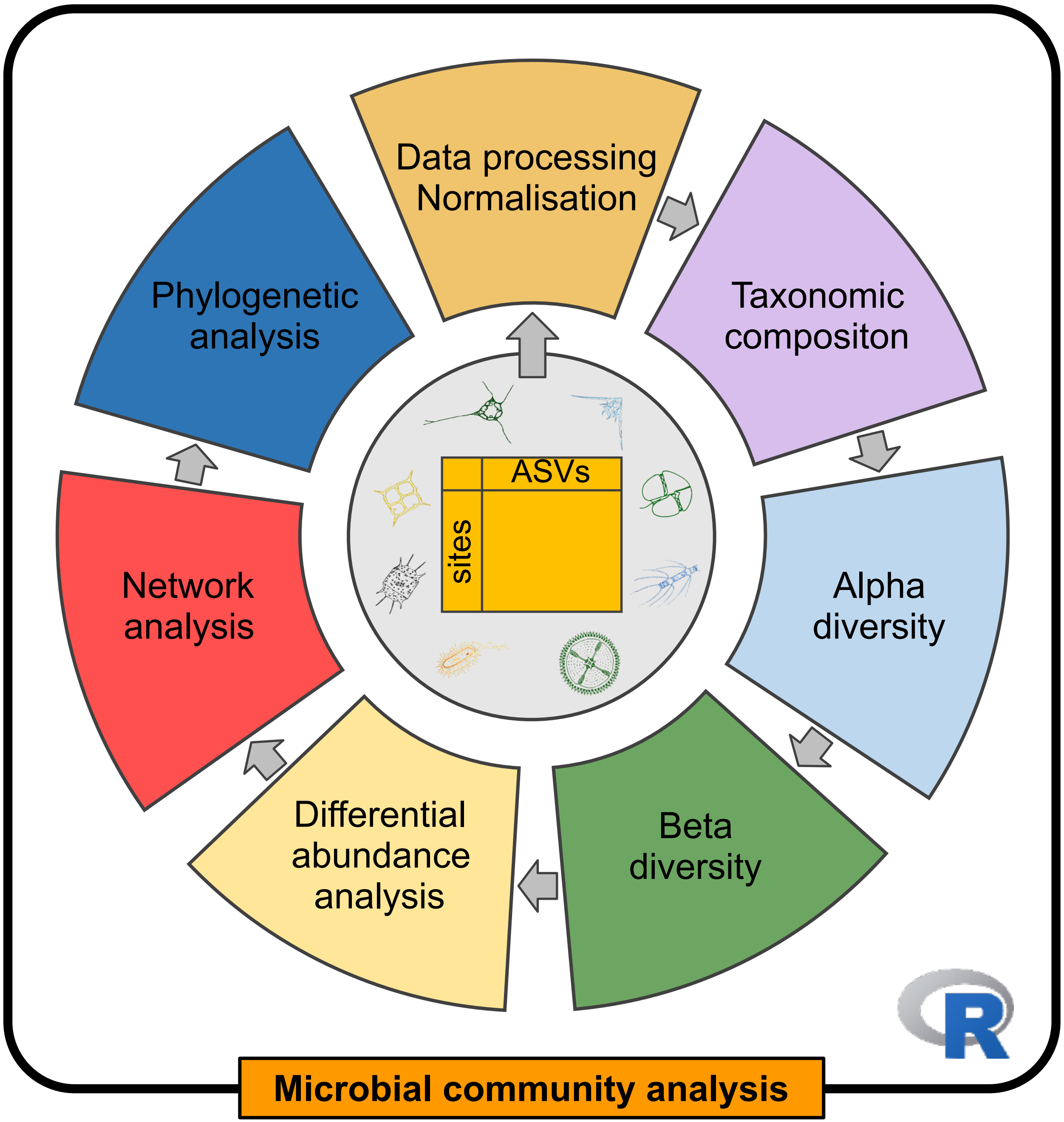
Ecological analysis of metabarcoding data

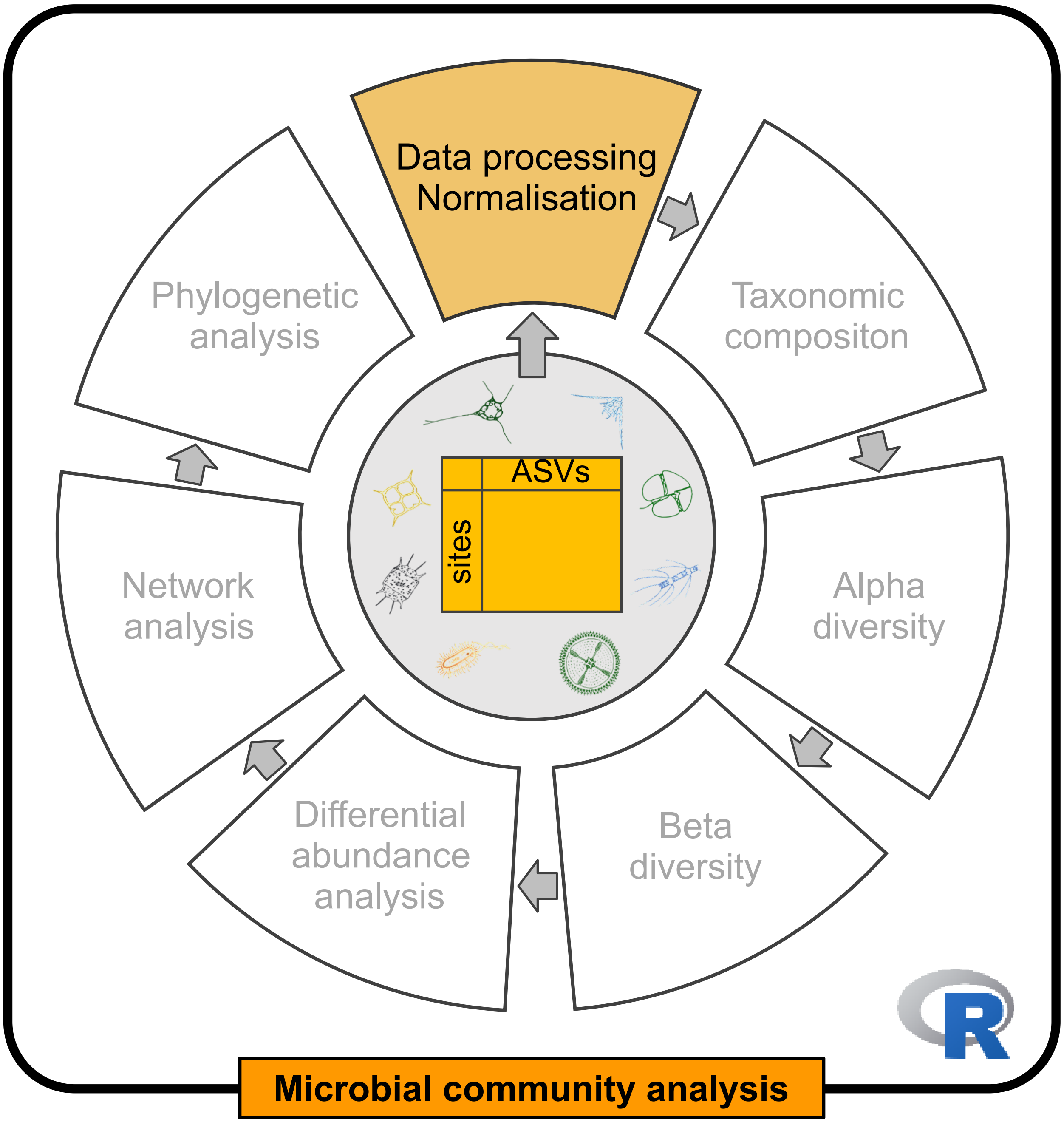
Data preparation

Clarisse Lemonnier

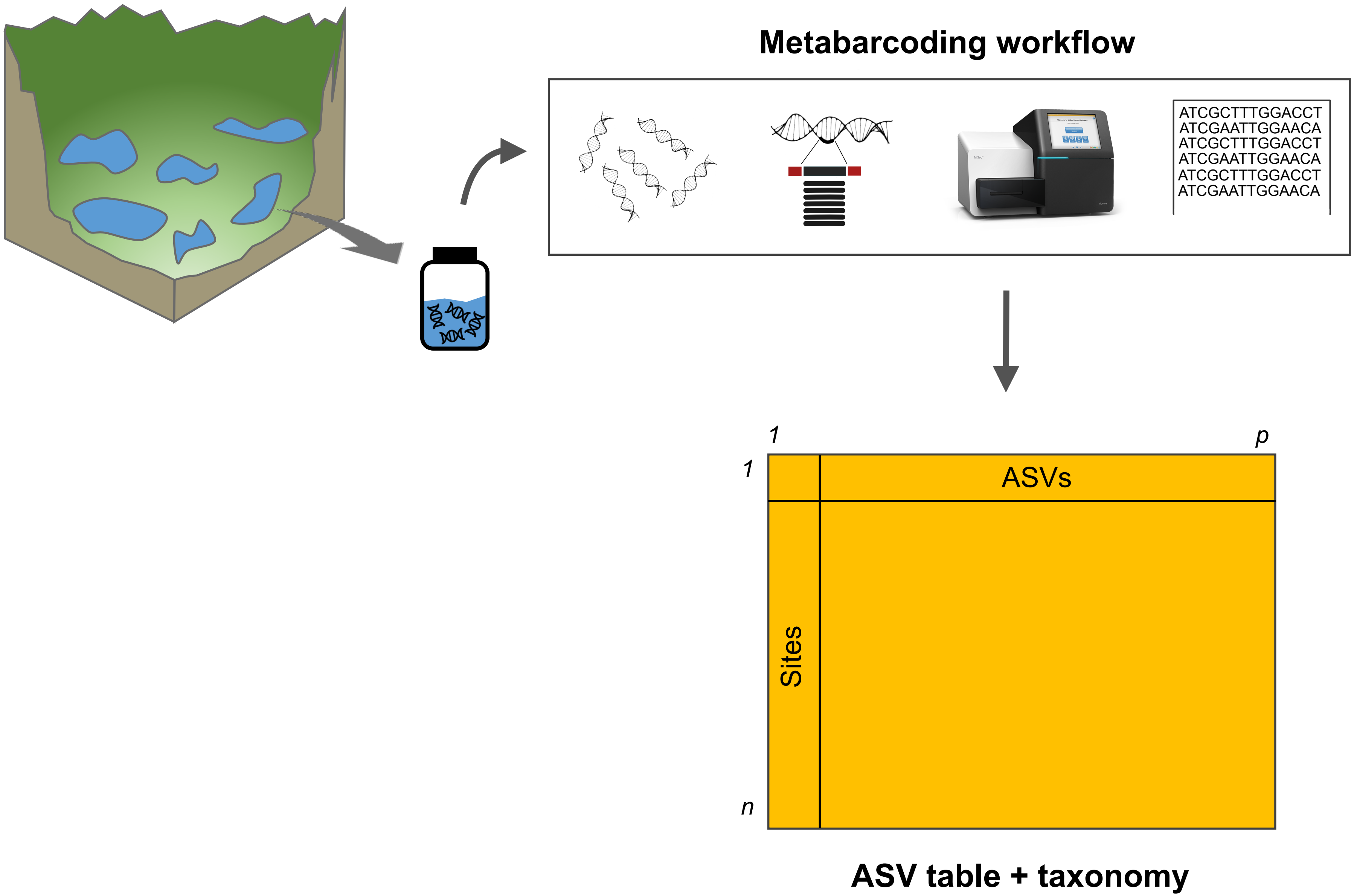
Statistical analysis workflow from metabarcoding data



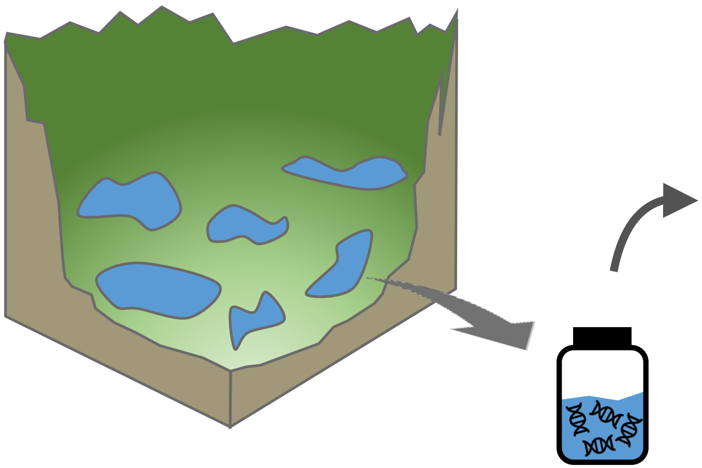
Statistical analysis workflow from metabarcoding data



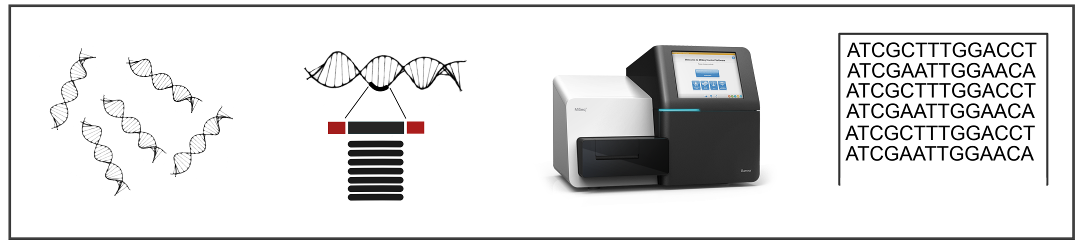
Data processing, normalisation



Data processing, normalisation



Metabarcoding workflow



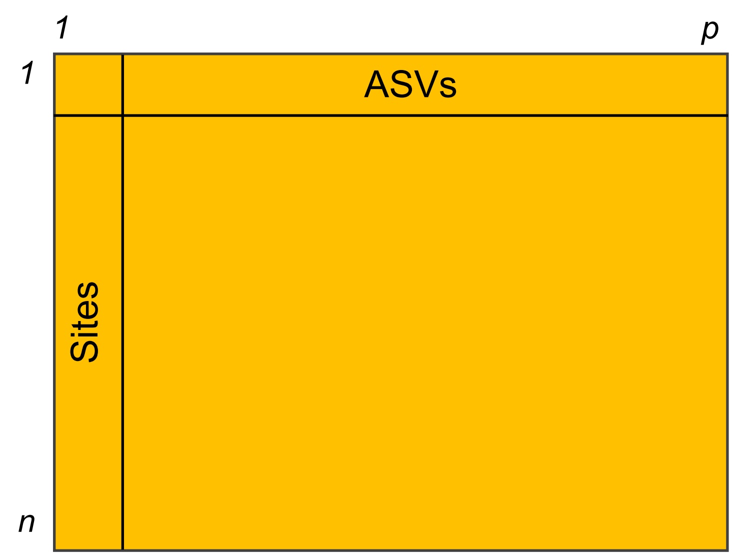
Metabarcoding specificities

Unwanted taxa

Uneven sequencing depth

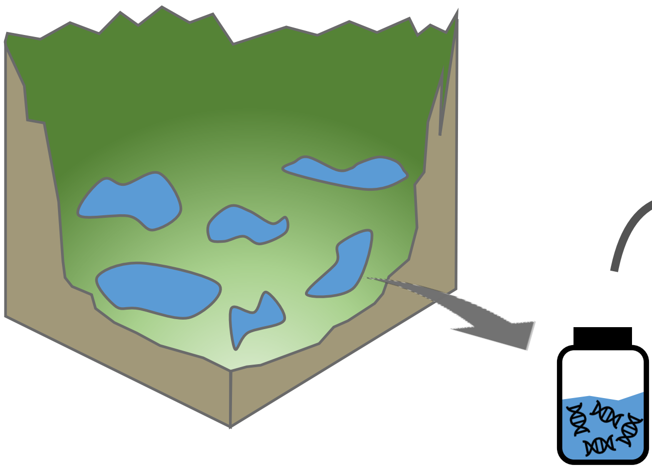
Compositionality

Sparsity

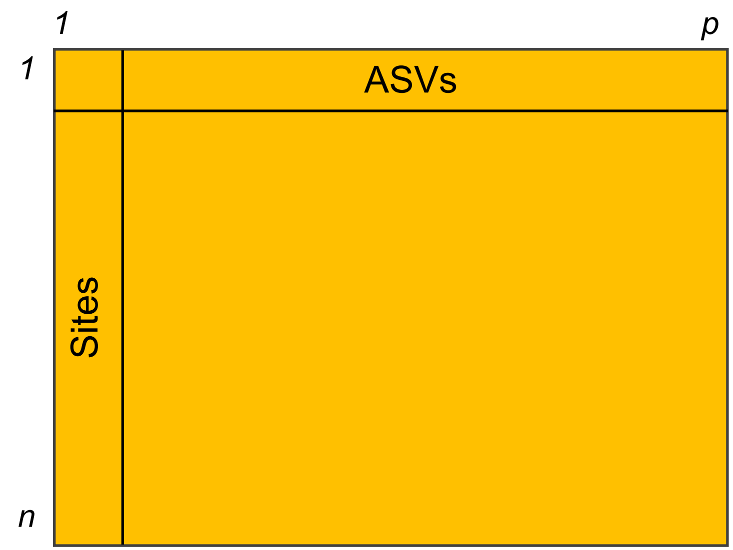


ASV table + taxonomy

Data processing, normalisation



Metabarcoding workflow



ASV table + taxonomy

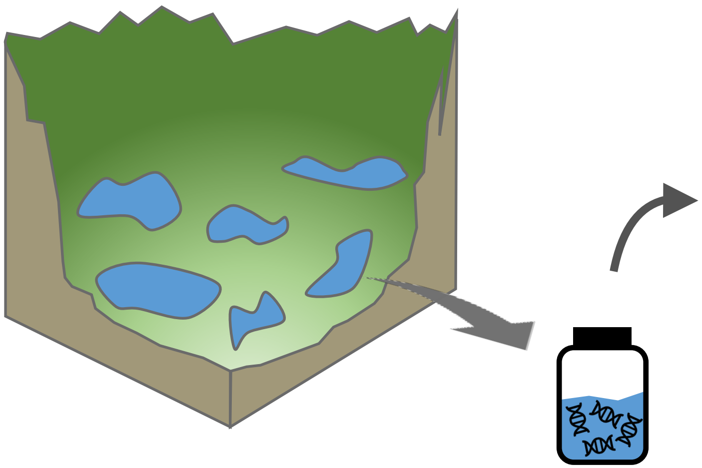
Metabarcoding specificities

→ Unwanted taxa

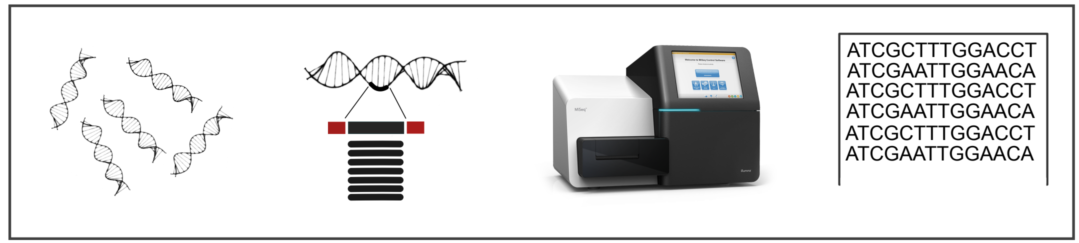
Uneven sequencing depth

Compositionality

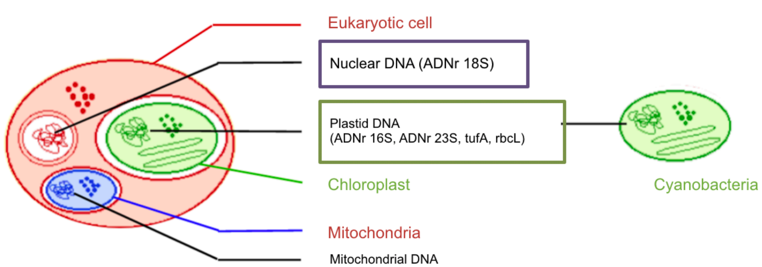
Sparsity



Metabarcoding workflow

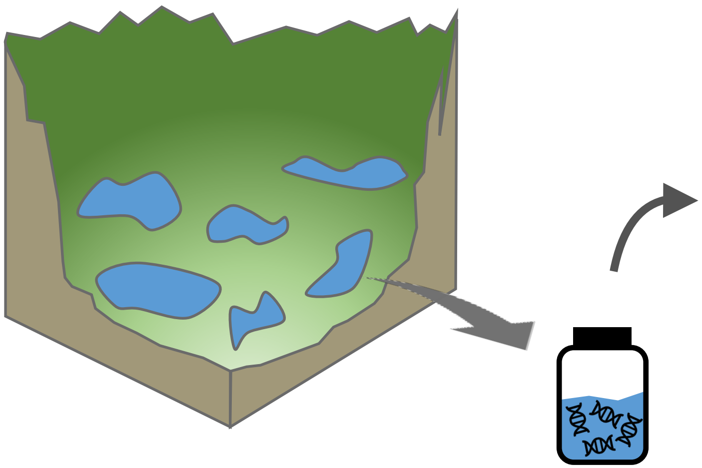


Exemple : you study bacterial communities through 16S metabarcoding



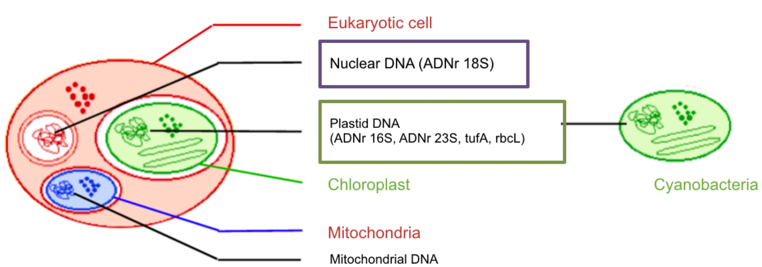
Bacteria + *Chloroplasts* *Mitochondria* *Archaea*

Data processing, normalisation



Metabarcoding workflow

Exemple : you study bacterial communities through 16S metabarcoding



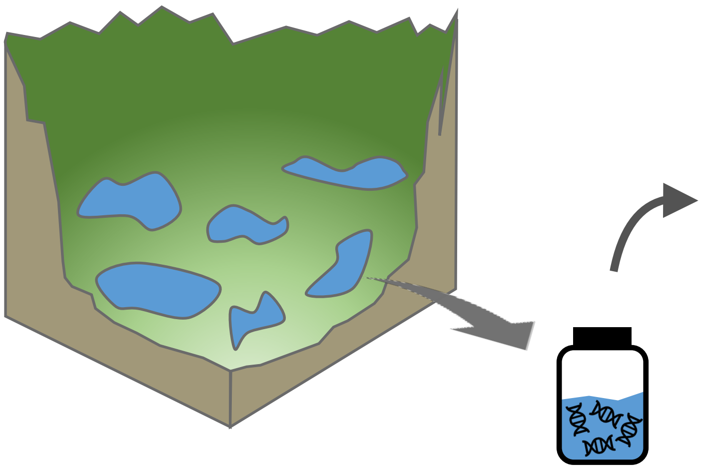
Bacteria + *Chloroplasts* *Mitochondria* *Archaea*

Presence of reads affiliated to non-targeted taxa can be due to several reasons :

Sequencing errors

Aspecific amplification

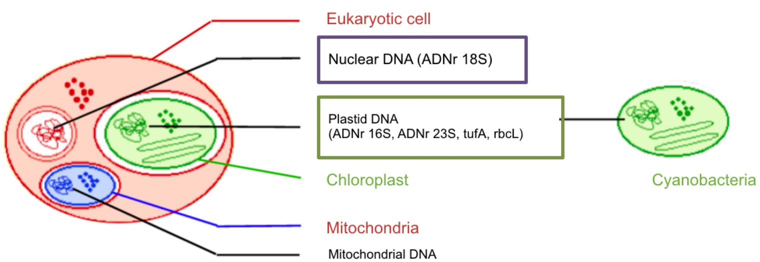
…

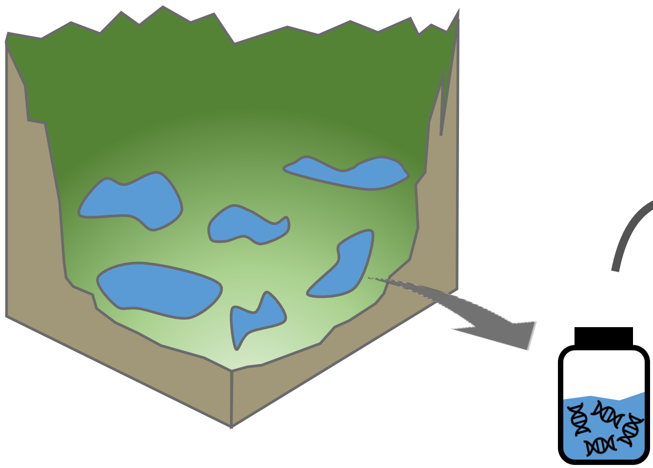


Metabarcoding workflow

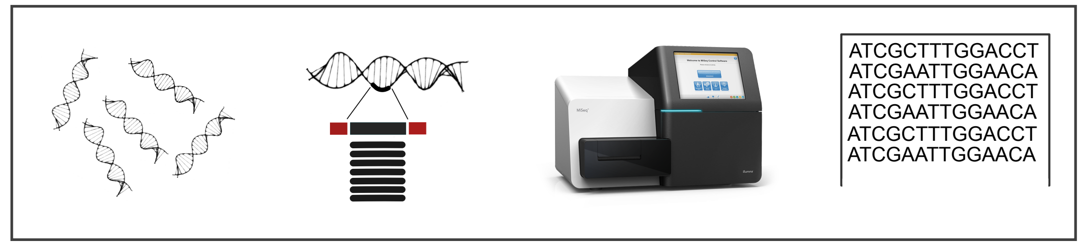
And for 23S sequencing of phytoplankton, what unwanted taxa could we get?

Let's see during the practice !





Metabarcoding workflow



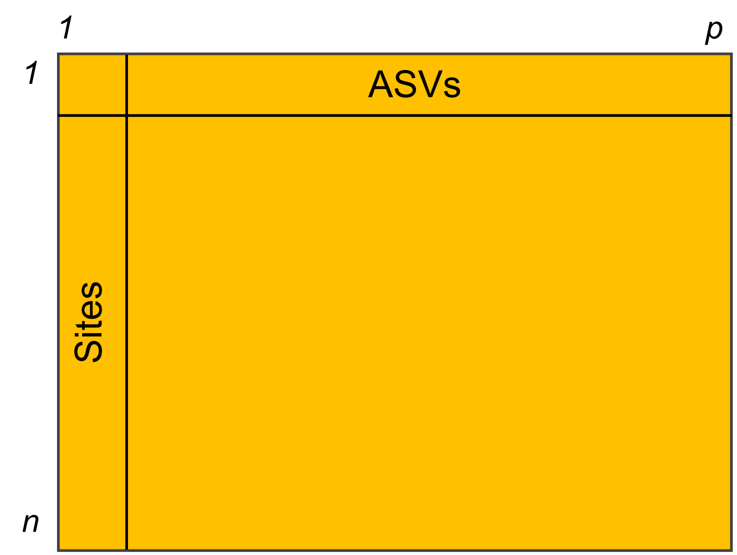
Metabarcoding specificities

Unwanted taxa

→ Uneven sequencing depth

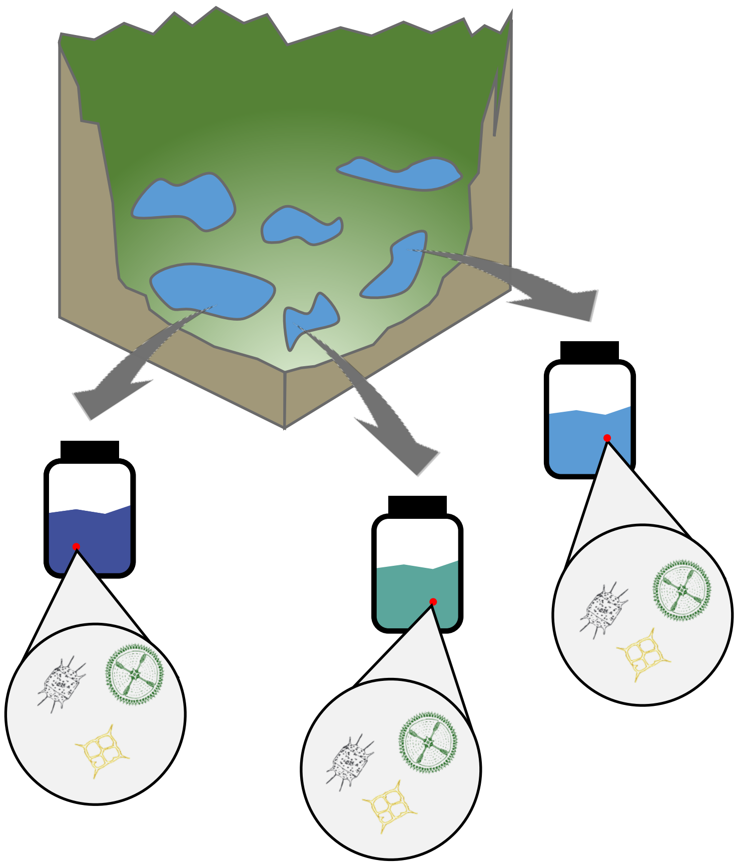
Compositionality

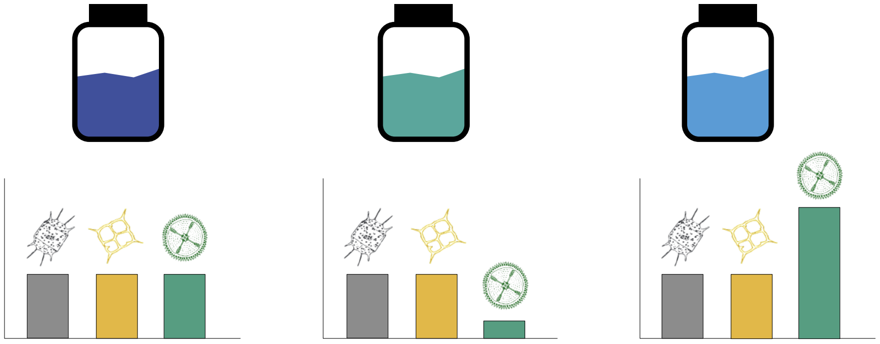
Sparsity



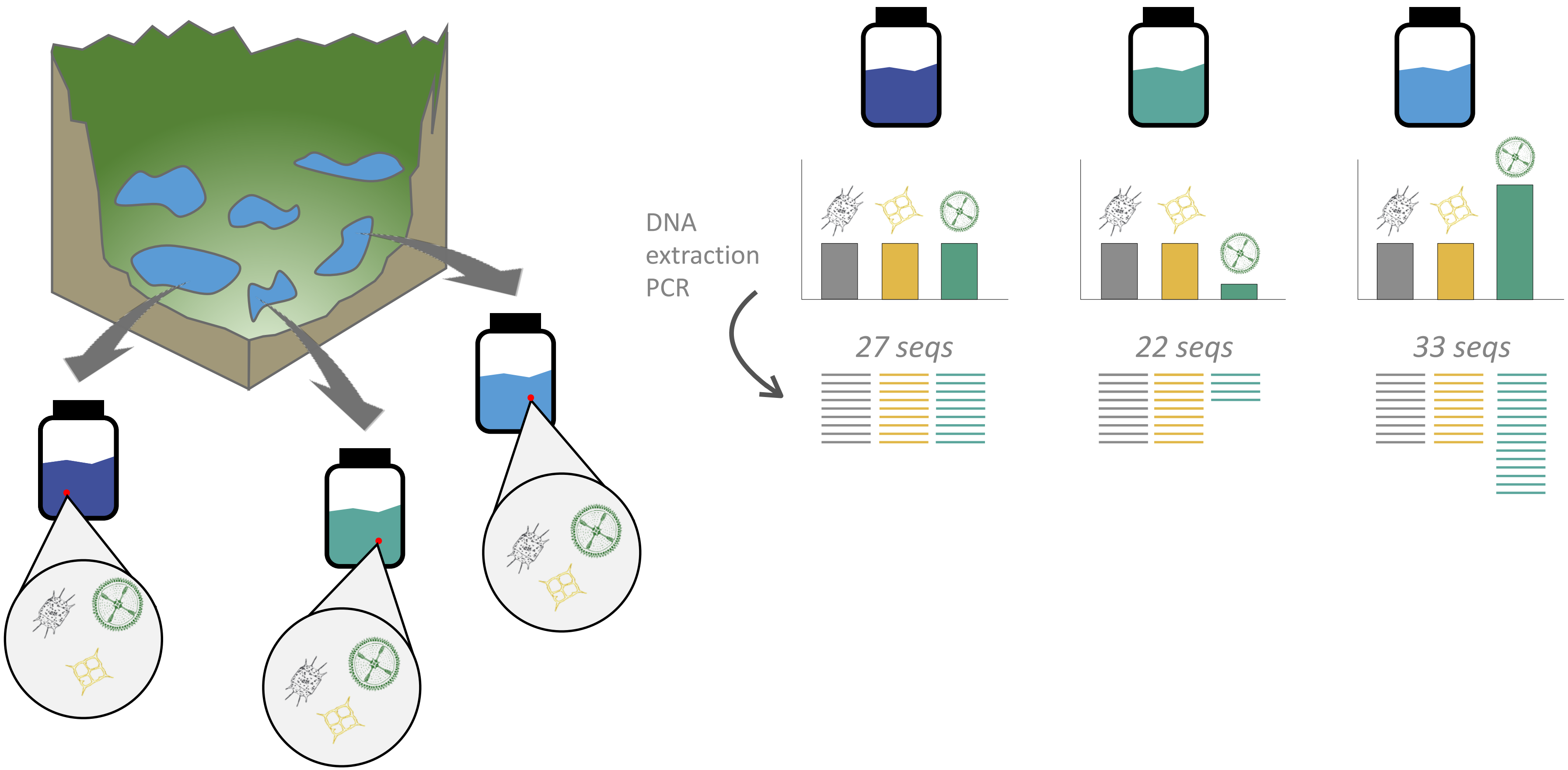
ASV table

Data processing, normalisation

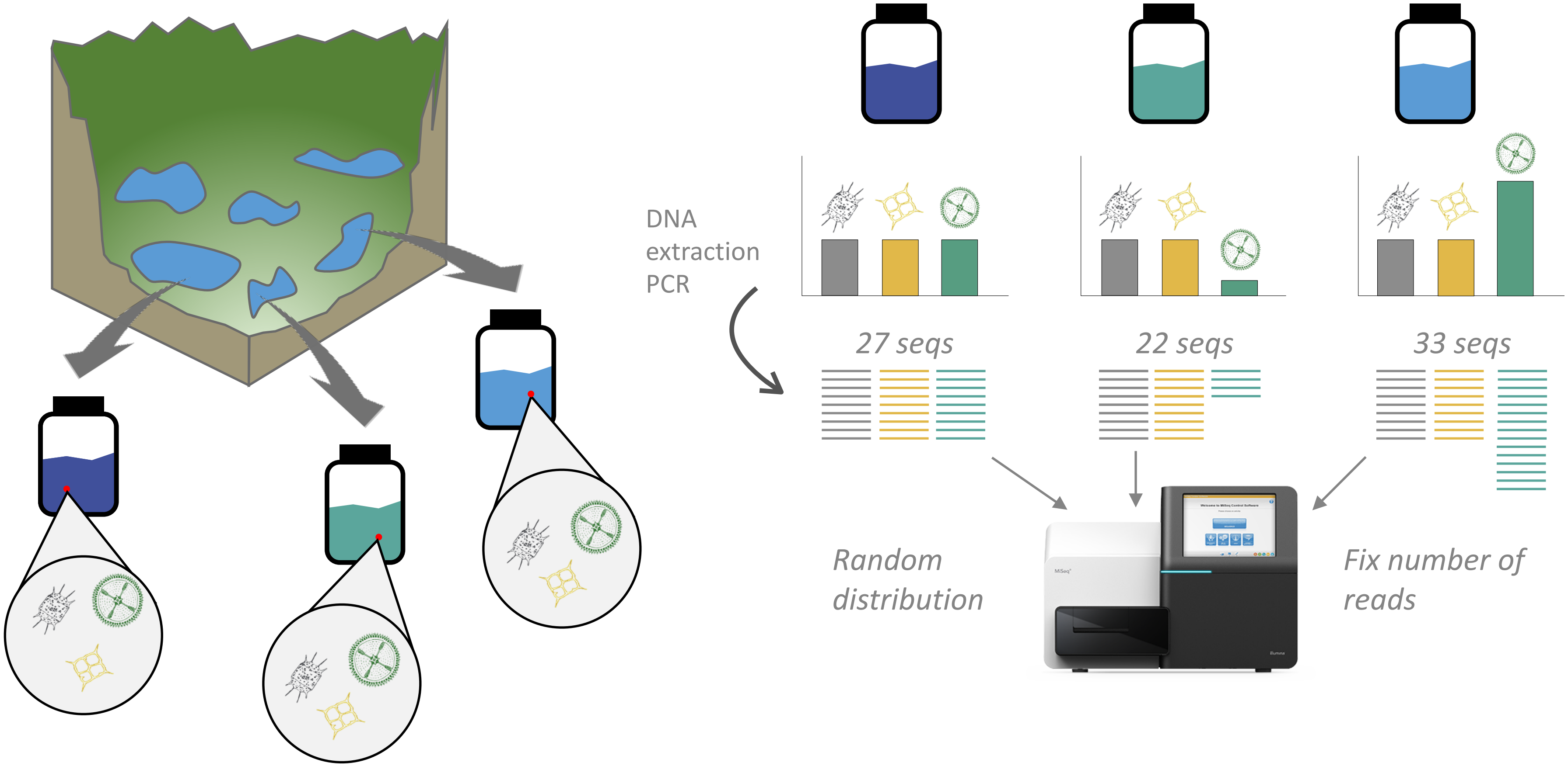




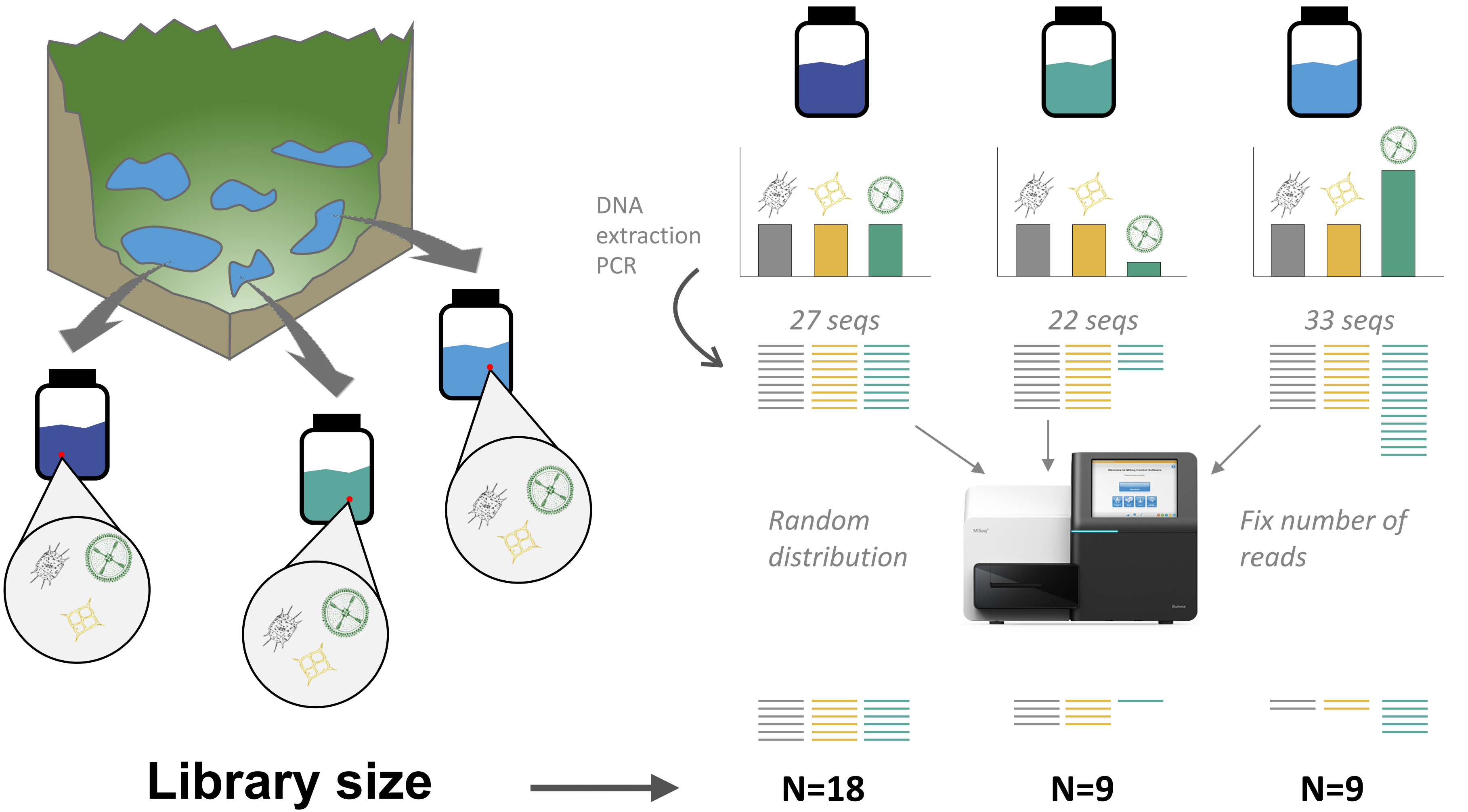
Data processing, normalisation

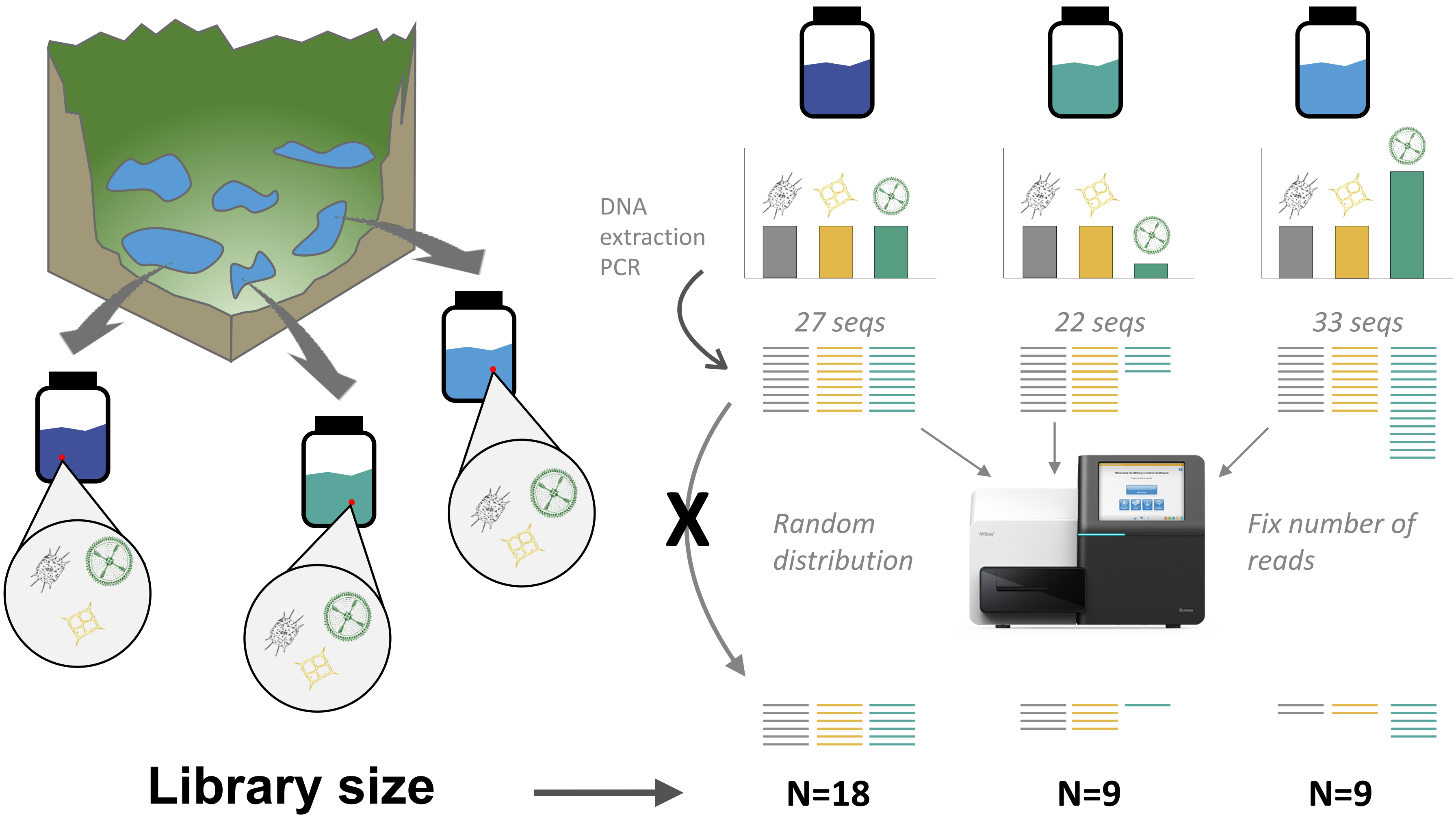


Data processing, normalisation



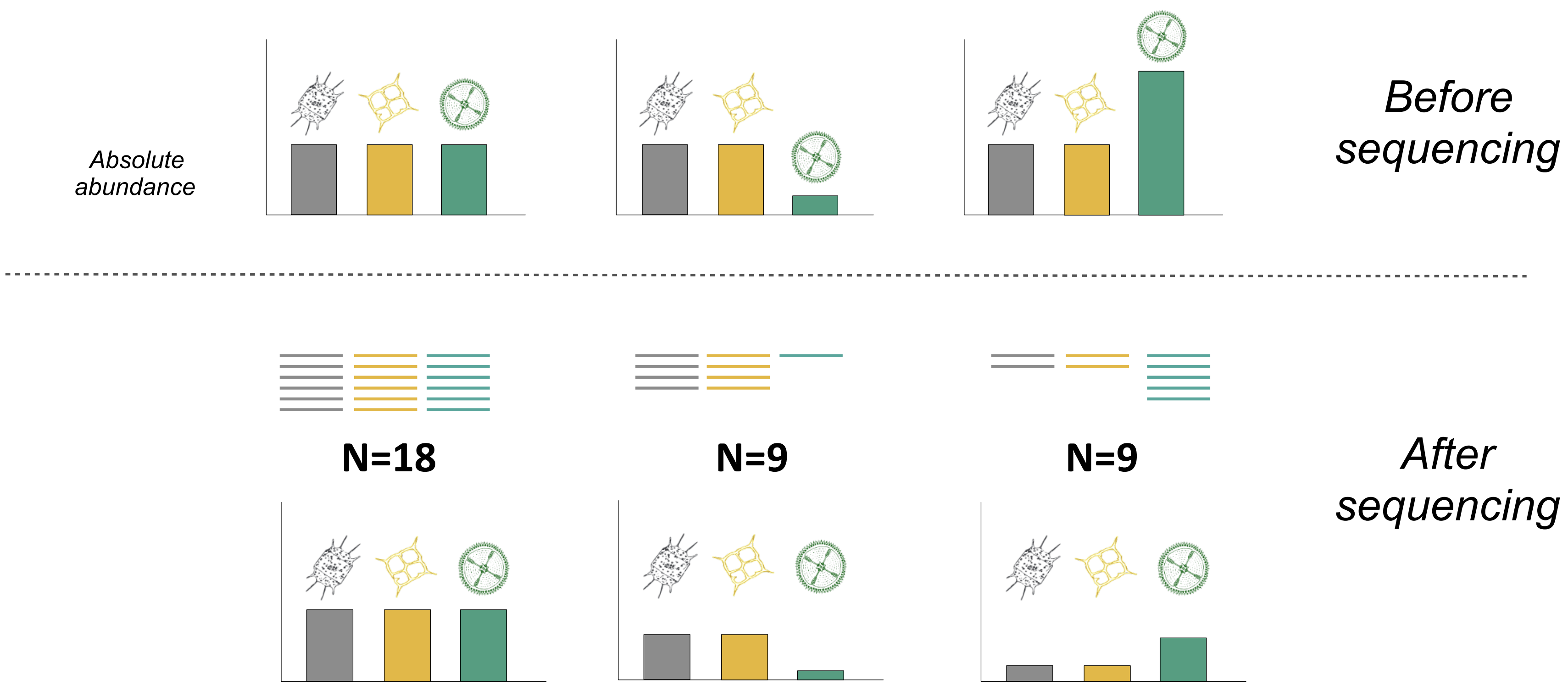
Data processing, normalisation



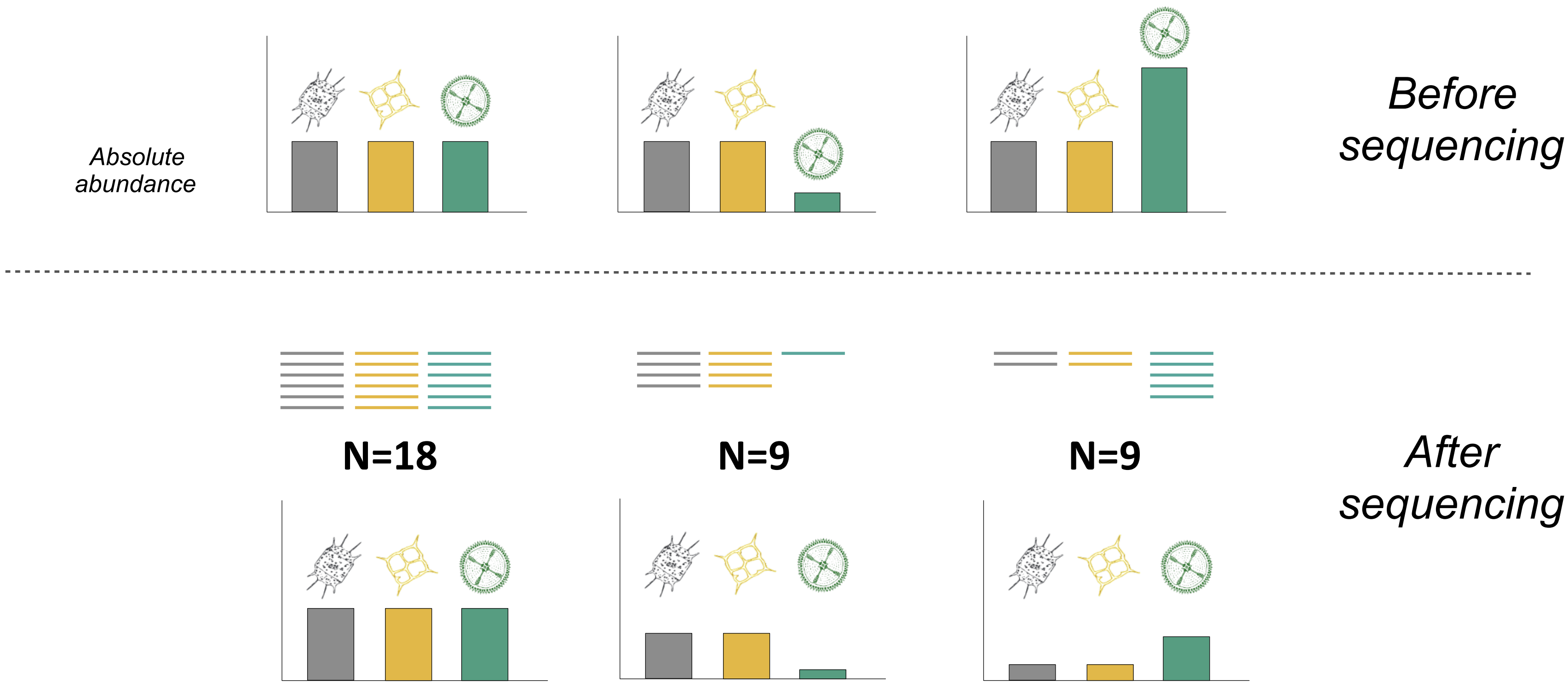


The library size is uneven due to the sequencing technology. This is not related to actual biological differences. It can change of more than 10s fold !!

Data processing, normalisation

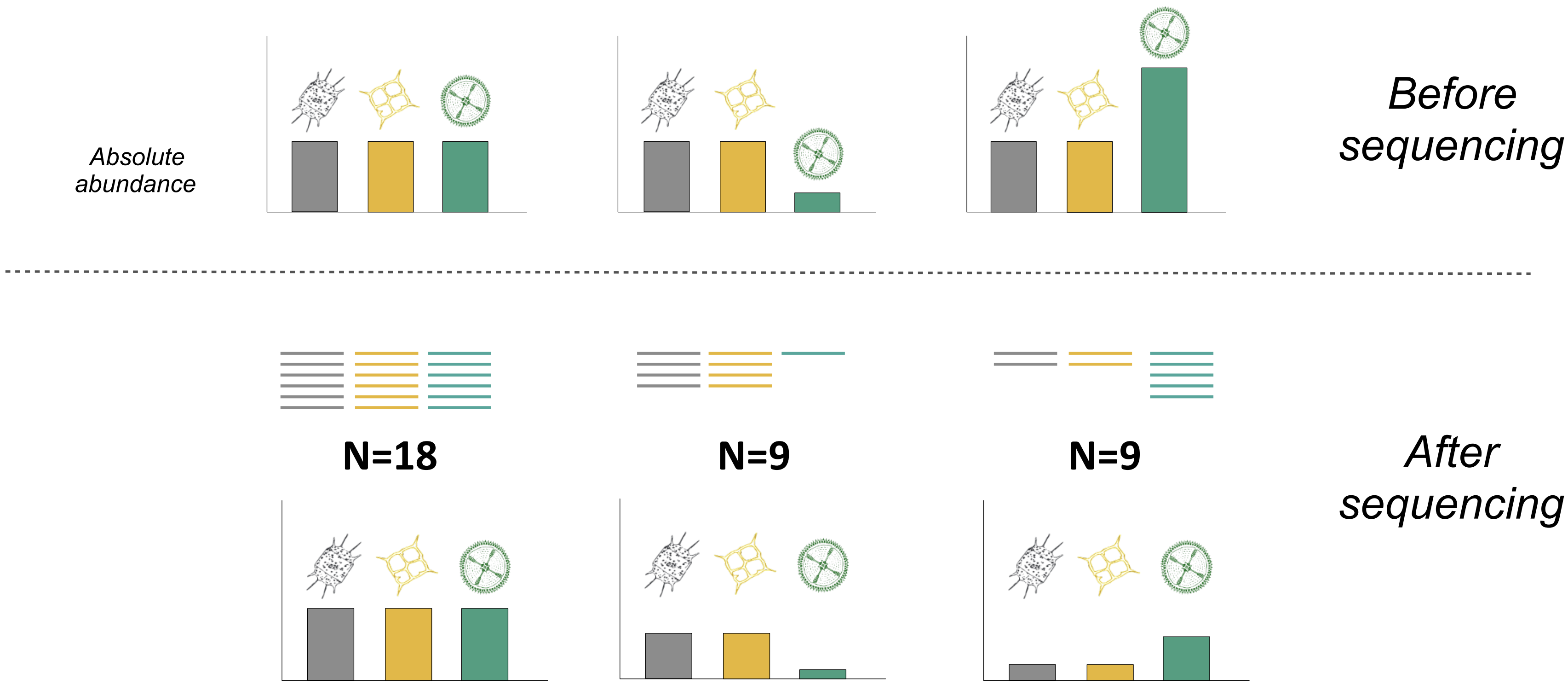


Data processing, normalisation



Samples cannot be compared directly based on read counts

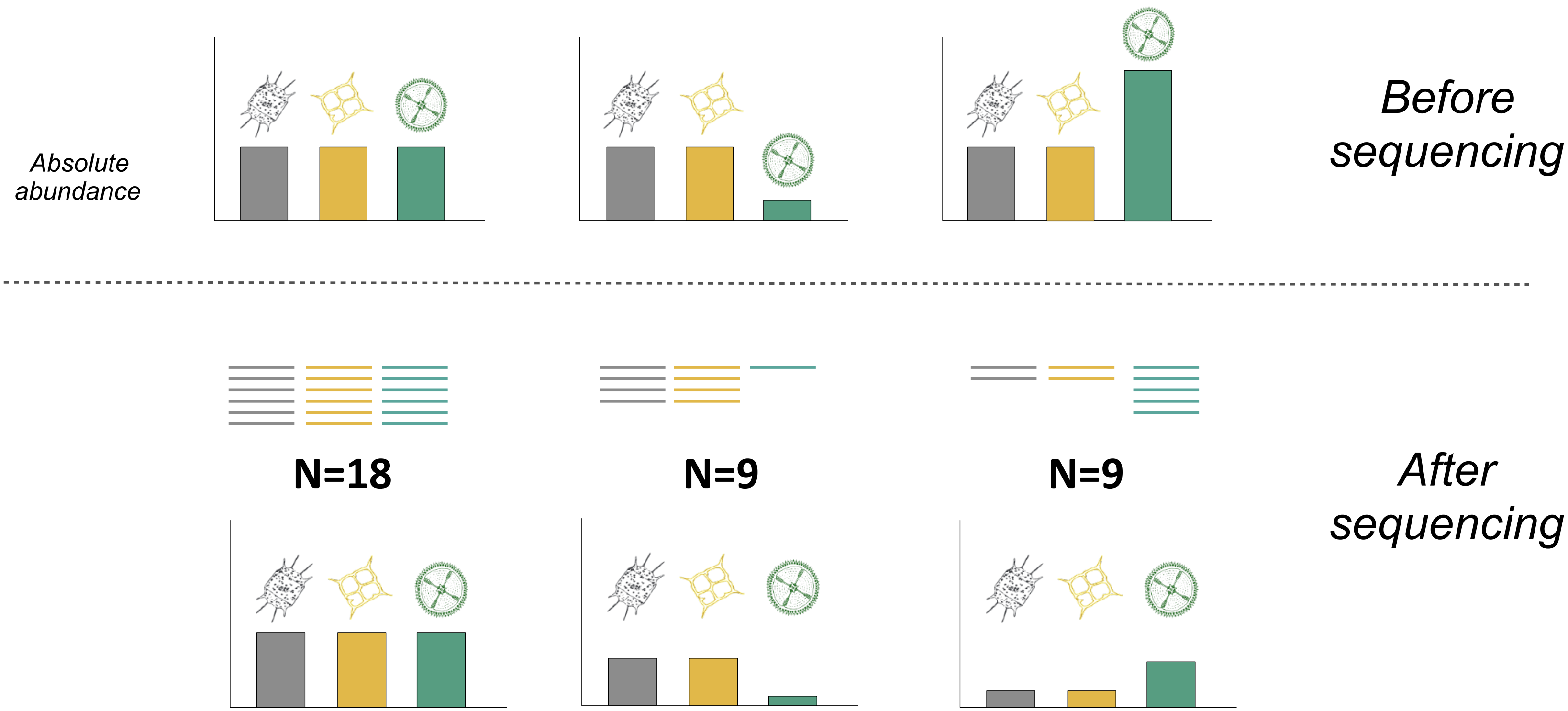
Data processing, normalisation



Samples cannot be compared directly based on read counts

The more you sequence, the more taxa you will possibly detect

Data processing, normalisation

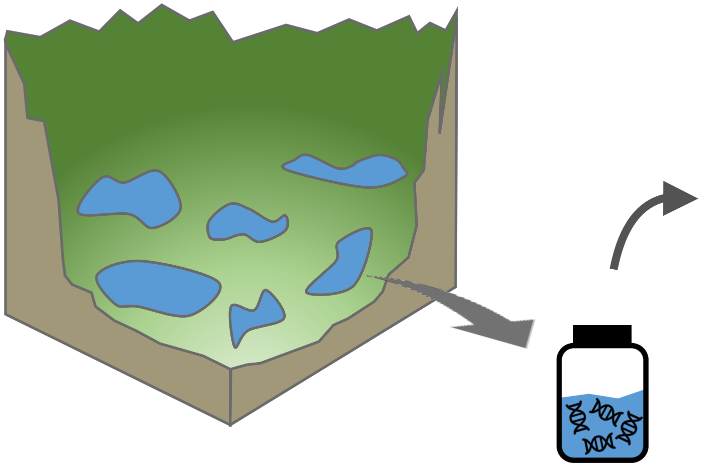


Samples cannot be compared directly based on read counts

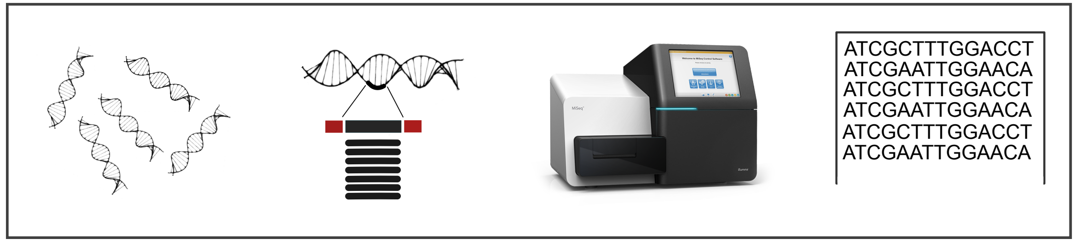
The more you sequence, the more taxa you will possibly detect

A normalisation step is mandatory before doing any ecological analysis

Data processing, normalisation



Metabarcoding workflow



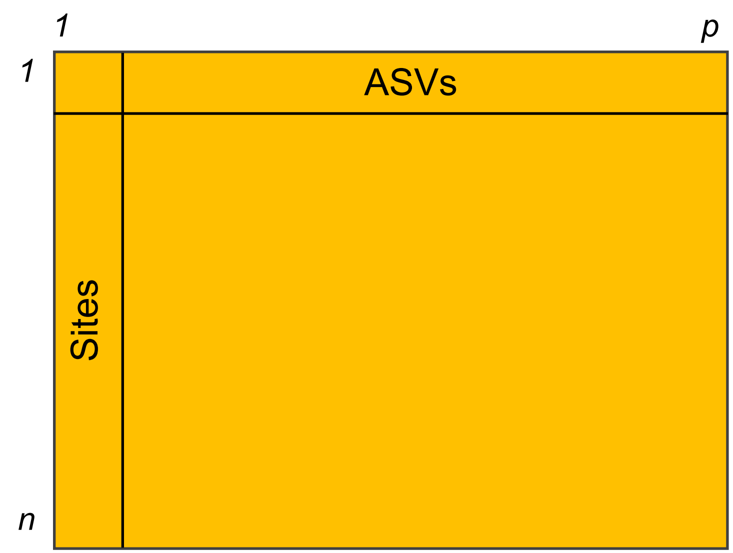
Metabarcoding specificities

Unwanted taxa

Uneven sequencing depth

→ Compositionality

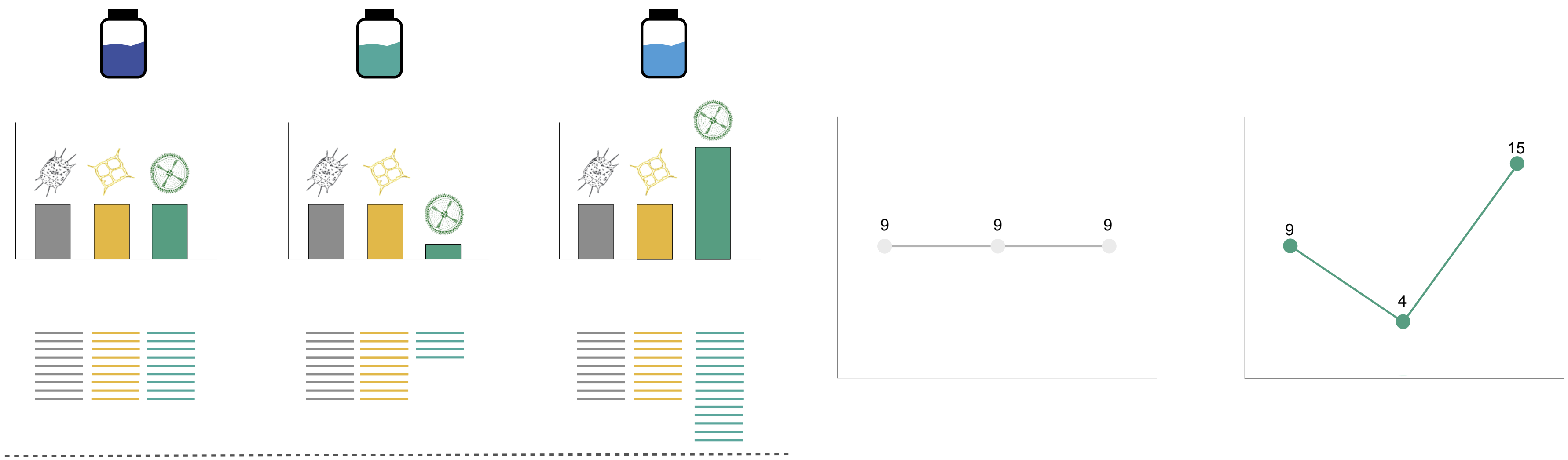
Sparsity



ASV table

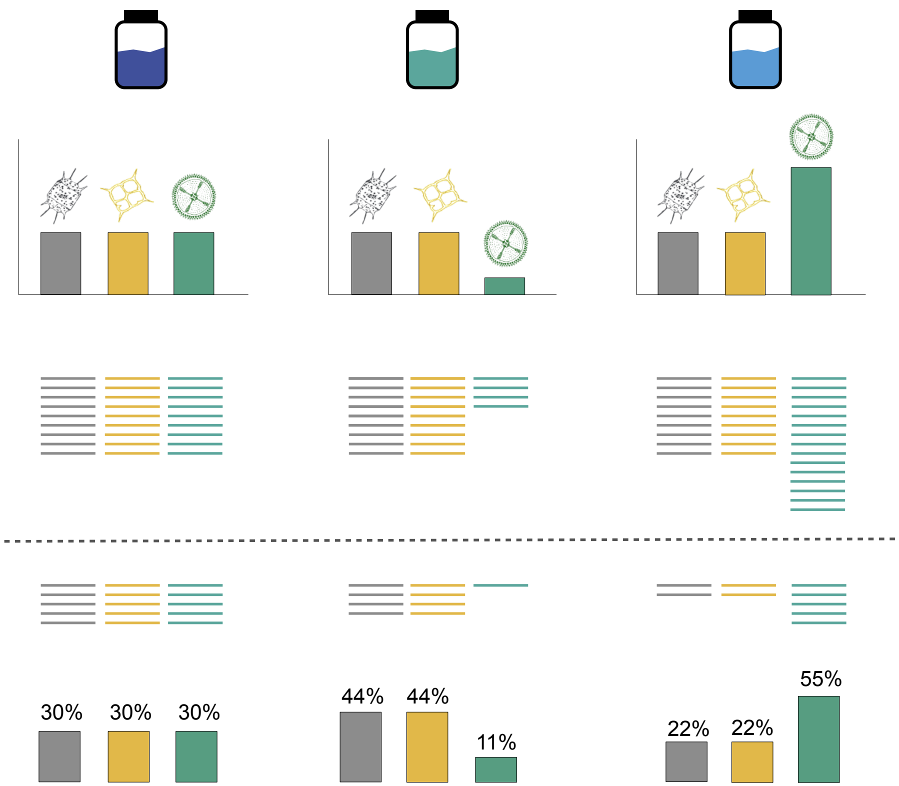
Data is compositional

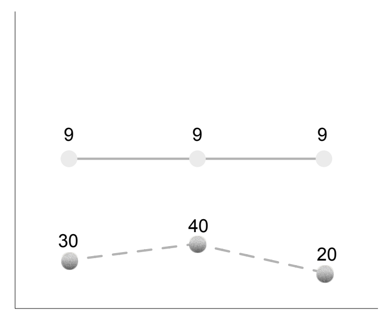
« A data set is compositional when the parts in each sample have an arbitrary or non-informative sum »

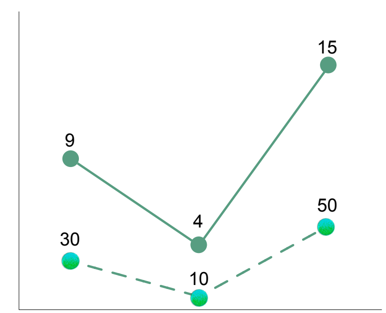


Data is compositional

« A data set is compositional when the parts in each sample have an arbitrary or non-informative sum »

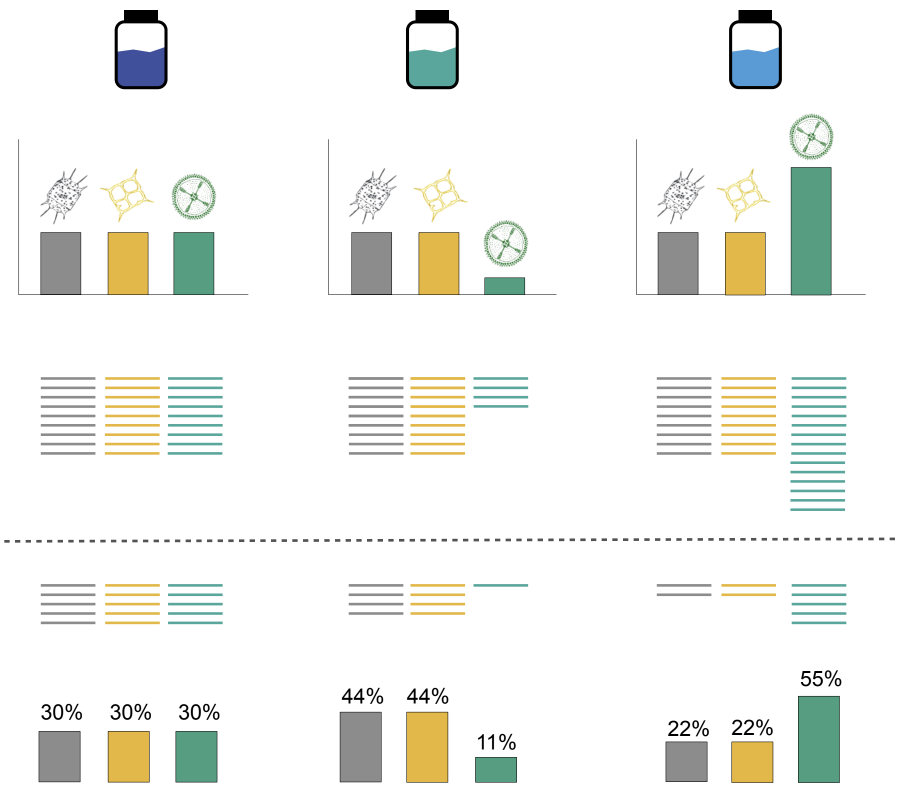


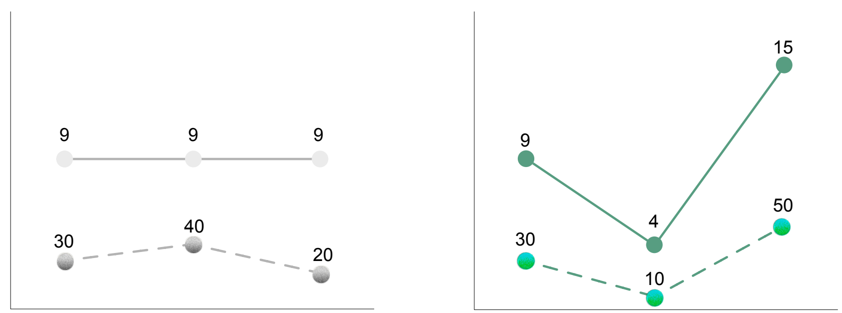




Data is compositional

« A data set is compositional when the parts in each sample have an arbitrary or non-informative sum »

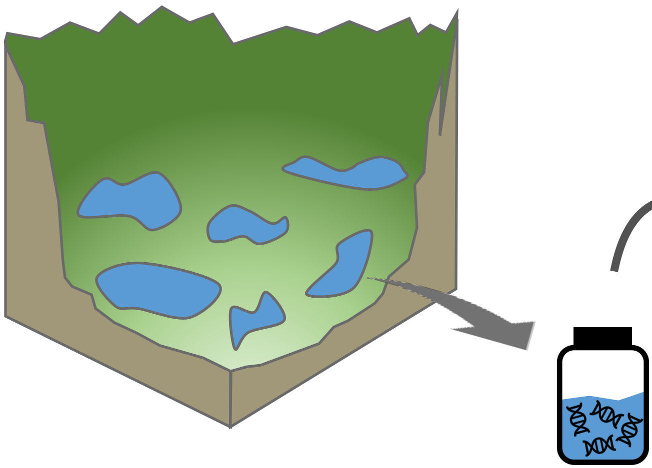




- Induce spurious correlations
- Problematic for **differential abundance** analysis
- Problematic for **network** analysis and correlations between taxas

Data processing, normalisation

BIOLAWEB



Metabarcoding workflow

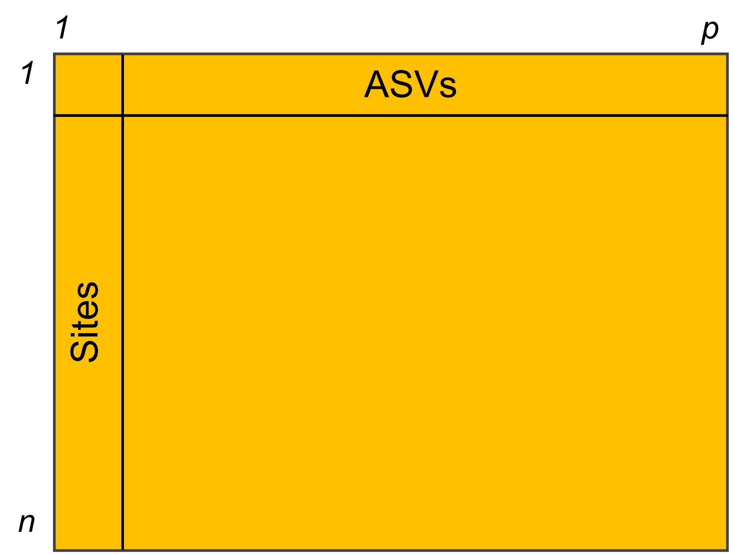
<u>Metabarcoding specificities</u>

Unwanted taxa

Uneven sequencing depth

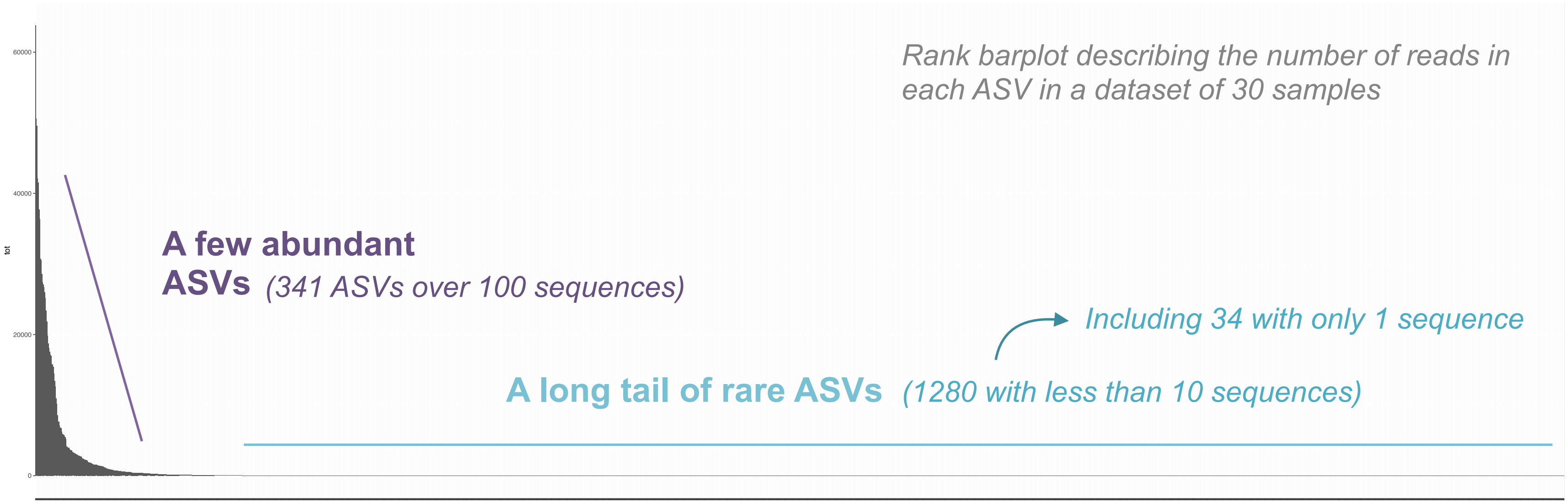
Compositionality

→ Sparsity



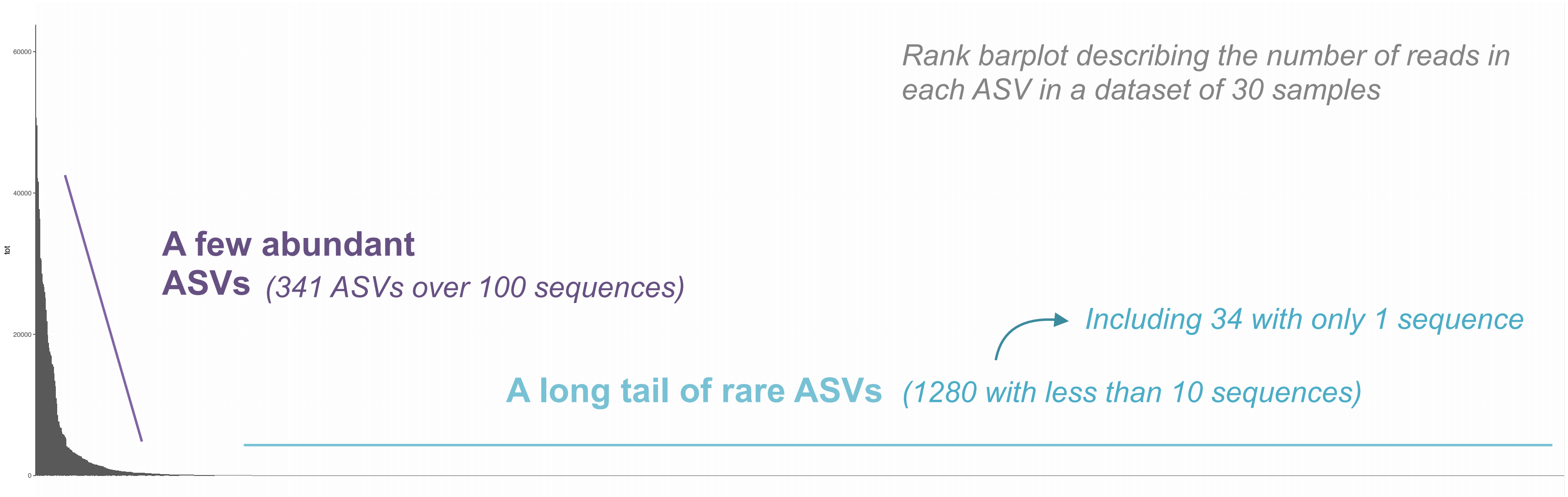
ASV table

Data is sparse



In some studies, data can be composed of more than 80% of 0s !

Data is sparse



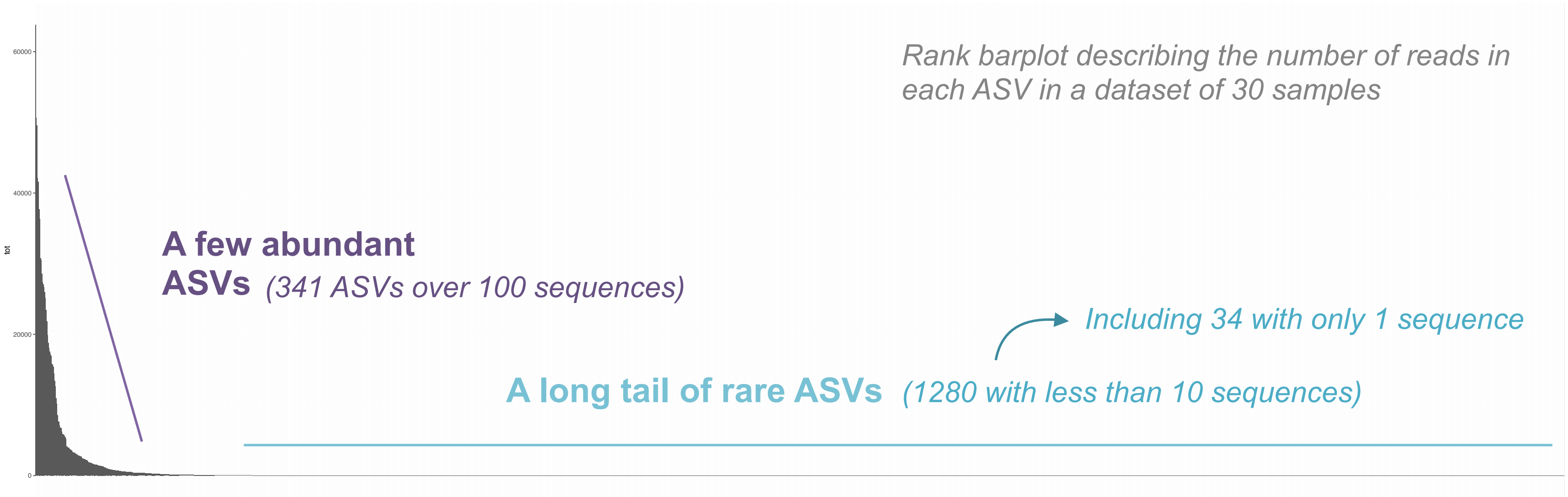
Why ?

Because metabarcoding is very sensitive and can detect rare species or variants

Because of sequencing errors

Because microbial communities are highly diverse

Data is sparse

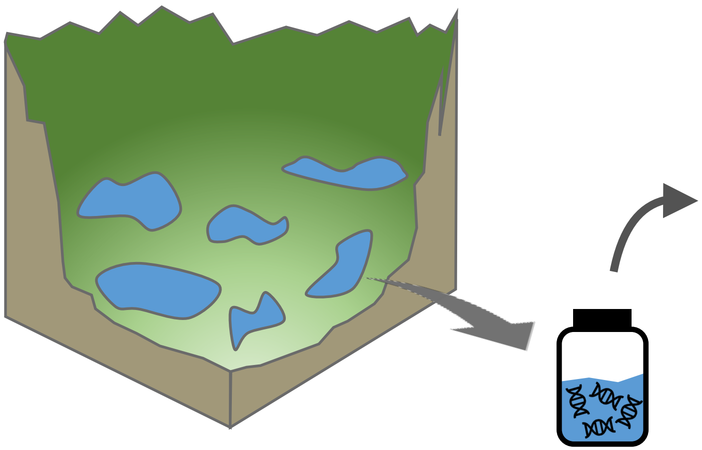


Problematics

A 0 does not mean the **absence** of a species (*e.g.* low sequencing depth).

Some statistical analysis can be impacted (due to « **double zeros** »)

It is complex to decipher true rare species from sequencing errors



Metabarcoding workflow

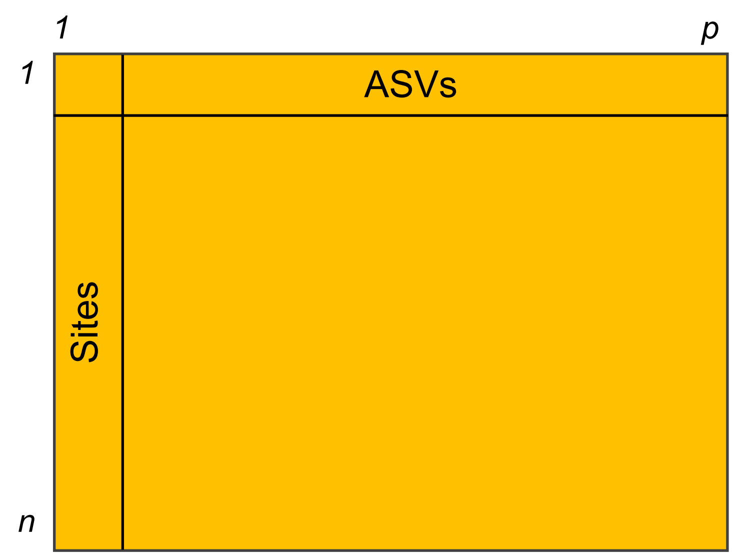
Metabarcoding specificities

Unwanted taxa

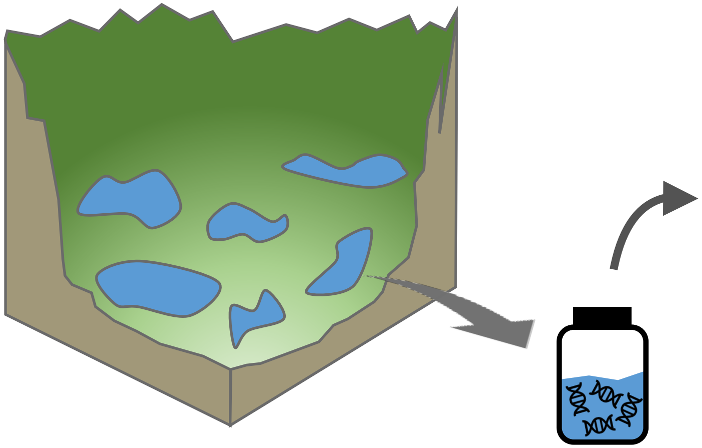
Uneven sequencing depth

Compositionality

Sparsity



ASV table

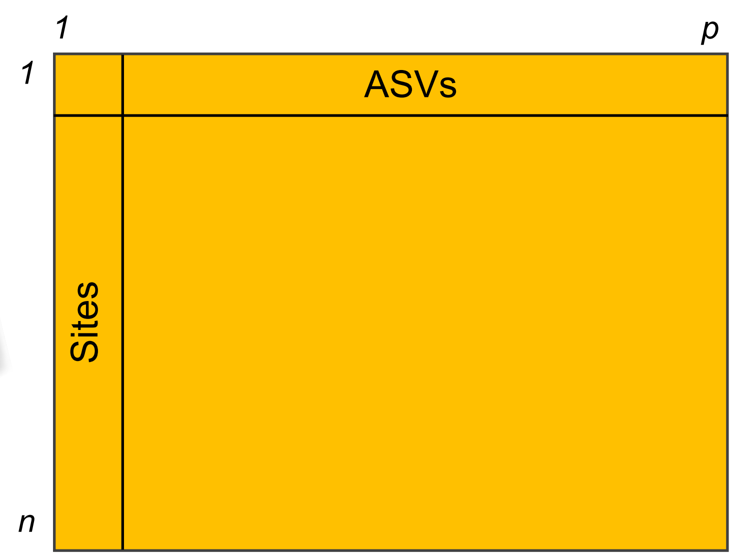


Metabarcoding workflow

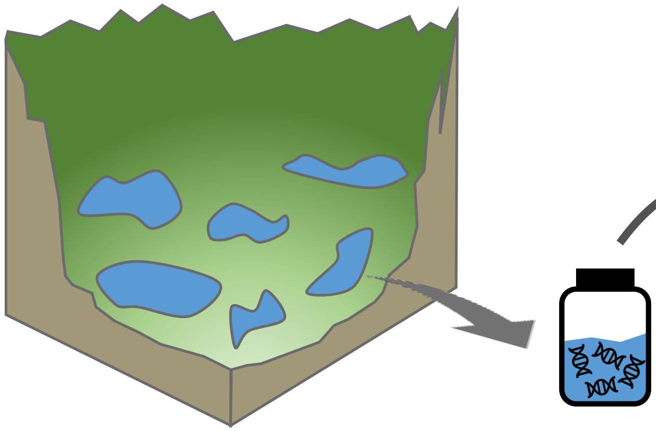
Metabarcoding sp

Metabarcoding data needs to be normalized and prepared before doing any ecological analysis

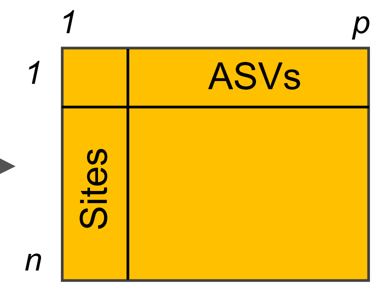
Sparcity



ASV table



Metabarcoding workflow

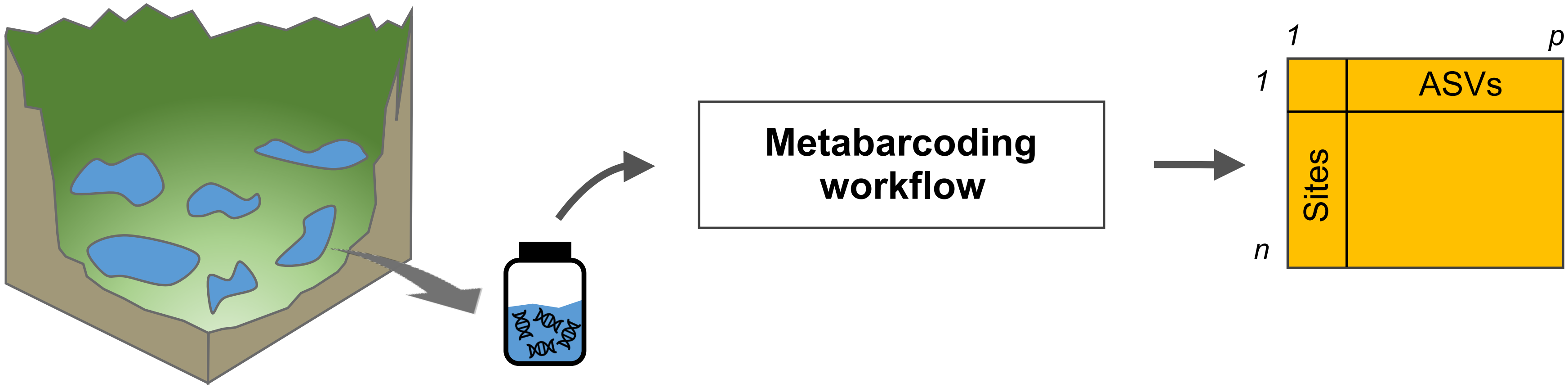


1

Filter ASVs based on taxonomy

Keep only ASVs assigned to the targeted organisms

(e.g. Diatoms)



1

Filter ASVs based on taxonomy

Keep only ASVs assigned to the targeted organisms

(e.g. Diatoms)

2

Filter ASVs that are not abundant

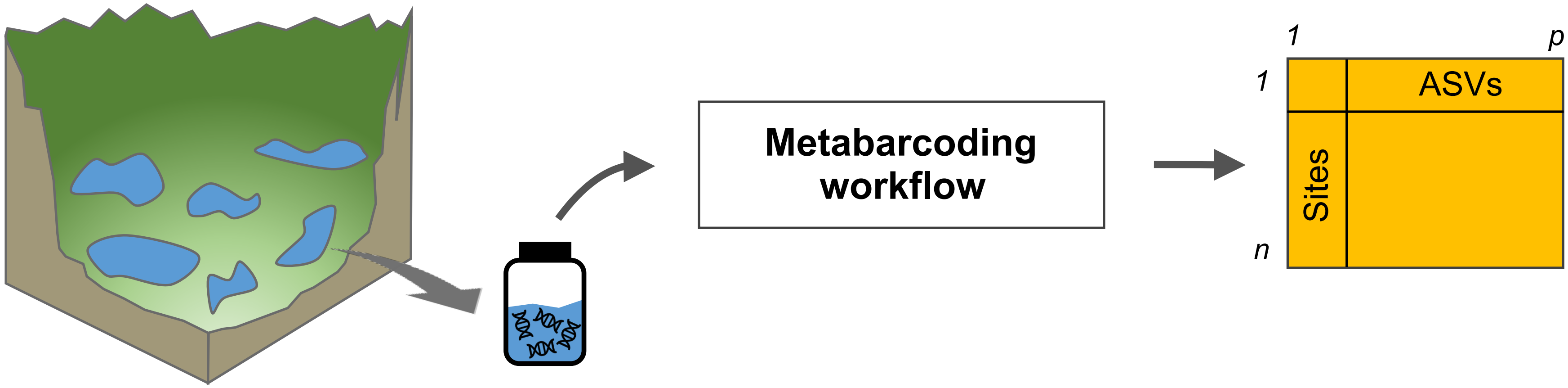
(optional)

Singletons

That have less than *x* sequences

That occurs in less than *x* samples

Data processing, normalisation



1

Filter ASVs based on taxonomy

Keep only ASVs assigned to the targeted organisms

(e.g. Diatoms)

2

Filter ASVs that are not abundant

(optional)

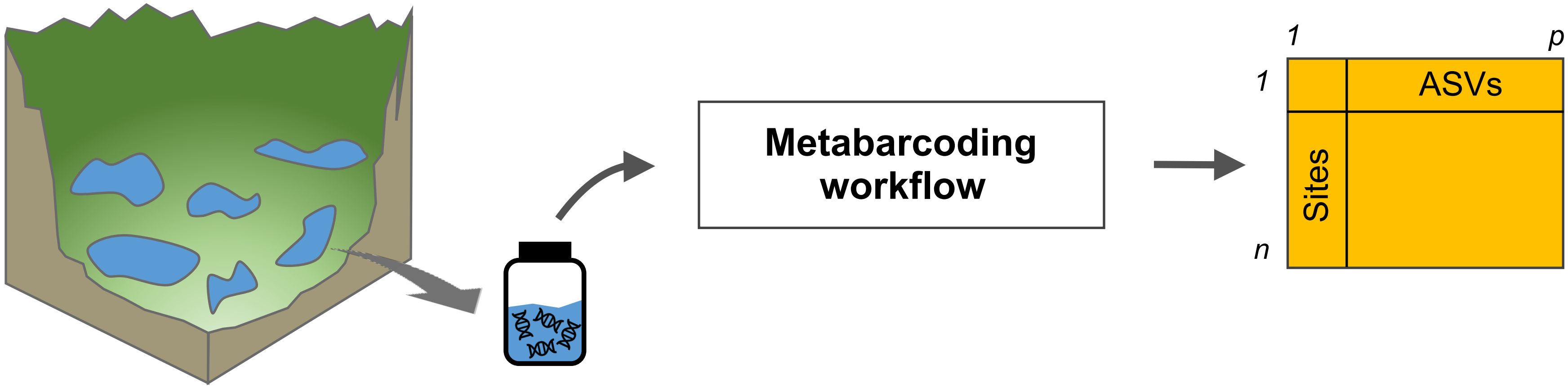
Singletons

That have less than *x* sequences

That occurs in less than *x* samples

Limit sparsity and sequencing errors

Data processing, normalisation



1

Filter ASVs based on taxonomy

Keep only ASVs assigned to the targeted organisms

(e.g. Diatoms)

2

Filter ASVs that are not abundant

(optional)

Singletons

That have less than *x* sequences

That occurs in less than *x* samples

3

Normalize/transform

Rarefaction

Scaling

Log ratio

…

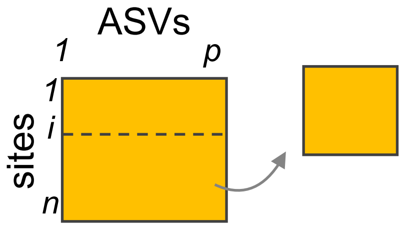
Limit sparsity and sequencing errors

Deal with library size and/or compositionallity

Different strategies exist to deal with uneven sequencing depth in metabarcoding data

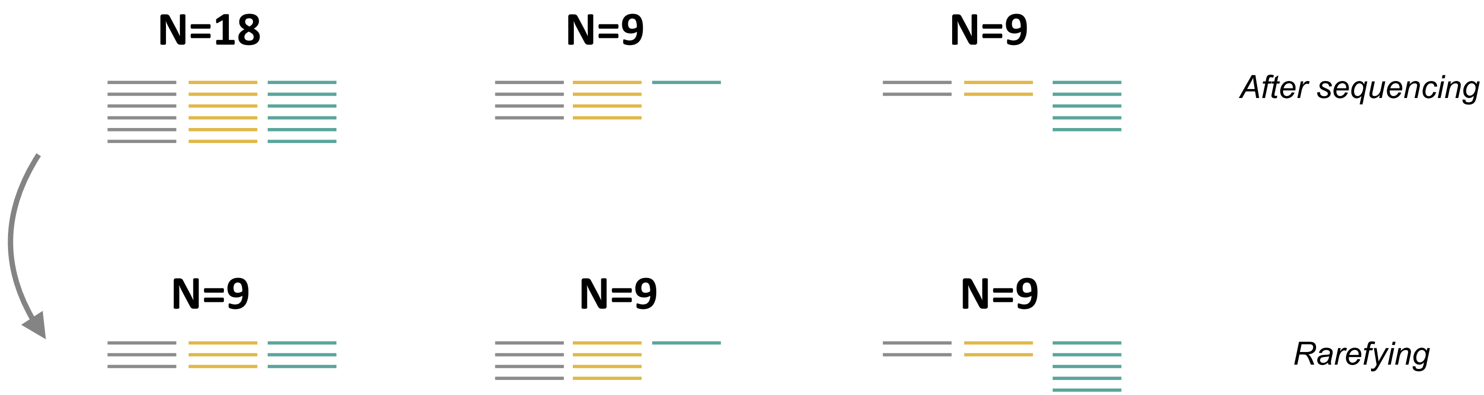
1

Rarefy so that all samples present the same number of reads

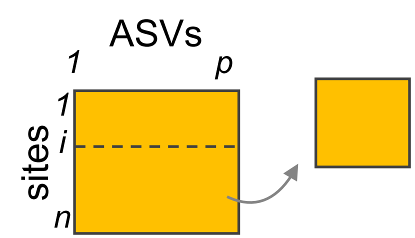


Data processing, normalisation

1. Rarefying and rarefaction



Rarefying : randomly subsample read counts of each sample to a common read depth (often taken from the smallest sample)



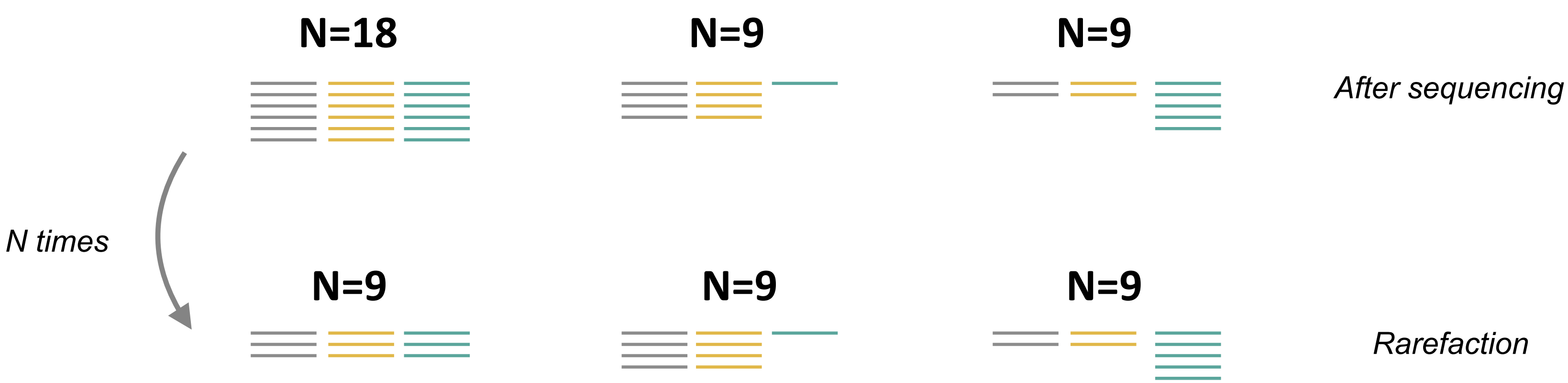

```
rrarefy(ASV_table, N)
```

vegan R package

McMurdie and Holmes, 2014

Data processing, normalisation

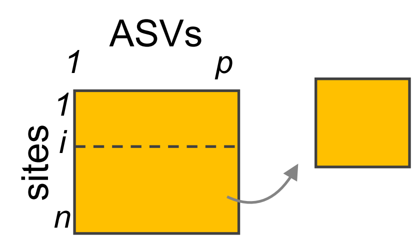
1. Rarefying and rarefaction



Rarefying : randomly subsample read counts of each sample to a common read depth (often taken from the smallest sample)

≠

Rarefaction : repeat the subsampling a high number of time (e.g. 100, 1000 times) and calculate the mean of the alpha or beta diversity metrics (more robust !!).



`rarefy(ASV_table, N)`

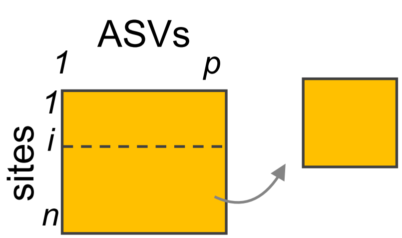
vegan R package

Schloss, 2023

Different strategies exist to deal with uneven sequencing depth in metabarcoding data

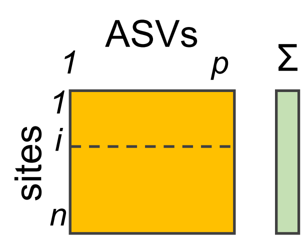
1

Rarefy so that all samples present the same number of reads

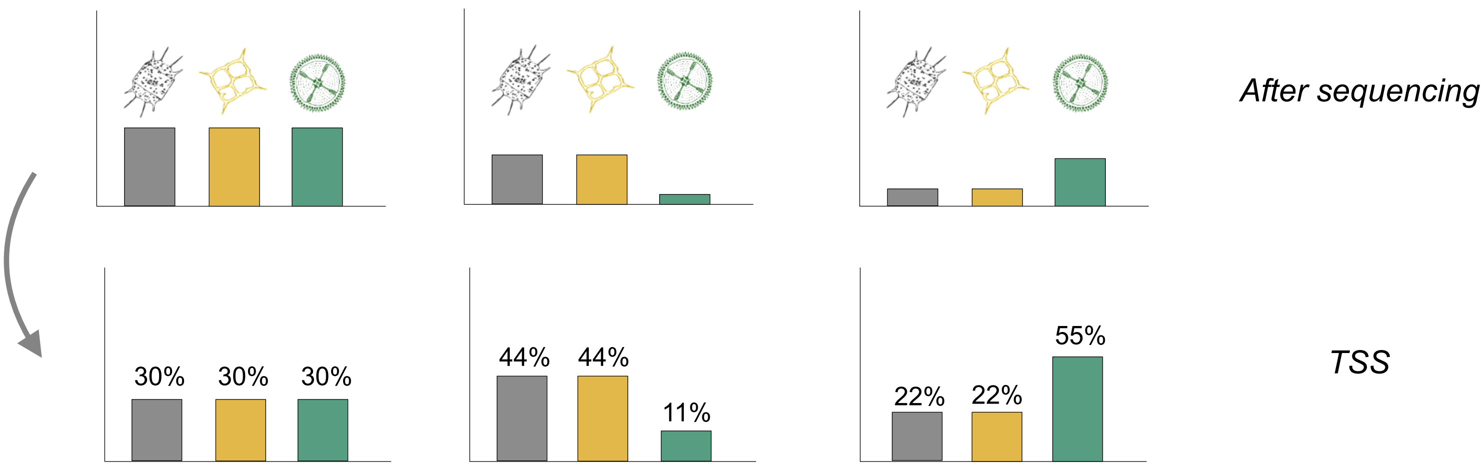


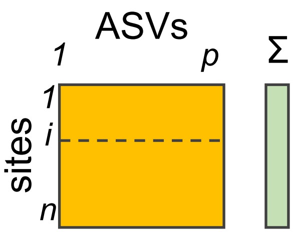
2

Re-scale the read counts in each sample (*i.e.* TSS, CSS)



2. Total sum scaling



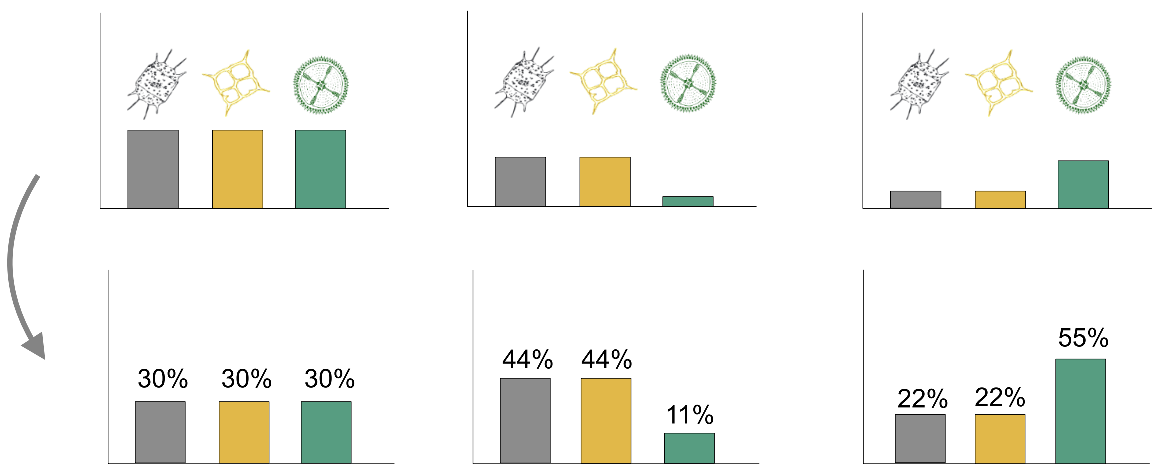


Divide ASVs read counts by the total number of reads in each samples

All values are summed up to 100%

```
ASV_table/rowSums(ASV_table)
```

3. Cumulative Sum Scaling (metagenomeSeq)

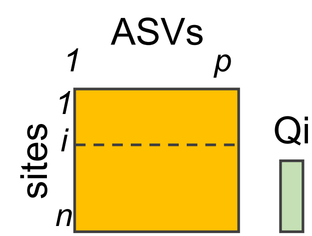


After sequencing

CSS

Re-scales the samples by using a subset of lower abundant taxa (quantile), thereby excluding the impact of highly abundant taxa

Number of reads in each samples are kept different




```
cumNorm(ASV_table,
p=cumNormStatFast(ASV_table))
```

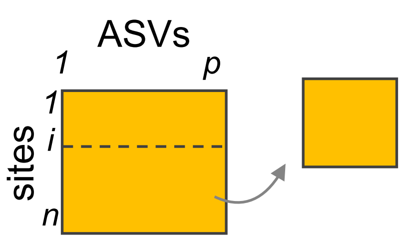
Paulsen et al., 2013

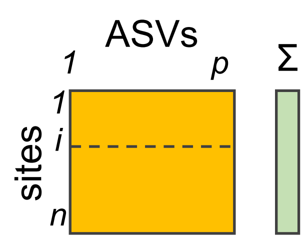
Data processing, normalisation

Different strategies exist to deal with uneven sequencing depth in metabarcoding data

1. Rarefy so that all samples present the same number of reads

2. Re-scale the read counts in each sample (*i.e.* TSS, CSS)



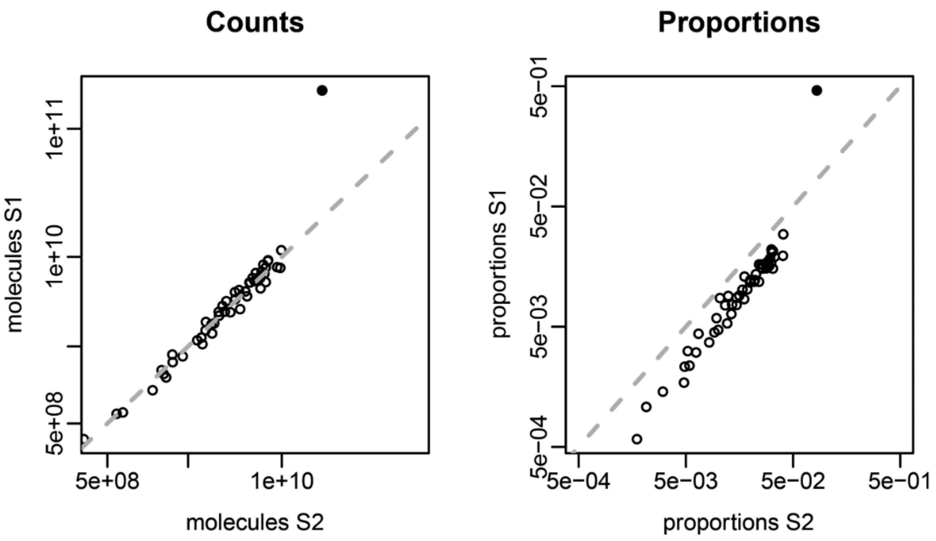


Deal with uneven sequencing depth

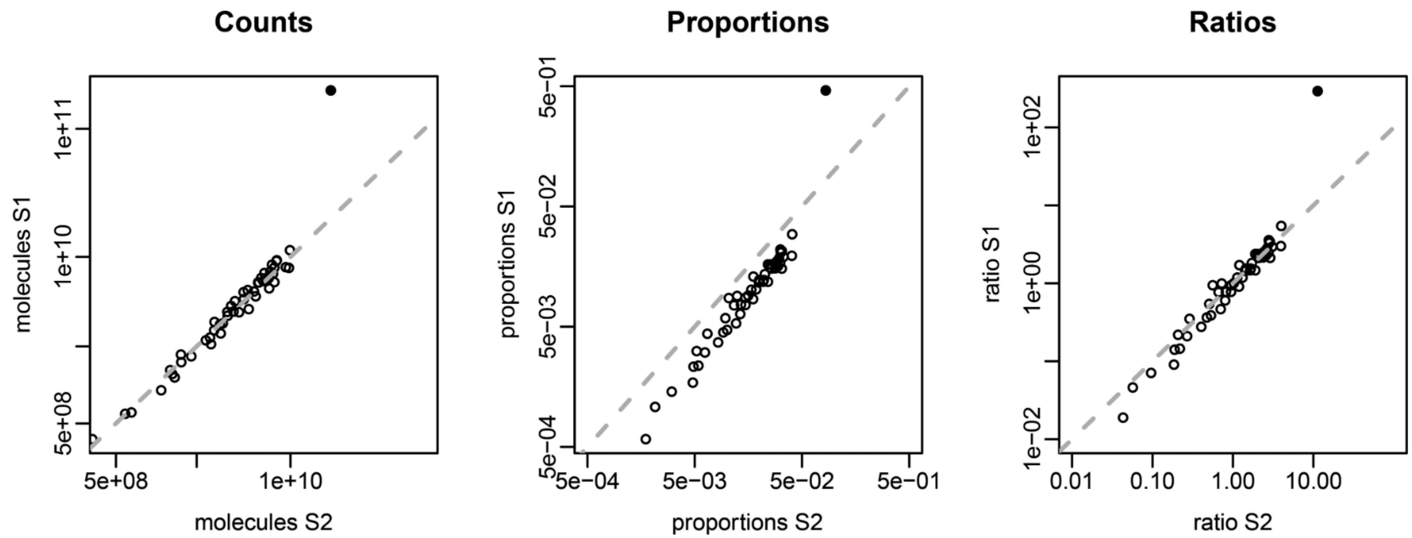
~~Compositionality~~

~~Sparsity~~

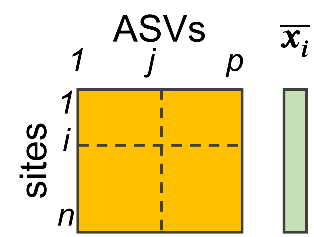
One main strategy to deal with compositional data : use ratios



One main strategy to deal with compositional data : use ratios



Ratios are conserved, **regardless of the library size**



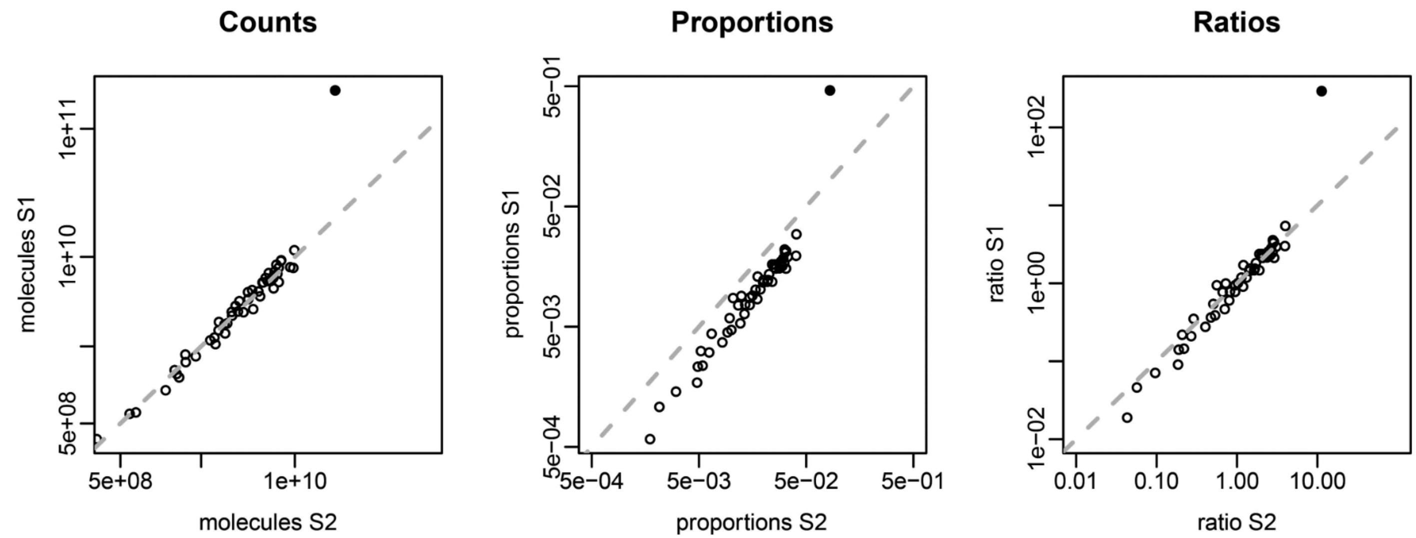
For metabarcoding data, the most widely used transformation is the **center log ratio**

```
decostand(ASV_table, "clr")
```

vegan R package

Data processing, normalisation

One main strategy to deal with compositional data : use ratios



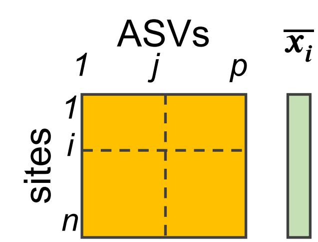
BUT, log(0) = infinity

And metabarcoding are sparse !

Different solutions :

Either remove all 0s (not recommended)

Re-estimate their abundance using pseudocounts




```
decostand(ASV_table, "clr")
```

vegan R package

Different strategies exist to deal with uneven sequencing depth in metabarcoding data

1

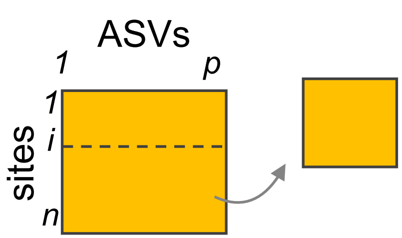
Rarefy so that all samples present the same number of reads

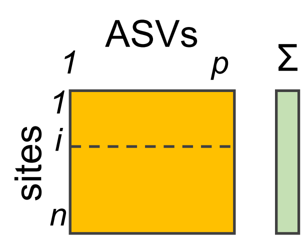
2

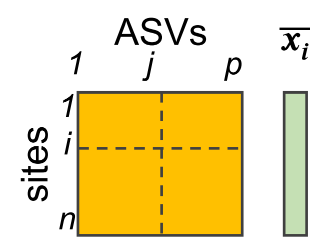
Re-scale the read counts in each sample (*i.e.* TSS, CSS)

3

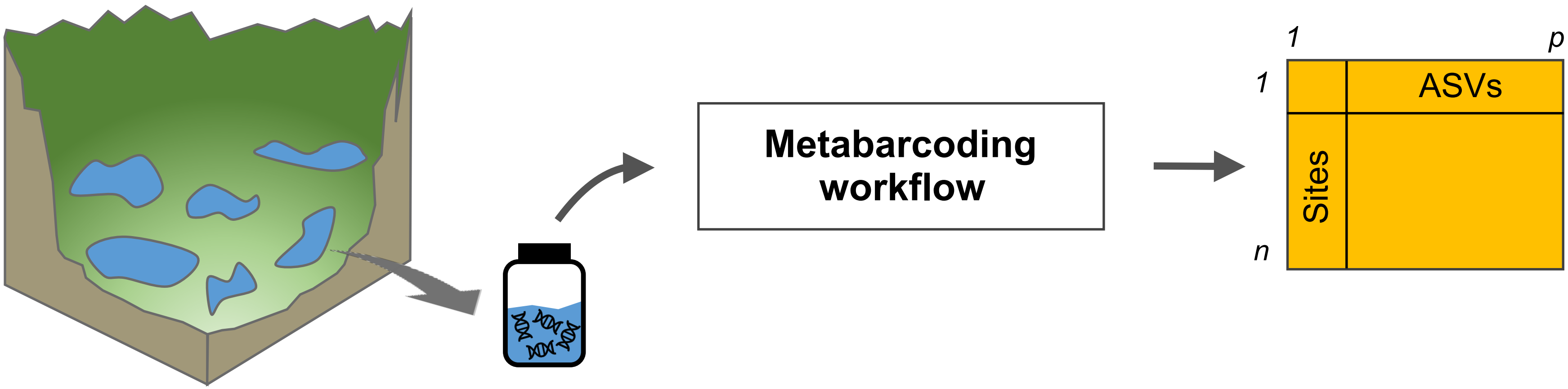
Do Log ratio transformation







Data processing, normalisation



1

Filter ASVs that have wrong taxonomic assignation

Keep only ASVs assigned to the targeted organisms

(e.g. Diatoms)

2

Filter ASVs that are not abundant

Singletons

That have less than *x* sequences

That occurs in less than *x* samples

3

Normalize/transform

ction

Which one?

Log ratio

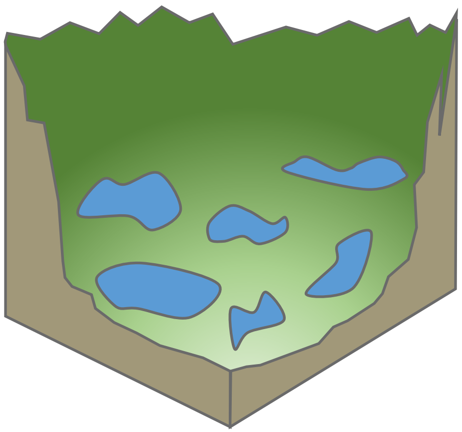
…

Limit sparsity and sequencing errors

Deal with library size and/or compositionallity

Data processing, normalisation

BIOLAWEB



It is important to consider that normalization is a highly debated topic and there is currently no consensus from experts on which normalization method is better

Methods for normalizing microbiome data: An ecological perspective

Donald T. McKnight[1] | Roger Huerlimann[1] | Deborah S. Bower[1,2] | Lin Schwarzkopf[1] | Ross A. Alford[1] | Kyall R. Zenger[1]

Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1*] Jean M. Macklaim[1] Vera Pawlowsky-
Juan J. Egozcue[3]

Alpha and beta-diversities performance comparison between different normalization methods and centered log-ratio transformation in a microbiome public dataset

David Bars-Cortina[1,2,3]

A review of normalization and differential abundance methods for microbiome counts data

Dionne Swift[1] | Kellen Cresswell[2] | Robert Johnson[1] | Spiro Stilianoudakis[1] | Xingtao Wei[1]

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes

Published: April 3, 2014 • https://doi.org/10.1371/journal.pcbi.1003531

Rarefaction is currently the best approach to control for uneven sequencing effort in amplicon sequence analyses

Patrick D. Schloss

Data processing, normalisation

BIOLAWEB

Developping transformation tools to account for problematics in sequencing data is an active domain of research

More and more sophisticate tools are developed, initially for RNAseq data / metagenomic data and differential abundance analysis…

… wich are not always optimal for metabarcoding data and alpha/beta diversity analysis

TABLE 1 Summary of normalization methods

Methods	Scale factor	Normalizing covariate/method	Availability (bioconductor/R)	Correction	
TSS	$S_j = \frac{Y_{ij}}{n_j}$	Total number of sample reads	Topics	Total reads	
CSS	$S_j = \frac{\sum_{i:Y_{ij} \le q_j^i} Y_{ij}+1}{N}$	Cumulative sum of counts (up to threshold q)	metagenomeSeq	Sequencing depth	2013
TMM	$\log_2(S_j) = \sum_{i \in G*} w_{ij} \log_2\left(\frac{X_{ij}}{X_{ir}}\right)$	Trimmed mean of logged expression ratios/ Inverse Variance	edgeR	Sequencing depth	2010
DESEq2	$S_j = med_i \frac{Y_{ij}}{\left(\prod_{j'=1}^{N} Y_{ij'}\right)^{1/N}}$	Median ratio of gene counts relative to geometric mean per gene	DESeq2	Sequencing depth and compositionality	2010
GMPR	$S_j = \left(\prod_{k=1}^{n} Median_{i \mid Y_{ij}Y_{ik} \neq 0}\left\{\frac{Y_{ij}}{Y_{ik}}\right\}\right)^{1/N}$	Geometric mean of the pairwise ratio of nonzero counts	GMPR	Sequencing depth, compositionality, and sparsity	2018
Wrench	$S_j = \frac{1}{p}\sum_{ij} w_{ij} \frac{X_{ij}}{X_{i\cdot}}$	Group-wise and sample-wise compositional bias factor	Wrench	Sequencing depth, compositionality, and sparsity	2023
ANCOM-BC	$\log(S_j) = \frac{1}{p}\sum_{i=1}^{p}\left(y_{ij} - x_j^T \widehat{\beta}_i\right)$	Ratio of expected absolute abundance to ratio of library size to microbial load	ANCOM-BC	Sequencing depth and compositionality	2020
CLR[a] transformation	$\log\left(\frac{Y_{ij}}{\left[\prod_i Y_{ij}\right]^{1/p}}\right)$	Log ratio of observed values and their geometric means		Compositionality	

Metabarcoding   RNASeq

Swift et al., 2022

Regardless which method you choose, you need to normalize before doing ecological analysis

1

Filter ASVs based on taxonomy

Keep only ASVs assigned to the targeted organisms

(e.g. Diatoms)

2

Filter ASVs that are not abundant

(optional)

Singletons

That have less than *x* sequences

That occurs in less than *x* samples

3

Normalize/transform

Rarefaction

or

Scaling

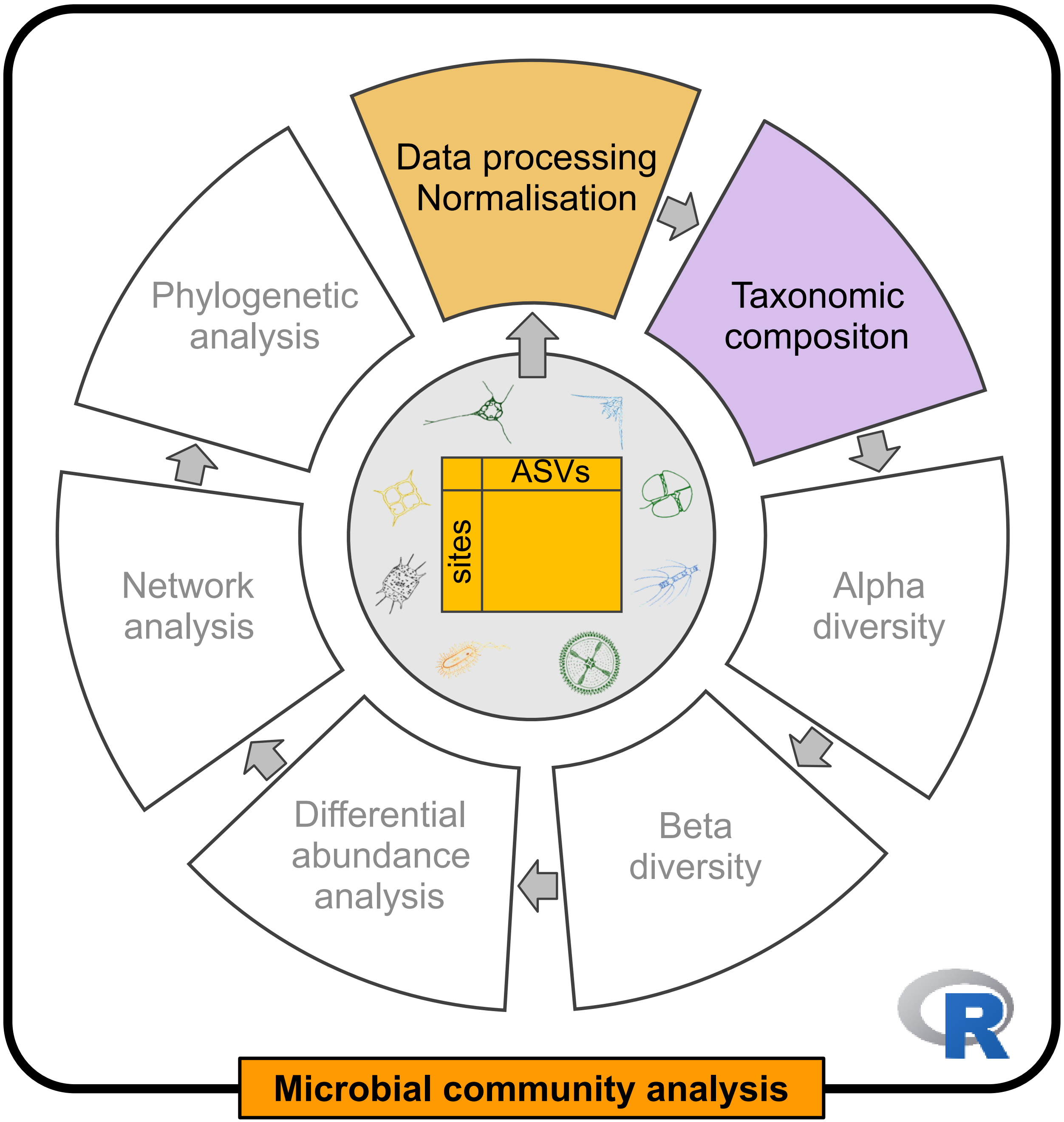
or

Log ratio

…

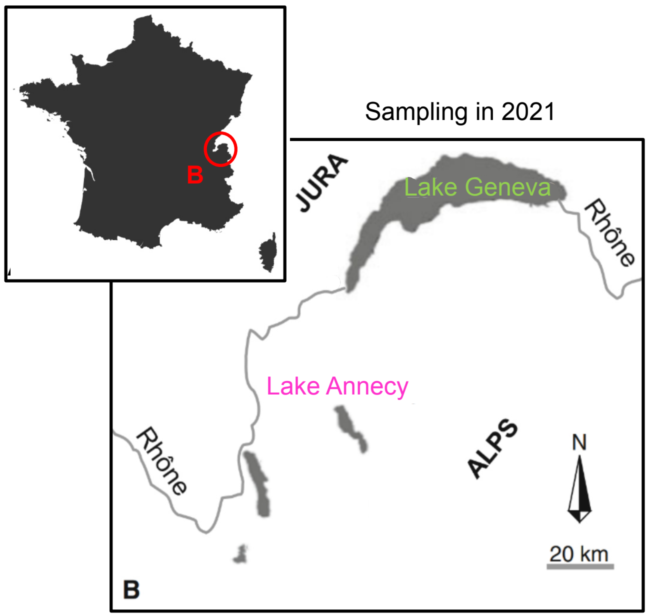
Limit sparsity and sequencing errors

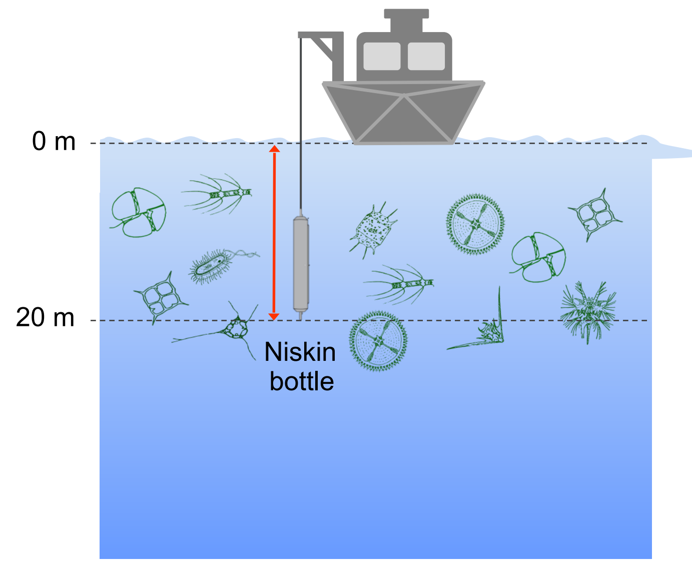
Deal with library size and/or compositionallity

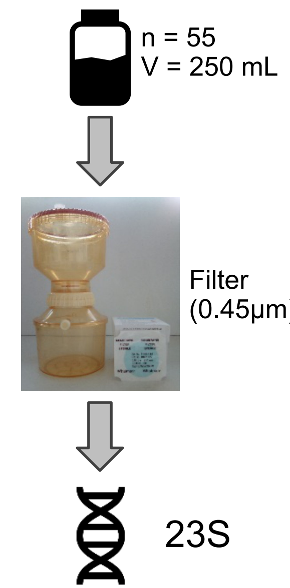

Data processing
Normalisation
Taxonomic
compositon
Alpha
diversity
Beta
diversity
Differential
abundance
analysis
Network
analysis
Phylogenetic
analysis
ASVs
sites
Microbial community analysis

Taxonomic composition

To illustrate alpha and beta-diversity analysis we will use a dataset of phytoplankton dynamic over one year (2021) from 2 alpine lakes

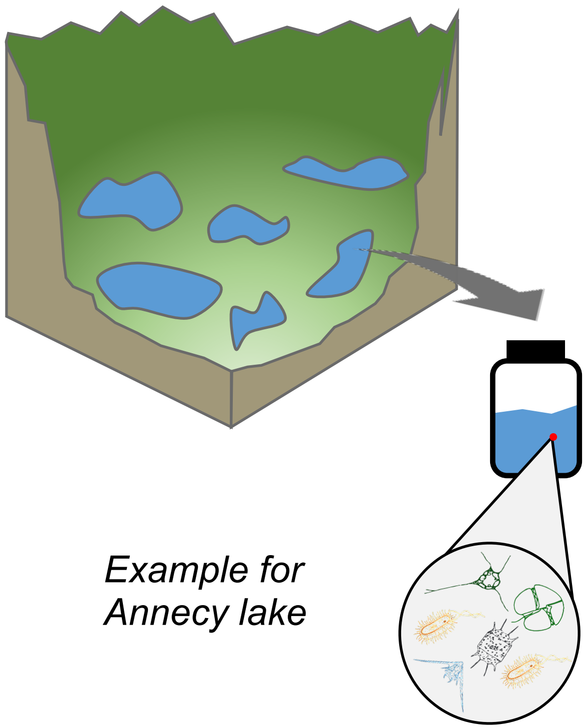




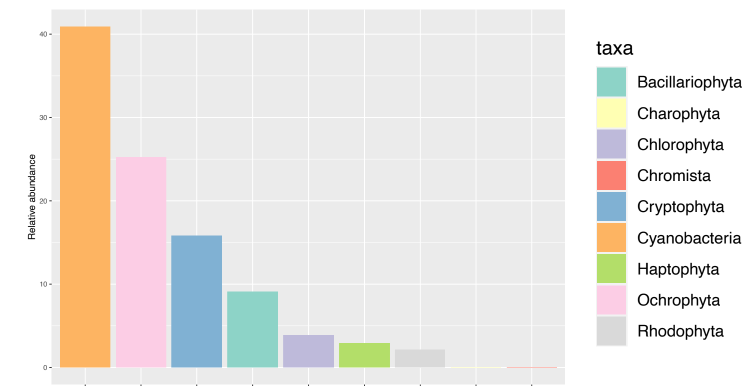


- Sampling 1 or 2 time per month
- 250mL of water filtered on 0.45µm filters
- Targeting 23S barcode
- Illumina MiSeq
- Data analysed with DADA2 pipeline (ASVs)

Taxonomic composition



Example for Annecy lake

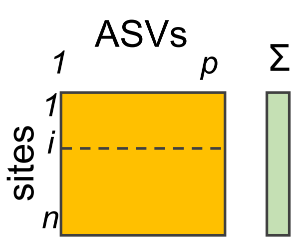



```
gg <- ggplot(dASV, aes(x=reorder(taxa, perc, decreasing=T), y=perc)) +
    geom_bar(stat="identity",position="stack", aes(fill=taxa))
    scale_fill_brewer(palette="Set3") + xlab("Lake") + ylab("Relative abundance") +
    theme(strip.text.y = element_text(angle = 0), axis.text.x = element_text(angle = 90))
```

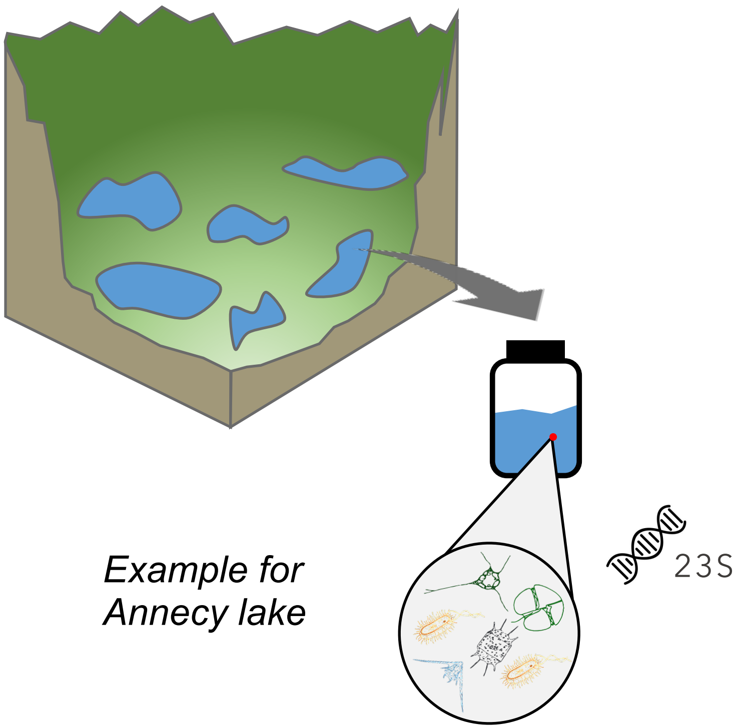
Who’s there?

And when?

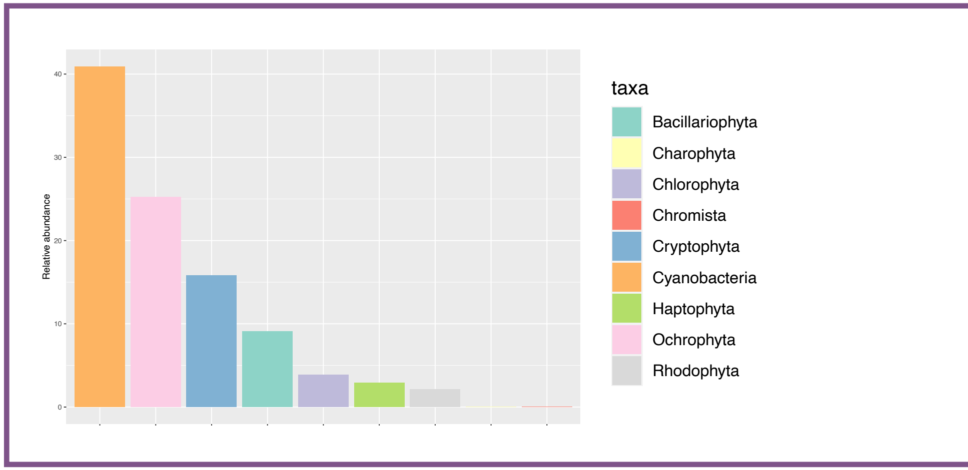
Using relative abundance



Taxonomic composition



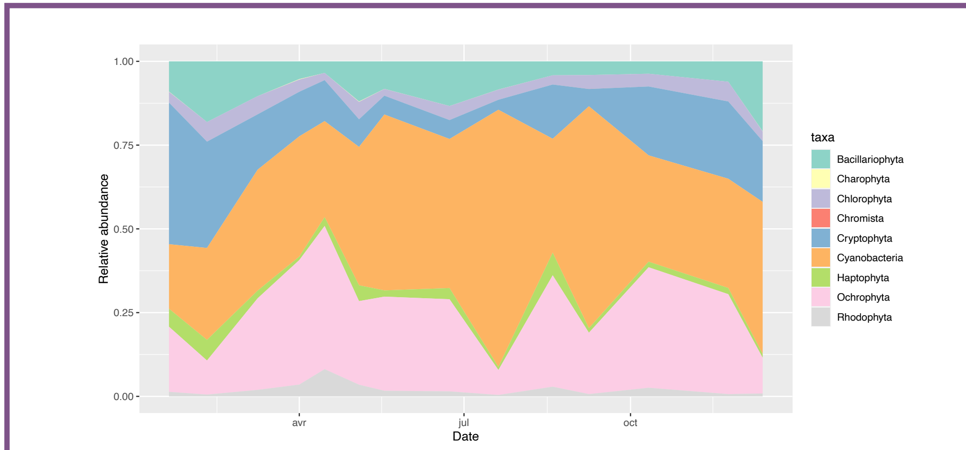
Example for Annecy lake



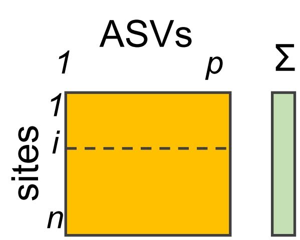

```
gg <- ggplot(dASV, aes(x=reorder(taxa, perc, decreasing=T), y=perc)) +
    geom_bar(stat="identity",position="stack", aes(fill=taxa))
    scale_fill_brewer(palette="Set3") + xlab("Lake") + ylab("Relative abundance") +
    theme(strip.text.y = element_text(angle = 0), axis.text.x = element_text(angle = 90))
```

Who's there?

And when?

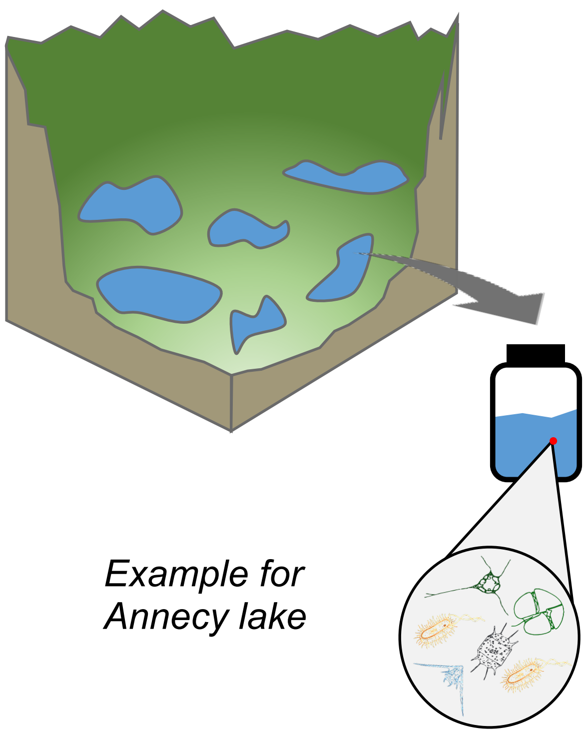


Using relative abundance

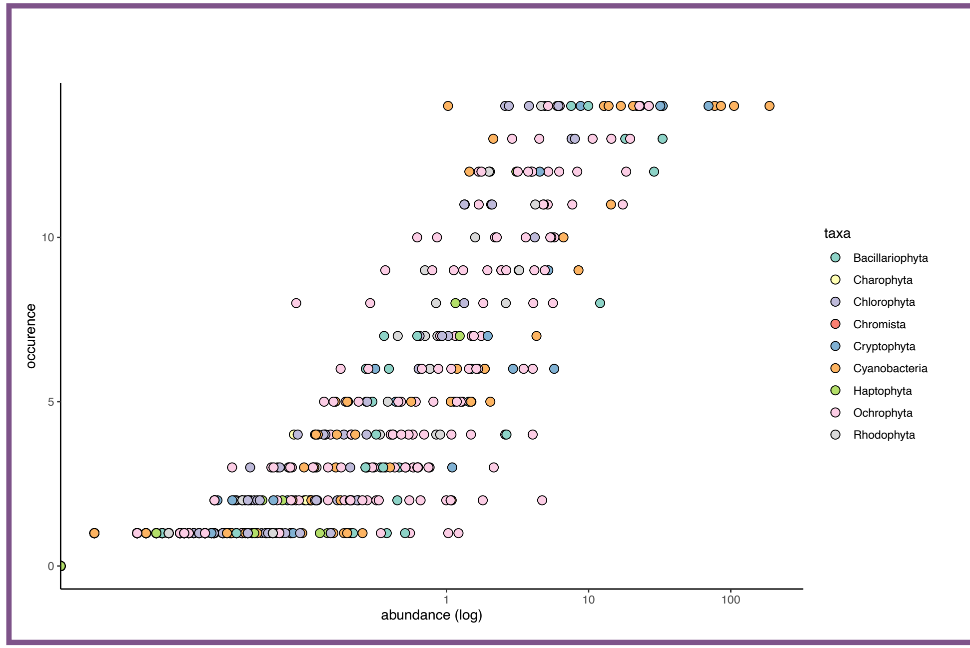



```
gg <- ggplot(dASV, aes(x=as.Date(Date,format="%d/%m/%Y"), y=value), group=taxa) +
geom_area(stat="identity", position="fill", aes(fill=taxa)) +
  scale_fill_brewer(palette="Set3") + xlab("Date") + ylab("clr abund")
```

Taxonomic composition



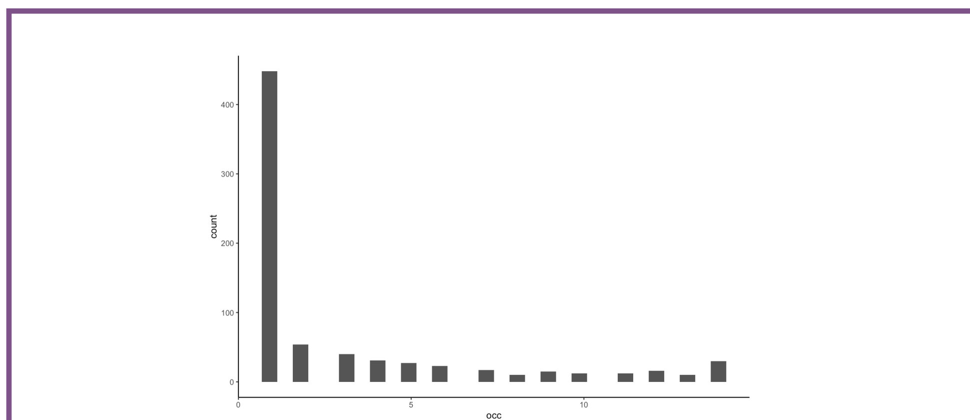
Example for Annecy lake



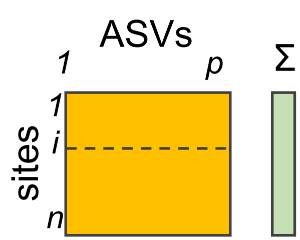

```
gg <- ggplot(ASV_table_occ, aes(x=occ, y=tot)) +
 geom_point(aes(fill=taxa), shape=21, size=3) +
 scale_y_continuous(trans="log", breaks=c(1,10,100)) + theme_classic() +
 scale_fill_brewer(palette="Set3") + xlab("occurence") + ylab("abundance (log)")
```

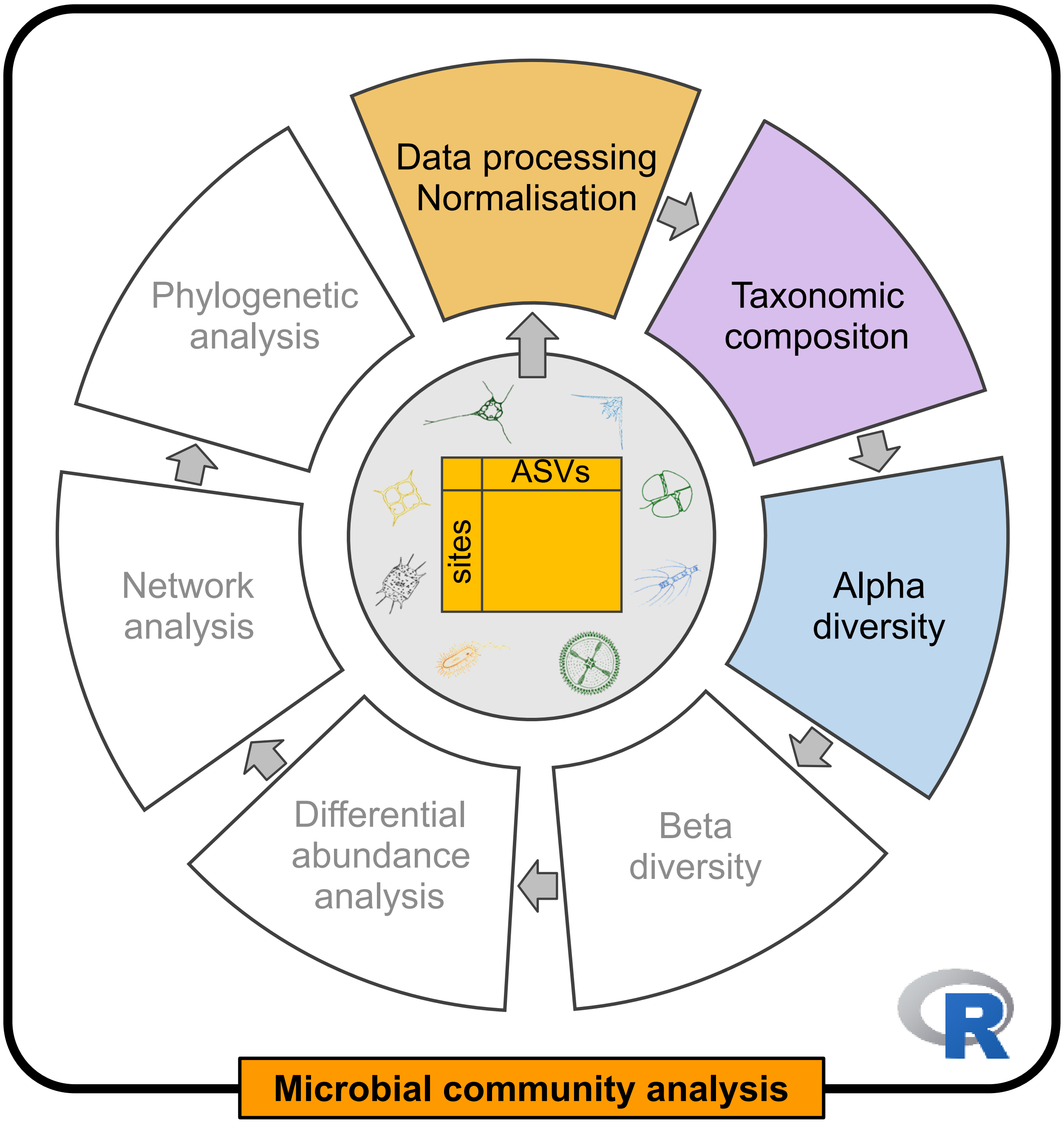
Who’s there?

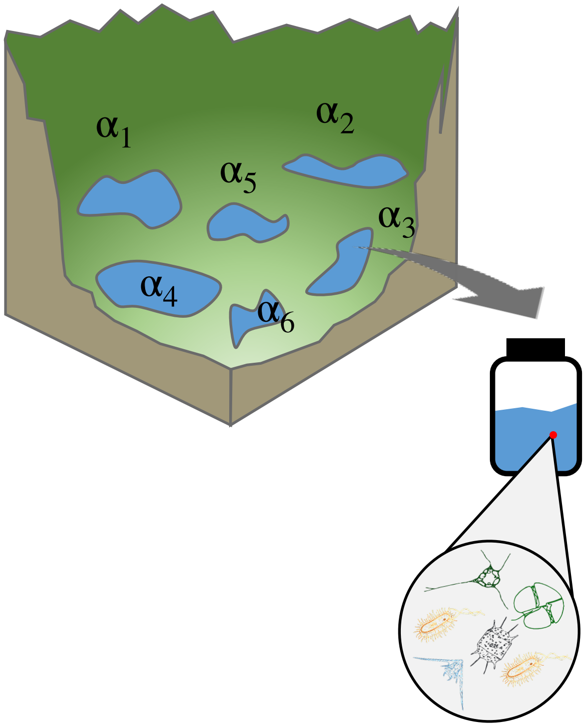
And when?



Using relative abundance



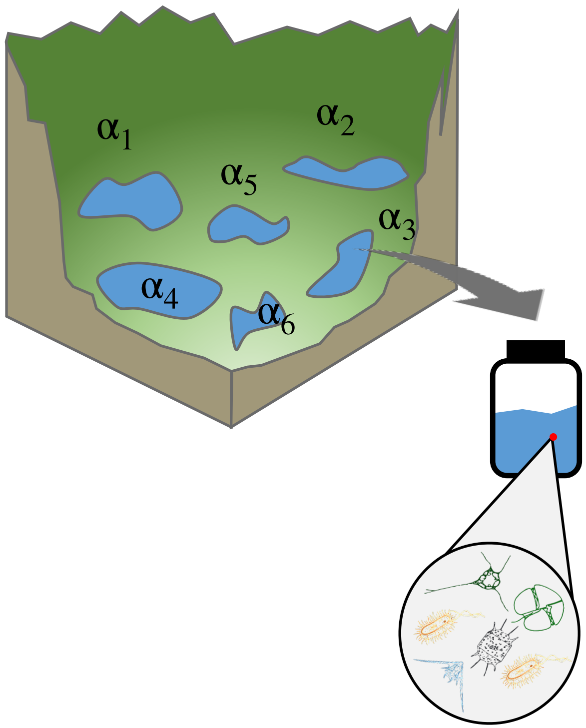

Data processing
Normalisation
Taxonomic
compositon
Alpha
diversity
Beta
diversity
Differential
abundance
analysis
Network
analysis
Phylogenetic
analysis
ASVs
sites
Microbial community analysis



What is the species diversity inside each sample ?

Alpha diversity

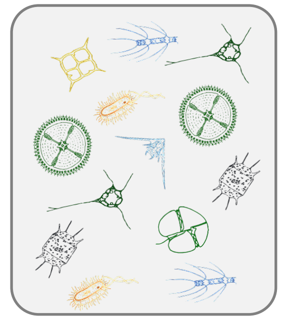
Alpha diversity



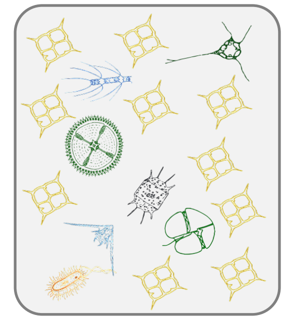
What is the species diversity inside each sample ?

Alpha diversity

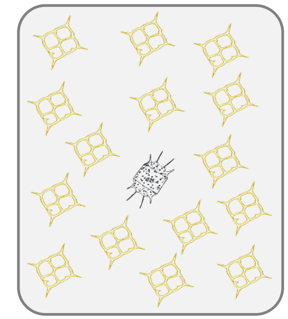
There are 2 component of diversity : richness and evenness



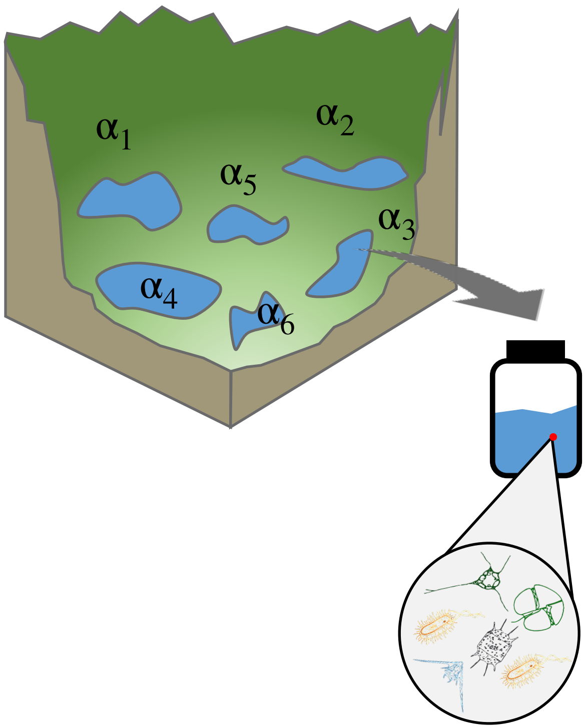
Rich and even



Rich, not even



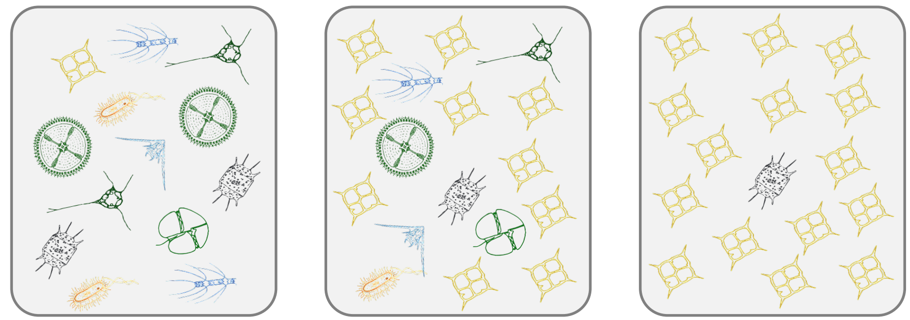
Not rich, not even



What is the species diversity inside each sample ?

Alpha diversity

There are 2 component of diversity : richness and evenness



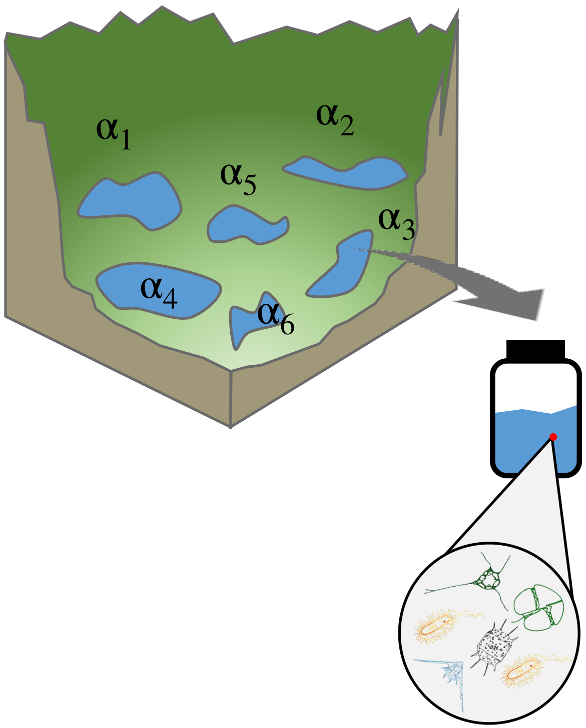
Rich and even — *Rich, not even* — *Not rich, not even*

There are different index

	Richness	Evenness	R function*
Observed	X		specnumber(ASV_table)
Shannon	**X**	X	diversity(ASV_table, « shannon »)
Simpson	X	**X**	diversity(ASV_table, « simpson »)
Pielou		X	shannon/log(specnumber(ASV_table))

* *Package vegan*

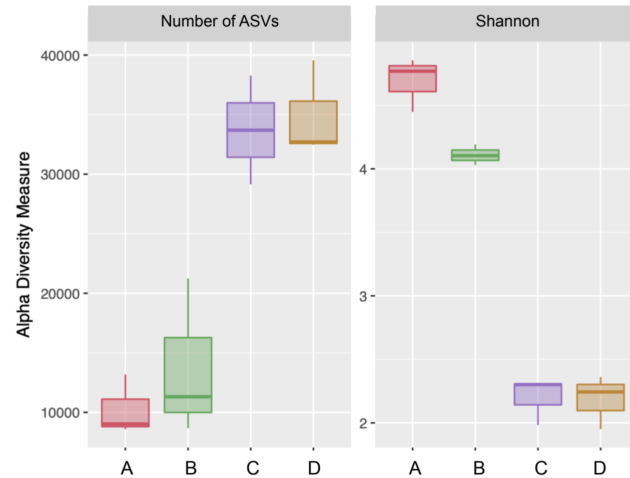
Alpha diversity



What is the species diversity inside each sample ?

Alpha diversity

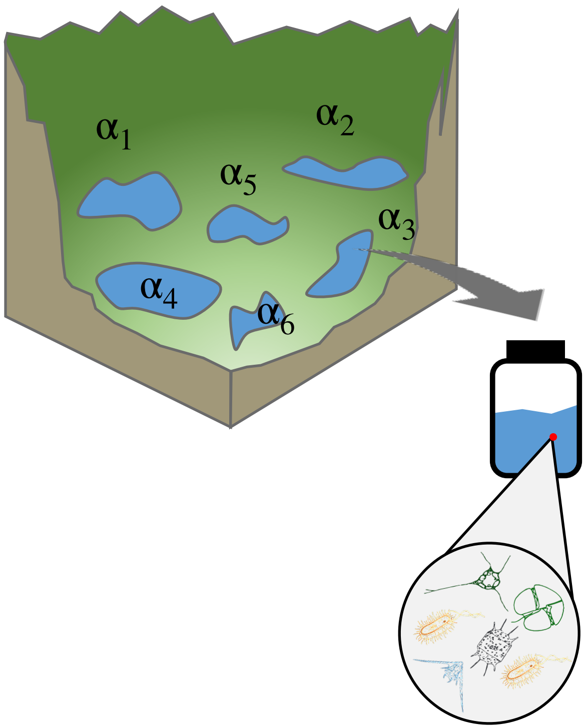
Exemple with bacterial communities composition of biofilm in seawater



Alpha-diversity measured based on richness only or with Shannon index gives opposite results…

Trigodet *et al.*, 2019

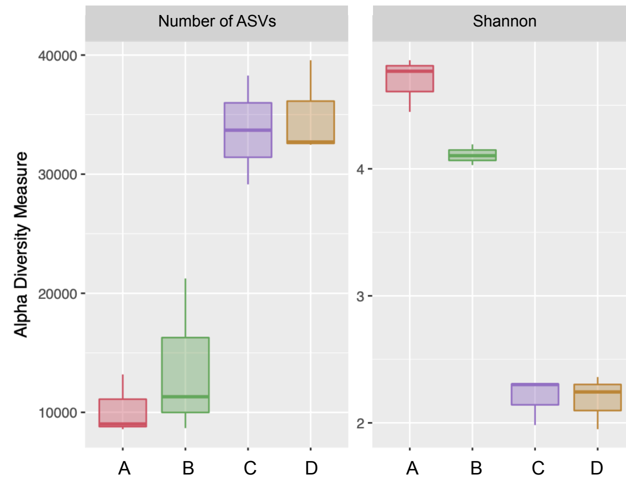
Alpha diversity



What is the species diversity inside each sample ?

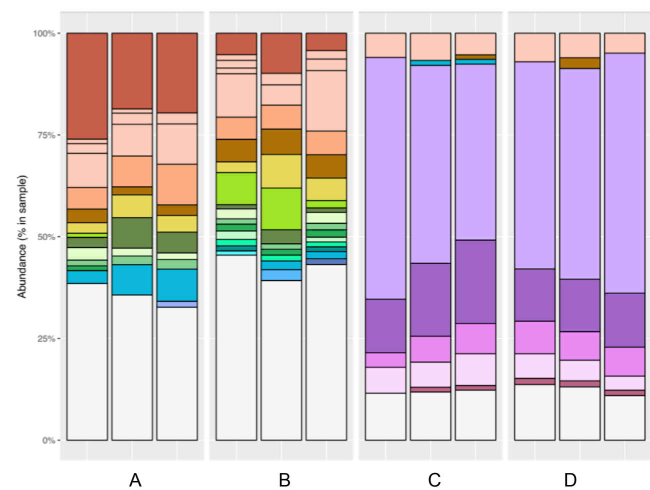
Alpha diversity

Exemple with bacterial communities composition of biofilm in seawater



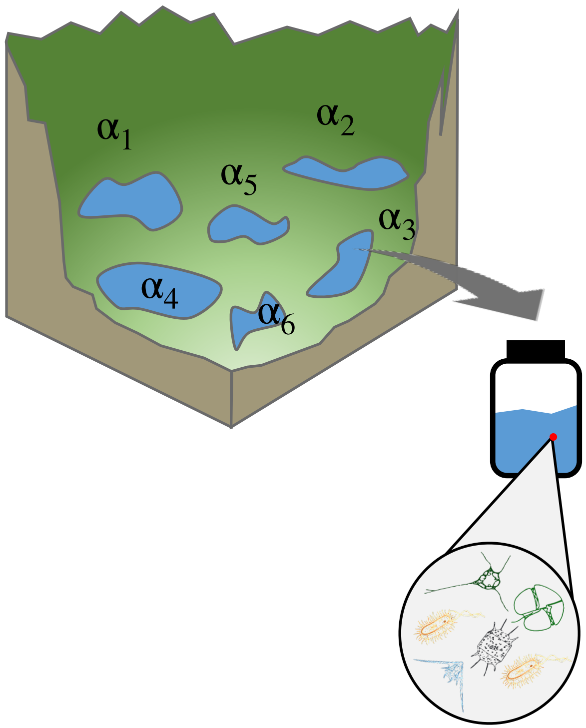
Alpha-diversity measured based on richness only or with Shannon index gives opposite results…

…This is because samples of condition A and B are even while samples of conditions C and D have highly dominant taxa (low evenness)



Trigodet *et al.*, 2019

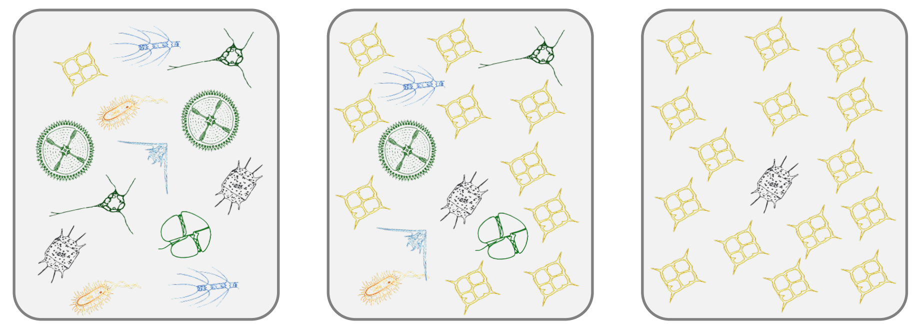
Alpha diversity



What is the species diversity inside each sample ?

Alpha diversity

There are 2 component of diversity : richness and evenness

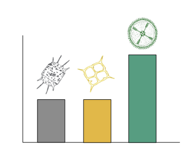


Rich and even | *Rich, not even* | *Not rich, not even*

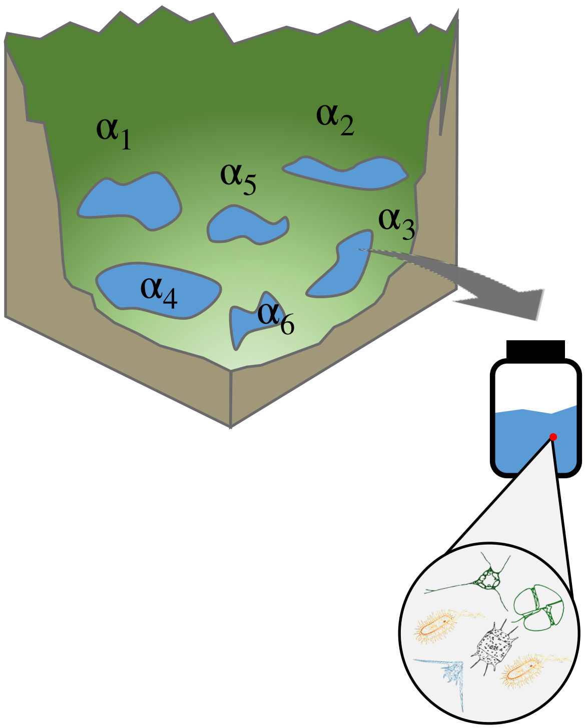
Two components are essential in the calcul of alpha diversity indexes

Number of ASVs

Abundance of ASVs



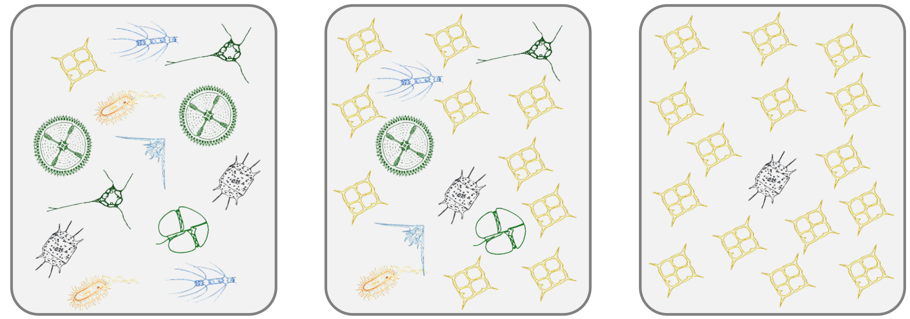
Alpha diversity



What is the species diversity inside each sample ?

Alpha diversity

There are 2 component of diversity : richness and evenness

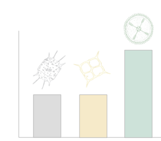


Rich and even | *Rich, not even* | *Not rich, not even*

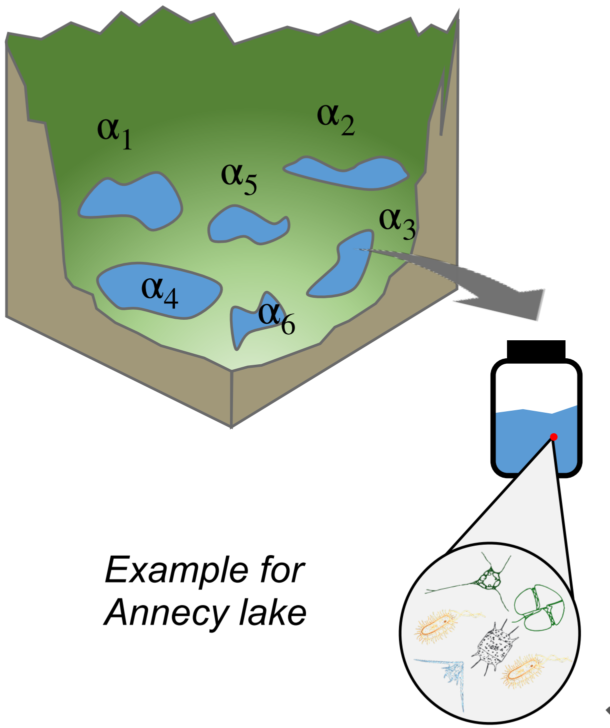
Two components are essential in the calcul of alpha diversity indexes

Number of ASVs

Abundance of ASVs



Influence of library size !



Example for Annecy lake

What is the species diversity inside each sample ?

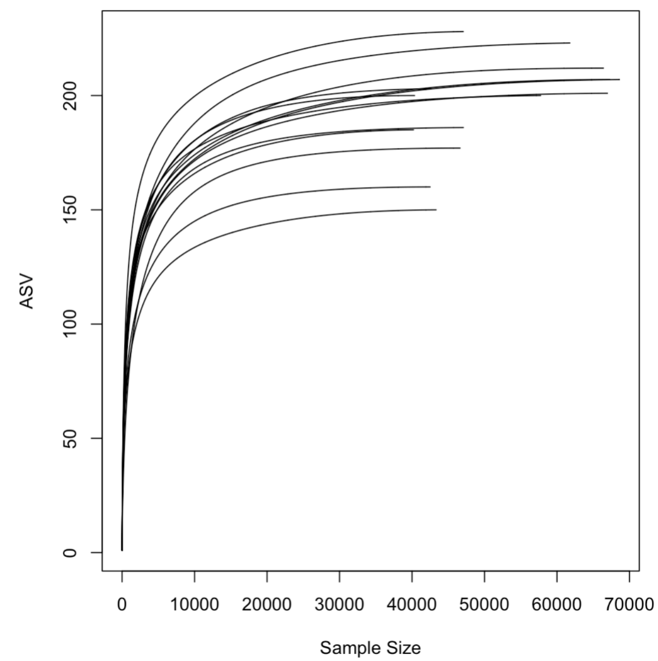
Alpha diversity

Problem to estimate **richness** with metabarcoding data

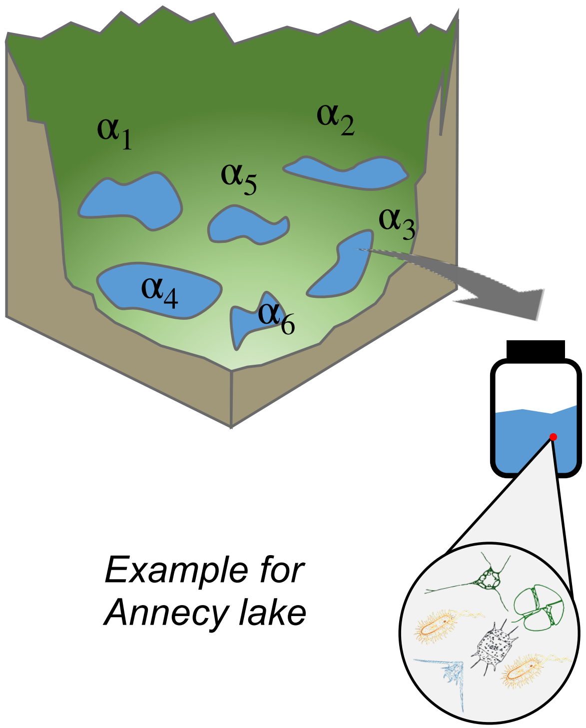
Different library size

Rarefaction curve :

Is the sequencing effort enough to recover phytoplankton diversity ?




```
vegan::rarecurve(ASV_table, step=30, xlab="Sample Size", ylab="ASV", label=T)
```



Example for Annecy lake

What is the species diversity inside each sample ?

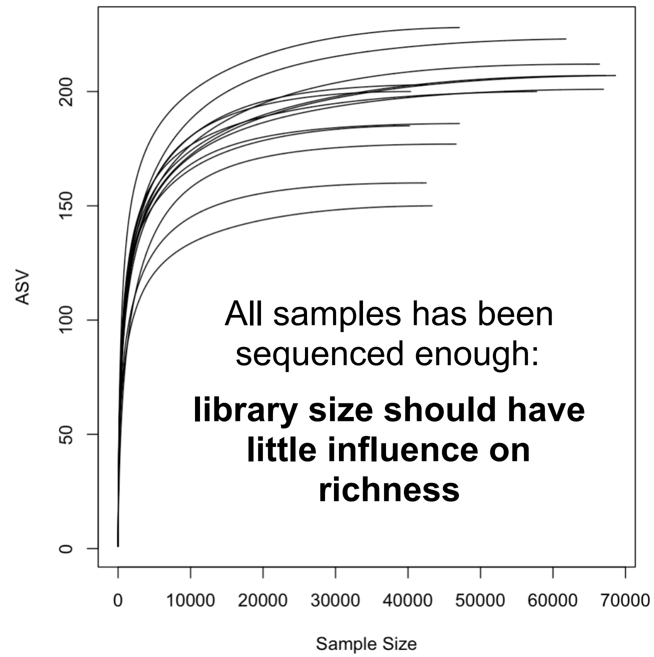
Alpha diversity

Problem to estimate **richness** with metabarcoding data

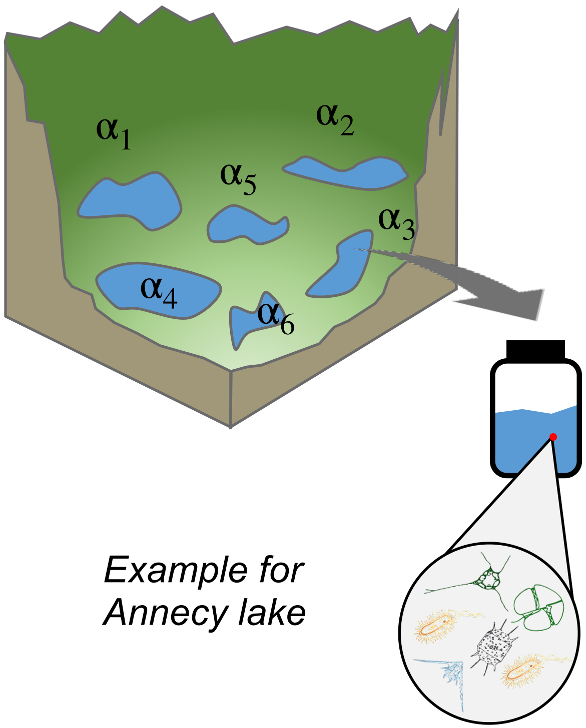
Different library size

Rarefaction curve :

Is the sequencing effort enough to recover phytoplankton diversity ?




```
vegan::rarecurve(ASV_table, step=30, xlab="Sample Size", ylab="ASV", label=T)
```



Example for Annecy lake

What is the species diversity inside each sample ?

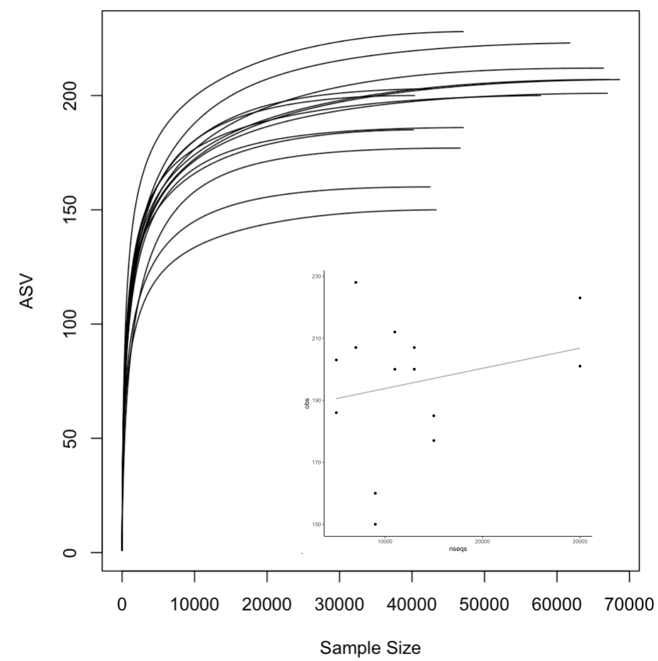
Alpha diversity

Problem to estimate **richness** with metabarcoding data

Different library size

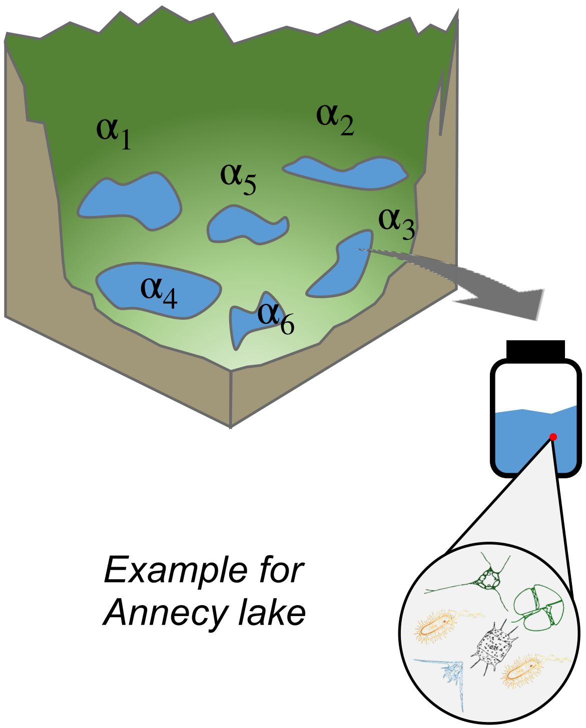
Rarefaction curve :

Is the sequencing effort enough to recover phytoplankton diversity ?




```
vegan::rarecurve(ASV_table, step=30, xlab="Sample Size", ylab="ASV", label=T)
```

Alpha diversity



Example for Annecy lake

What is the species diversity inside each sample ?

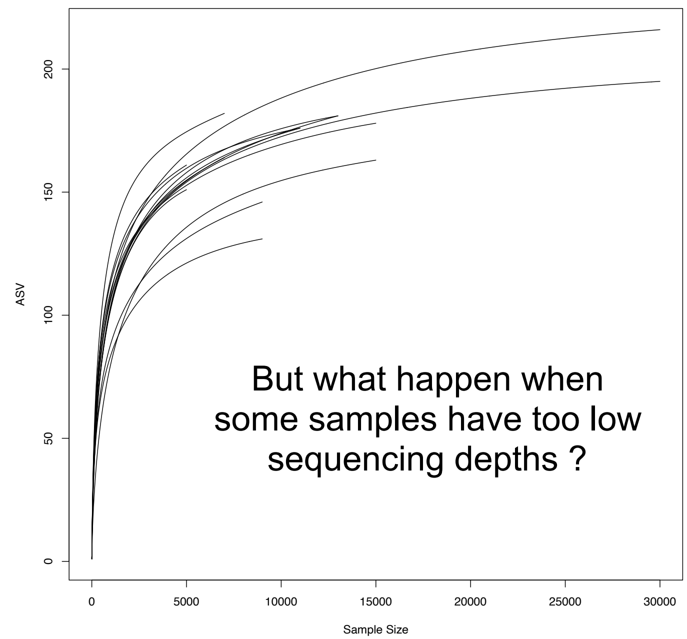
Alpha diversity

Problem to estimate **richness** with metabarcoding data

Different library size

Rarefaction curve :

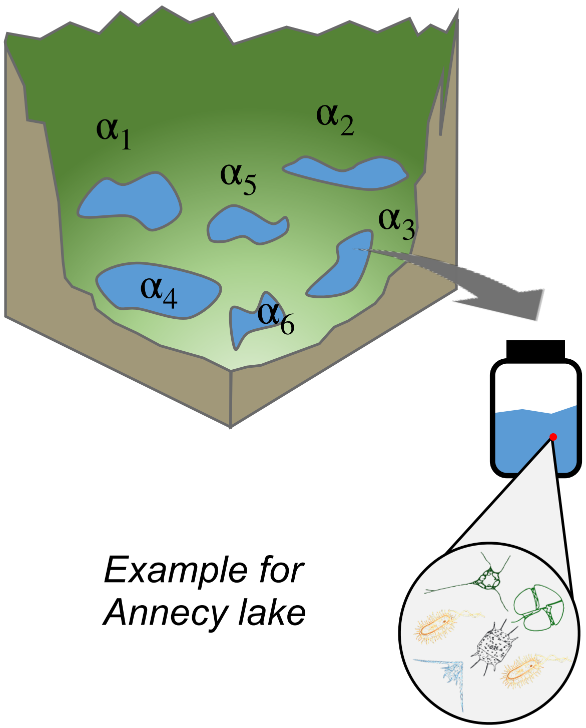
Is the sequencing effort enough to recover phytoplankton diversity ?



But what happen when some samples have too low sequencing depths ?

```
vegan::rarecurve(ASV_table, step=30, xlab="Sample Size", ylab="ASV", label=T)
```

Alpha diversity



Example for Annecy lake

What is the species diversity inside each sample ?

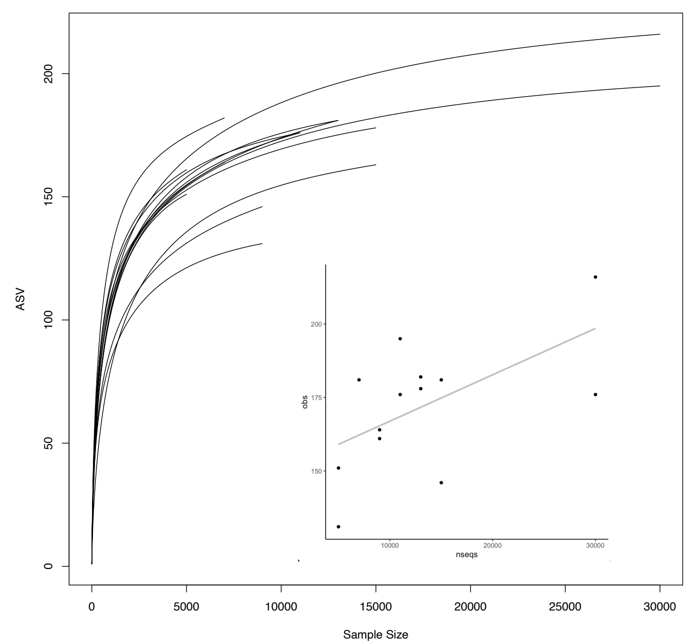
Alpha diversity

Problem to estimate **richness** with metabarcoding data

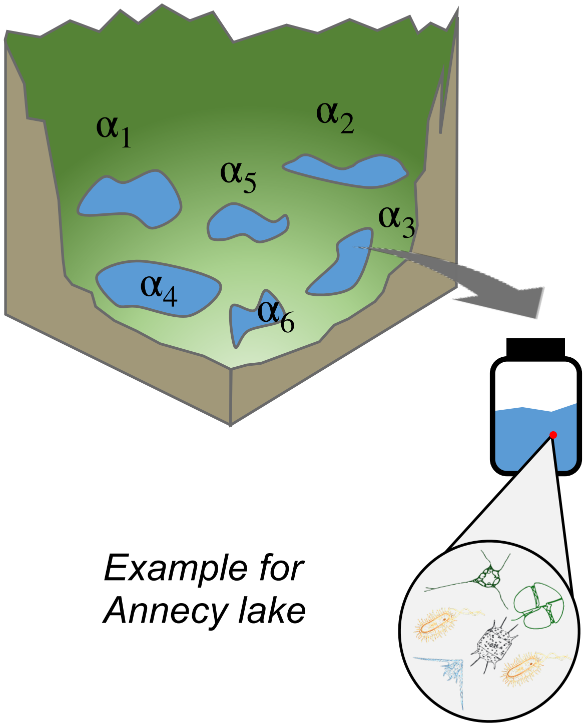
Different library size

Rarefaction curve :

Is the sequencing effort enough to recover phytoplankton diversity ?




```
vegan::rarecurve(ASV_table, step=30, xlab="Sample Size", ylab="ASV", label=T)
```



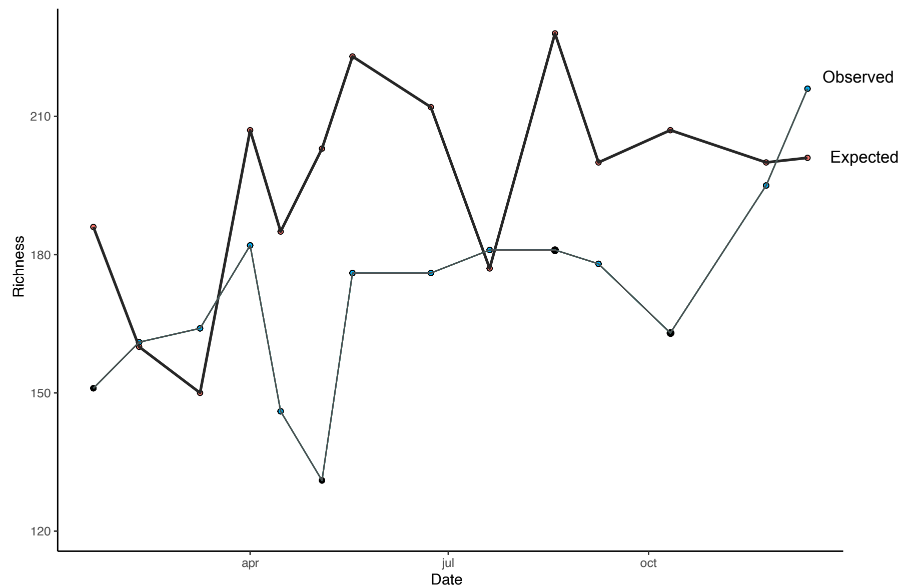
Example for Annecy lake

What is the species diversity inside each sample ?

Alpha diversity

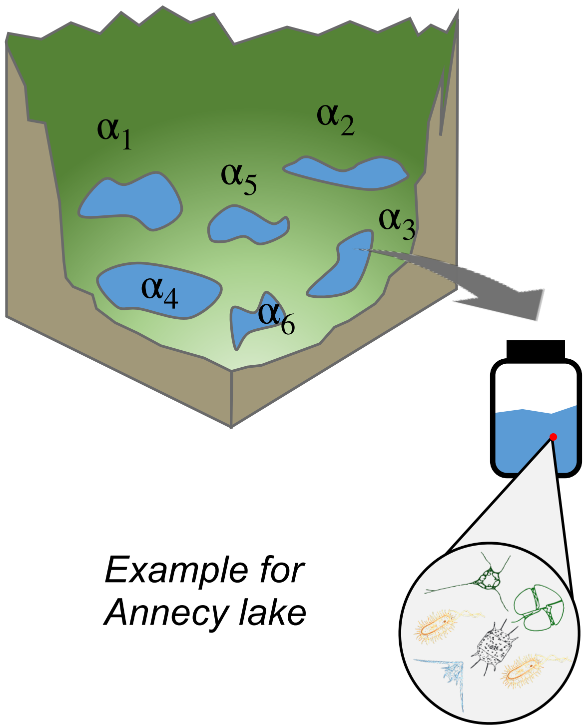
Problem to estimate **richness** with metabarcoding data

Different library size



Estimation of richness is incorrect

Alpha diversity



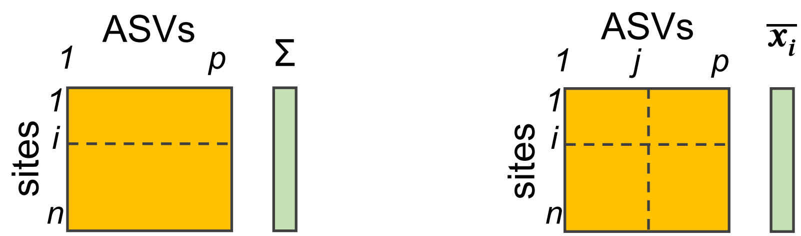
Example for Annecy lake

What is the species diversity inside each sample ?

Alpha diversity

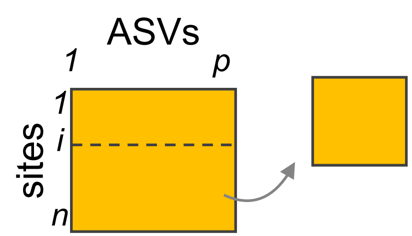
Problem to estimate **richness** with metabarcoding data

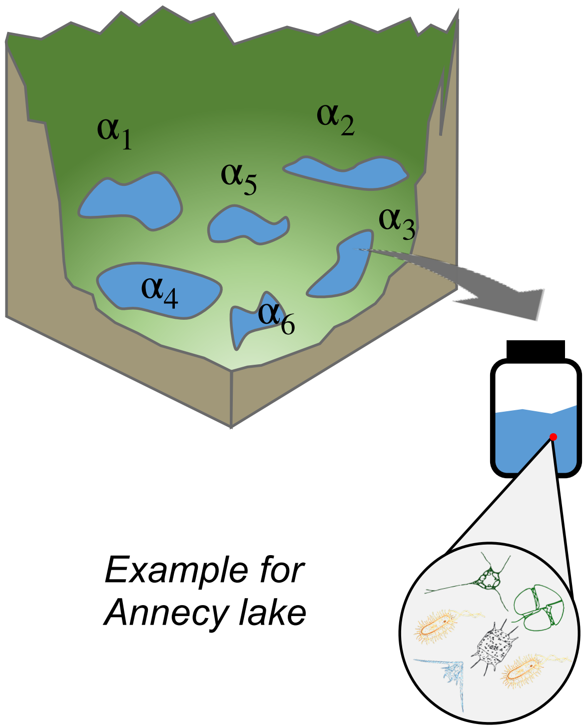
Different library size



Most normalisation methods (TSS, CSS) won't change anything in the number of ASVs **within a sample.**

Rarefaction is often suggested to be a good way to prepare data for alpha-diversity analysis, particularly if you have high variation in library size (more than 10x)



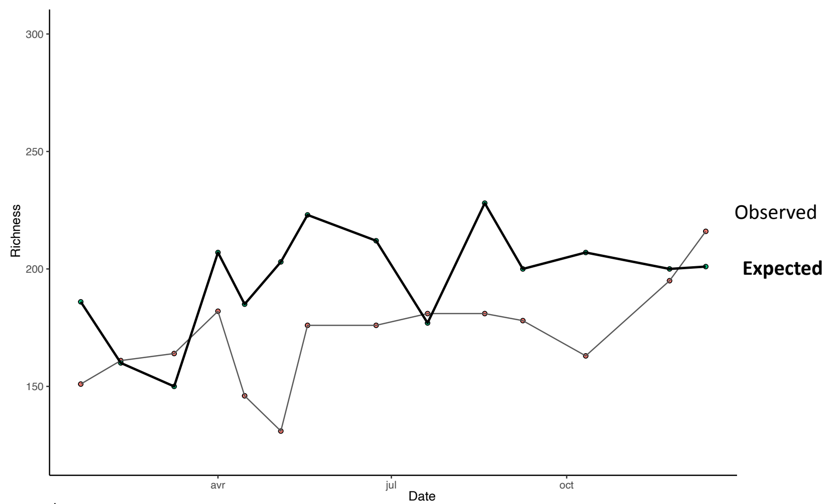


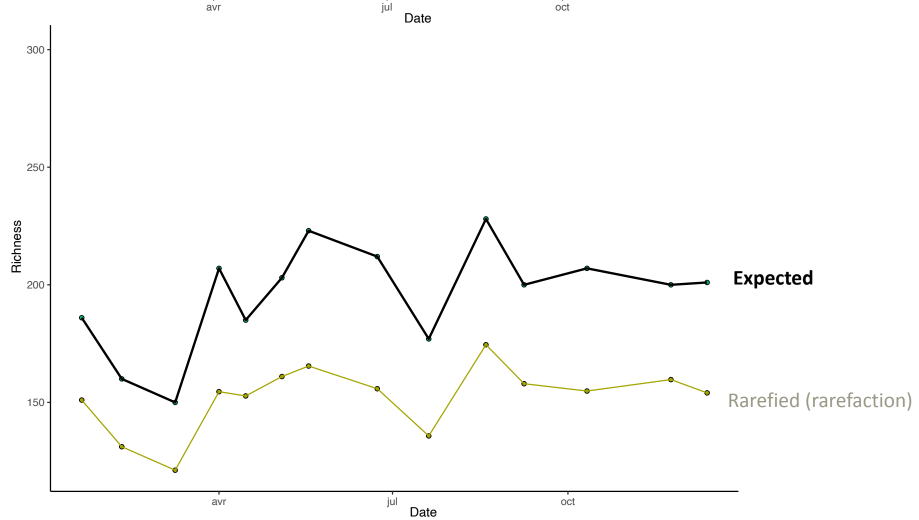
Example for Annecy lake

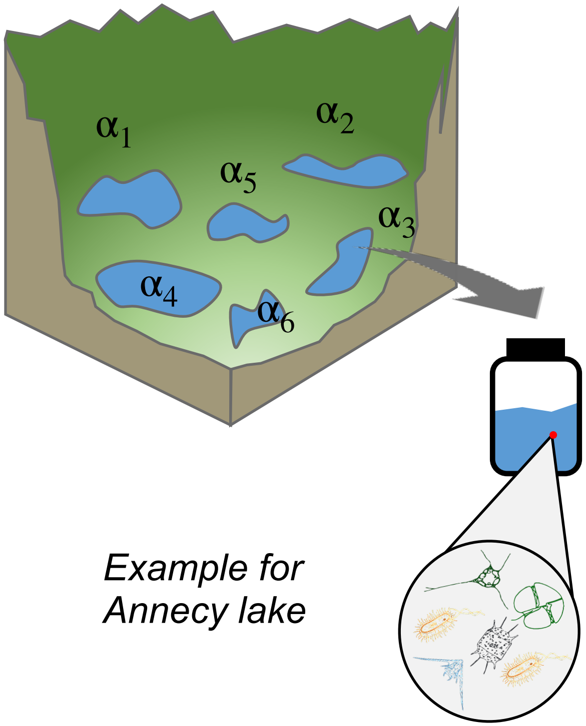
What is the species diversity inside each sample ?

Alpha diversity

Problem to estimate **richness** with metabarcoding data







Example for Annecy lake

What is the species diversity inside each sample ?

Alpha diversity

Problem to estimate **richness** with metabarcoding data

Different library size

Another possibility : **estimating diversity**

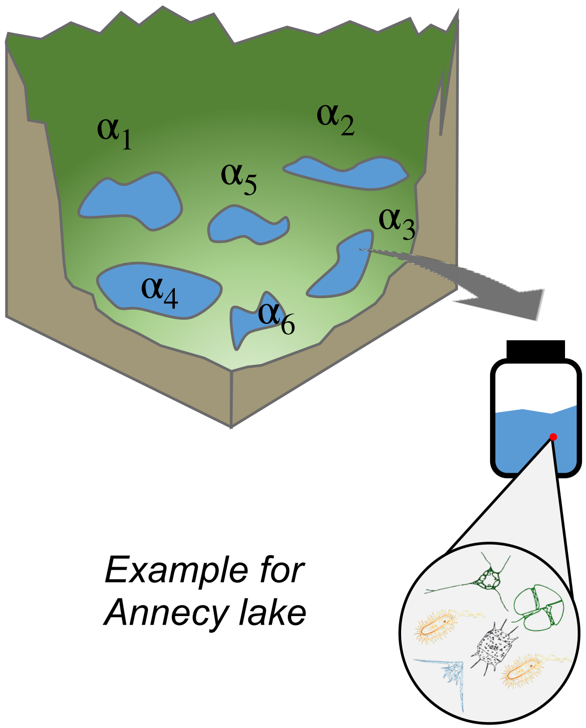
« *I encourage ecologists to use estimates of diversity that account for non-observed species* » Willis, 2019

A very famous richness estimator is Chao1 which is a non-parametric analysis that extrapolate the number of species present in a sample, based on the number of rare species.

(Chao, 2004)

Amy Willis also developed another index, parametric this time available in breakaway R package

(Willis, 2019)



Example for Annecy lake

What is the species diversity inside each sample ?

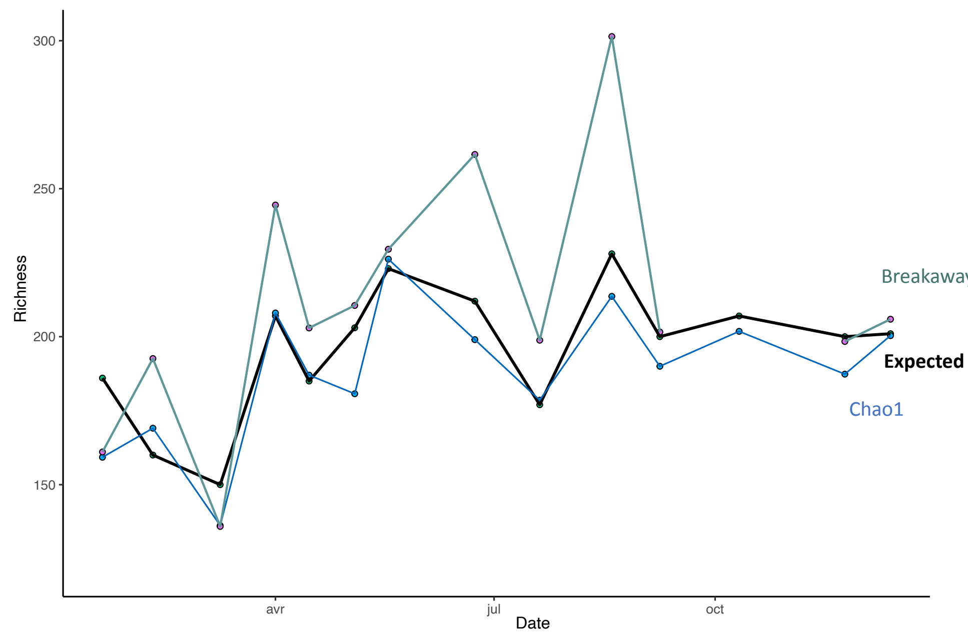
Alpha diversity

Problem to estimate **richness** with metabarcoding data

Different library size

Another possibility : **estimating diversity**

« *I encourage ecologists to use estimates of diversity that account for non-observed species* » Willis, 2019



Questions?