

Master degree in Physics of Data - Academic Year 2024/2025

Final Project for the course of:

Laboratory of Computational Physics - mod A

Teacher: Marco Zanetti



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Data Analysis of Mice Gut Microbiota

Group ID: 08

Supervisor: Samir Simon Suweis | samir.suweis@unipd.it

Students name | Student ID | Student email

Bortolato, Angela | 2156562 | angela.bortolato.2@studenti.unipd.it

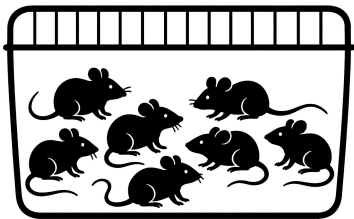
Fasiolo, Giorgia | 2159992 | giorgia.fasiolo@studenti.unipd.it

Volpi, Luca | 2157843 | luca.volpi@studenti.unipd.it

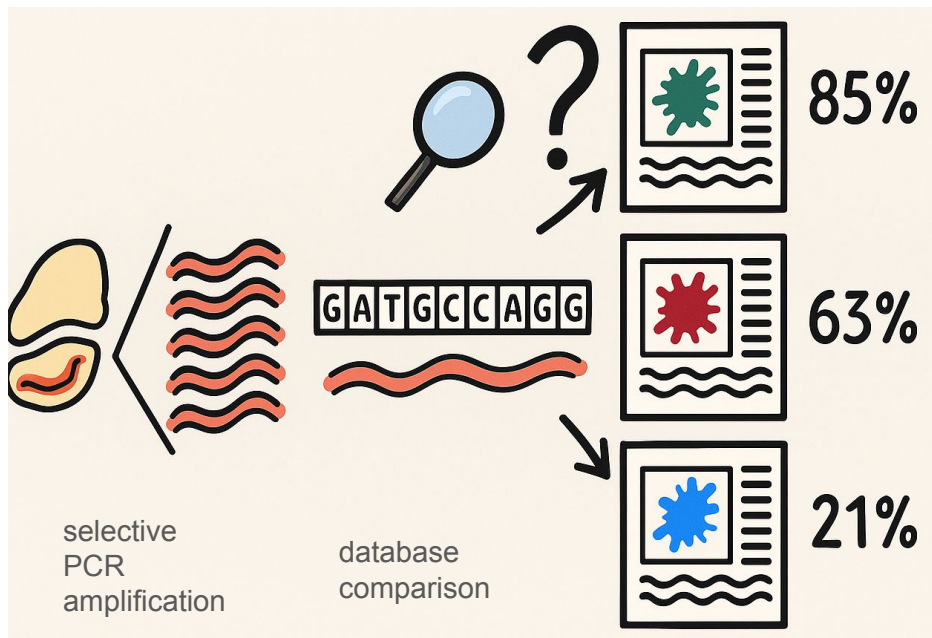
Zara, Miriam | 2163328 | miriam.zara@studenti.unipd.it

Familiarize with the Data

- 8 mice, born from the same parents and raised in the same cage



- A fecal sample taken from each of them every few days (~ 4-7)
- Bacteria in it are identified with **16s rRNA sequencing technique**

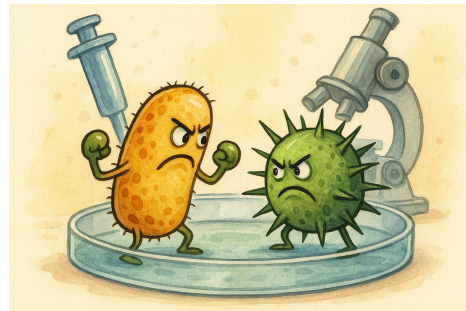
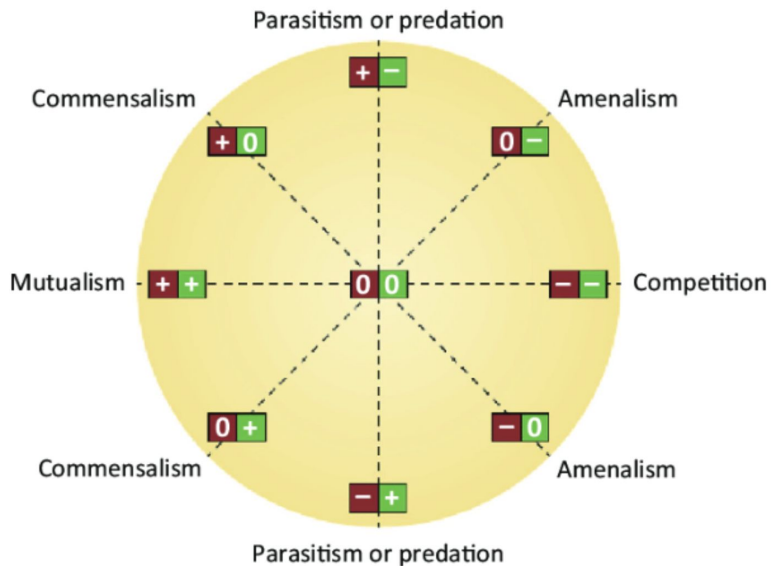
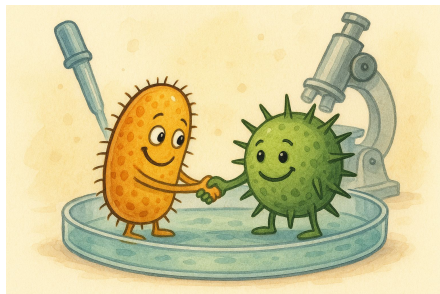


Learning Ecological Interactions

Treating *Clostridium difficile* Infection With Fecal Microbiota Transplantation

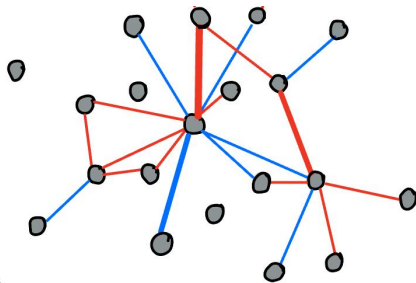
Bakken, Johan S. et al.

Clinical Gastroenterology and Hepatology, Volume 9, Issue 12,
1044 - 104, 2011 DOI: [10.1016/j.cgh.2011.08.0149](https://doi.org/10.1016/j.cgh.2011.08.0149)



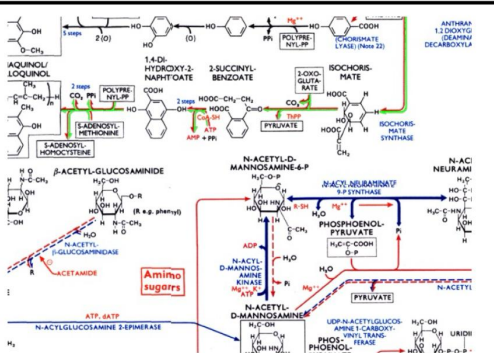
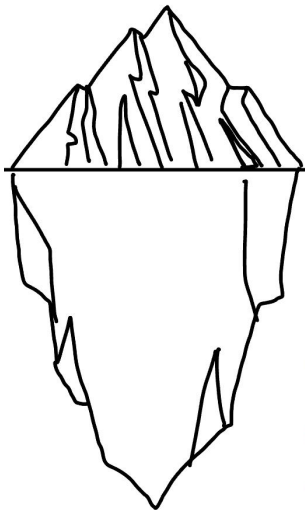
Learning Ecological Interactions

Species - species
effective
interaction network



- can ecological interactions really be inferred from the data?
- do the time series exhibit significant serial cross-correlation?
- are inter - species interactions a justified assumption - or does a “single species model” suffice to explain the observations?

metabolic
pathways

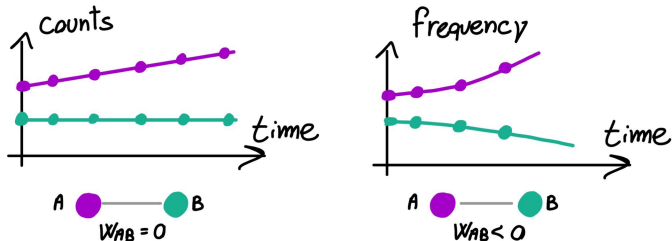


Criticalities

finite sequencing depth:

a species that is rare could still have a big influence on the others. By imposing a threshold (explicitly or implicitly) we exclude possibly vital information.

the sample is fixed in size: measures are frequencies, not counts
-> correlations may arise as statistical artifacts but have no correspondent physical reality



- 16s rRNA sequences was found to be efficient at identifying the high-order taxonomy, but less efficient at low-level taxonomy





- 16s rRNA sequences was found to be efficient at identifying the high-order taxonomy, but less efficient at low-level taxonomy

query	Phylum	Class	Order	Family	Genus	Species
OTU00001	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	Prevotella sp. Smarlab 121567 (79.62%)
OTU00002	Firmicutes,	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	Lactobacillus taiwanensis (100%)
OTU00003	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonada ceae	Parabacteroides	Parabacteroides distasonis



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

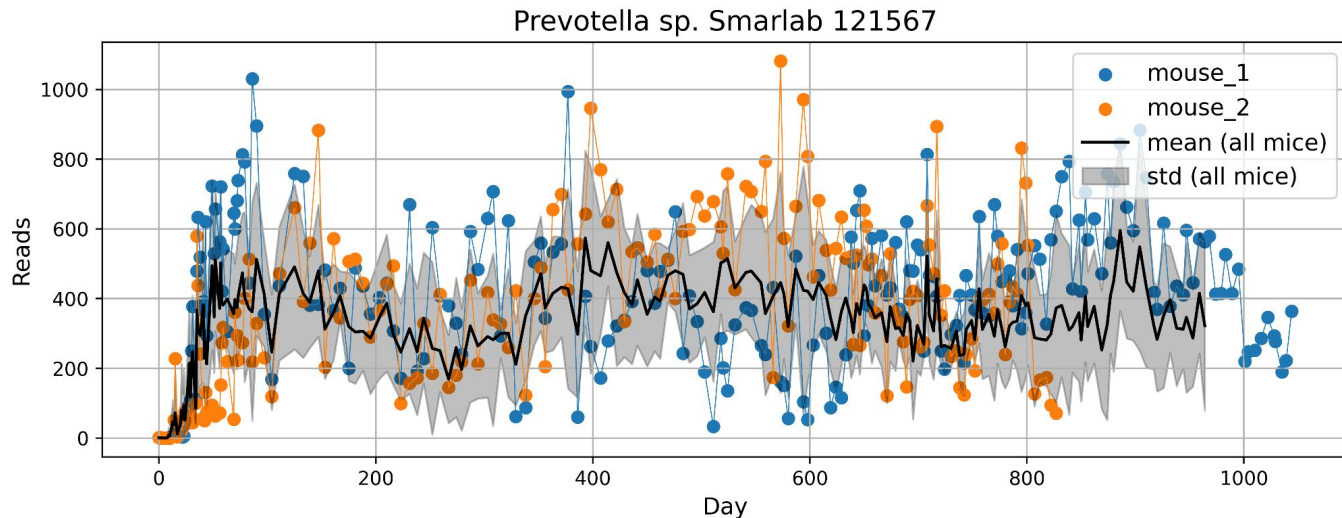
Data Preliminary Analysis



Familiarize with the Data

Preprocessing step:
aggregate the reads
for OTUs assigned to
the same species

OTU queries: 21.768
Species: 1.260
Genus: 412
Family: 141
Order: 66
Class: 37



Data is time series of the populations evolution - from birth to death of the host.

Threshold?



Sample Composition - 1

OTU queries: 21.768

Species: 1.260

Genus: 412

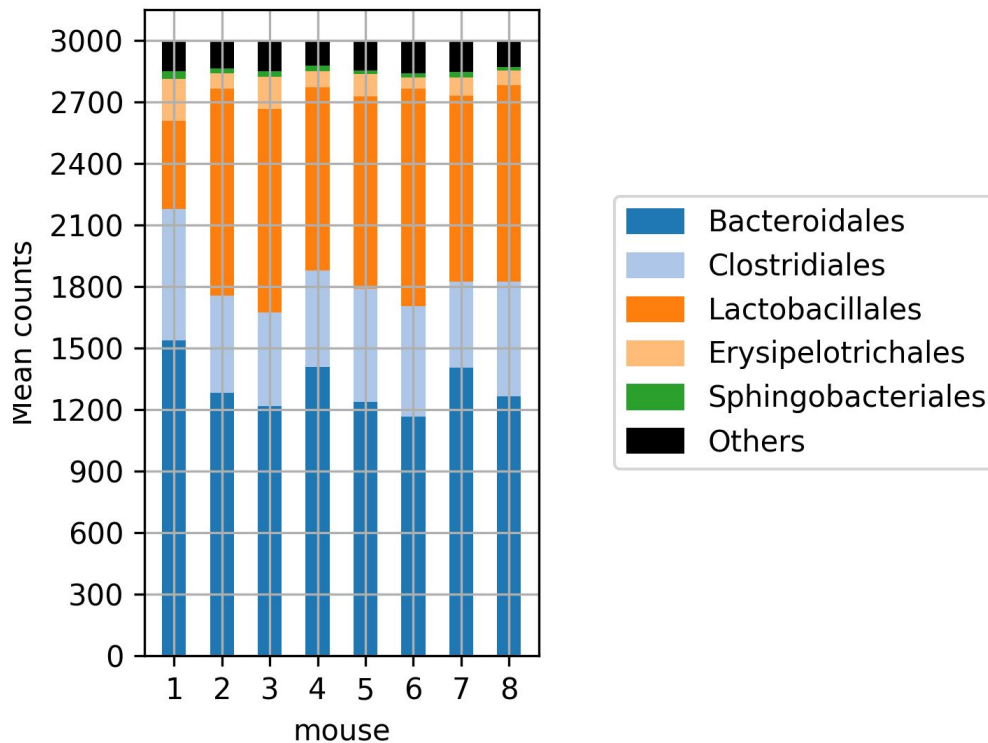
Family: 141

Order: 66

Class: 37

Composition is
homogeneous across the
subjects, at different levels
of taxonomic classification

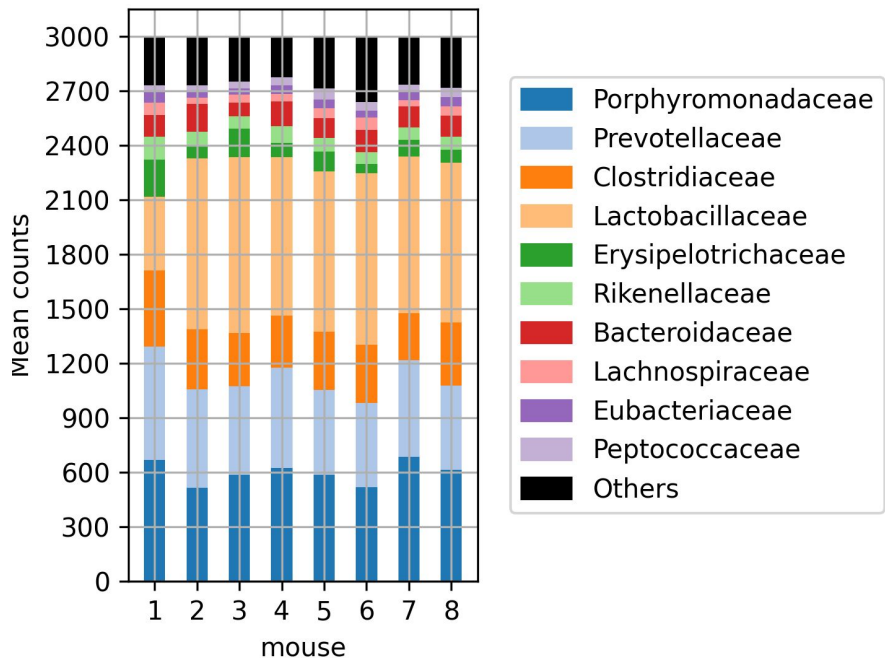
Order Abundances (Top 5 + Others)



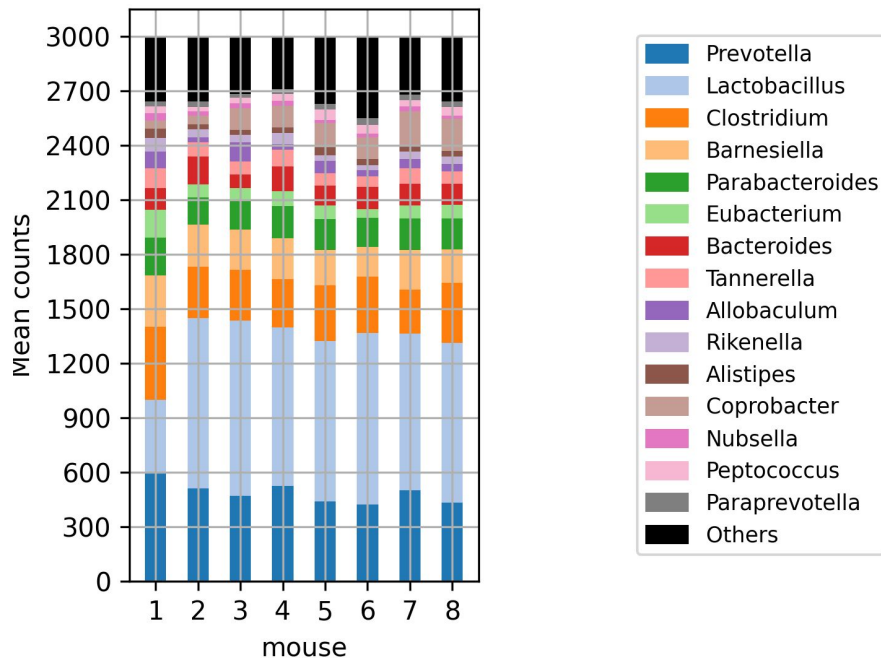


Sample Composition - 1

Family Abundances (Top 10 + Others)



Genus Abundances (Top 15 + Others)

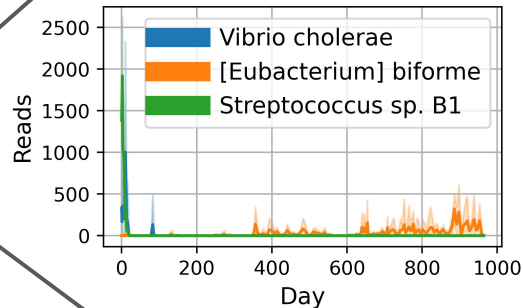
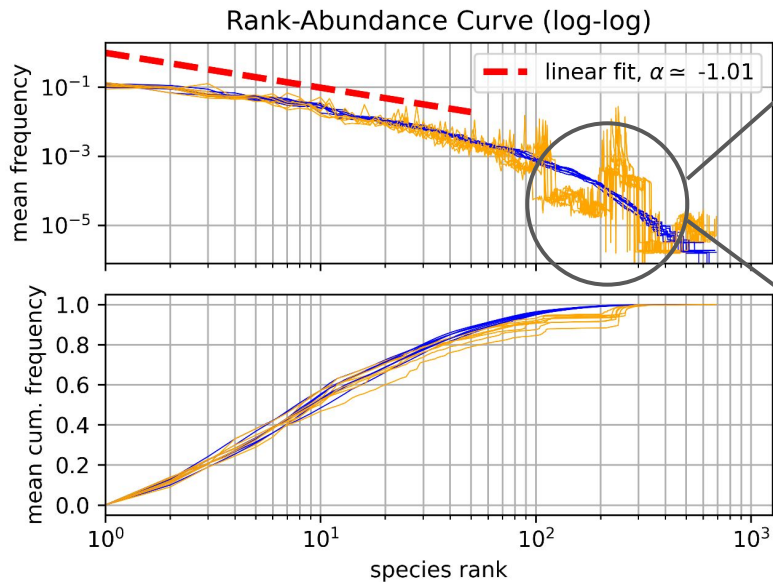
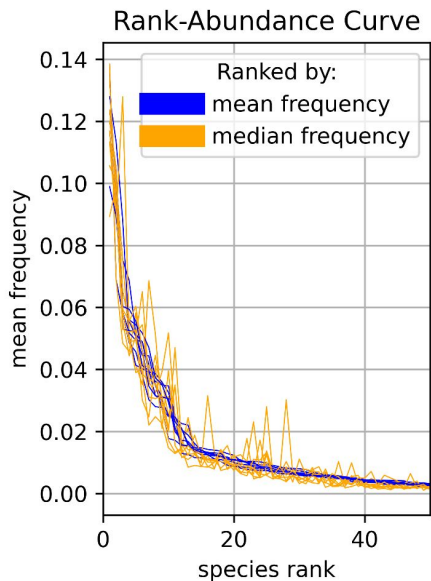




RAD : Rank-Abundance Distribution

Power law:

$$\text{frequency} = c \cdot \text{rank}^{\alpha}$$



non-stationary
series

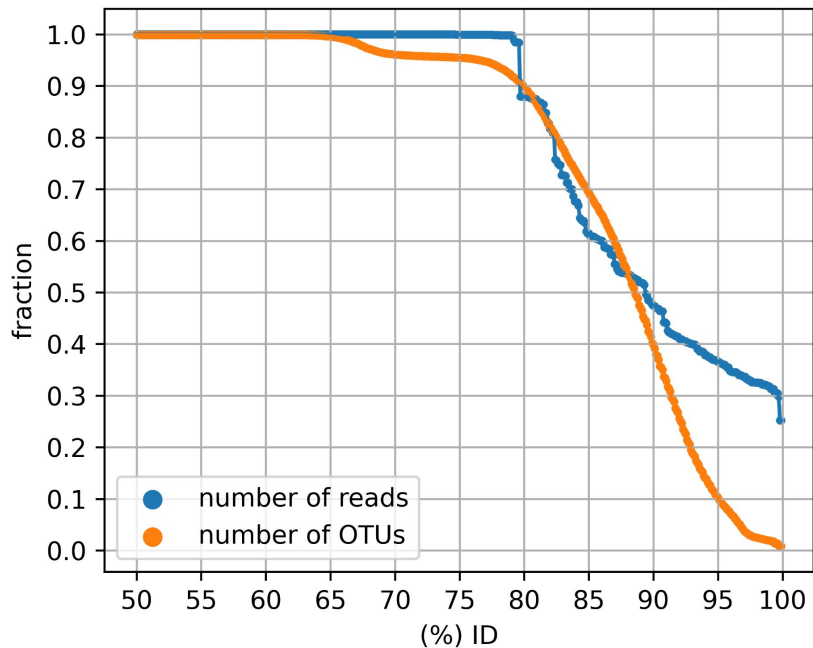


Data Analysis - Measure uncertainties

How reliable is *Species* assignation ?

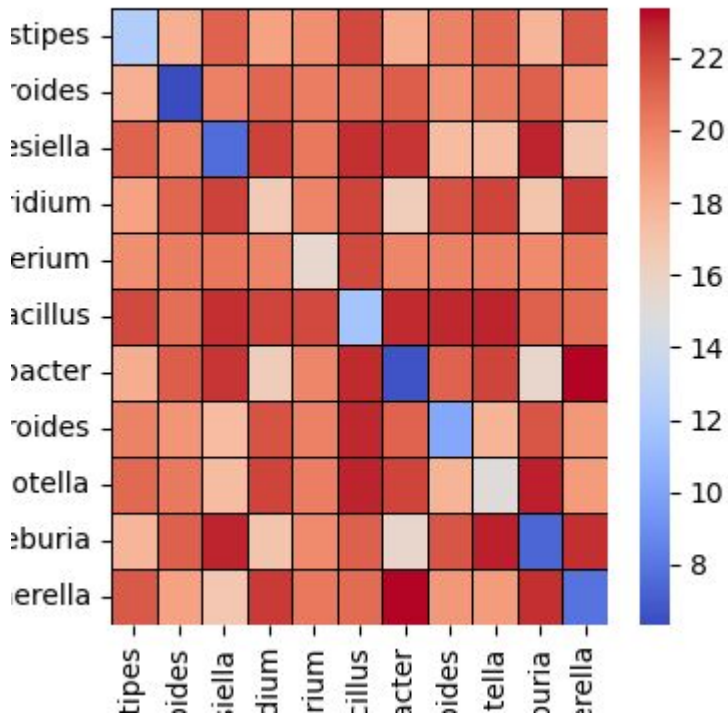
OTU queries: 21.768

Species: 1.260
Genus: 412





Data Aggregation by “Genus”





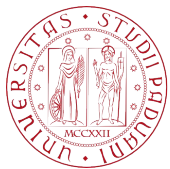
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Time Series Analysis

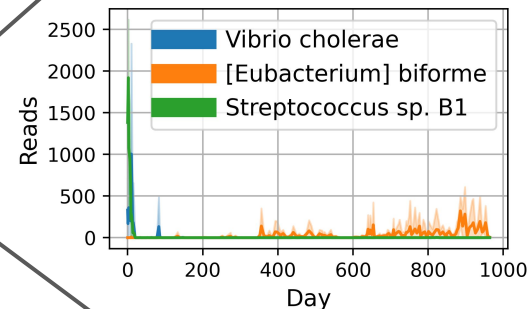
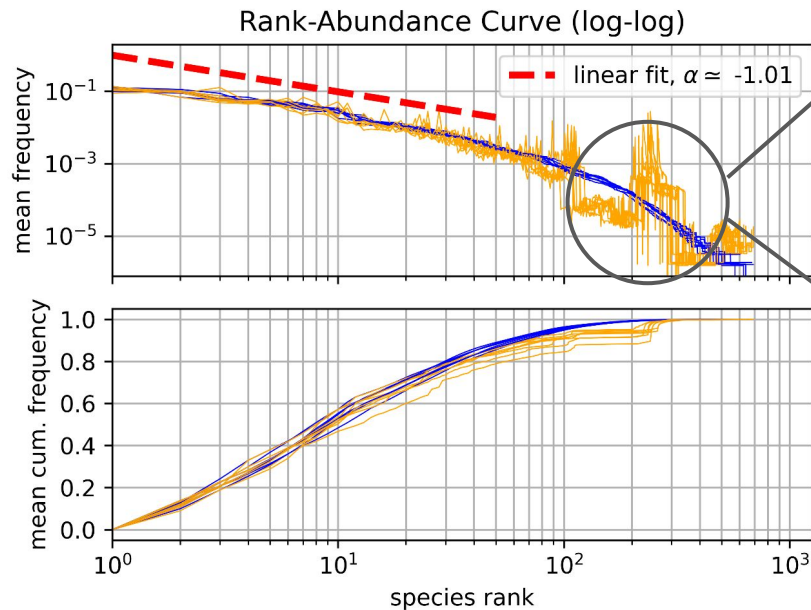
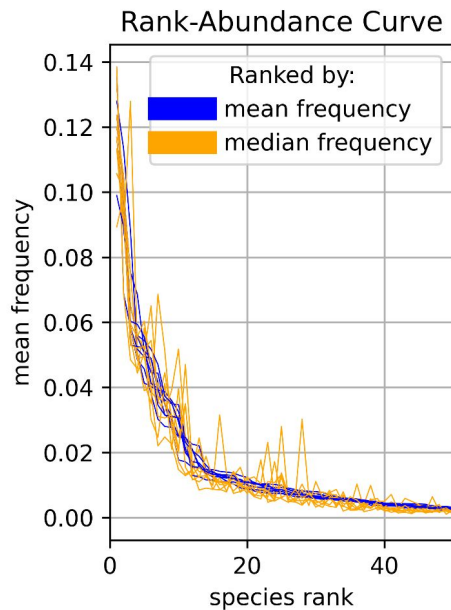


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Autocorrelation Function (ACF)



Data Analysis - RAD

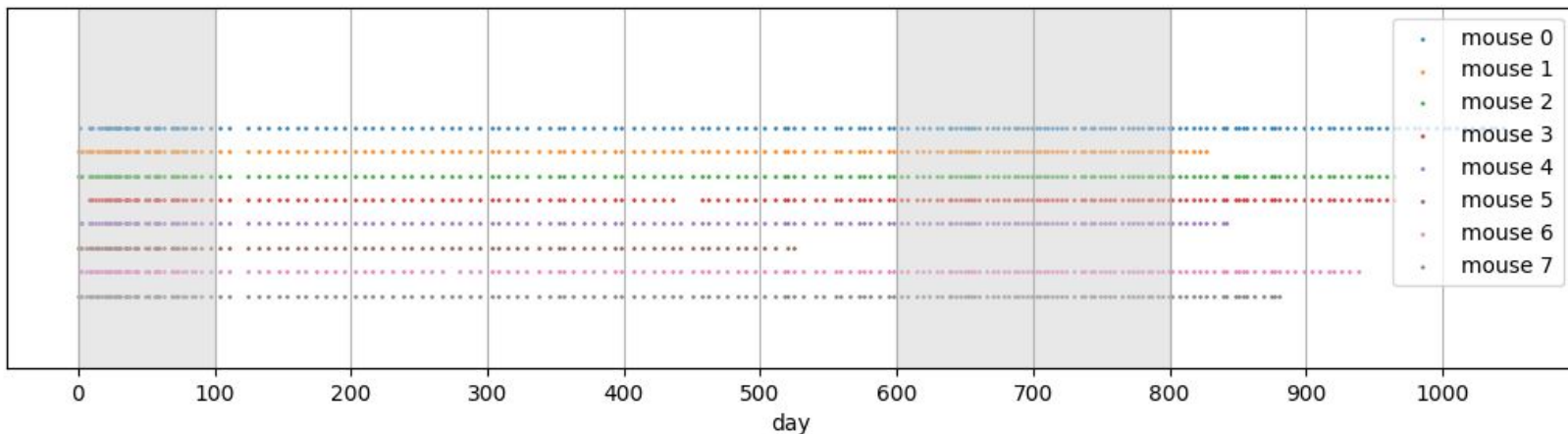


Power law:

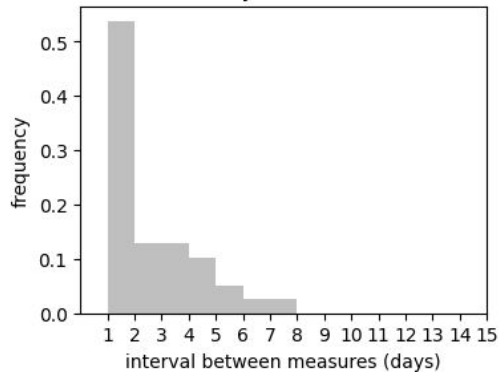
$$\text{frequency} = c \cdot \text{rank}^{\alpha}$$



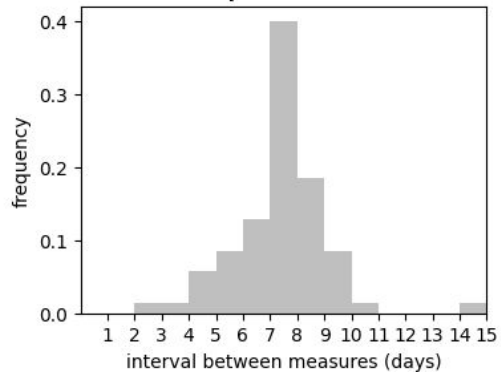
Data Preprocessing



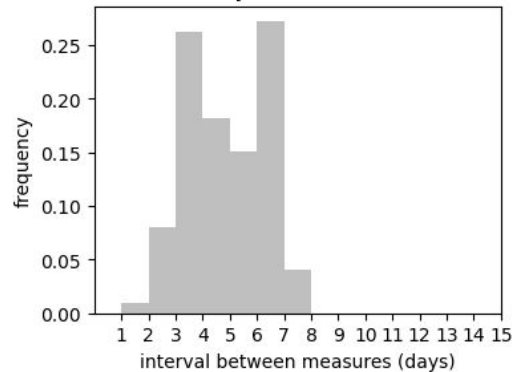
day 0 to 100



day 100 to 600



day 600 to 1044





UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Stationarity tests



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Logistic Model
