ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Exploring the DNA Micro-World: Data Science Tools for Public Health

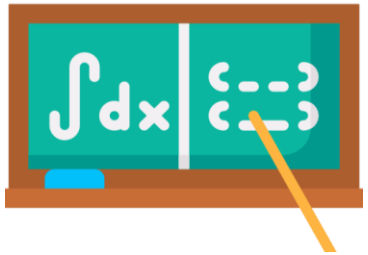
Alessandro Fuschi

University of Bologna – Department of Physics

Alessandro Fuschi

Applied Physicist | University of Bologna

Metagenomics



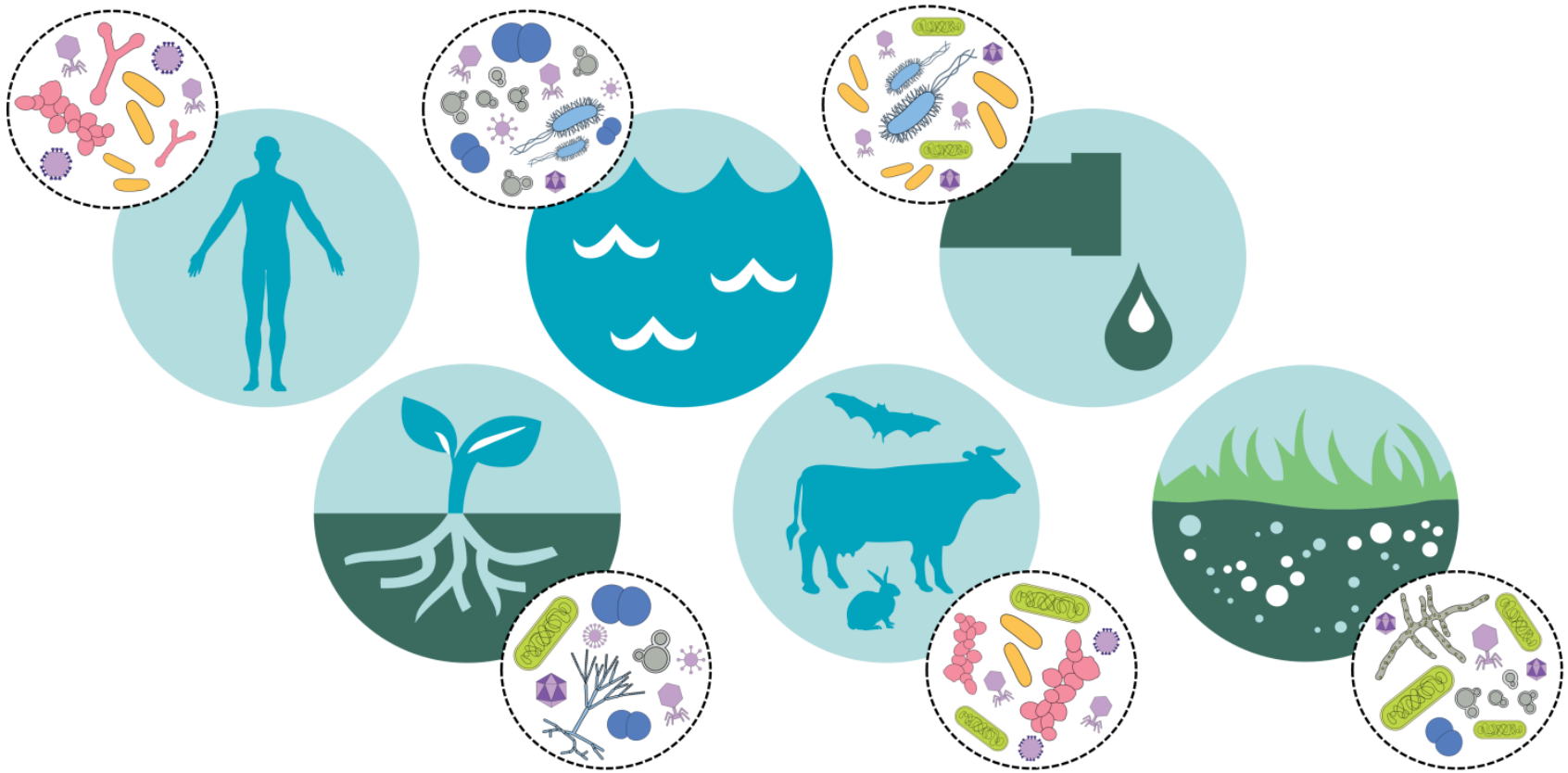
Mathematical Modeling

Networks

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

What is Metagenomics?

Reading all the DNA from all micro-organisms directly from their environments.



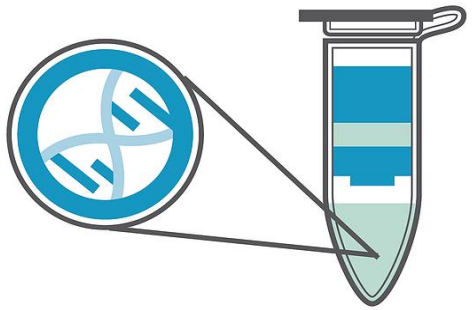
🧬 Found **everywhere:**
Human body, animals, water, soil, plants, wastewater...

🎯 The goal:
To uncover how the invisible microbial world works — who is there, what they do, and how they interact with us and the environment.

How Do We Unlock the Microbial World?

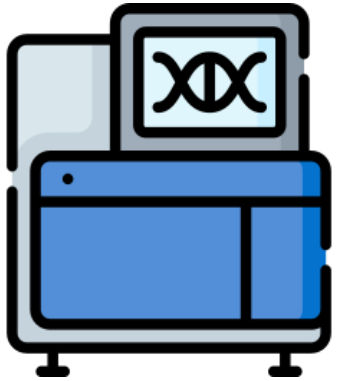
DNA Extraction

All DNA is extracted from the sample (soil, water, gut, etc.)



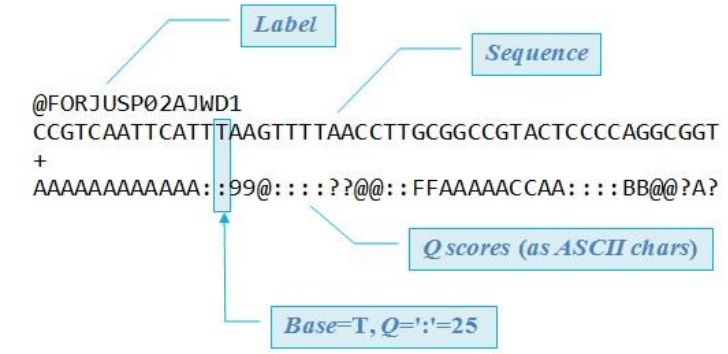
DNA Sequencing

Sequencing machines read the DNA base by base, producing millions of fragments



DNA Reads

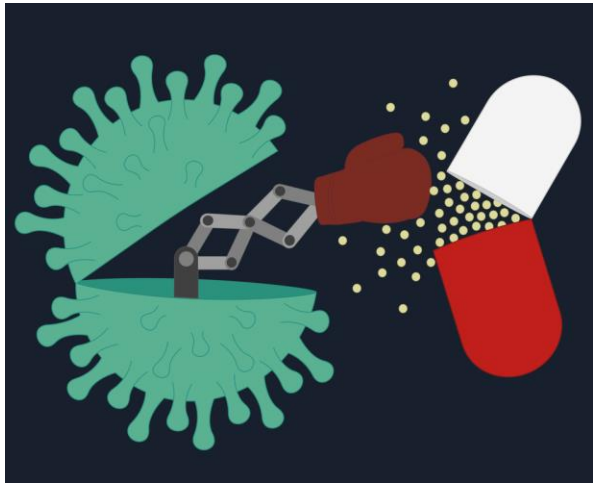
Short DNA sequences representing all organisms in the sample

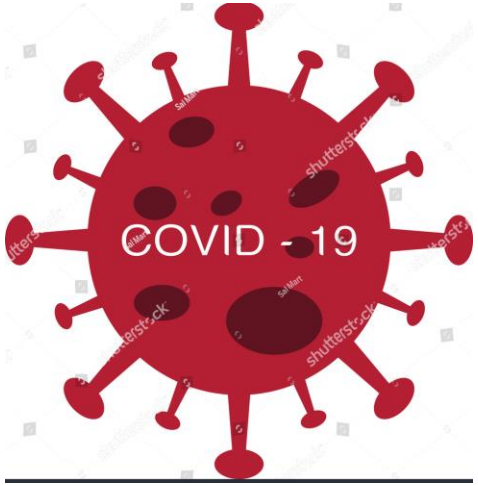


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Why Should We Care?

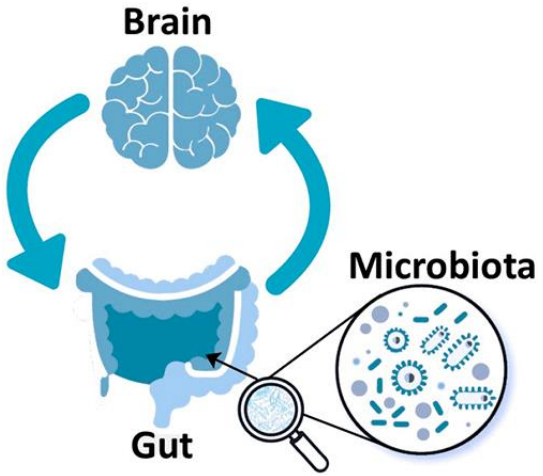
Antibiotic Resistance
Monitoring the spread of resistance genes in hospitals, farms, and the environment is essential to prevent future health crises

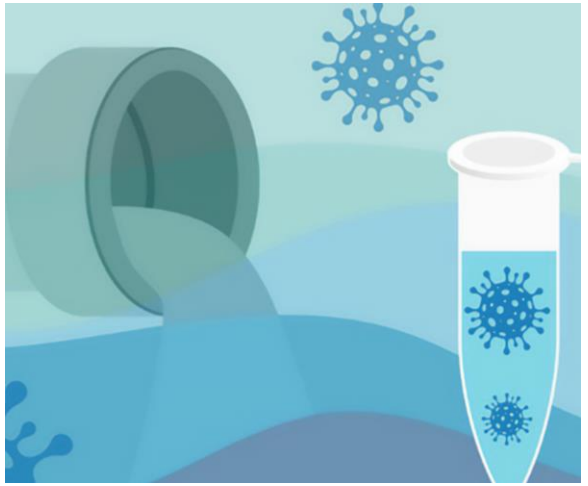




COVID-19
Sequencing revealed new variants early and helped track the evolution of the virus during the pandemic.

Human Microbiota
Gut microbes influence immunity, brain function, and chronic diseases — understanding their role is key to personalized medicine.





Wastewater Detection
Sewage surveillance can spot outbreaks before clinical cases emerge, offering a powerful early-warning system.

Who Does What in a Metagenomic Study?

🔬 **Clinicians & Biologists**

- Design the study and select subjects/sampling sites
- Collect and prepare biological/environmental samples

💻 **Bioinformaticians**

- Process raw DNA sequences into usable formats
- Use HPC systems and massively parallel pipelines (often described as “embarrassingly parallel”) to efficiently handle large-scale data

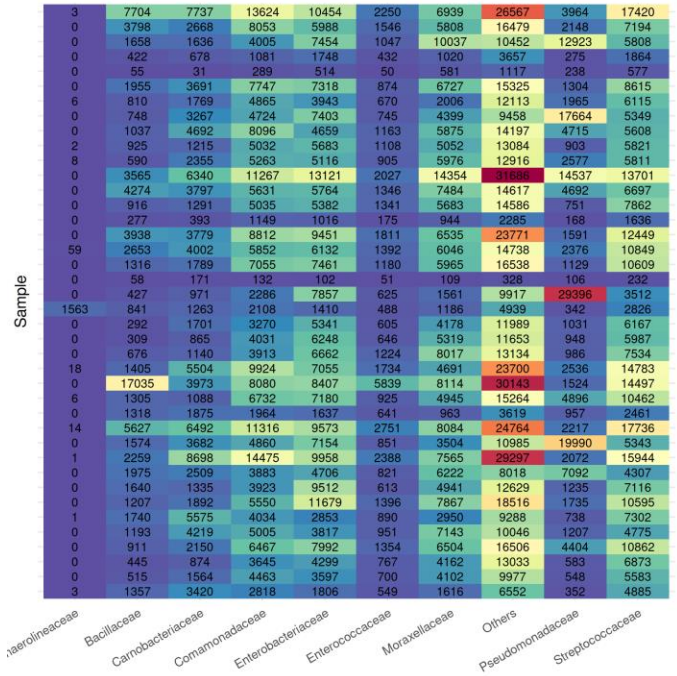
📊 **Data Scientists & Statisticians**

- Analyze abundance profiles or embeddings
- Uncover patterns, associations, and biological insight

How Do We Analyze Metagenomic Data?

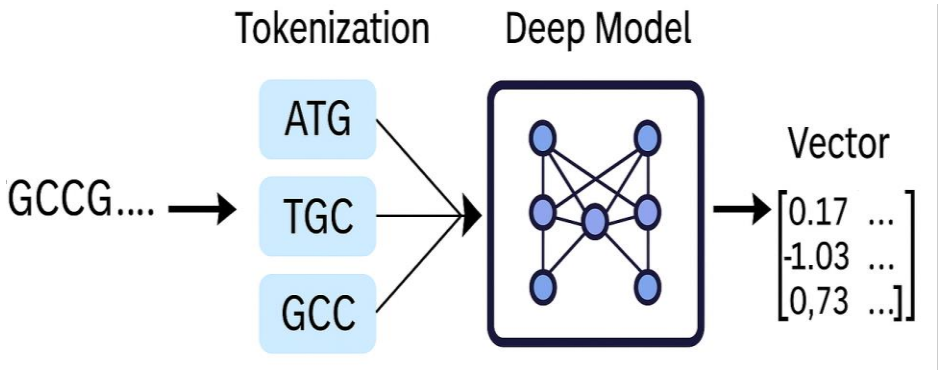
Once we have the DNA reads, how we can extract useful information?

Profile-Based Analysis



- **Quantify** known or reconstructed elements (e.g., species, genes, MAGs)
- Produces sample × feature **matrices** for downstream analysis

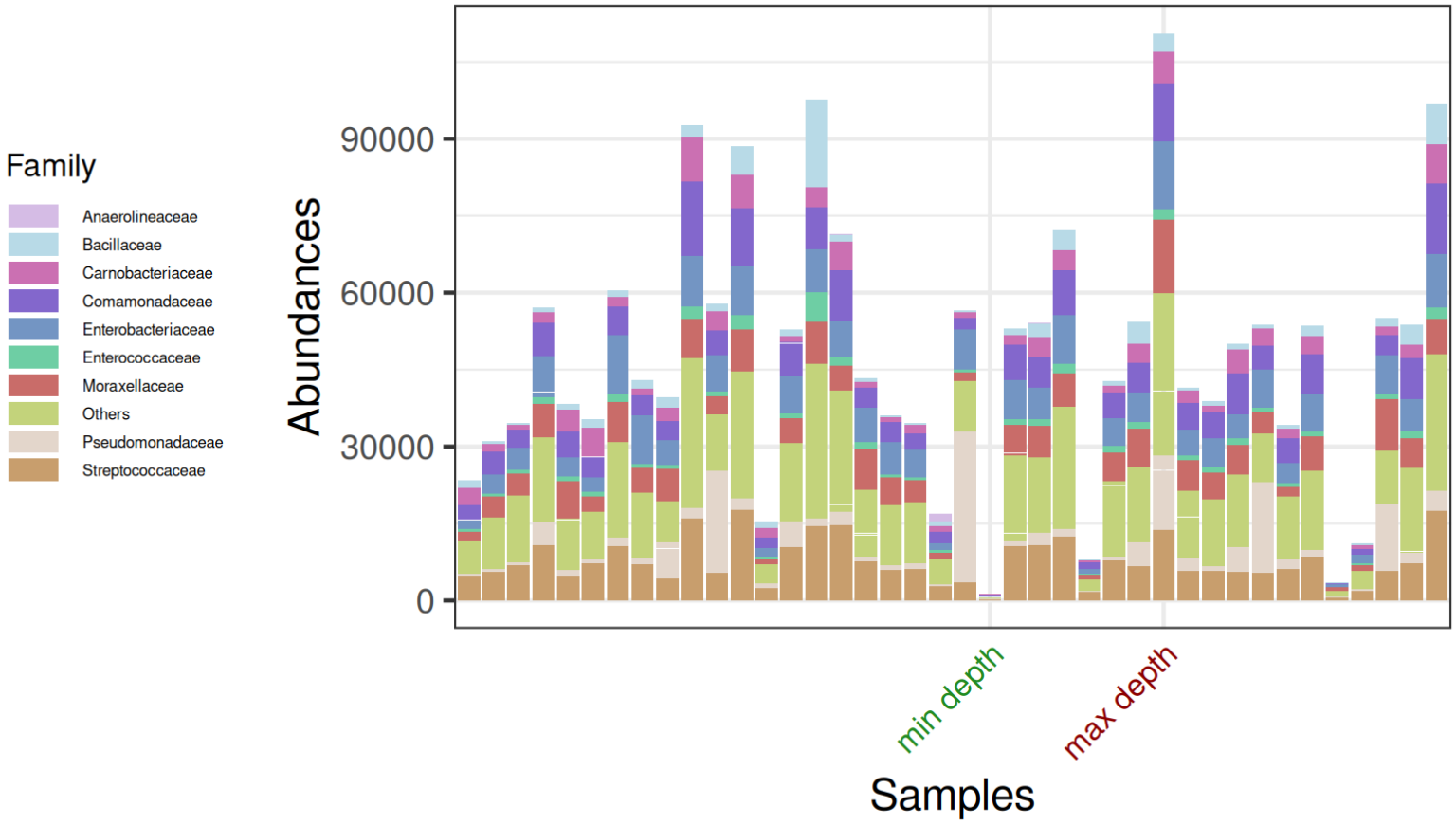
Language-Based Modeling

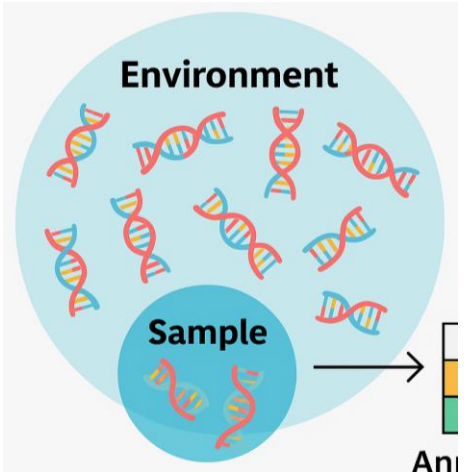


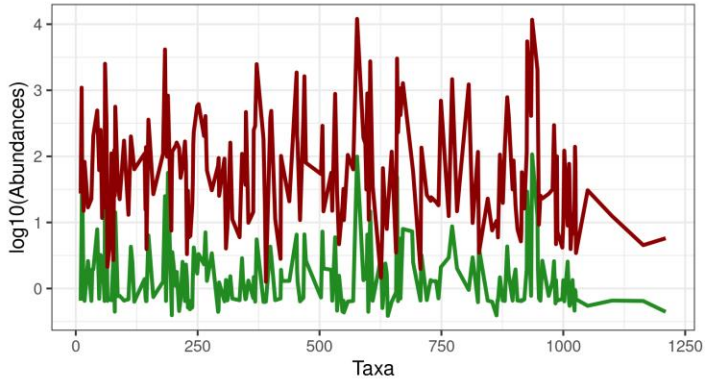
- Tokenize DNA into little pieces and **embed** with deep models.
- Captures structure and semantics, similar to **natural language.**

Absolute Counts Don't Tell the Whole Story

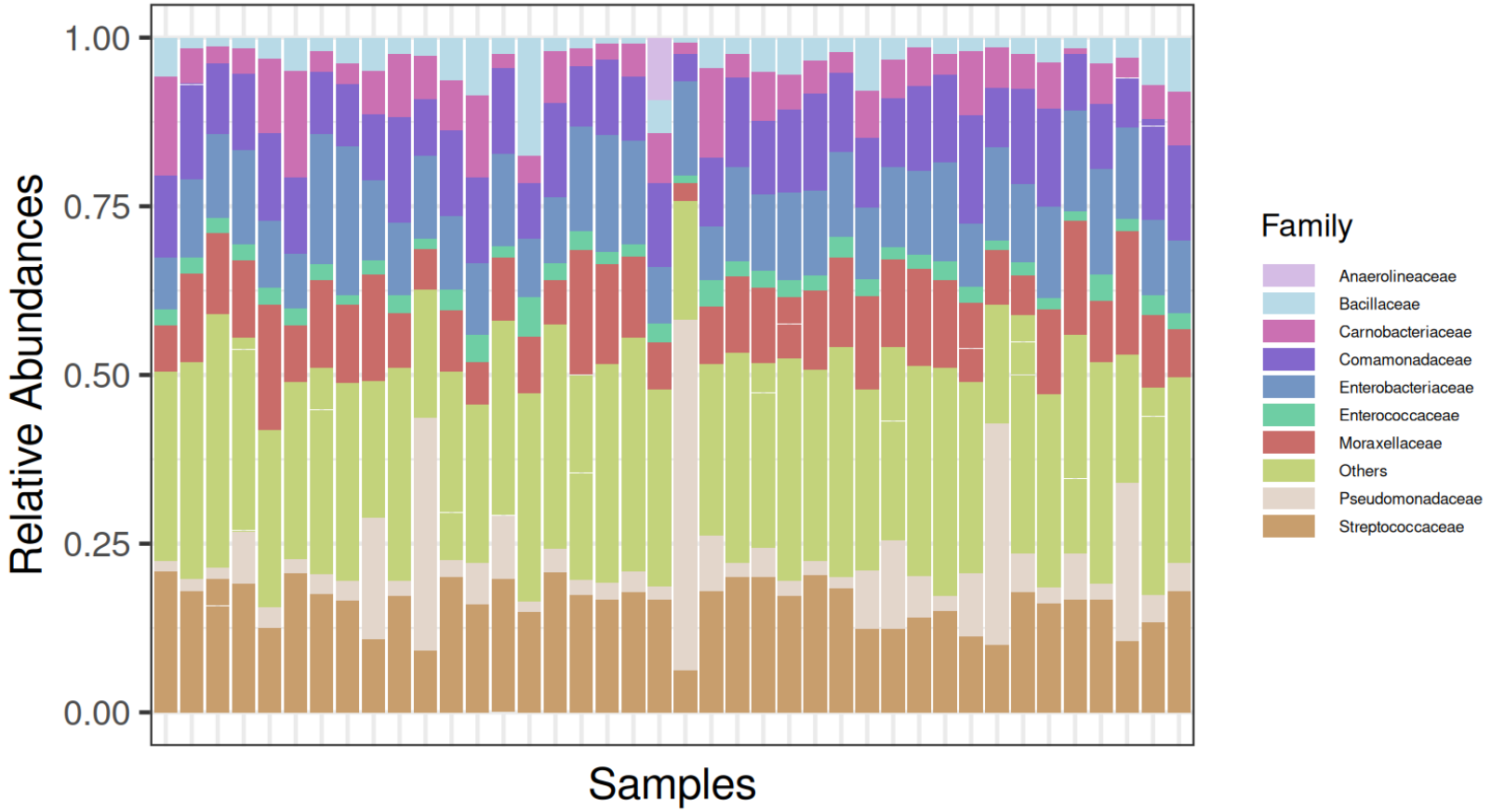
How variations in sequencing depth affect the interpretation of metagenomic profiles







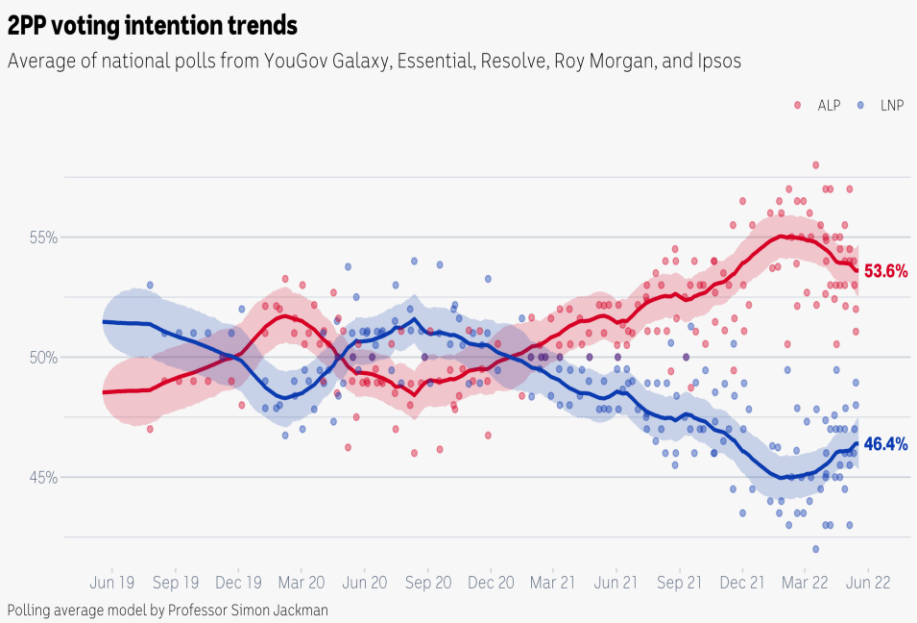
From Counts to Proportions



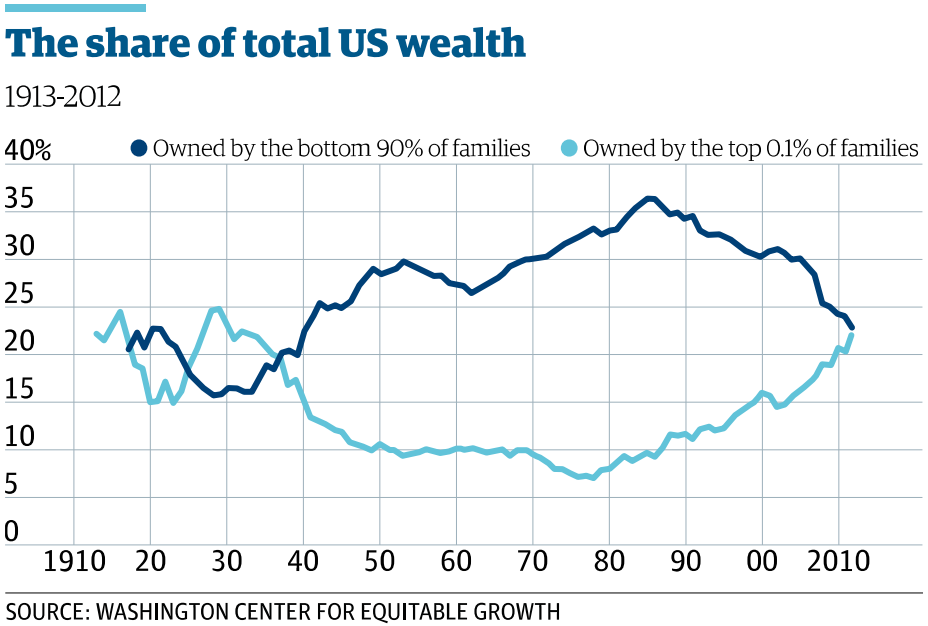
- **Normalization**: Converts absolute abundances to relative values (sum = 1)
- **Comparability**: Makes samples comparable, regardless of sequencing depth

Compositional Datasets: Proportions, Not Absolute Values

Just like voting trends and wealth distribution, metagenomic data represents **proportions**



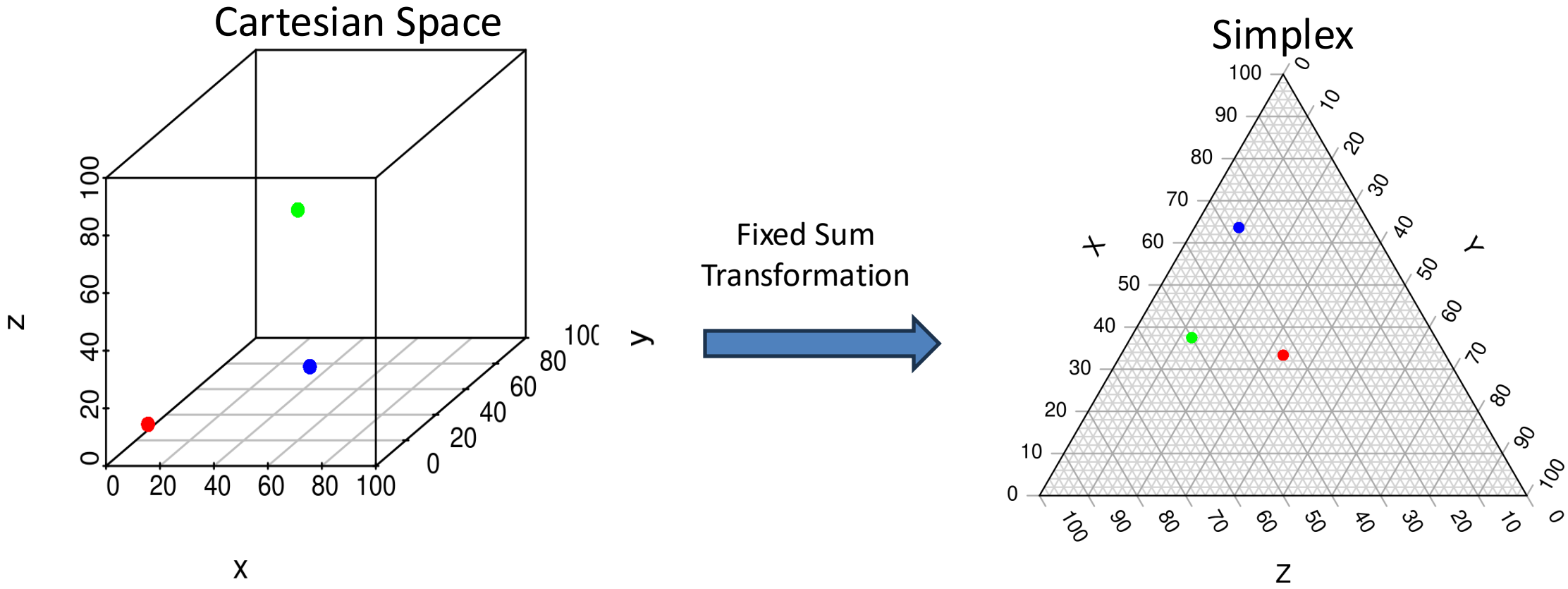
Political voting trends as proportions of the total electorate, where relative changes matter, not absolute numbers.



The share of total wealth held by different income groups, illustrating how proportions reflect the inequality, not total wealth.

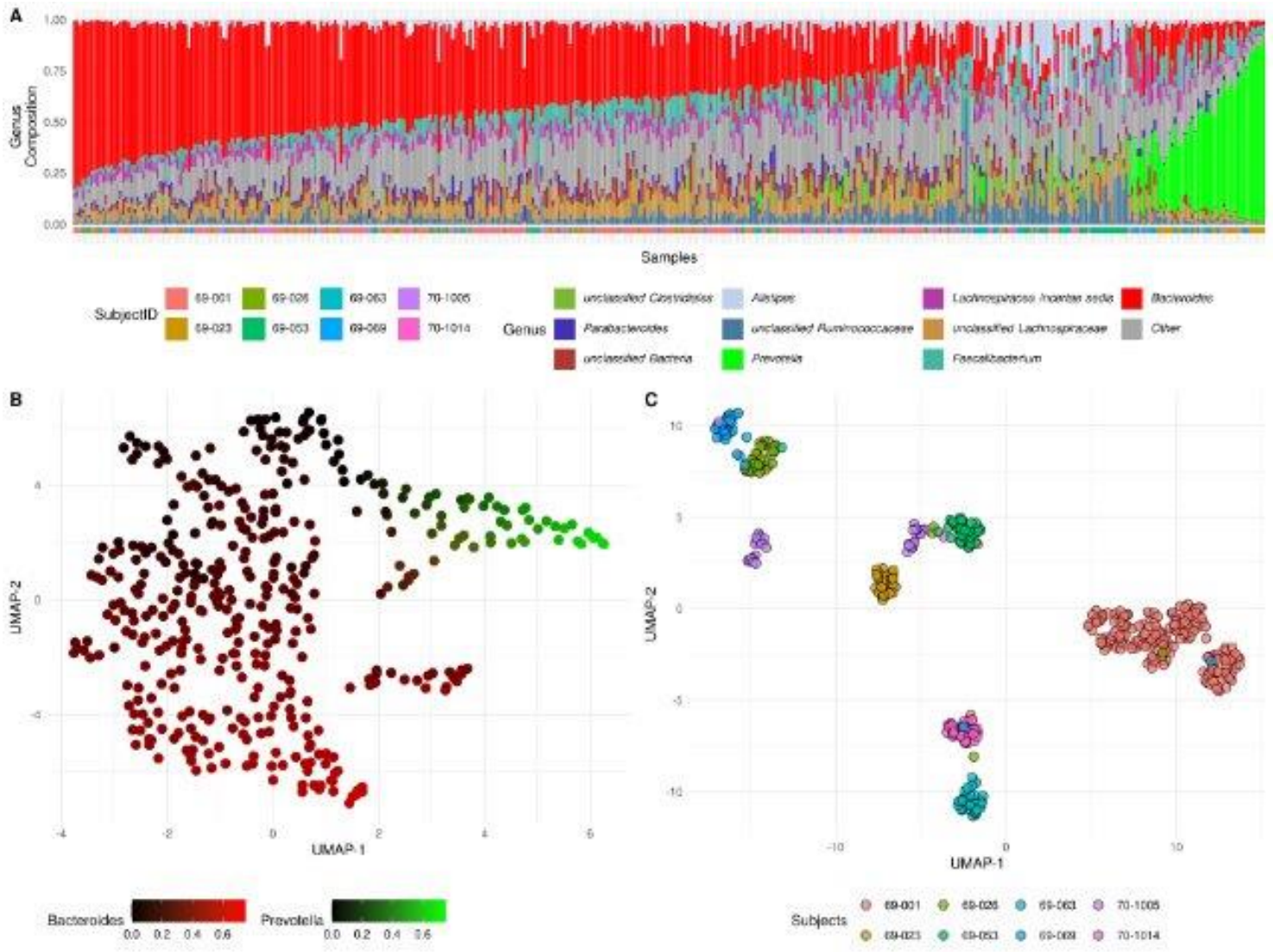
Metagenomics Data Lives on the Simplex

Euclidean Metrics and Compositional Data: A Bad Match



After the sum constraint transformationthe data lie on a simplex and it can be **extremely dangerous** to use **Euclidean** metrics for proximity and correlation estimations.

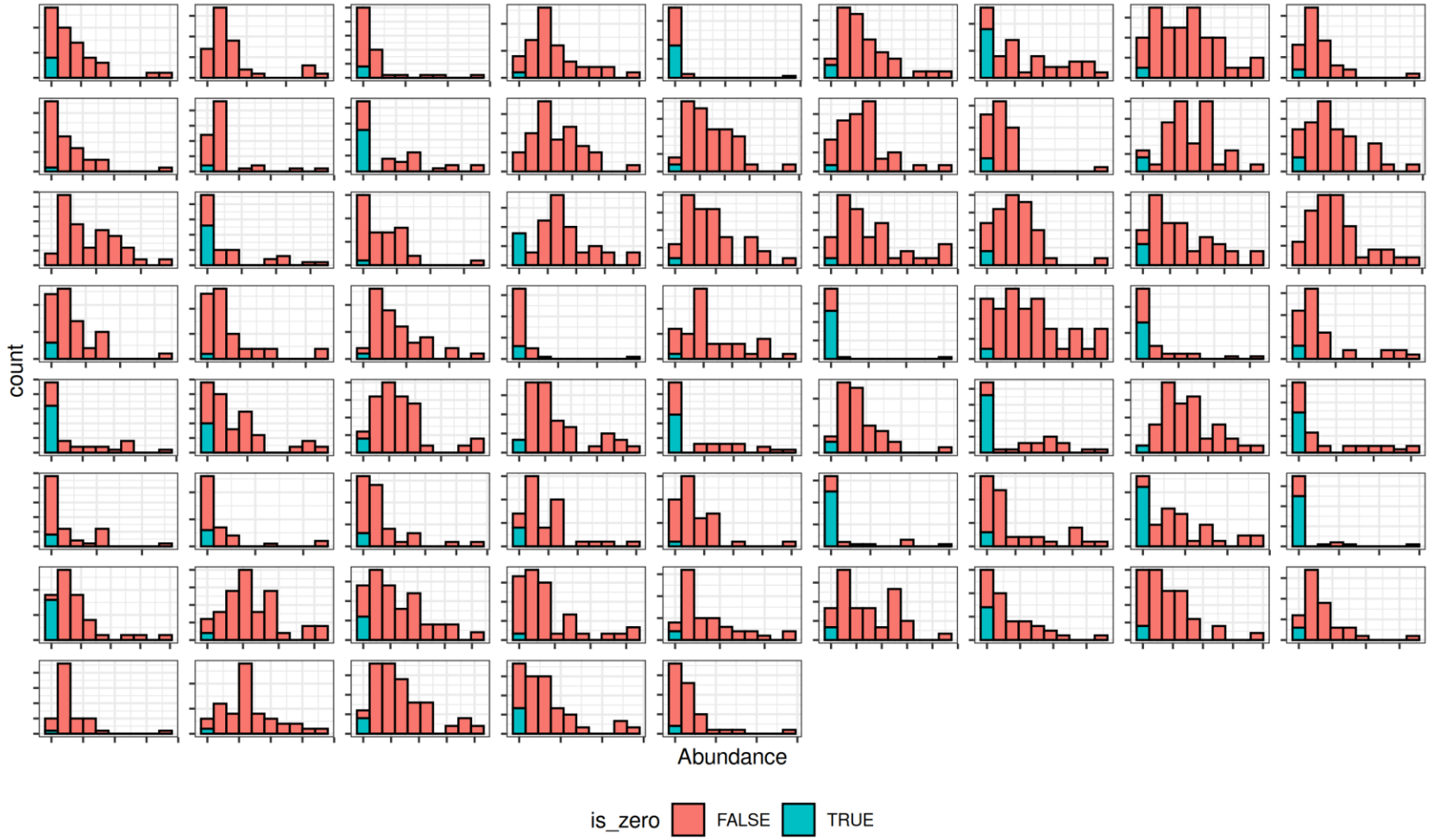
Choosing the Right Metric Changes Everything



Different metrics lead to different structures in microbiome data.
Be cautious: standard distances can distort the real biological variation when applied to compositional data.

Horrible Distributions in Metagenomic Data

Heavy tails, positive values and many zeros completely violate the assumption of normality.

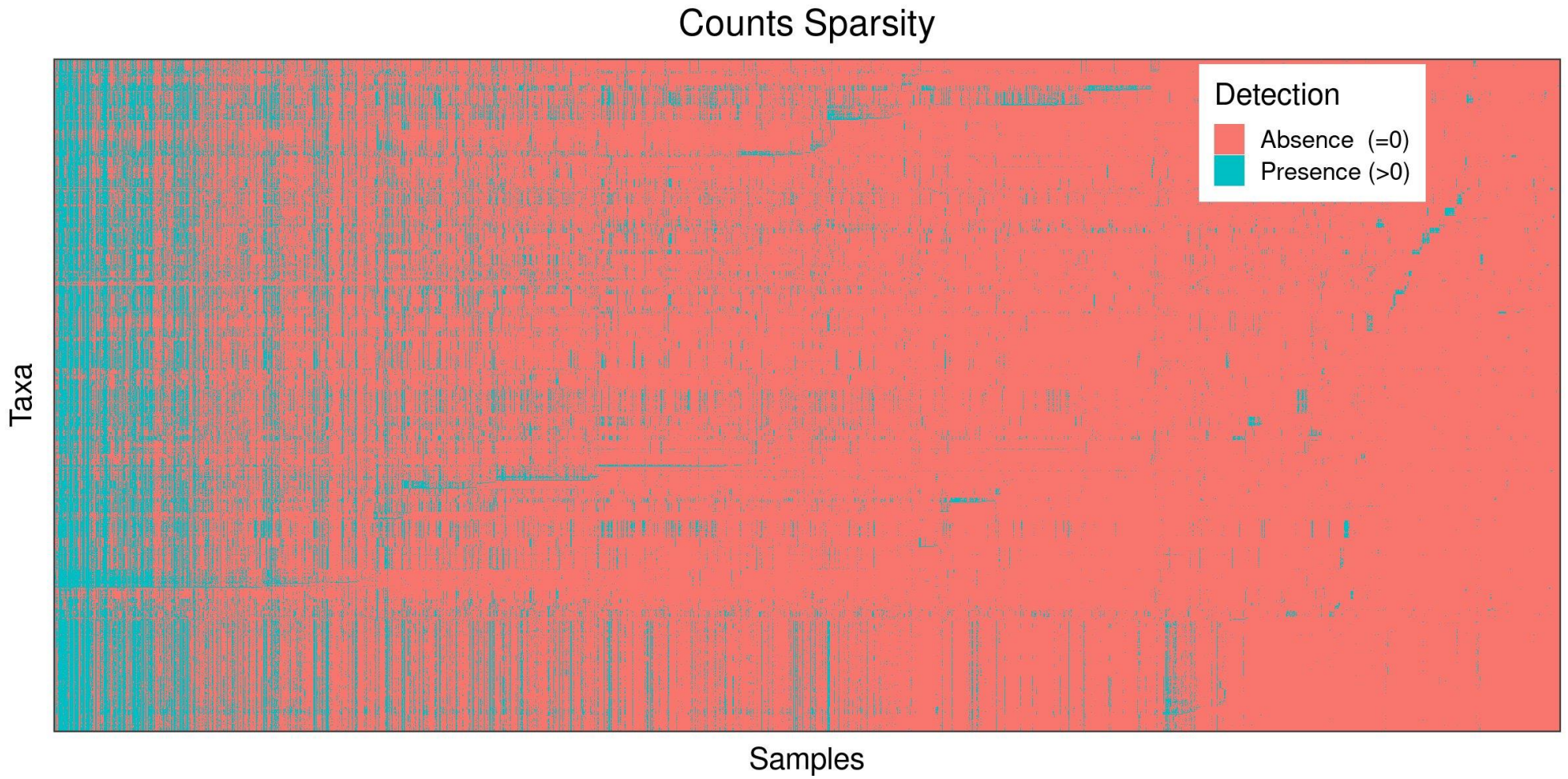


Frequent **zeros** represent values **below the detection** threshold, not the absence of taxa. They should be treated as **missing values** due to **technical limitations** in measurement.

Zeros and heavy tails invalidate the assumptions of many statistical

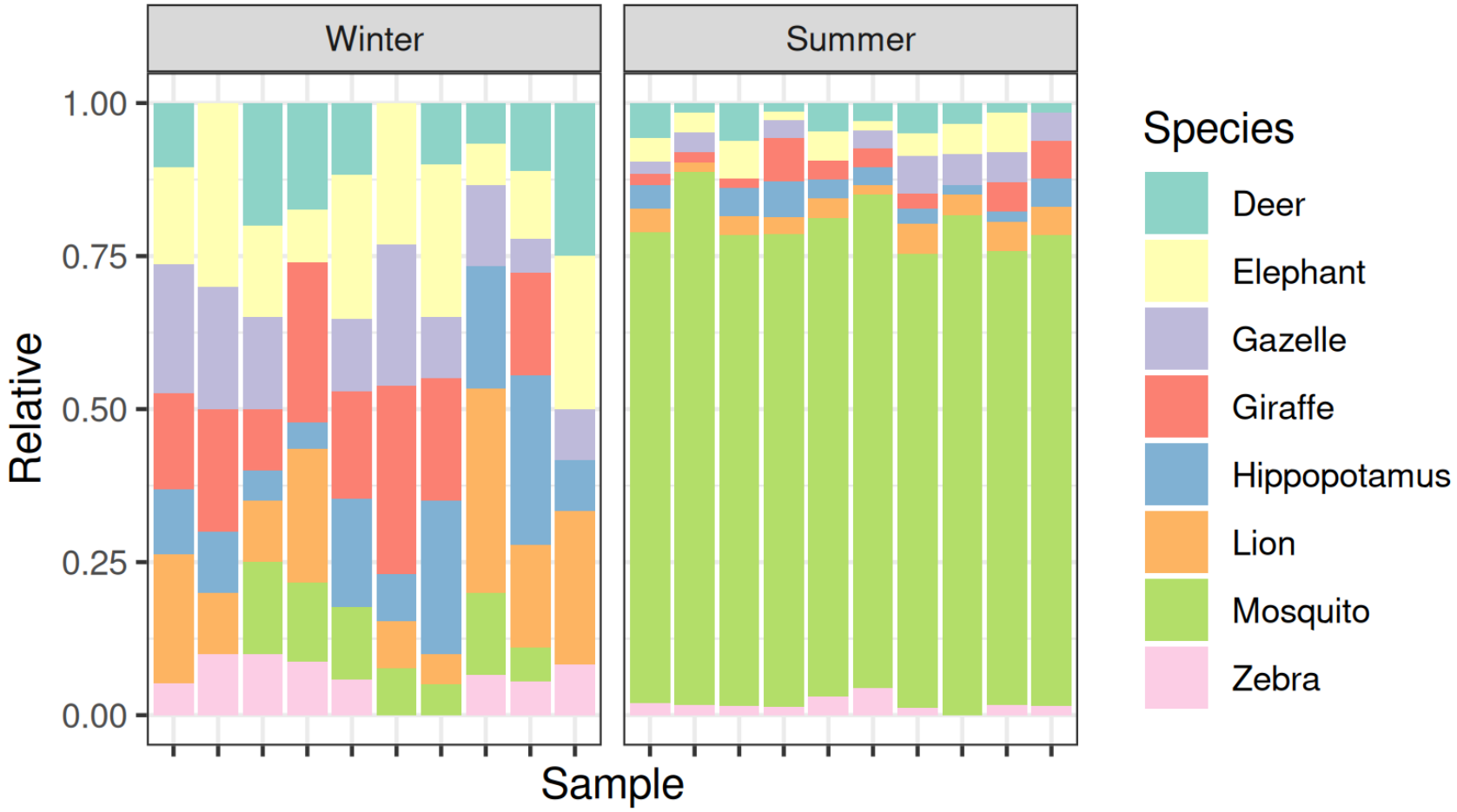
Metagenomics Data Could Be Really Sparse

About 90% of the elements of the count matrix are 0!



Misleading Relationships in Compositional Data

How the fixed-sum constraint can create false relationship between variables.

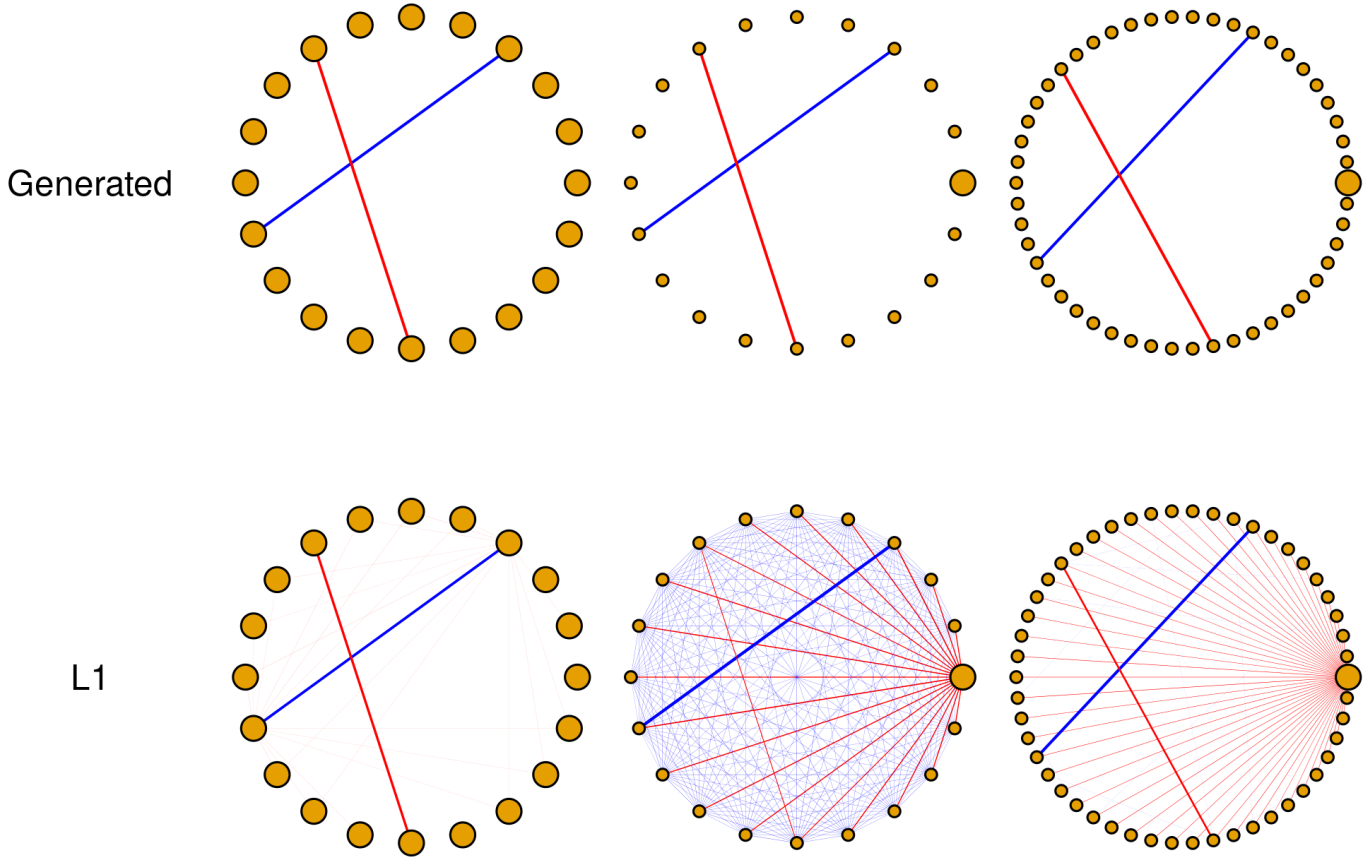


Proportionality Issue: The increase in mosquito abundance doesn't mean other species have decreased. The fixed sum constraint leads to a shift in proportions, not actual decreases.

Interpretation Caution: Relative changes in compositional data are influenced by the fixed sum constraint and may not reflect true biological changes.

False Correlations from Relative Data

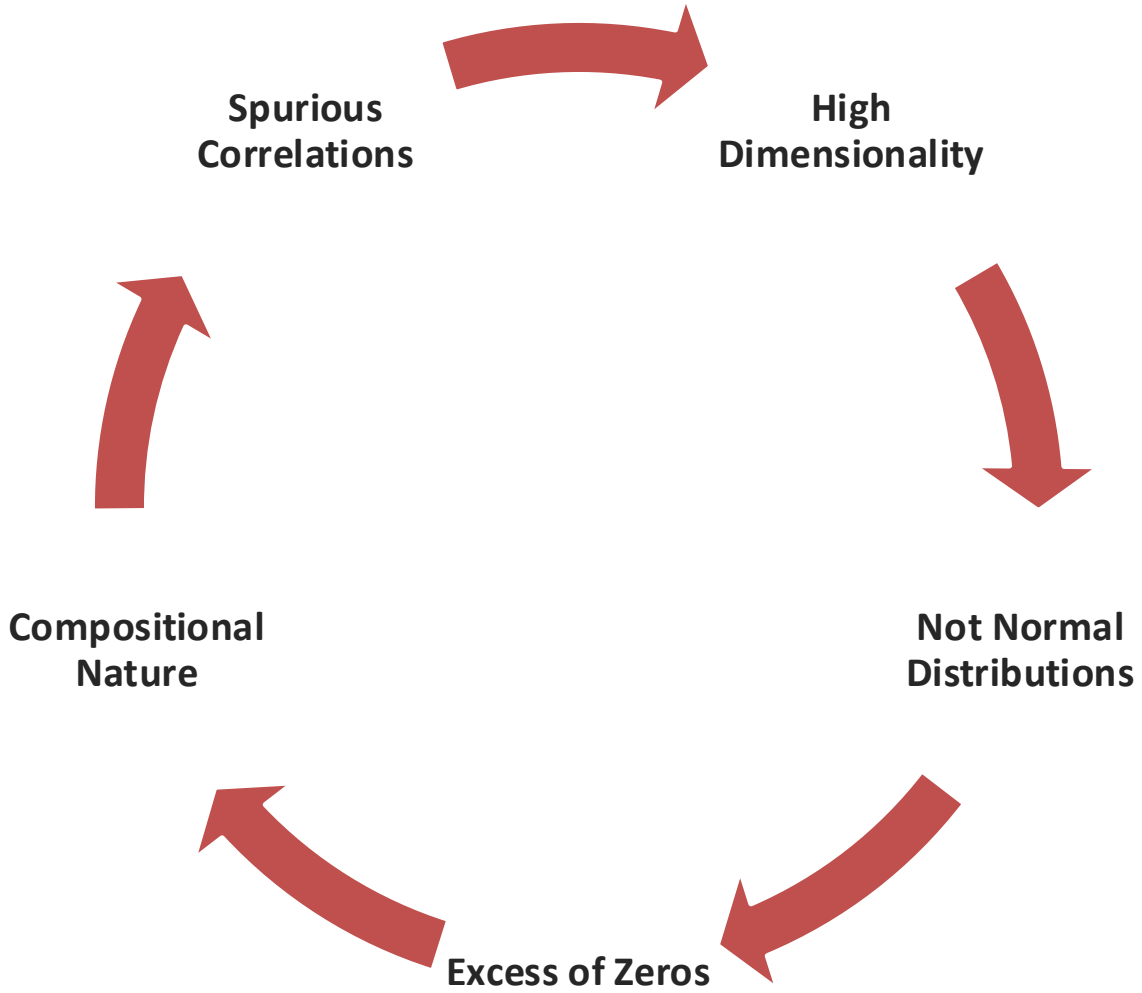
The fixed-sum constraint ties all variables together—one change alters the rest.



Increasing the abundance of a single variable (top row) in compositional data leads to widespread false correlations when applying standard correlation (bottom row). These spurious links are not real but induced by the fixed-sum constraint.

This phenomenon was already described by **Karl Pearson** in 1897 as "spurious correlation" due to common denominators.

Recap On Metagenomic Data Issues

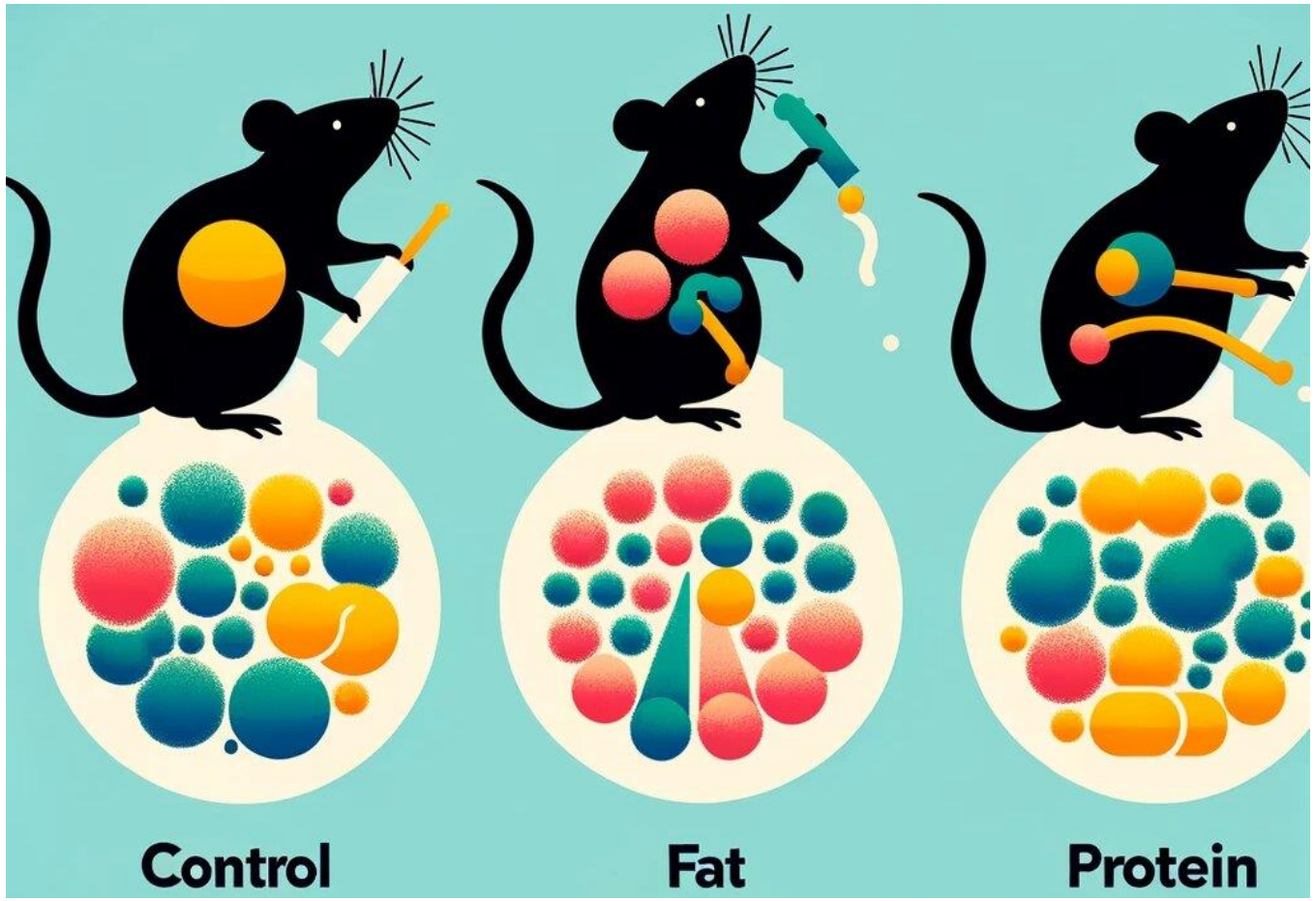


- Requires **filtering** out rare features with excessive zeros to reduce sparsity.
- Requires **normalization** to address both the compositional nature and the heavily skewed distribution of the data.

To properly deal with compositional data, I rely on the theory developed by **John Aitchison** in the 1980s. This framework uses **log-ratio transformations** to correctly handle the constant-sum constraint.

While this is a widely accepted approach, other strategies also exist depending on the context and research goals.

Real Case Study – POMP DO



Objective: To investigate microbiome differences in mice subjected to three different diets.

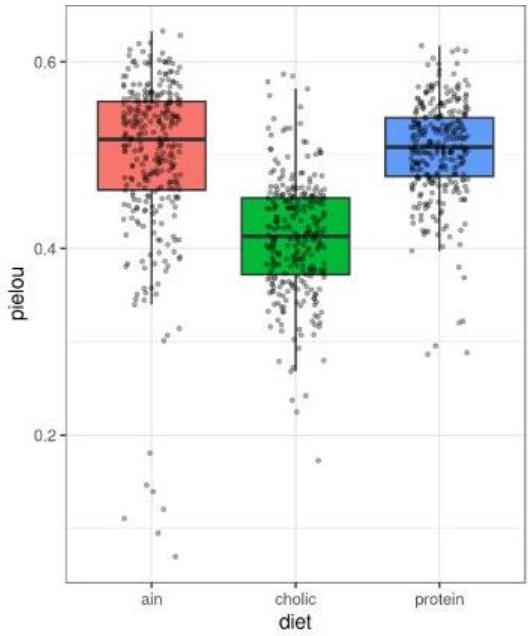
Diets:

- Control: AIN
- High Fat: CHOLICH
- High Protein: PROTEIN

Cohort Size: Over 800 samples.

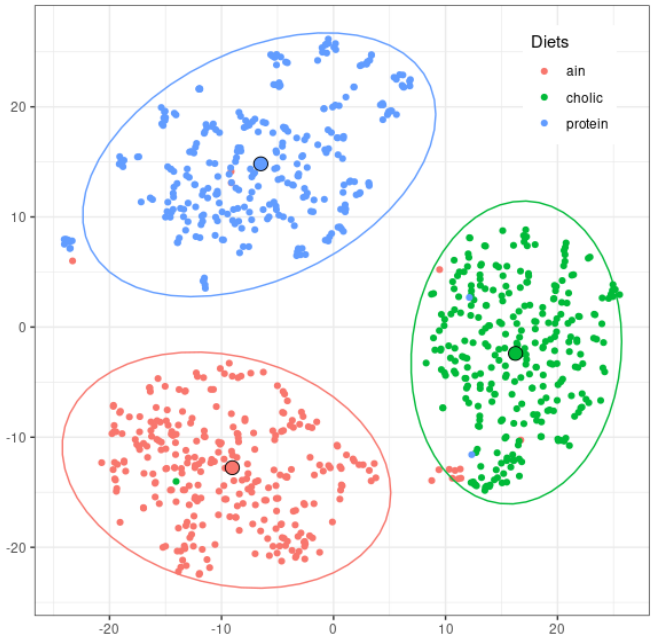
Overview

Alpha Diversity



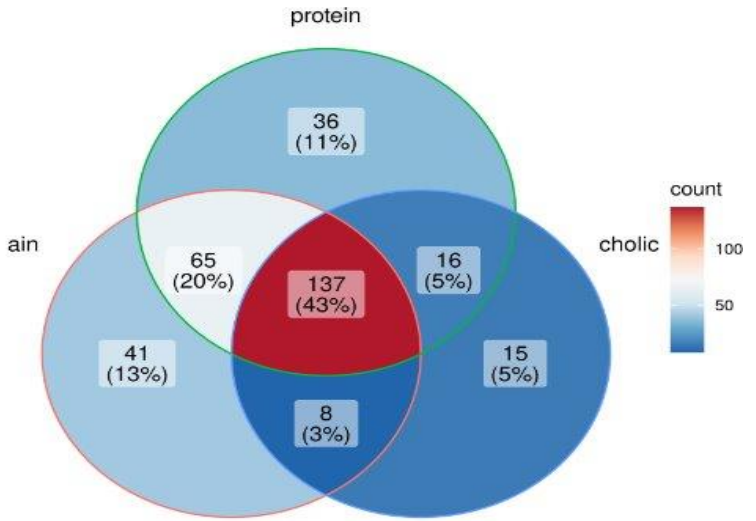
In Cholic diet there is a loss of diversity

Beta Diversity



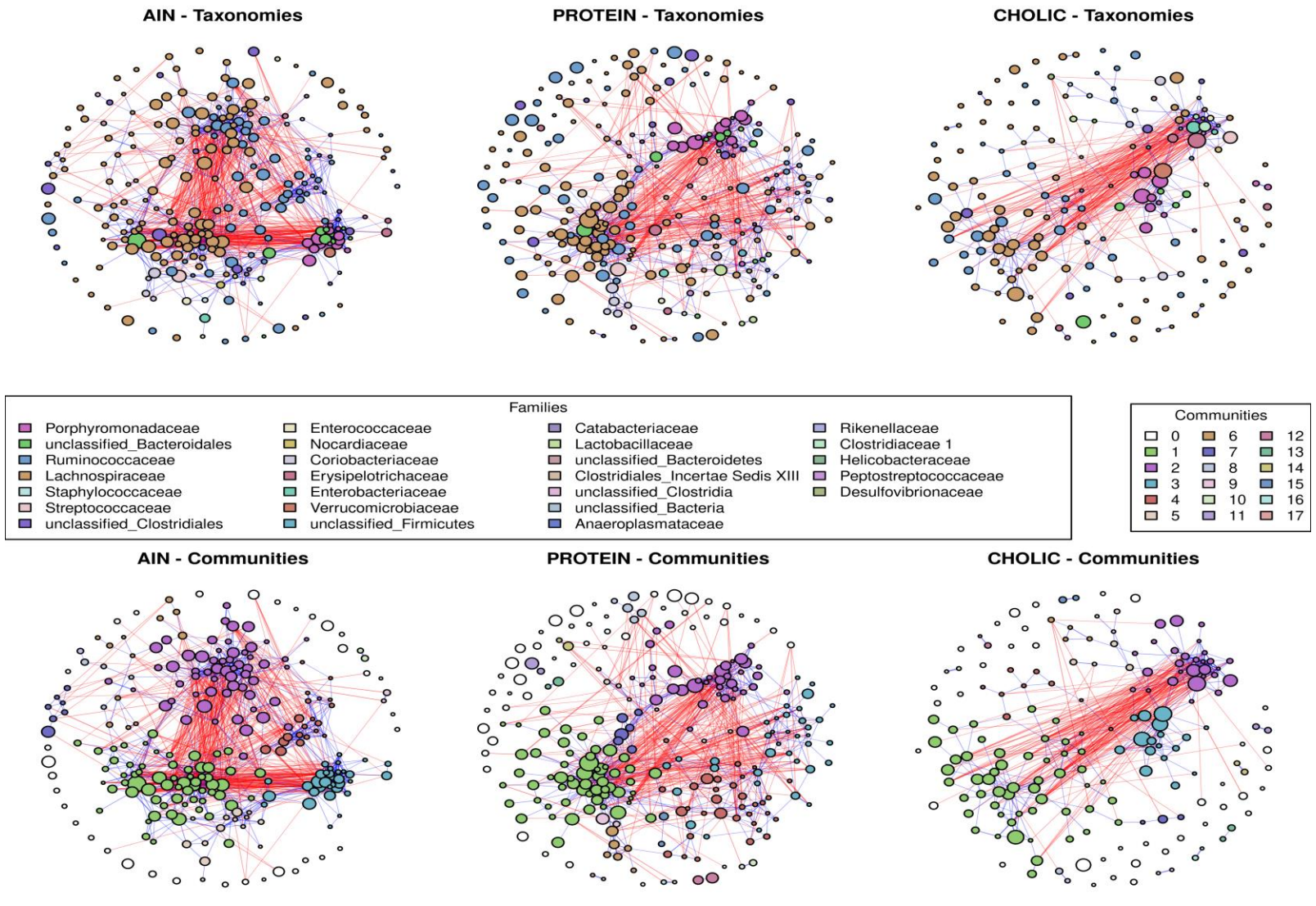
Tsne with Aitchison distance shows a clear distinction between the three diets

Filtering Step



After the filtering step there is a core of shared OTUs (43%) but also others specific for the diets with a larger overlap between control and protein

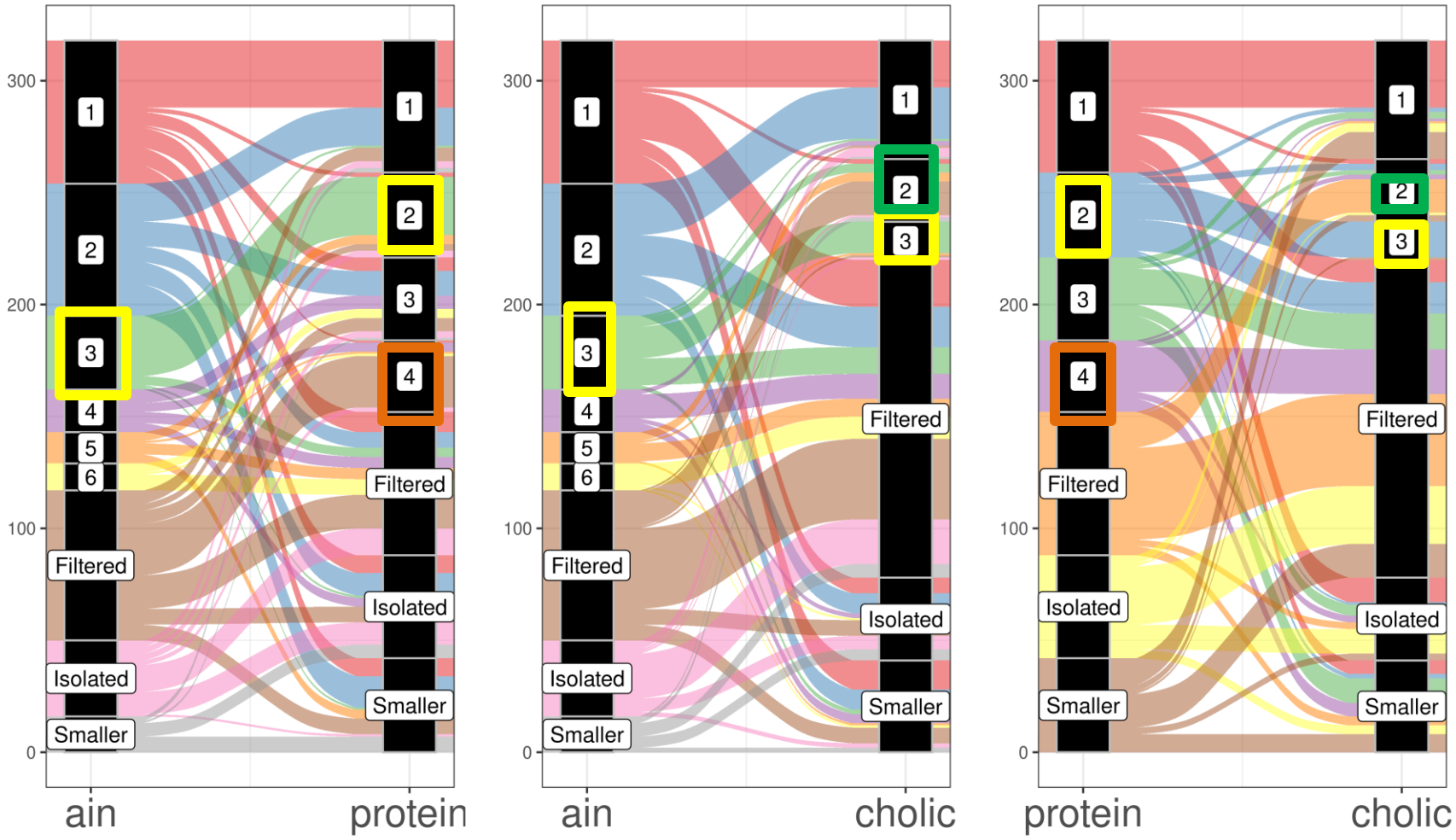
Bacterial Networks



- Top row networks with vertex color associated to the taxonomical families.
- Bottom row same networks but the colors are associated to communities.
- All networks show clear distinction in communities with a modularity >= 0,6.

Bacterial Comunities Across the Diets

- AIN 3, PROTEIN 2, CHOLICH 3 highly conserved
- PROTEIN 4 specific
- CHOLIC 2 Specific



From *Who's There?* to *What Are They Doing?*

Predicting Metabolic Functions from Microbial Communities

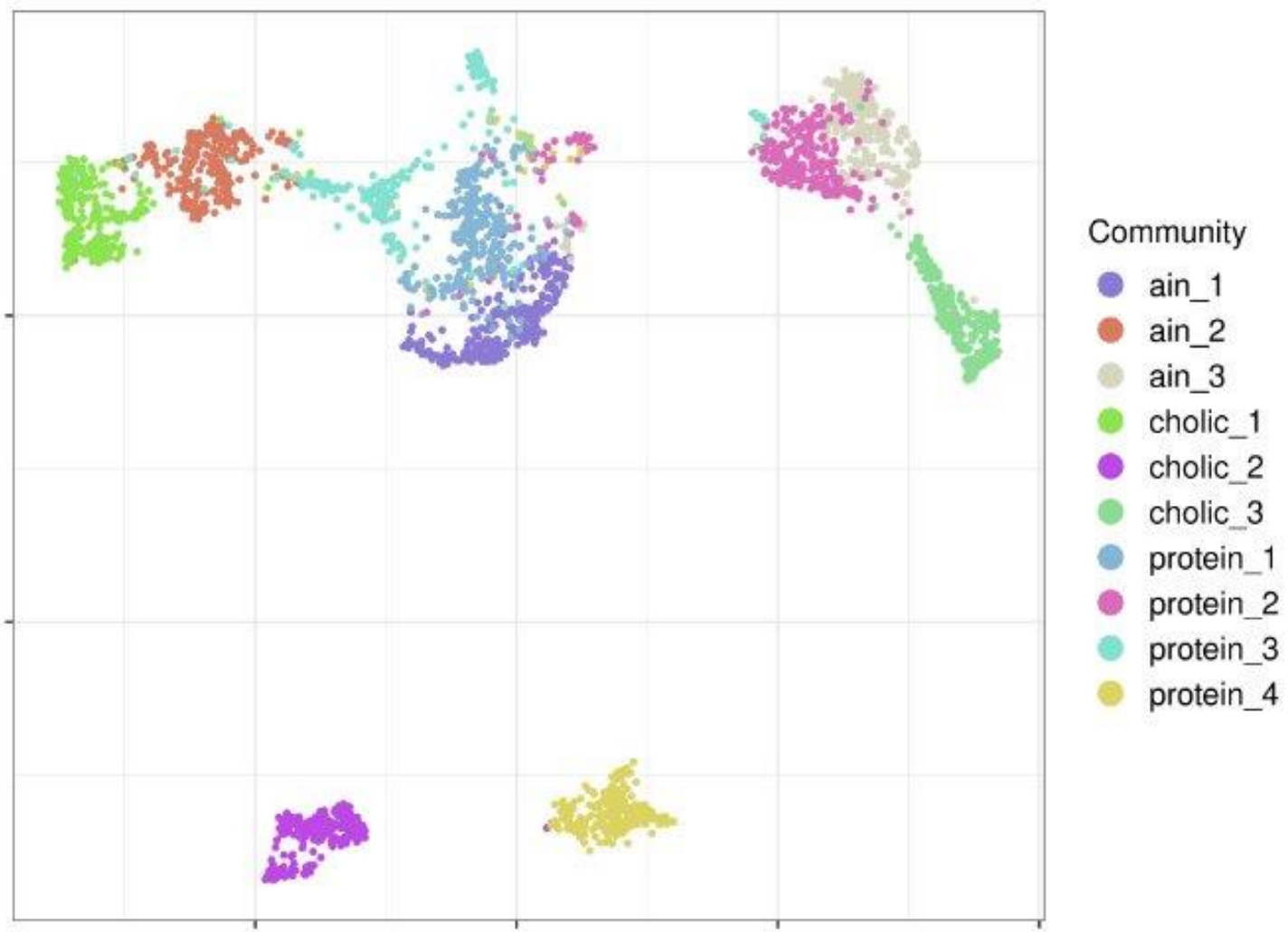
- After identifying the microbial composition, I inferred **metabolic pathways** to understand the **functional potential** of each community.

- This marks a conceptual shift from **"Who are you?"** (bacterial profiles) to **"What do you do?"** (metabolic functions).

- I used **PICRUSt2** to predict pathways, running the analysis in parallel on the **OPH HPC cluster** at UniBo.

- Thanks to parallelization, it took just **1 hour** on the cluster — versus **a few days** on a standard laptop.

Real Case Study – POMP DO Metabolic Pathways

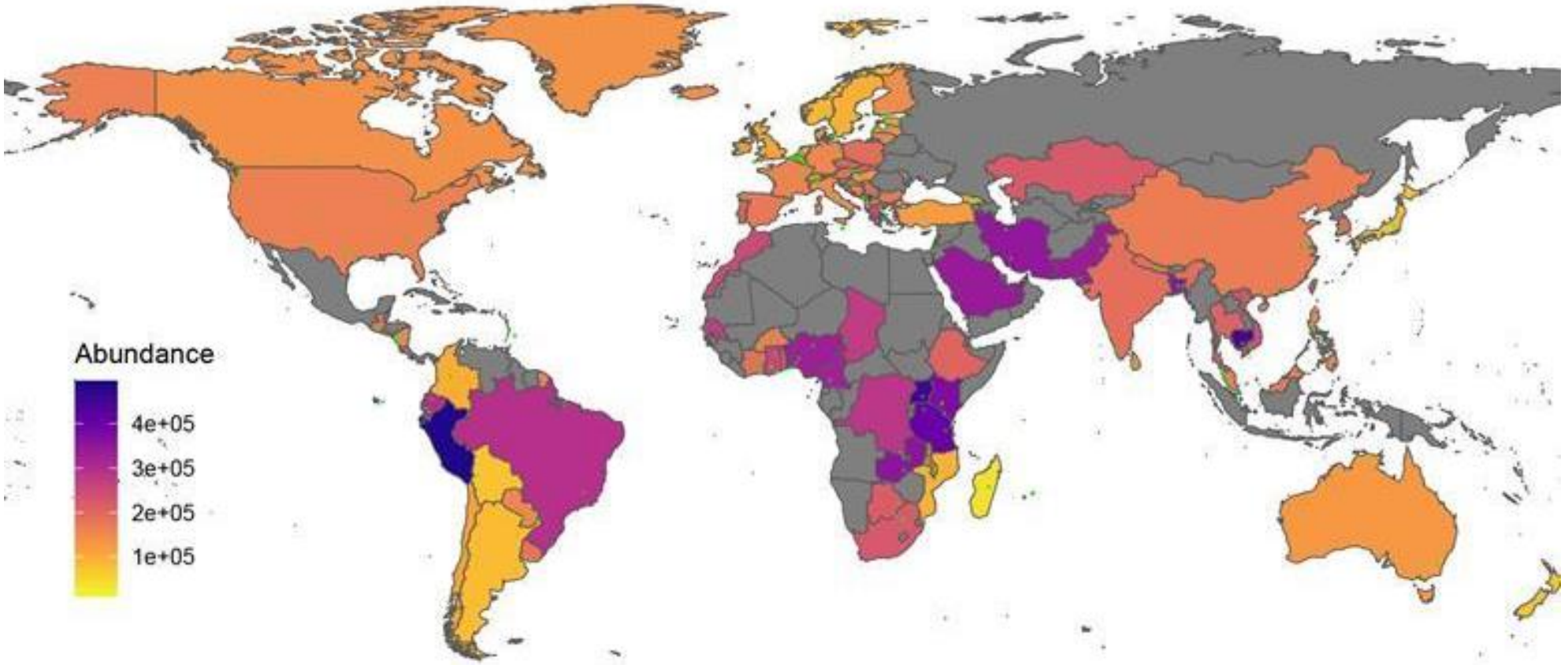


Methodology: Extracted metabolic pathways from bacterial communities using PICRUSt2.

Key Findings:

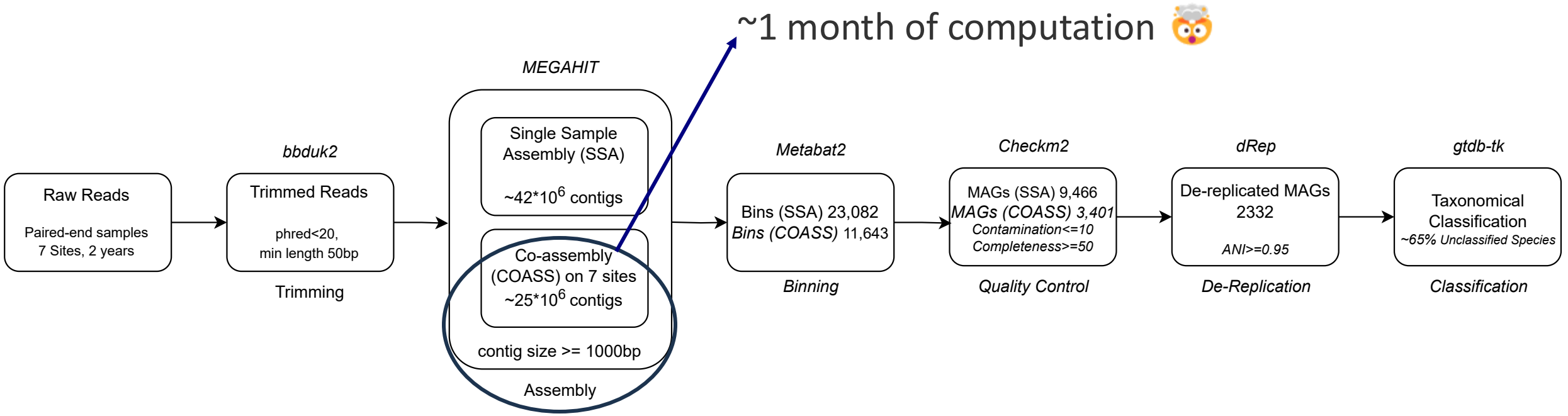
- **Isolated Communities:** Two distinct islands represent unique communities specific to the network diet.
- **Major Island:** One large island with three highly conserved communities across all diets.
- **Macro Island:** A significant grouping of various communities, each linked to different dietary conditions.

Real Case Study – Map Of Antimicrobial Resistance Genes



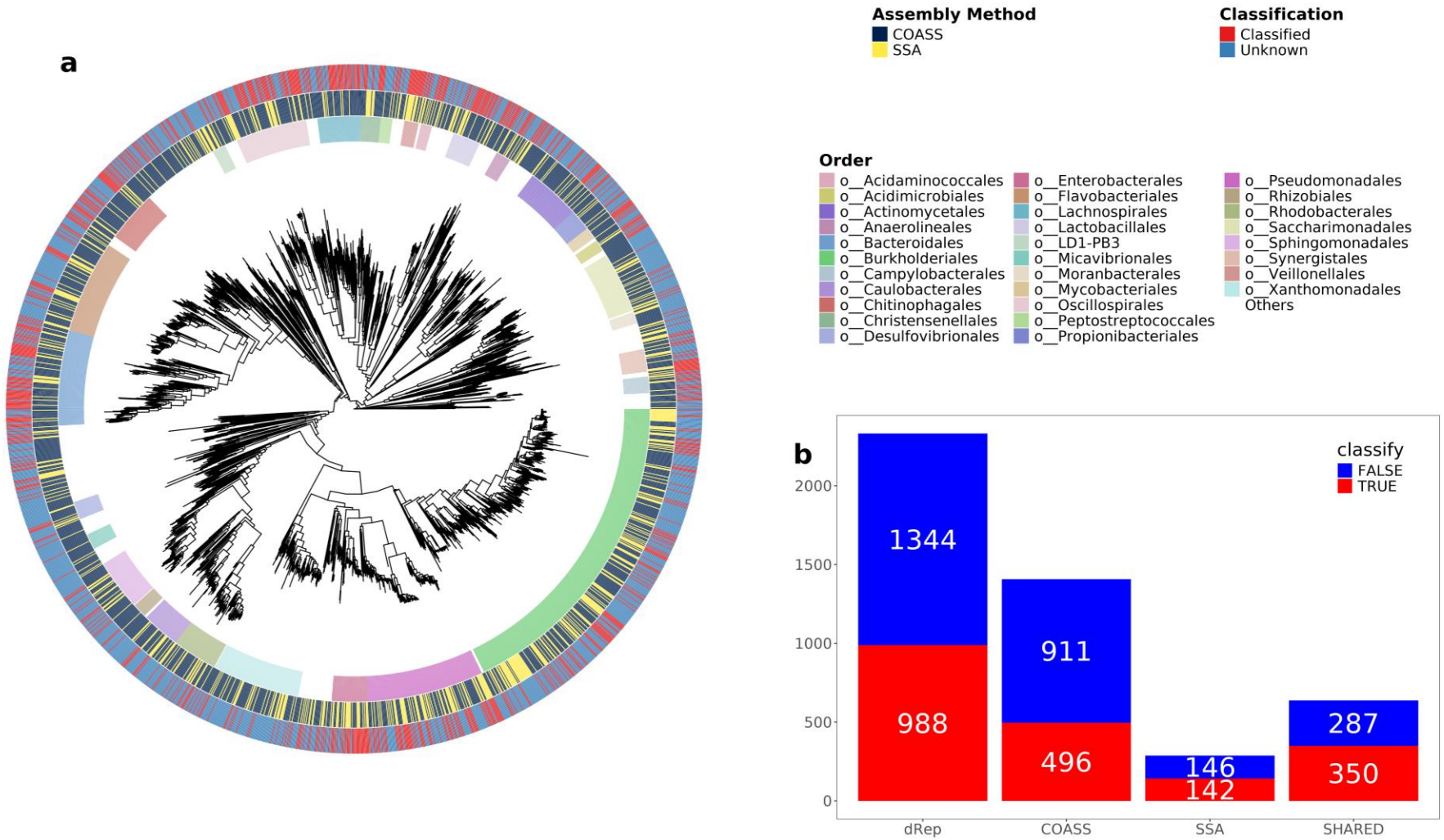
Reconstructed Bacterial Genomes from Sewage

A glimpse into the bioinformatics behind large-scale genome recovery



In metagenomics, **assembly** is the process of merging millions of short sequencing reads into longer contiguous sequences (*contigs*), aiming to reconstruct microbial genomes. It's a standard yet computationally intensive step. In this project, the assembly alone took **over one month** on **40 CPU cores** and **1 TB of RAM** using the **Computerome supercomputer** in Denmark.

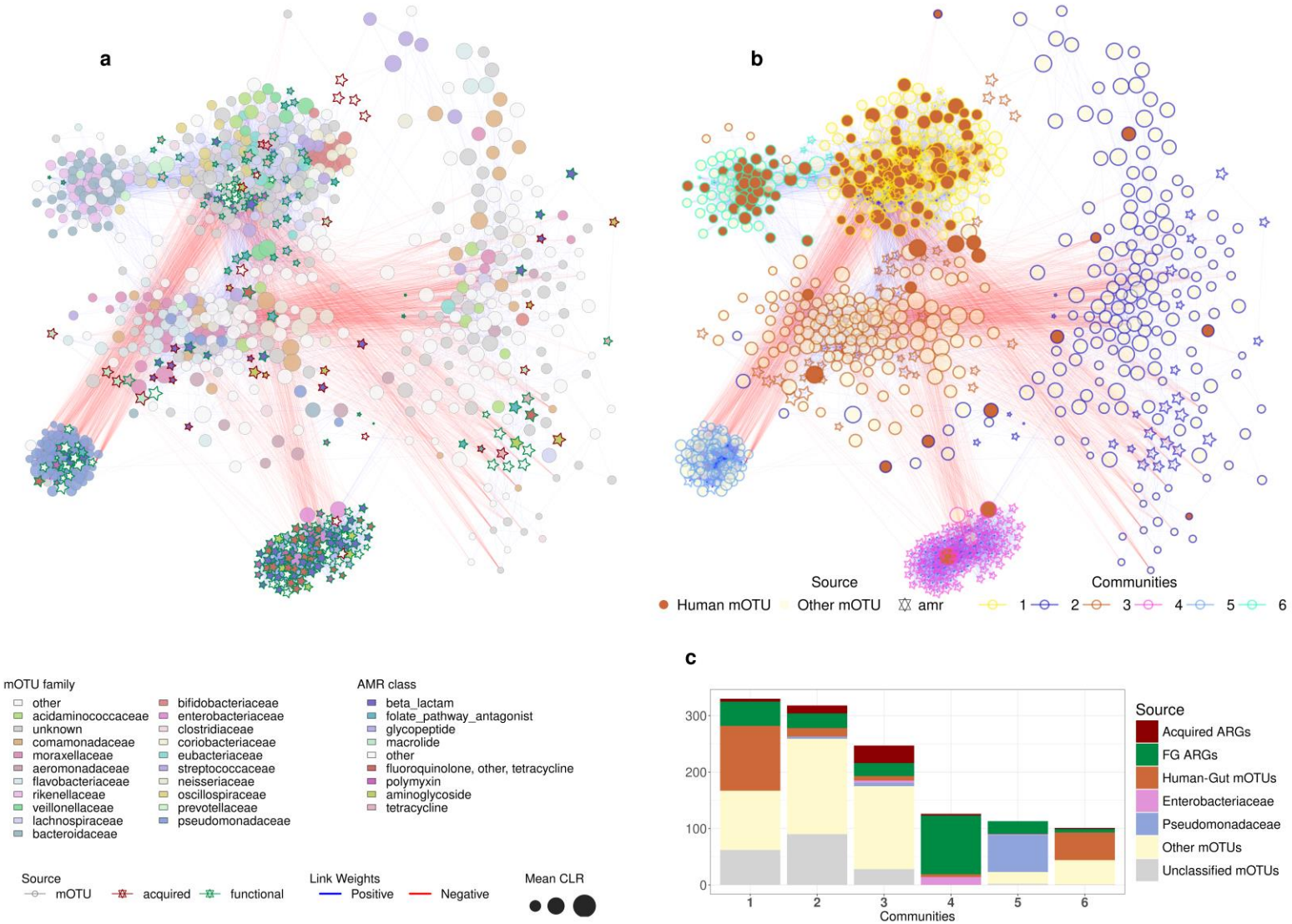
2,332 MAGs Reconstructed — Many Still Unknown



- ~65% of MAGs are unclassified species
- Genomic diversity reflects the complexity of urban wastewater

Bacteria–AMR Network of the Global Sewage

A community of Enterobacteriaceae shows strong links with resistance genes



- Network linking bacterial taxa and antimicrobial resistance (AMR) genes.
- A dense **Enterobacteriaceae** community is clearly associated with multiple AMR genes.
- Highlights possible hotspots of resistance transmission and co-selection.

Key Insights from Today

- Metagenomic data are **compositional** — absolute counts can be misleading
- Proper **normalization** and modeling are crucial for meaningful analysis
- **Dimensionality reduction** and **networks** help uncover biological patterns
- **Bioinformatic pipelines** and HPC enable large-scale processing
- **Interpretation requires awareness** of both technical and biological constraints

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Thanks for your attention

Questions?