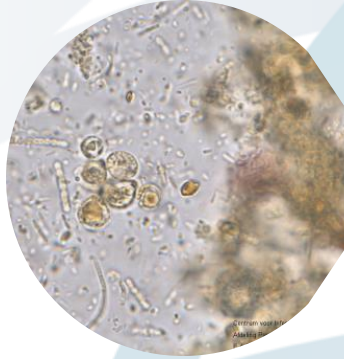
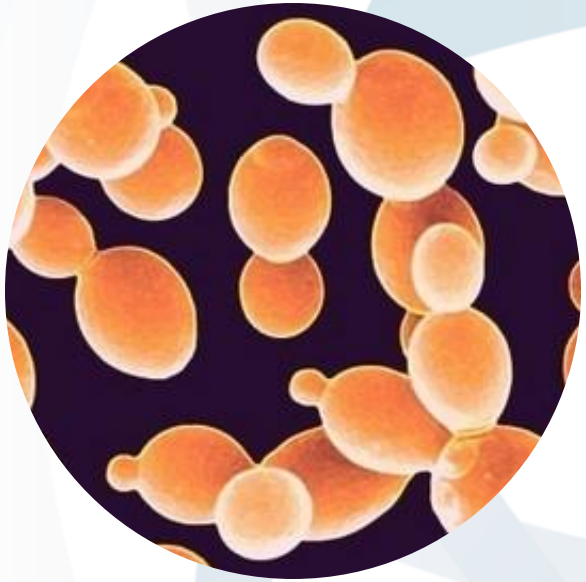
MUNI | RECETOX

MUNI FI

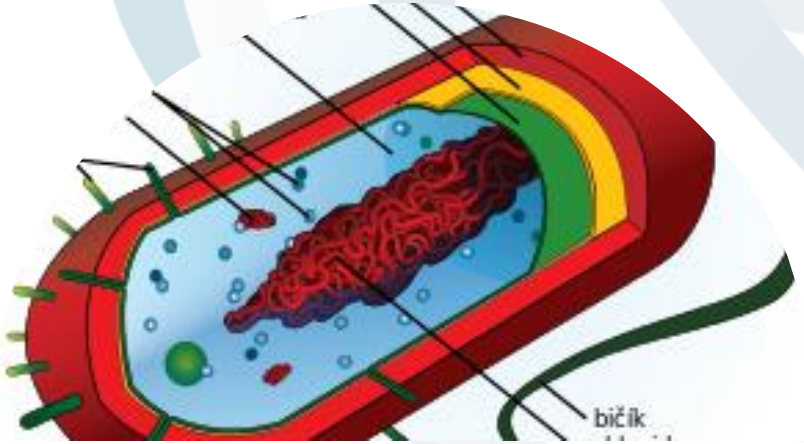


Metagenomics data analysis



IV110 Projekt z bioinformatiky I
IV114 Projekt z bioinformatiky a systémové biologie
E4014 Projekt z Matematické biologie a biomedicíny - biomedicínská bioinformatika

Mgr. Eva Budinská, Ph.D.
Ing. Vojtěch Bartoň



eva.budinska@recetox.muni.cz, vojtech.barton@recetox.muni.cz

Metagenomics is the study of genetic material recovered directly from environmental samples.

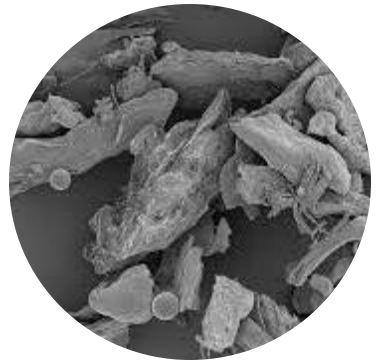
skin

soil

permafrost

stool

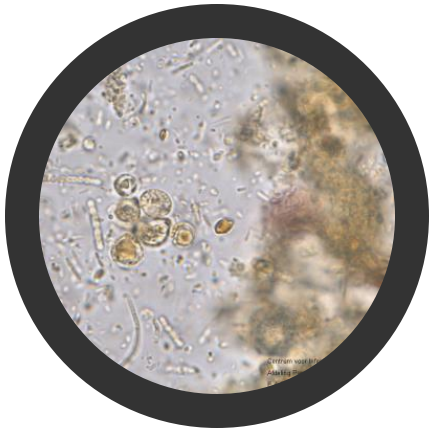
water

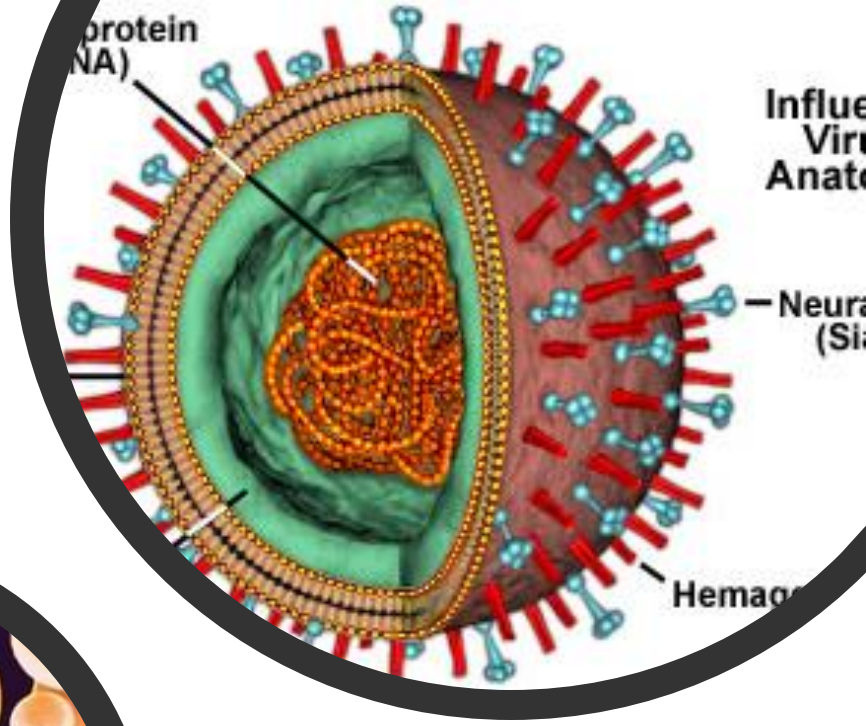


indoor dust

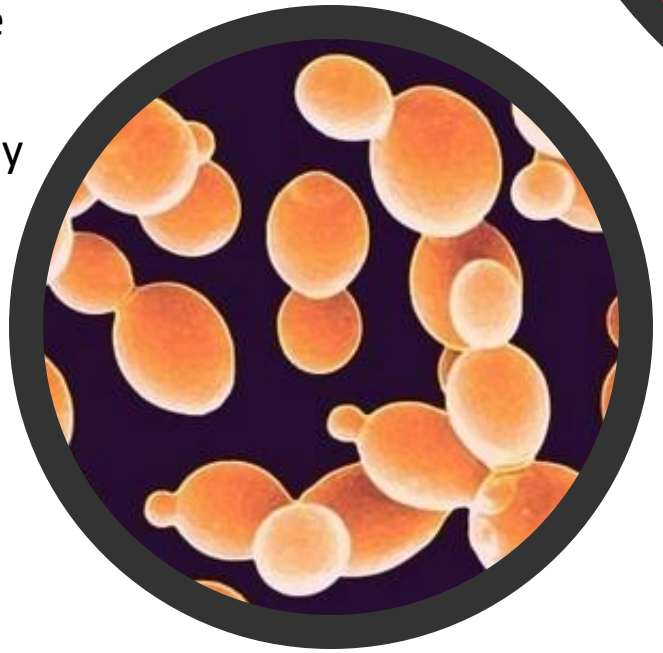
Metagenomics hence studies **microbiome**

Microbiome

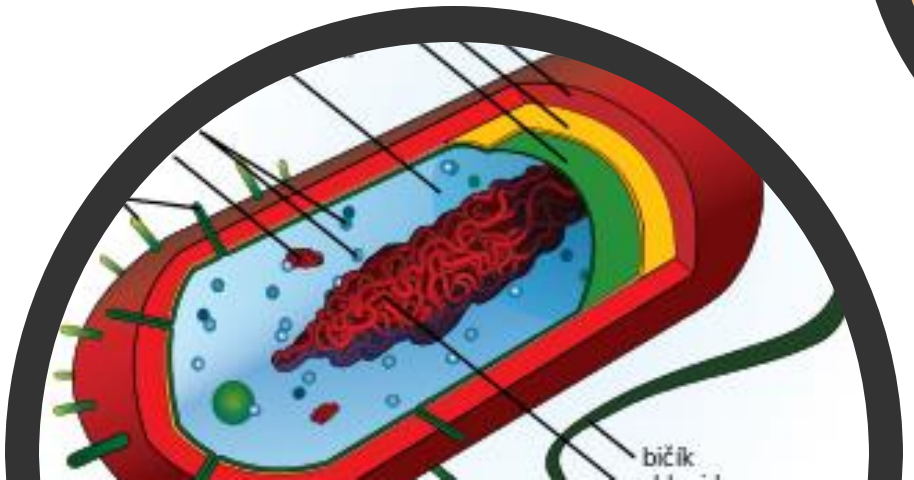


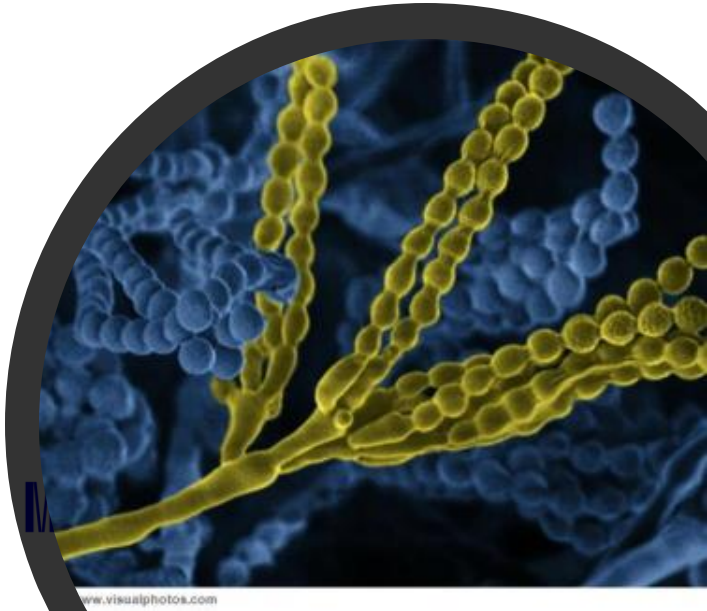


- is a community of microorganisms that can usually be found living together in a specific environment
- Microorganism - a single-celled organism that can only be seen under a microscope



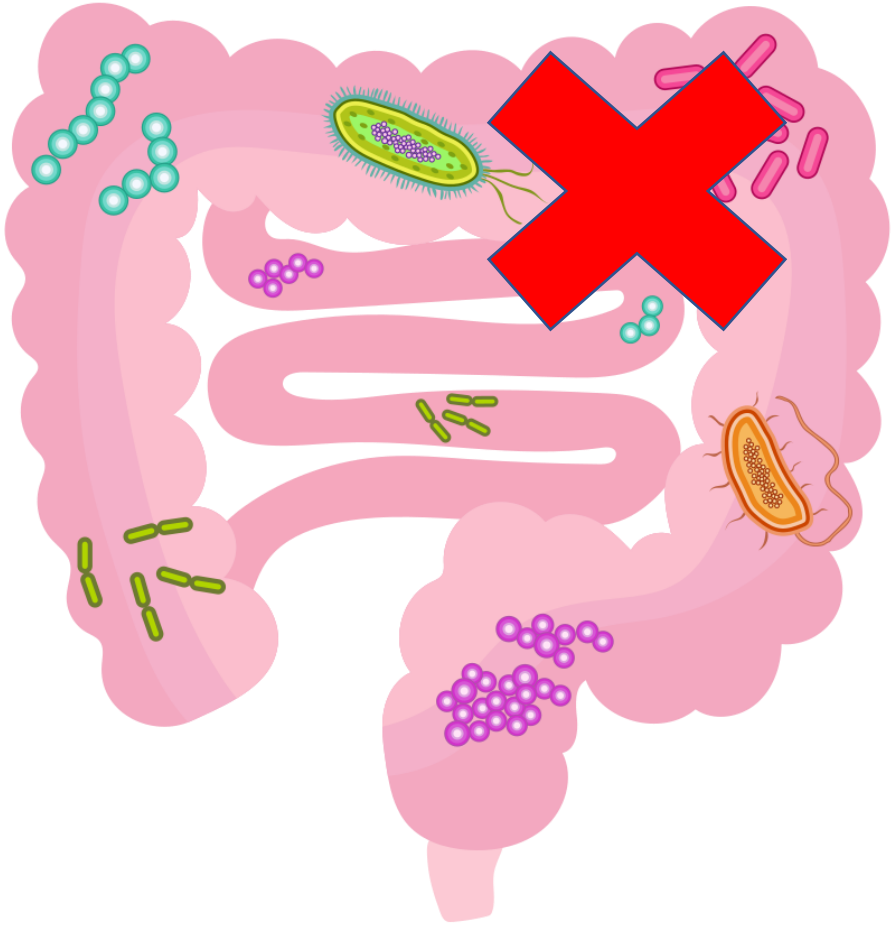
- Bacteria, Viruses, Fungi, Yeast, Algae, ...





Dysbiosis - when something goes wrong

- Microbiome out of balance
- Associated with <u>many diseases,</u> including **cancer**



Specifics of metagenomic data analysis

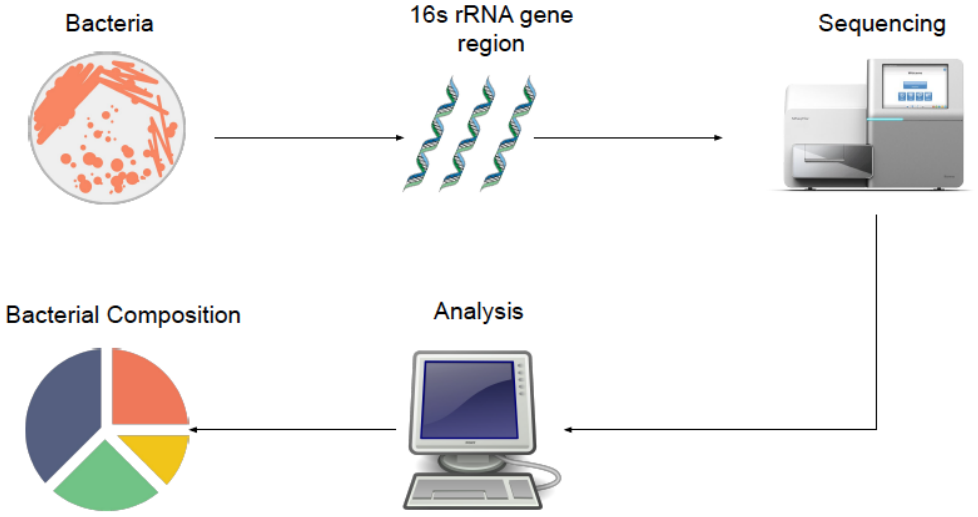
- Metagenomics studies **sample DNA from the whole community**
- A metagenomic sample often contains reads from a **huge number of (micro)organisms** of which **many are unknown**
- The sequences are often **incomplete** and hard to assemble to individual genes or recover full genomes of each organism

Two main approaches

Marker-gene metagenomics (targeted sequencing)

Sequencing specific target genes (16S rRNA, 18S rRNA, ITS, rpoB…)

Result: A quick **estimate of taxonomic diversity** and **composition.**



Two main approaches

Marker-gene metagenomics (targeted sequencing)

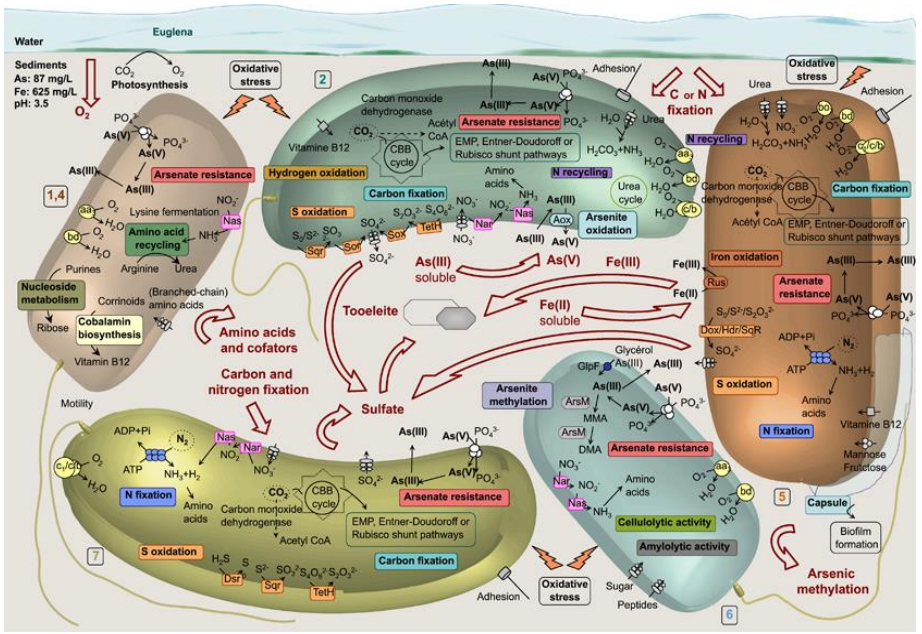
Sequencing specific target genes (16S rRNA, 18S rRNA, ITS, rpoB…)

Result: A quick estimate of taxonomic diversity and composition.

Shotgun metagenomics (whole-genome sequencing)

Sequence all genomic fragments from the sample.

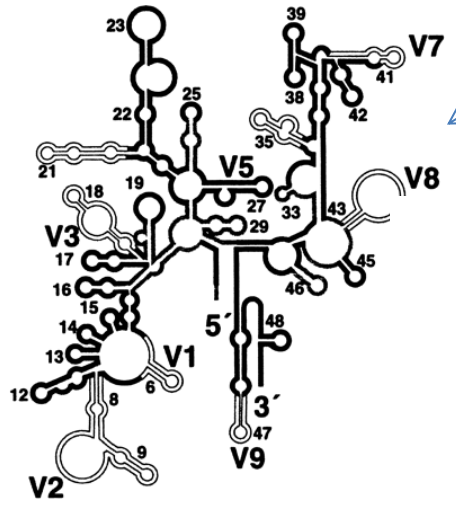
Result: Insights into **composition** and **function** of the microbiome of the sample.



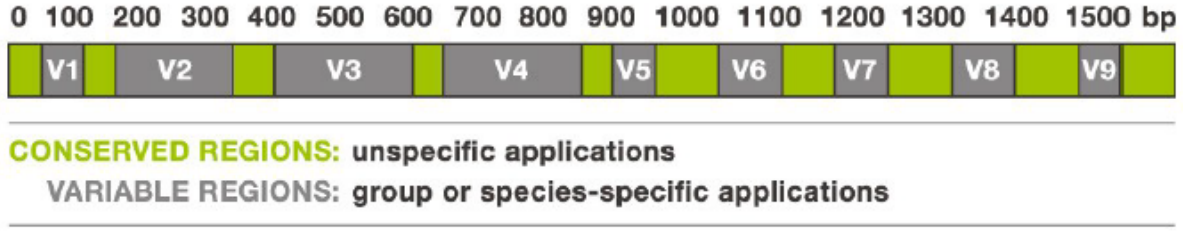
Marker-gene metagenomics (targeted sequencing)

Marker-gene (targeted) metagenomics

- Aim: obtain taxonomical representation of microbiome in the sample using **specific target genes**

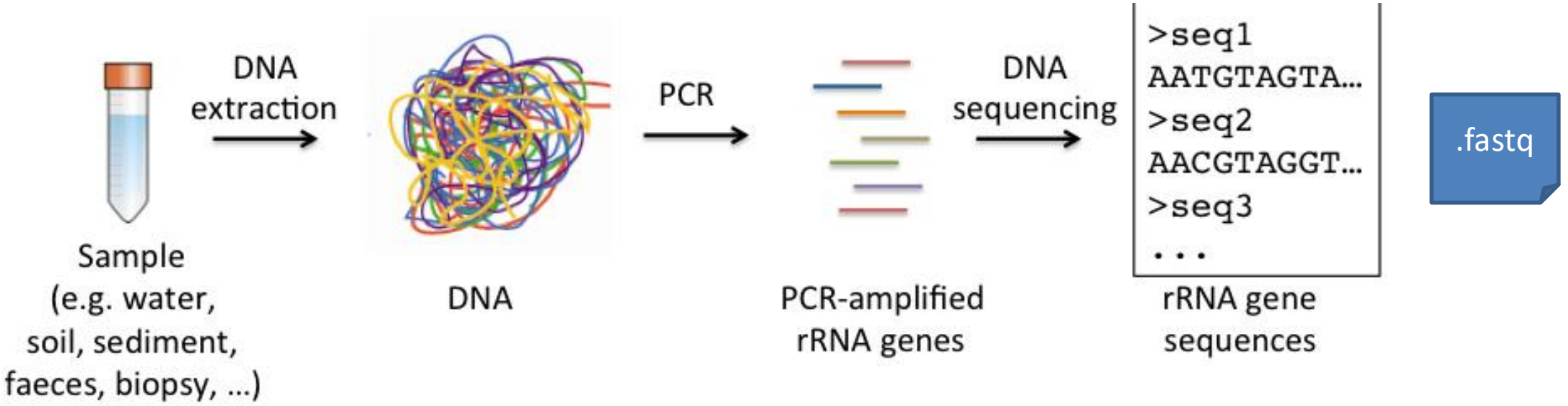


For **bacteria**: gene for **16s rRNA** and its selected variable regions



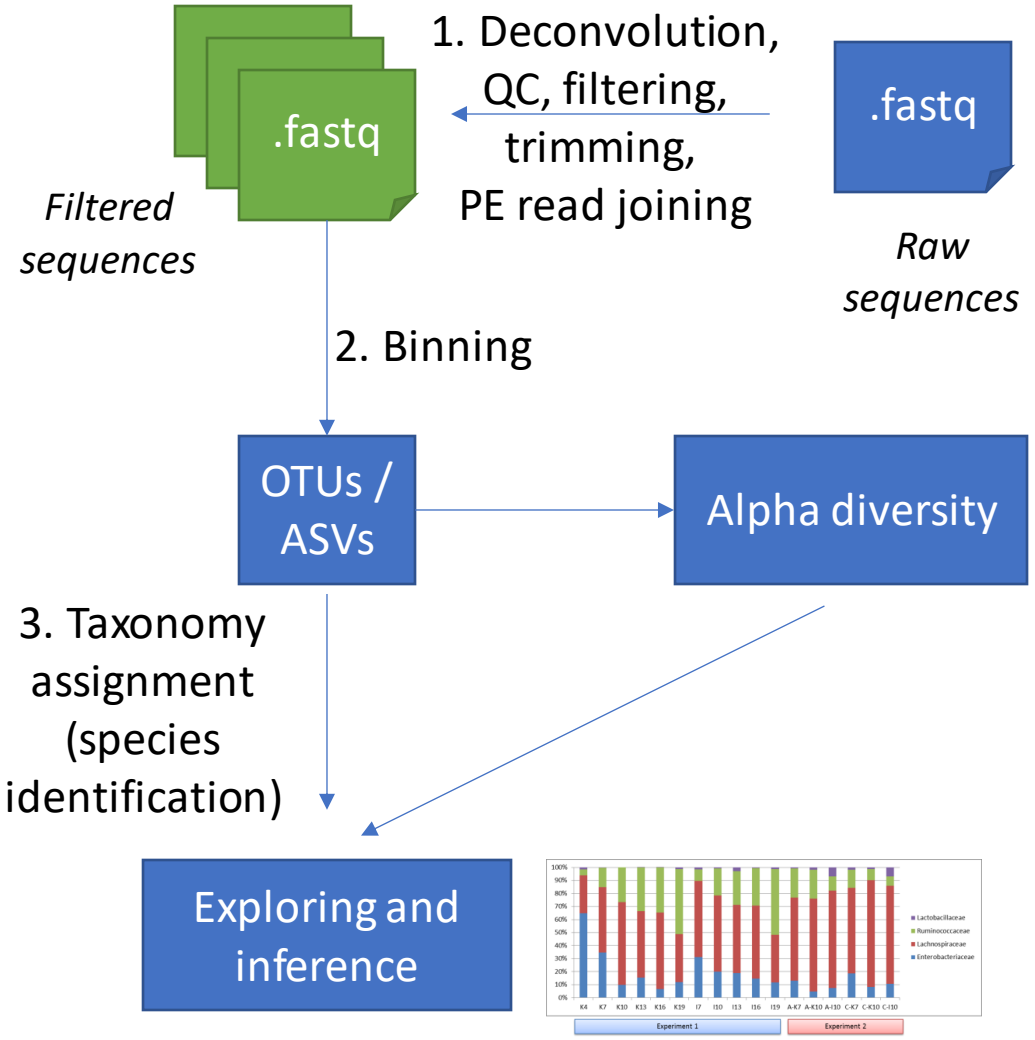
It is possible to identify species from mRNA – this is done usually from RNAseq experiments of different human tissues.

Marker-gene metagenomics (targeted sequencing) - basic workflow



The usual data analysis pipeline

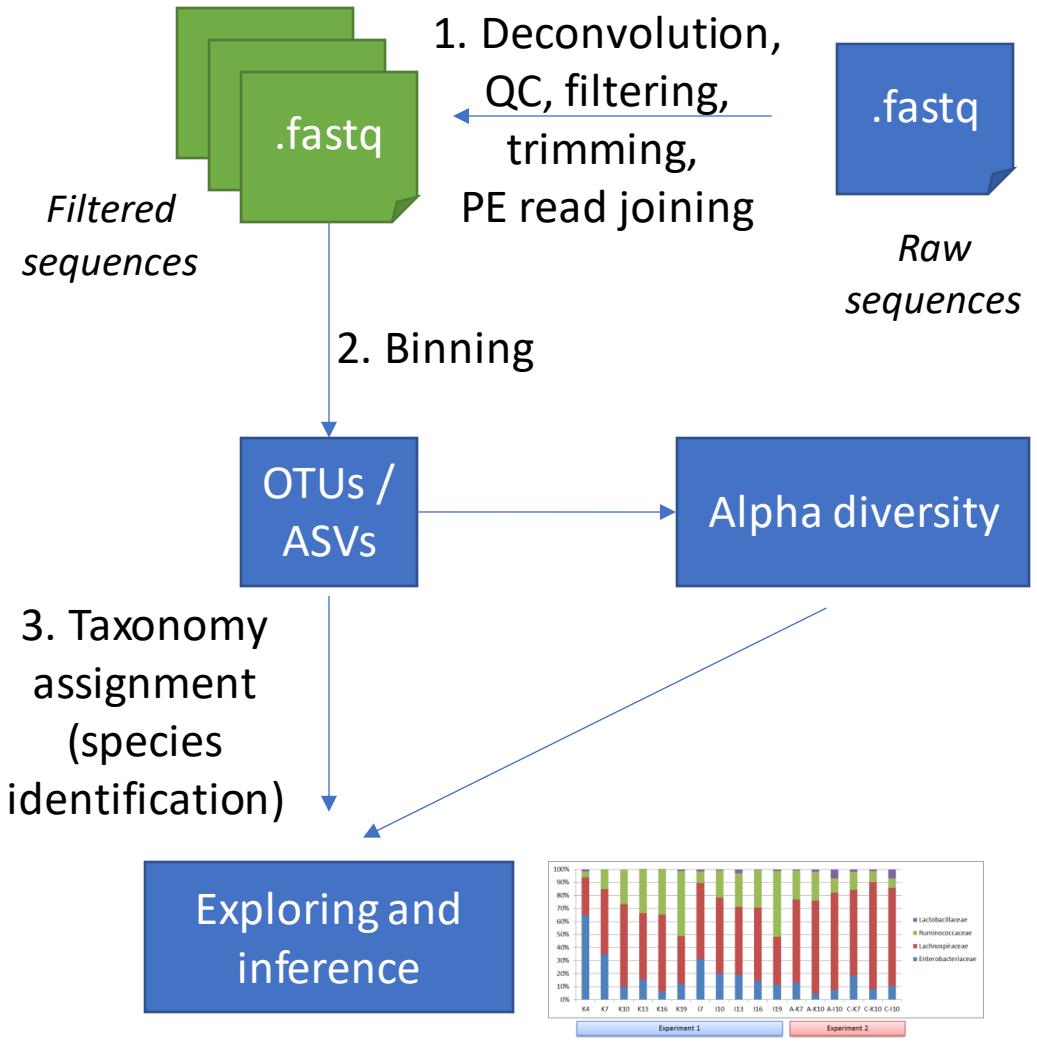
1. **Preprocessing:** Fastq files preprocessing, deconvolution, QC, trimming, joining reads
2. **Identification of sequences representative of potential species**
3. **Taxonomy assignment**
4. Exploratory analysis
   - Analysis of diversity measures and their visualization
5. Inference analysis
   - Associating composition with variables of interest



MUNI | RECETOX

The usual data analysis pipeline

1. **Preprocessing:** Fastq files preprocessing, deconvolution, QC, trimming, joining reads
2. Identification of sequences representative of potential species
3. Taxonomy assignment
4. Exploratory analysis
   - Analysis of diversity measures and their visualization
5. Inference analysis
   - Associating composition with variables of interest



MUNI | RECETOX

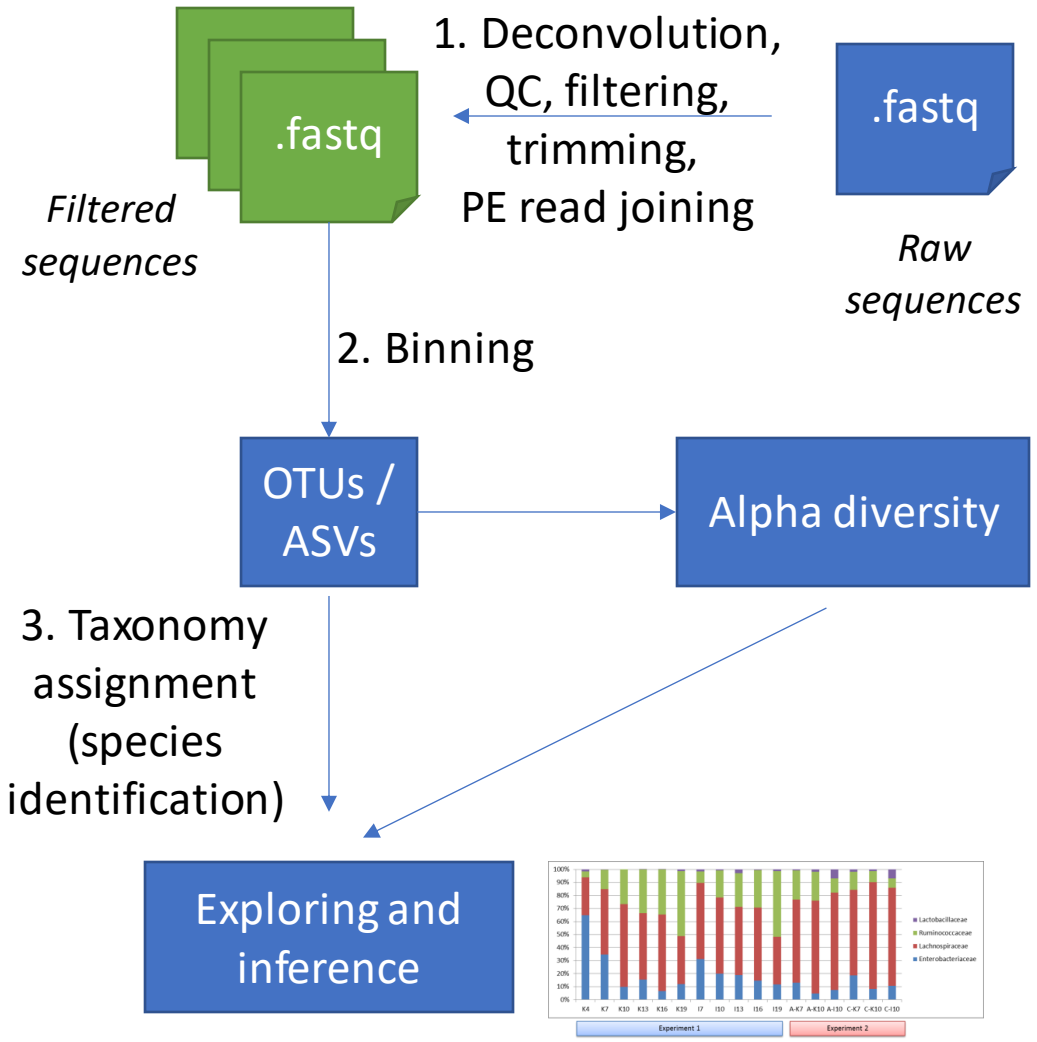
Marker-gene metagenomics

Step 1. Preprocessing and QC

- Marker gene metagenome is small => many samples are combined within a single run
- Not uncommon to have all the reads in one fastq file
- Samples need to be barcoded and **demultiplexed**
- **Followed by standard QC, trimming, joining PE reads**

The usual data analysis pipeline

1. **Preprocessing:** Fastq files preprocessing, deconvolution, QC, trimming, joining reads
2. **Identification of sequences representative of potential species**
3. **Taxonomy assignment**
4. Exploratory analysis
   - Analysis of diversity measures and their visualization
5. Inference analysis
   - Associating composition with variables of interest

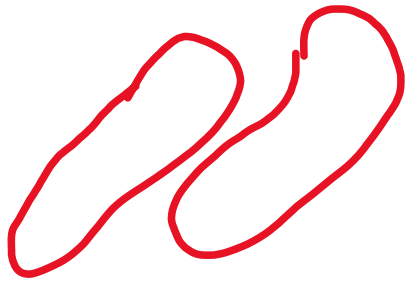


MUNI | RECETOX

Marker-gene metagenomics

Step 2. Identification of sequences representative of potential species

- **Aim:** Organize reads according to their individual/organism of origin



- **Theory:** one read ~ one gene ~ one individual ~ one species
- **Reality:**
  - one species can have **multiple and different copies of a gene**

AACCGTCGACGGTCAT
AACCGTCGACGGTCAT
AACCGTCGACGGTCAT

  - one sequence can be **shared** by **multiple** species

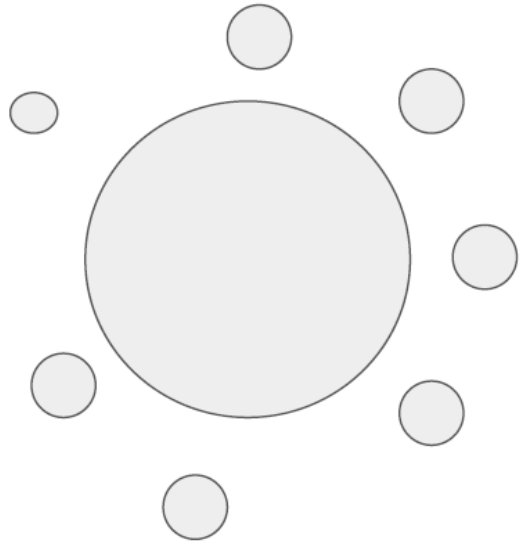
TTGCCATGACGATATA
TTGCCATGACGATATA

  - problem to **distinguish** sequencing **error** from a real **change** between species

Marker-gene metagenomics

Step 2. Identification of sequences representative of potential species

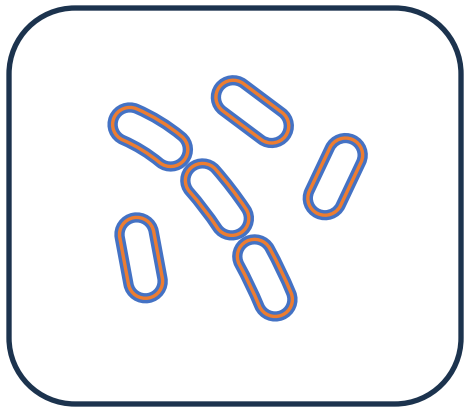
- Solution: Clustering/binning of **similar sequences** into an **OTU – operational taxonomic unit**
- Similarity: 97% or less or more...
- "OTU picking"
- Final representative sequence of an OTU is a **consensus sequence** (average, …)

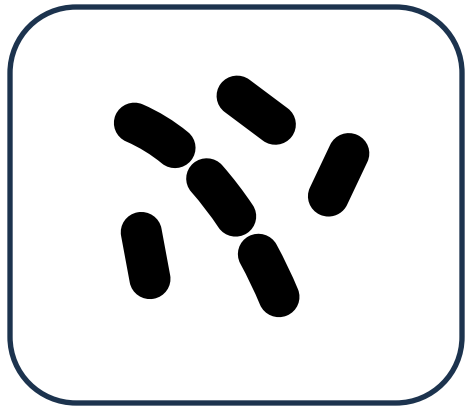


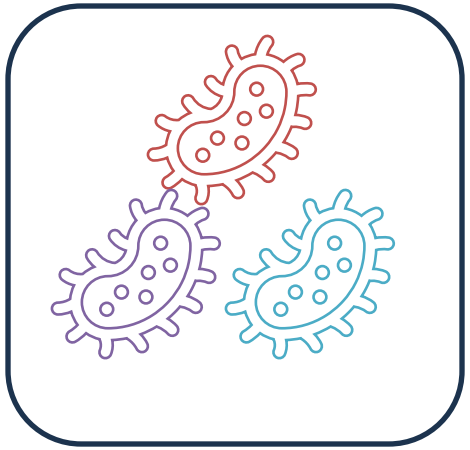
Marker-gene metagenomics

Clustering *de novo*

- Clustering based on similarity of sequencing **without taking into account reference database**





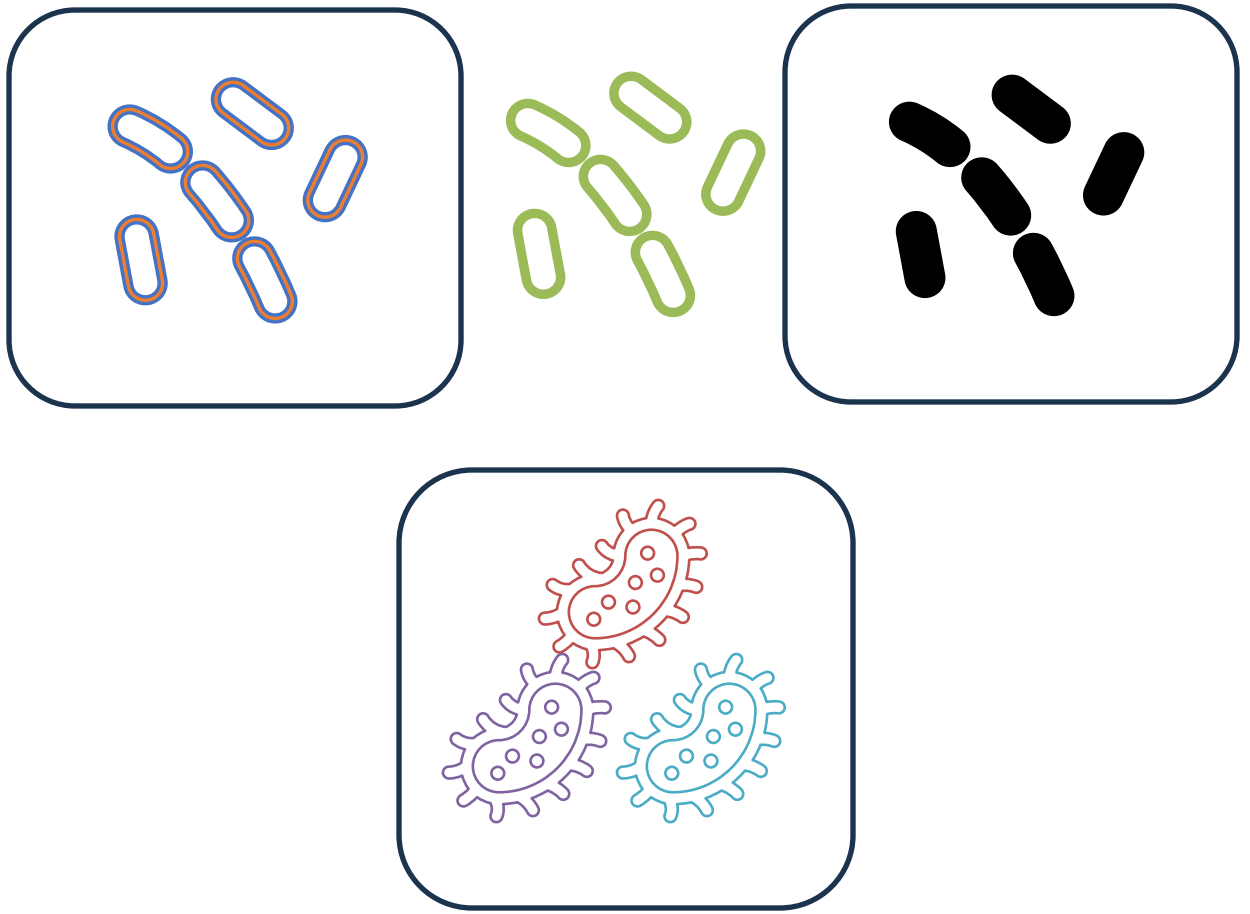


Marker-gene metagenomics

Clustering *de novo*

- Clustering based on similarity of sequencing **without taking into account reference database**

- Disadvantages:
  - if new samples added, we need to **recluster** (reanalyze)

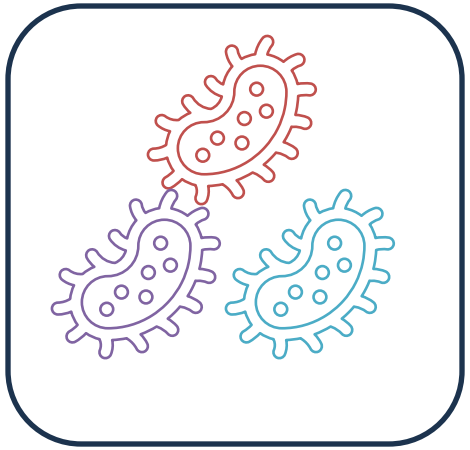


Marker-gene metagenomics

Clustering *de novo*

- Clustering based on similarity of sequencing **without taking into account reference database**

- Disadvantages:
  - if new samples added, we need to recluster (reanalyze), this means, **clustering can change**

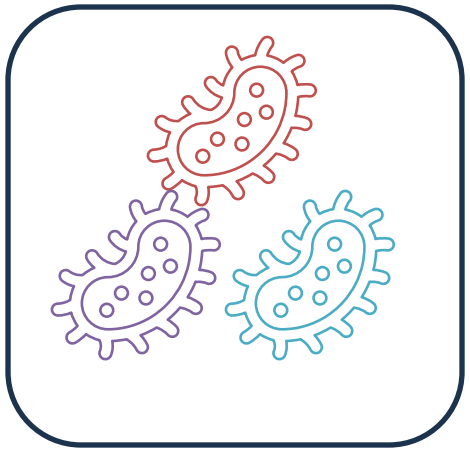


Marker-gene metagenomics

Clustering *de novo*

- Clustering based on similarity of sequencing **without taking into account reference database**

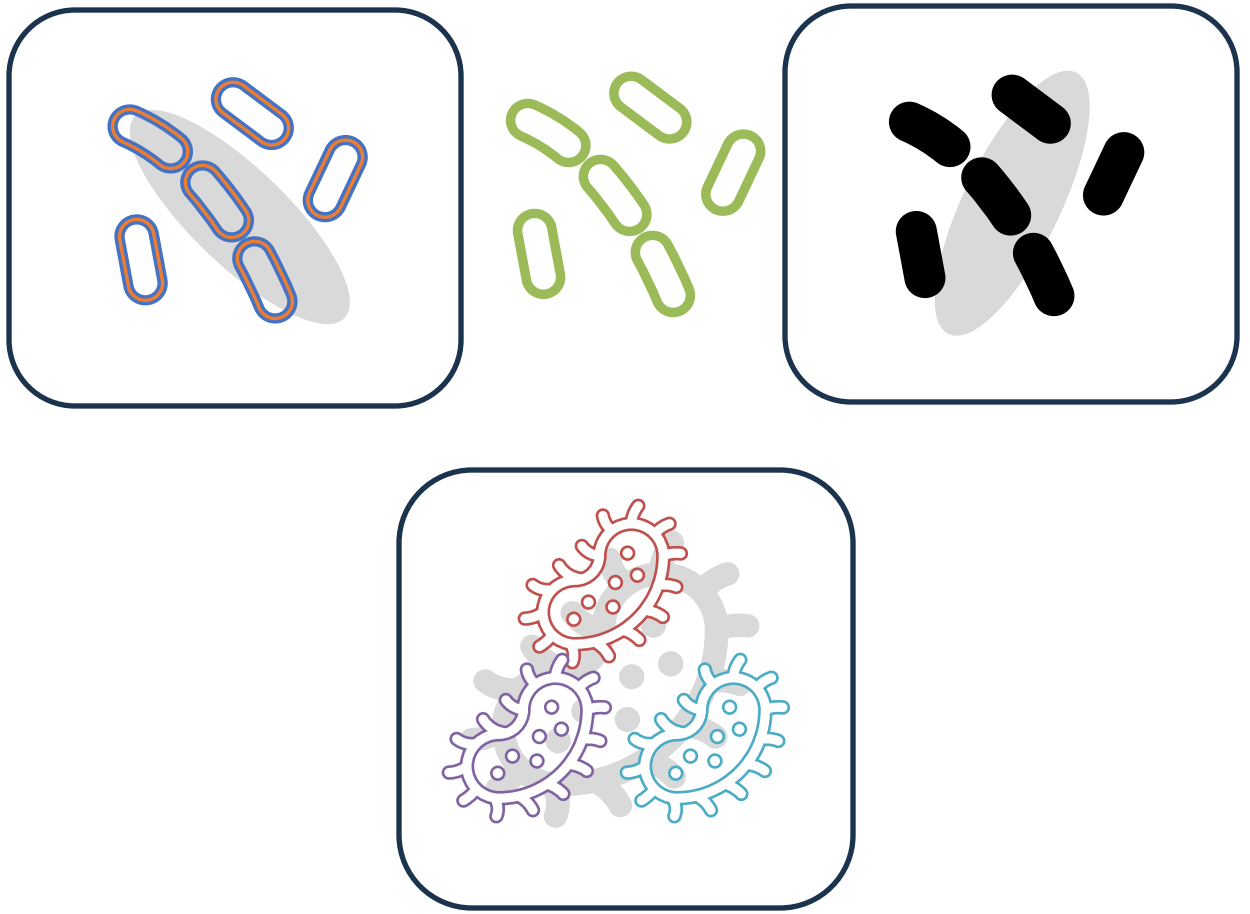
- Disadvantages:
  - if new samples added, we need to recluster (reanalyze), this means, **clustering can change**
  - computationally expensive
  - only way if reference is unknown



Marker-gene metagenomics

Clustering - *closed reference*

- Using reference databases, we cluster around **known sequences**

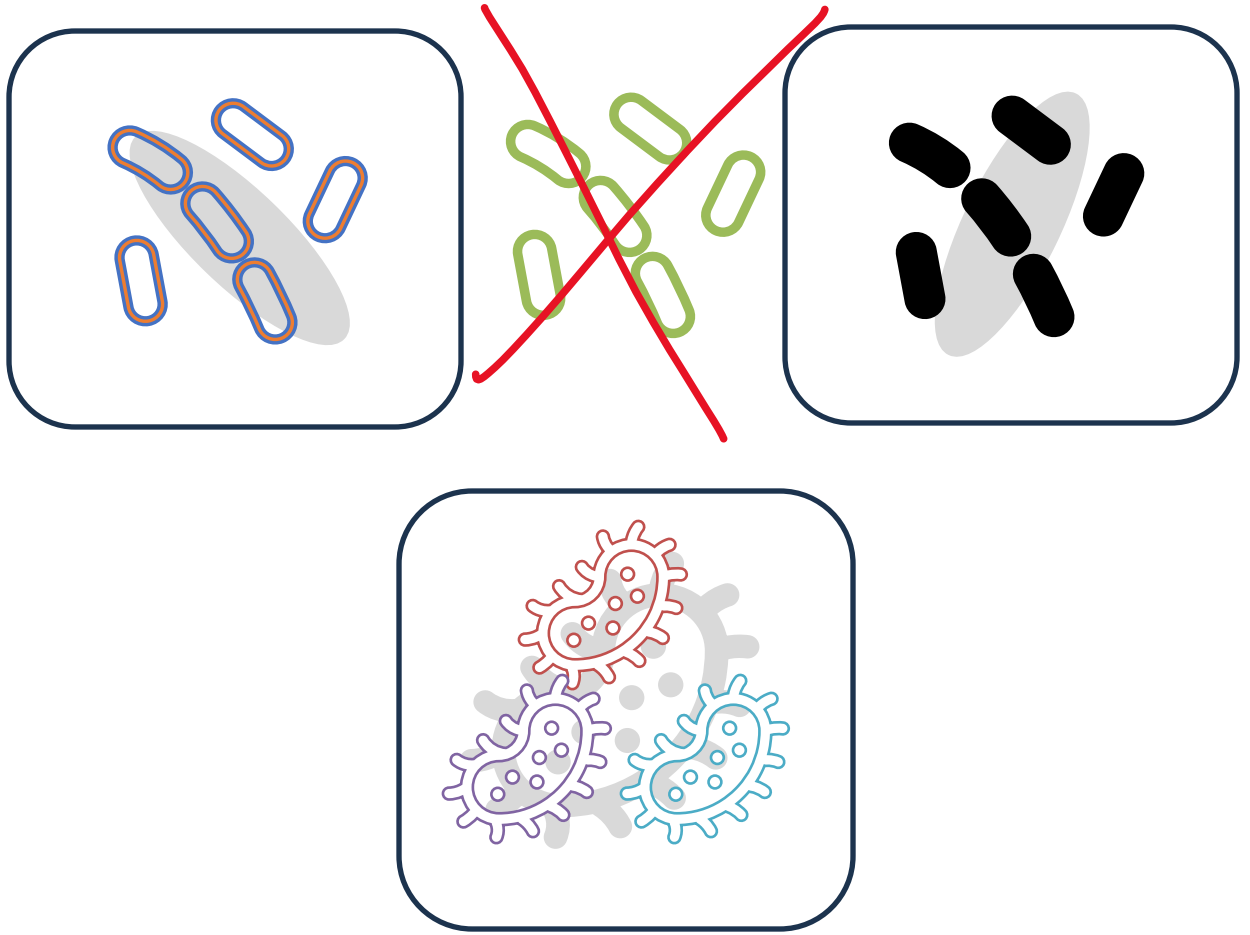


Marker-gene metagenomics

Clustering - *closed reference*

- Using reference databases, we cluster around **known sequences**

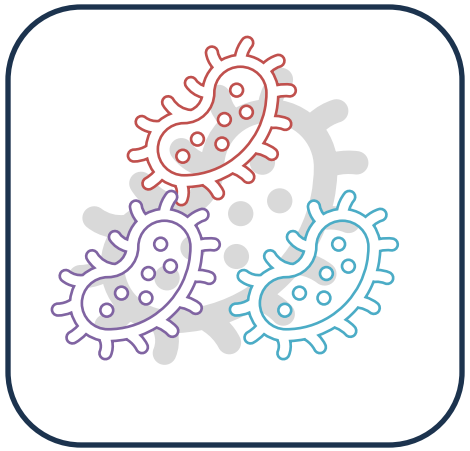
- Disadvantages:
  - only for **well characterized** types of samples (stool?)
  - **discarding** unknown



Marker-gene metagenomics

Clustering – *open reference*

- Using reference databases, we cluster around **known sequences**
- **Those that did not cluster with reference are clustered de novo**

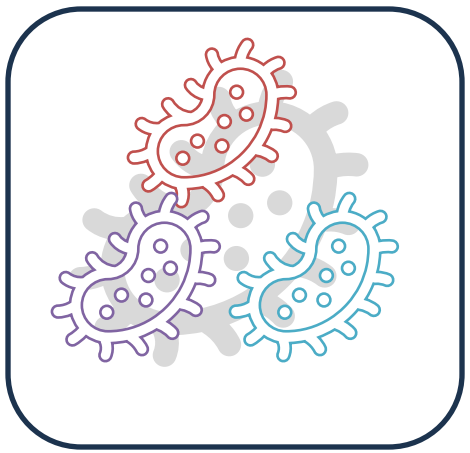


Marker-gene metagenomics

Clustering – *open/close reference*

- PROBLEM: **REFERENCE BIAS**

well studied microbes have hundreds to thousands sequences as opposed to few or none of the less studied ones



Method	Function supported	Alignment method	Clustering method[a]	Using reference database	Generating distance matrix	Computational complexity	Space complexity
DOTUR	Clustering	N/A	HC	N	Y	$O(N^2)$	$O(N^2)$
Mothur	Sequence alignment + clustering	Profile based MSA method	HC	Y	Y	$O(N^2)$	$O(N^2)$
ESPRIT	Sequence alignment + clustering	PSA	HC	N	Y	$O(N^2)$	$O(N^2)$
ESPRIT-Tree	Sequence alignment + clustering	PSA	HC	N	N	$O(N^{1.2})$	$O(N)$
NAST[b]	Sequence alignment	Profile based MSA method	N/A	Y	Y	$O(N)$	$O(N^2)$
RDP/Pyro	Sequence alignment + clustering	Infernal aligner	HC	Y	Y	$O(N^2)$	$O(N^2)$
CD-HIT	Sequence alignment + clustering	PSA	Greedy heuristic clustering	N	N	$O(N^{1.2})$	$O(N)$
UCLUST	Sequence alignment + clustering	PSA	Greedy heuristic clustering	N	N	$O(N^{1.2})$	$O(N)$
MUSCLE	Sequence alignment	MSA	N/A	N	Y	$O(N^4)$	$O(N^2)$

[a]Complete linkage is the default method in DOTUR, mothur, ESPRITand RDP/Pyro. ESPRIT-Tree supports only average linkage. [b]NASTonly supports the sequence-alignment step. By aligning query sequences against a database, its computational complexity grows linearly with respect to the number of sequences. However, according to the NASTwebsite, it aligns at a rate of approximately I0 sequences per minute. N/A = not applicable; N = no; Y = yes.

From Sun Y1, Cai Y, Huse SM, Knight R, Farmerie WG, Wang X, Mai V. (2011) **A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis.** Brief Bioinform. 2012 Jan;13(1):107-21

Marker-gene metagenomics

Bunch of problems...

Substitutions

ACTGCTAGC

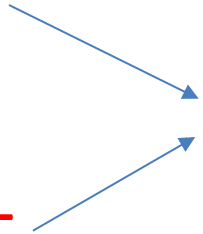
ACTGATAGC

- PCR
- sequencing

Chimeras

ACTGTAGC

AGACGGCT



ACTGGGCT

- PCR

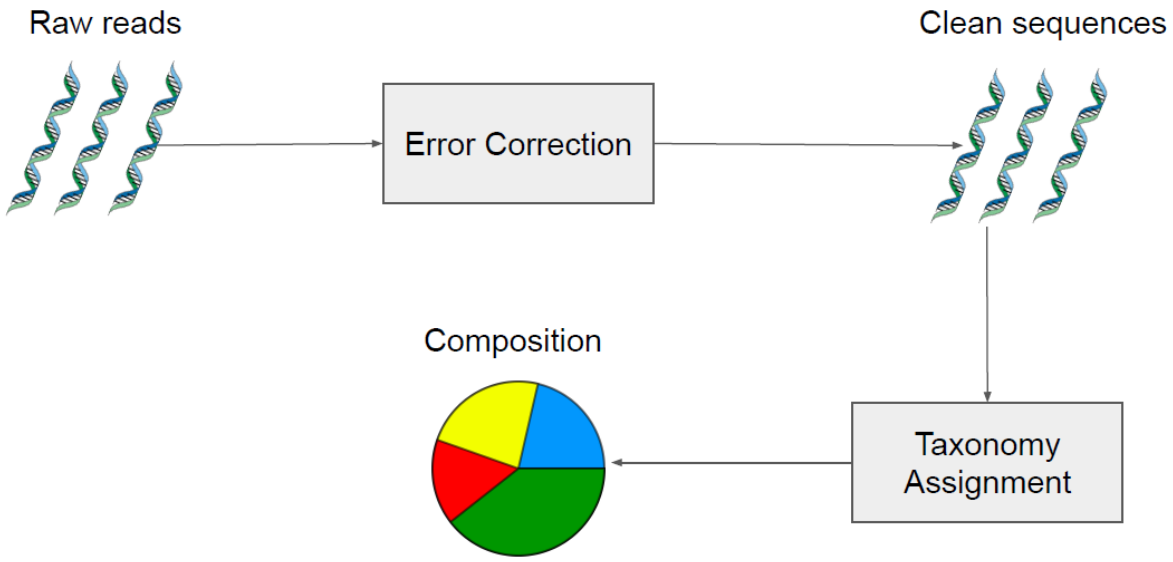
How can we tell errors from real changes?

Marker-gene metagenomics

Correcting errors

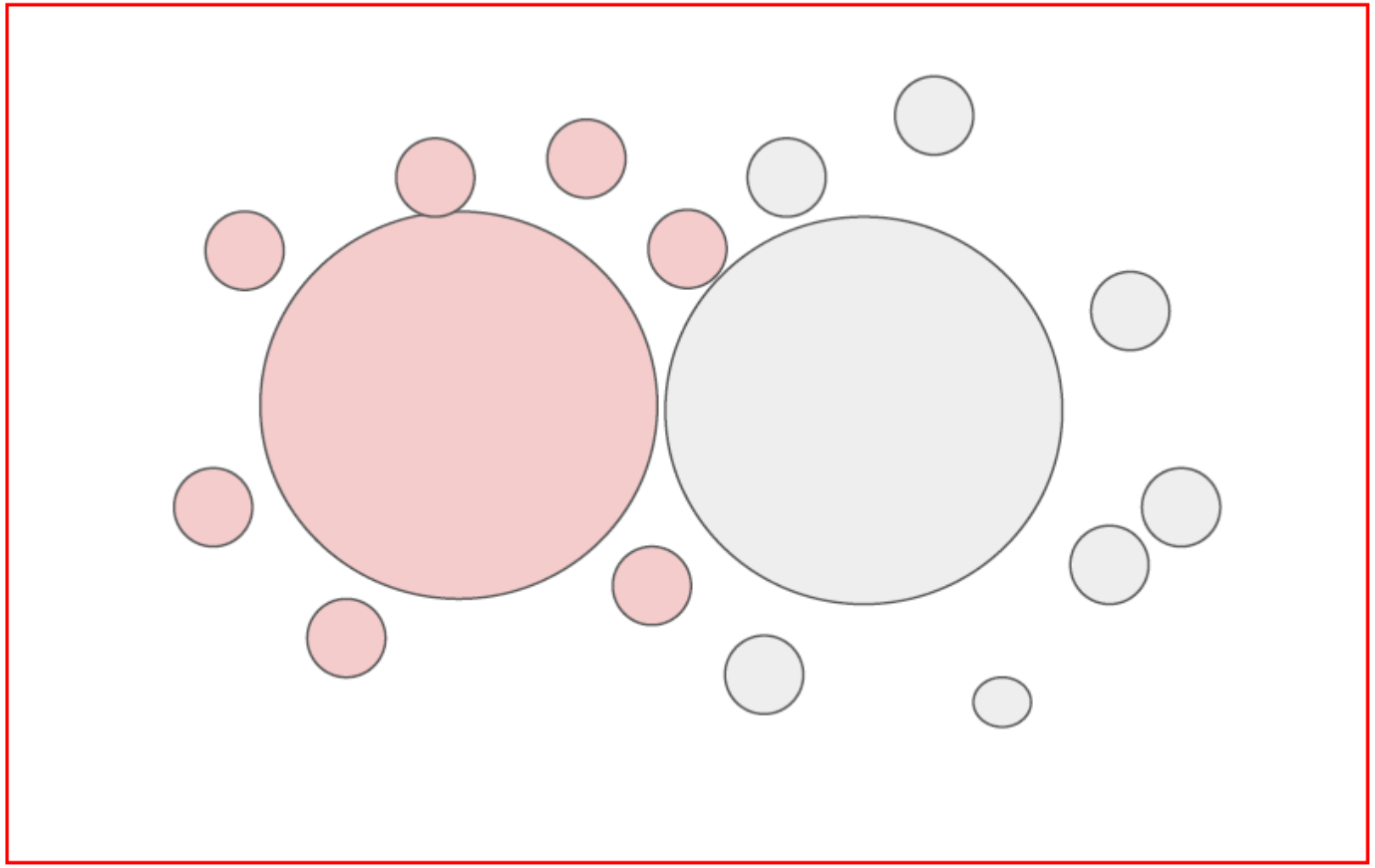
- **Typical:** QC, trimming, N-filtering, length-filtering, adapter removal

- **OTU picking** per se helps correcting some errors, but the problem persists



Marker-gene metagenomics

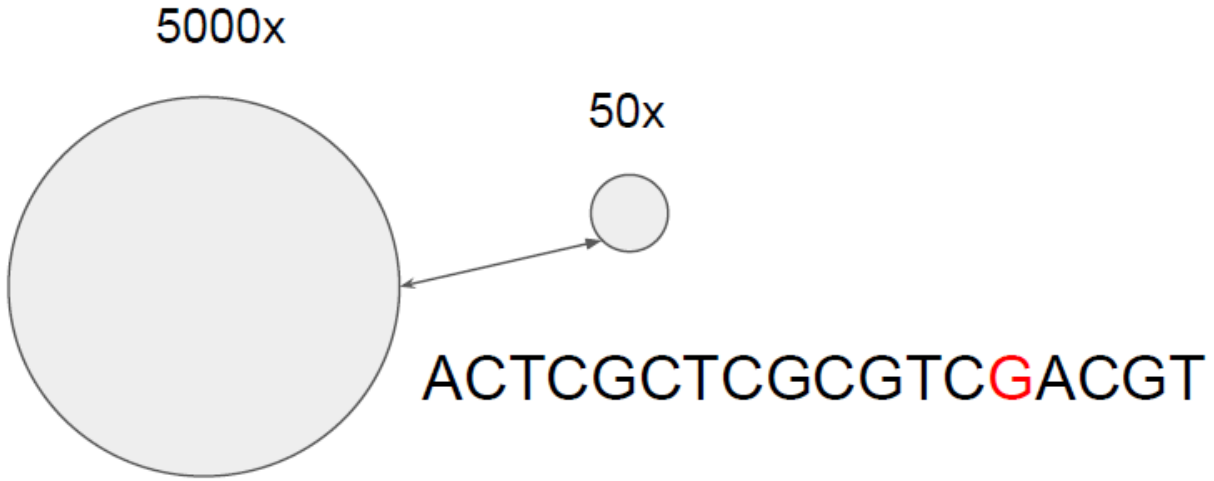
OTU picking



97% Threshold is arbitrary

Marker-gene metagenomics

Let's be smarter - build an error model



We can calculate posterior probability of a sequence being real vs artefact, based on our knowledge of sequence errors and their frequencies.

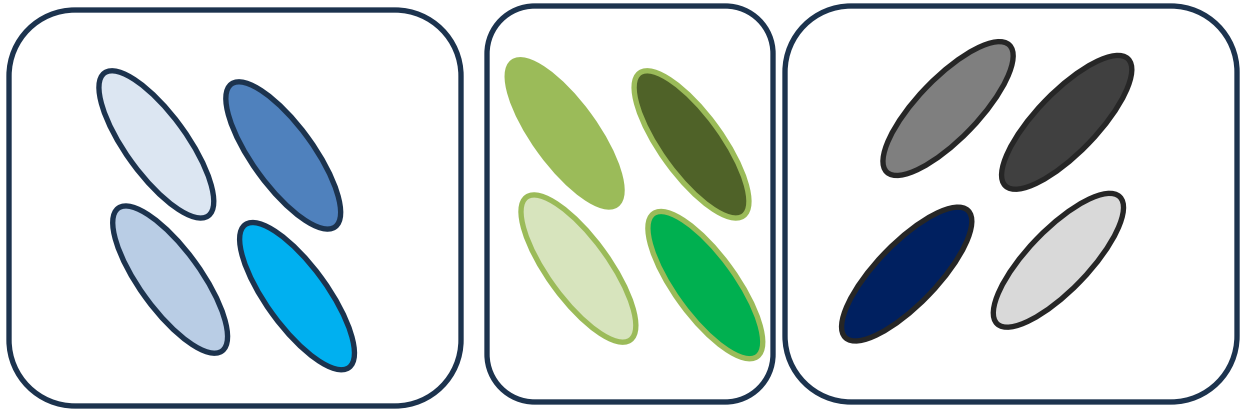
We can test whether our observed frequency is larger than expected frequency and get a p-value.

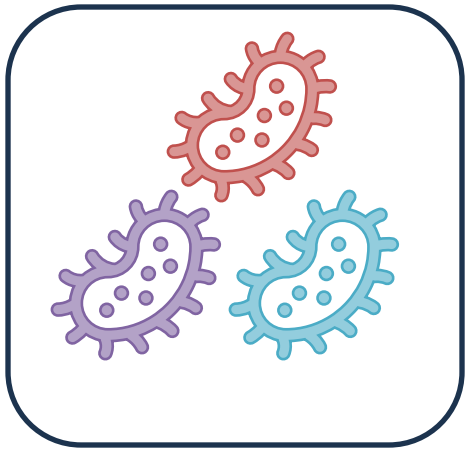
If we reject the hypothesis, we keep that sequence!

Marker-gene metagenomics

Applying error model

- I discard improbable sequences

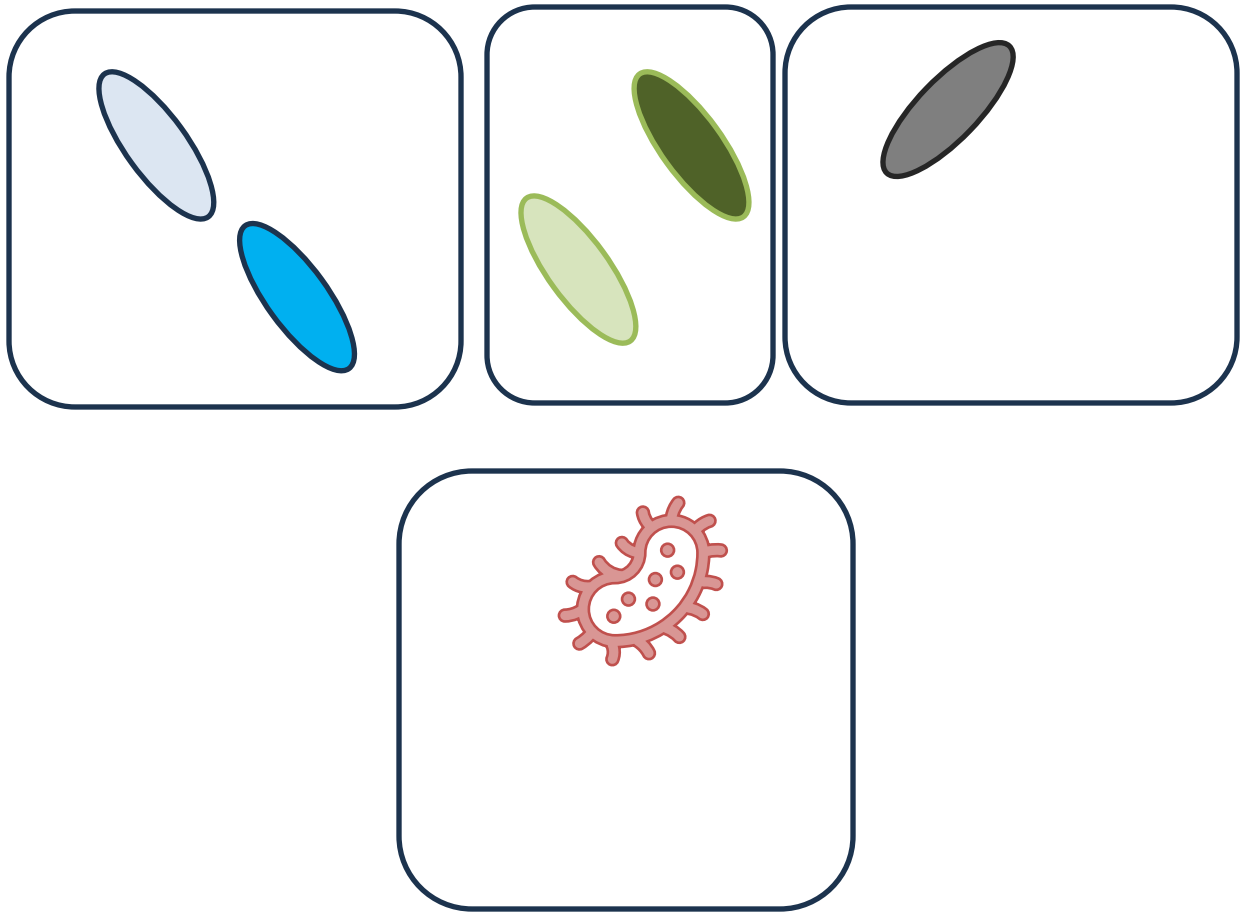




Marker-gene metagenomics

Applying error model

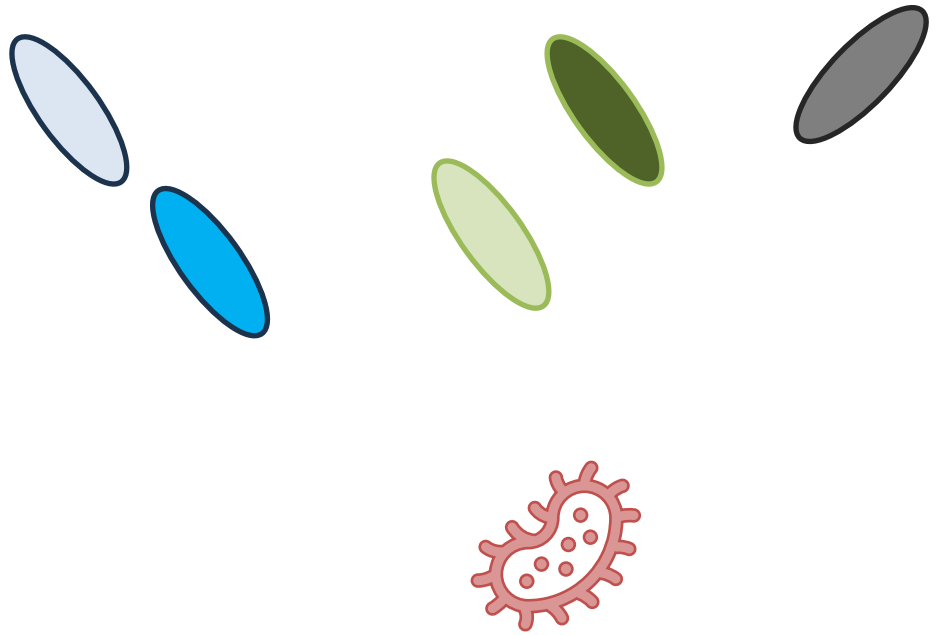
- I discard improbable sequences



Marker-gene metagenomics

Applying error model

- I discard improbable sequences
- Each sequence now represents a taxonomical unit



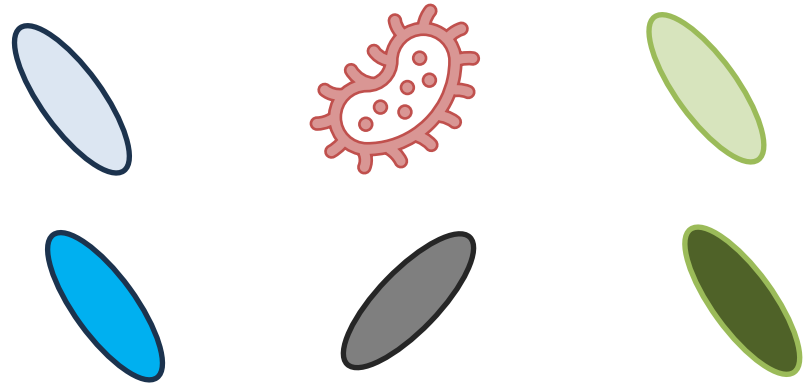
Marker-gene metagenomics

Applying error model – getting ASVs

- I discard improbable sequences

- Each sequence now represents a taxonomical unit called

Amplicon **S**equence **V**ariant – **ASV**



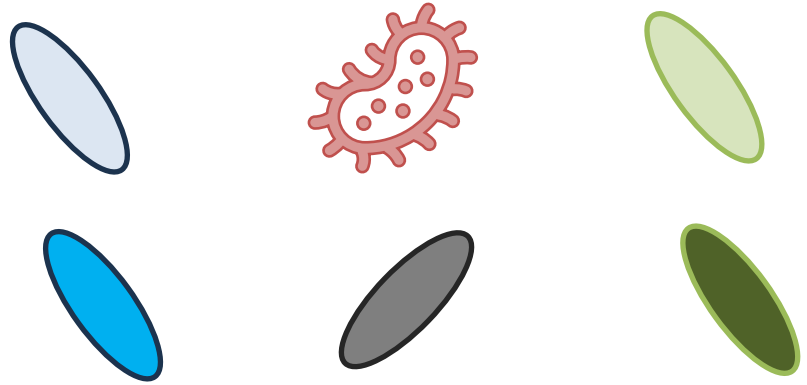
- **Only 1 sequence represents the bacteria**

Also known as ESV (exact sequence variant) or zOTU (zero-radius OUT)

Marker-gene metagenomics

Applying error model – getting ASVs

- Already existing ASVs are not changed if new samples are added

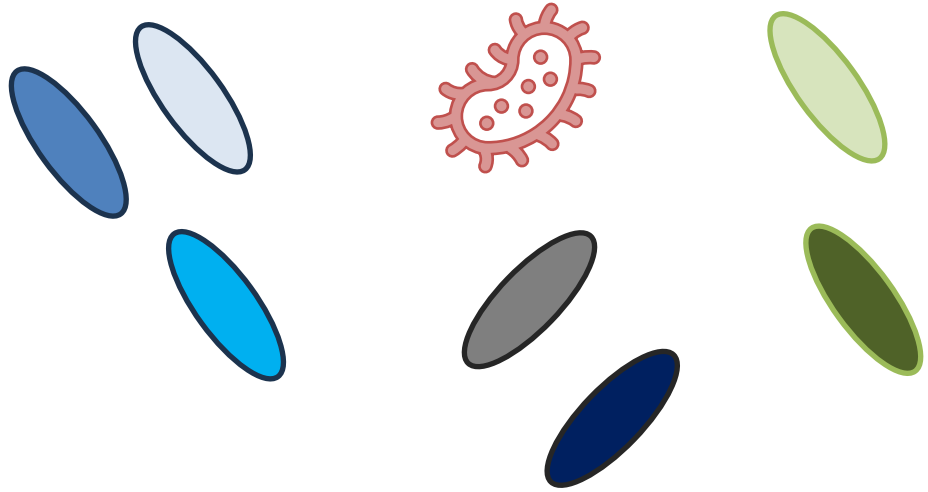


Marker-gene metagenomics

Applying error model – getting ASVs

- Already existing ASVs are not changed if new samples are added

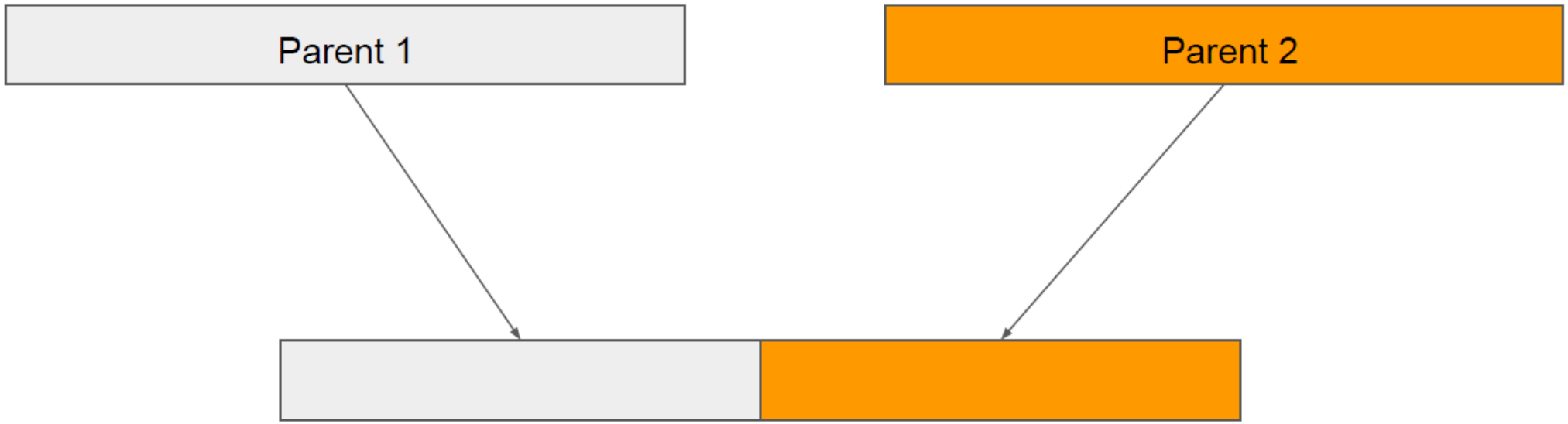
- However! Adding new samples can result in previously discarded sequences becoming more frequent and become ASVs!



Marker-gene metagenomics

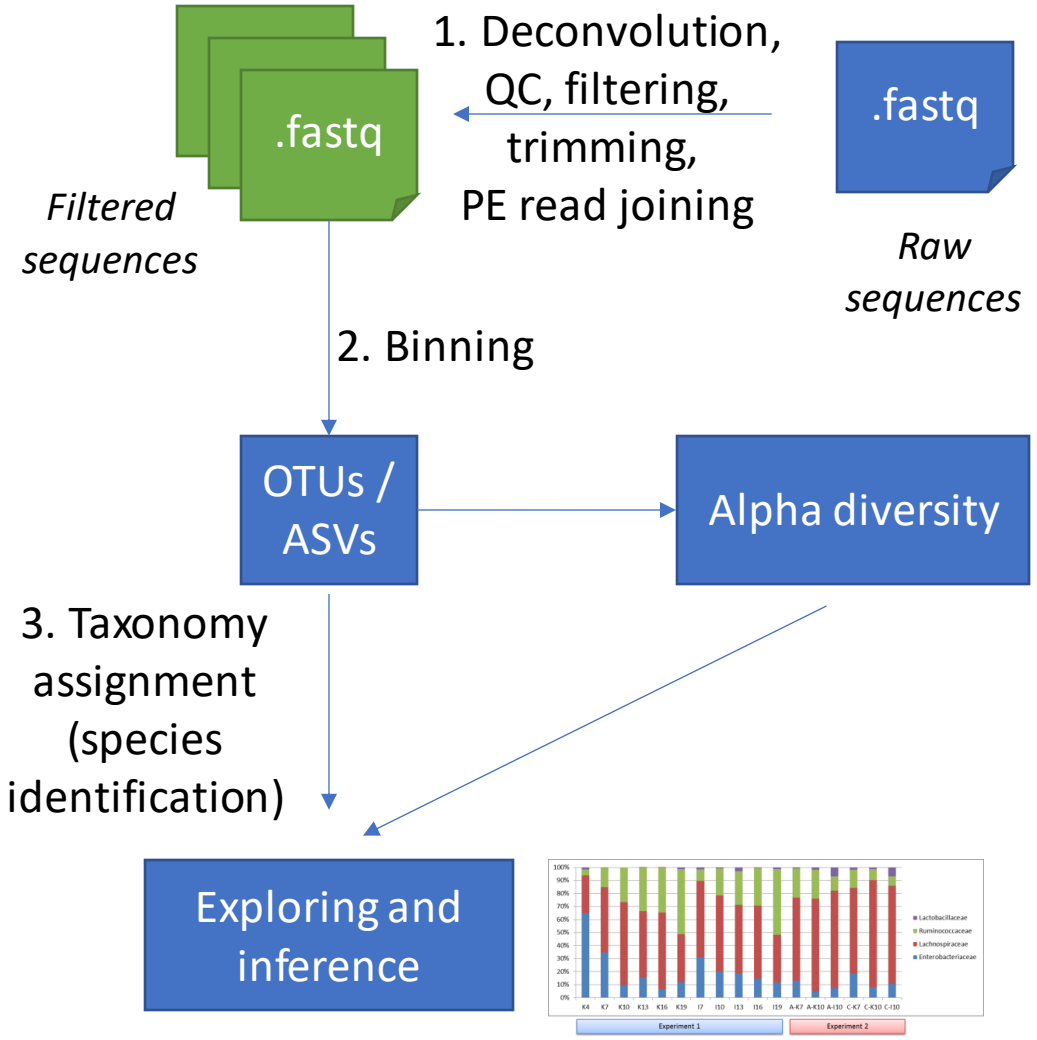
Chimera removal

1. Chimera must originate from other sequences in sample
2. Chimera will be low-abundant
3. Chimera will align well to combination of two parent sequences



The usual data analysis pipeline

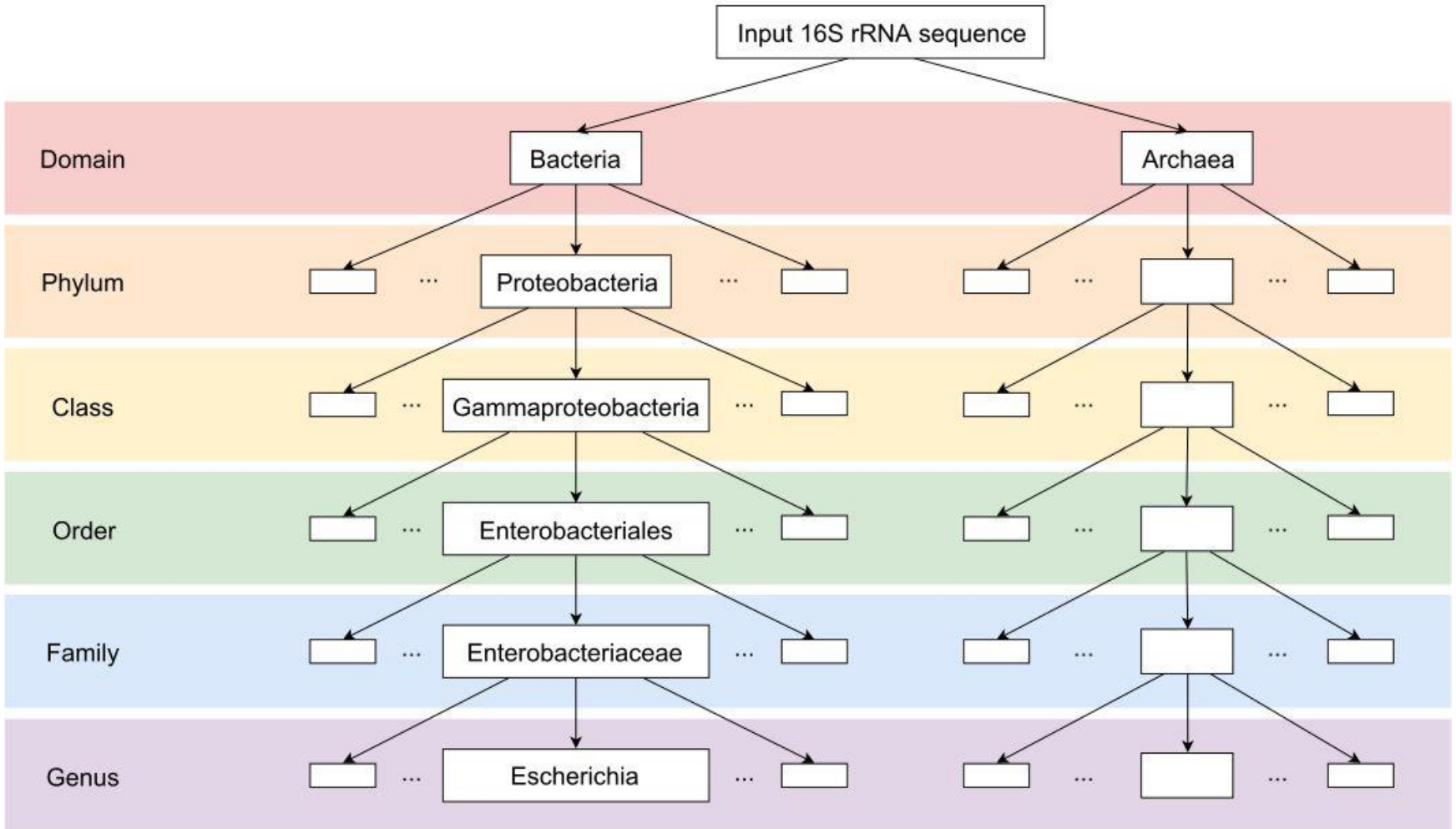
1. **Preprocessing:** Fastq files preprocessing, deconvolution, QC, trimming, joining reads
2. **Identification of sequences representative of potential species**
3. **Taxonomy assignment**
4. Exploratory analysis
   - Analysis of diversity measures and their visualization
5. Inference analysis
   - Associating composition with variables of interest



MUNI | RECETOX

Marker-gene metagenomics

Step 3. Taxonomy assignment



Marker-gene metagenomics

Step 3. Taxonomy assignment

- A relatively easy task
- Just alignd ASVs to the reference db!
- BLAST, HMMER or USEARCH
- **But – how to deal with multiple results?**

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Colibacter massiliensis strain Marseille-P2911 16S ribosomal RNA, partial sequence	2830	2830	98%	0.0	100%	NR_147373.1
Colibacter massiliensis partial 16S rRNA gene, type strain Marseille-P2911T	2830	2830	98%	0.0	100%	LT576403.1
Megasphaera sp. oral clone CS025 16S ribosomal RNA gene, partial sequence	2787	2787	99%	0.0	99%	AF287784.1
Megasphaera sp. oral clone BS073 16S ribosomal RNA gene, partial sequence	2756	2756	97%	0.0	99%	AF287785.1
Uncultured organism clone ELU0026-T115-S-NIPCRAMqANa_000399 small subunit riboso	2706	2706	95%	0.0	99%	HQ749314.1
Uncultured organism clone ELU0026-T115-S-NIPCRAMqANa_000160 small subunit riboso	2700	2700	95%	0.0	99%	HQ749075.1
Uncultured organism clone ELU0026-T115-S-NIPCRAMqANa_000396 small subunit riboso	2693	2693	95%	0.0	99%	HQ749311.1
Uncultured organism clone ELU0026-T115-S-NIPCRAMqANa_000369 small subunit riboso	2693	2693	95%	0.0	99%	HQ749284.1
Uncultured organism clone ELU0026-T115-S-NIPCRAMqANa_000274 small subunit riboso	2691	2691	95%	0.0	99%	HQ749189.1
Uncultured organism clone ELU0025-T109-S-NIPCRAMqANa_000450 small subunit ribosc	2678	2678	95%	0.0	99%	HQ748787.1
Uncultured organism clone ELU0040-T218-S-NIPCRAMqANa_000085 small subunit ribosc	2665	2665	95%	0.0	99%	HQ757389.1
Uncultured organism clone ELU0026-T115-S-NIPCRAMqANa_000128 small subunit riboso	2663	2663	95%	0.0	99%	HQ749043.1

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

Ok, all agree

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

Ok, all agree

RECETOXBACTERIALES

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

RECETOXBACTERIALES

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

RECETOXBACTERIALES

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

Ok, most agree

RECETOXBACTERIALES

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

Ok, most agree

RECETOXBACTERIALES **RECETOXBACTERIA**

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

RECETOXBACTERIALES **RECETOXBACTERIA**

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

RECETOXBACTERIALES **RECETOXBACTERIA**

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

???

RECETOXBACTERIALES **RECETOXBACTERIA**

Marker-gene metagenomics

LCA – Least common ancestor

Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

???

RECETOXBACTERIALES **RECETOXBACTERIA** **UNASSIGNED**

Marker-gene metagenomics

LCA – Least common ancestor

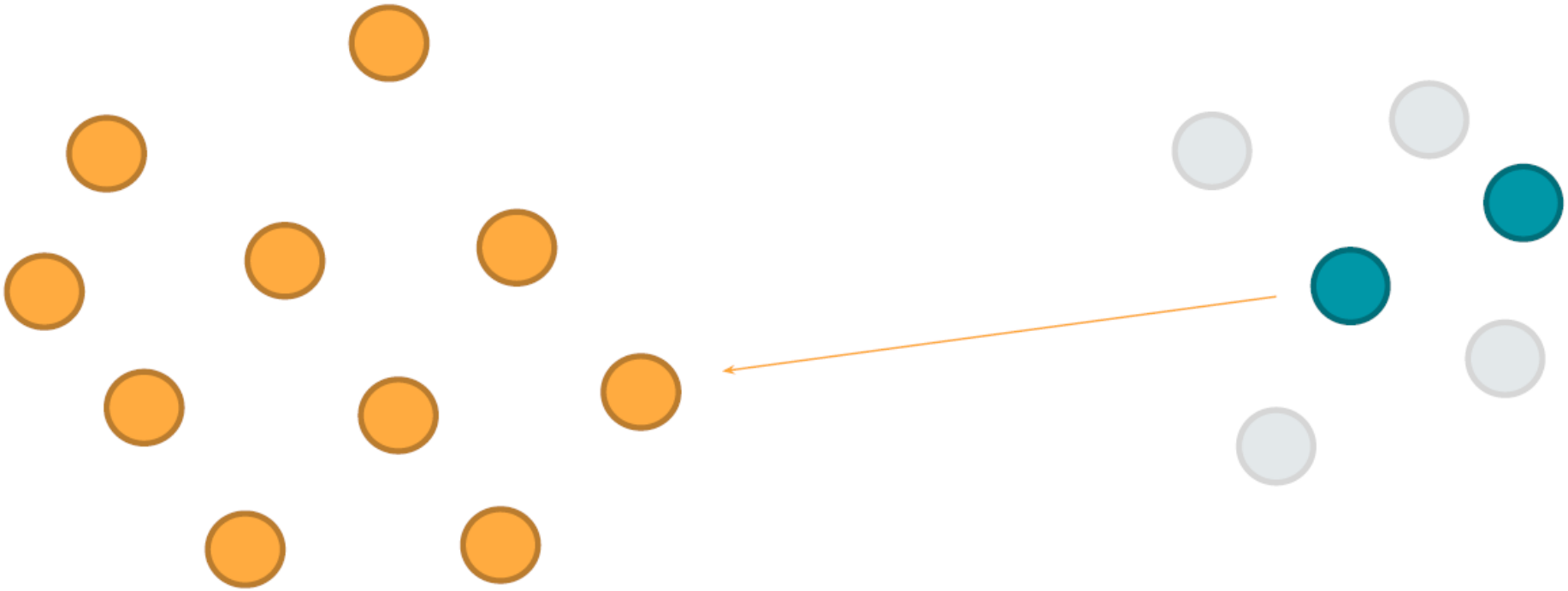
Hits:

RECETOXBACTERIALES	RECETOXBACTERIA	Smatanobacter
RECETOXBACTERIALES	RECETOXBACTERIA	Zwinseria
RECETOXBACTERIALES	RECETOXBACTERIA	E. Budinski
RECETOXBACTERIALES	INBITAE	Micenkobacterium

RECETOXBACTERIALES **RECETOXBACTERIA** **UNASSIGNED**

Marker-gene metagenomics

Database bias



Marker-gene metagenomics

Reference databases for 16SrRNA

- SILVA
- GreenGenes
- RDP

They also use 16s rRNA sequences from metagenomic studies. These were **already classified by classifier.**

- RDP Training set
- BLAST 16s rRNA

They contain 16s rRNA of well-characterized species. However, are **incomparably smaller.**

Beware of regularly updated versions of the dbs

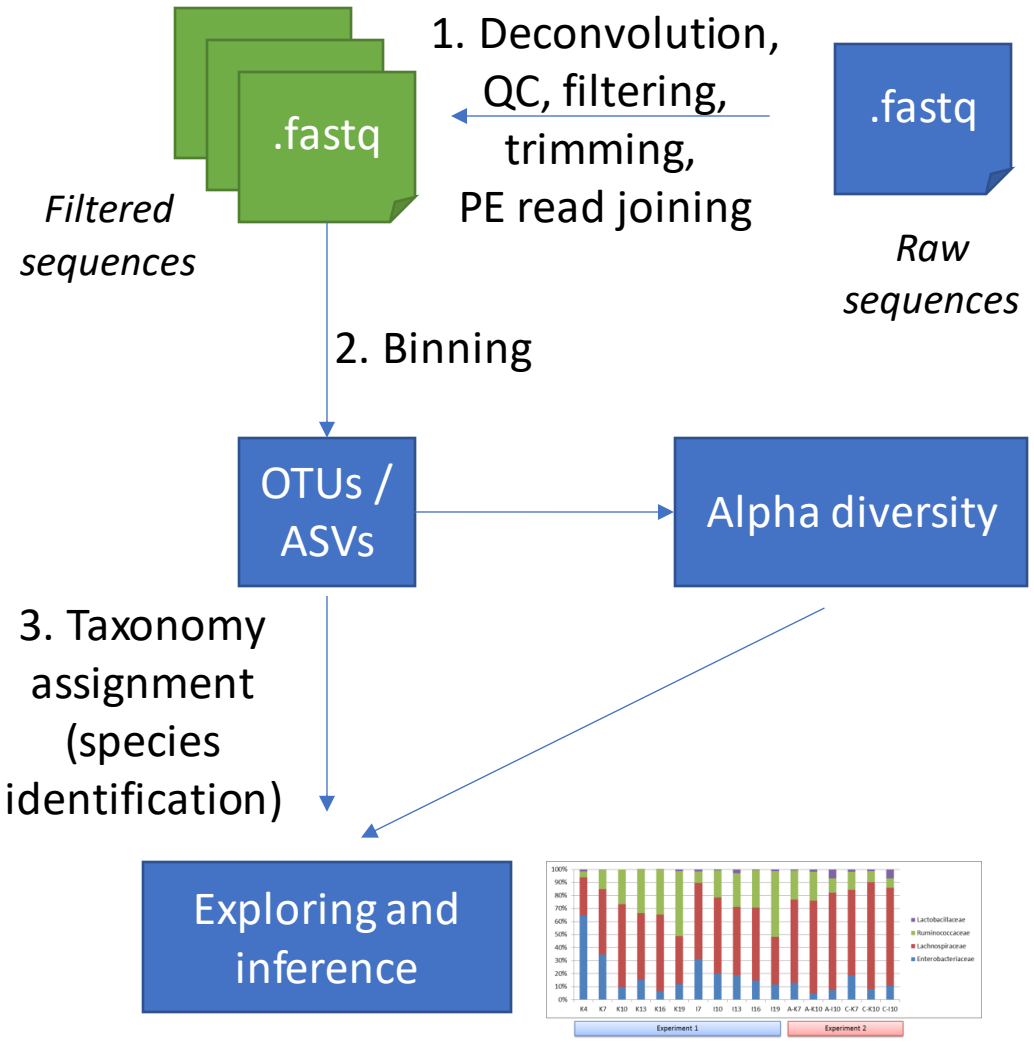
Marker-gene metagenomics

Taxonomy classifiers for 16S rRNA gene sequences

	16s rRNA
Error correction - clustering	Uclust, cd-hit
Error correction - denoising	DADA2, unoise
Taxonomy assignment	Usearch, BLAST, RDP
Pathways, metabolic potential	Picrust 2 (prediction)
Toolsets/pipelines:	QIIME 1, QIIME 2

The usual data analysis pipeline

1. **Preprocessing:** Fastq files preprocessing, deconvolution, QC, trimming, joining reads
2. **Identification of sequences representative of potential species**
3. **Taxonomy assignment**
4. Exploratory analysis
   - Analysis of diversity measures and their visualization
5. Inference analysis
   - Associating composition with variables of interest



MUNI | RECETOX

Alpha diversity (*within-sample* diversity)

- **Used in ecology, a measure** of **how diverse** a single sample is

1. Shannon index

$$H = -\sum[(p_i) \times \log(p_i)]$$

p_i - proportion of individuals of i-th species in a whole community;

2. Simpson index

$$D = \frac{\sum n_i(n_i - 1)}{N(N-1)}$$

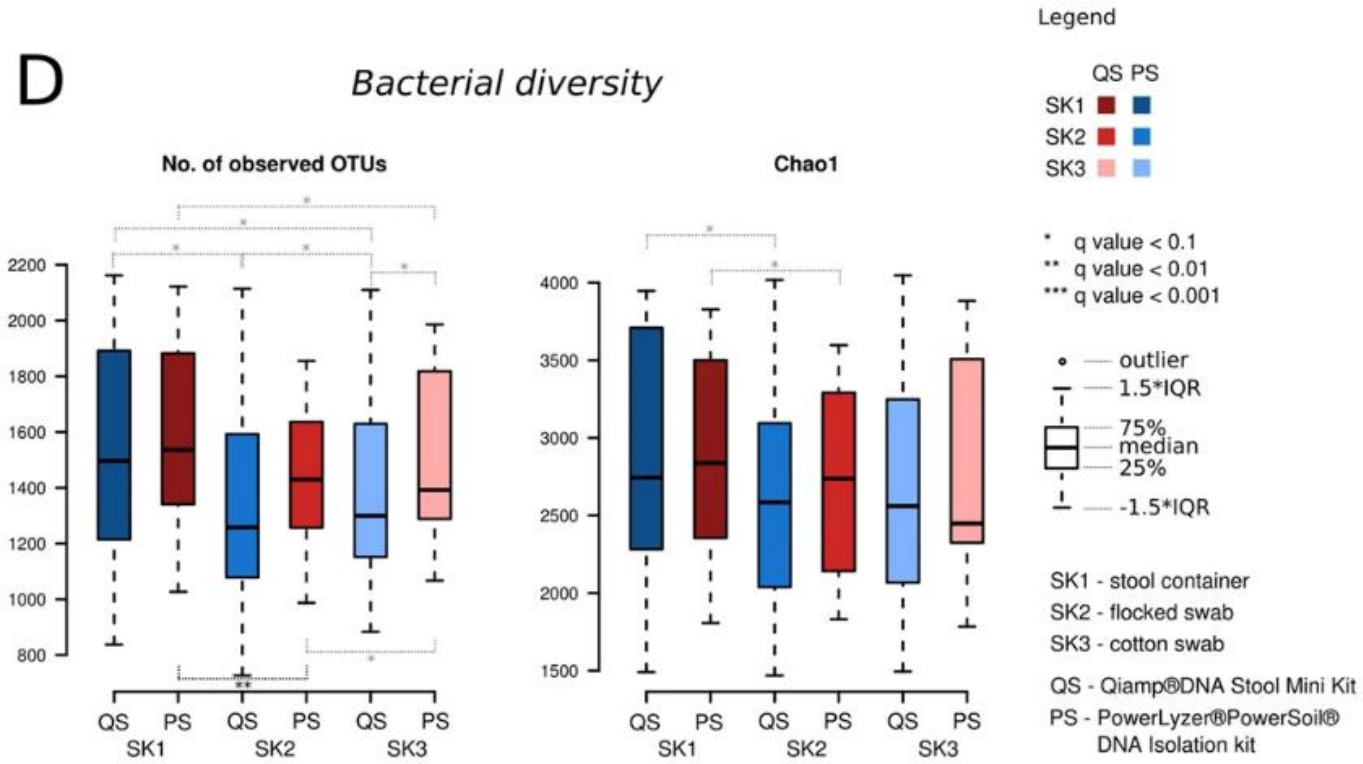
- n_i — Number of individuals in the i-th species; and
- N — Total number of individuals in the community.

3. Number of ASVs

MUNI | RECETOX

Alpha diversity (*within-sample* diversity)

3. Number of OTUs/ASVs



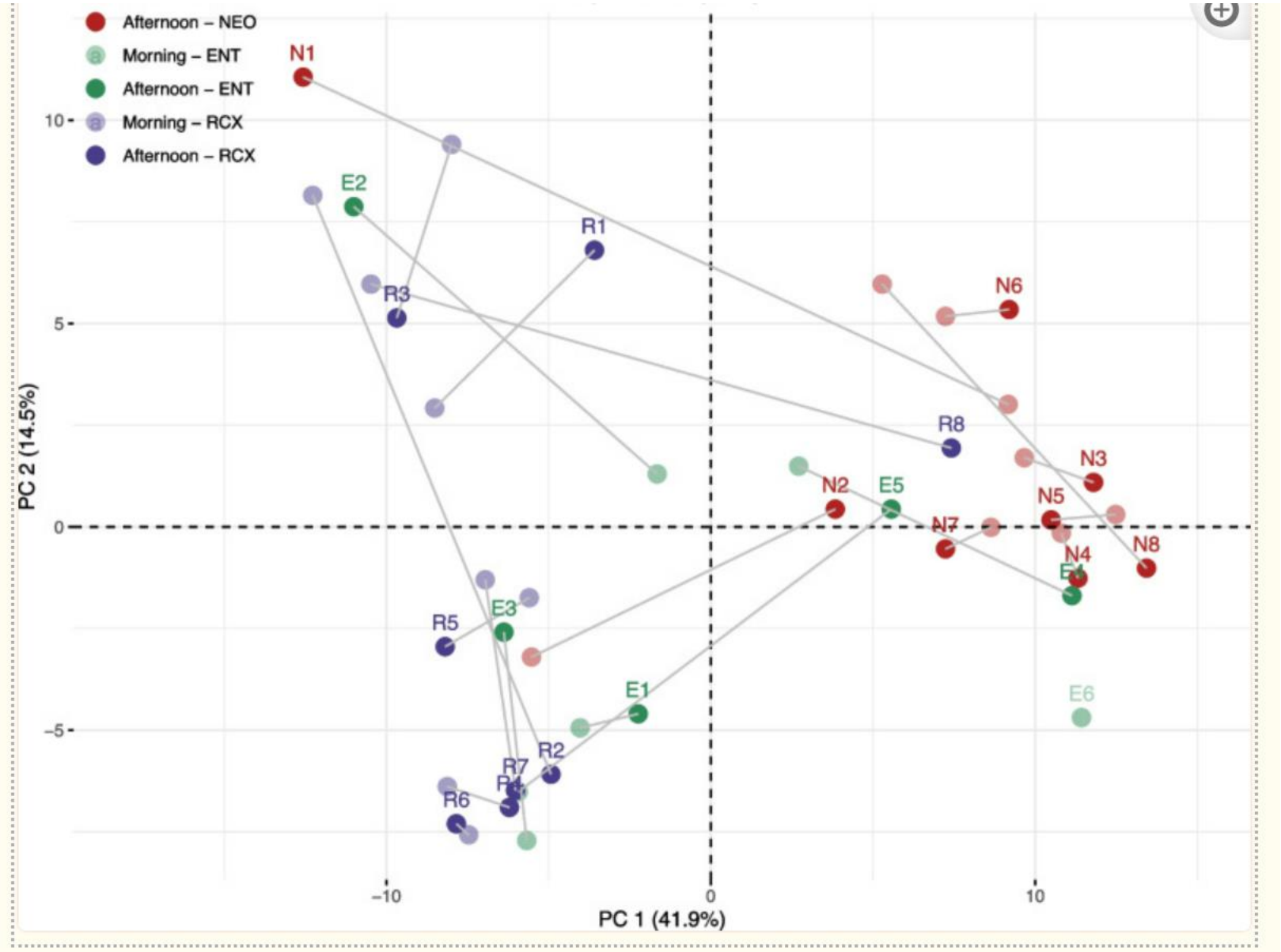
4. Chao 1

$A = N + S^2 / (2\,D)$

- N is the number of OTUs/ASVs
- S is the number of singleton OTUs/ASVs
- D is the number of doublet OTUs, i.e. OTUs with abundance 2.

MUNI | RECETOX

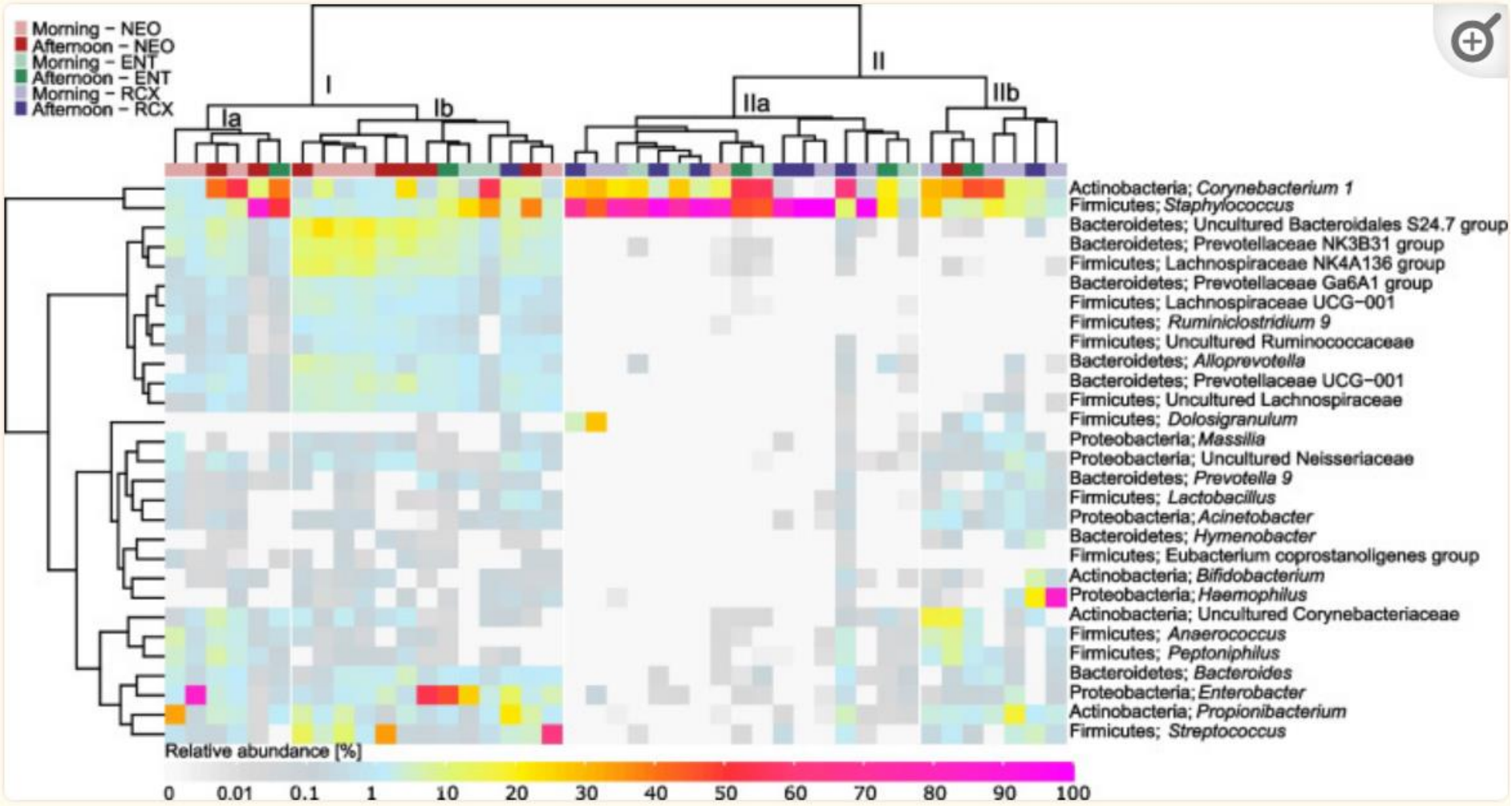
PCA – principal component analysis



Based on the normalized c ompositional profiles

MUNI | RECETOX

Clustering

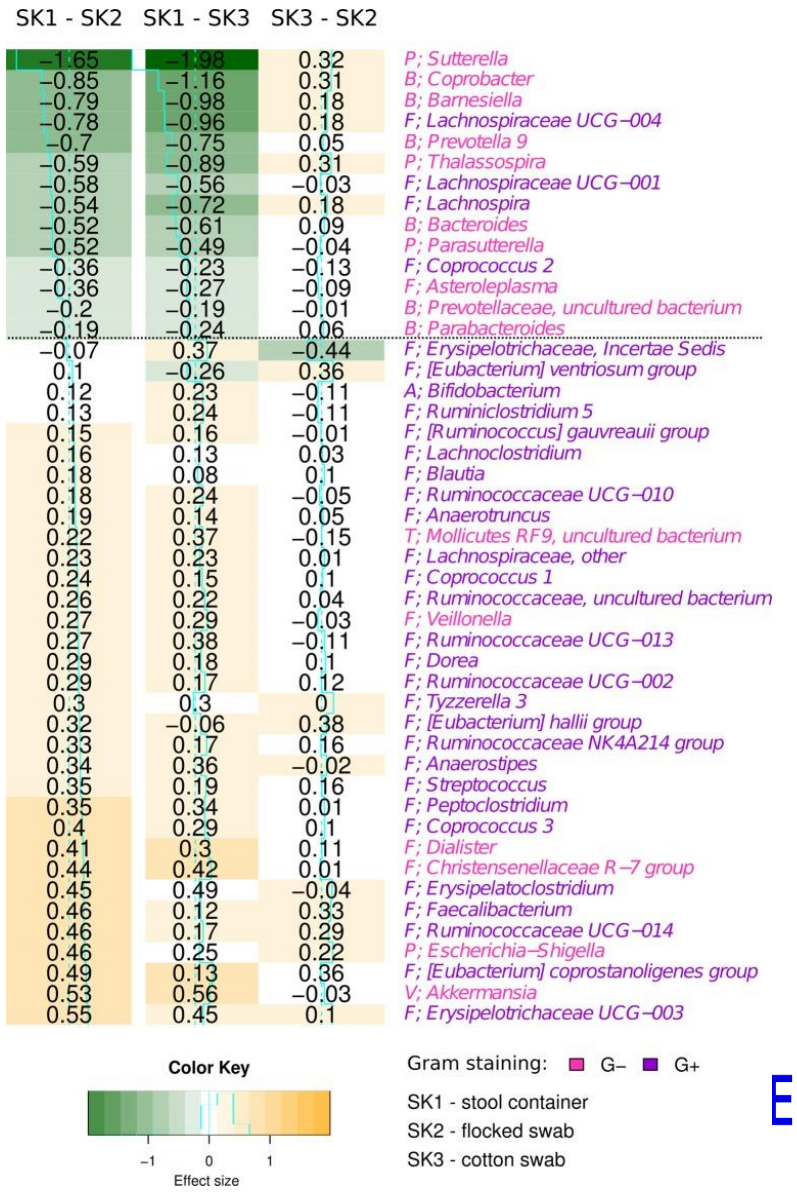


Based on the normalized c ompositional profiles

MUNI | RECETOX

Group comparison – comparing differences between groups

Applying statistical testing on each bacteria to determine difference in their abundance



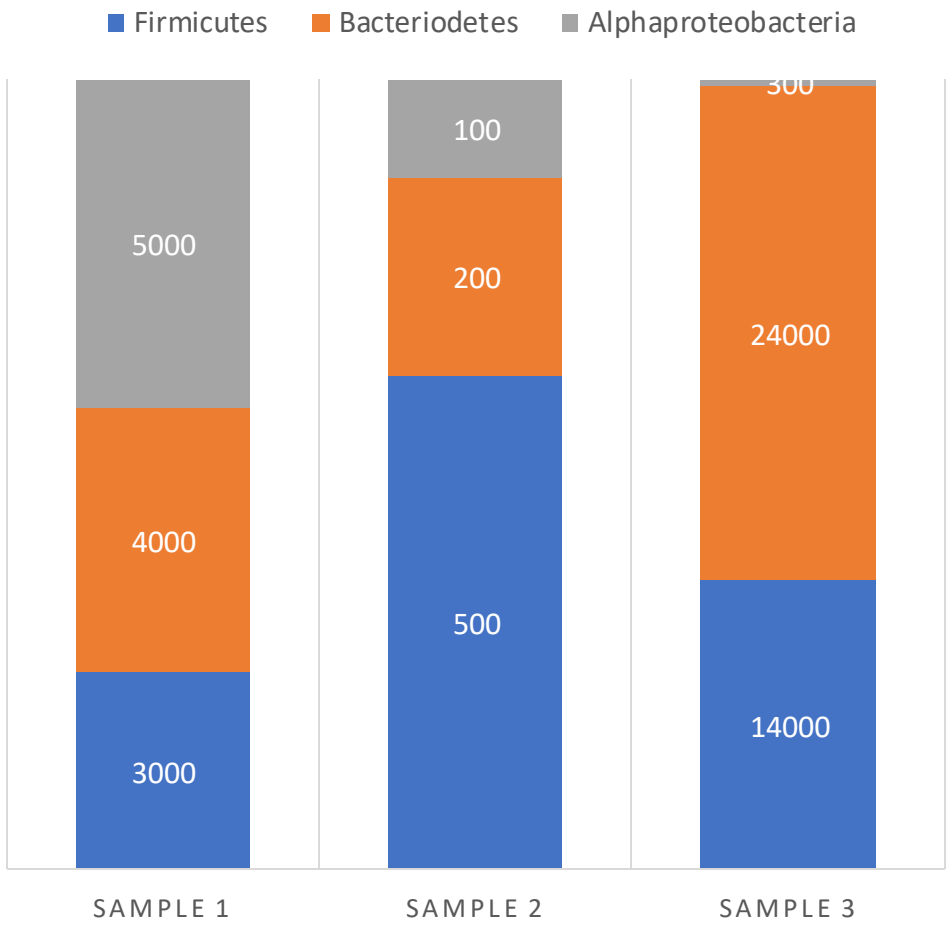
ECETOX

Statistical considerations

Compositional nature of the metagenomic data

- The microbiome abundancies (read counts) are **constrained by the maximum number of DNA reads** that the sequencer can provide (the total count constraint)

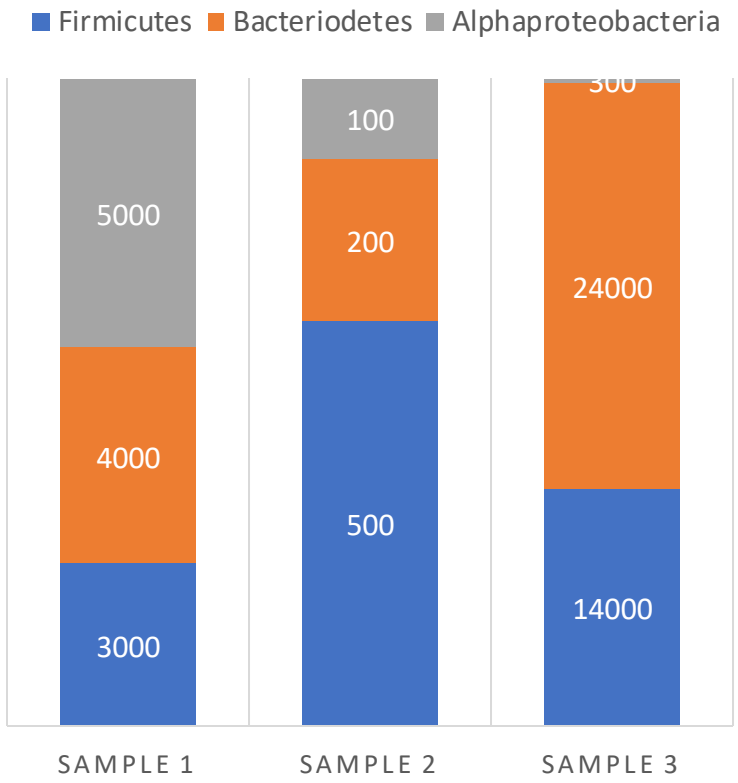
- Hence the data represents in fact a **proportion (composition) of taxa**!



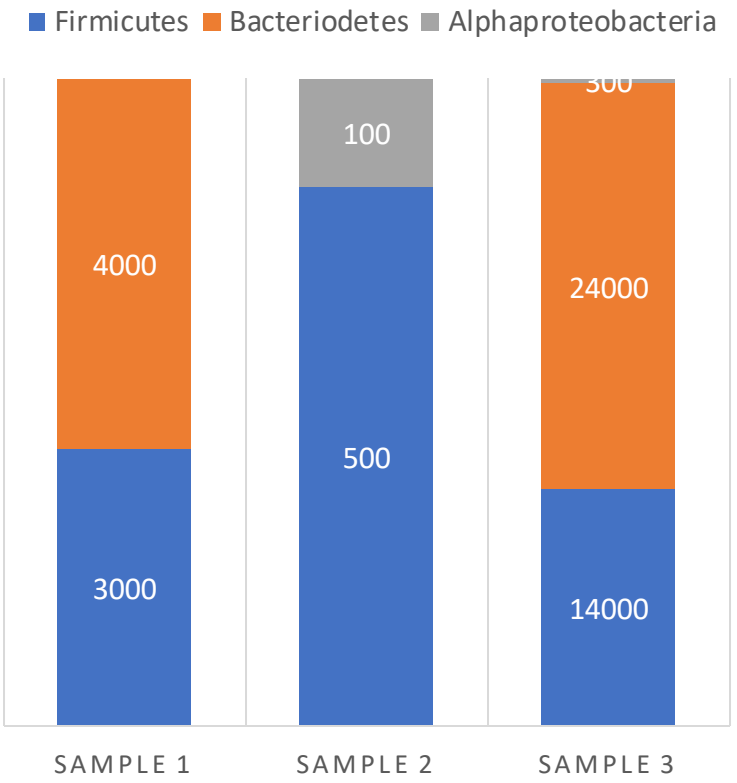
MUNI | RECETOX

The data is compositional – so what?

- The compositional nature of the data induces **strong dependencies** among the abundances of the different taxa:
  - an increase in the abundance of one taxon implies the decrease of the observed number of counts – hence proportions - for other taxa and vice versa



Before treatment



After treatment

MUNI | RECETOX

The data is compositional – so what?

In a composition the value of each component is not informative by itself and the **relevant information is contained in the ratios between the components**

The most known: **Firmicutes / Bacteroides ratio**

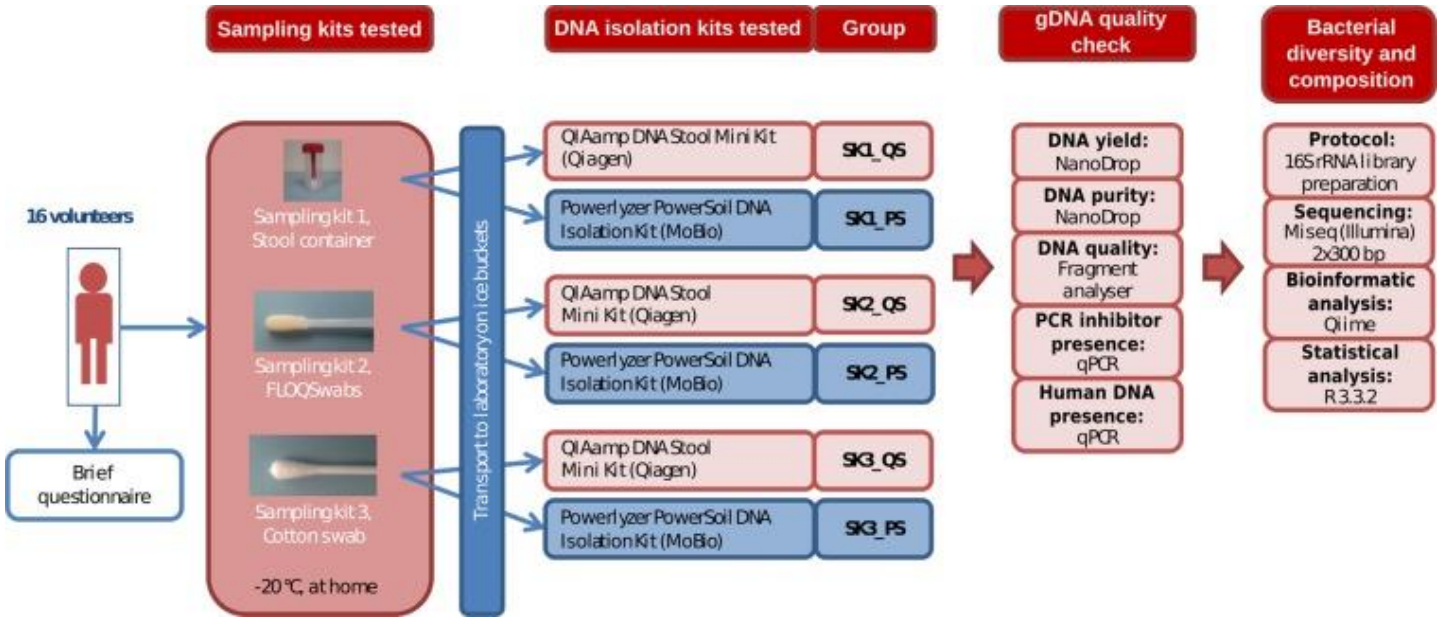
Firmicutes/Bacteroides ratio?
Effect of DNA isolation kit on Gram+ vs Gram- bacteria

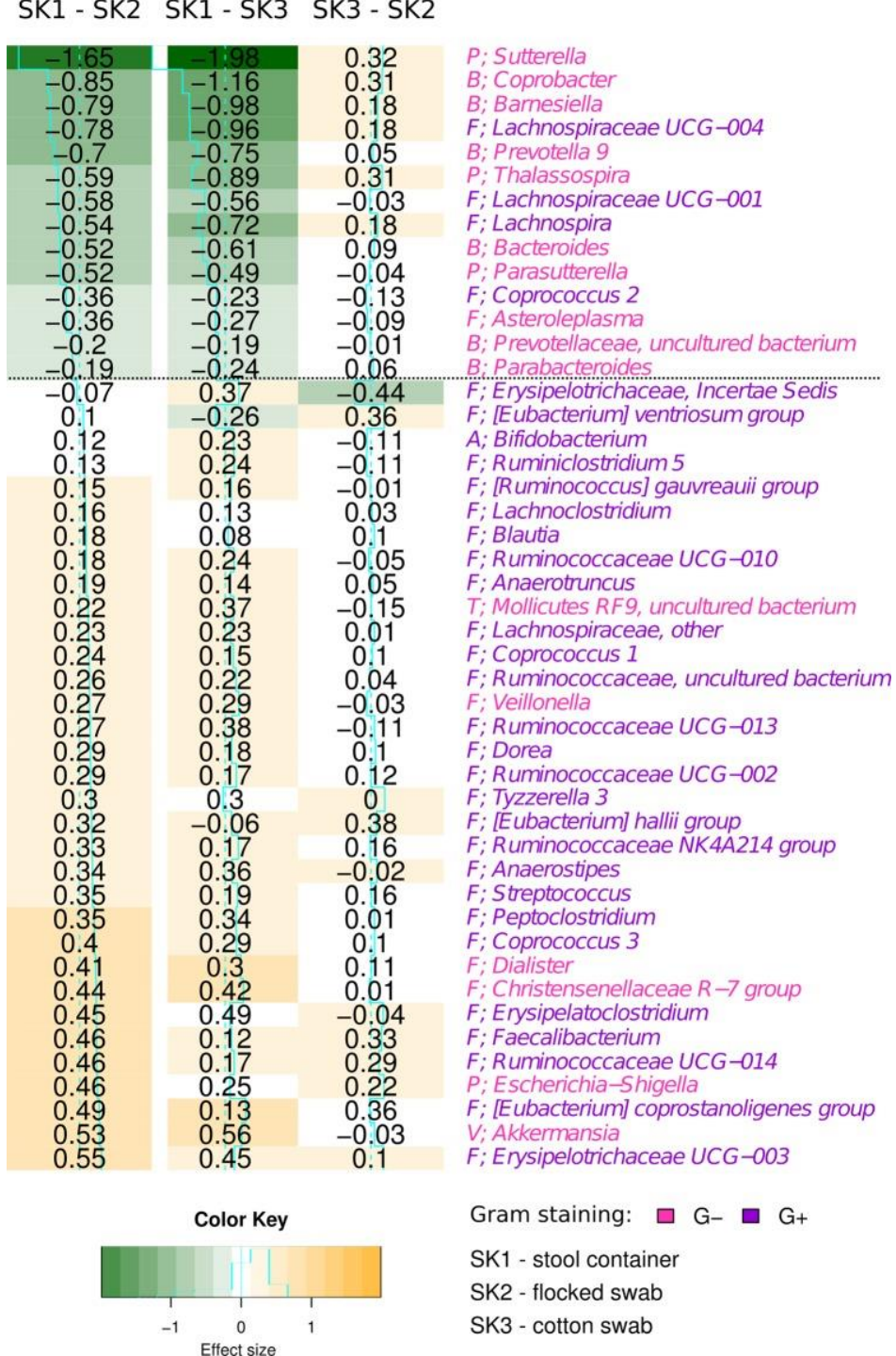
Taxa (show ten most abundant)	q-value		Sign of the estimated effect size of the isolation or sampling kit				Relative abundance: total sum %	Gram stain
	isolation kit effect	sampling kit effect	PS to QS	SK2 to SK1	SK3 to SK1	SK3 to SK2		
Firmicutes	**1.27E-15**	**4.43E-11**	+	−	−	+	68.3	G+
Bacteroidetes	**3.42E-02**	**1.81E-02**	−	+	+	+	18.5	G−
Actinobacteria	**3.67E-17**	**5.13E-04**	+	−	−	−	7.1	G+
Proteobacteria	**5.05E-08**	**3.47E-04**	−	+	+	+	1.1	G−
Verrucomicrobia	**1.69E-03**	**2.39E-03**	−	−	−	+	0.5	G−
Tenericutes	**2.08E-03**	4.05E-01	−	−	−	−	0.1	G−

Videnska et al. (2019) Stool sampling and DNA isolation kits affect DNA quality and bacterial composition following 16S rRNA gene sequencing using MiSeq Illumina platform, Scientific Reports

MUNI | RECETOX

Effect of sampling kit on G+ and G- bacteria

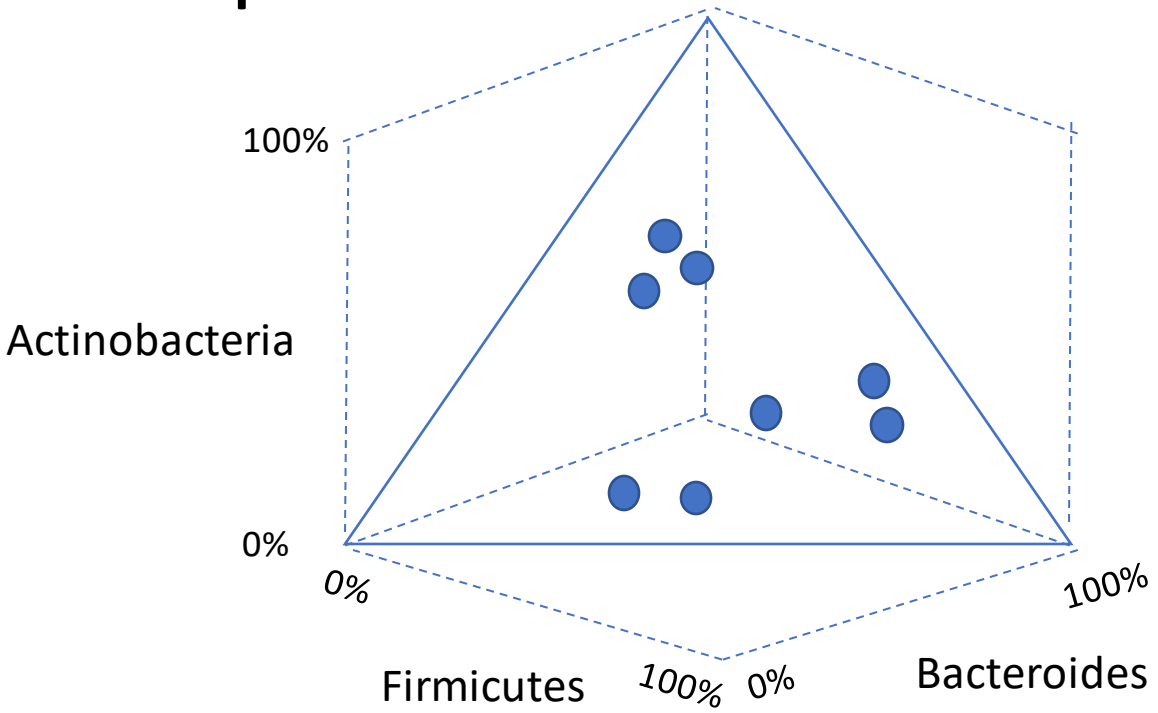




Videnska et al. (2019) Stool sampling and DNA isolation kits affect DNA quality and bacterial composition following 16S rRNA gene sequencing using MiSeq Illumina platform, Scientific Reports

The data is compositional – so what? / part 2

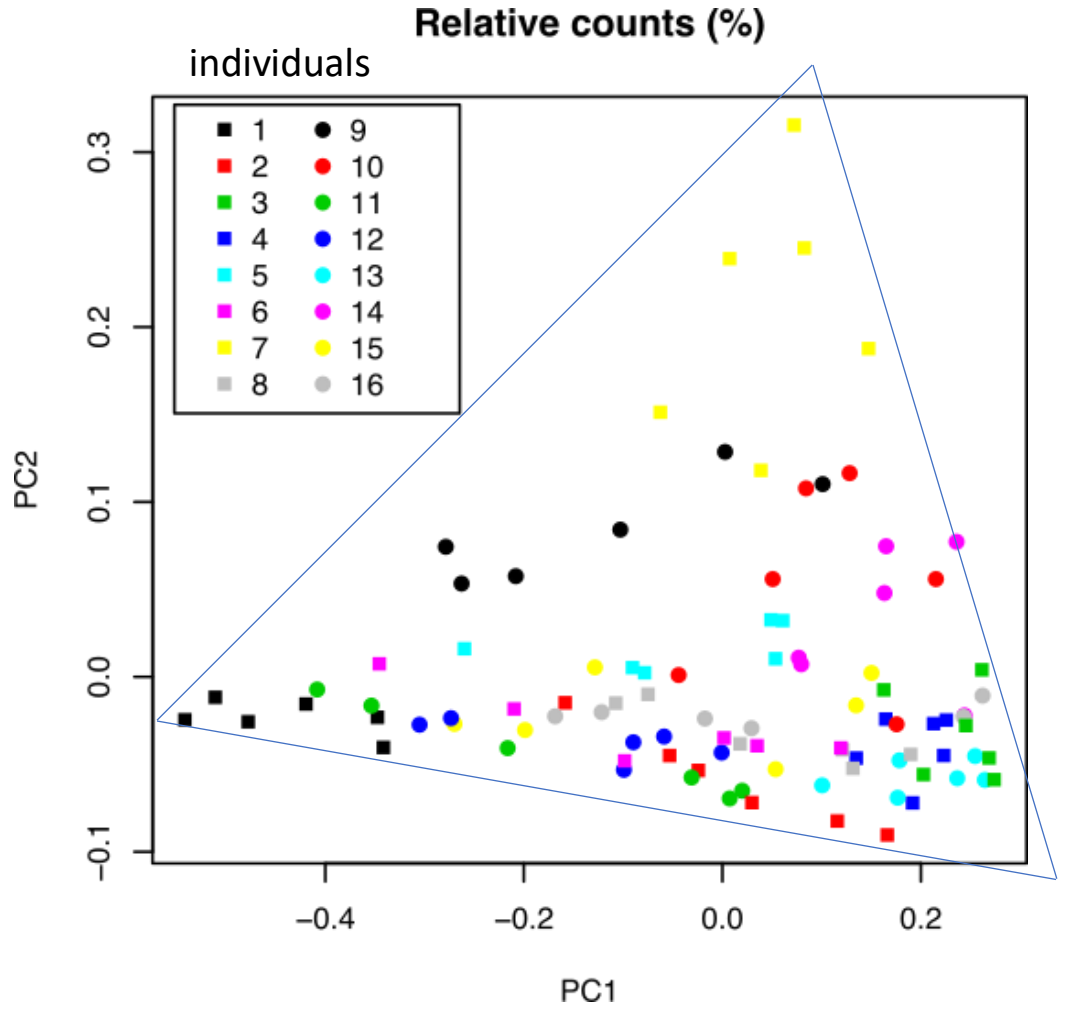
- Compositional data **do not exist in the Euclidean space**, but in a special constraint space called the **simplex**



- Hence it is incorrect to apply any multivariable techniques that are dependent on this space without proper data transformation (e.g. PCA, clustering….)

MUNI | RECETOX

PCA on compositional data (without proper transformation)



MUNI | RECETOX

Statistical methods for analysis of compositional data need to fulfill these criteria:

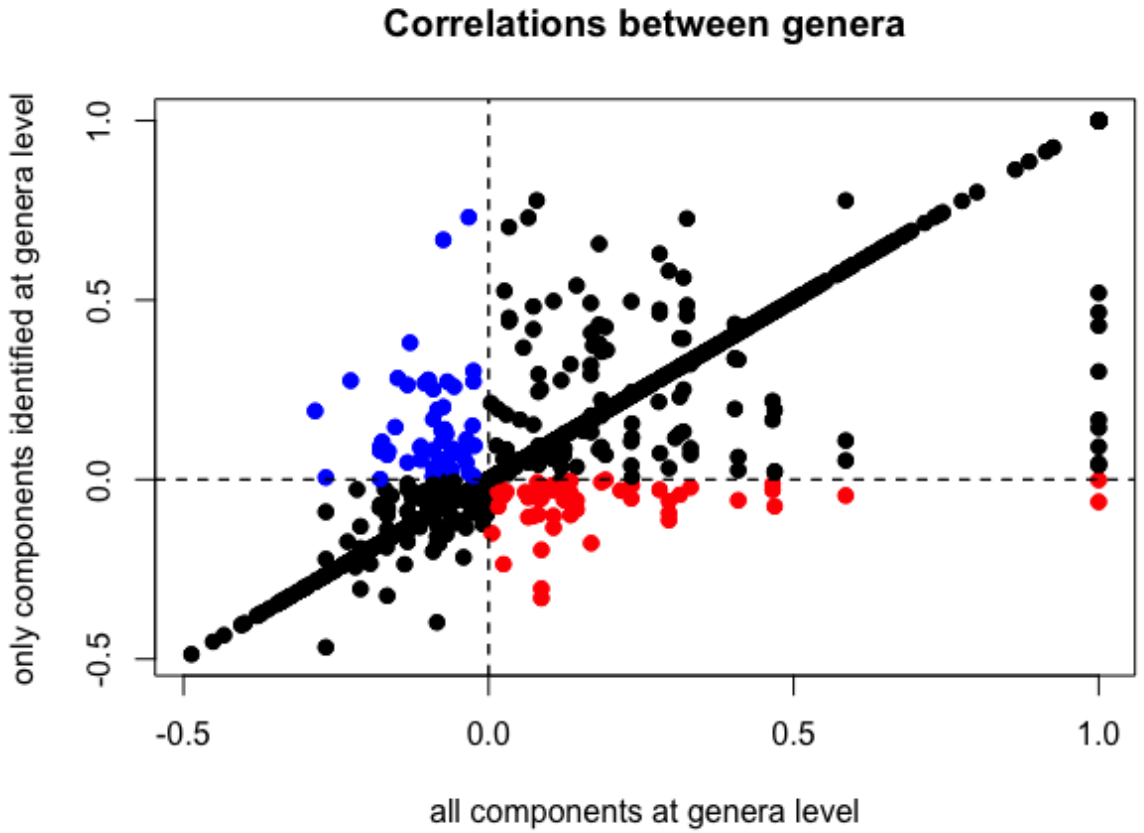
1. **Scale Invariance** (e.g. the result should be the same regardless of the scale of the measurement)
   - Example: how similar are these two samples?

	%		Absolute read counts	
	A	B	A	B
Fusobacteria	10	11	700	11000
Proteobacteria	15	14	1050	14000
Bacterioides	25	20	1750	20000
Firmicutes	50	55	3500	55000
Euclidean distance	7.2		57088.6	

Statistical methods for analysis of compositional data need to fulfill these criteria:

2. Subcompositional coherence (e.g. the analyses should lead to the same conclusions regardless of which components of the data are included)

This is especially a problem for correlations between taxa, which tend to be more negative when we remove some taxa and recalculate the proportions.



MUNI | RECETOX

Statistical methods for analysis of compositional data need to fulfill these criteria:

2. **Subcompositional coherence** (e.g. the analyses should lead to the same conclusions regardless of which components of the data are included)

Alternative(s) to correlation:

$$VLR\left(\boldsymbol{x}_g, \boldsymbol{x}_h\right) = var\left(\ln\frac{x_g^1}{x_h^1} + \ln\frac{x_g^2}{x_h^2} + \ldots + \ln\frac{x_g^n}{x_h^n}\right)$$

1. phi (Φ) = var(Ai -Aj)/var(Ai)
2. rho (ρ) = var(Ai -Aj)/(var(Ai) + var(Aj))
3. phis (Φs) = var(Ai -Aj)/var(Ai +Aj)

Aitchison, 1982, J.R.Statist. Soc.
Lovell et al., 2015, PLoS Comp Biol
Quinn et al, 2017, Scientific Reports 7

Ai is the log-transformed values for a metagenomic component 'i' in the data

MUNI | RECETOX

Data transformation (normalization)

- Compositional data can be normalized in order to make them suitable for existing statistical techniques
- Aitchinson, 1982 - build a theory and concepts of analysis of compositional data and suggested normalizations
- Basic concept – make log-ratios between components

ALR (additive log-ratio transformation)

$$\mathrm{alr}(x) = \left[\log \frac{x_1}{x_D} \cdots \log \frac{x_{D-1}}{x_D}\right]$$

+ good for most statistical techniques
– needs careful selection of one component, we are working with k-1 taxa, more difficult to interpret

CLR (centered log-ratio transformation)

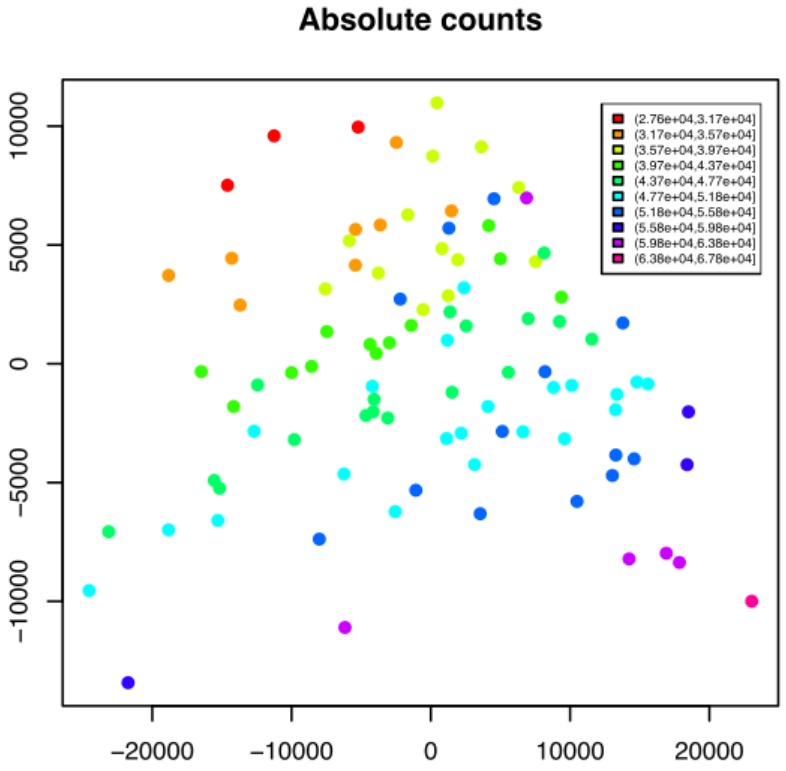
$$\mathrm{clr}(x) = \left[\log \frac{x_1}{g(x)} \cdots \log \frac{x_D}{g(x)}\right]$$

+ ratio to geometric mean, preserves all taxa, no need to select one
– creates singular covariance matrix

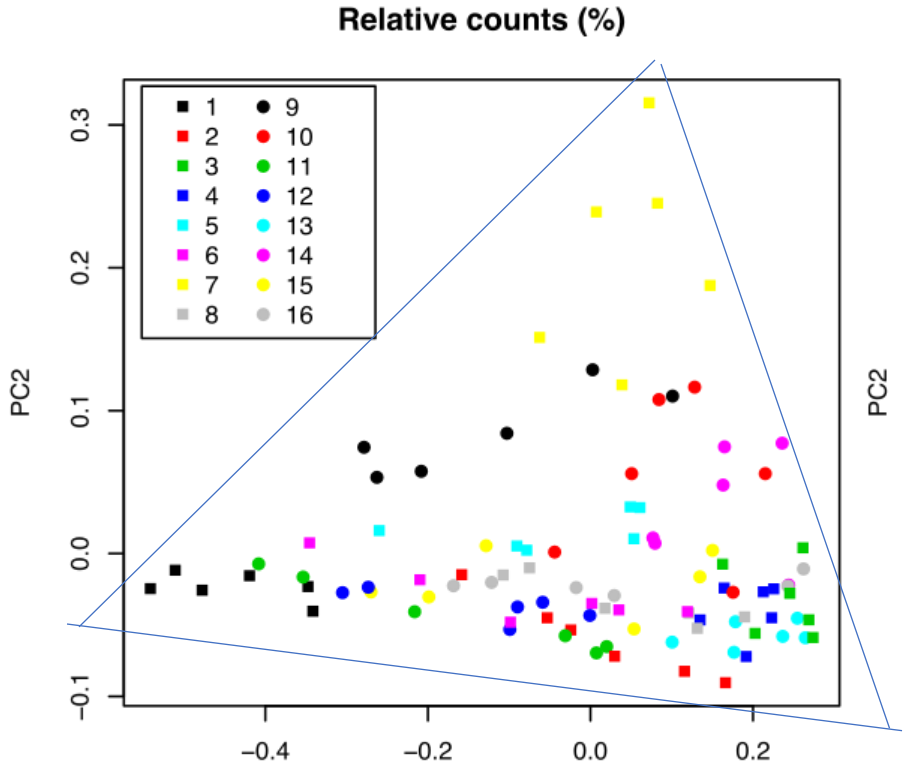
ILR (isometric log-ratio transformation) [Egozcque, 2003]

PhILR (phylogenetic partitioning based ILR transform) [Silverman et al, 2017]

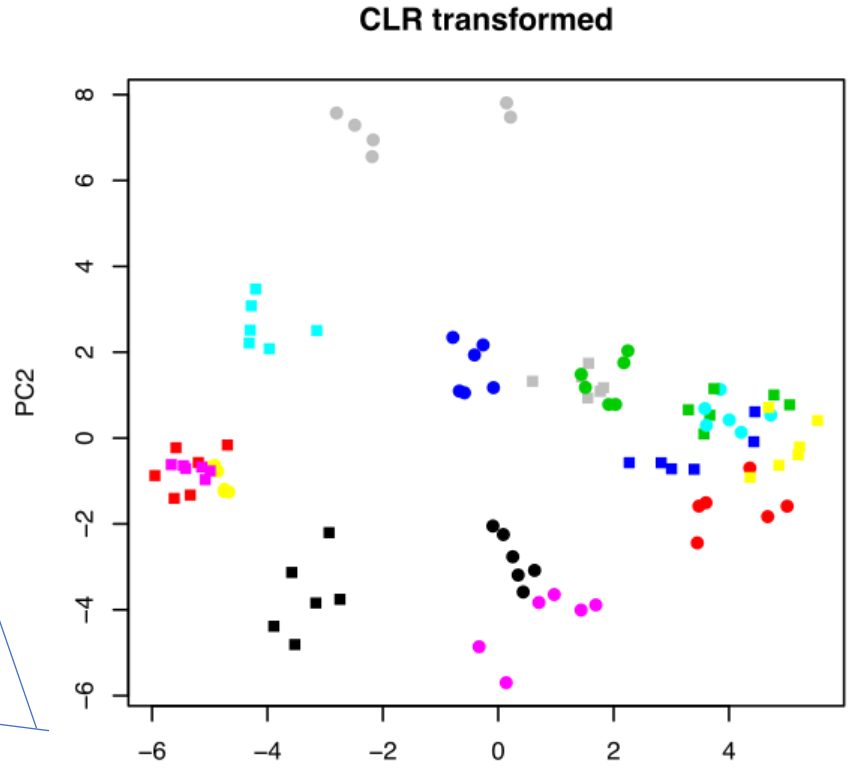
Compositional data - PCA before and after normalization



PCA on absolute counts – main variance lies in the sequencing depth



Relative data – problem with simplex, colour by individual



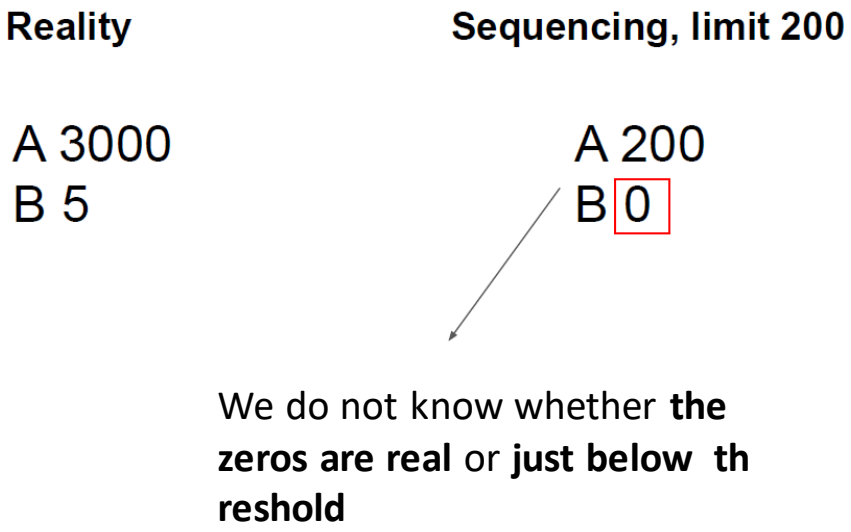
CLR transformed data – colour by individual

MUNI | RECETOX

The excess zero problem

- Log-ratio transformations require data with **positive values**, any statistical analysis of count compositions must be preceded by a proper **replacement of the zeros**

- What to do?
  - We need to fill in the zeroes....
  - E.g. Bayesian multiplicative treatment of count zeros [Martín-Fernandez,2014,Statistical Modelling]



The 16SrRNA gene copy number problem

- Different taxa have **different number of copies** of the 16S rRNA gene - this
- Some algorithms exist that try to normalize this count, but for many taxa this is unknown, hence estimation takes place.
- Should we normalize for count or not?

Take home messages

- Metagenomic data are compositional and it is not optional!
- Compositional data do not exist in Euclidean space
- Transformations of compositional data and specific statistical approaches are needed for data analysis
- It is not easy to make correlations between the taxa
- Normalization to 16SrRNA gene copy number change is necessary, but still quite tricky and the benefits are not clear

Resources to study

- OUT vs ASV explained VIDEO
- Dan Knight: microbiome discovery – series of videos

(however, uses QIIME and OTUs, which is outdated now, still very worth it!)

- PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples | Microbiome | Full Text (biomedcentral.com)

Metagenomic pipelines

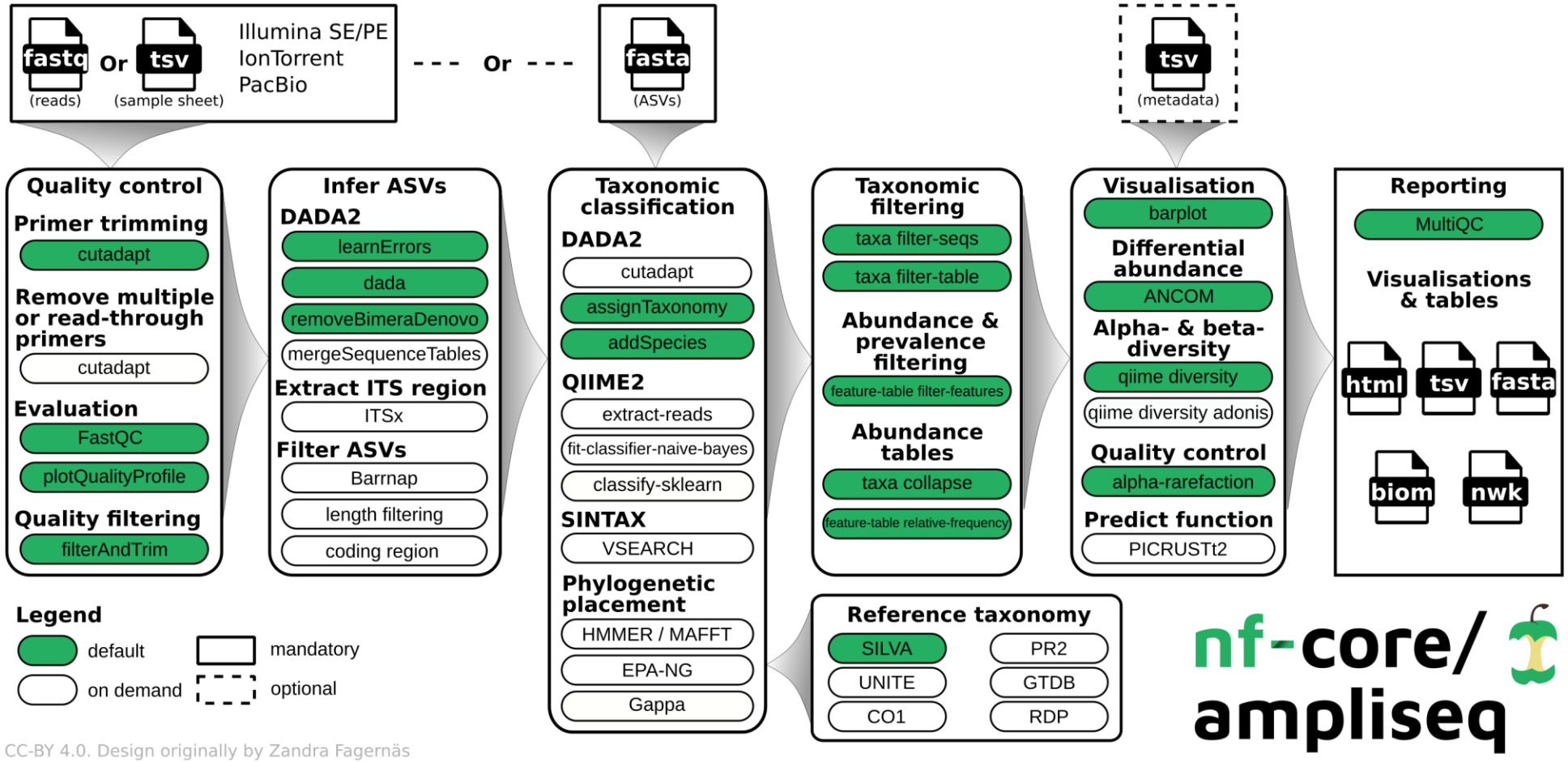
Pipelines

- DADA2
- QIIME2
- Mothur
- PathoScope 2.0
- Kraken

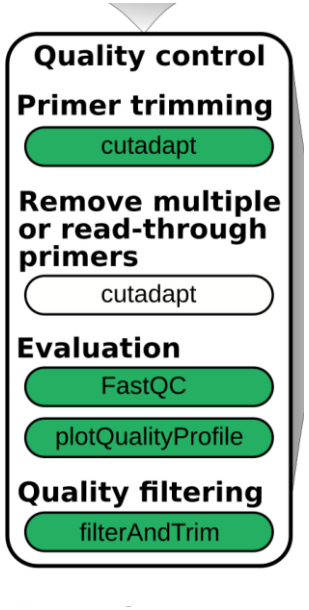
Databases

- SILVA 138
- GreenGenes 13_8
- RefSeq2020
- Kraken
- Blast nt/rna

RCX pipeline – nf-core/ampliseq

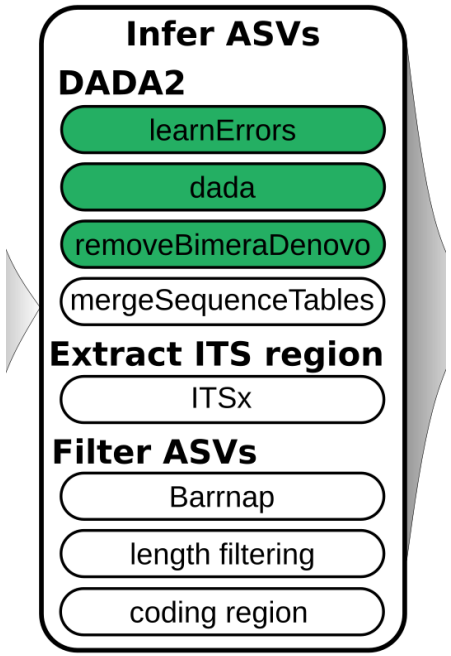


Quality control



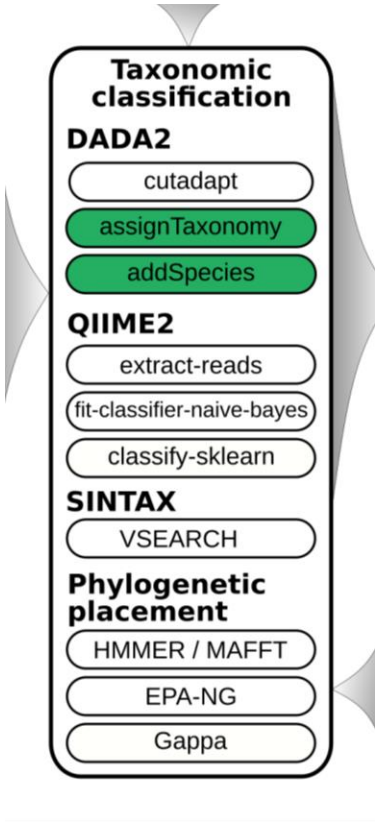
- Data preprocessing
- Check reads quality
- Perform filt&trim

ASV calculation



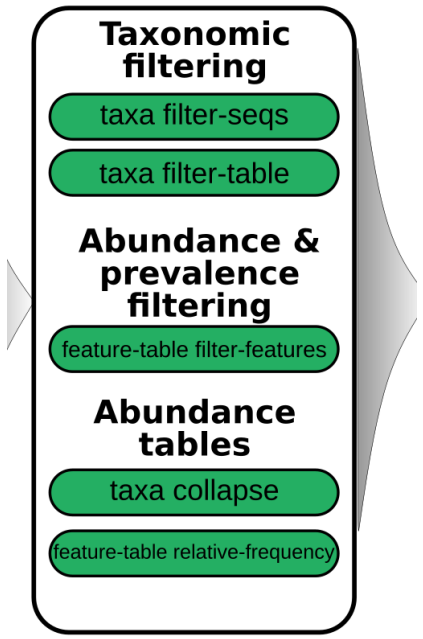
- DADA2
- Error estimation
- Chimera removal
- Contamination removal
- Filtering

Assign taxonomy



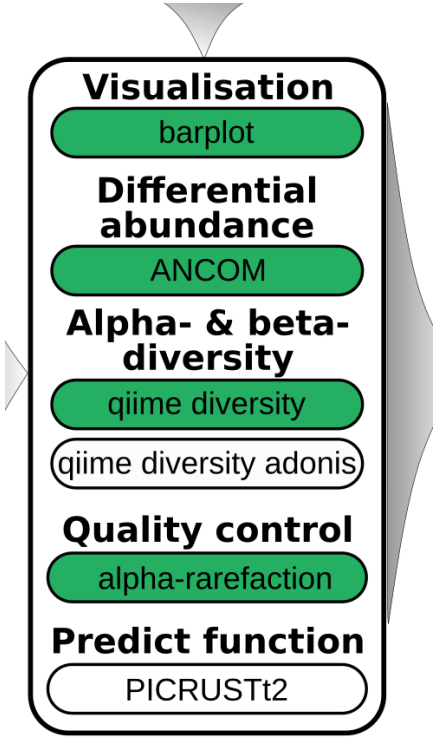
- Database dependent
- Infer species
- Confidence intervals
- Multiple assignment

Taxonomic filtering



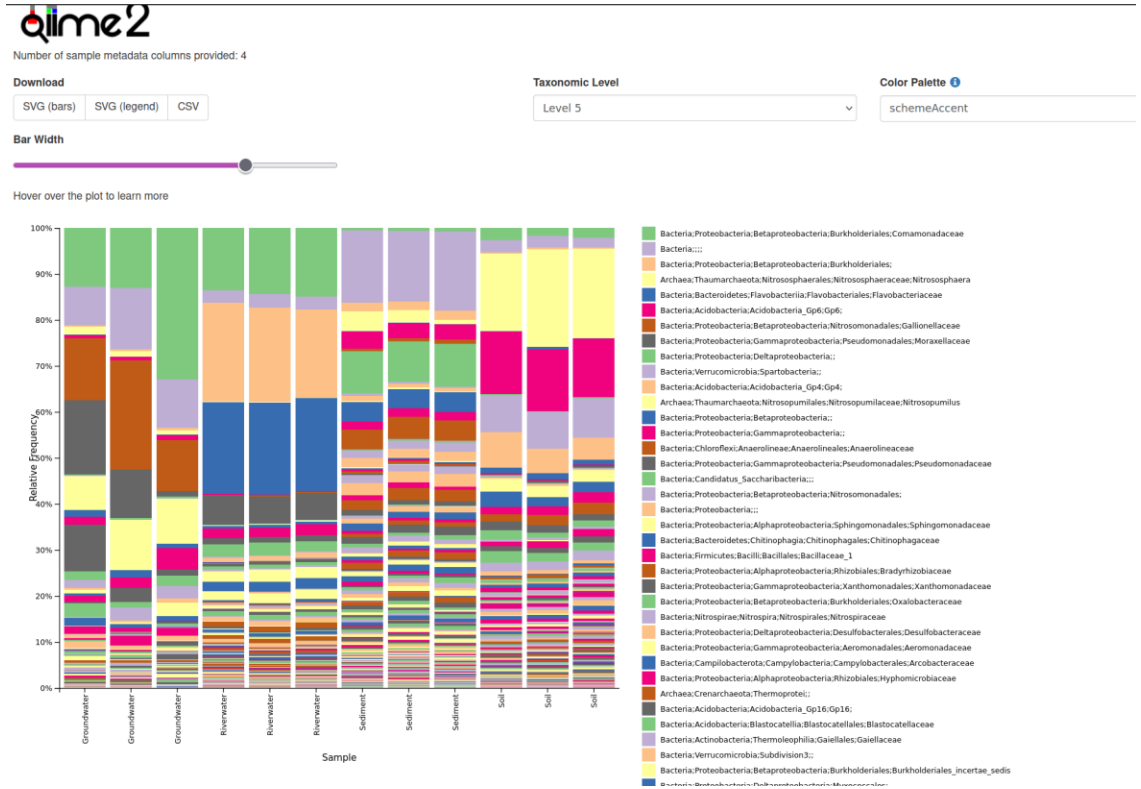
- Filter specific taxa
- Abundance filtering

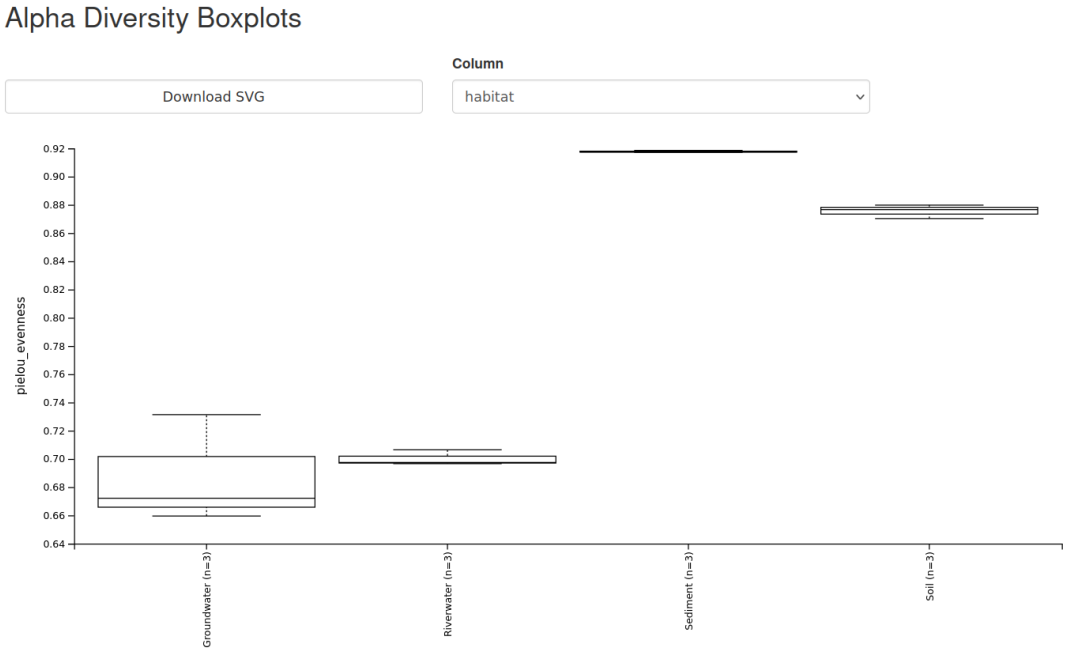
Post processing

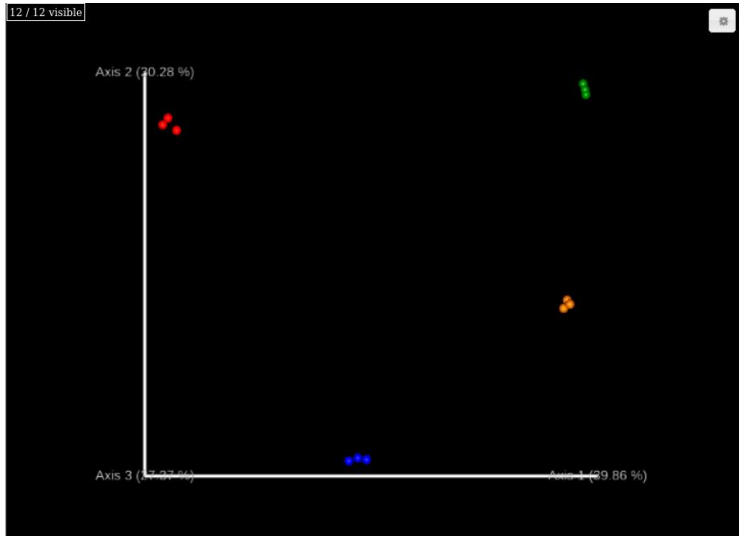


- Visualisation
- Diversity computation
- Functional analysis

Visualisations







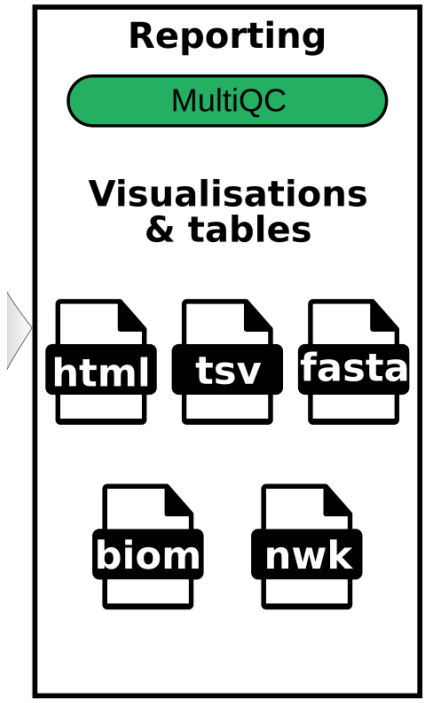
Picrust2

- *Phylogenetic Investigation of Communities by Reconstruction of Unobserved States*
- Functional analysis
- Gene families

- KEGG and COG database
- Based on phylogeny

- Genes present in microbial genomes are similar amongst relatives
- When sufficient genome sequences are available, it is possible to predict which gene families are present in a given microbial OTU from phylogeny alone.

Reporting



- Output files

- Additional analysis
- In-house post-processing