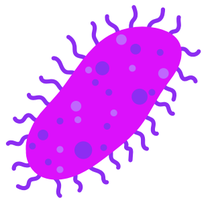
Contamination removal and taxonomic classification

Jeferyd Yepes García

SIB Swiss Institute of Bioinformatics

UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Overview



Removal of contaminants

- What are contaminants?
- Why remove them?
- Tools to remove sequences.

Today’s data

- Experiments.
- Pipeline.

Taxonomic classification

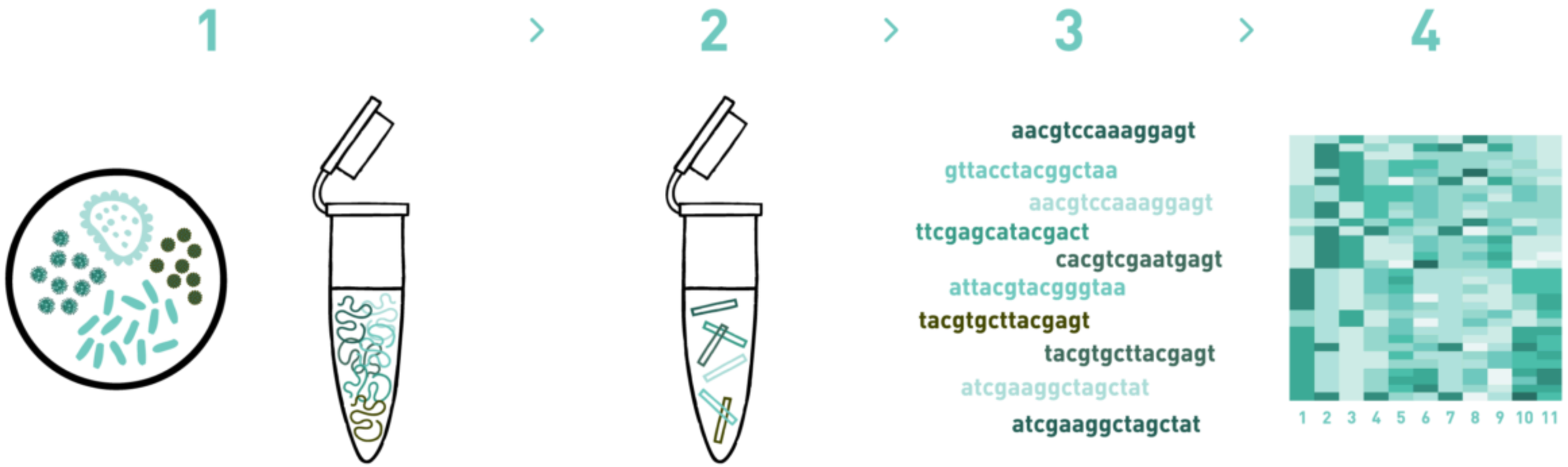
- Tools.
- Kraken2.
- Bracken.

Downstream analysis

- BIOM table.
- Phyloseq object.
- α-diversity.
- β-diversity.
- Ordination methods.

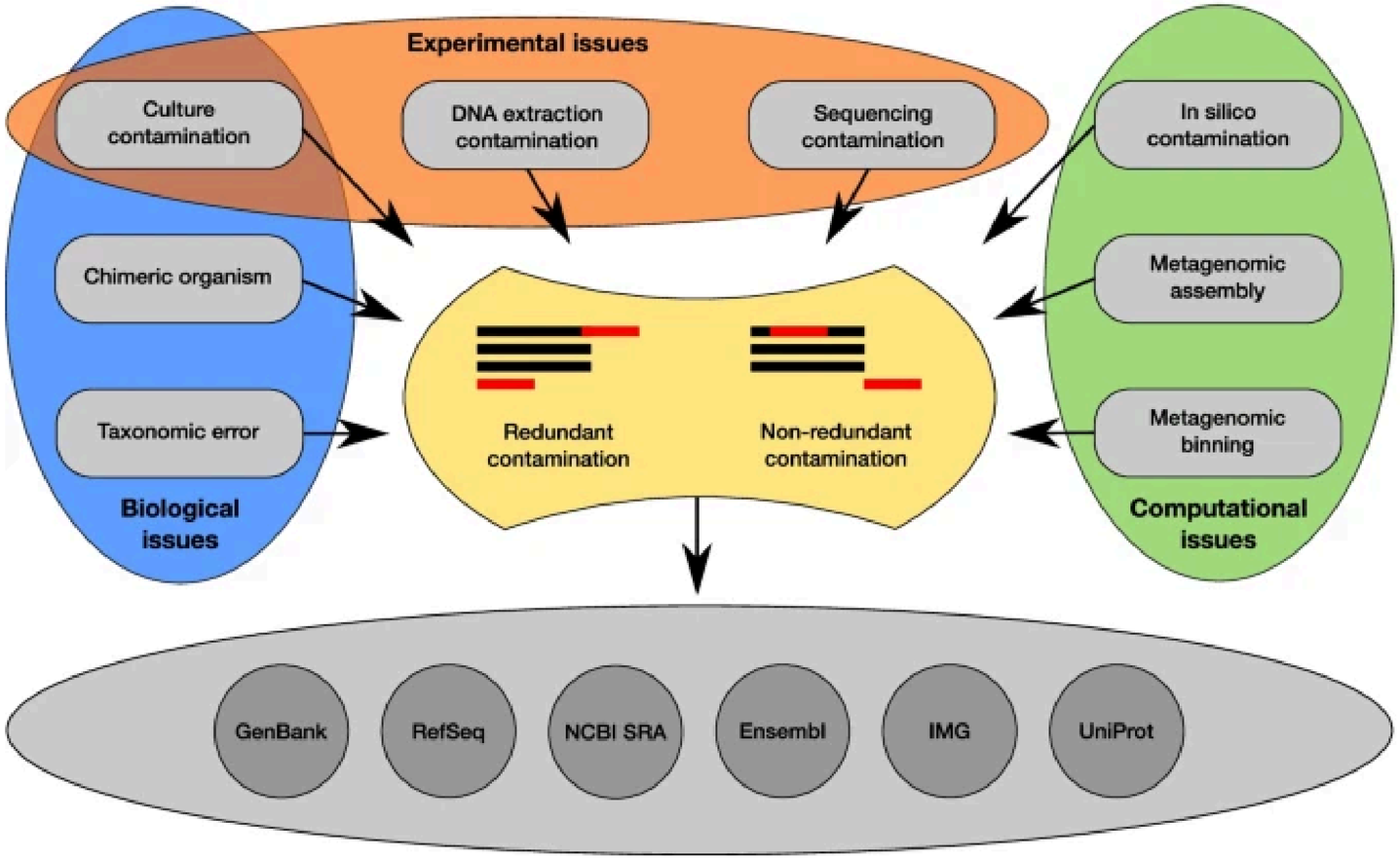
Removal of contaminants

What are contaminants?



Removal of contaminants

What are contaminants?



Removal of contaminants

Problems caused by these sequences:

- **Misleading Taxonomic Profiles:** False positives, inflated diversity.

Removal of contaminants

Problems caused by these sequences:

- **Misleading Taxonomic Profiles:** False positives, inflated diversity.
- **Erroneous Functional Profiles:** Misassignment/overestimation of functions.

Removal of contaminants

Problems caused by these sequences:

- **Misleading Taxonomic Profiles:** False positives, inflated diversity.
- **Erroneous Functional Profiles:** Misassignment/overestimation of functions.
- **Biased Abundance Estimates:** Skewed abundances, false dominance.

Removal of contaminants

Problems caused by these sequences:

- **Misleading Taxonomic Profiles:** False positives, inflated diversity.
- **Erroneous Functional Profiles:** Misassignment/overestimation of functions.
- **Biased Abundance Estimates:** Skewed abundances, false dominance.
- **Increased Computational Load:** Resource waste, algorithm performance.

Removal of contaminants

Problems caused by these sequences:

- **Misleading Taxonomic Profiles:** False positives, inflated diversity.
- **Erroneous Functional Profiles:** Misassignment/overestimation of functions.
- **Biased Abundance Estimates:** Skewed abundances, false dominance.
- **Increased Computational Load:** Resource waste, algorithm performance.
- **Money thrown away.**

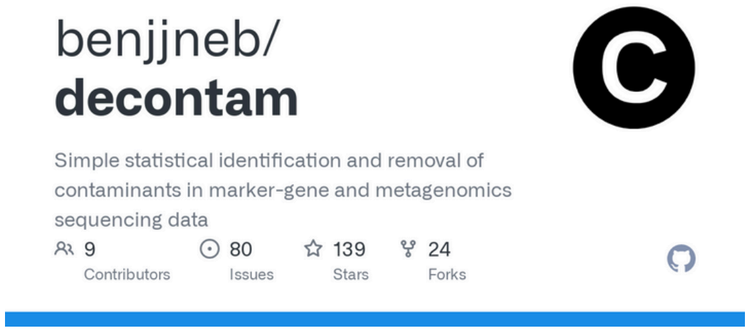
Removal of contaminants

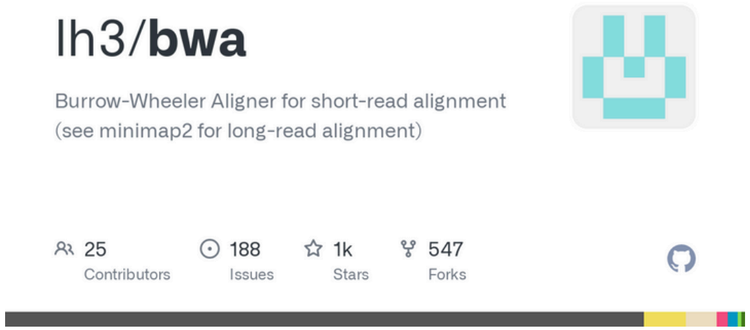
Problems caused by these sequences:

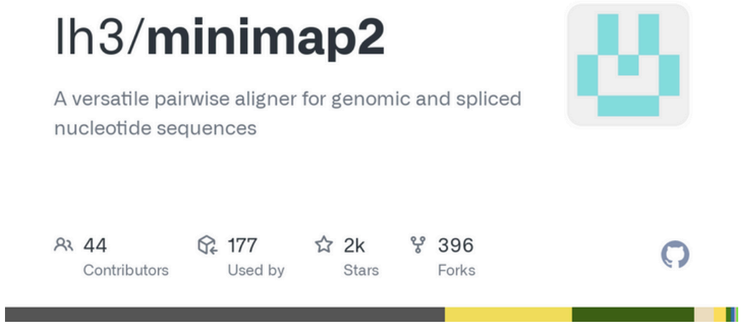
- **Misleading Taxonomic Profiles:** False positives, inflated diversity.
- **Erroneous Functional Profiles:** Misassignment/overestimation of functions.
- **Biased Abundance Estimates:** Skewed abundances, false dominance.
- **Increased Computational Load:** Resource waste, algorithm performance.
- **Money thrown away.**
- **Reproducibility issues.**

Removal of contaminants

Tools to remove sequences:

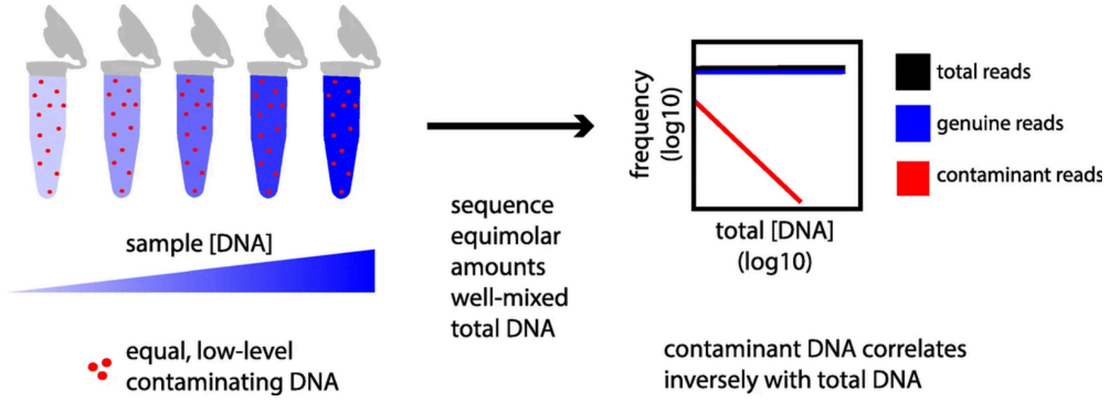




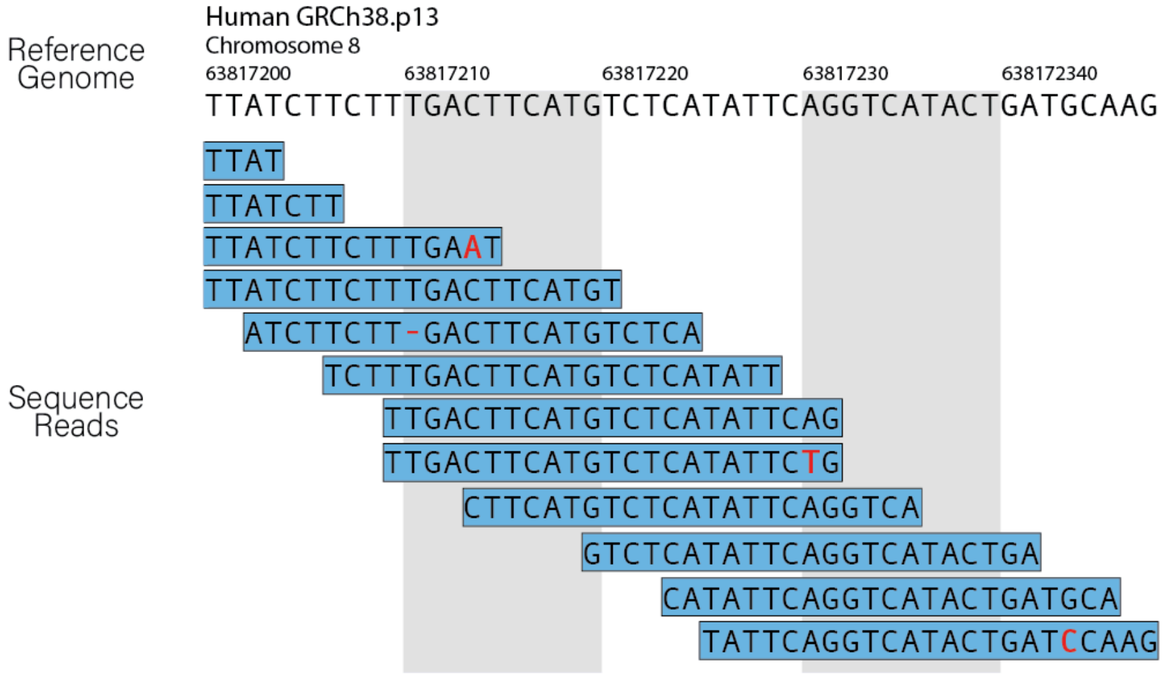


Removal of contaminants

Tools to remove sequences:



Statistical approach



Alignment against indexed genomes

Taxonomic classification

Tools:

biobakery/
MetaPhlAn

MetaPhlAn is a computational tool for profiling the composition of microbial communities from metagenomic shotgun sequencing data

19 Contributors · 11 Used by · 277 Stars · 77 Forks

Unique clade-specific marker genes

DaehwanKimLab/
centrifuge

Classifier for metagenomic sequences

8 Contributors · 124 Issues · 234 Stars · 73 Forks

Indexed genomes

KAIJU

Reference protein database

MEGAN6

Analysis and visualization

NCBI or SILVA taxonomy

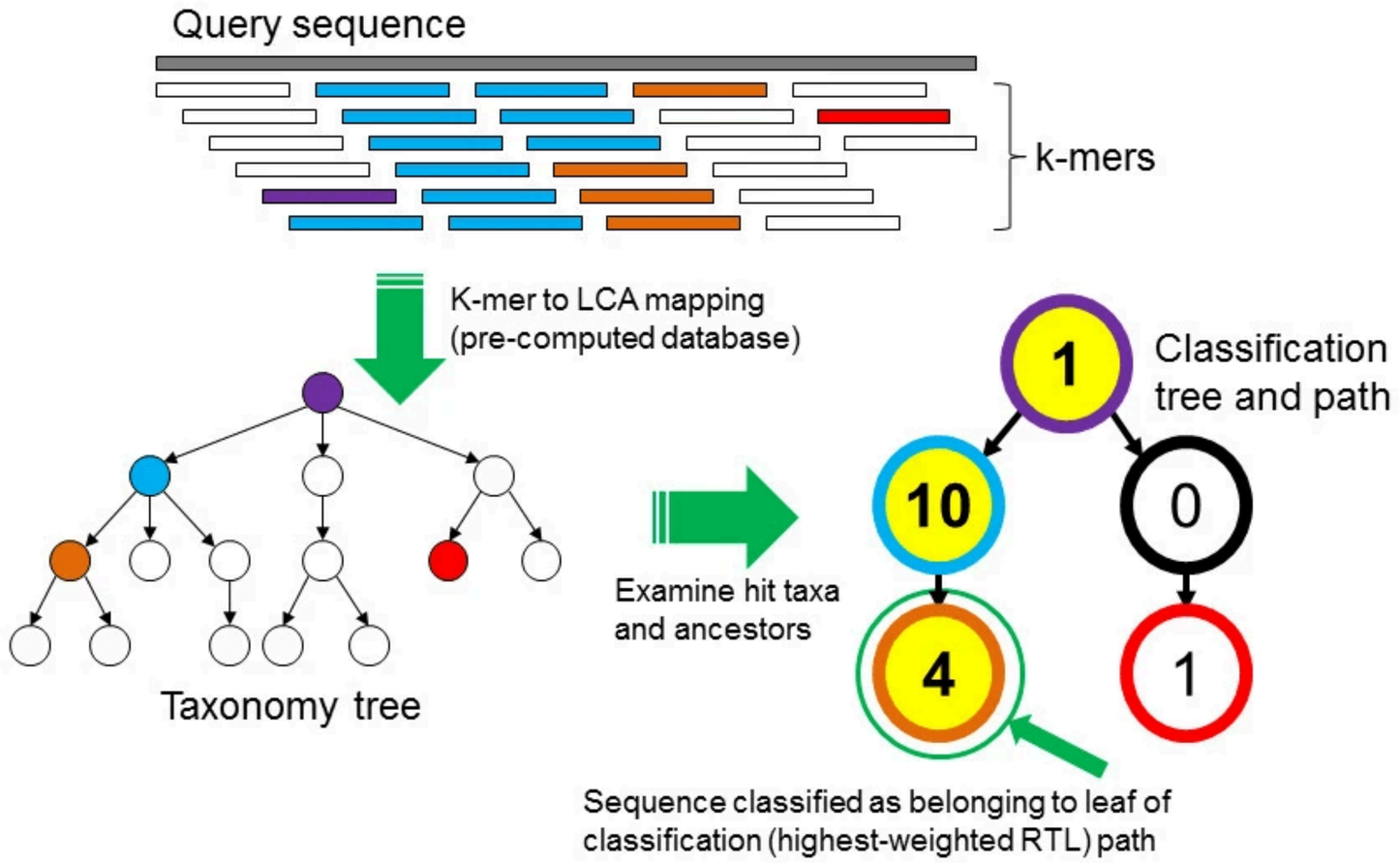
CLARK

Fast, accurate and versatile sequence classification system

Discriminative k-mers

Taxonomic classification

Kraken2: Exact k-mer matches.



Taxonomic classification

Kraken2: Output

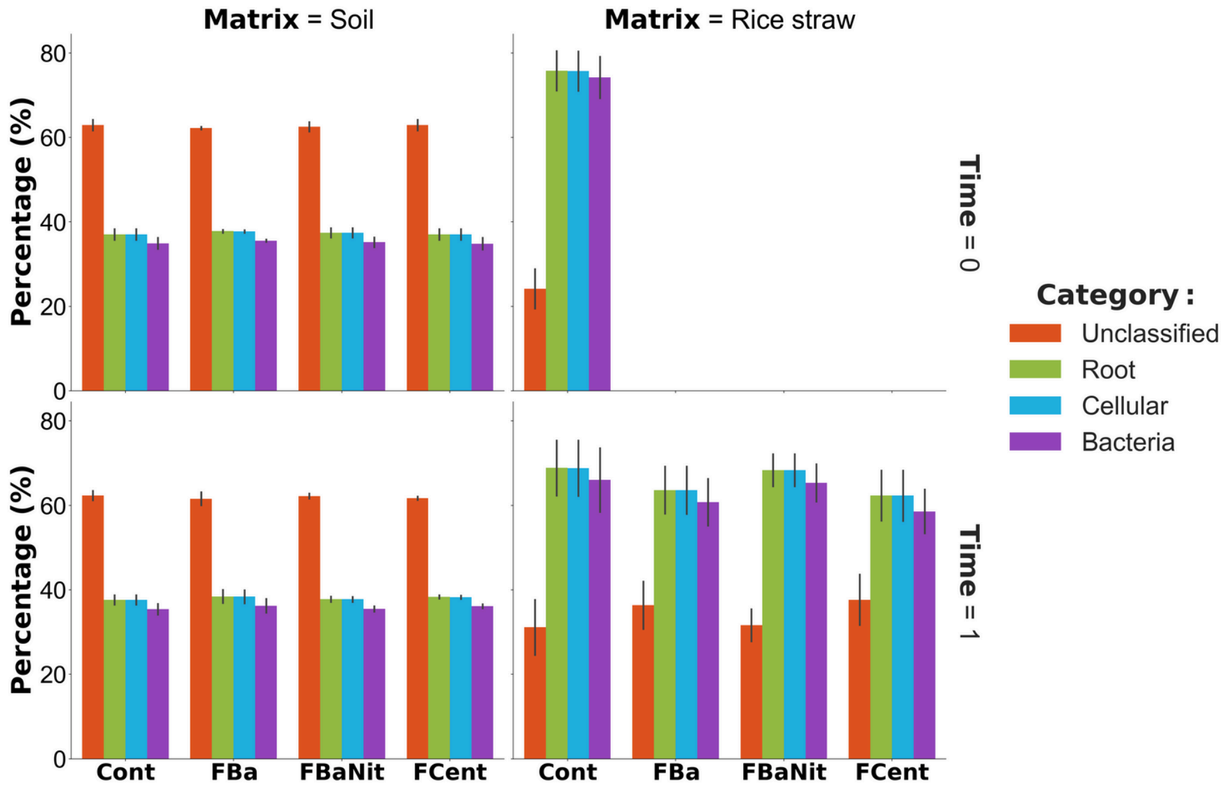
Output

```
 78.13  587119  587119  U       0       unclassified
 21.87  164308  1166    R       1       root
 21.64  162584  0       R1      131567    cellular organisms
 21.64  162584  3225    D       2           Bacteria
 18.21  136871  3411    P       1224          Proteobacteria
 14.21  106746  3663    C       28211           Alphaproteobacteria
  7.71  57950   21      O       204455            Rhodobacterales
  7.66  57527   6551    F       31989               Rhodobacteraceae
  1.23  9235    420     G       1060                  Rhodobacter
  0.76  5733    4446    S       1063                    Rhodobacter sphaeroides
```

1. Percentage of reads covered by the clade rooted at this taxon.
2. Number of reads covered by the clade rooted at this taxon.
3. Number of reads assigned directly to this taxon.
4. A taxonomy rank code.
5. NCBI taxonomy ID.
6. Indented scientific name.

Taxonomic classification

Example:

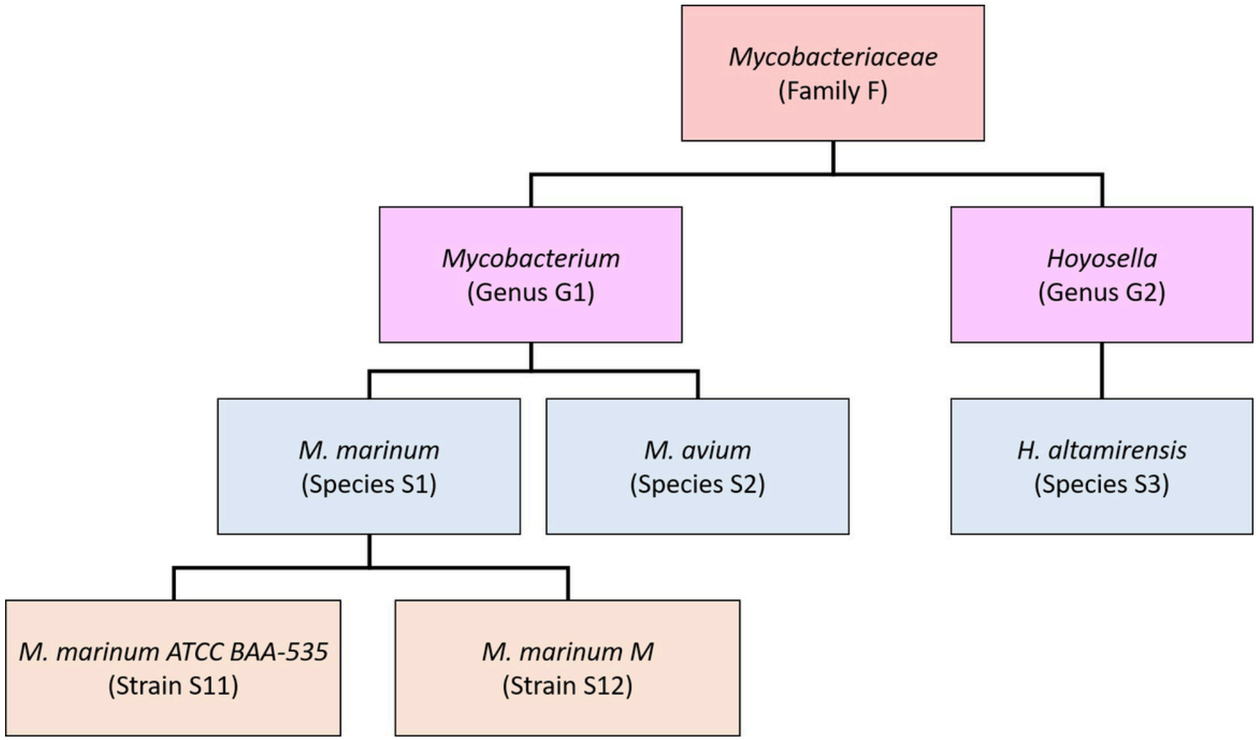


Problems with databases:

- Gap with the real world diversity.
- Overrepresentation of species.
- Low-abundance species.
- Database corruption.

Taxonomic classification

Bracken: Bayesian Reestimation of Abundance with KrakEN



$$P\left(S_i|G_j\right) = \frac{P\left(G_j|S_i\right)P\left(S_i\right)}{P\left(G_j\right)}.$$

$P(S_i|G_j)$: probability that a read classified at genus *Gj* belongs to the genome *Si*. (**Posterior**)

$P(G_j|S_i)$: probability that a read from genome *Si* is classified by Kraken as the parent genus *Gj*. (**Likelihood**)

$P(S_i)$: probability that a read in the sample belongs to genome *Si*. (**Prior**)

$P(G_i)$: probability that a read is classified by Kraken at the genus level *Gj*. (**Marginalization**)

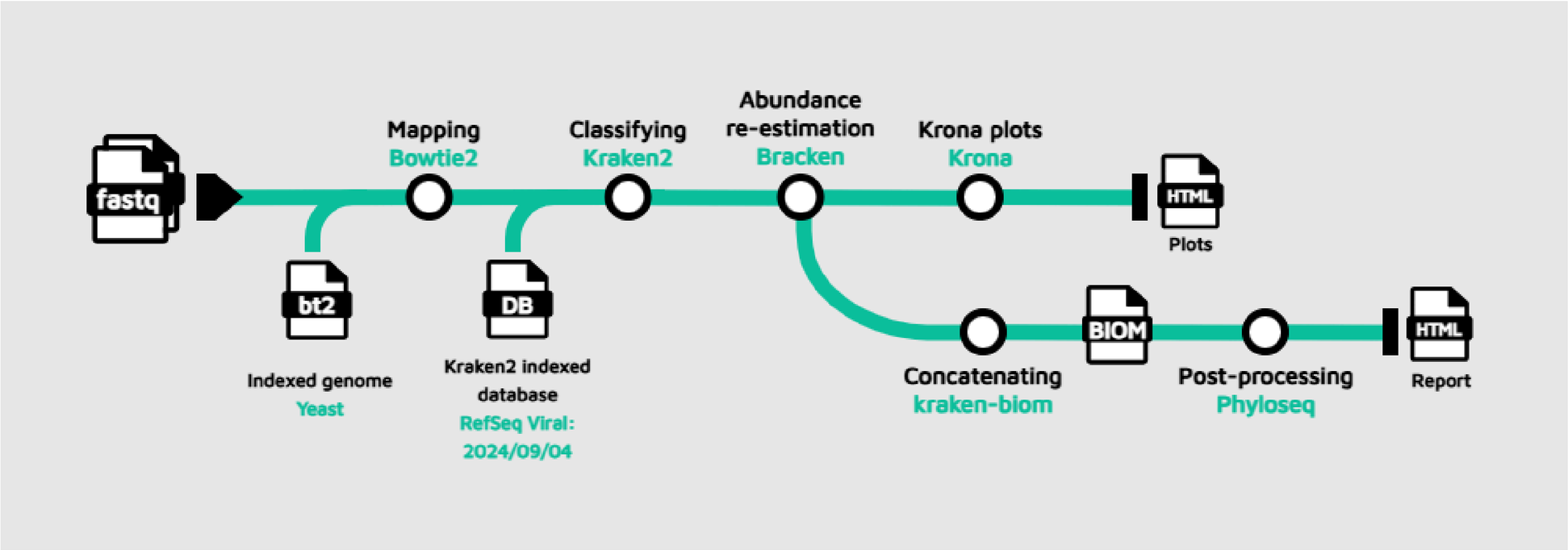
Today's Data

The dataset we will be using for this part of the workshop was collected in Cuatro Ciénegas (México). Cuatro Ciénegas is an oasis in the Mexican desert whose environmental conditions are often linked to the ones present in ancient seas, due to a higher-than-average content of sulfur and magnesium but a lower concentrations of phosphorus and other nutrients (Okie et al. 2020); the BioProject accession number: PRJEB22811.

Name	Group	Replicate
ERR2143758	Cont	1
ERR2143759	Cont	2
ERR2143760	Cont	3
ERR2143771	Unenriched	1
ERR2143772	Unenriched	2
ERR2143774	Unenriched	3
ERR2143769	Fertilized	1
ERR2143770	Fertilized	2
ERR2143773	Fertilized	3

Pipeline

For this course, we propose to wrap with Nextflow the protocol published by Lu et al. (2022) wrapped with Nextflow:



Implementation: **GitHub Codespaces**

Downstream analysis

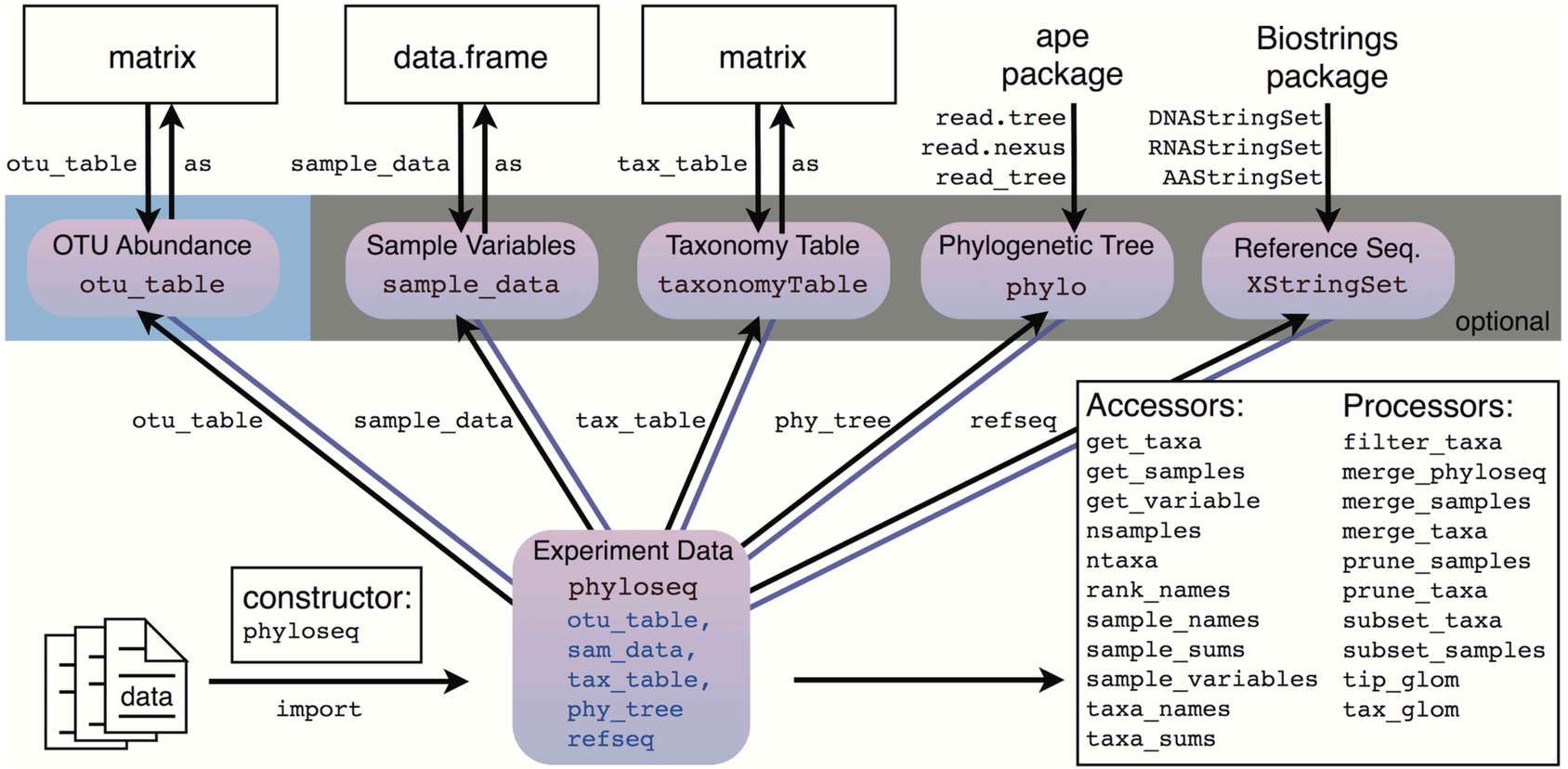
Biological Observation Matrix:

```
>>> import numpy as np
>>> from biom.table import Table
>>> data = np.arange(40).reshape(10, 4)
>>> sample_ids = ['S%d' % i for i in range(4)]
>>> observ_ids = ['O%d' % i for i in range(10)]
>>> sample_metadata = [{'environment': 'A'}, {'environment': 'B'},
...                    {'environment': 'A'}, {'environment': 'B'}]
>>> observ_metadata = [{'taxonomy': ['Bacteria', 'Firmicutes']},
...                    {'taxonomy': ['Bacteria', 'Firmicutes']},
...                    {'taxonomy': ['Bacteria', 'Proteobacteria']},
...                    {'taxonomy': ['Bacteria', 'Proteobacteria']},
...                    {'taxonomy': ['Bacteria', 'Proteobacteria']},
...                    {'taxonomy': ['Bacteria', 'Bacteroidetes']},
...                    {'taxonomy': ['Bacteria', 'Bacteroidetes']},
...                    {'taxonomy': ['Bacteria', 'Firmicutes']},
...                    {'taxonomy': ['Bacteria', 'Firmicutes']},
...                    {'taxonomy': ['Bacteria', 'Firmicutes']}]
>>> table = Table(data, observ_ids, sample_ids, observ_metadata,
...               sample_metadata, table_id='Example Table')
```

```
>>> table
10 x 4 <class 'biom.table.Table'> with 39 nonzero entries (97% dense)
>>> print(table)
# Constructed from biom file
#OTU ID S0  S1  S2  S3
O0  0.0 1.0 2.0 3.0
O1  4.0 5.0 6.0 7.0
O2  8.0 9.0 10.0    11.0
O3  12.0    13.0    14.0    15.0
O4  16.0    17.0    18.0    19.0
O5  20.0    21.0    22.0    23.0
O6  24.0    25.0    26.0    27.0
O7  28.0    29.0    30.0    31.0
O8  32.0    33.0    34.0    35.0
O9  36.0    37.0    38.0    39.0
>>> print(table.ids())
['S0' 'S1' 'S2' 'S3']
>>> print(table.ids(axis='observation'))
['O0' 'O1' 'O2' 'O3' 'O4' 'O5' 'O6' 'O7' 'O8' 'O9']
>>> print(table.nnz)  # number of nonzero entries
39
```

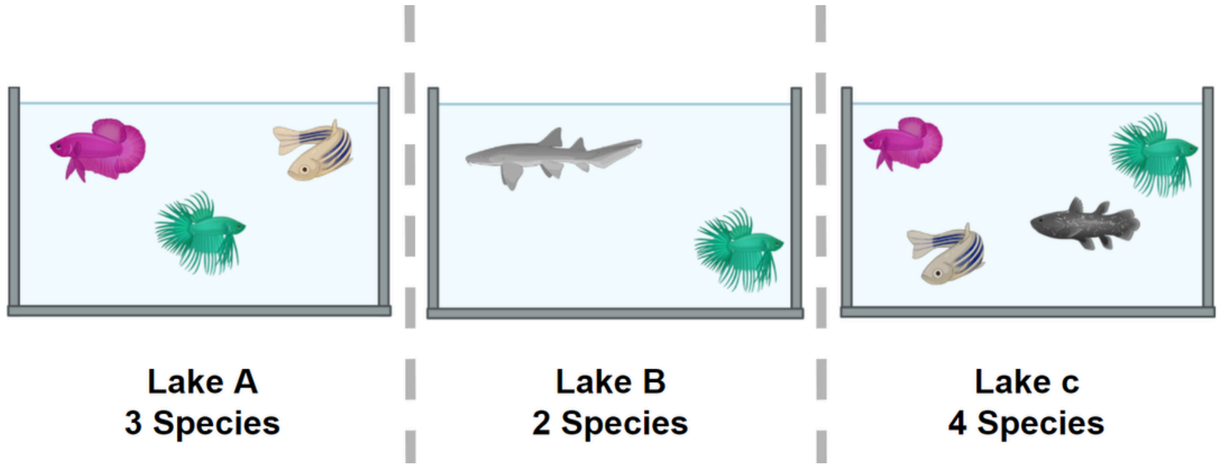
Downstream analysis

Phyloseq object:



Downstream analysis

α-Diversity:

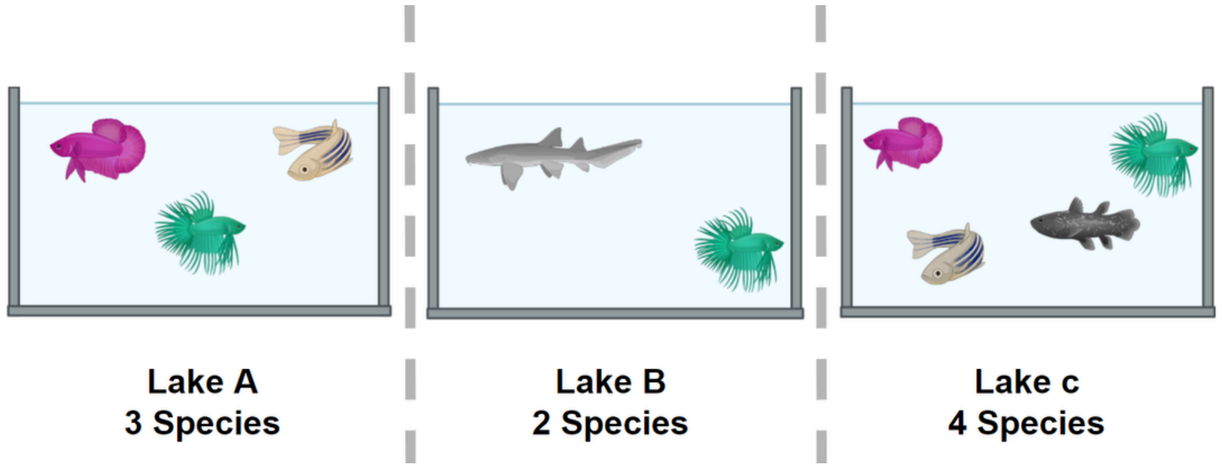


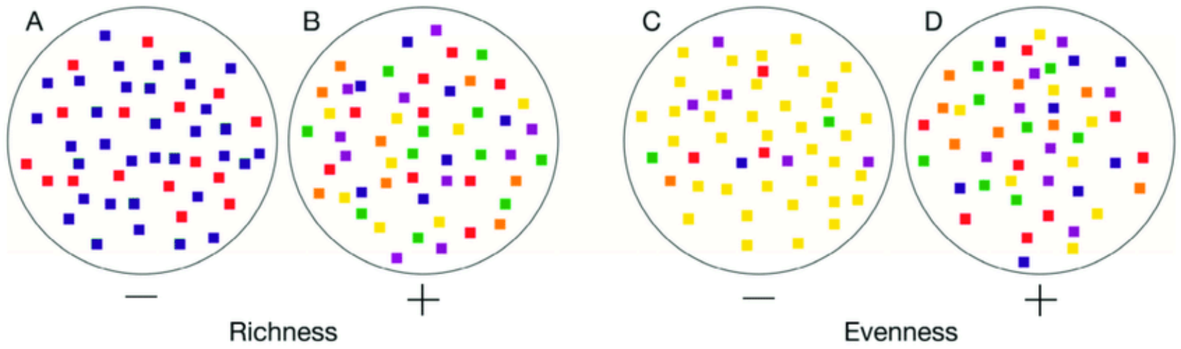
Diversity Indices	Description
Shannon (H)	Estimation of species richness and species evenness. More weight on richness.
Simpson's (D)	Estimation of species richness and species evenness. More weigth on evenness.
Chao1	Abundance based on species represented by a single individual (singletons) and two individuals (doubletons).

Zirión-Martínez et al., (2024). A Data Carpentry-Style Metagenomics Workshop. Journal of Open Source Education, 7(72), 209.

Downstream analysis

α-Diversity:



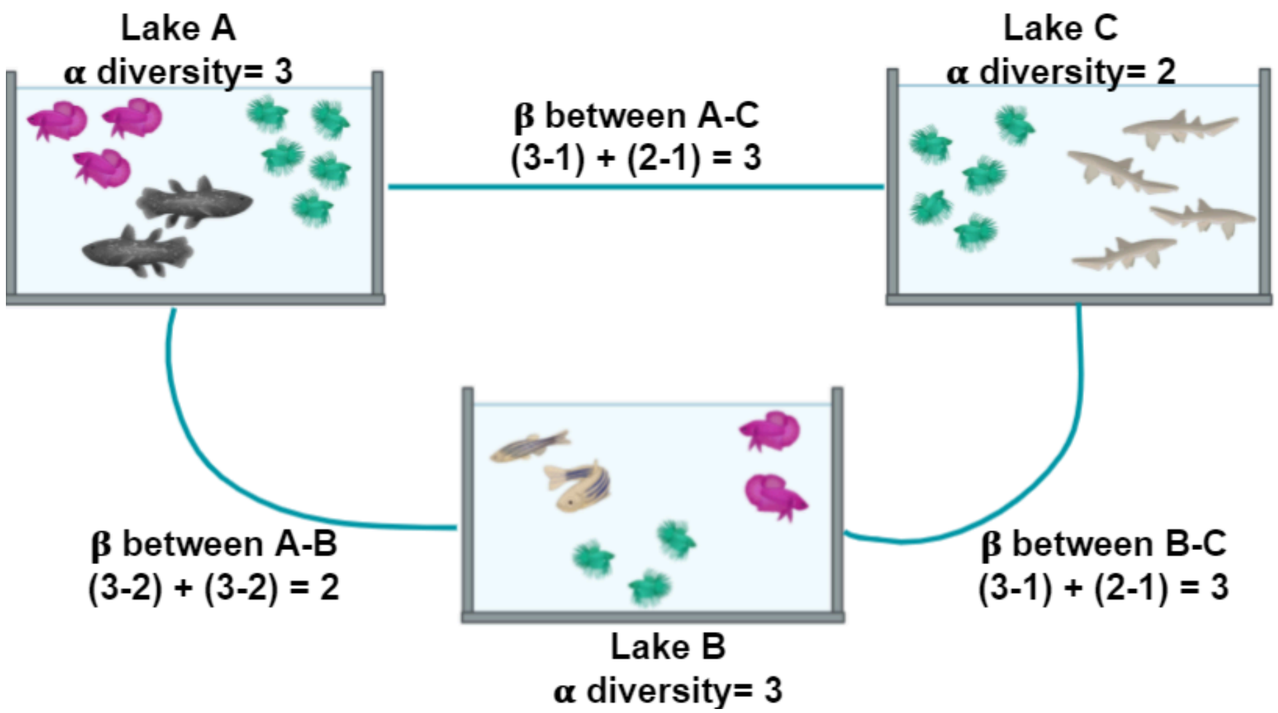


Diversity Indices	Description
Shannon (H)	Estimation of species richness and species evenness. More weight on richness.
Simpson's (D)	Estimation of species richness and species evenness. More weigth on evenness.
Chao1	Abundance based on species represented by a single individual (singletons) and two individuals (doubletons).

Zirión-Martínez et al., (2024). A Data Carpentry-Style Metagenomics Workshop. Journal of Open Source Education, 7(72), 209.

Downstream analysis

β-Diversity:



Zirión-Martínez et al., (2024).

- Bray-Curtis.

$$BC_d = \frac{\sum \left| x_i - x_j \right|}{\sum \left(x_i + x_j \right)}$$

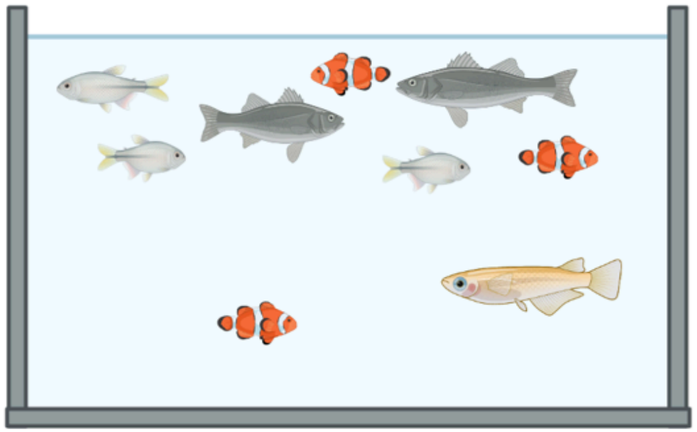
- Jaccard.
- UniFrac.
- Weighted Unifrac.
- Manhattan.
- Euclidean.

Downstream analysis

Exercise:

Lake A

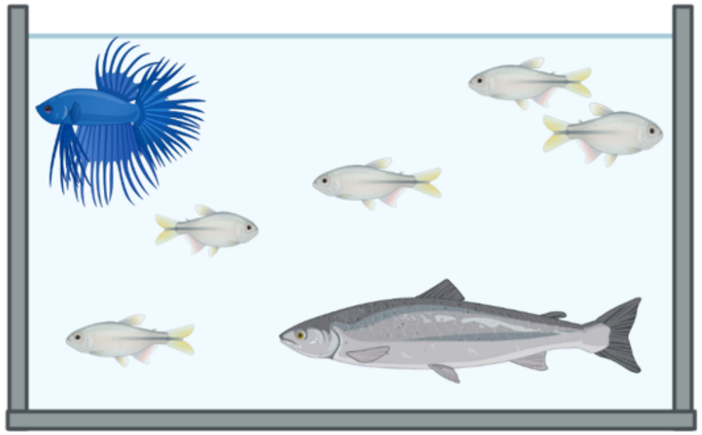
α-diversity= ?



A-B
β-diversity= ?

Lake B

α-diversity= ?



α-Diversity lake A?
α-Diversity lake B?
β-Diversity lake A-B?

Important considerations

Metagenomics data is compositional:

- The total number of counts per sample is highly variable and **constrained by the maximum number** of DNA reads that the sequencer can provide.

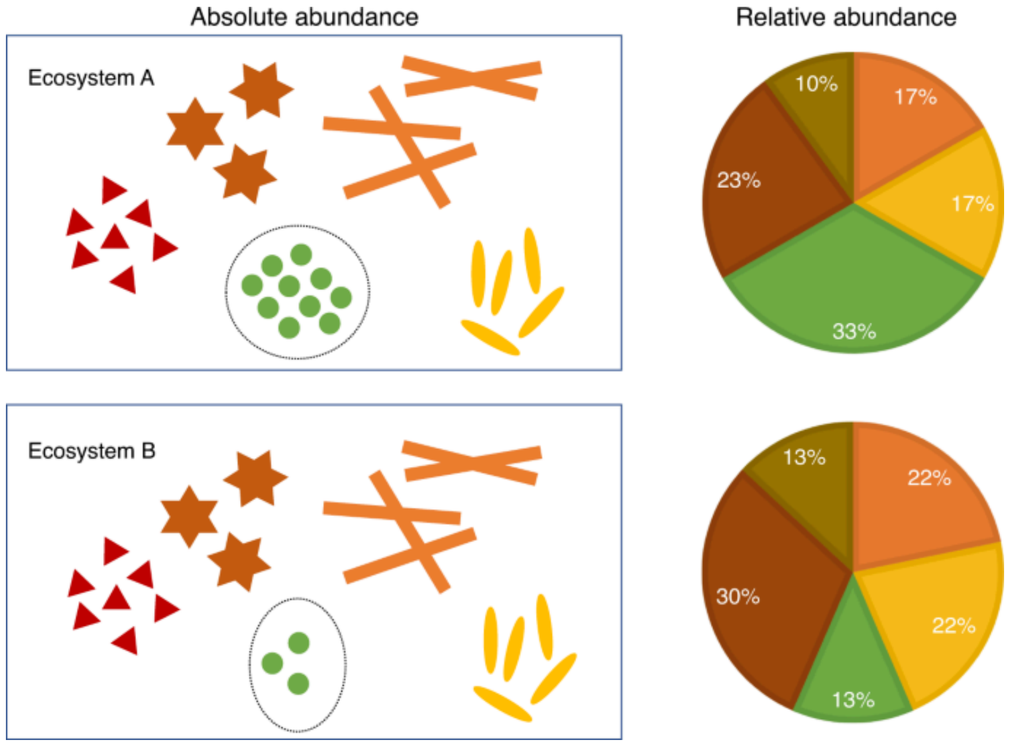
- An **increase** in the abundance of one taxon **implies** the **decrease** of the observed number of counts for some of the other taxa.

- Observed raw abundances and the total number of reads per sample are non-informative since they represent **only a fraction** or random sample of the **original DNA content in the environment**.

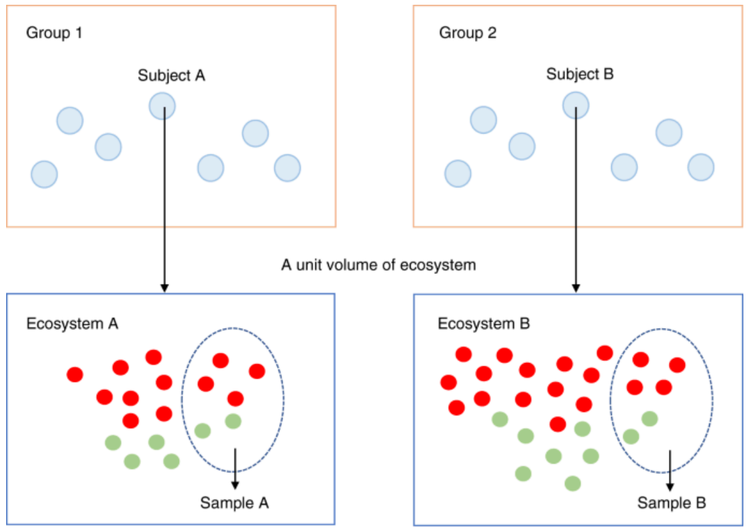
Calle (2019)

Important considerations

Metagenomics data is compositional:



Lin & Peddada (2020)



	Sample		Ecosystem	
	A	B	A	B
●	4	4	12	18
●	2	2	6	9
Sum	6†	6†	18‡	27‡

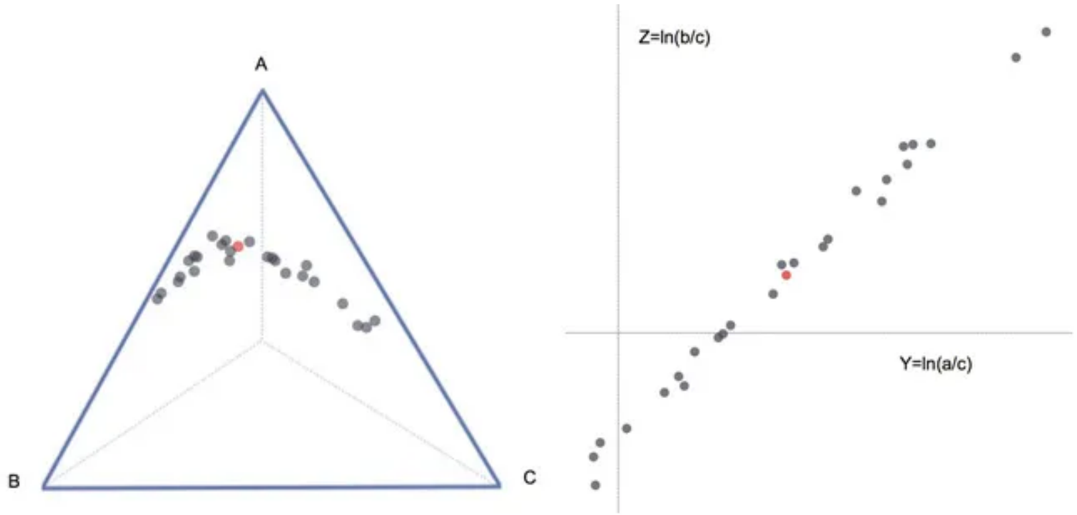
† Library size; ‡ microbial load.

Bias introduced by sampling fraction

Important considerations

Centered-log transformation*:

$$clr(x_1, \cdots, x_k) = \left(\log\left(\frac{x_1}{g(x)}\right), \ldots, \log\left(\frac{xk}{g(x)}\right)\right)$$

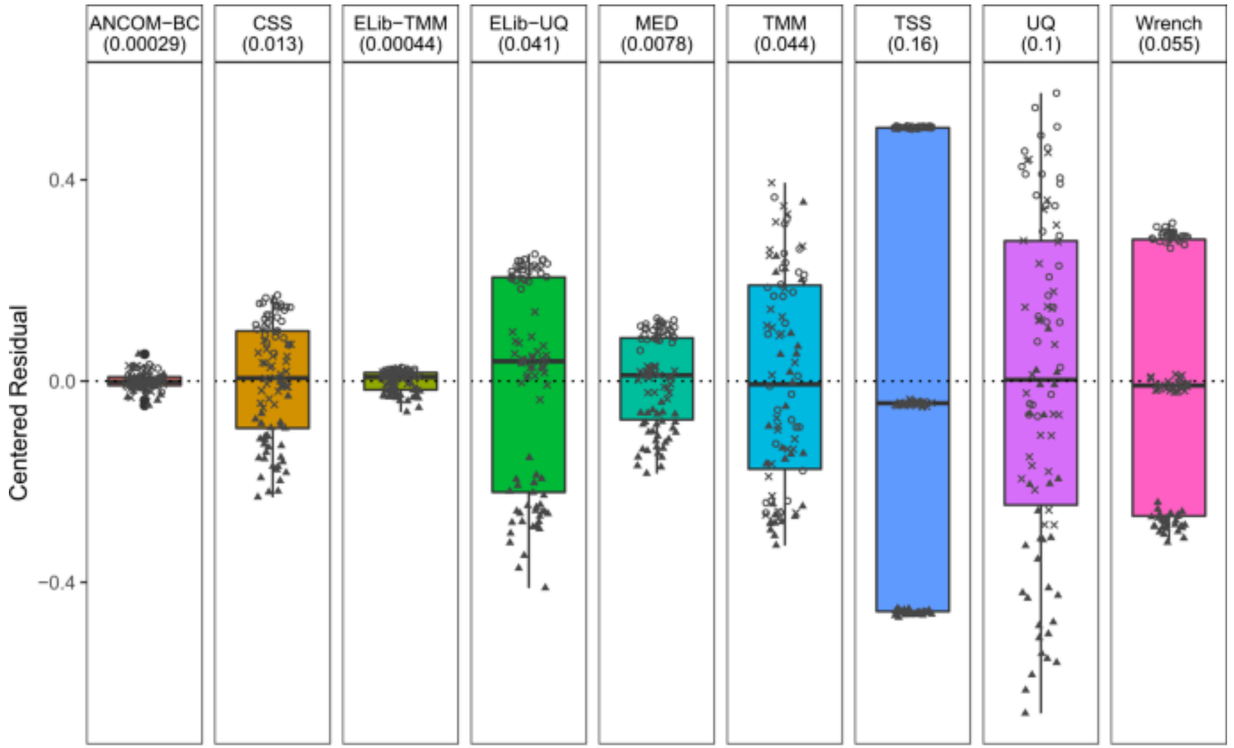


- Permutation invariance (**order**)
- Scale invariance (**relative relationship**)
- Sub-compositional coherence (**subset/whole**)

*Requires zero replacement.

Important considerations

Normalization and transformation:



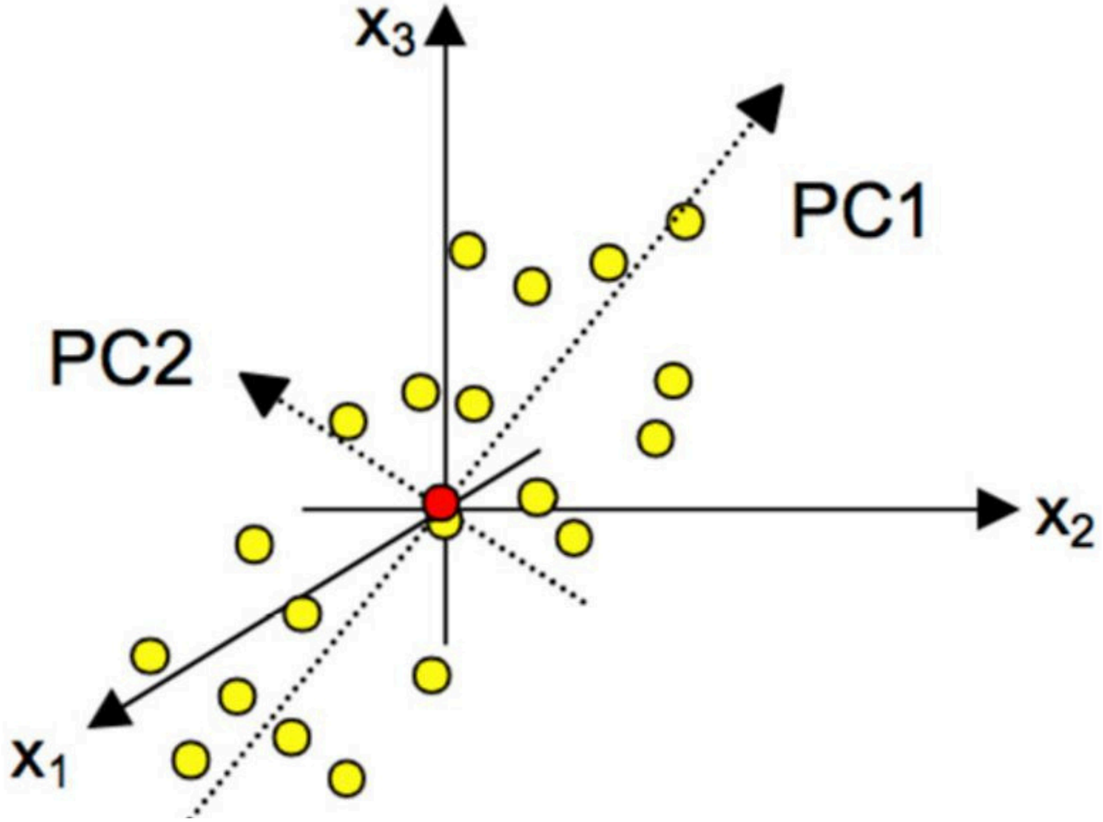
Lin & Peddada (2020)

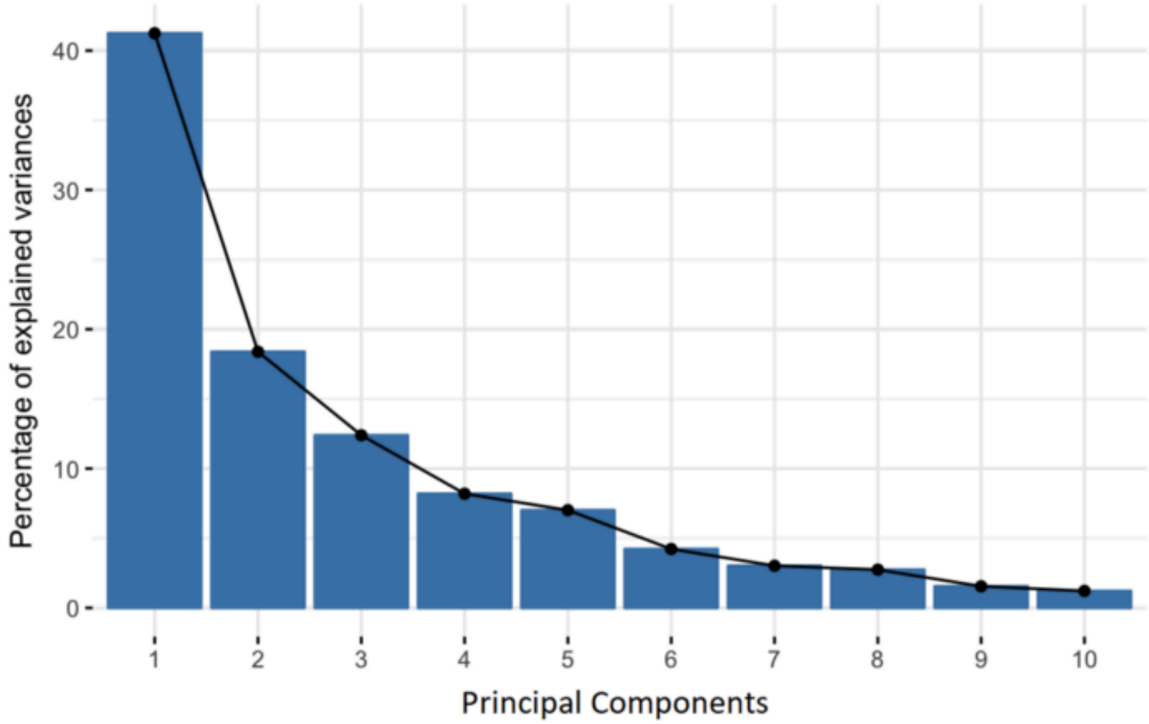
Operation	Standard approach	Compositional approach
Normalization	Rarefaction 'DESeq'	CLR ILR ALR
Distance	Bray-Curtis UniFrac Jenson-Shannon	Aitchison
Ordination	PCoA (Abundance)	PCA (Variance)
Multivariate comparison	perManova ANOSIM	perMANOVA ANOSIM
Correlation	Pearson Spearman	SparCC SpiecEasi ϕ ρ
Differential abundance	metagenomSeq LEfSe DESeq	ALDEx2 ANCOM

Gloor et al. (2017)

Downstream analysis

Principal Components Analysis (PCA) or Principal Coordinate Analysis (PCoA):





Downstream analysis

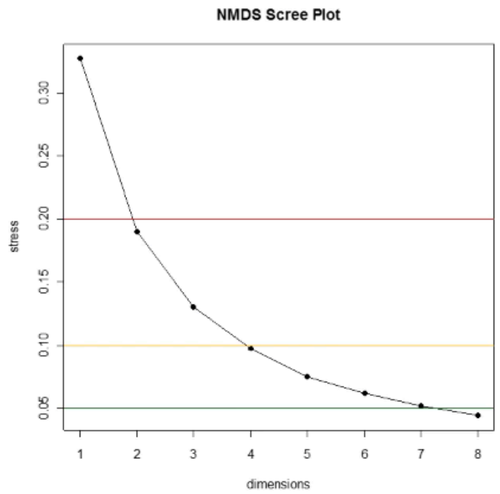
Non-metric Multidimensional Scaling (NMDS):

Multivariate data

	Sp A	Sp B	Sp C
Site 1	...	...	...
Site 2	...	...	...
Site 3	...	...	...

User-defined distance matrix (real space)

	Site 1	Site 2	Site 3
Site 1	...		
Site 2	...	...	
Site 3	...	...	...



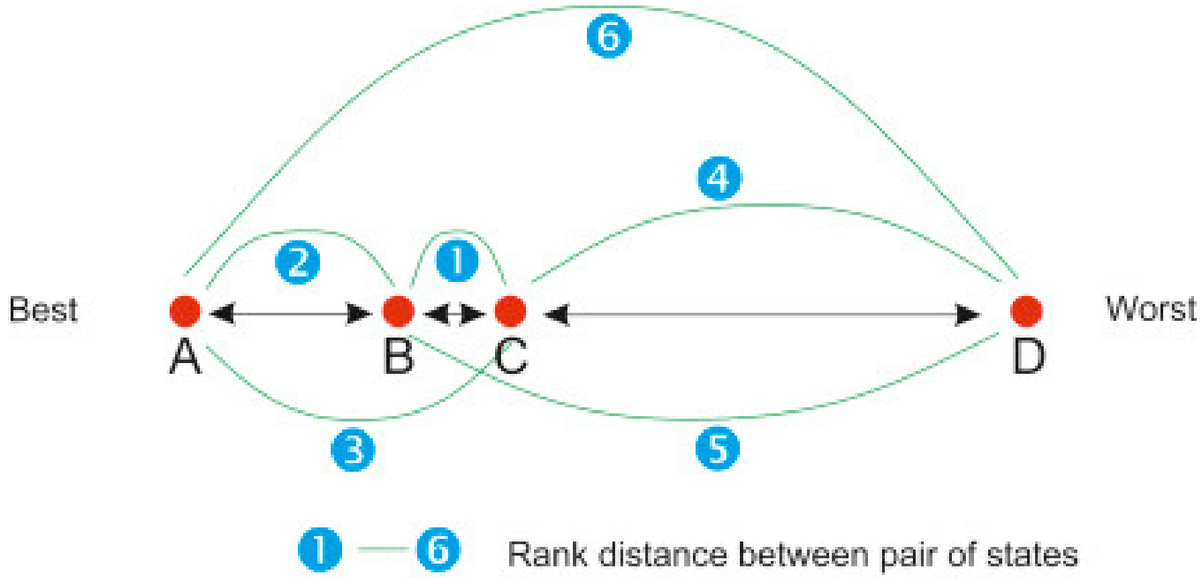
Goodness of fit:

>0.2 Poor (risks in interpretation)

0.1-0.2 Fair (some distances misleading)

0.05-0.1 Good (inferences confident)

<0.05 Excellent



Downstream analysis

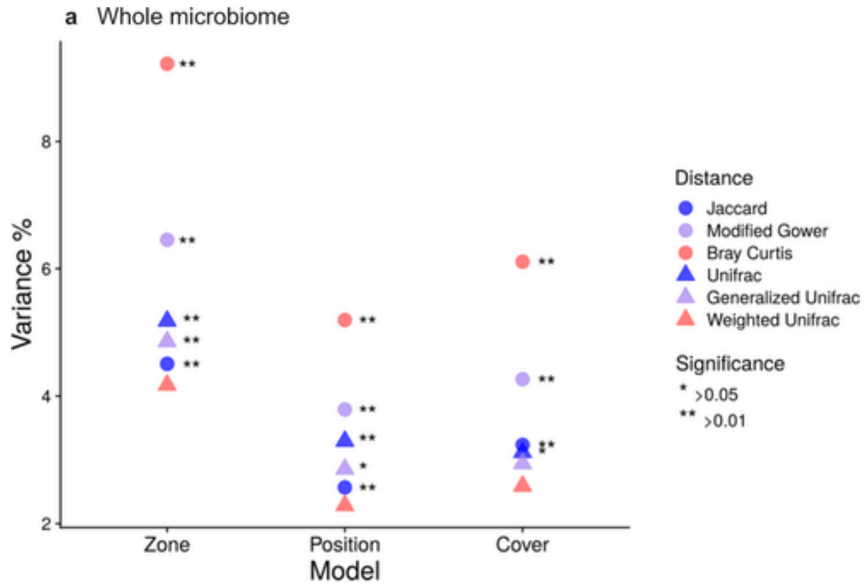
Multivariate differential abundance testing:

Distance/Dissimilarity ~ Group (covariates)

PERMANOVA

Compartment	Variable	PERMANOVA		
		F. Model	R^2	*p*
L	Timepoint	2.978	0.090	0.006
	Root section	13.681	0.137	0.001
	Tree	3.637	0.292	0.001
	Timepoint * Root section	1.935	0.058	0.076
T	Timepoint	2.980	0.119	0.018
	Root section	7.259	0.096	0.001
	Tree	3.248	0.345	0.002
	Timepoint * Root section	1.680	0.067	0.112

Becker et al. (2022)

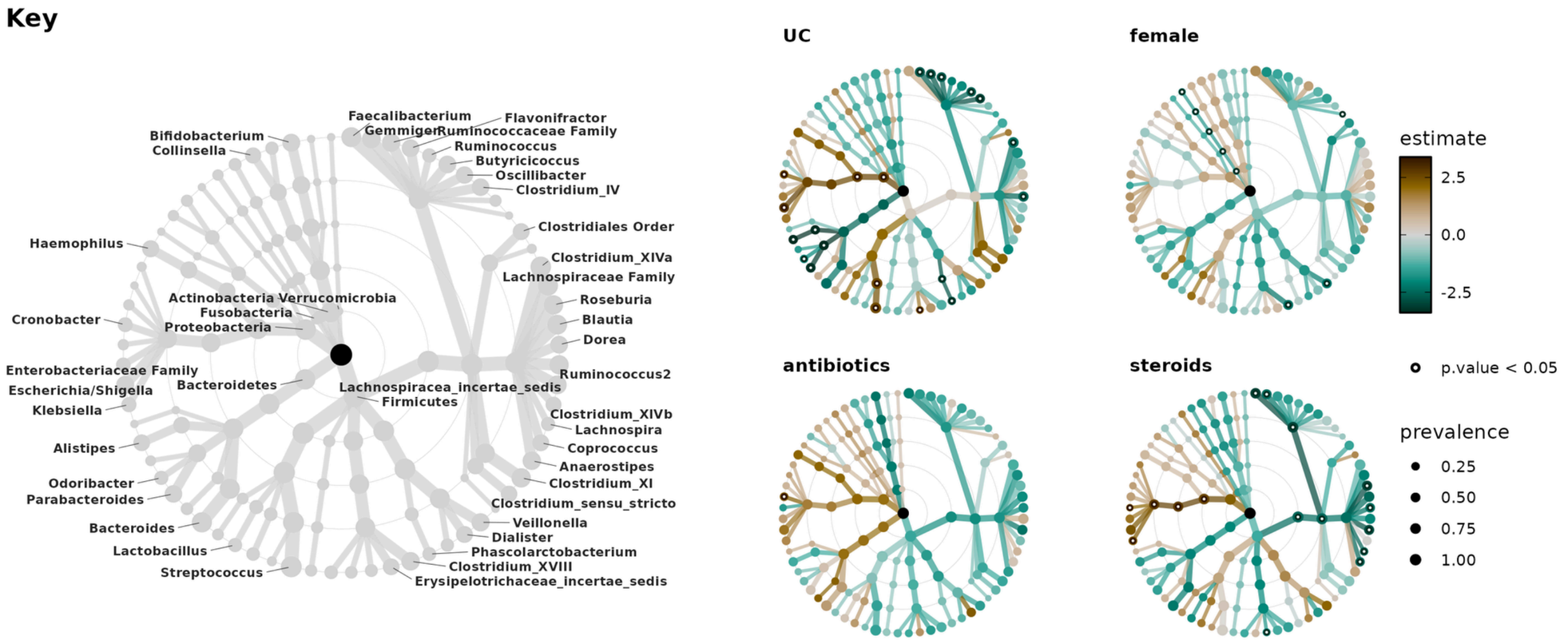


Clever et al. (2022)

Downstream analysis

Statistical modeling:

Taxa ~ Variables



Take-home messages

- **Contaminants removal** is a key step during read-based metagenomics data analysis.
- There are several methods to characterize the community composition. It is advisable to verify the results with **as many methodologies as possible**.
- Use **different metrics** to corroborate initial or exploratory observations.
- “Microbiome datasets are **compositional**: And this is not optional ”.
- Be aware of the multiple existing **methods to establish differentially abundant taxa**, and the implications of using each of them.

Additional resources

- Phyloseq tutorial
- Data Processing and Visualization for Metagenomics
- Integrated analysis with microViz
- ANCOM-BC2 tutorial