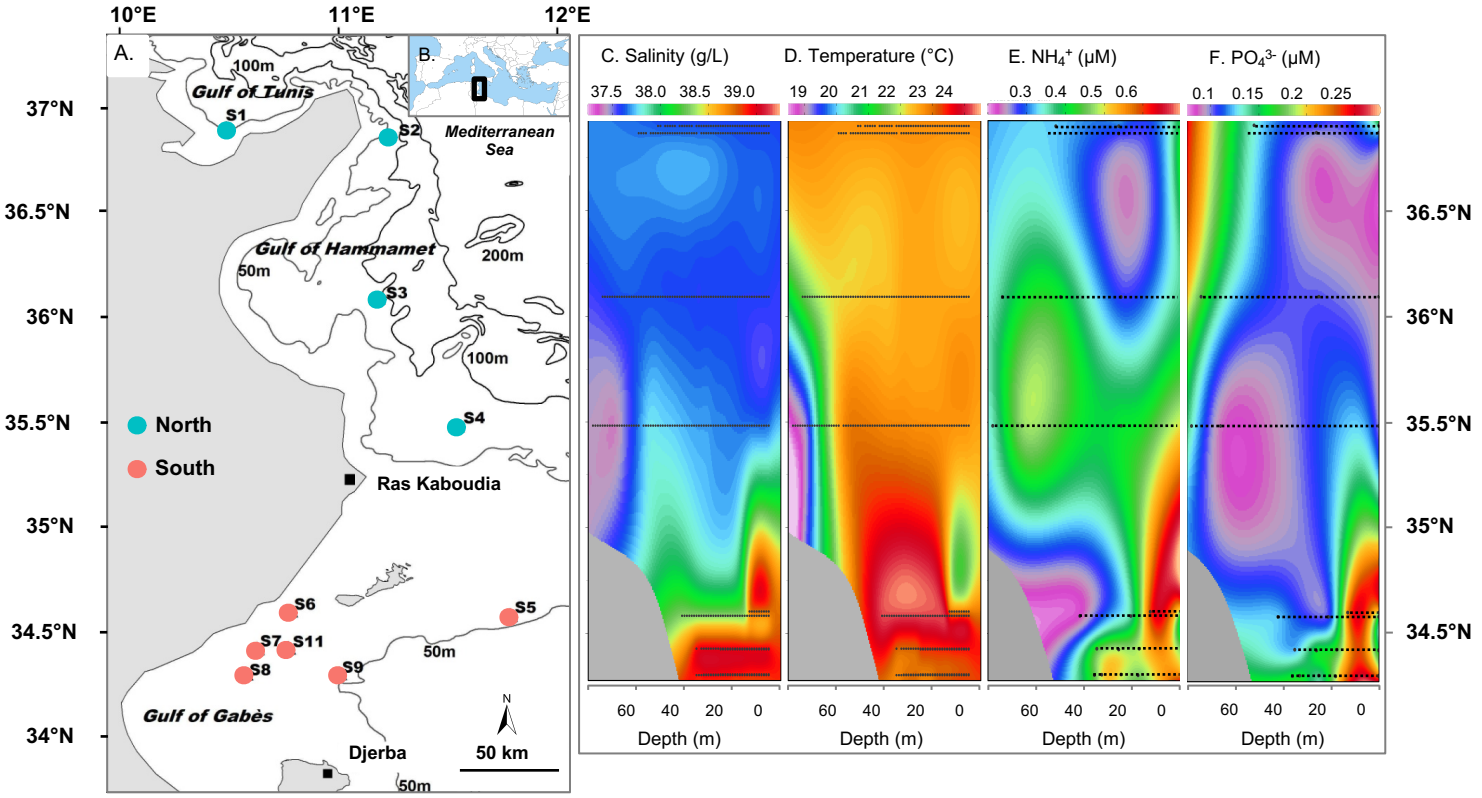
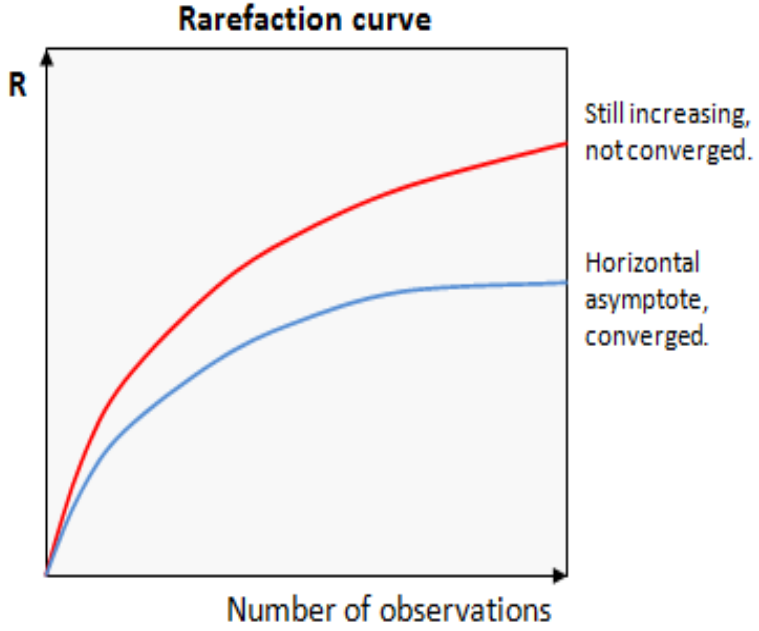
MetaData

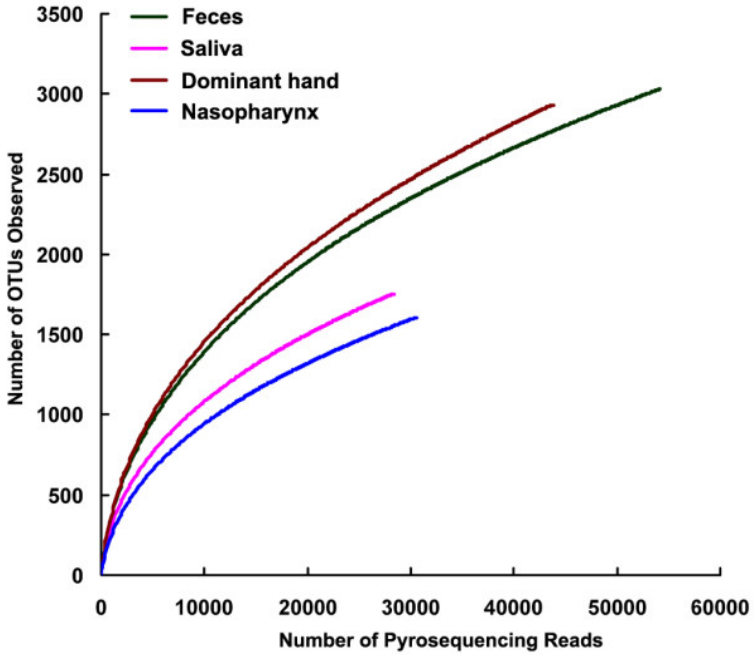


Add data describing your experiment (e.g. env param)

Rarefaction Curves

« Is the sequencing effort performed (sequencing depth) for a sample (s) sufficient for the number of species observed ? »



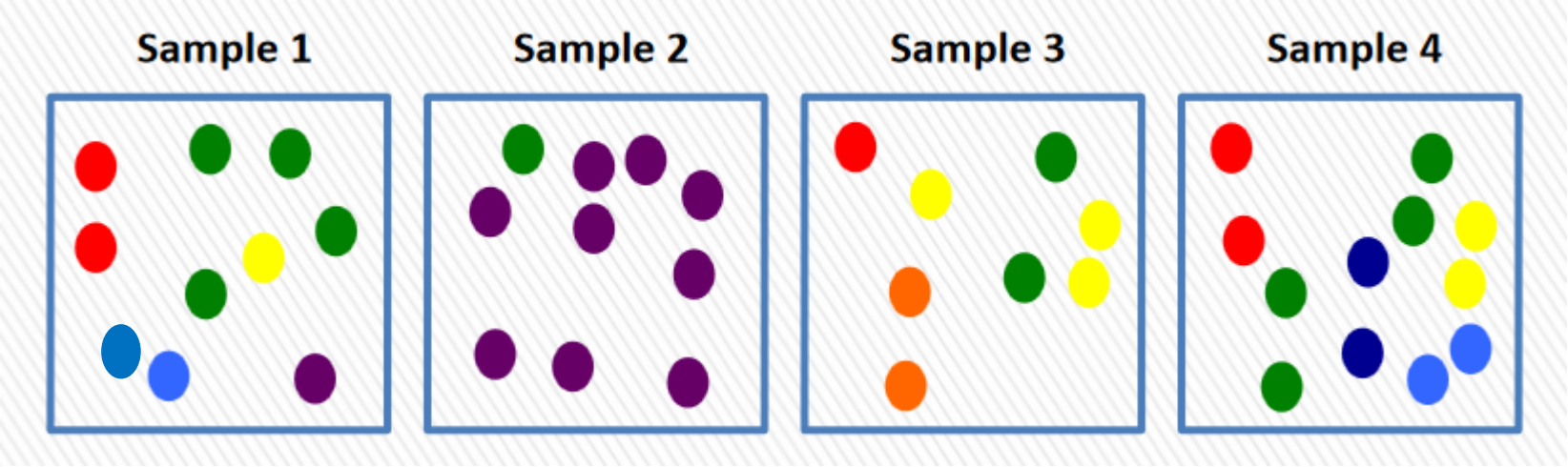


→Reach the asymptote ???

Asymptote means that sequencing more (depth), will not increase your number of OTU/ASVs observed

Alpha Diversity

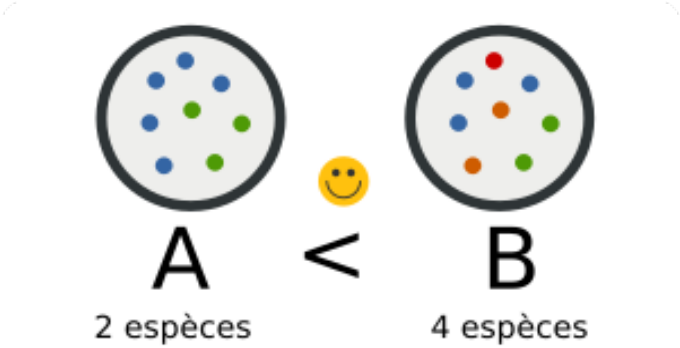
Alpha diversity is a measure of species diversity in a particular area/ecosystem → richness within a sample



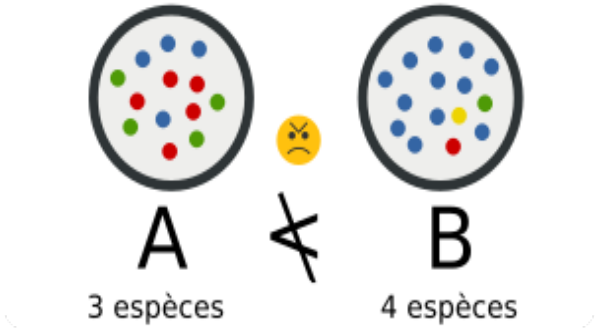
Can you rank the samples according to their diversity (from the lowest to greatest)??

Diversity: species richness

Global Richness (R) indice: Richness is the number of uniq taxa/ASVs observed in a sample



B has more species than A



B has more species than A
But seems less diversified than A

How to deal with?

Use indices of alpha diversity such as **Shannon, Simpson which reflect the taxa richness and their relative abundance (distribution, evenness)**

Shannon-Weaver Index: Combine taxa richness & evenness

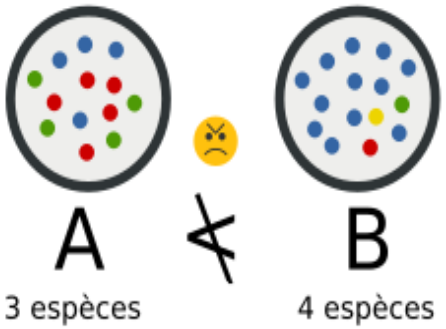
For each species : proportion of the taxa multiplicates by log (proportion)

$$H(X) = H_2(X) = -\sum_{i=1}^{n} P_i \log_2 P_i.$$

- A consist of 3 species, of which 4 green, 5 red & 4 blue

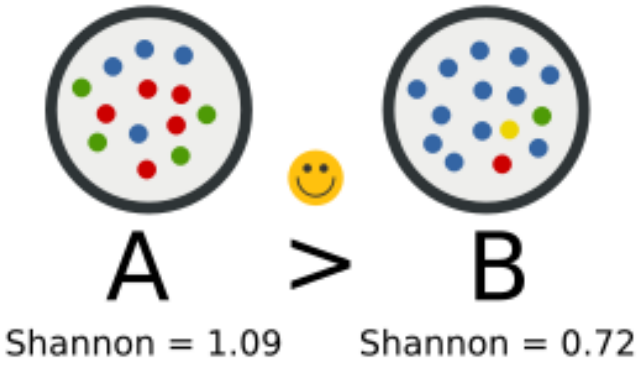
The Shannon indice will be :

$$-(\frac{4}{13}log(\frac{4}{13}) + \frac{5}{13}log(\frac{5}{13}) + \frac{4}{13}log(\frac{4}{13})) = 1.09$$



- B consist of 4 species, of which 1 reen, 1 red, 1 yellow & 11 blue

Finally, after estimating Shannon for B sample



→ Influenced by low abundant taxa
→ Greater weight on evenness

Equitability Pielou index… with Shannon scores

- **Allow comparison between samples (Normalization)**

Shannon is dependent on species richness!

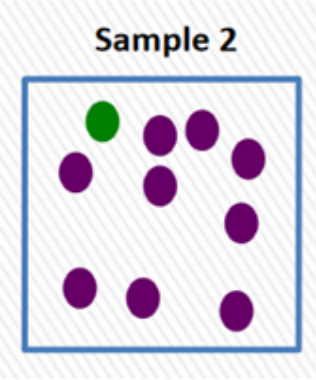
$\rightarrow$ **Pielou index is independent of species richness**

$$\textit{Pielou index} = \frac{Shannon\ Index\ (H)}{\ln(S)}$$

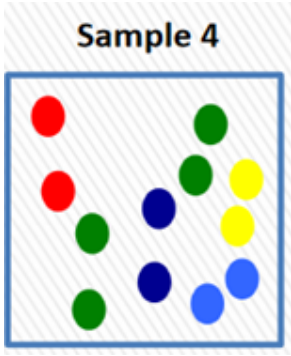
S is Species Richness

Simpson's diversity index: Combine taxa richness & evenness

Idea : Indicates the taxa dominance and gives the probability of two individuals that belong to the same taxa being randomly chosen



A value of 0.8 …
2 sequences randomly selected
have 80% chance to belong to the same ASV!



Simpson index = D

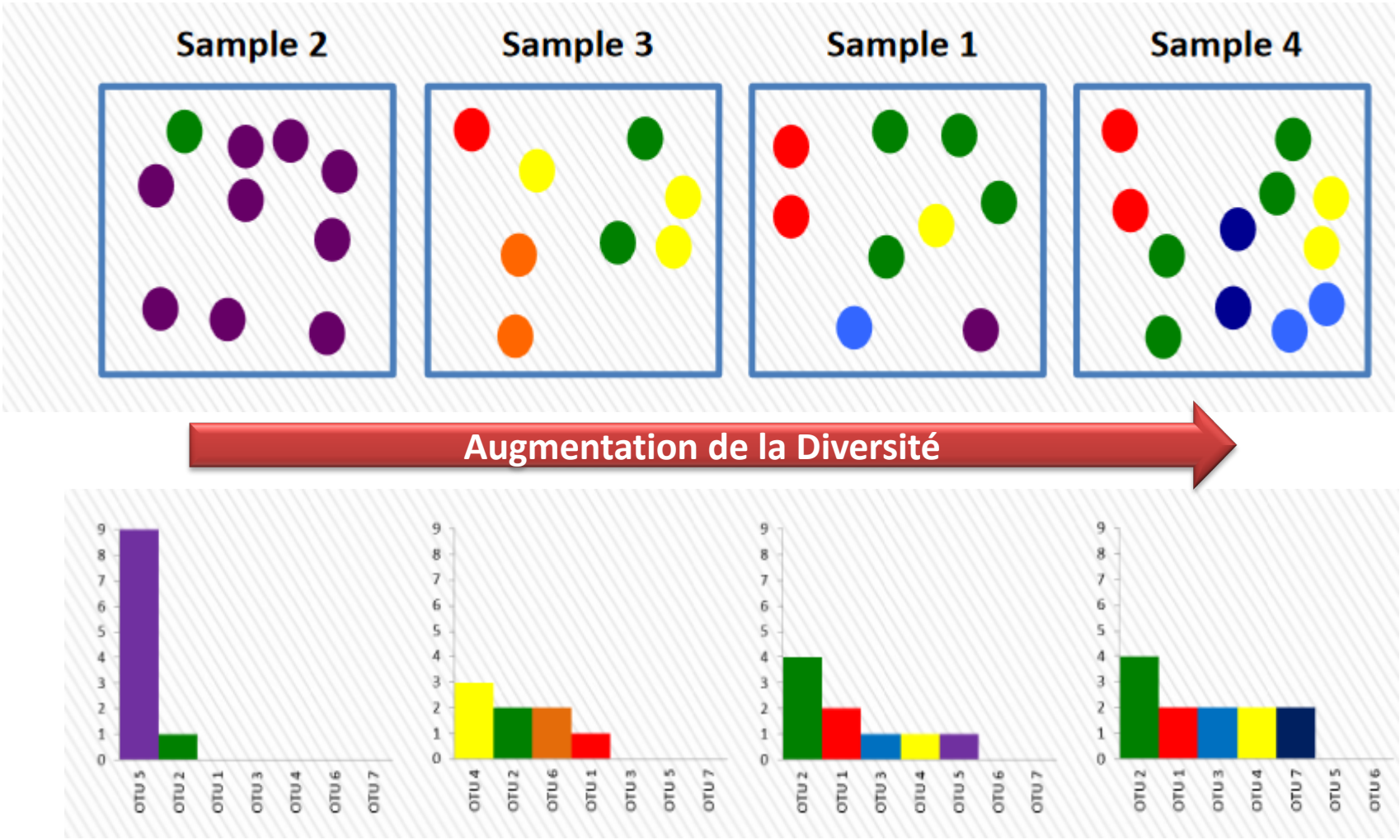
$$D = \sum_{i=1}^{S} p_i^2$$

P = proportion of the species

Gini-Simpson = 1 - D

$$E = 1 - \sum_{i=1}^{S} p_i^2$$

- Influenced by highly abundant Taxa
- Greater weight on evenness
- Range 0 to 1 (high)



Diversity =Richness + evenness

Diversity Estimators

- Chao1 & ACE are non-parametric estimators of taxa richness

- Sampling at infinity

- Good sampling gives you a total number of ASV/OTU observed not far from the Chao1 / ACE value (predicted for the sampled environment)

Chao1= S_{obs} + Adjustment (linked to the rare)

Chao1 adjustment

Idea : Rare taxa bring most information about the number of missing taxa

$$\frac{F_1(F_1-1)}{2(F_2-1)}$$

Singletons = F1

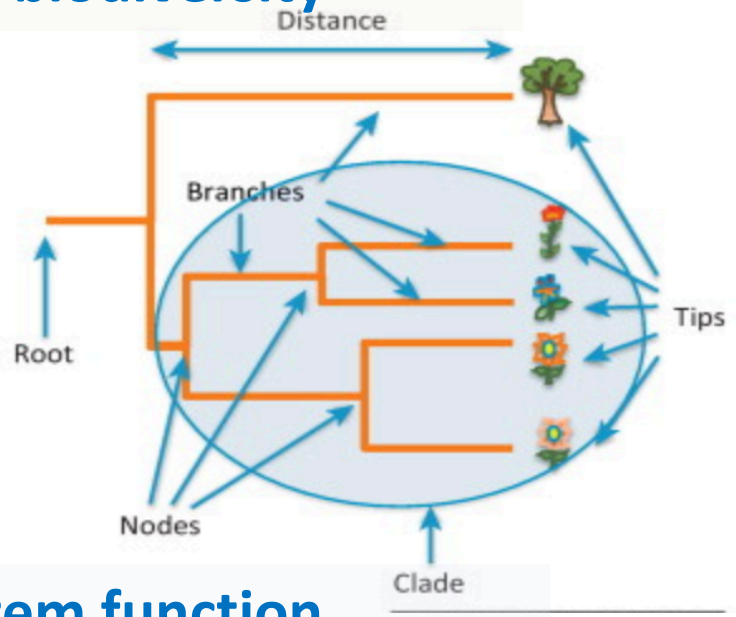
Doubletons = F2

Phylogenetic Indices

Phylogenetic Diversity (PD) measure of the evolutionary history within a set of species :

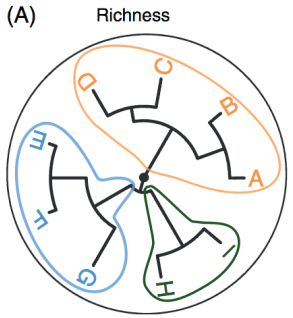
- Relatedness, speciation, events …

→ describes a fundamental aspect of biodiversity



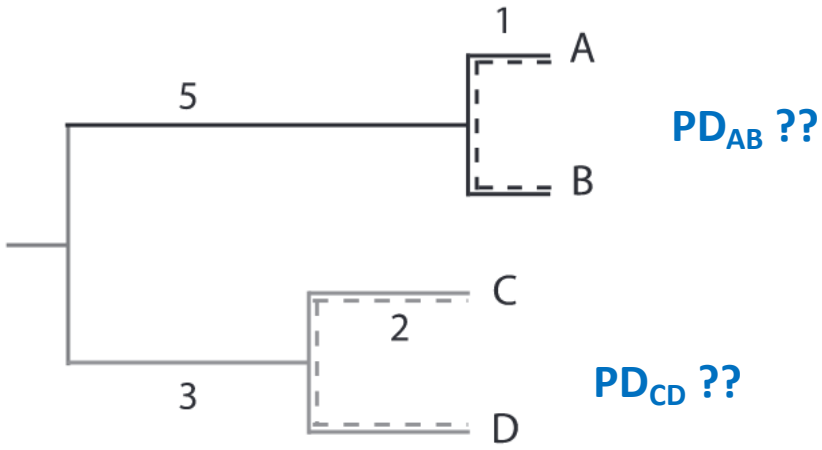
A better predictor of ecosystem function than species richness & evenness

Richness = How much ?



- **PD** $_{faith}$ (Faith's Phylogenetic Diversity)

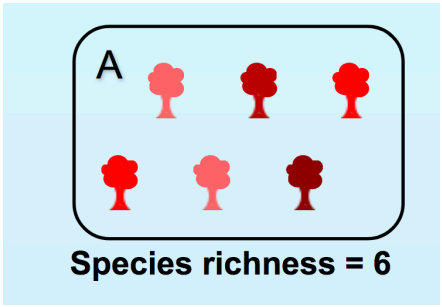
→ As the sum up of branch lenghts beetween the root to the tips! (Related to SR)



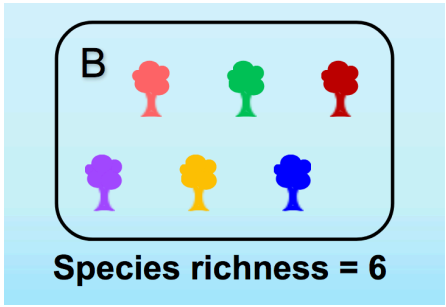
PD_{AB} ??

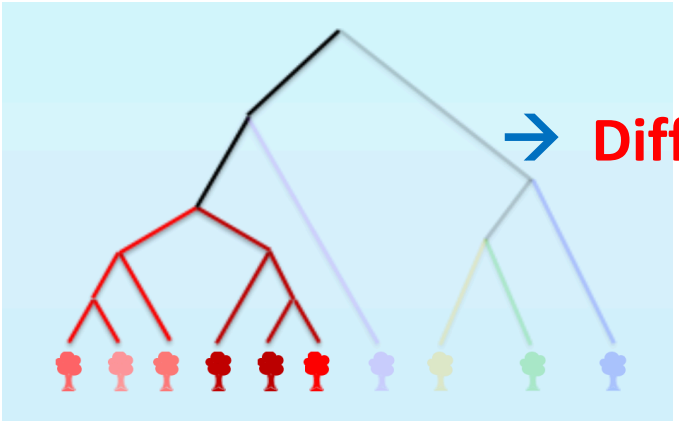
PD_{CD} ??

NB: Minimum spanning tree…

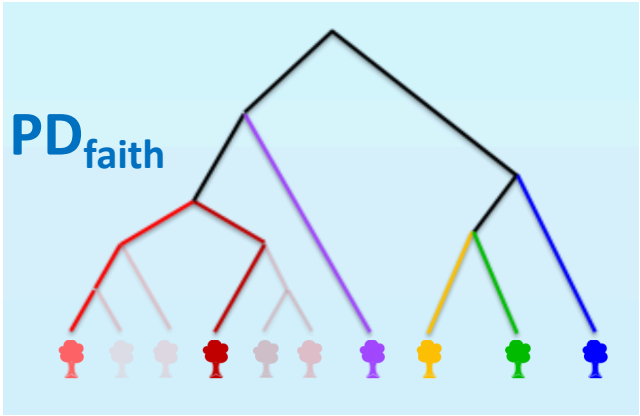


Same SR





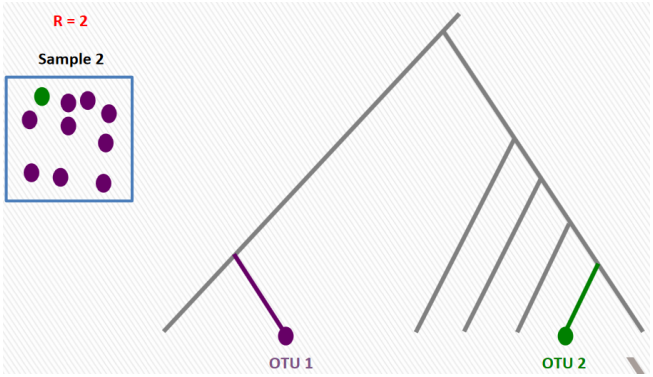
→ Different PD_{faith}

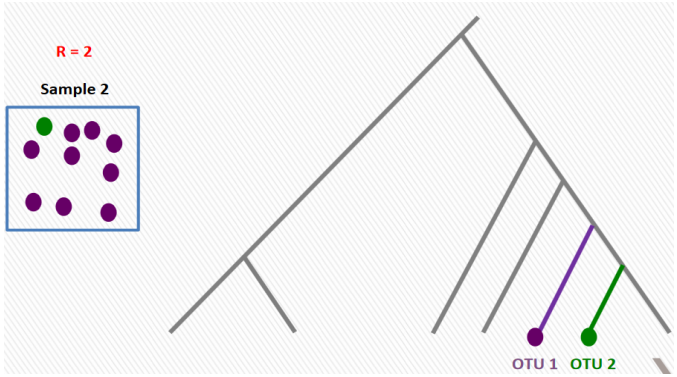


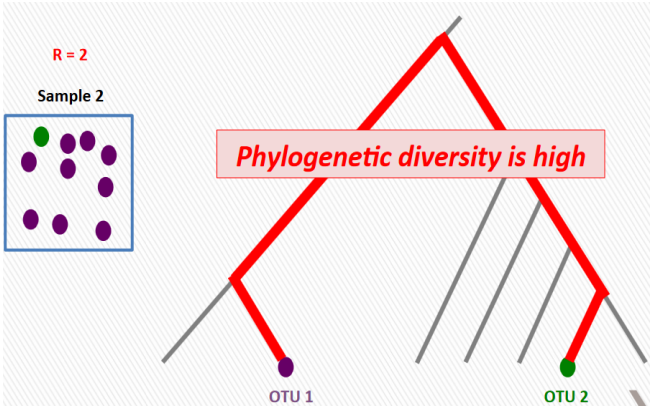
EnvB : Sum up branch lengths

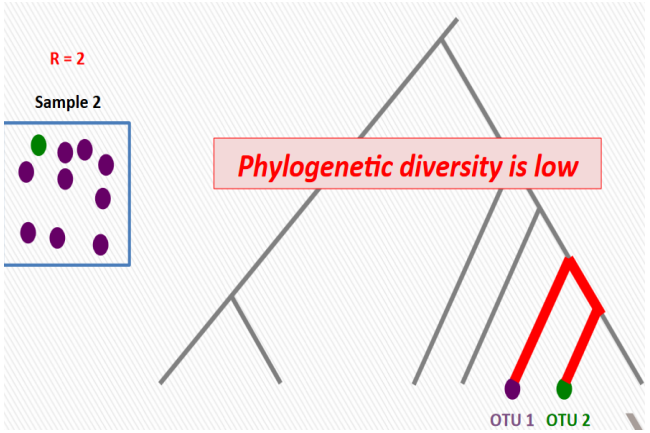
EnvA : Sum up branch lengths

Same SR & same evenness = Same Shannon/simpson



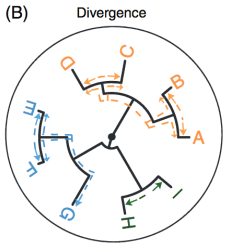






BUT different PD_{faith}!!!!

Divergence : Quantify the phylogenetic difference…



Why PD is a proxy of functional diversity, niche/community dissimilarity :

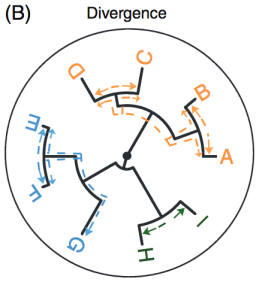
- **Closely related species tend to have similar functions/traits** (similar habitat requirement)
- **Distant related species tend to have greater complementary functions**

Meaning for the community assemblage :

Dominance of **closely related** species → **Clustering pattern**
Dominance of **distant related** species → **Overdispersion pattern**

Given species richness, does the **phylogenetic diversity** in **AN** assemblage is greater or less than that expected?

Divergence : How different?



Two commonly used metrics were used to quantify:
the Net Relatedness Index (NRI) and Nearest Taxon Index (NTI)

Highlight phylogenetic structure of assemblages **at different evolutionary depths**

NRI : Net Relatedness Index

- Based on the **Mean Phylogenetic Distances (MPD)** in each community.

→ **Average** phylogenetic **distance** of species (**to every other species)**

→ « Basal measure » : Clade representation. **Strongly influenced by the 'basal' structure** of the phylogenetic tree

$$NRI = -1 \bullet \frac{MPD - MPD_{rnd}}{sdMPD_{rnd}}$$

NTI : Nearest Taxon Index

- Based on **MNTD** the Mean Nearest phylogenetic Neighbor Distance

→ **Average** phylogenetic **distance** to the **nearest neigbour**

→ **Reflect Phylogenetic structure of the tree tips**

MNTD same as MNND

$$NTI = -1 \bullet \frac{MNND - MNND_{rnd}}{sdMNND_{rnd}}$$

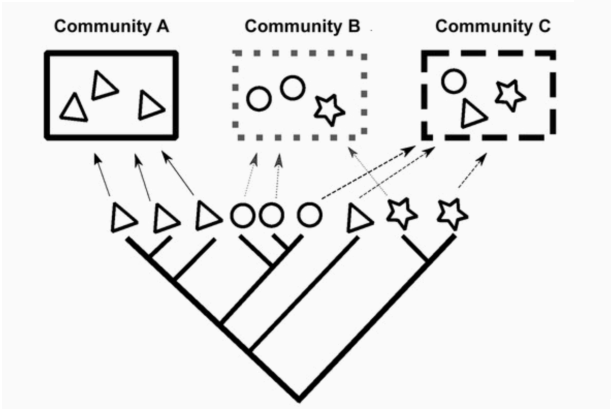
So get NRI /NTI values .. And so what ????

The « Null model » : Phylogeny randomization…

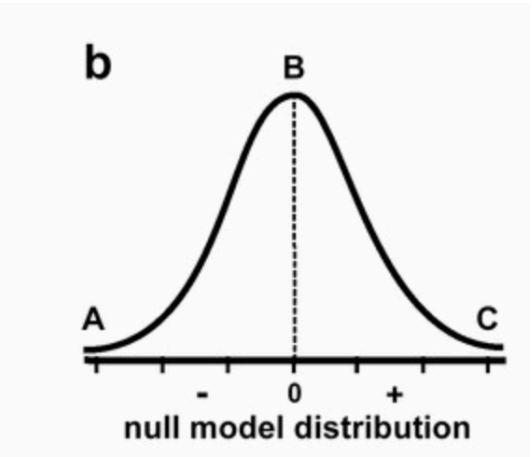
Need a **reference** for comparison → Absence/overdispersion/Clustering!

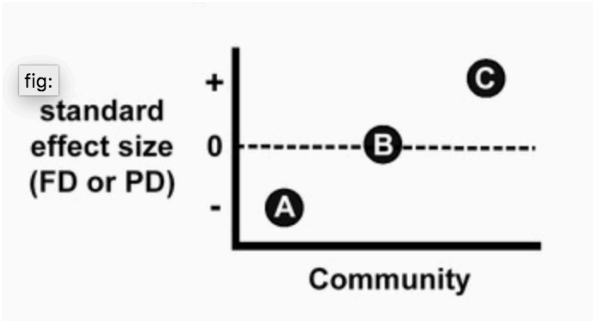
→ A distribution « Null Model » based on random taxa positions within tree

→Is the measure for a specific community is more or less expected by chance ?



Randomization





SES for your values

NTI and NRI are Z scores!!

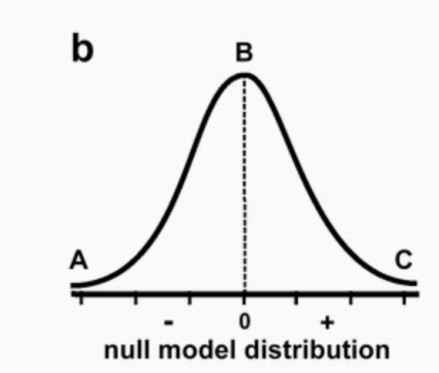
Interpretation

- A negative NRI/NTI value indicates an overdispersed phylogeny where taxa are less related to each other than expected by chance

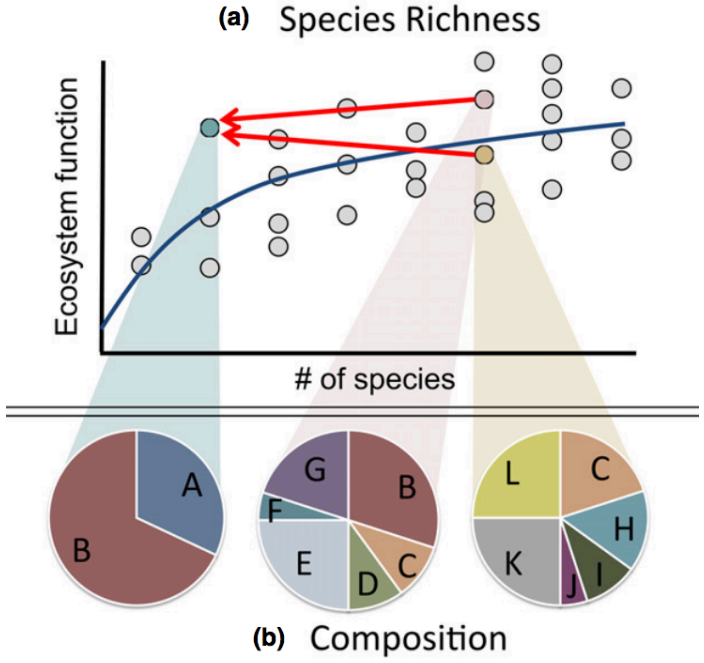
→ Significance < -1.96 (why this score?)

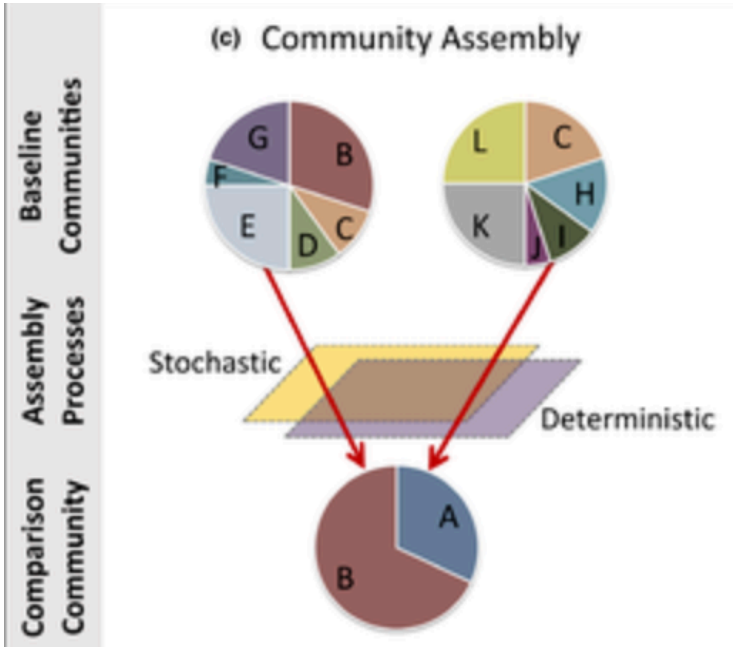
- Positive NRI/NTI values indicate a clustered phylogeny where taxa are more related to each other than expected by chance

→ Significance > 1.96



How a community assembles?
Which drivers determine patterns of species diversity/composition?





Community assembly : Spatial and temporal processes

- Niche-driven = Deterministic

- **Selection**

- Biotic interactions (taxa interactions)

- Environmental filtering

(=Abiotic conditions, physiochemical)

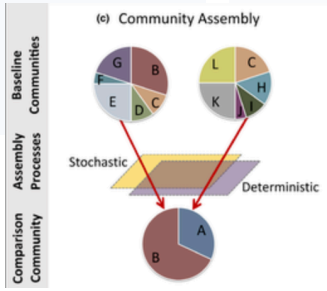
- Neutral = Stochastic process

- **Unpredictable**

- Random profileration, dispersal

- Random birth-death events

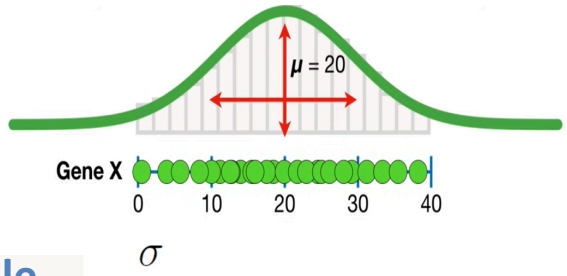
Ecological drift (loss diversity, small pop)



→ See β NTI/NRI

Z-score transformation = How far from the mean Am I ?

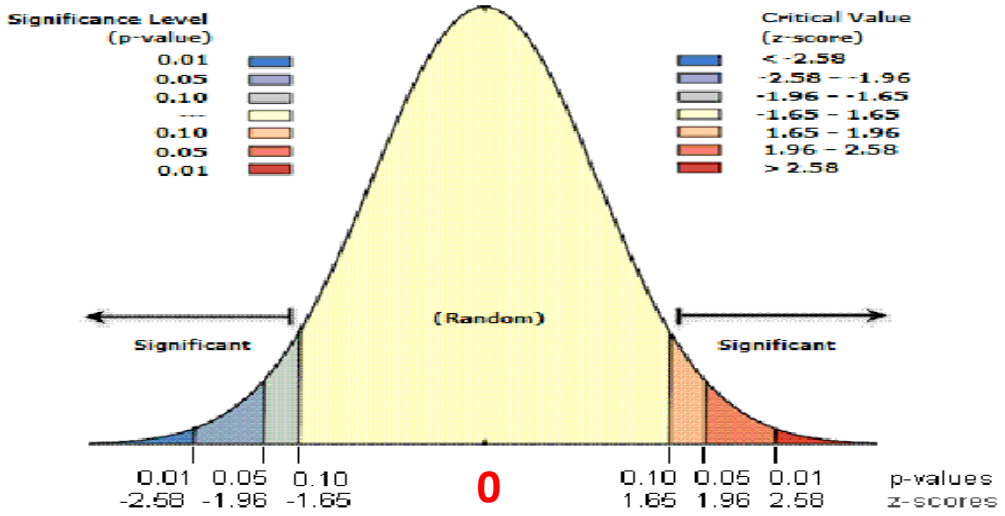
Heterogeneity of Units : environmental data
→ Fonction *scales* dans package Vegan.



Centering = substitute the mean to each value of the variable
Reduction = divide each value by the SD
→ Z-Scores

μ = Mean
σ = Standard Deviation

$$z = \frac{x - \mu}{\sigma}$$



What is Compositional Data (CoDA)?

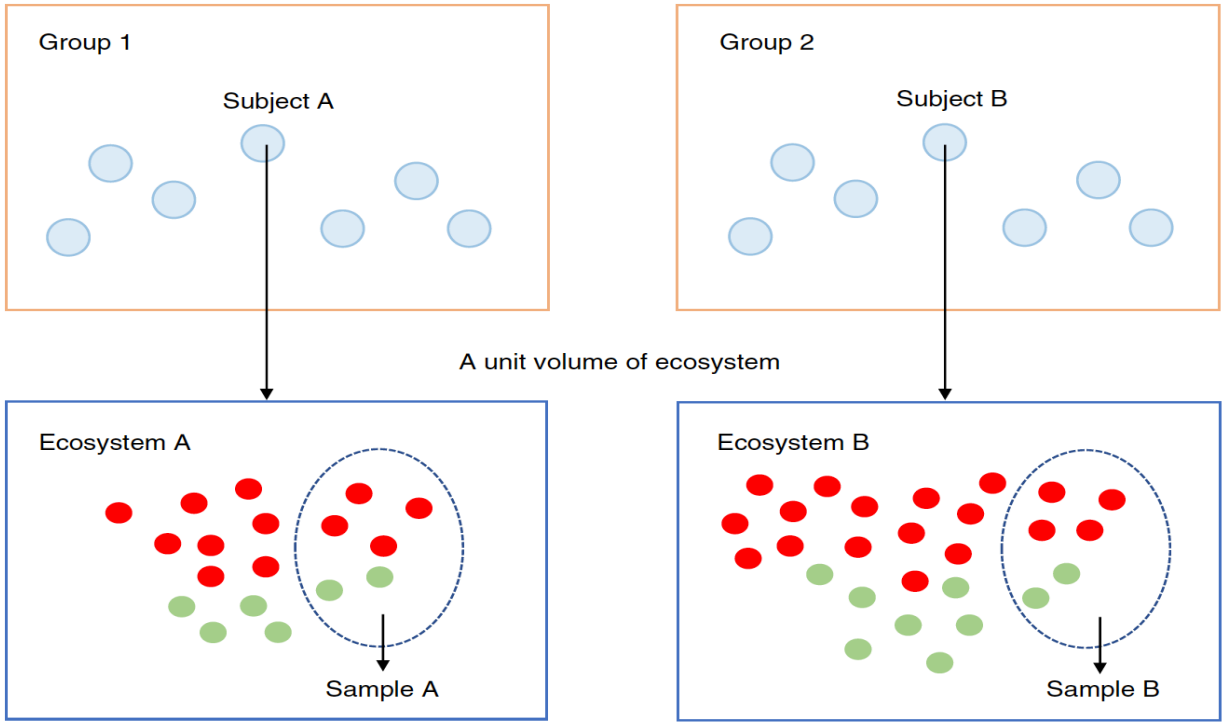
Describe a data set in which the parts in each sample have an arbitrary/constant sum (high-throughput DNA sequencing, pourcentage, probalities…) = A closed System

known as problematic, multivariate data analysis approaches such as ordination and clustering and univariate methods that measure differential abundance are invalid

BUT in the facts (publications)
→ CoDA is still in its infancy, and requieres sometimes strong mathematical background to understand it! Most of studies do not apply CoDA…

Pionneer addressing CoDa: Pearson & Aitchison

Sampling Fraction issue



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

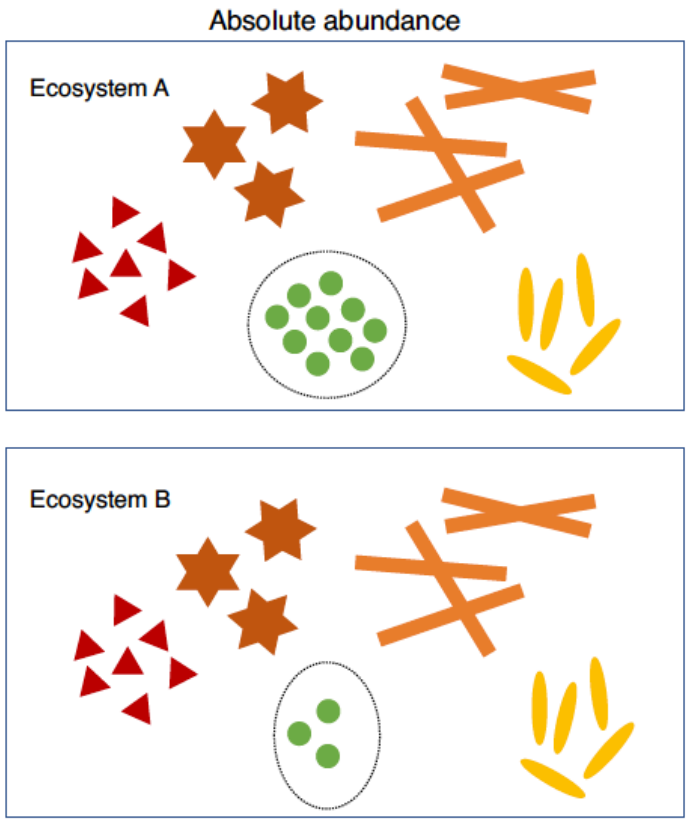
† Library size; ‡ microbial load.

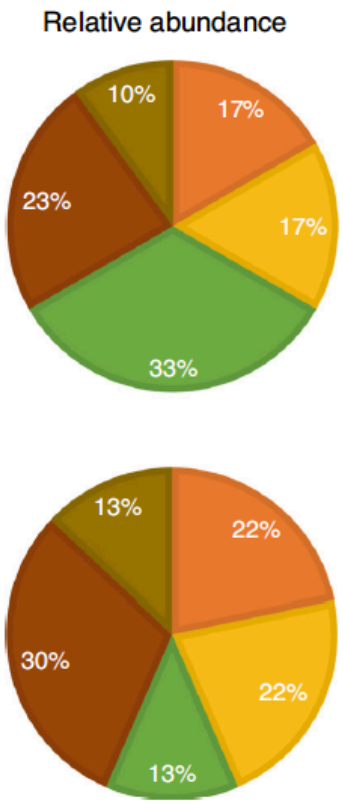
Sampling Fraction : ratio

Means→ Library Size/Microbial load

$$\frac{\text{Expected Absolute abondance of taxa in Sample}}{\text{Absolute Abundance in the Ecosystem}}$$

Absolute abundance vs. Relative Abundance





Absolute Ab. → Only Green taxa differ

Relative Ab. → All taxa differ

Changing one taxon modifies all the other!!!!!

Consequences of fraction size & relative abundances

- Known as « closed » system (a part of a sum) = Compositional Data
- Relative Ab. of one taxa impact all the others : not independent

→ Increase **false positive** in **Differential Abundance analyses**
→ Increase **spurious correlations!!**

Solutions

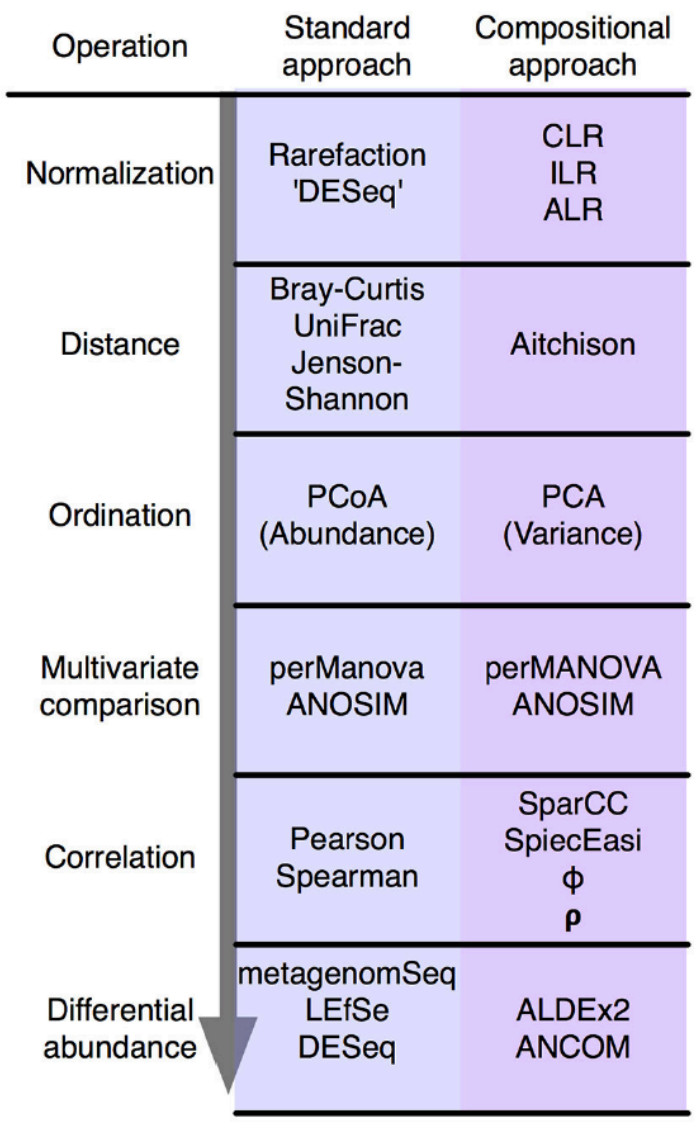
- Normalized the data: **Sampling fraction** and not only library size

- Deal with sampling fraction and Compositional data → CoDA
  → Use Linear regression, log ratio transformation (CLR, ALR, ILR, PhyLR)

DA: ANCOM, Aldex2, Deseq2 …
Correlations: SparCC, SpiecEasi

→ Ask me for bibliography If you want to go further

Standard approach vs. Coda



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