Introduction to 16S rRNA amplicon sequencing

Unraveling microbial diversity

Anna Carratalà, PhD
anna.carratala@epfl.ch
Environmental Chemistry Laboratory (LCE, ENAC)

Research questions

What bacteria are found in my samples?

What do they do?

What is a healthy community? And a healthy ecosystem?

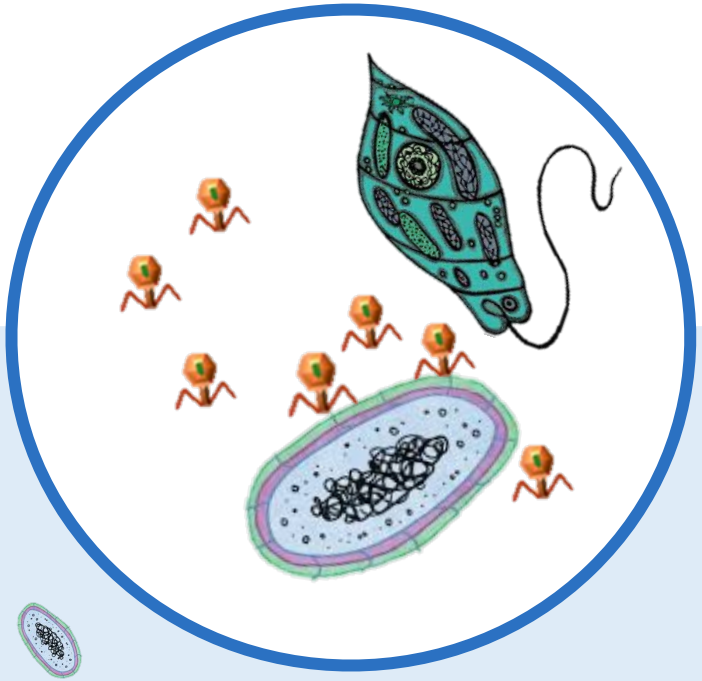
How do they affect ecosystem functioning?

What are their dynamics in space and time?

How to they respond to environmental changes?

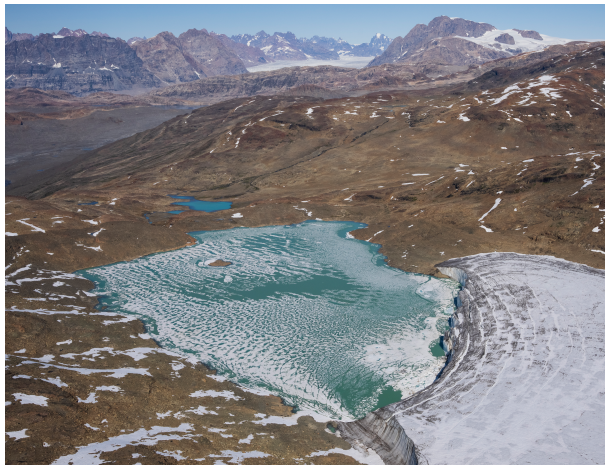
Sample pre-processing steps

10^9-10^{10} viruses/liter
10^7-10^9 bacteria/liter
10^2-10^4 protists/liter

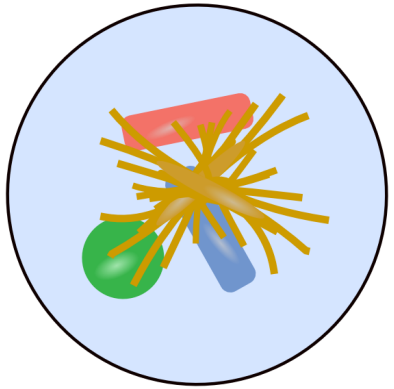


- Filtration (0.8, 0.22 µm)
- Fixation
- Culturing, enrichments…
- Measuring environmental variables or other metadata
- **Appropriate sample conservation**

Sample pre-processing steps



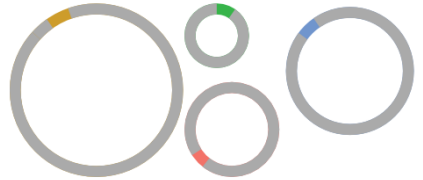
Mixed microbial community



DNA Extraction

- **In-house protocols**
- **Commercial kits**
- **Extraction controls to seq**

Amplicon sequencing



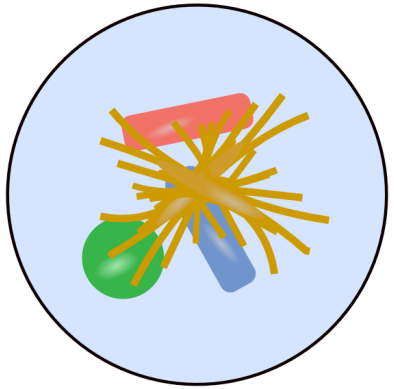
Multiple copies of fragments from 1 target gene

Metagenomics sequencing

Short sequence fragments from "all" DNA

Sample pre-processing steps

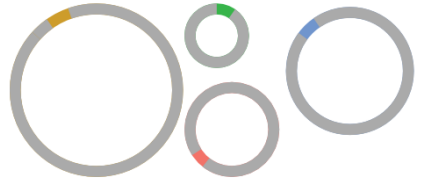
Mixed microbial community



DNA Extraction

- In-house protocols
- Commercial kits
- Extraction controls to seq

Amplicon sequencing



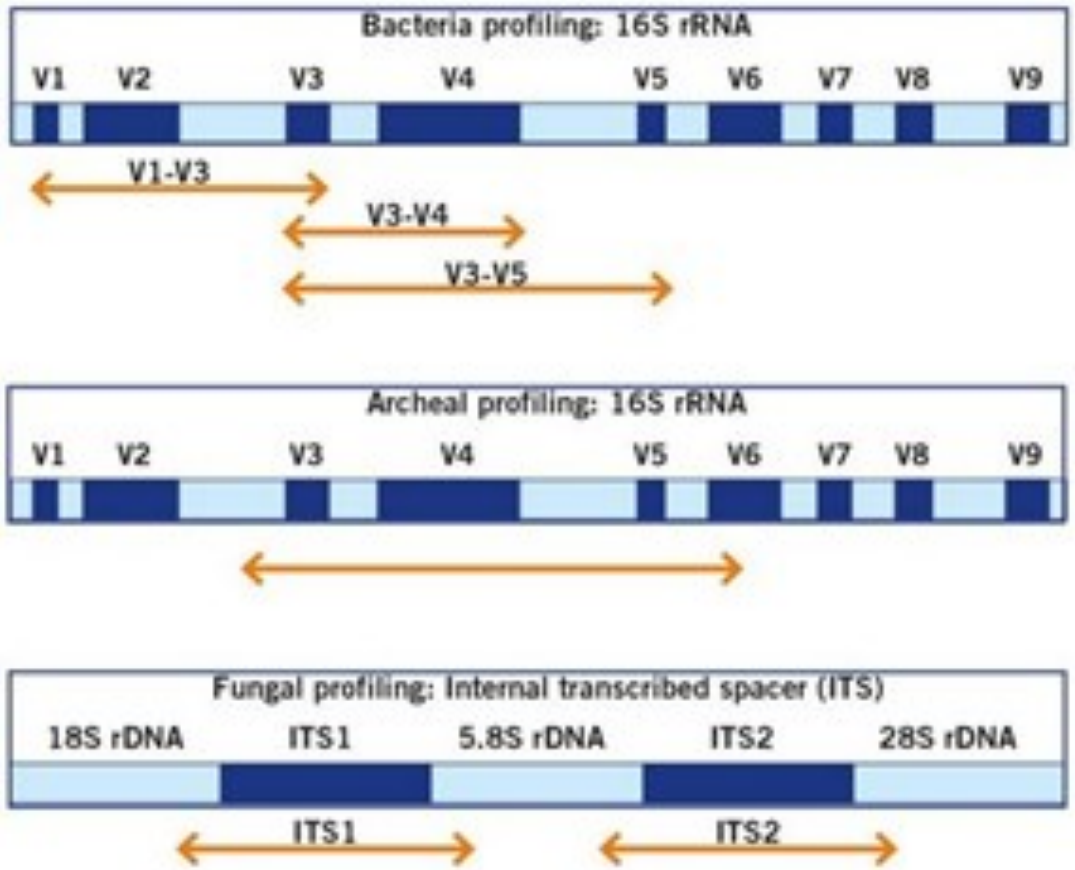
Multiple copies of fragments from 1 target gene

Metagenomics sequencing

Short sequence fragments from "all" DNA

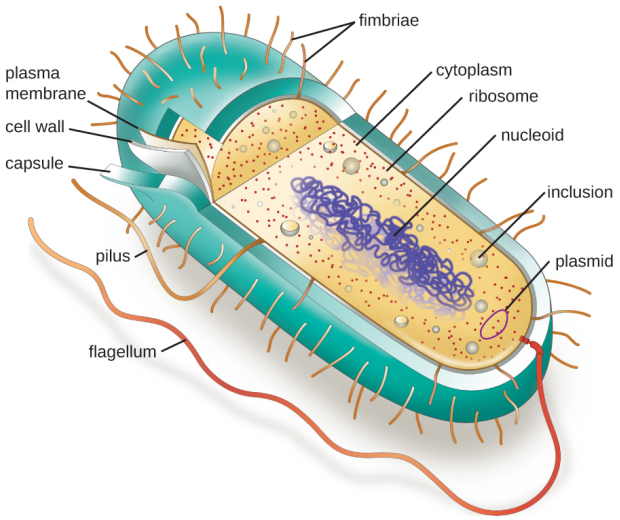
Amplicon sequencing:

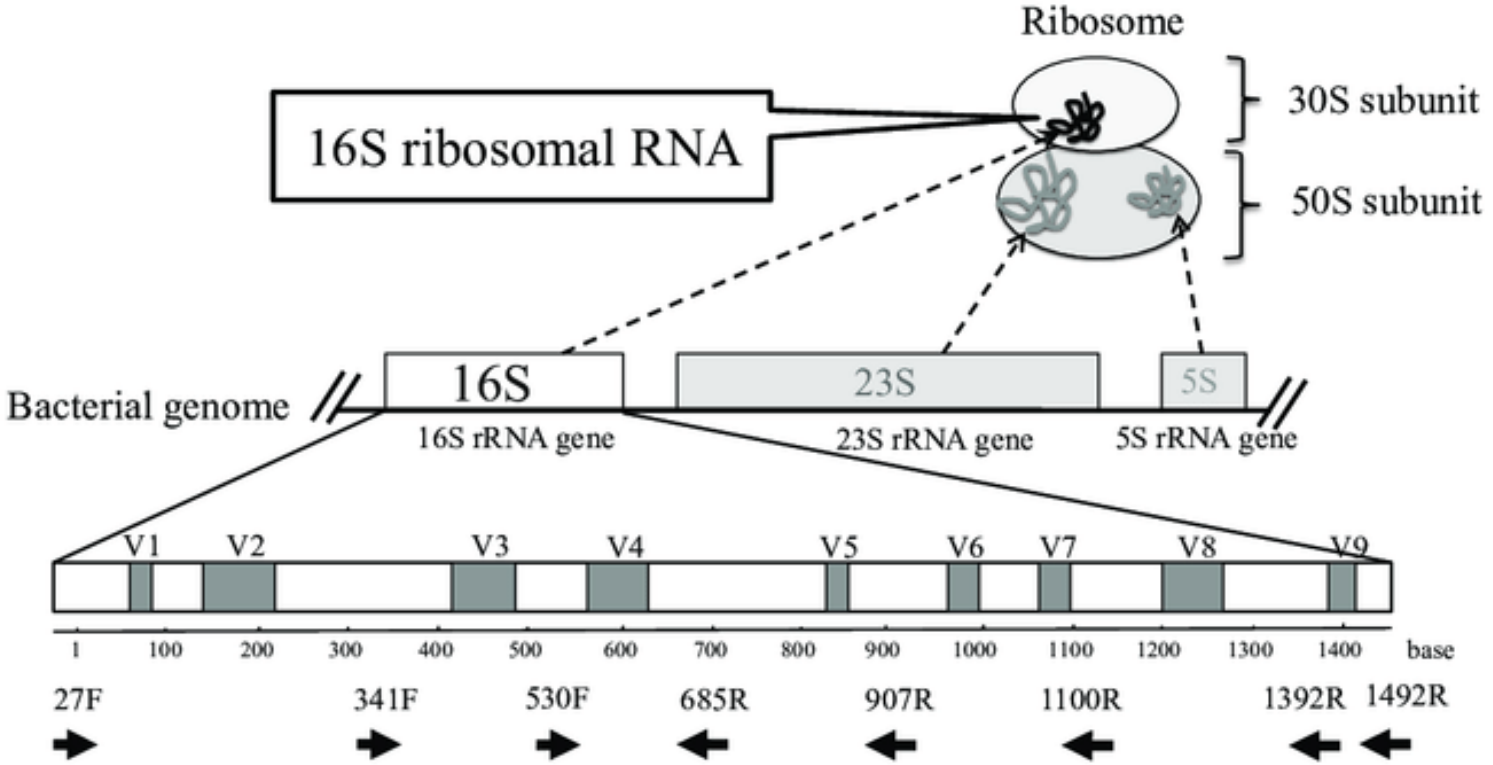
- Sequence a well-conserved gene
- Based on "universal" (-ish) primers for **PCR**
- Gene sequence is a barcode for species
- Usually, ribosomal RNA
- Bacteria/Archaea: 16S rRNA
- Fungi: Internal Transcribed Spacer (ITS)
- Other genes occasionally used, e.g. COI



What is 16S rRNA?

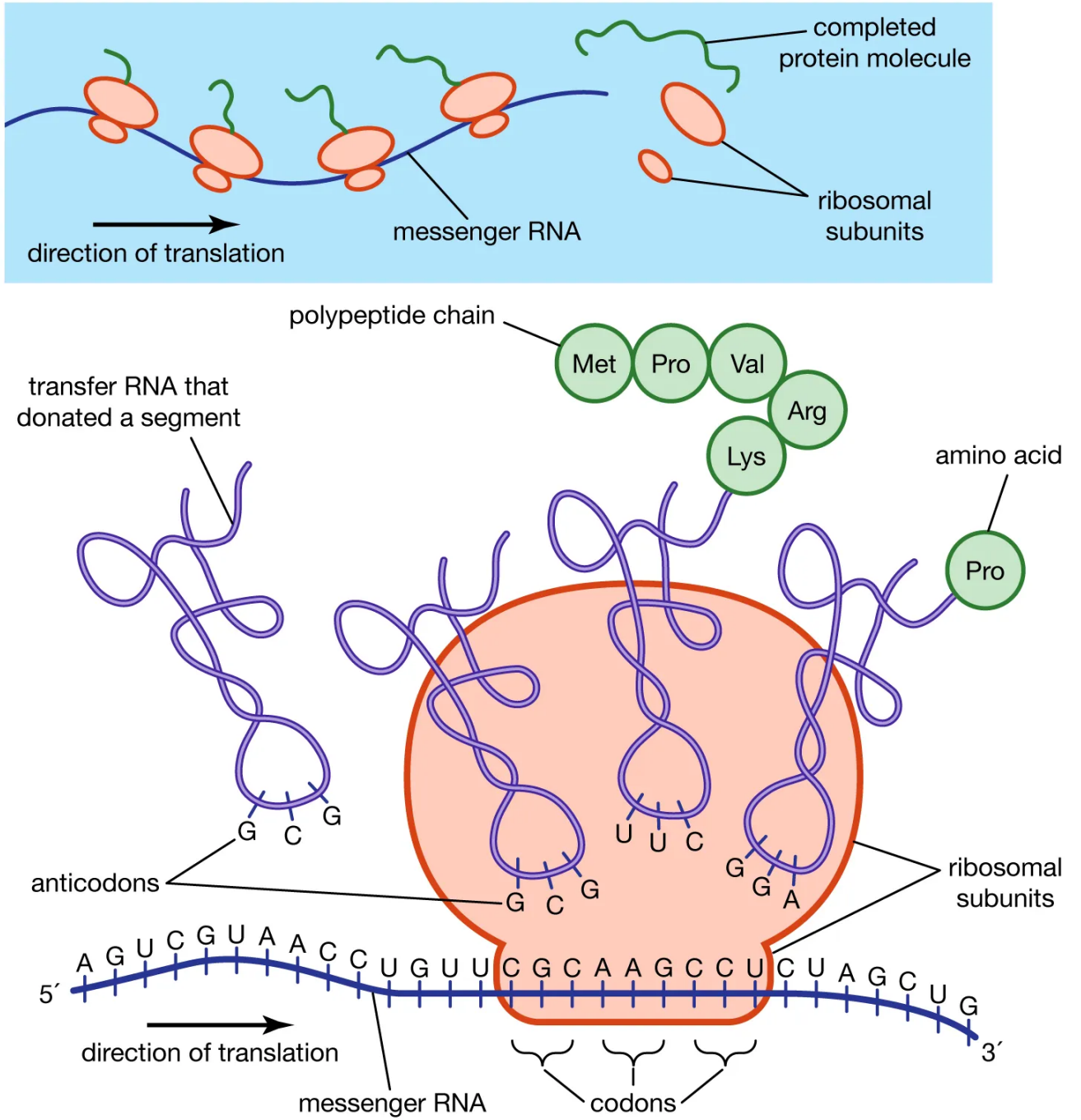
- 16S rRNA is a type of ribosomal RNA, a key structural and functional component of ribosomes—the cellular machinery responsible for protein synthesis.
- Found in the small subunit of prokaryotic ribosomes, 16S rRNA is highly conserved across different bacterial and archaeal species.





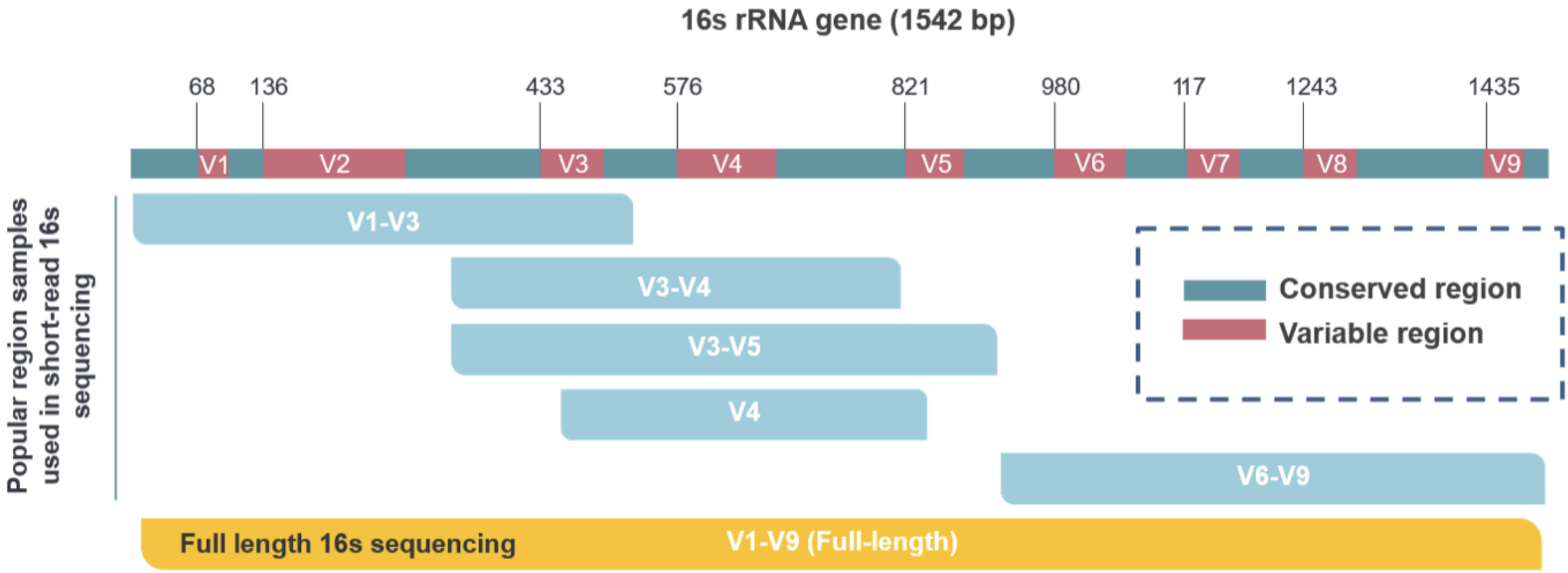
Functional significance

- The primary function of 16S rRNA is to help align and stabilize the ribosomal components during protein synthesis, facilitating the decoding of mRNA (messenger RNA) and the binding of tRNA (transfer RNA) molecules.



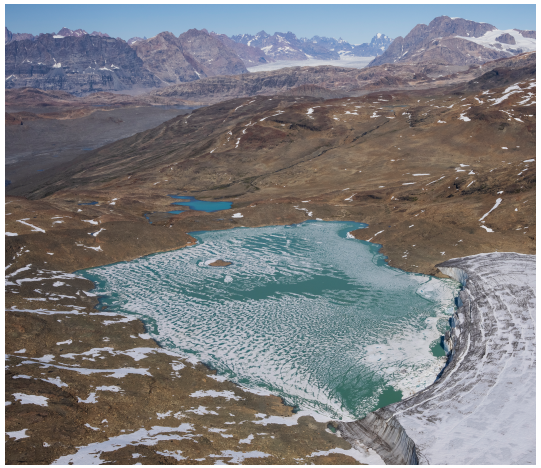
© Encyclopædia Britannica, Inc.

16s rRNA gene regions

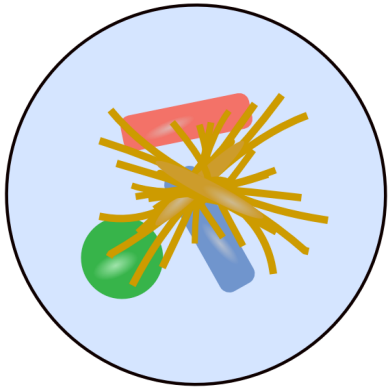


Workflow of 16S rRNA Sequencing

- **Sample collection**
- **DNA extraction**
- **16s rRNA PCR amplification**
- Library preparation
- Sequencing
- Data analysis

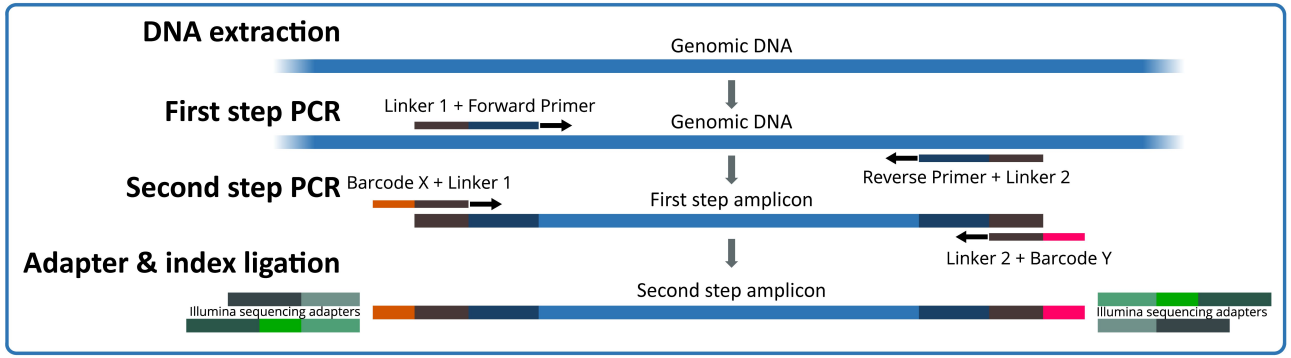


Mixed microbial community



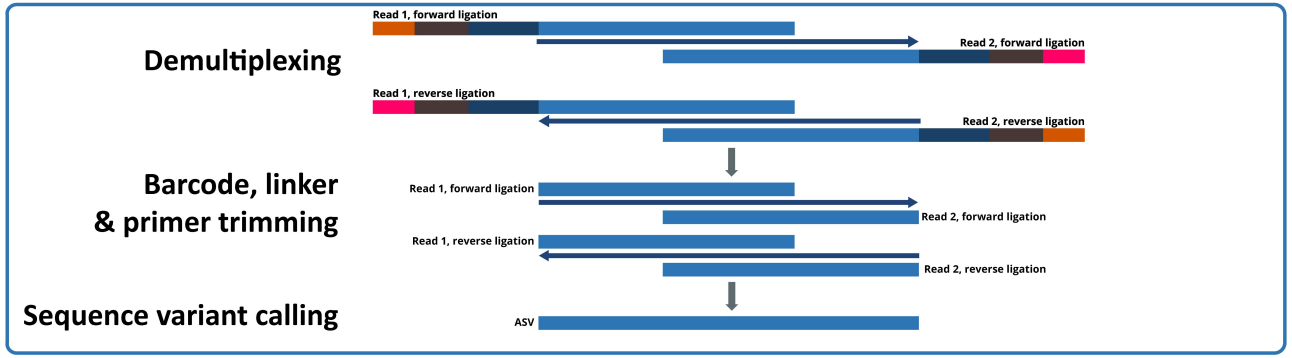
DNA Extraction

1. Amplicon generation

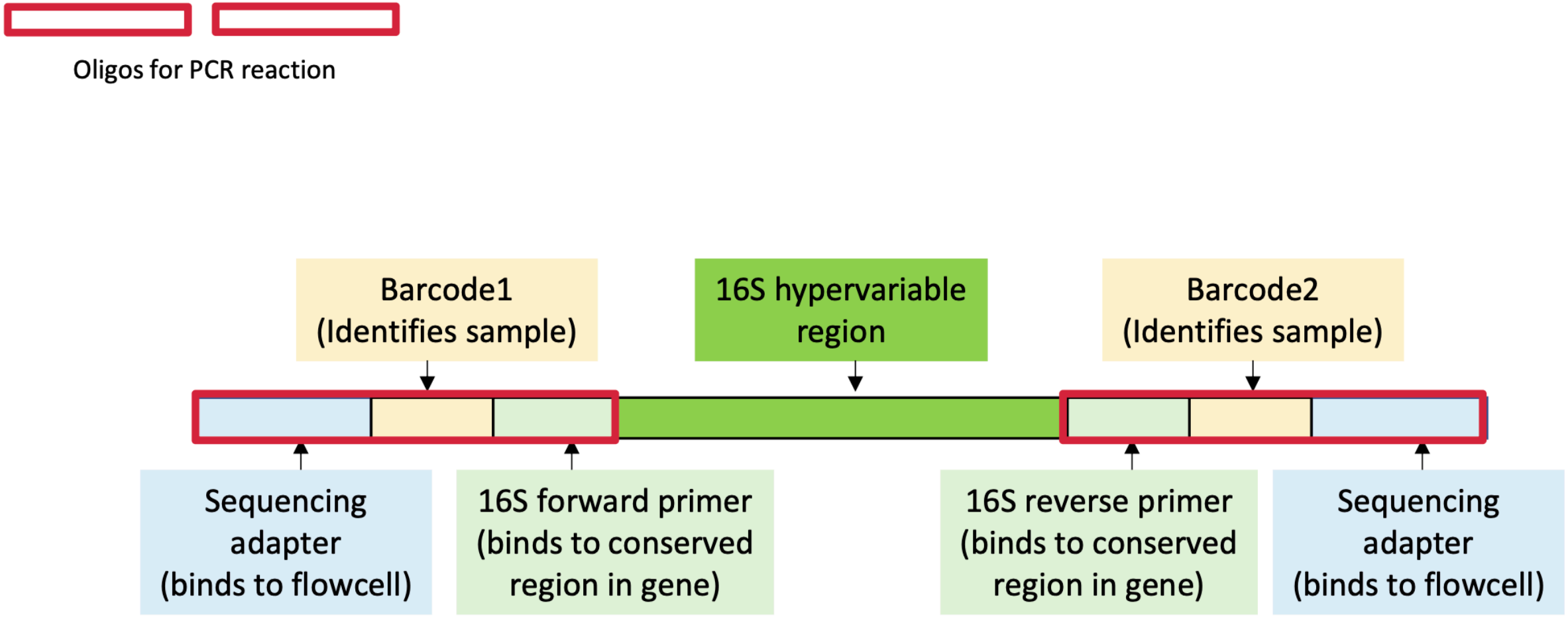


2. Sequencing

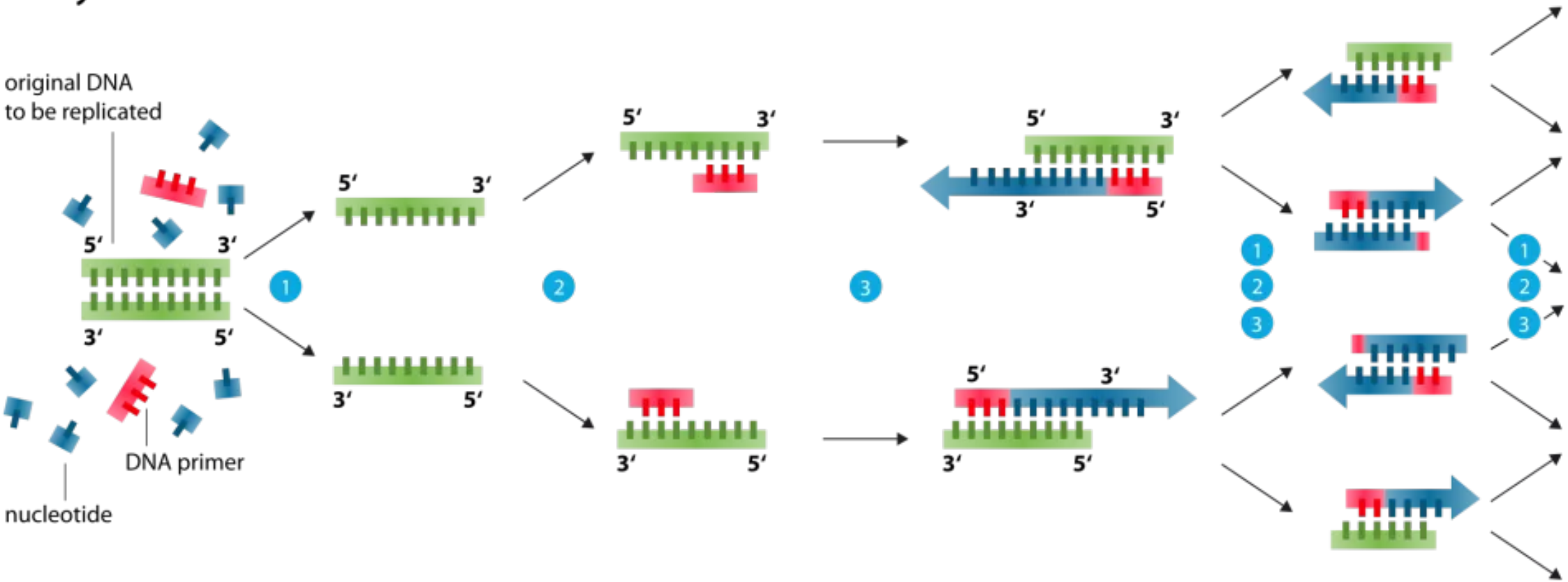
3. Amplicon sequence analysis



16s rRNA PCR: amplicon generation



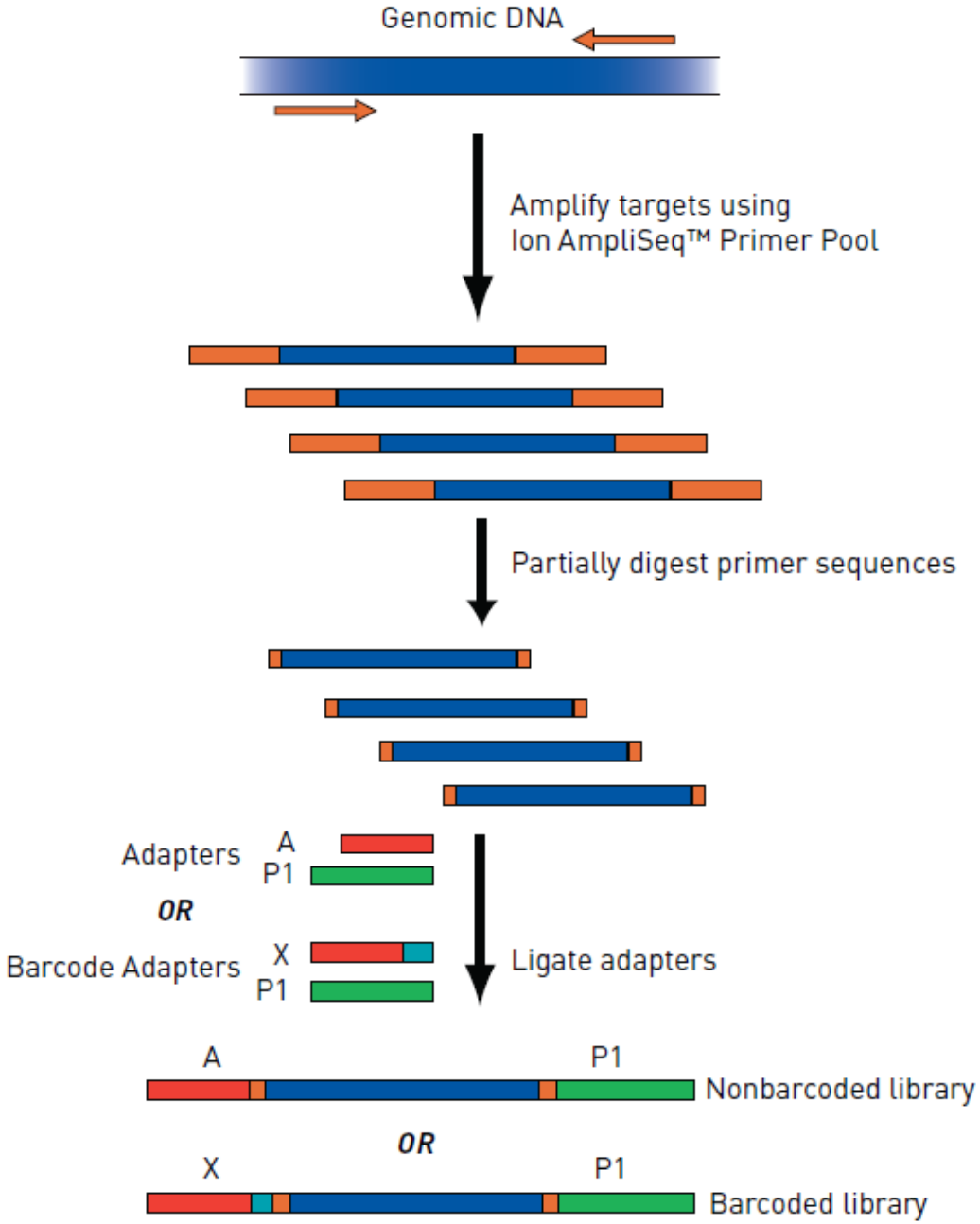
Polymerase chain reaction - PCR



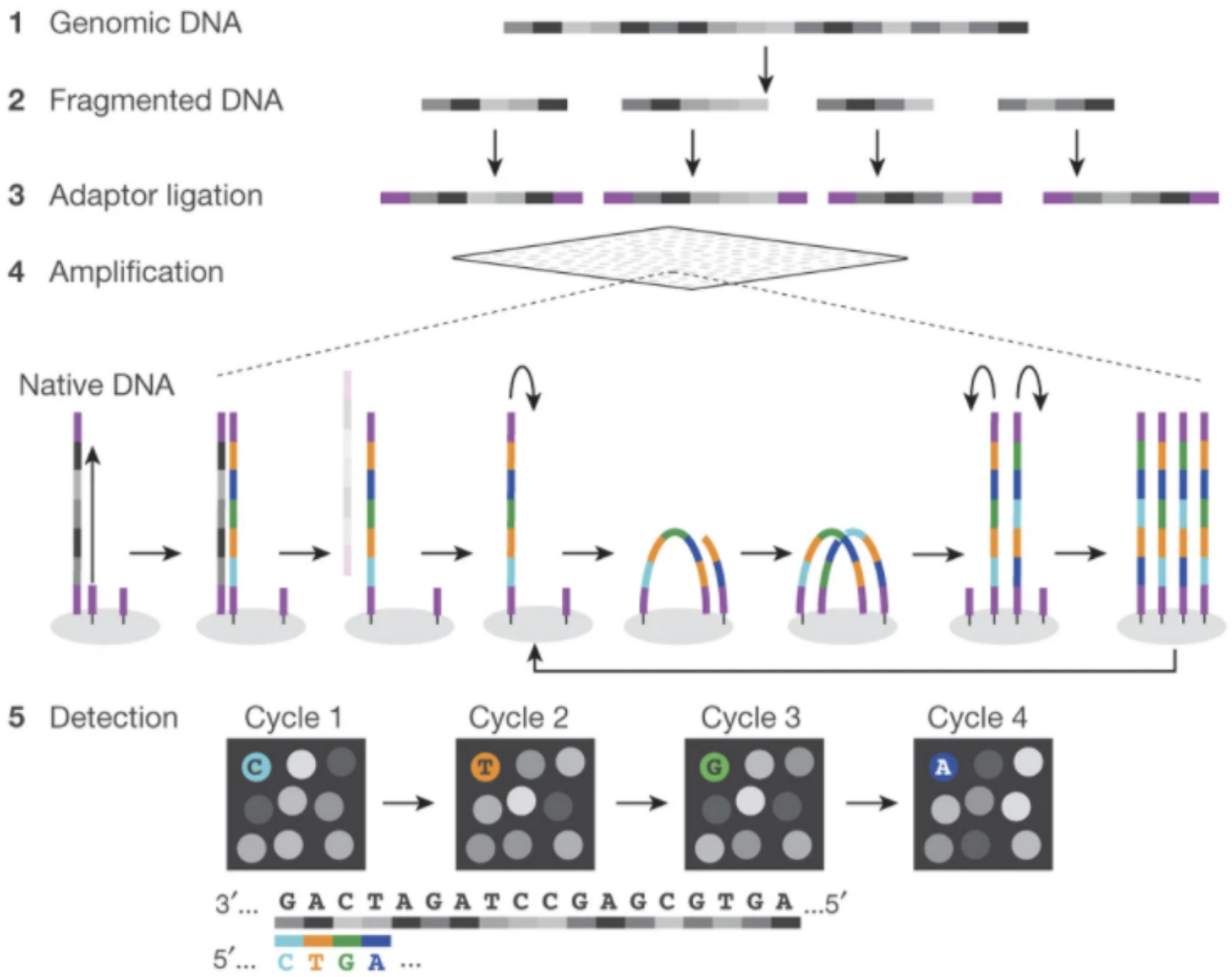
1. **Denaturation** at 94-96°C
2. **Annealing** at ~68°C
3. **Elongation** at ca. 72 °C

Library preparation

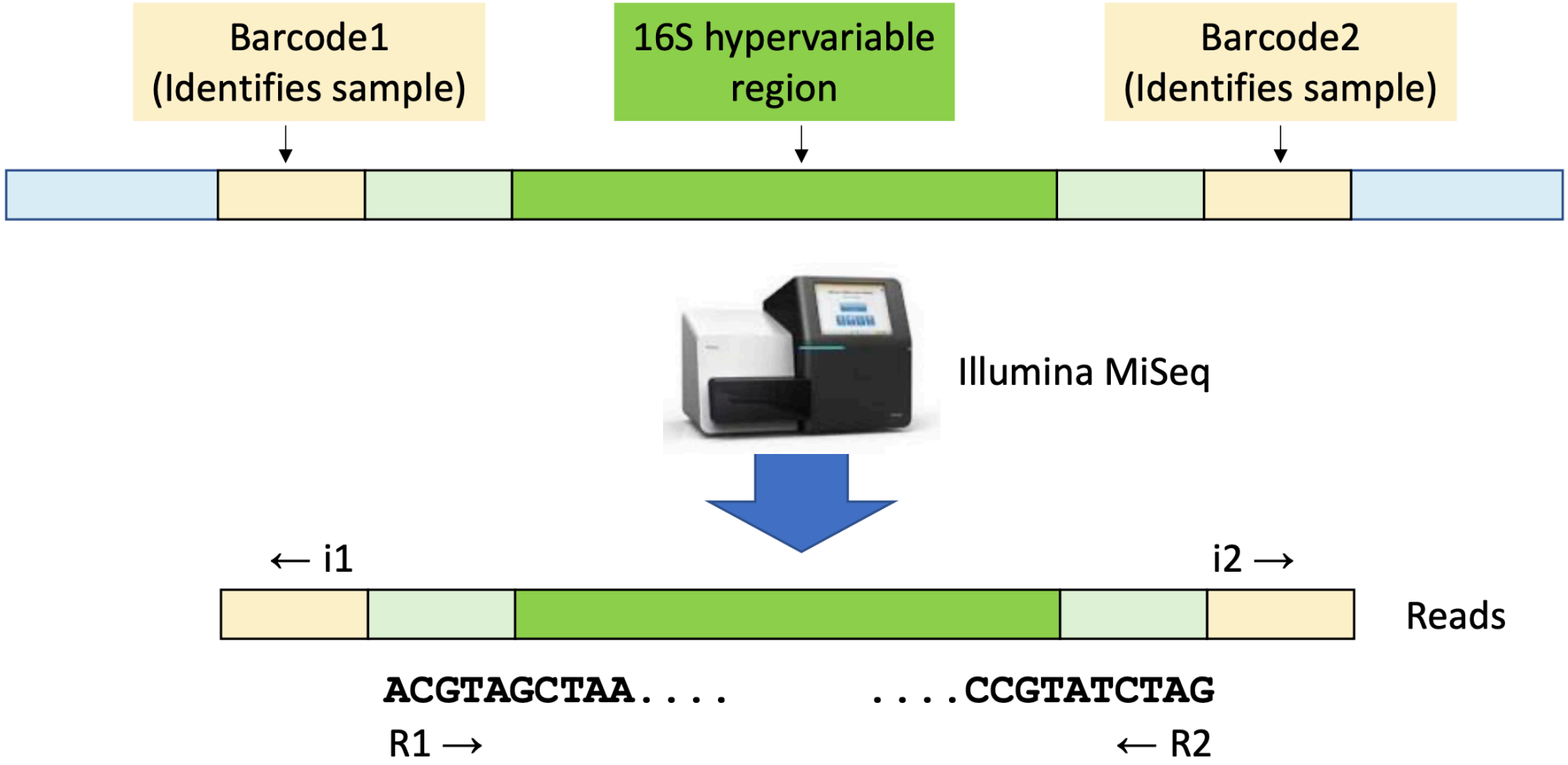
•Quantify the concentration of each library using quantitative PCR (qPCR) or another suitable method.
•Normalize the concentration of each library to ensure equal representation in the sequencing pool.



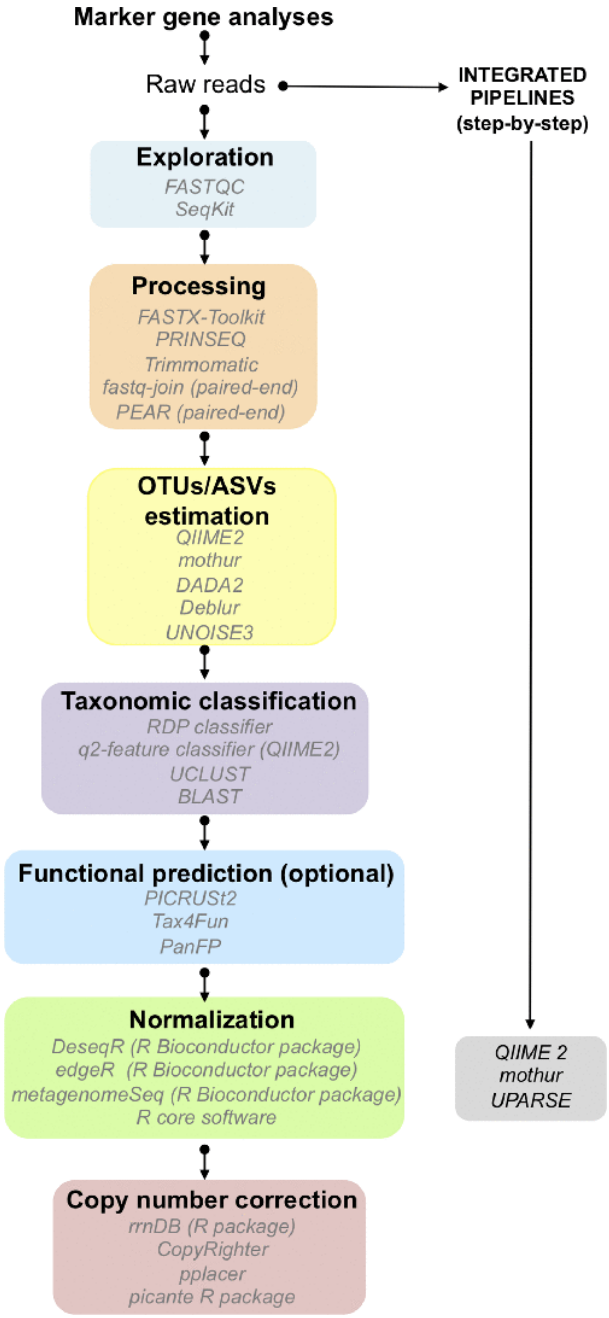
Sequencing


1 Genomic DNA
2 Fragmented DNA
3 Adaptor ligation
4 Amplification
Native DNA
5 Detection
Cycle 1
Cycle 2
Cycle 3
Cycle 4
C
T
G
A
3′... G A C T A G A T C C G A G C G T G A ...5′
5′... C T G A ...

Raw reads


Barcode1
(Identifies sample)
16S hypervariable
region
Barcode2
(Identifies sample)
Illumina MiSeq
← i1
i2 →
Reads
ACGTAGCTAA. . . .
. . . .CCGTATCTAG
R1 →
← R2

Amplicon sequencing analysis



FastQC: evaluating quality of reads

- Standard text-based format used to store both nucleotide sequence and corresponding quality score information generated from high-throughput sequencing experiments.
- FASTQ files are plain text files containing lines of sequence data and quality scores.
- A single entry in a FASTQ file consists of four lines:
  - **Sequence Identifier Line (Header):** Begins with '@' followed by a unique identifier for the sequence.
  - **Sequence Line:** Contains the nucleotide sequence of the DNA or RNA.
  - **Quality Identifier Line:** Begins with '+' and usually has the same identifier as the sequence identifier line.
  - **Quality Scores Line:** Contains ASCII-encoded quality scores corresponding to each base in the sequence.

Quality Scores:

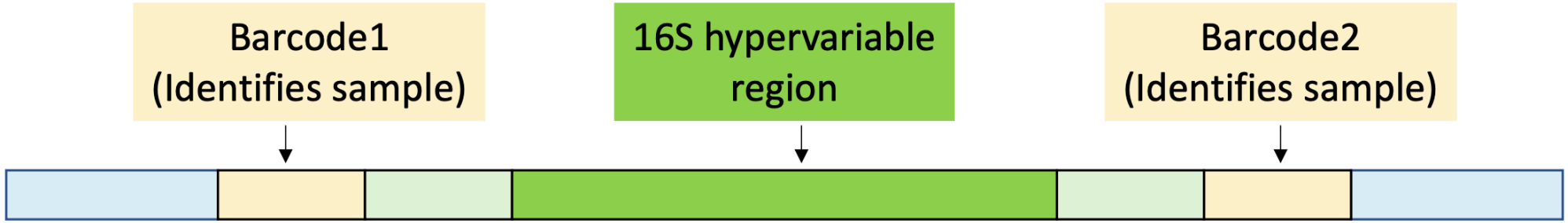
- Quality scores represent the confidence or accuracy of each base call in the sequence. They are encoded using ASCII characters, where a higher ASCII value corresponds to a higher quality score. Common encoding schemes include Phred scores.

Phred Scores:

- Phred scores are logarithmically scaled quality scores. A Phred score of 10 corresponds to a 1 in 10 chance of an incorrect base call, and a Phred score of 20 corresponds to a 1 in 100 chance, and so on.

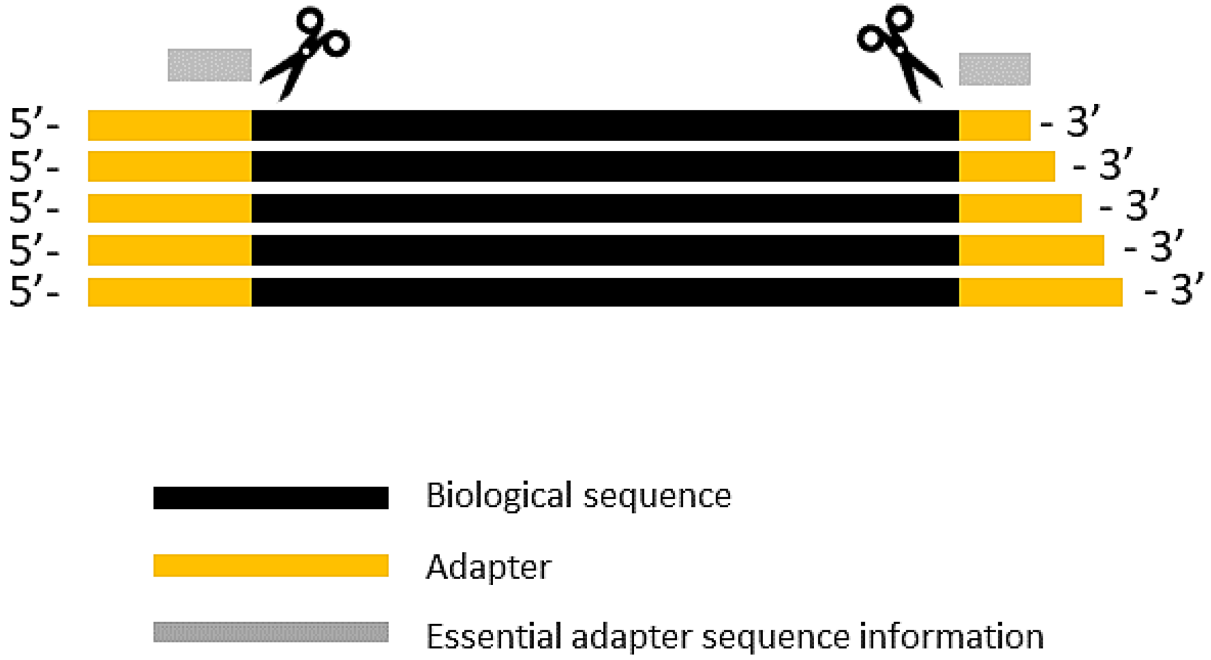
Demultiplexing

- Demultiplexing is a critical step in the bioinformatics pipeline for high-throughput sequencing data analysis. It allows to associate each sequence read with its source sample, enabling accurate and meaningful interpretation of the data in the context of individual experiments or samples.
- Sorting and assigning sequence reads from a single sequencing run to their respective samples or individuals based on barcode sequences.



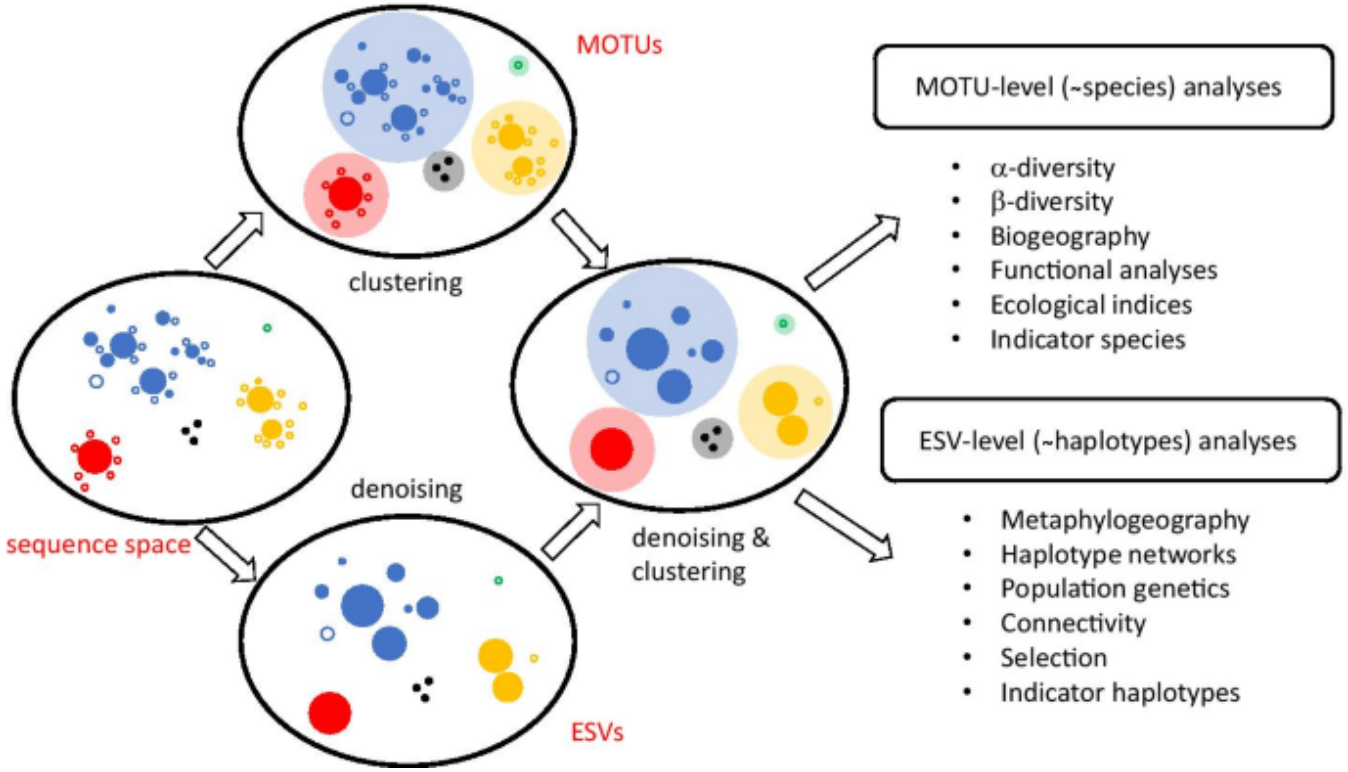
Trimming

- Process of removing unwanted or low-quality bases from the ends of sequence reads. Trimming is a crucial step in the data pre-processing pipeline, and it is typically performed to improve the accuracy and reliability of downstream analyses.



Denoise/clustering

- **Clustering**: an OTU sequence should be at least a given percentage different from all other OTUs and should be the most abundant sequence compared to similar sequences. People traditionally chose to cluster at 97%, which means that the variation between sequences should be at least 3%.
- **Denoising**: to identify all correct biological sequences in the dataset, which is visualized in the figure below. This schematic shows a clustering threshold at 100% and trying to identify errors based on abundance differences. The retained sequences are called ZOTU or Zero-radius Operational Taxonomic Unit. In other software programs they might also be called ASVs.



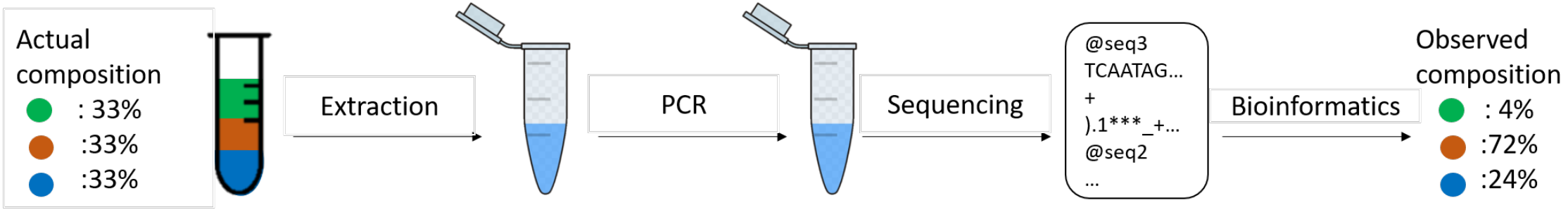
Taxonomic assignment

- Process of assigning an Operational Taxonomic Unit (OTU, that is, groups of related individuals) to sequences that can be reads or contigs. Sequences are compared against a database constructed using complete genomes. When a sequence finds a good enough match in the database, it is assigned to the corresponding OTU. The comparison can be made in different ways.
- **Strategies for taxonomic assignment**
- There are many programs for doing taxonomic mapping, and almost all of them follow one of the following strategies:

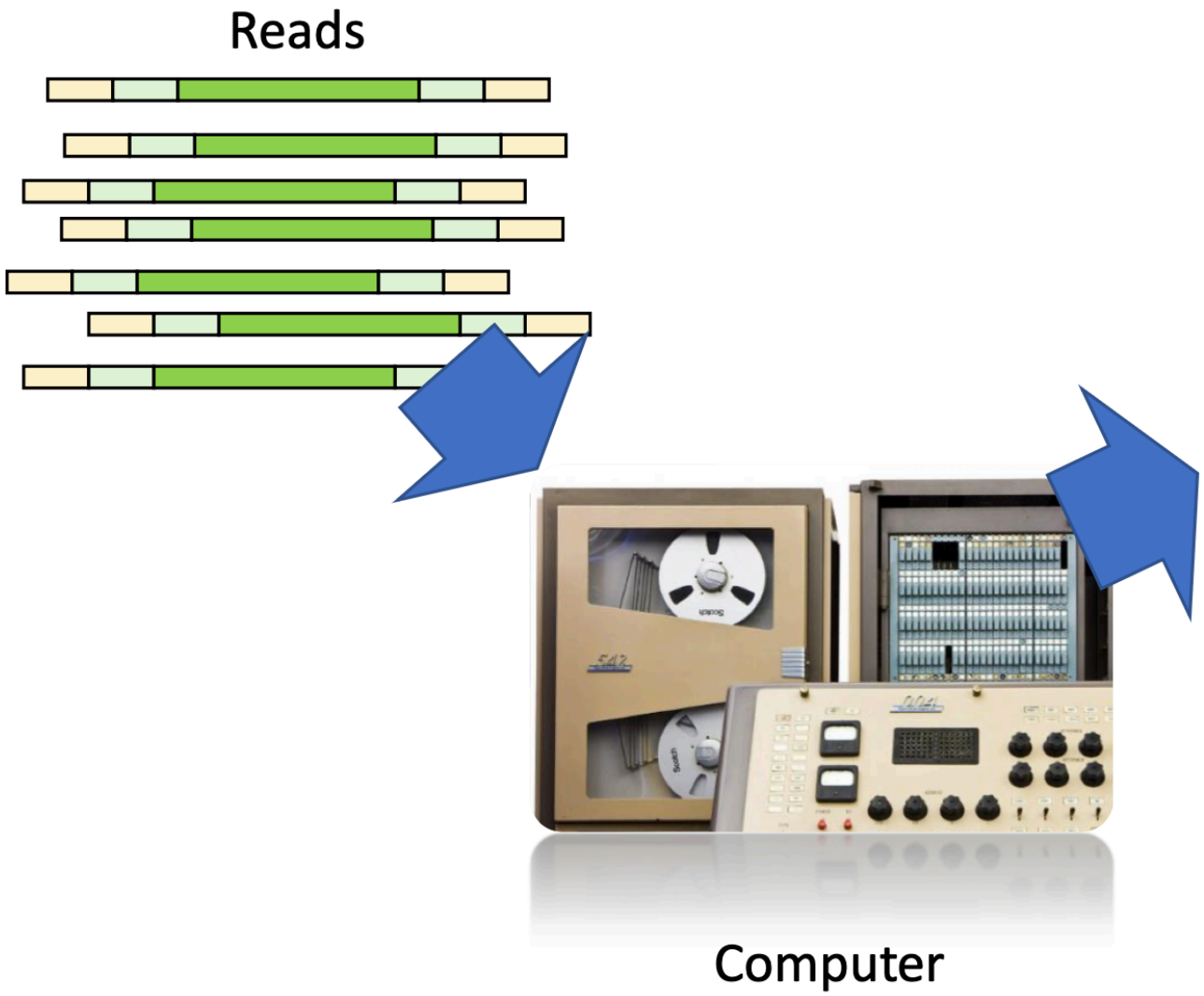
1. **BLAST**: Using BLAST or DIAMOND, these mappers search for the most likely hit for each sequence within a database of genomes (i.e., mapping). This strategy is slow.
2. **Markers**: They look for markers of a database (Greengenes, SILVA, eg.) made a priori in the sequences to be classified and assigned the taxonomy depending on the hits obtained.
3. **K-mers**: A genome database is broken into pieces of length k to be able to search for unique pieces by taxonomic group, from a lowest common ancestor (LCA), passing through phylum to species. Then, the algorithm breaks the query sequence (reads/ contigs) into pieces of length k, looks for where these are placed within the tree and make the classification with the most probable position.

Taxonomic assignment: challenges

- **Sequence Variability:** High sequence diversity within microbial taxa can complicate accurate taxonomic assignment.
- **Database bias:** Incomplete or biased reference databases may impact the accuracy of assignments, especially for novel or less-studied taxa.
- **Abundance bias**: The absolute abundance of a taxon is the number of sequences (eg. reads) assigned to it. Moreover, its relative abundance is the proportion of sequences assigned to it. It is essential to be aware of the many biases that can skew the abundances along the metagenomics workflow, shown in the figure, and that because of them, we may not be obtaining the actual abundance of the organisms in the sample.



OTU tables

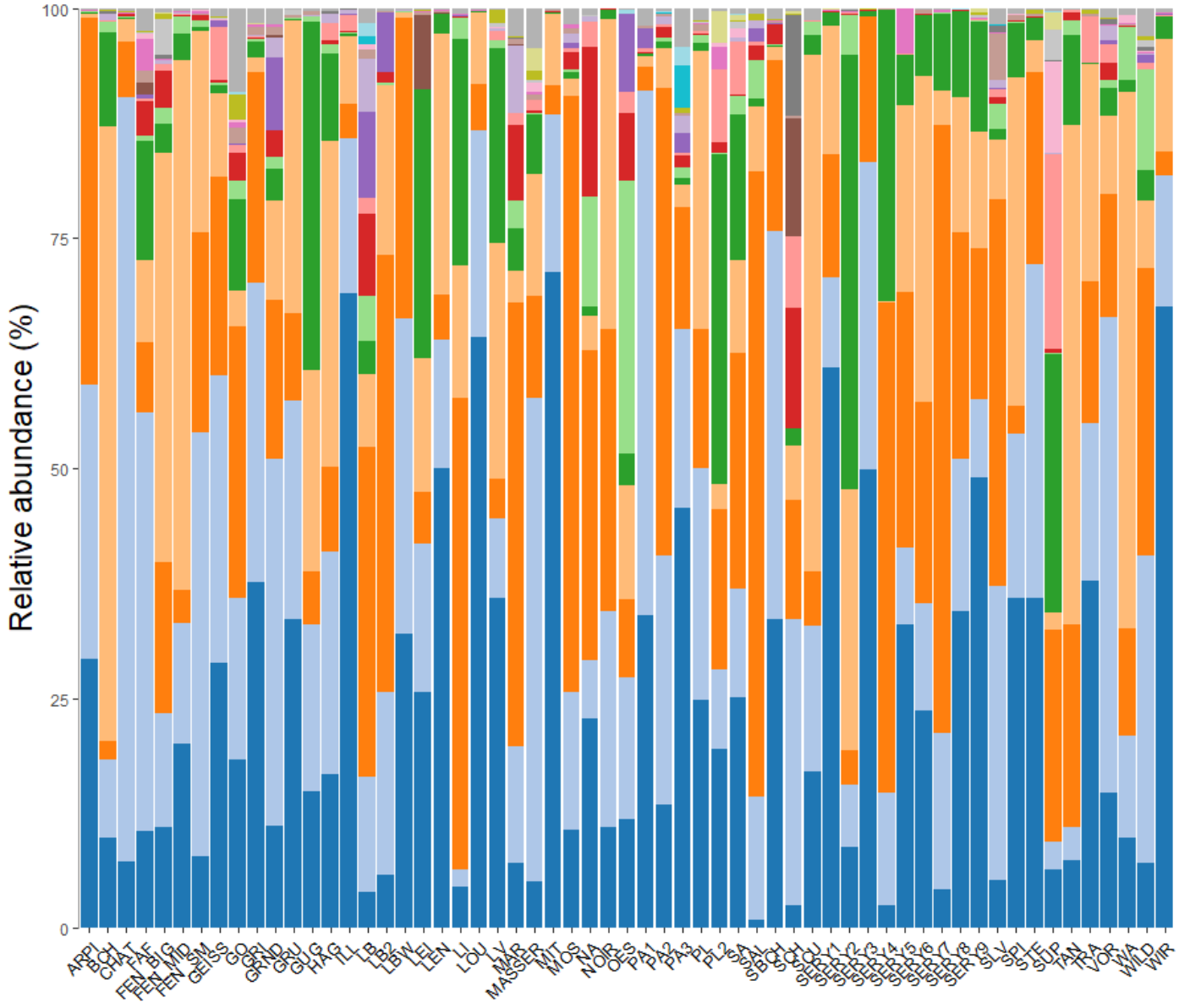


OTU table

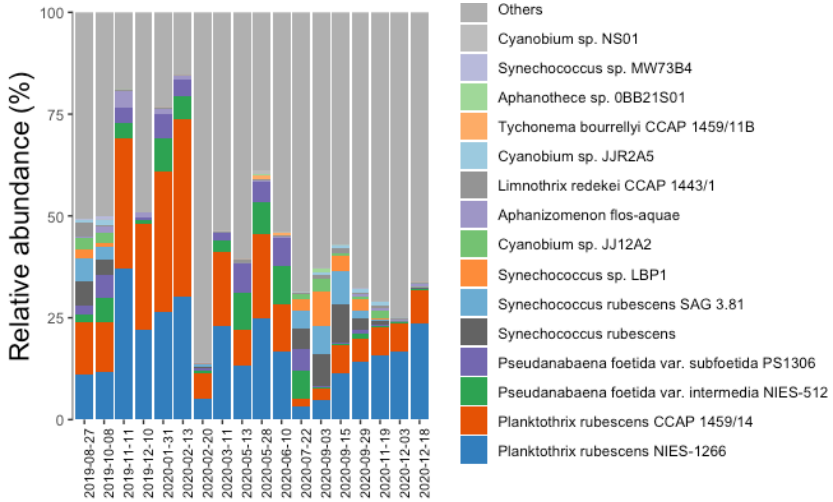
	Sam1	Sam2	Sam3
Otu1	0.34	0.32	0.29
Otu2	0.12	0.17	0.10
Otu3	0.07	0.03	0.11
Otu4	0.06	0.02	0.09

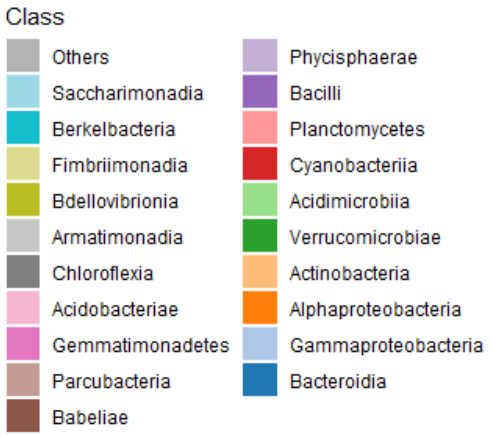
Rows are "OTUs" ~ species
Columns = samples
Values = frequencies

Data visualization: relative abundance barplots

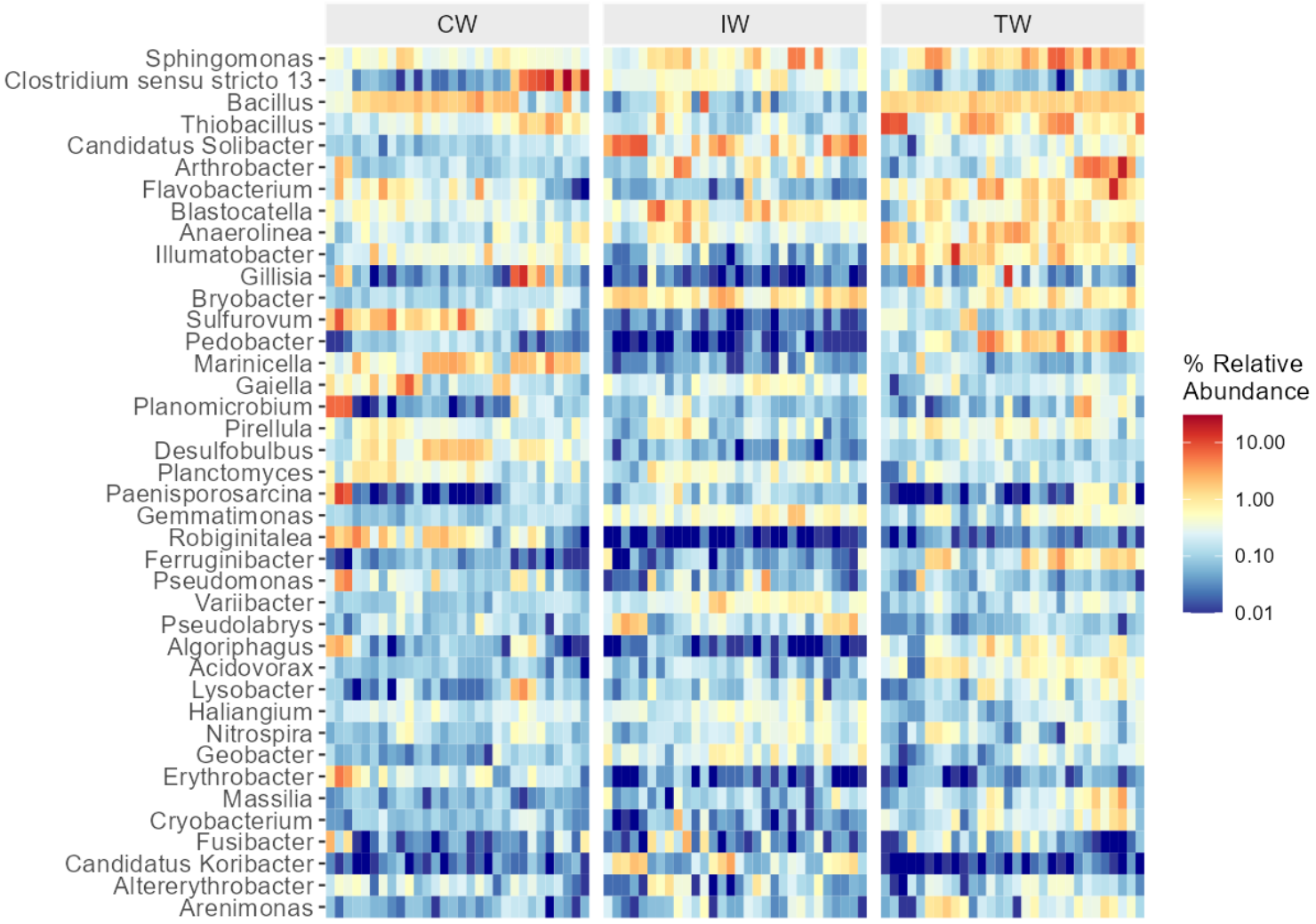


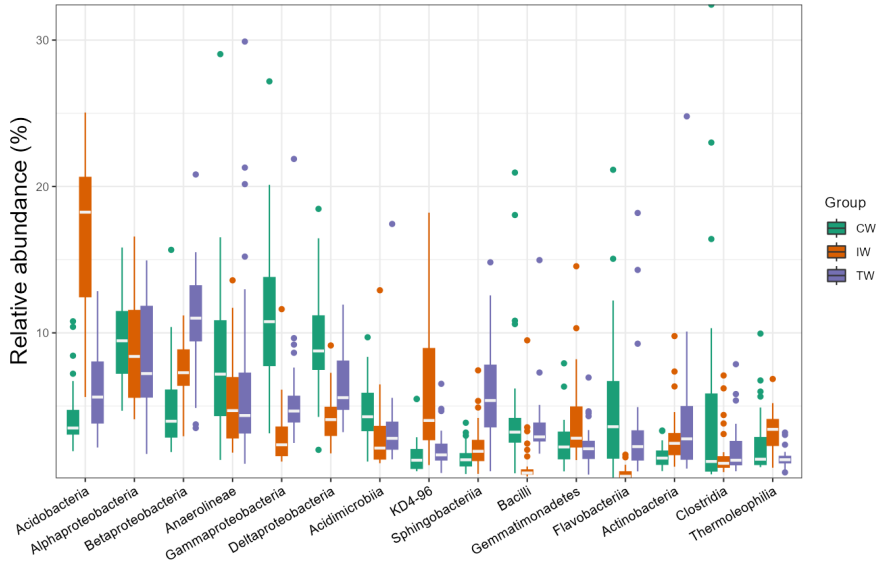
Lake Geneva



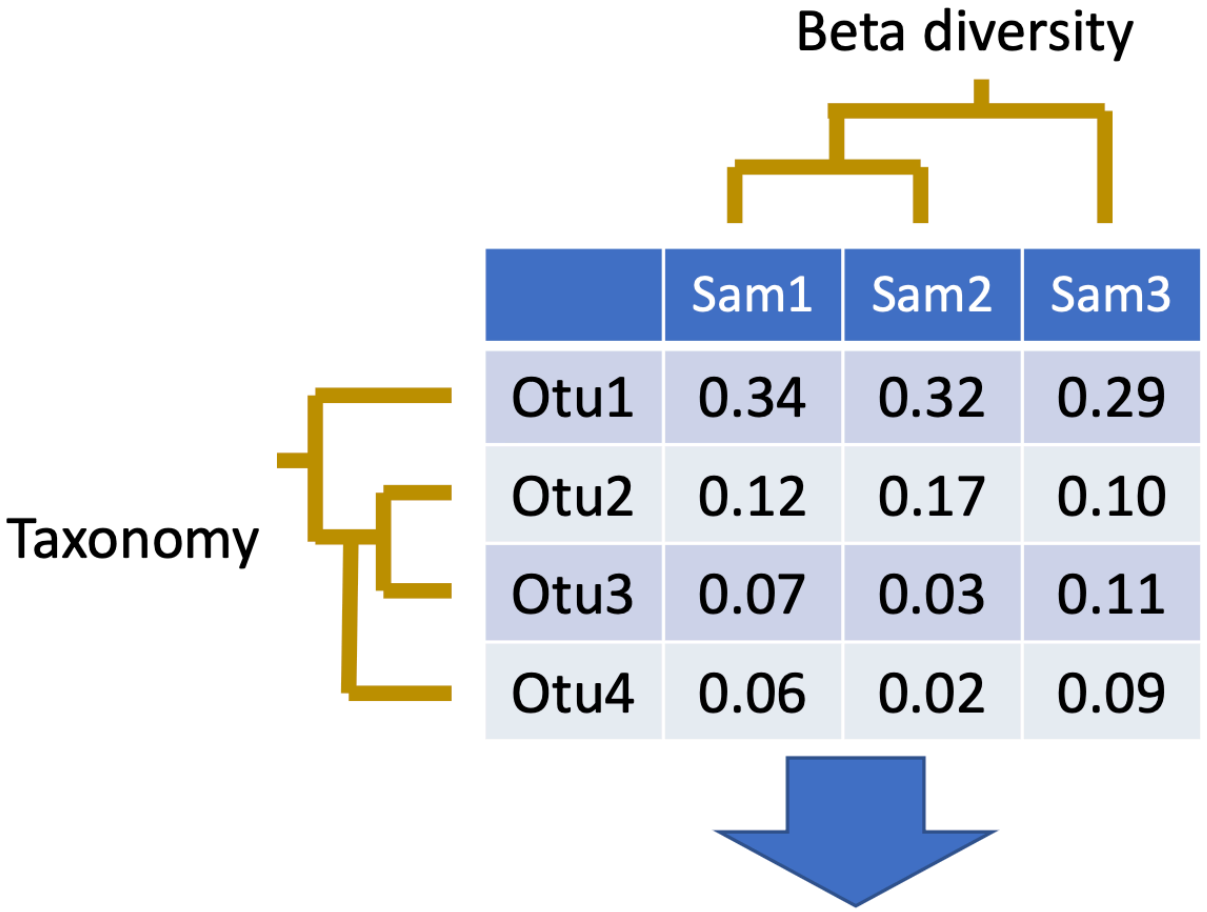


Data visualization: relative abundance as heatmaps and boxplots





Diversity analysis

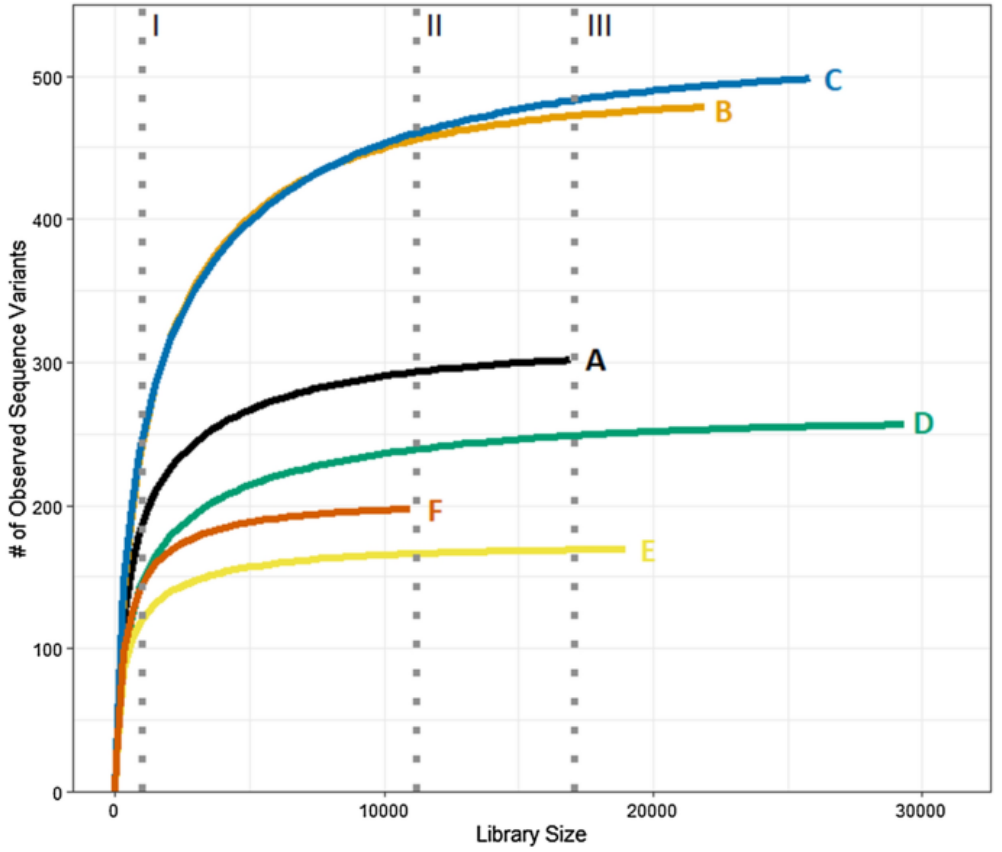


	Sam1	Sam2	Sam3
Otu1	0.34	0.32	0.29
Otu2	0.12	0.17	0.10
Otu3	0.07	0.03	0.11
Otu4	0.06	0.02	0.09

Alpha diversity = how many species in one sample
Beta diversity = how similar / different is each pair of samples
Taxonomy = classification and relationships between OTUs

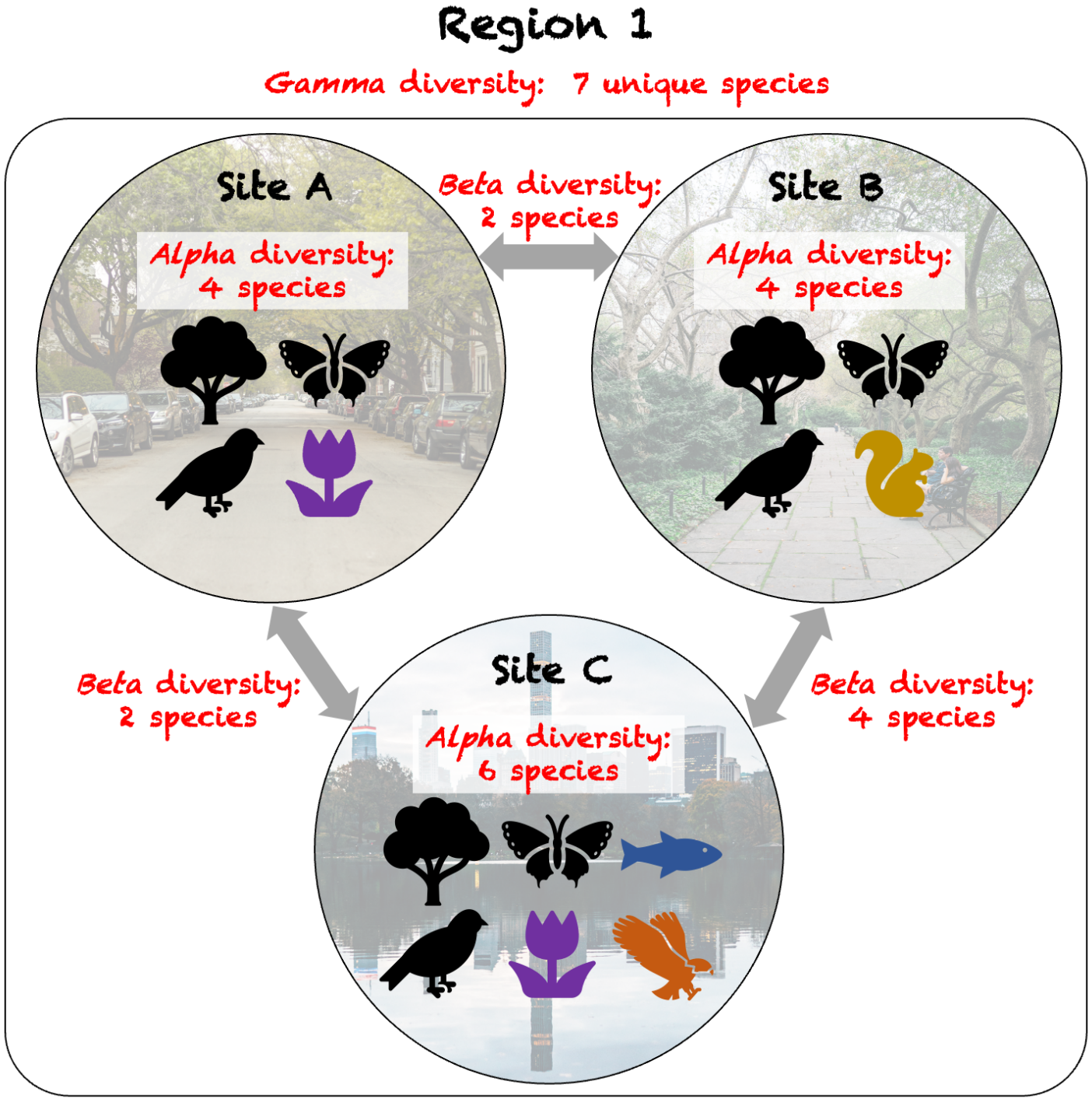
Rarefaction: to be or not to be

- Rarefaction is a common yet strongly criticized method developed to assess the coverage of detected sequences in a sample by plotting the number of observed sequences (or taxa, i.e. genera or species) as a function of sequencing depth.

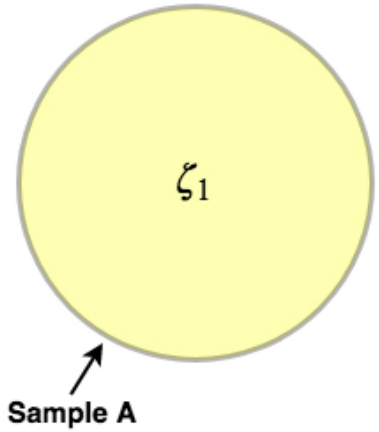
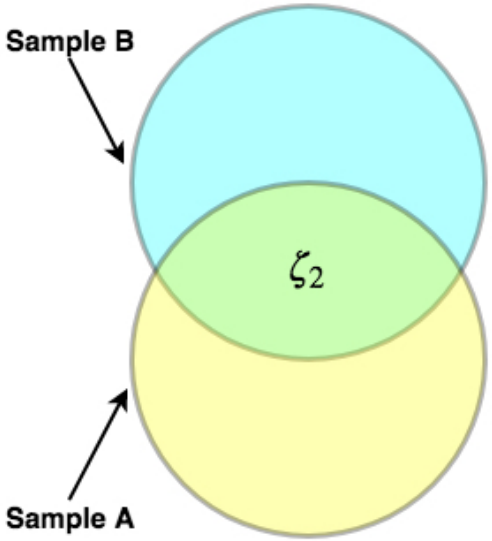
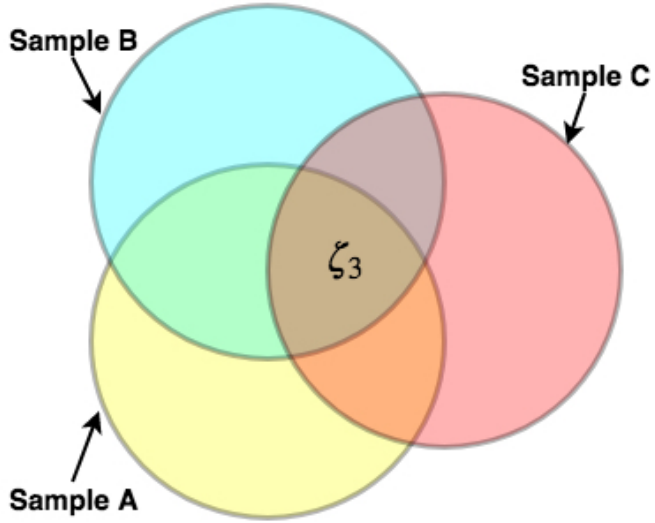


Types of diversity

- Alpha diversity (sample)
- Beta diversity (2 samples)
- Gamma diversity (region)
- Zeta diversity (overlap)



- Zeta diversity (overlap)

		
$\zeta_1 = A$	$\zeta_1 = \frac{1}{2} \times (A + B)$	$\zeta_1 = \frac{1}{3} \times (A + B + C)$
	$\zeta_2 = A \cap B$	$\zeta_2 = \frac{1}{3}(A \cap B + B \cap C + C \cap A)$
		$\zeta_3 = A \cap B \cap C$

- Chao1: alpha diversity richness

$$Chao1 = S + (a^2/2b)$$

- Bray- Curtis dissimilarity: beta diversity richness

"In the first case you subtract the abundance of one species in a sample from its counterpart in the other sample but ignore the sign. The second component is the abundance of a species in one sample added to the abundance of its counterpart in the second sample. If a species is absent, then its abundance should be recorded as 0 (zero)".

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

- Gamma diversity:

$${}^{q}D_{\gamma} = \frac{1}{\sqrt[q-1]{\sum_{i=1}^{S} p_i p_i^{q-1}}}.$$

R packages for community ecology analysis

- Vegan: https://cran.r-project.org/web/packages/vegan/index.html

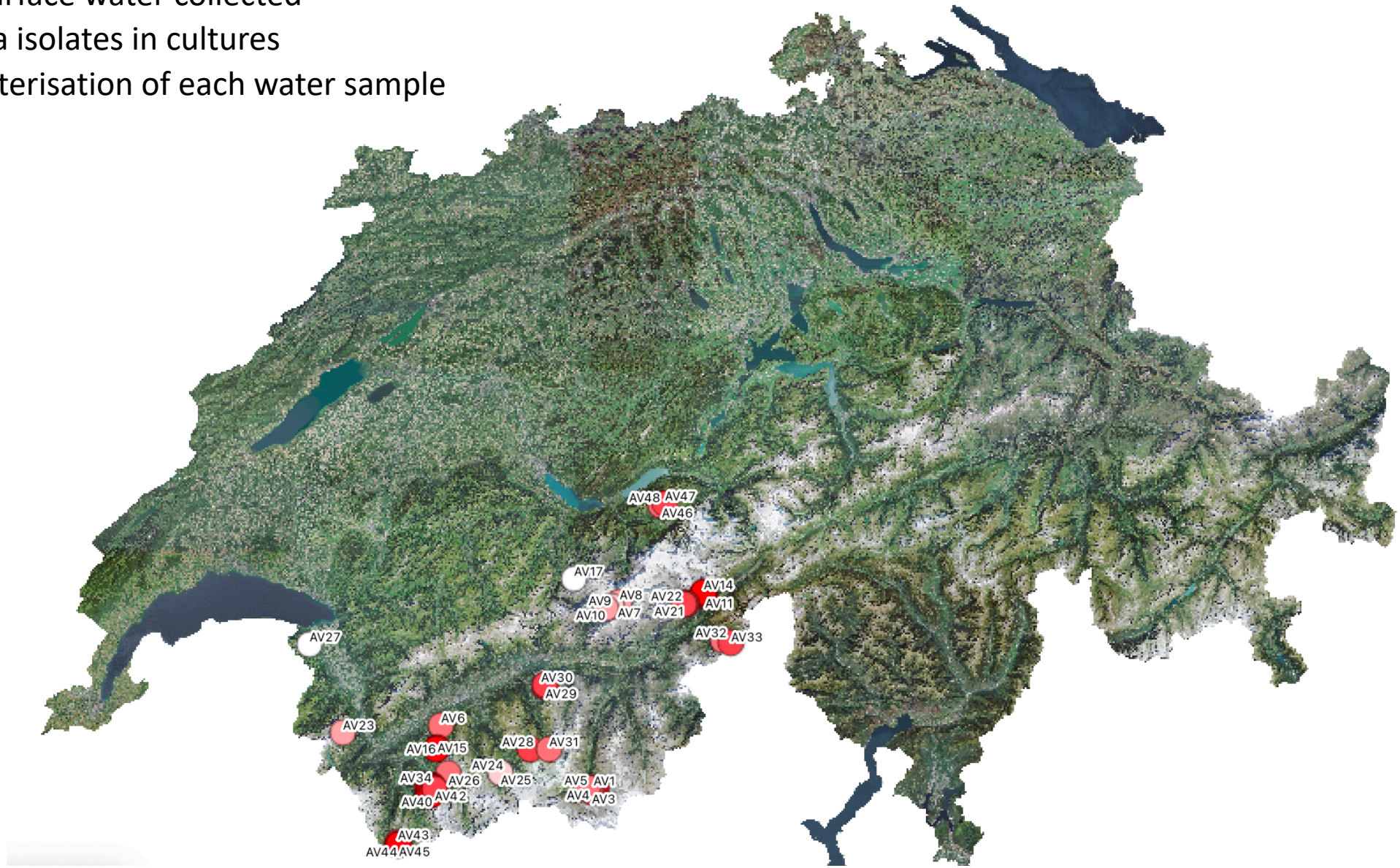
The vegan package in R provides functions for calculating a variety of descriptive statistics, including species richness, species diversity, and species evenness. These statistics can be calculated for the entire community or for specific subsets of the data.

- Microeco: https://chiliubio.github.io/microeco_tutorial/

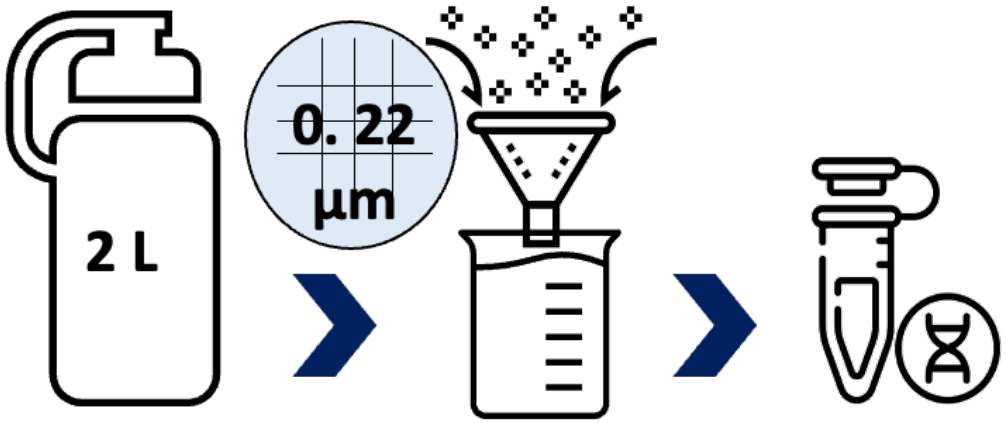
Data mining in microbial community ecology. Rapidly analyzing microbiome data, utilizing a range of cutting-edge and commonly adopted methodologies. To facilitate data mining, every component of the microeco package has been modularized to ensure that users can easily recall, search, and employ classes.

Case study: spatial distribution of bacteria diversity in swiss alpine lakes

- About 60 samples of surface water collected
- More than 220 bacteria isolates in cultures
- Biogeochemical characterisation of each water sample



Project’s methodological overlook



16/18S metagenomics
Illumina NextSeq

qPCR (16s rRNA, mcyA, Humbac, 18s…)

Taxonomic abundance
(reads/total sequencing reads)

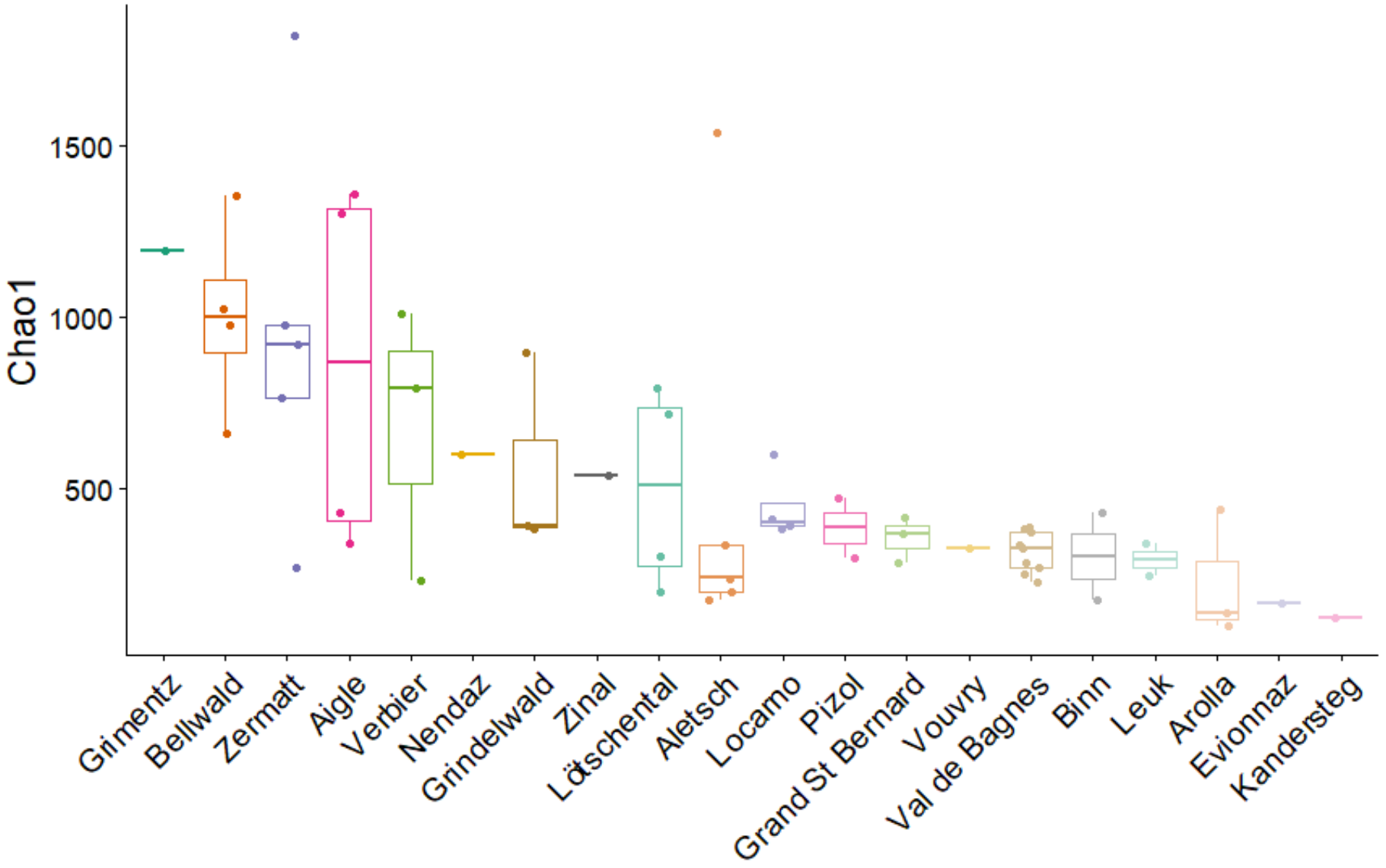
Diversity indices
(alpha as Chao1, beta as Bray-Curtis dissimilarity distance)

Functional predictions
(Picrust2)

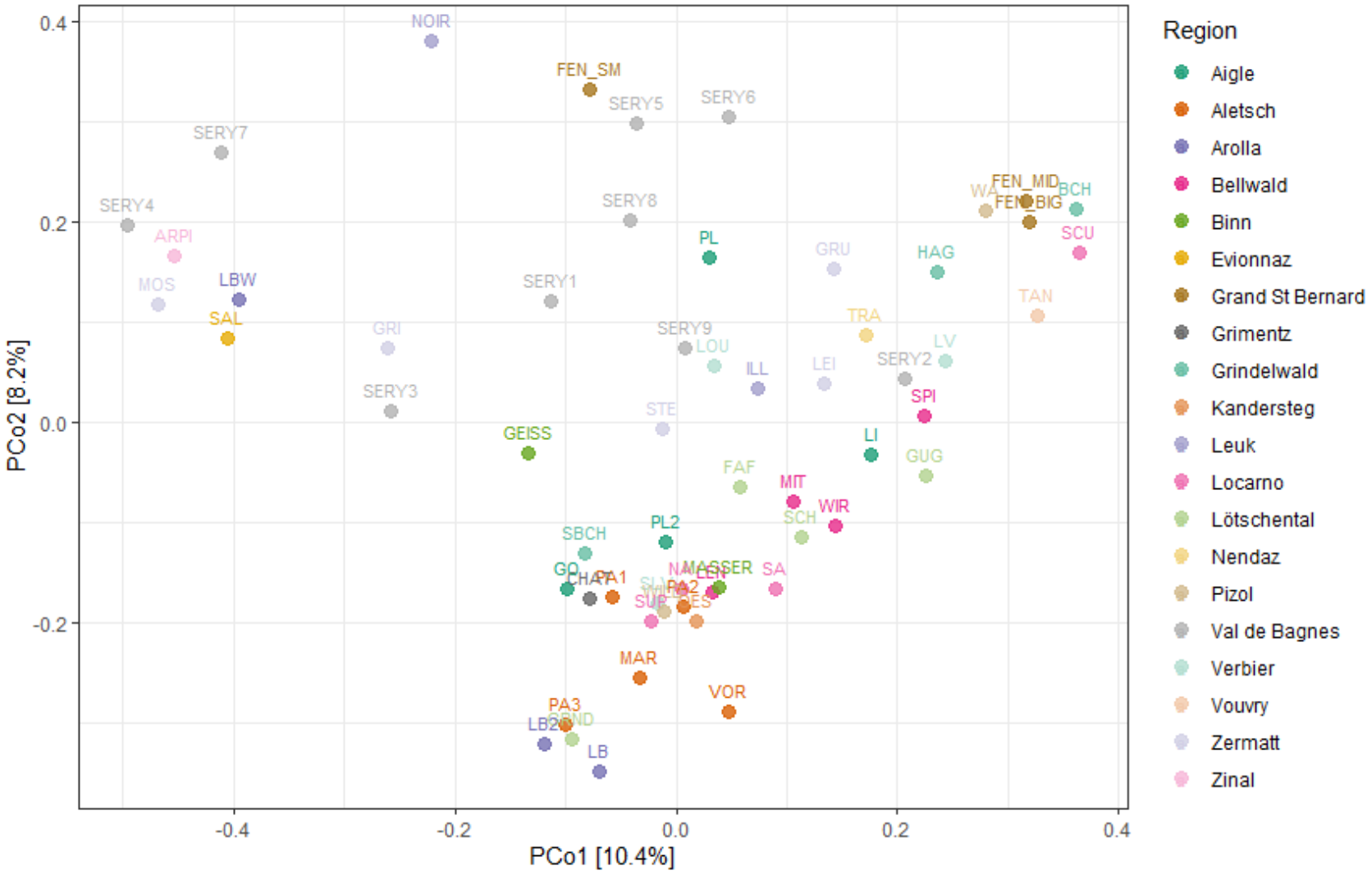
Correlation and redundancy analysis
(RDA plots, environmental associations)

Identification of biomarkers
(LefSe)

Alpha diversity case study: Bacteria diversity in Swiss alpine lakes



Beta diversity case study: Bacteria diversity in Swiss alpine lakes

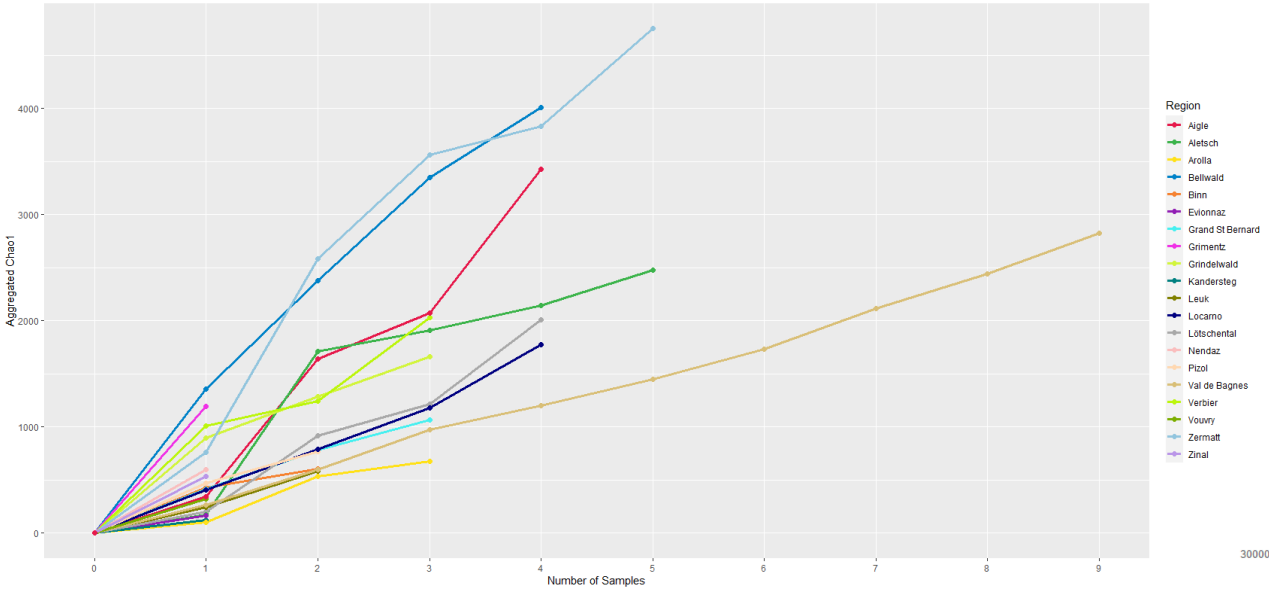


Region

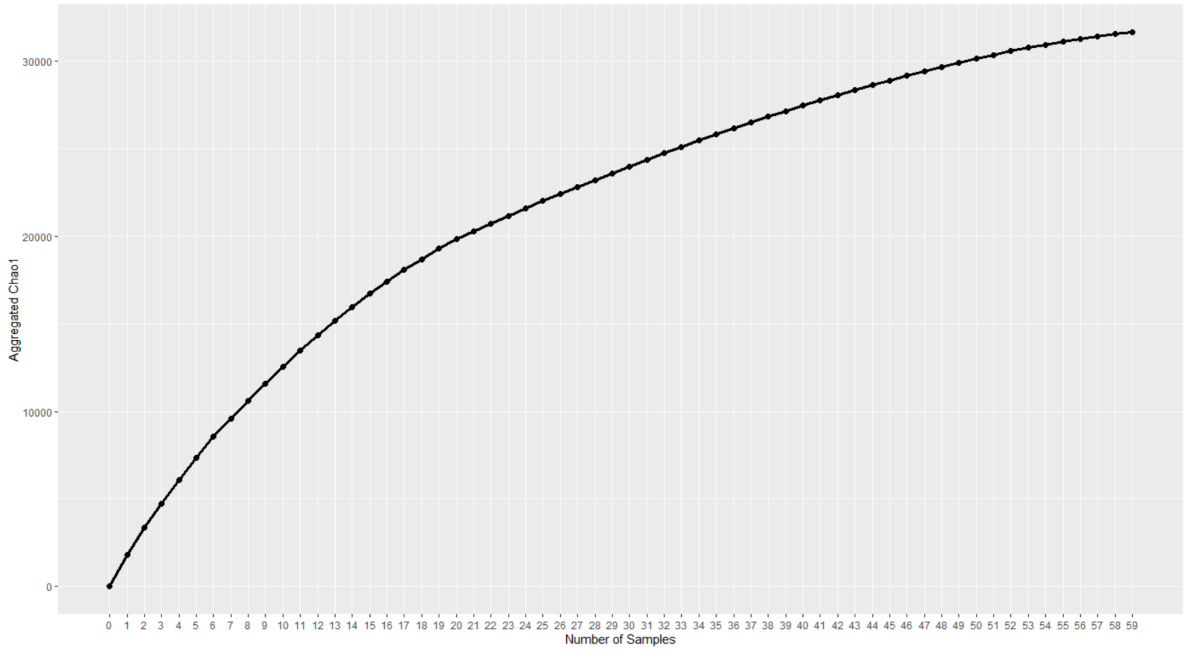
- Aigle
- Aletsch
- Arolla
- Bellwald
- Binn
- Evionnaz
- Grand St Bernard
- Grimentz
- Grindelwald
- Kandersteg
- Leuk
- Locarno
- Lötschental
- Nendaz
- Pizol
- Val de Bagnes
- Verbier
- Vouvry
- Zermatt
- Zinal

Gamma diversity case study: Bacteria diversity in Swiss alpine lakes

By subregion: identification of regions with higher diversity (Zermatt and Bellwald)



All samples in the Alps



Differential Abundance Testing

- **Definition:** Statistical approach used to identify taxa that significantly differ in abundance between conditions or treatments.
- **Methods:** DESeq2, edgeR, ANCOM, LefSe…
- **Application:** Pinpointing microbes associated with specific treatments or environmental factors.

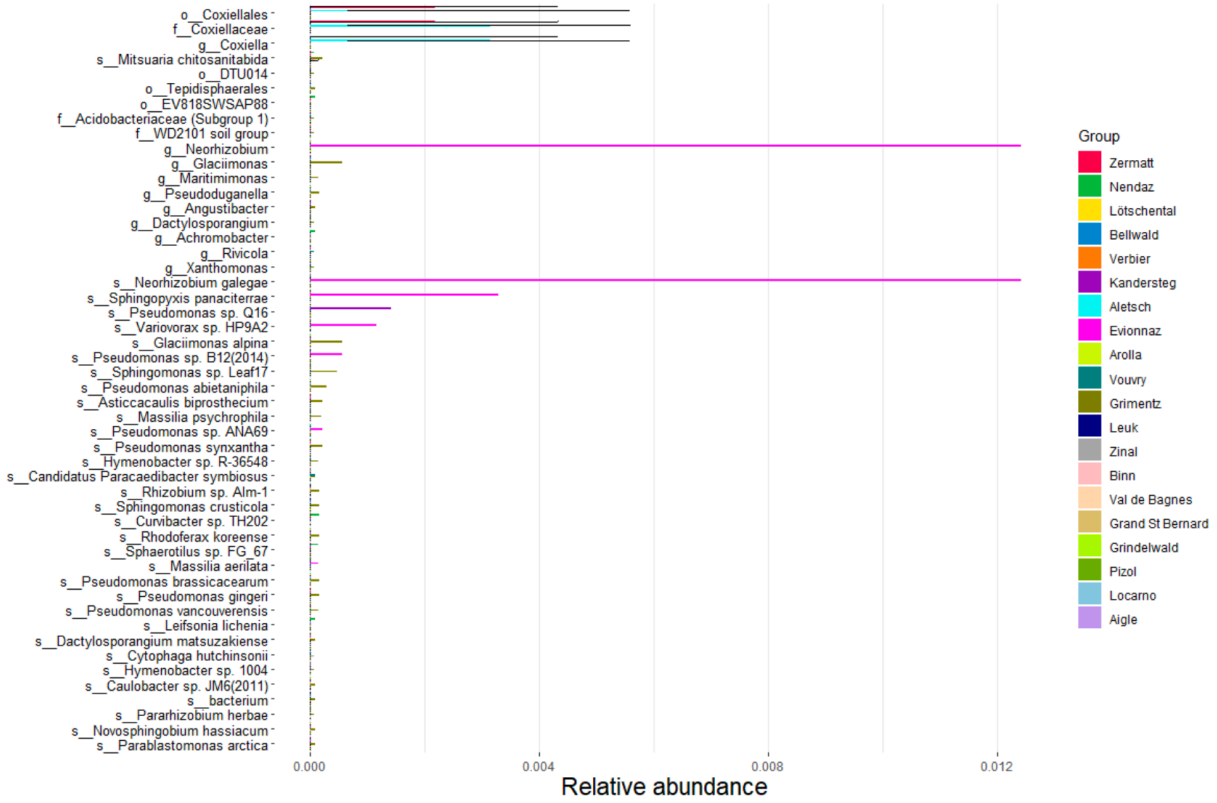
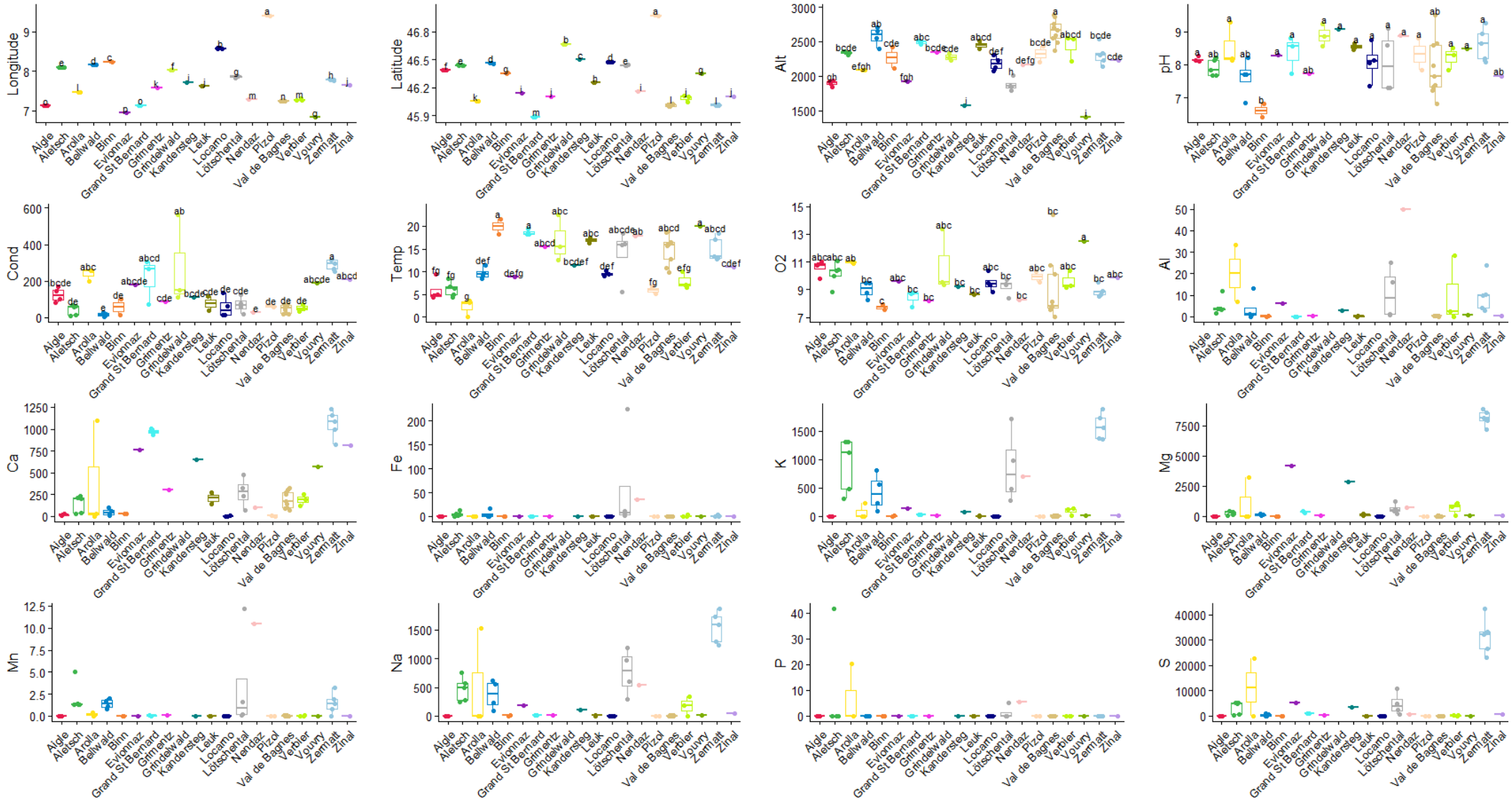


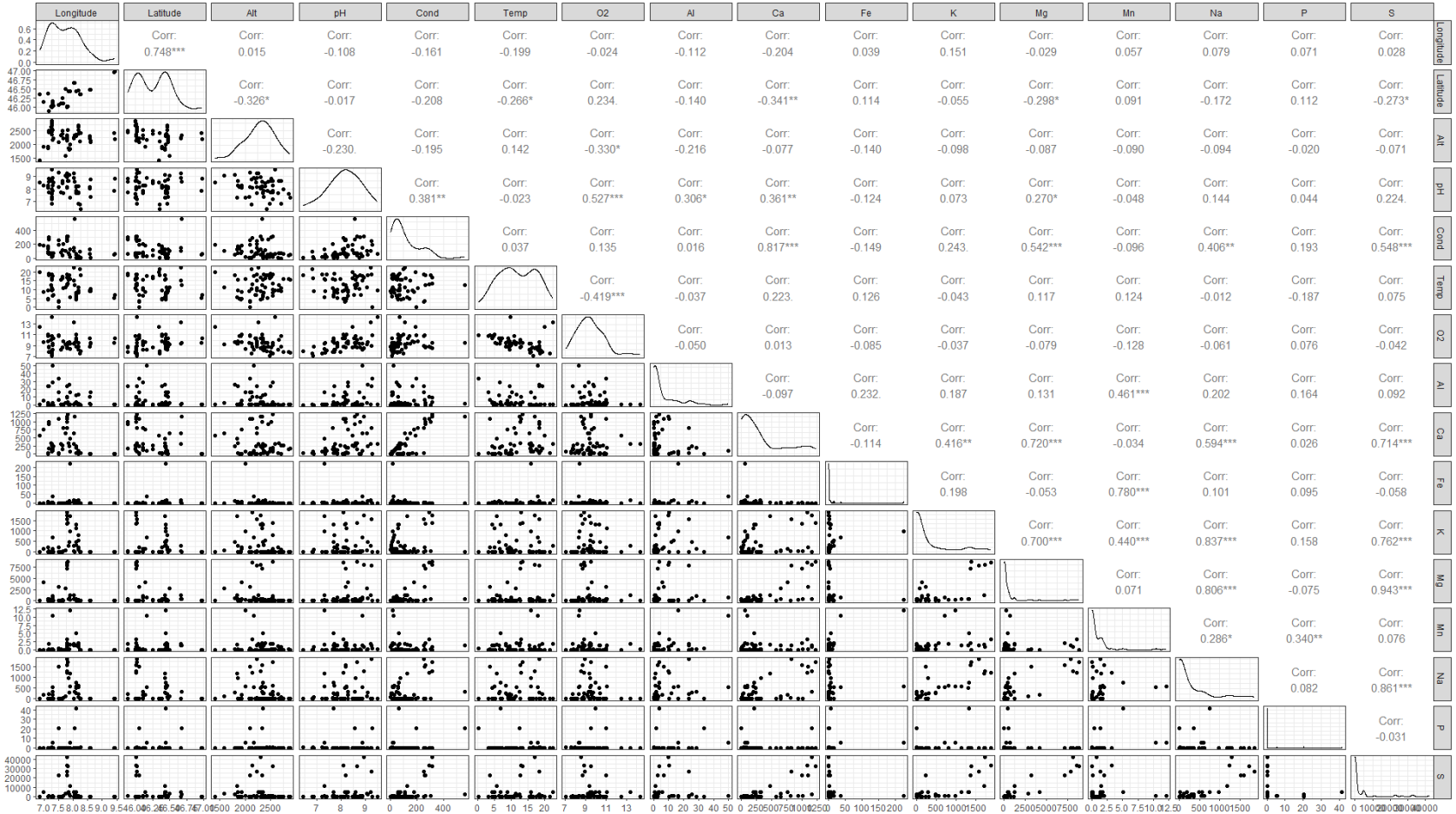
Figure 14: Difference in relative abundance between region using LEfSe and keeping the first 50 taxa

Environmental associations



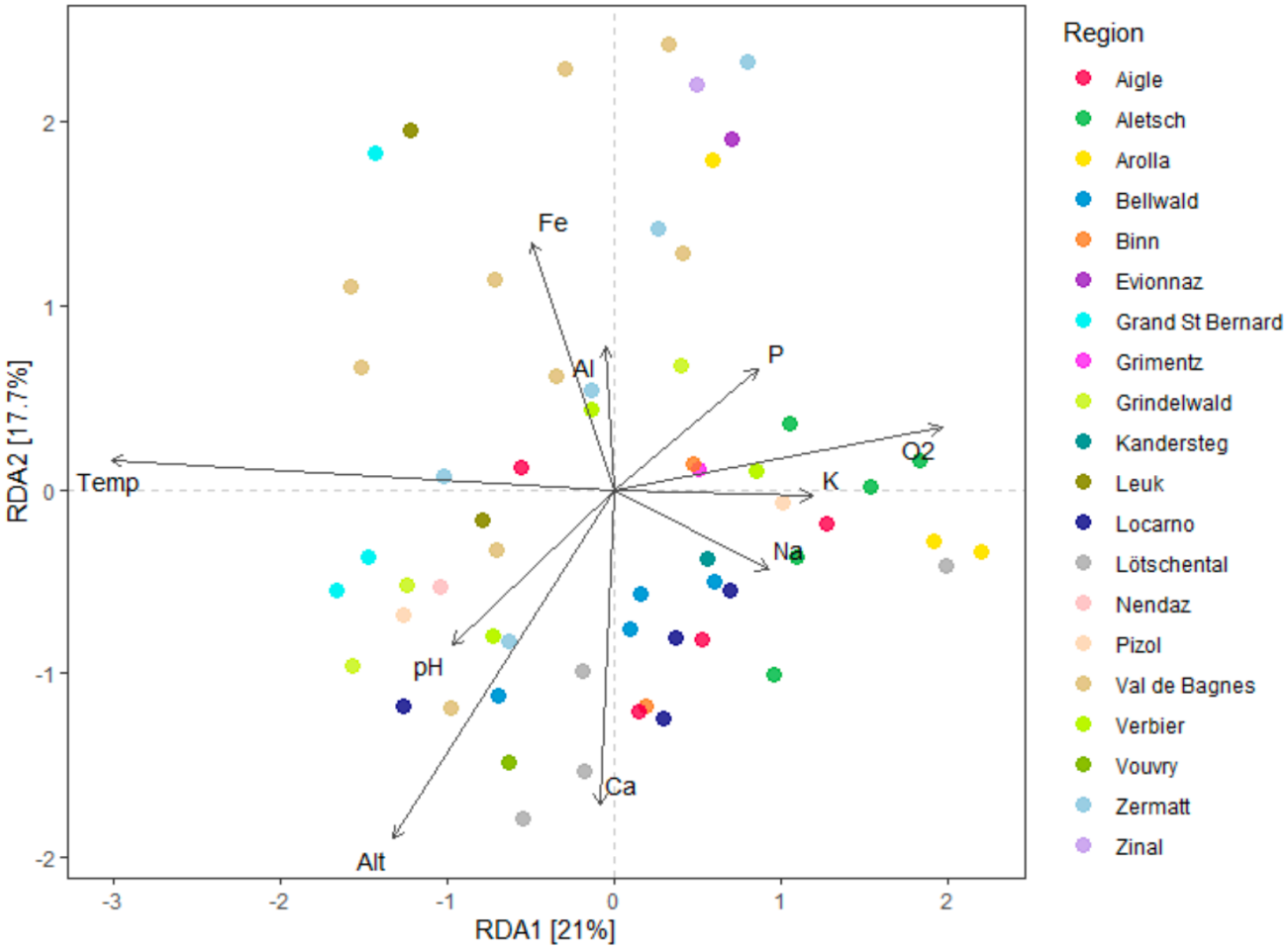
Environmental associations: avoiding co-correlations

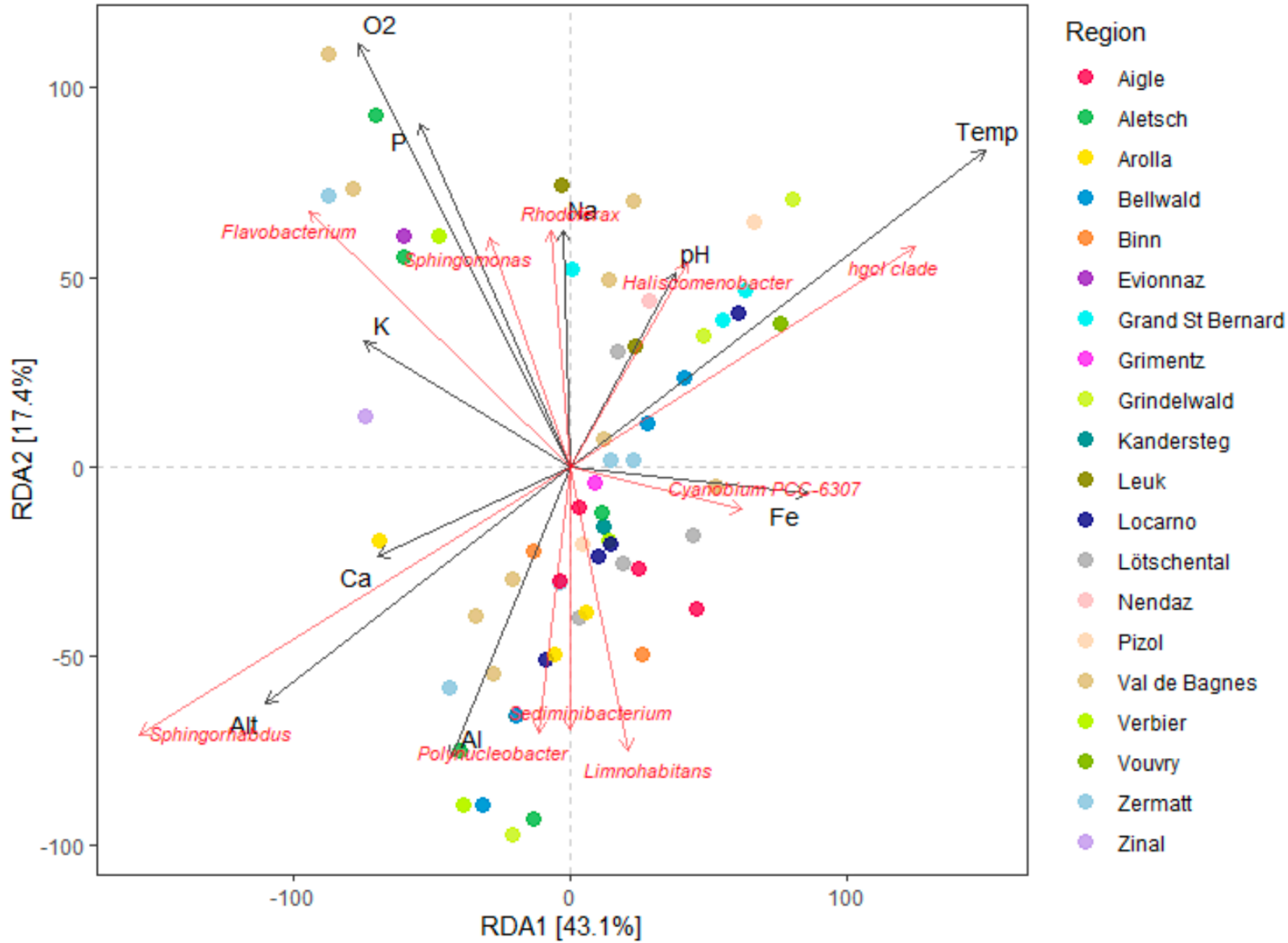
- Eliminate variables with high correlation coefficients (eg. > 0.7 or VIF > 5)
- Keep the most biologically significant of the pair, if evidence.



Environmental associations: redundancy analysis (RDA)

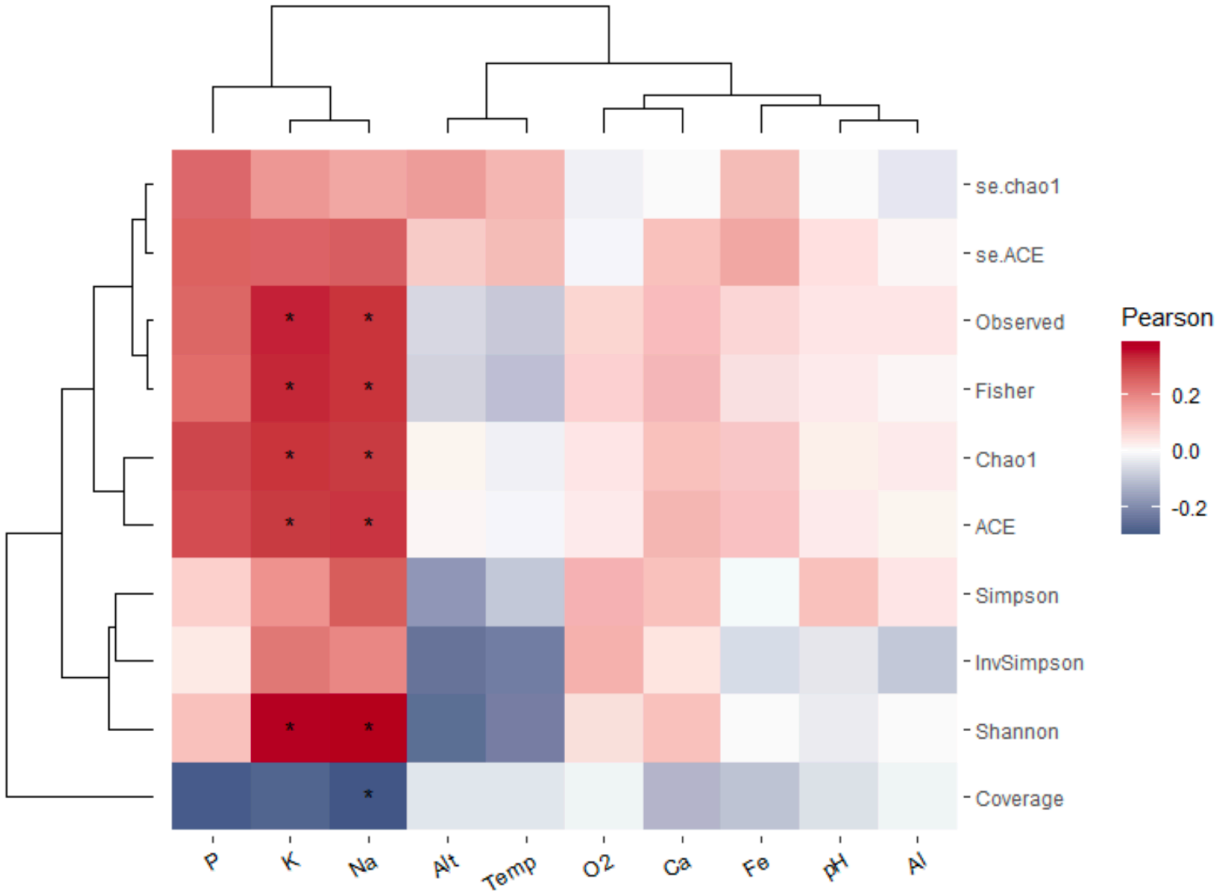
- Multivariate statistical technique commonly used in ecology to explore relationships between variables.
- Multiple regression considering linear combinations of the predictor variables that capture the most variation in both the response and predictor sets.
- Results often visualised in an ordination plot, representing the relationships between the observations and the variables in a reduced-dimensional space.




O2
P
Na
Temp
pH
K
Ca
Fe
Alt
Al
Flavobacterium
Sphingomonas
Rhodobacter
Haliscomenobacter
hgcI clade
Cyanobium PCC-6307
Sphingorhabdus
Polynucleobacter
Sediminibacterium
Limnohabitans
RDA1 [43.1%]
RDA2 [17.4%]
100
50
0
-50
-100
-100
0
100
Region
Aigle
Aletsch
Arolla
Bellwald
Binn
Evionnaz
Grand St Bernard
Grimentz
Grindelwald
Kandersteg
Leuk
Locarno
Lötschental
Nendaz
Pizol
Val de Bagnes
Verbier
Vouvry
Zermatt
Zinal

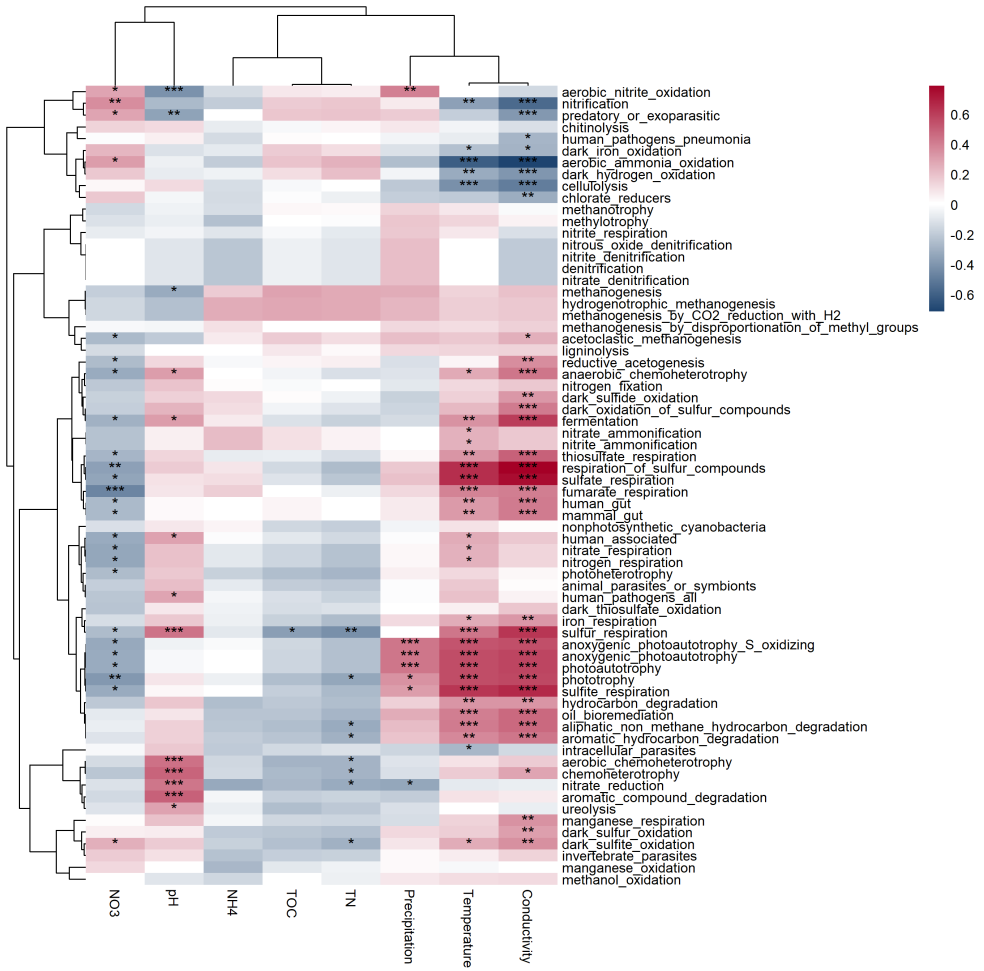
Spearman or Pearson Correlation Analysis

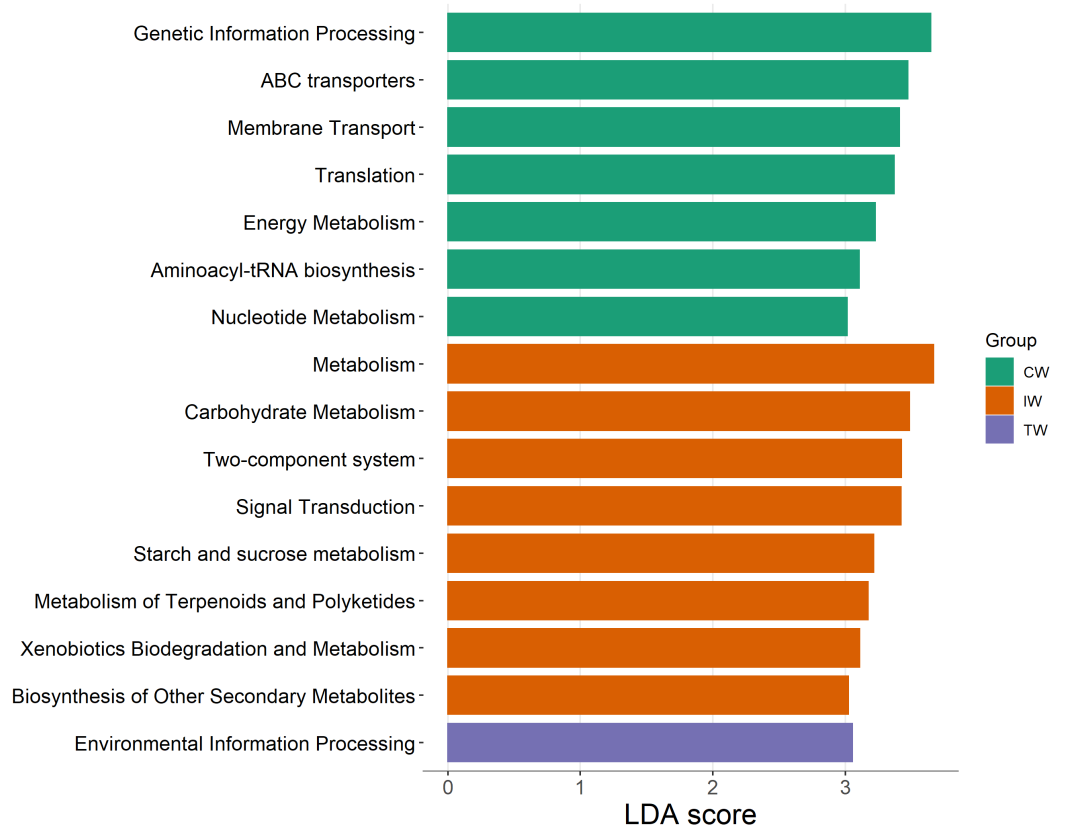
Description: Examines the correlation between the abundance of specific microbial taxa and continuous environmental variables
Application: Identifying taxa that show significant correlations with environmental gradients.



Functional predictions (Picrust2, Faprotax…)

Definition: Assessing the functional potential of microbial communities.
Methods: PICRUSt, HUMAnN, and other metagenomic pathway analysis tools.
Application: Uncovering functional differences between groups.





Other applications:

1.Random Forest Analysis:

1. **Description:** Machine learning approach that can predict environmental variables based on microbial community data.
2. **Application:** Building predictive models to understand the relationship between microbial composition and environmental factors.

2.Co-occurrence Network Analysis:

1. **Description:** Examines patterns of co-occurrence or mutual exclusion between microbial taxa and environmental variables.
2. **Application:** Identifying microbial interactions and their relationship with environmental conditions.

3.Mantel Test:

1. **Description:** Assesses the correlation between matrices of microbial community dissimilarity and environmental dissimilarity.
2. **Application:** Evaluating the overall congruence between microbial and environmental patterns.