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

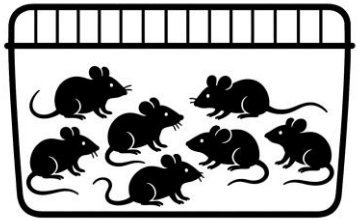
Supervisor: Samir Simon Suweis | samir.suweis@unipd.it

Students name | Student ID | Student email

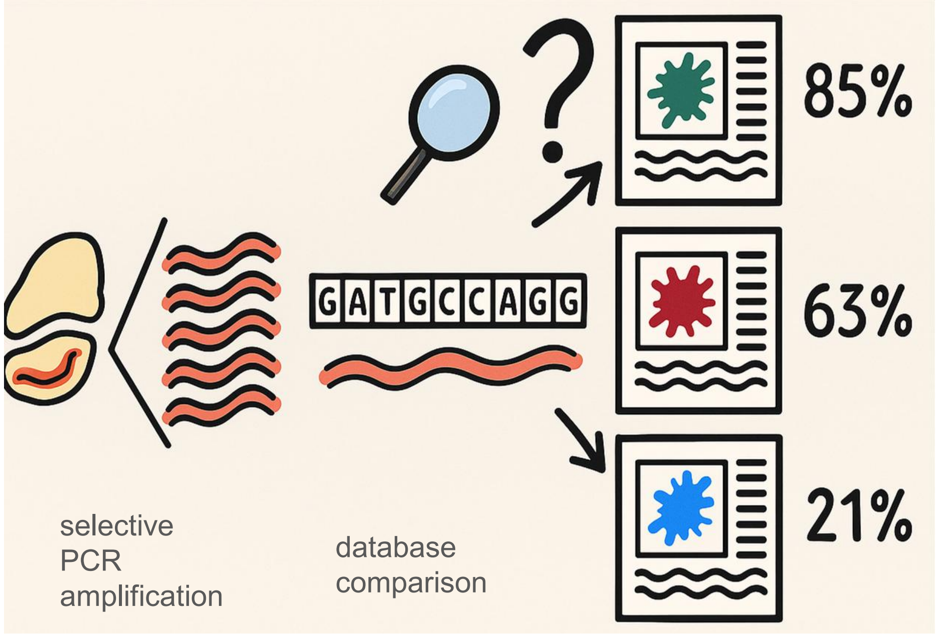
Bortolato, Angela | 2156562 | angela.bortolato.2@studenti.unipd.it
Fasiolo, Giorgia | 2159992 | giorgia.fasiolo@studenti.unipd.it
Volpi, Luca | 2157843 | luca.volpi@studenti.unipd.it
Zara, Miriam | 2163328 | miriam.zara@studenti.unipd.it

Familiarize with the Data

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

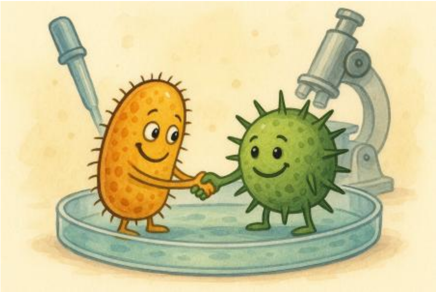


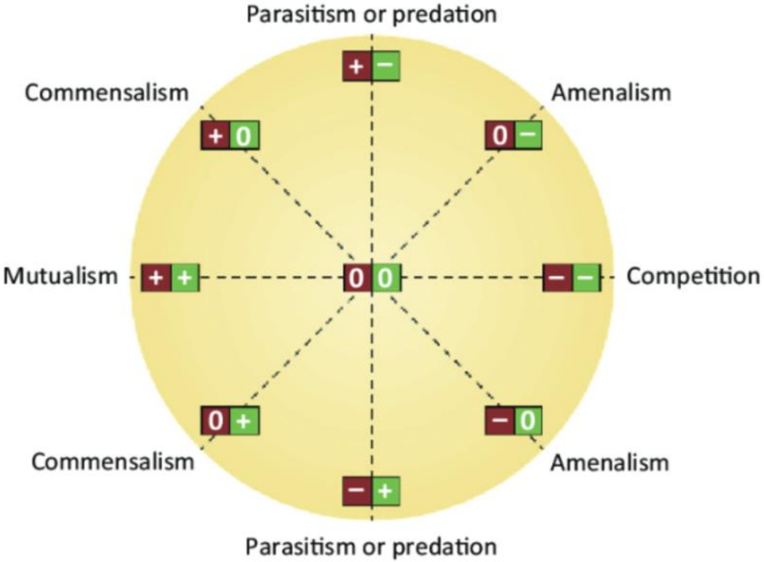
- A fecal sample taken from each of them every few days (~ 4-7)
- Bacteria in it are identified with **16s rRNA sequencing technique**

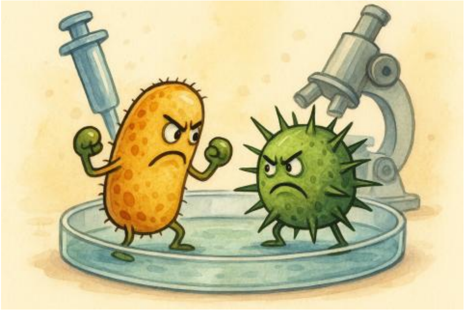


Learning Ecological Interactions

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





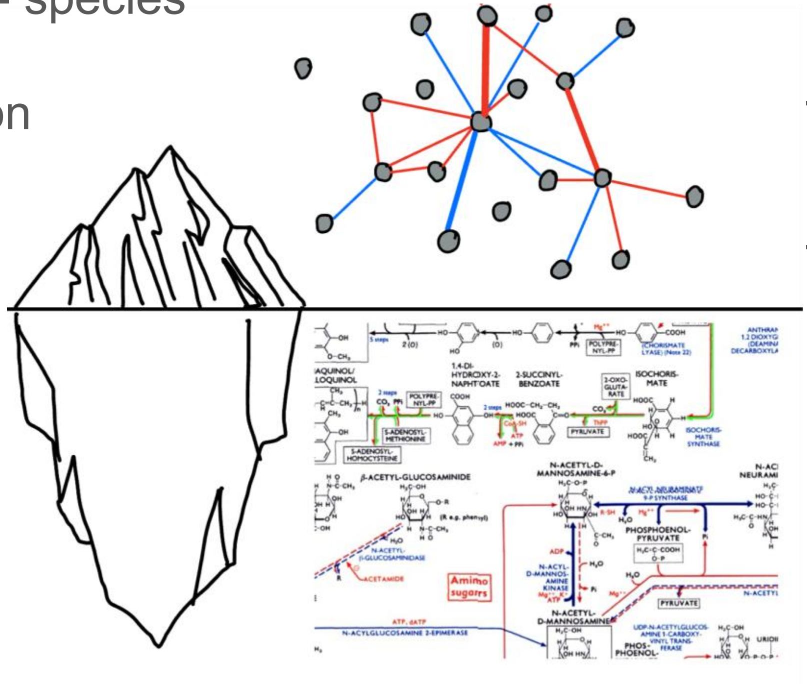


images generated with AI

UNIVERSITÀ DEGLI STUDI DI PADOVA

Learning Ecological Interactions

Species - species effective interaction network



metabolic pathways

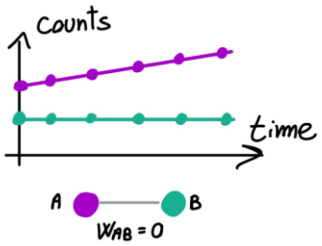
- can ecological interactions really be inferred from the data?
- do the time series exhibit significant serial cross-correlation?
- are inter - species interactions a justified assumption - or does a “single species model” suffice to explain the observations?

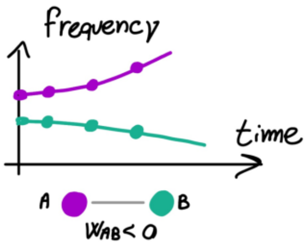
Criticalities

finite sequencing depth:

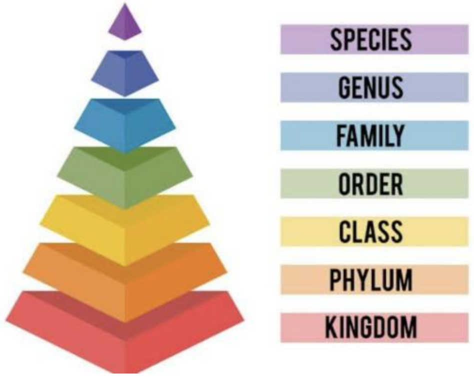
a species that is rare could still have a big influence on the others. By imposing a threshold (exiplicitly or implicity) we exclude possibly vital information.

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



Criticalities

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

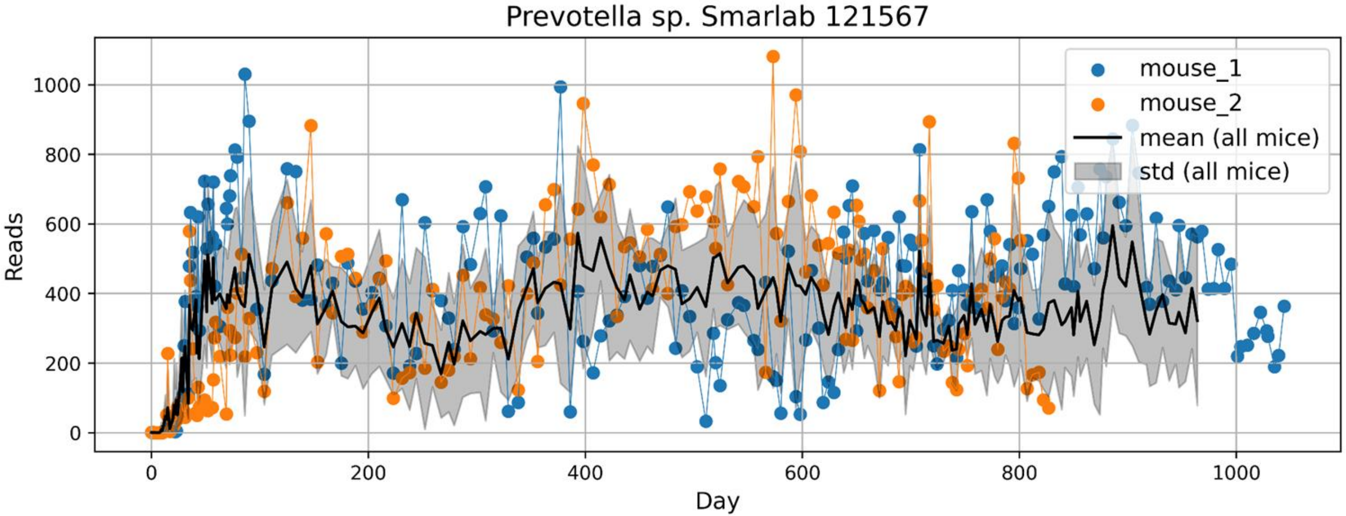
query	Phylum	Class	Order	Family	Genus	Species
OTU00001	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	**Prevotella**	Prevotella sp. Smarlab 121567 (79.62%)
OTU00002	Firmicutes,	Bacilli	Lactobacillales	Lactobacillaceae	**Lactobacillus**	Lactobacillus taiwanensis (100%)
OTU00003	Bacteroidetes	Bacteroidia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

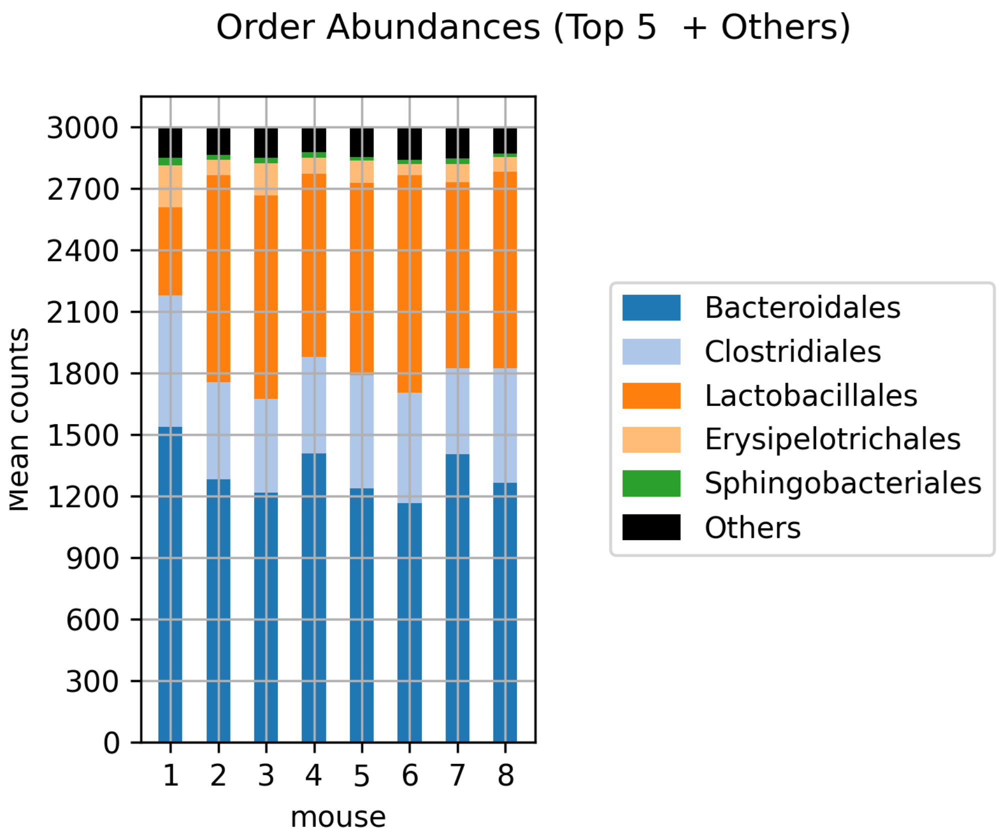
Threshold?

Università degli Studi di Padova

Sample Composition - 1

OTU queries: 21.768

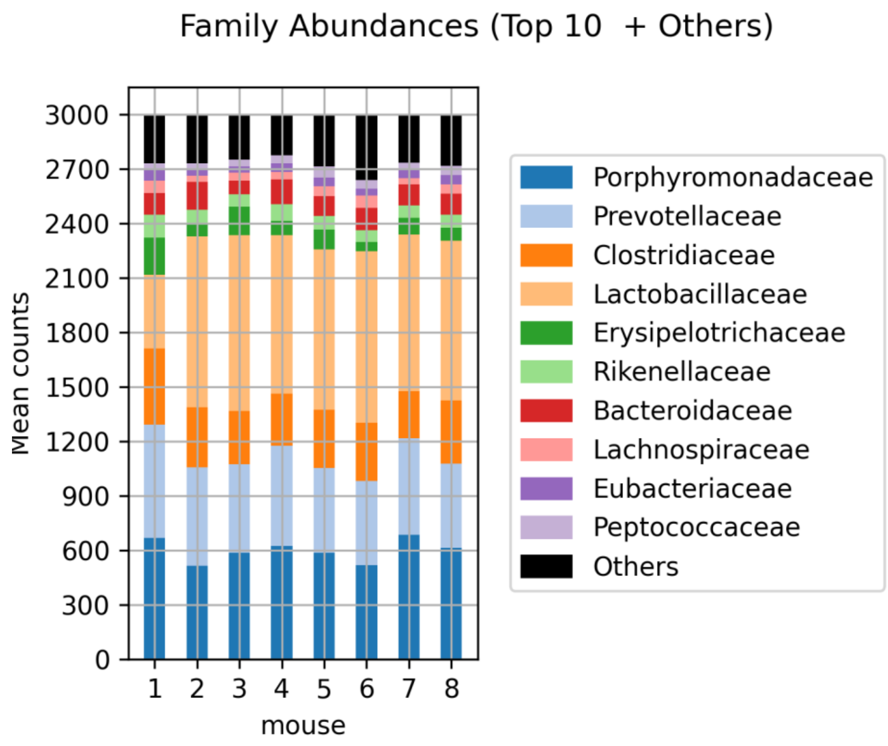
Species: 1.260
Genus: 412
Family: 141
Order: 66
Class: 37

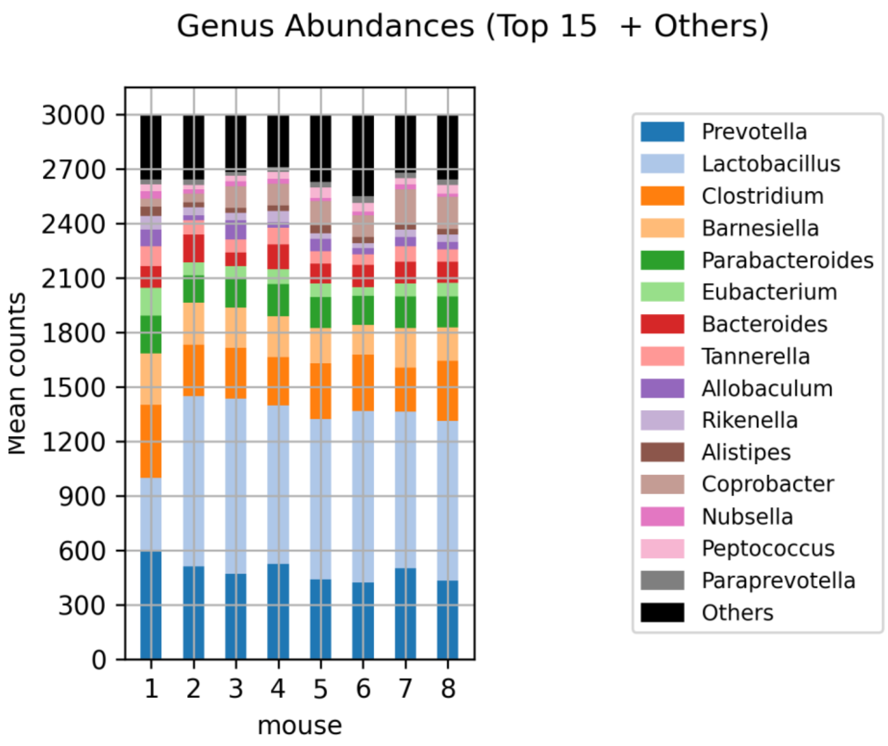


Composition is homogeneous across the subjects, at different levels of taxonomic classification

Università degli Studi di Padova

Sample Composition - 1

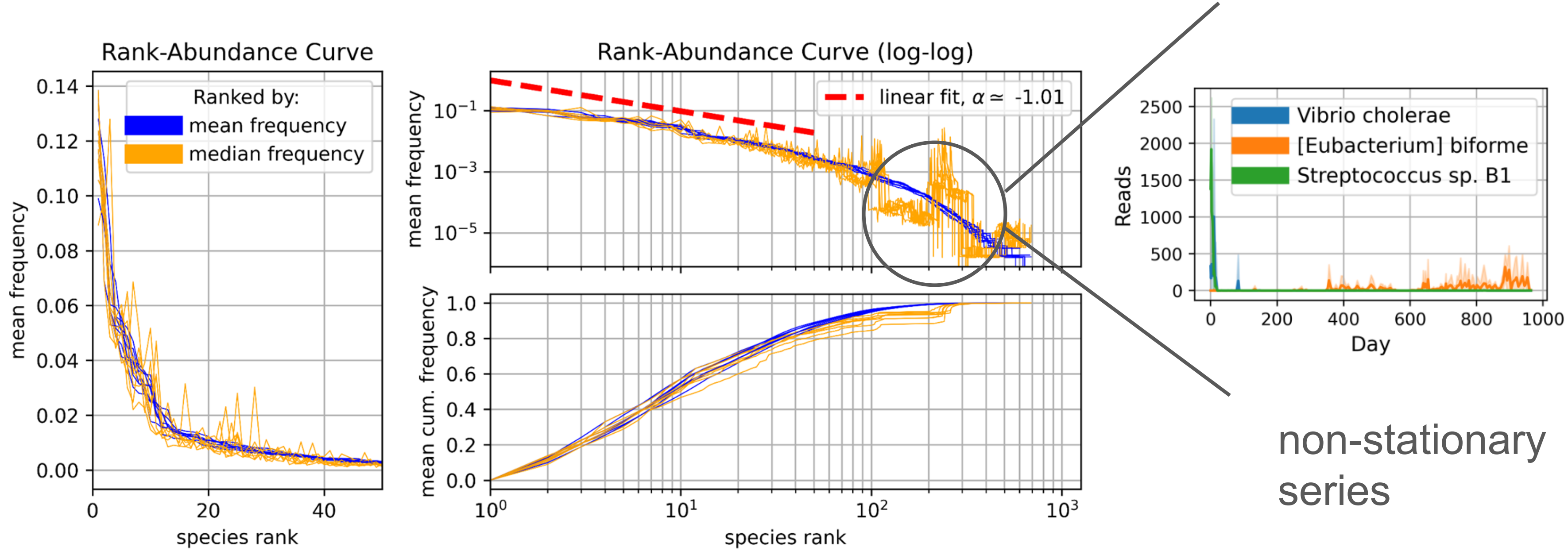




Sample composition - 2

RAD : Rank-Abundance Distribution

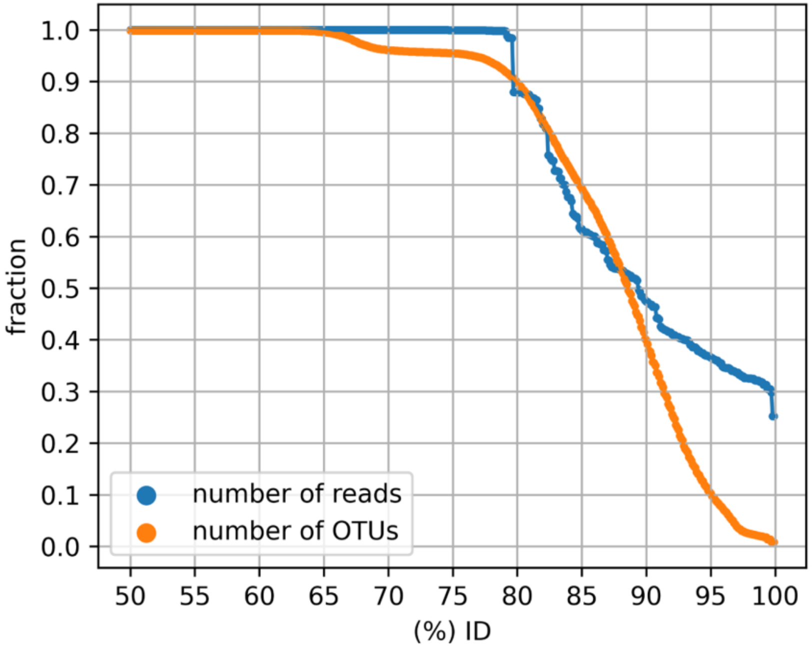
Power law: $$\text{frequency} = c \cdot \text{rank}^{\alpha}$$



non-stationary series

Data Analysis - Measure uncertainties

How reliable is *Species* assignation ?

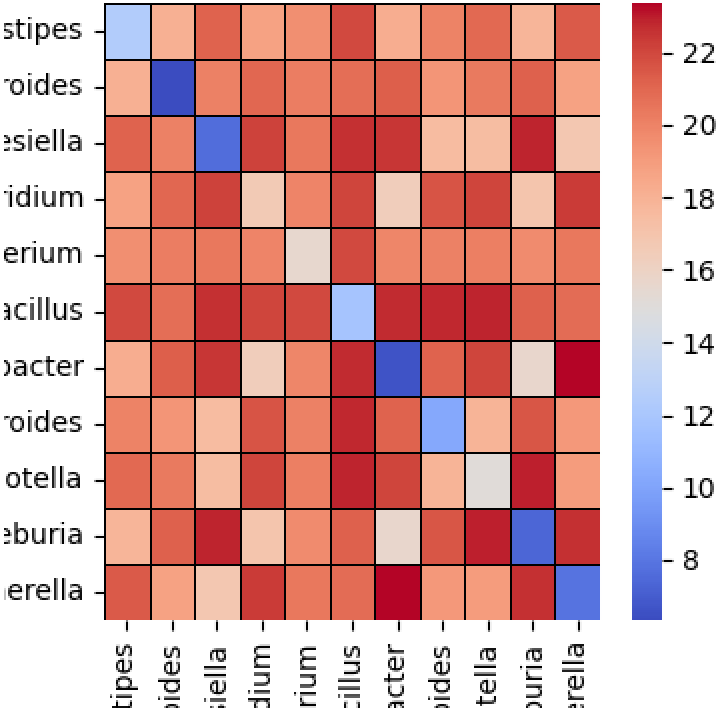


OTU queries: 21.768

Species: 1.260
Genus: 412

UNIVERSITÀ DEGLI STUDI DI PADOVA

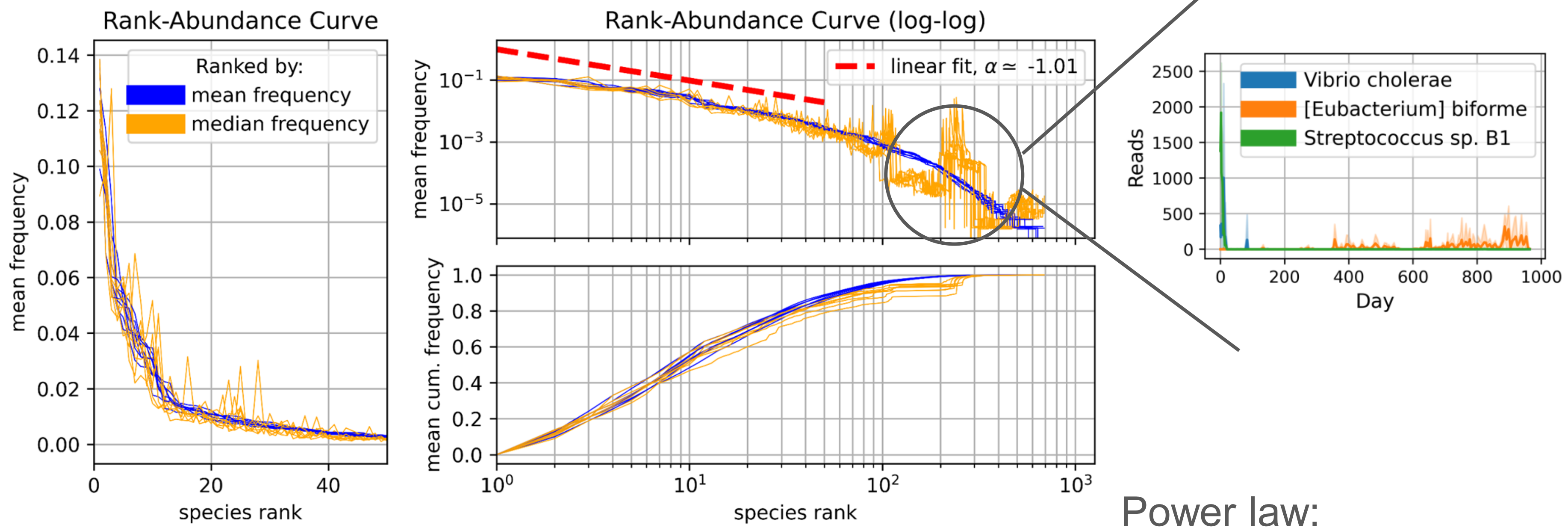
Data Aggregation by “Genus”



Time Series Analysis

Autocorrelation Function (ACF)

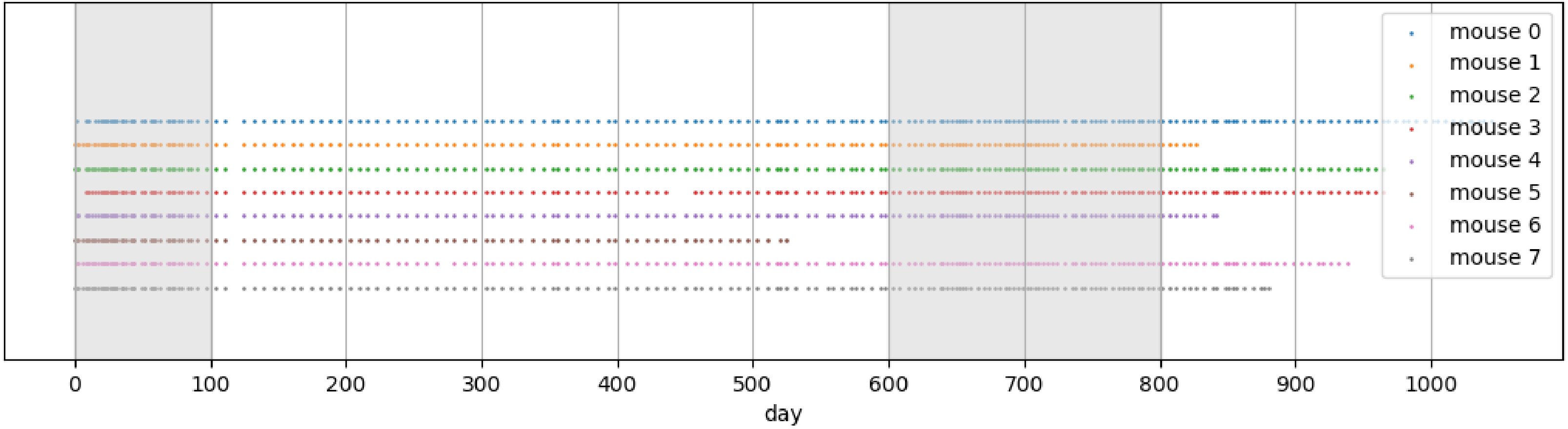
Data Analysis - RAD

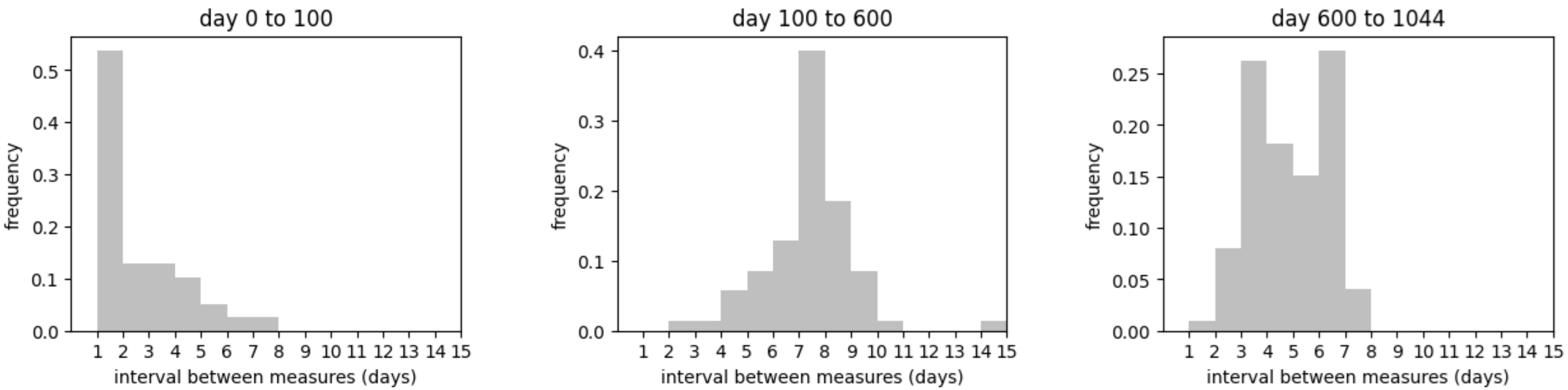


Power law:

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

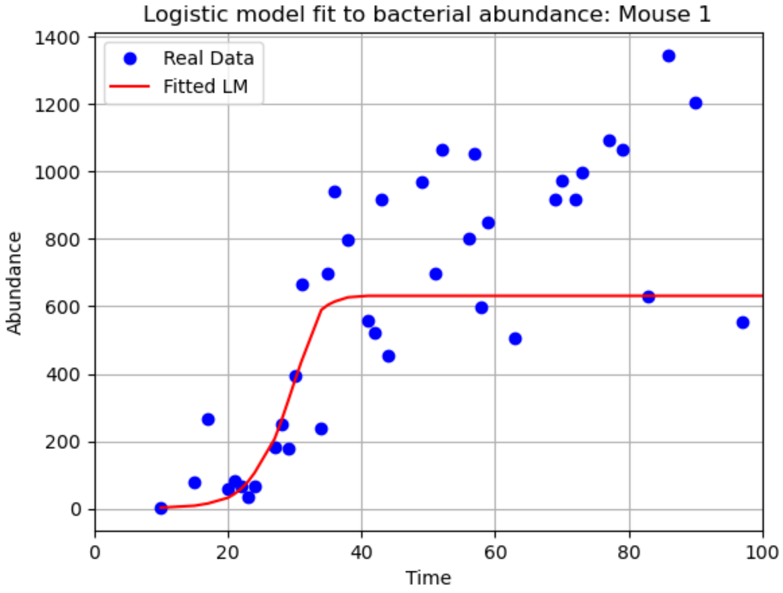
Data Preprocessing

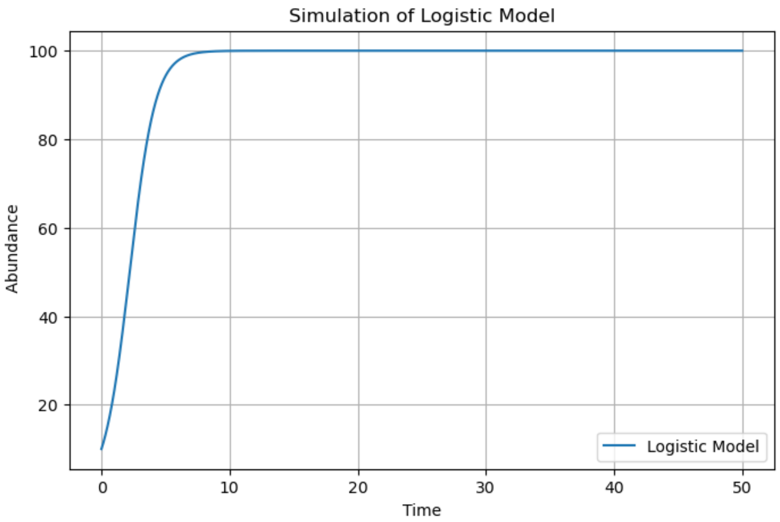




Università degli Studi di Padova

Logistic Model





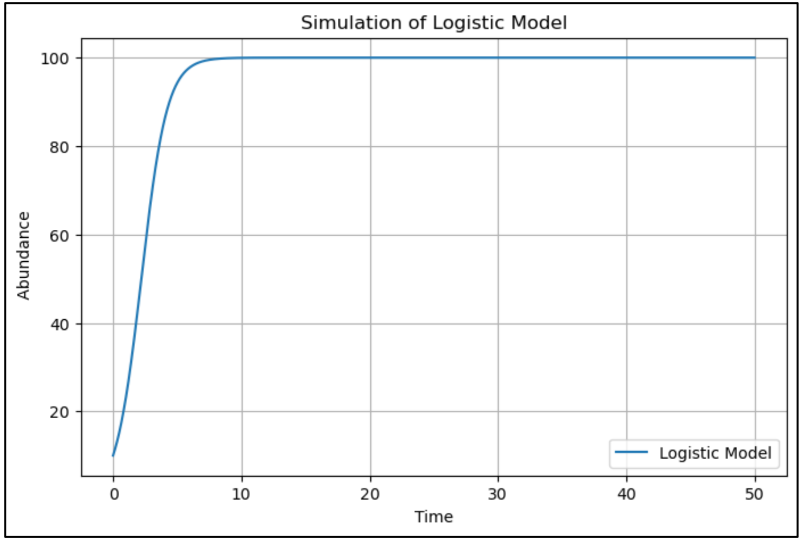
$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right)$$

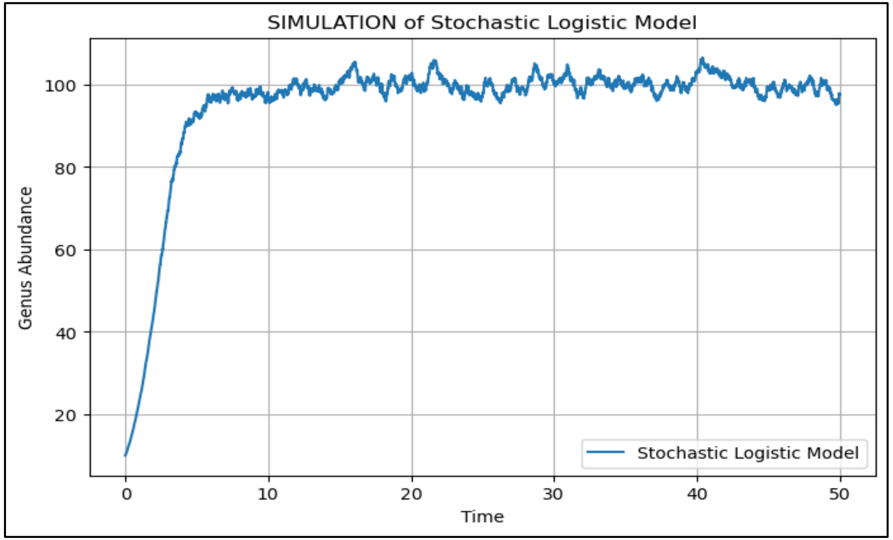
N = Population size
r = Growth rate
K= Carrying capacity

(Stochastic) Logistic Model

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) + Noise$$

N = Population size
r = Growth rate
K= Carrying capacity





UNIVERSITÀ DEGLI STUDI DI PADOVA

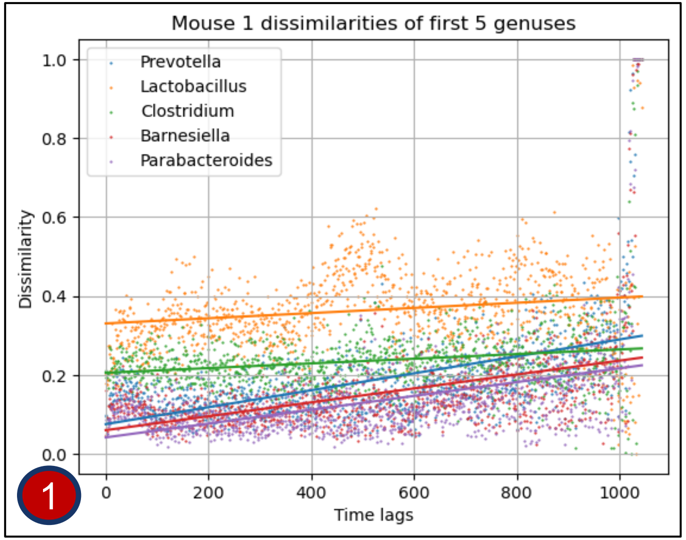
Stationarity tests

1 DISSIMILARITY

$$\Phi_i(t, T) = \left(\frac{\lambda_i(t) - \lambda_i(t+T)}{\lambda_i(t) + \lambda_i(t+T)}\right)^2$$

2 THRESHOLD

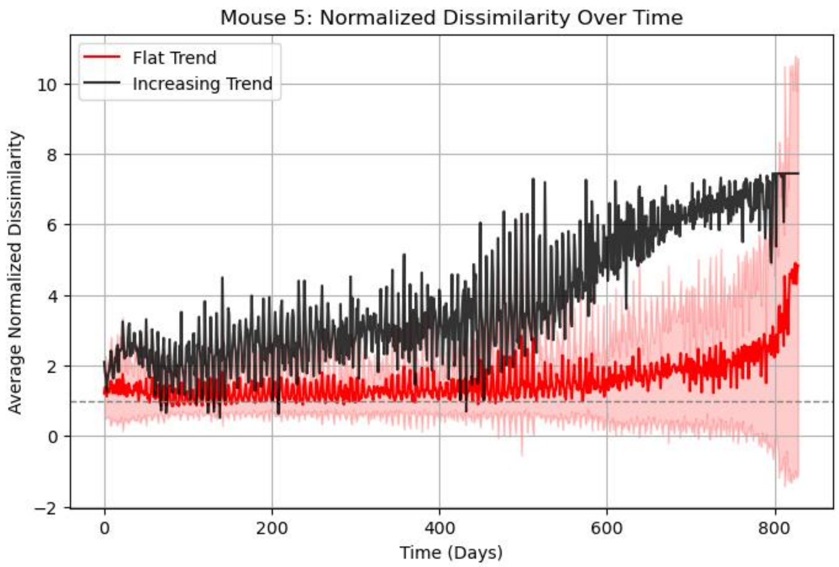
95th percentile of dissimilarity slopes

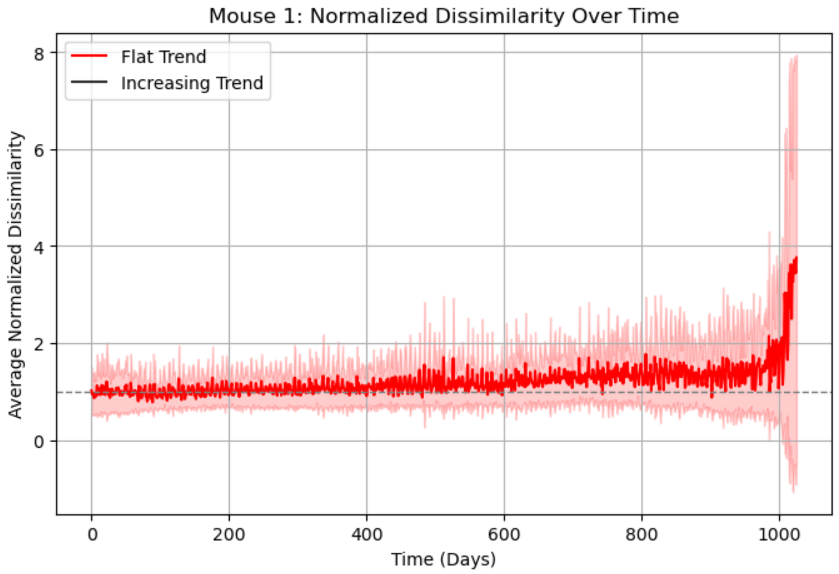


3 GAMMA DISTRIBUTION

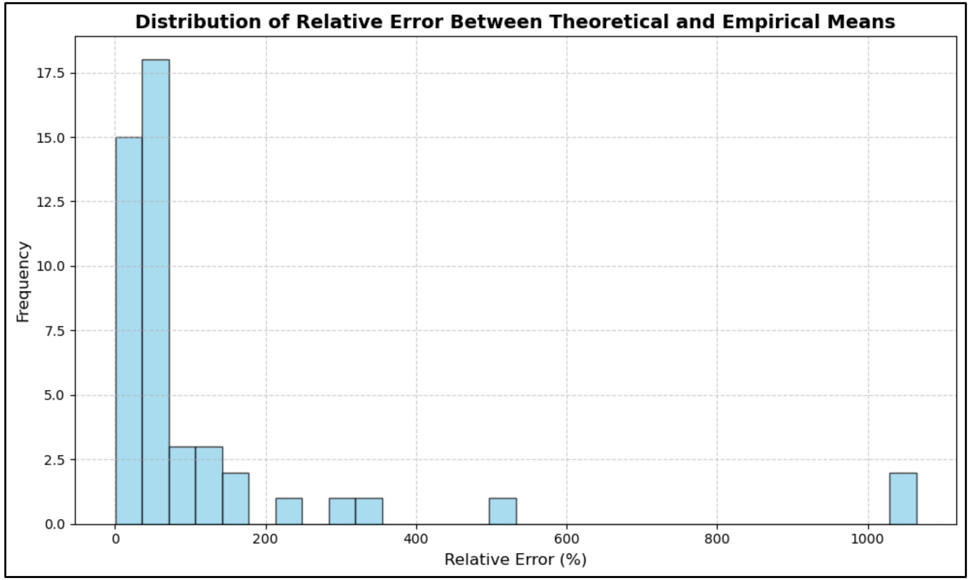
For stationary Genus

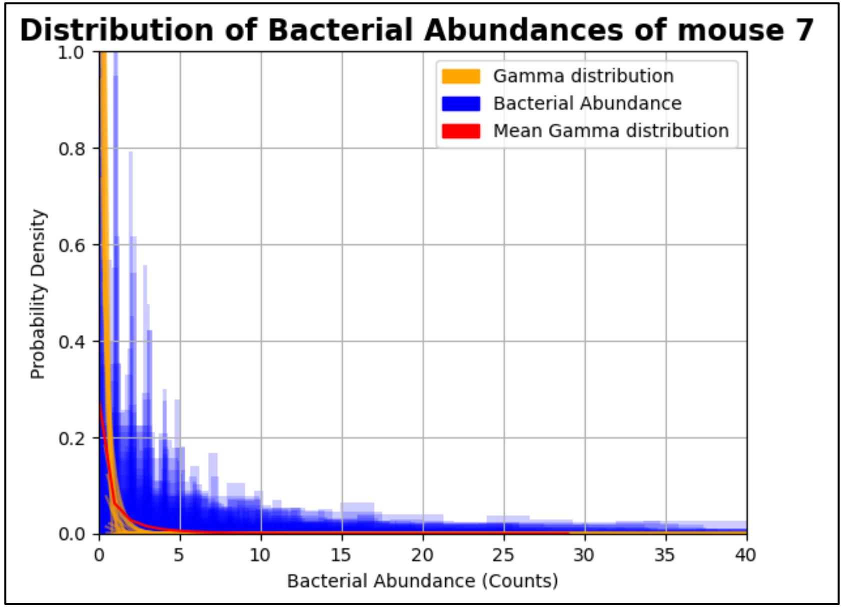
2 Stationarity tests





3 Gamma distribution





$$E[\Phi_\infty] = \frac{\sigma}{4-\sigma}$$

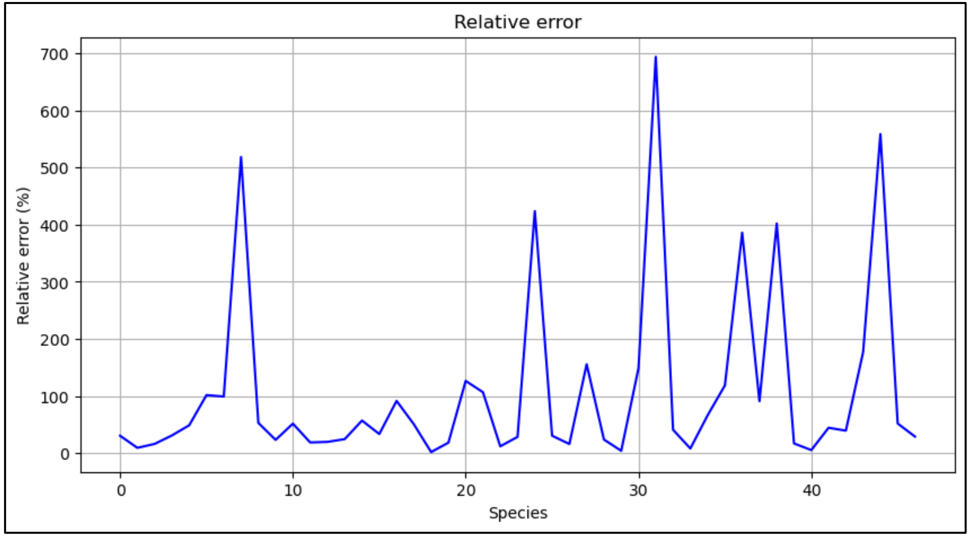
$$<\lambda> = K\left(\frac{2-\sigma}{2}\right)$$

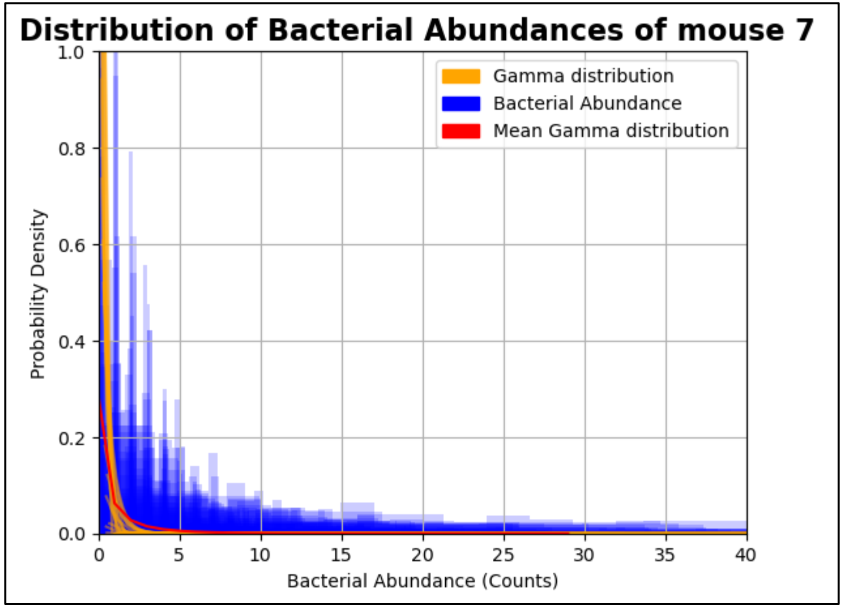
$$Var(\lambda) = \left(\frac{\sigma}{2-\sigma}\right) <\lambda>^2$$

$$P(\lambda; K, \sigma) = \frac{1}{\Gamma\left(\frac{2}{\sigma}-1\right)} \cdot \left(\frac{2-\sigma}{K}\right)^{\frac{2}{\sigma}-1} \cdot \lambda^{\frac{2}{\sigma}-2} \cdot \exp\left(-\frac{2}{\sigma K} \cdot \lambda\right)$$

UNIVERSITÀ DEGLI STUDI DI PADOVA

3 Gamma distribution





$$E[\Phi_\infty] = \frac{\sigma}{4-\sigma}$$

$$<\lambda> = K\left(\frac{2-\sigma}{2}\right)$$

$$Var(\lambda)=\left(\frac{\sigma}{2-\sigma}\right) <\lambda>^2$$

$$P(\lambda; K, \sigma) = \frac{1}{\Gamma\left(\frac{2}{\sigma}-1\right)} \cdot \left(\frac{2-\sigma}{K}\right)^{\frac{2}{\sigma}-1} \cdot \lambda^{\frac{2}{\sigma}-2} \cdot \exp\left(-\frac{2}{\sigma K} \cdot \lambda\right)$$