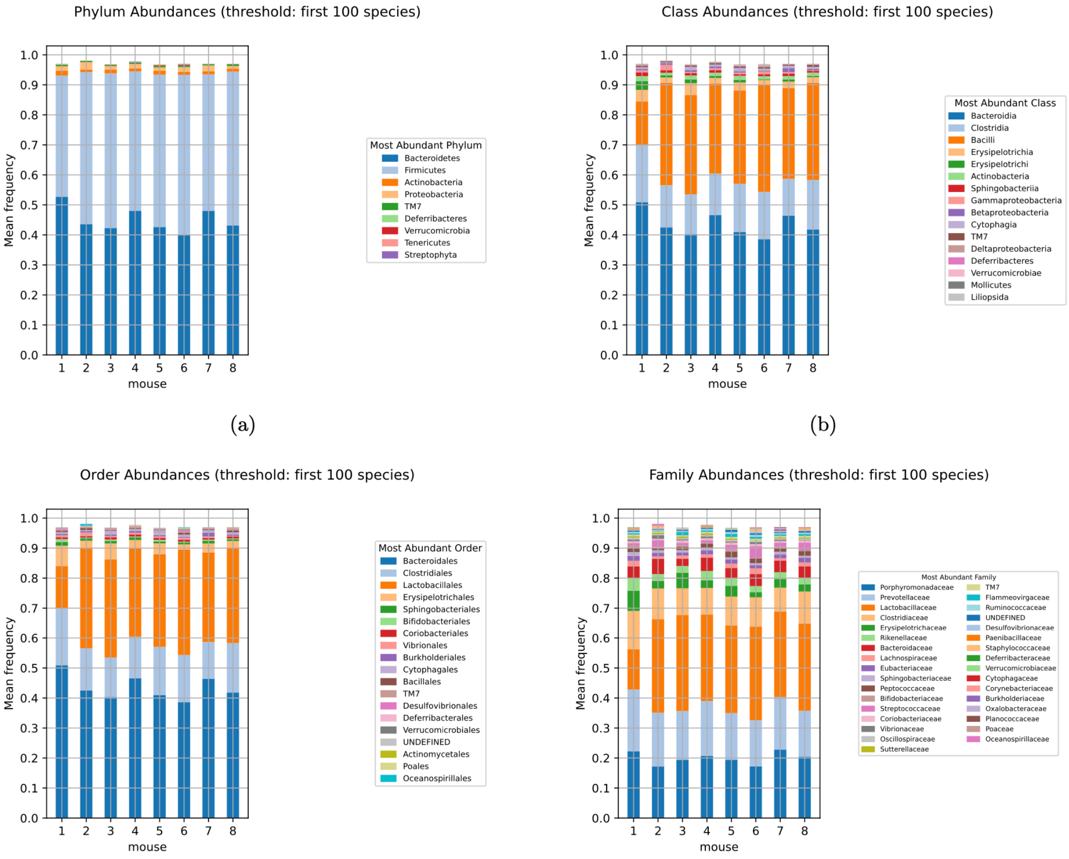
   WHAT MORE CAN BE DONE?

1.3 Outputs

[1]:

```python
from IPython.display import Image
Image("mice/stacked.png")
```

[1]:



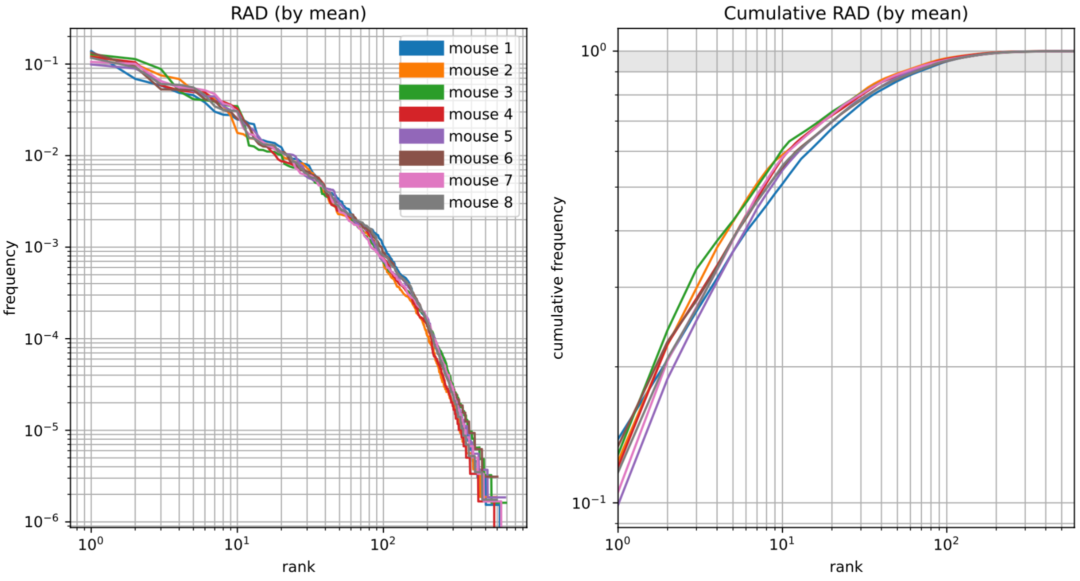
(a)

(b)

(c)

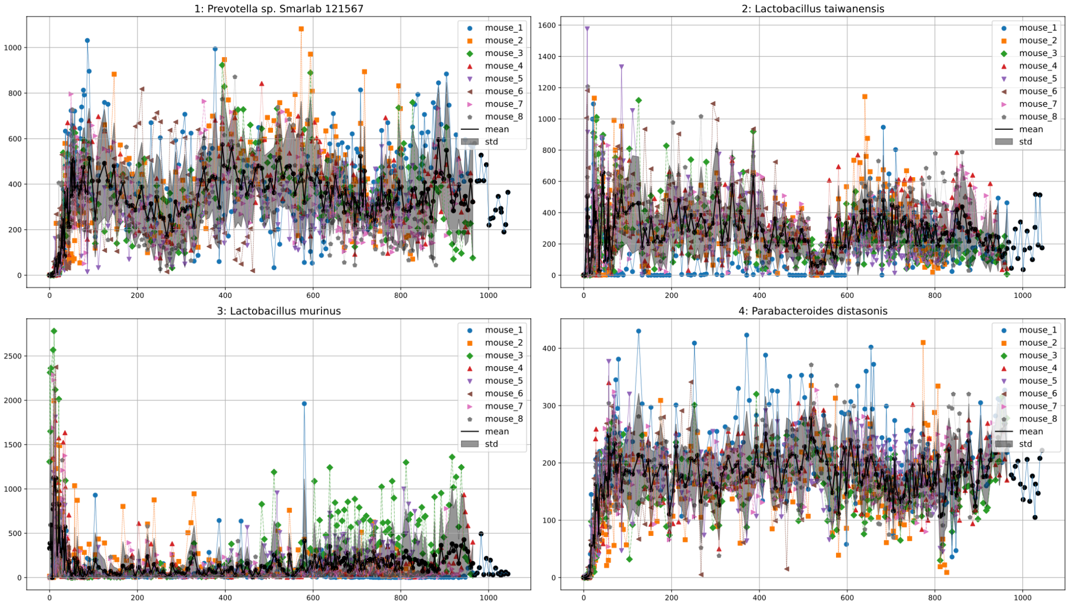
(d)

```
[2]: from IPython.display import Image
     Image("mice/rad.png")
[2]:
```



[3]:
```
from IPython.display import Image
Image("mice/timeseries.png")
```

[3]:


1: Prevotella sp. Smarlab 121567
2: Lactobacillus taiwanensis
3: Lactobacillus murinus
4: Parabacteroides distasonis
mouse_1
mouse_2
mouse_3
mouse_4
mouse_5
mouse_6
mouse_7
mouse_8
mean
std