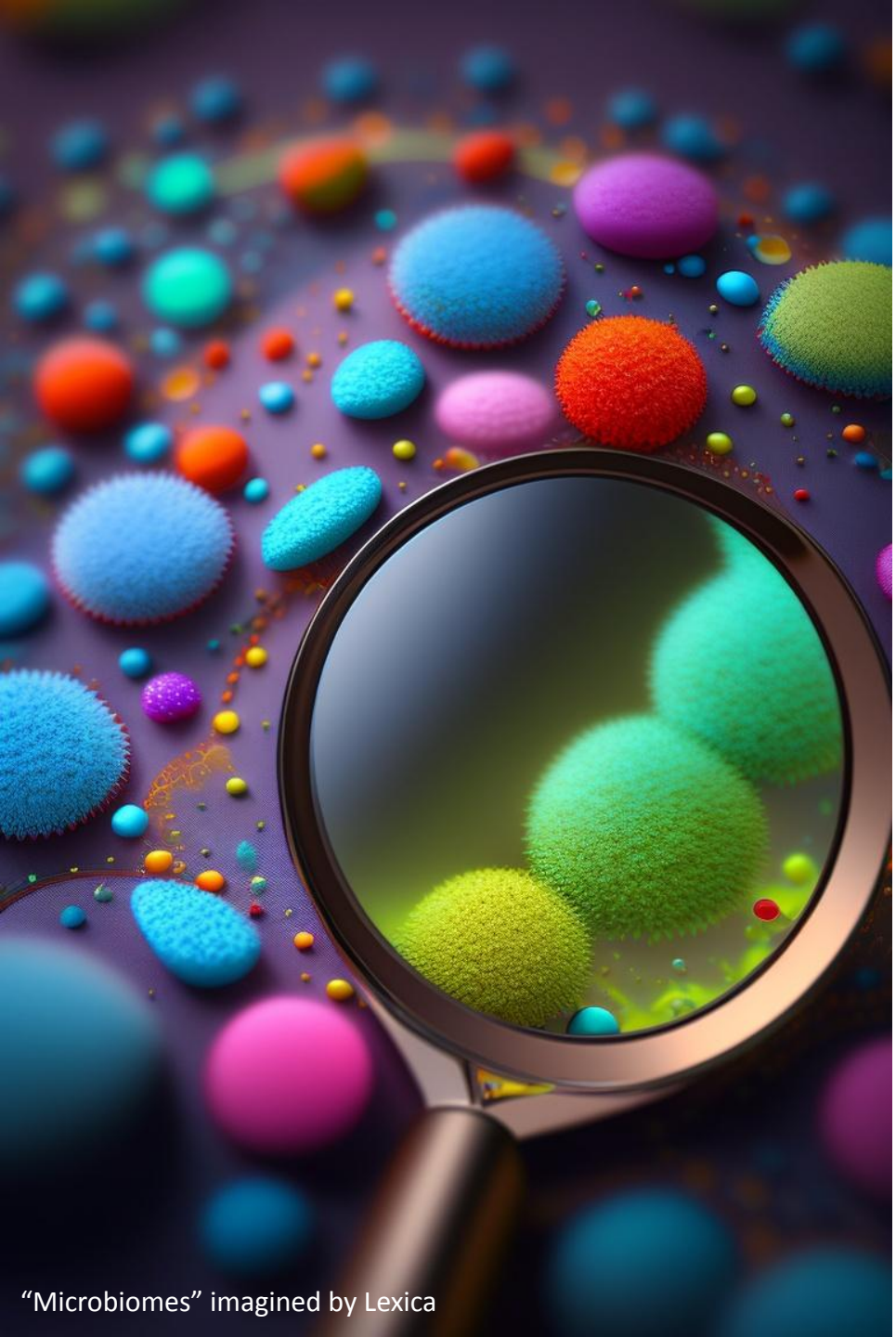


"Microbiomes" imagined by Lexica

Microbiome data analysis

Stats and Plots in R

Evomics - Krumlov

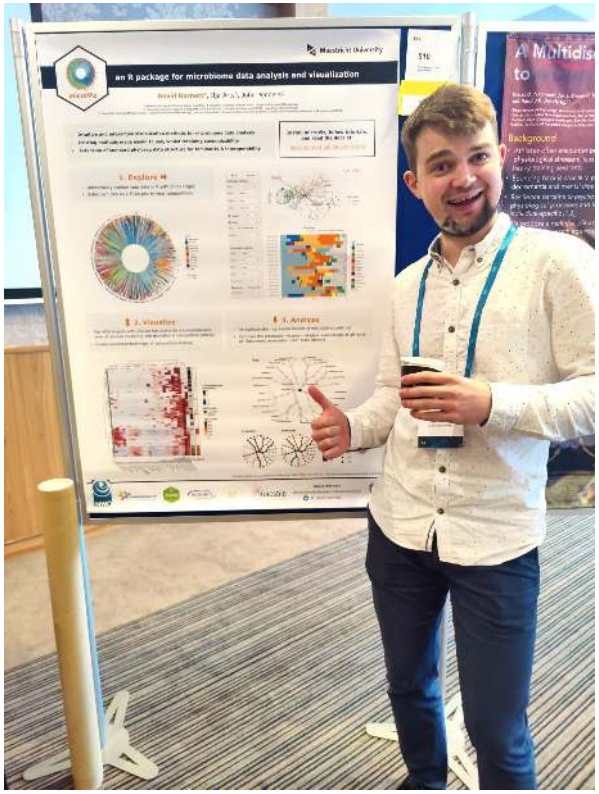
16th Jan 2024

David Barnett

github.com/david-barnett

david.barnett@maastrichtuniversity.nl

David



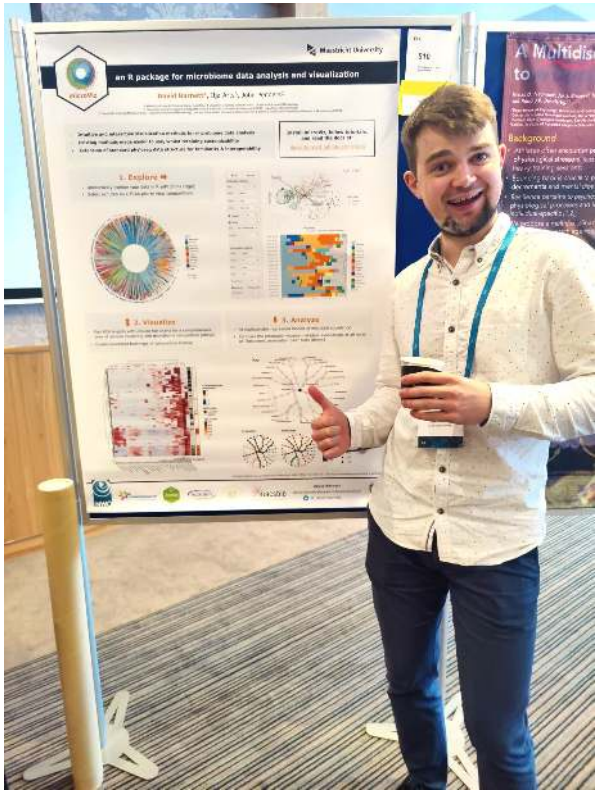
David Barnett

Bioinformatician

Maastricht University, NL

- Epidemiology MSc
- Infant gut microbiome PhD
- Bioinformatics “Postdoc”
  - Medical Microbiology

David

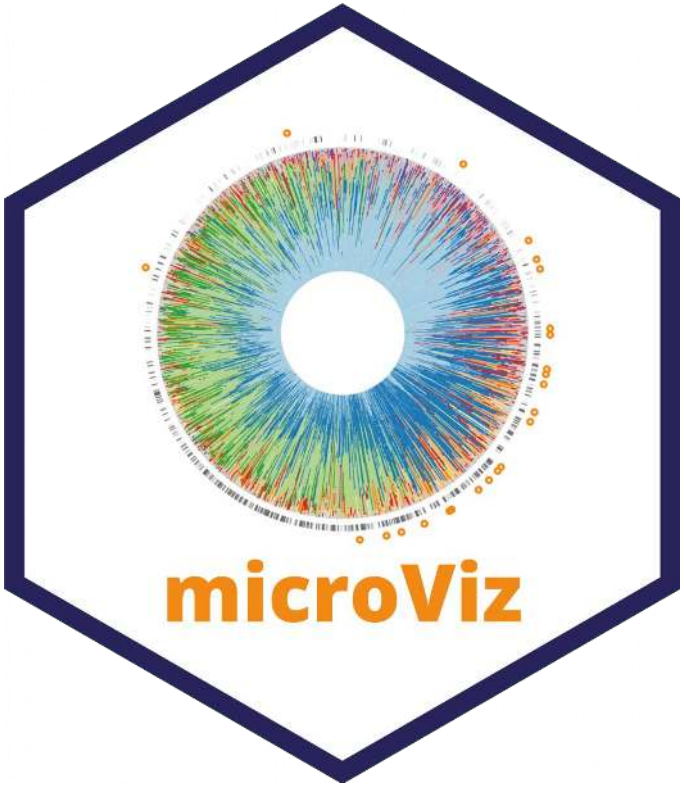


David Barnett

Bioinformatician

Maastricht University, NL

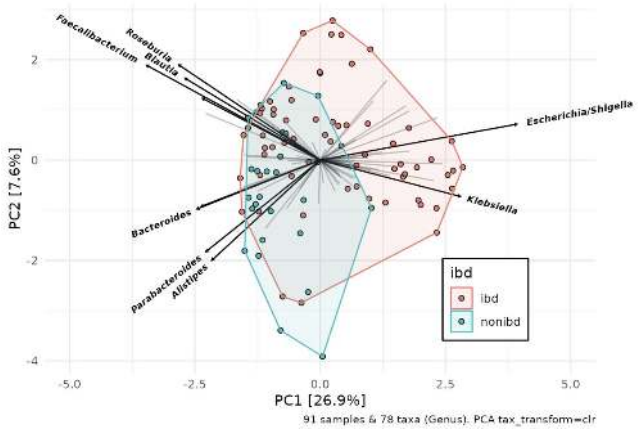
R package



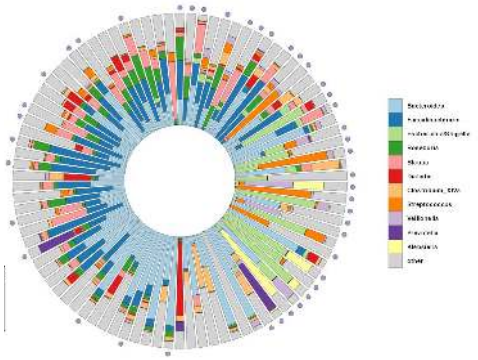
Tools for microbiome data visualization & statistics

microViz

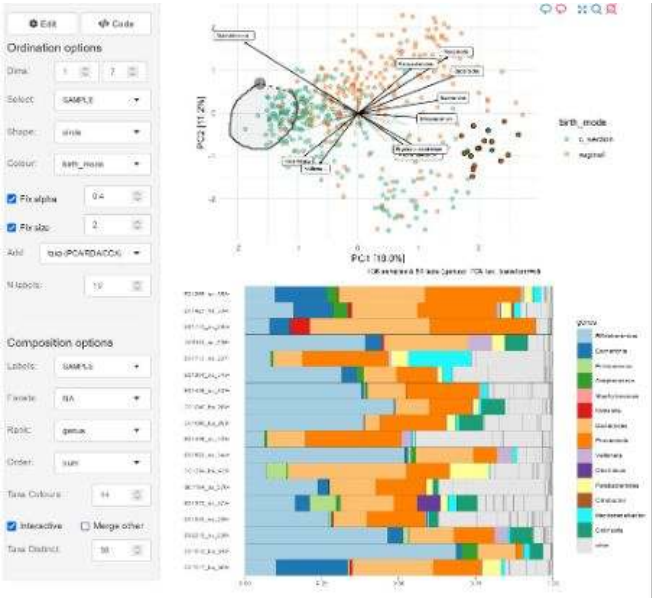
Ordination plots



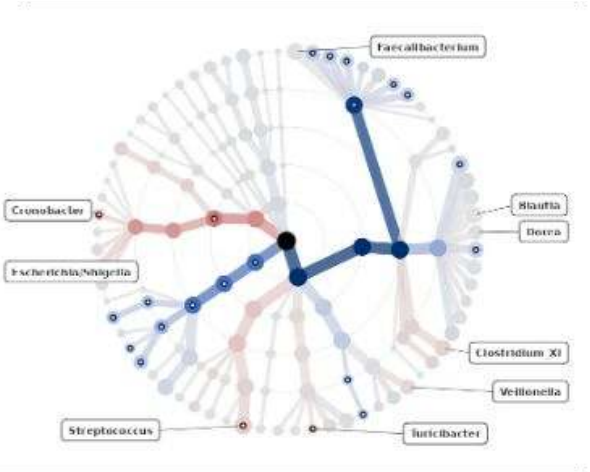
Bar charts



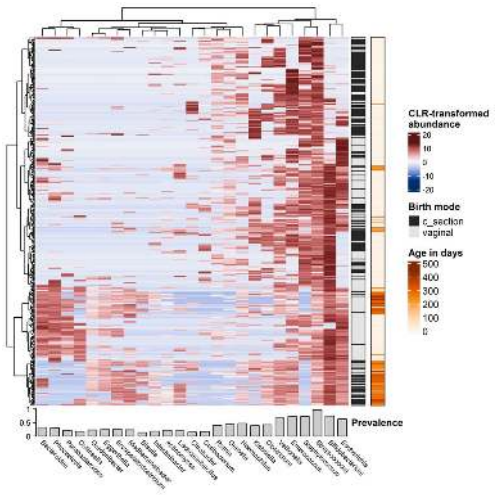
Interactive data exploration



Taxonomic association trees



Annotated heatmaps



The plan: 2 parts

A

- **Key concepts in microbiome taxonomic data analysis**

  19:00 - 19:30 - Lecture

- **Barcharts and Diversity - getting started in R**

  R Exercises - about 45 mins

- - - - - - - - - - - Take a break <u>before</u> 20:30 - - - - - - - - - - -

B

- **Dissimilarity, Ordination, & Differential Abundance**

  20:30 - 21:00 - Lecture

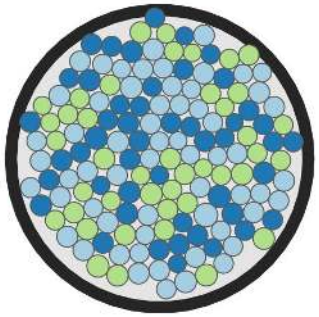
  R Exercises - until 22:00

Key concepts in microbiome taxonomic data analysis

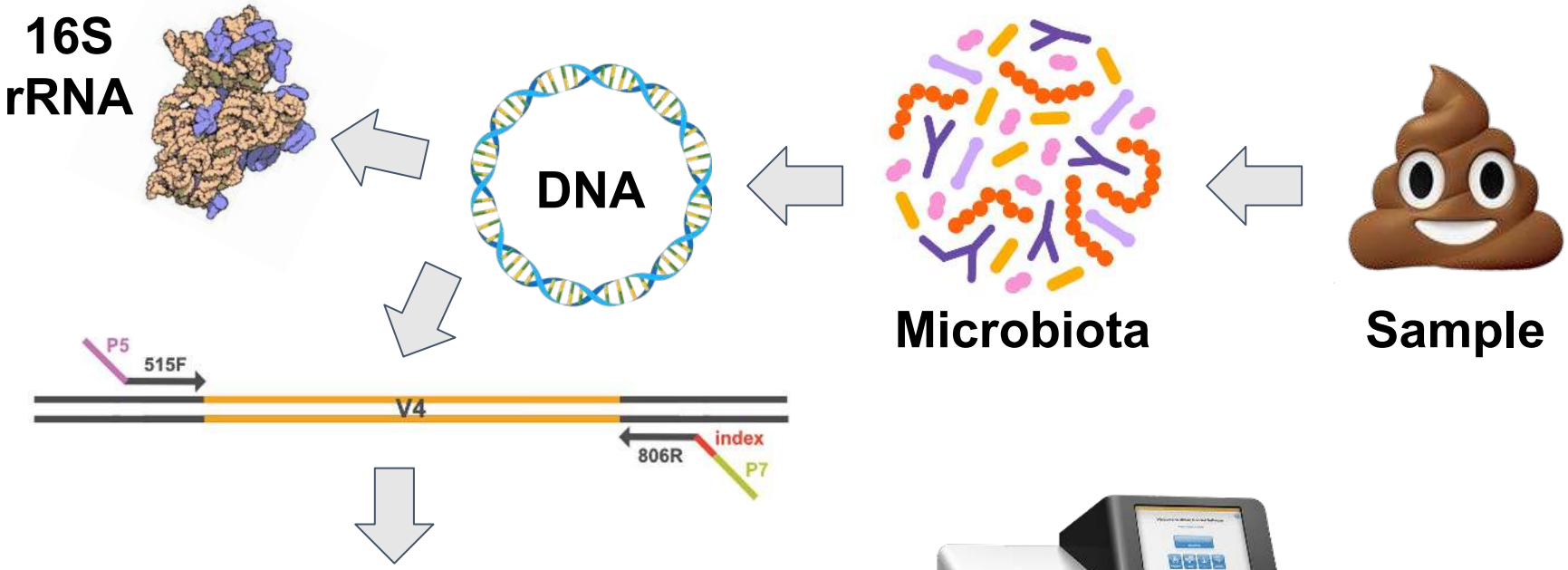
1. **Processing 16S gene amplicon sequencing data**

   - Denoising with: DADA2

2. **Analysing taxonomic info** (16S / ITS / Shotgun / etc...)



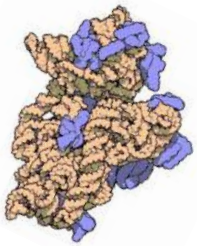
   - Diversity
   - Dissimilarity
   - Differential Abundance

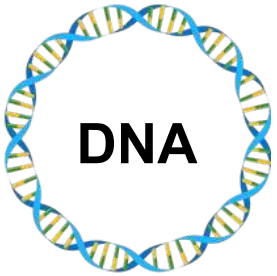


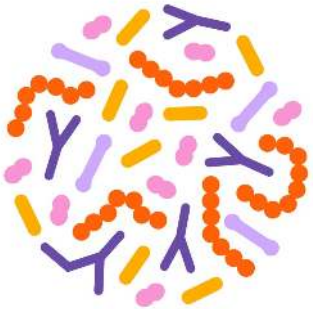
Sequencing: of 16S rRNA gene amplicons

Microbiota profiling - who's there?

16S rRNA

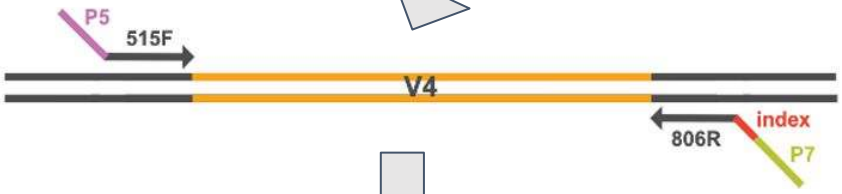






Microbiota

Sample



Sequencing: of 16S rRNA gene amplicons

DADA2

Amplicon Sequencing. **Exactly.**

Denoising: Infer Amplicon Sequence Variants

Taxonomic classification: map ASVs to database

silva

high quality ribosomal RNA databases

DA^2 vs. OTUs

Sample sequences
(Ground truth)

Sequencing reads
(Our raw data)

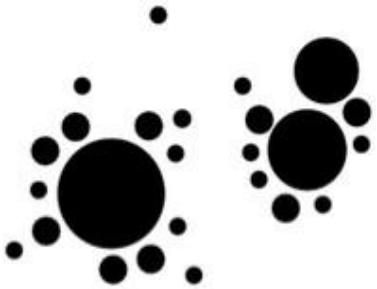
Operational Taxonomic Units
(~97% similar sequences)

Errors

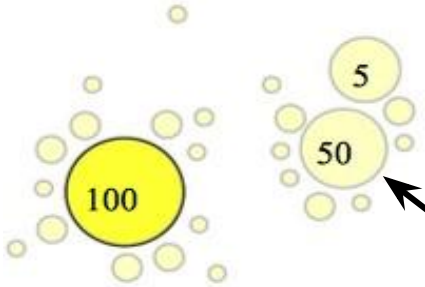
Cluster into OTUs

DADA2

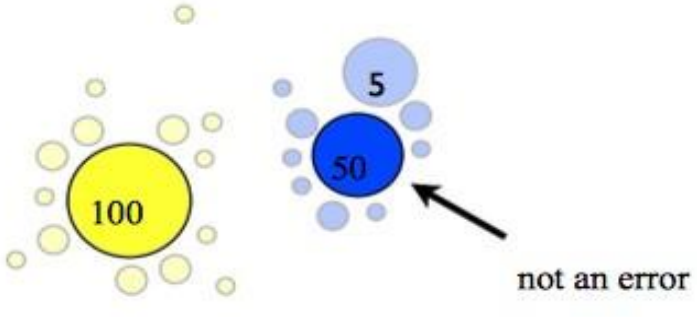
Divisive Amplicon Denoising Algorithm



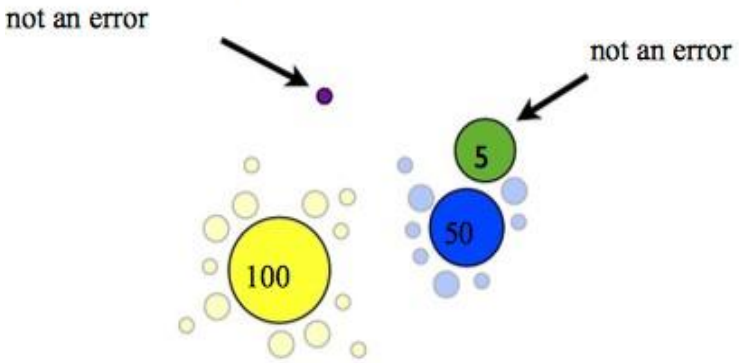
Count unique sequences



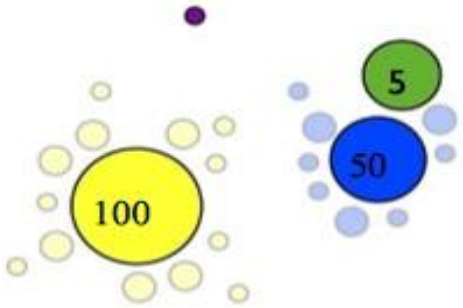
Initial error model
(one partition)



Reject sequence least likely to arise from errors
(too abundant and/or different)



Repeat: Reject more sequences and divide into further partitions

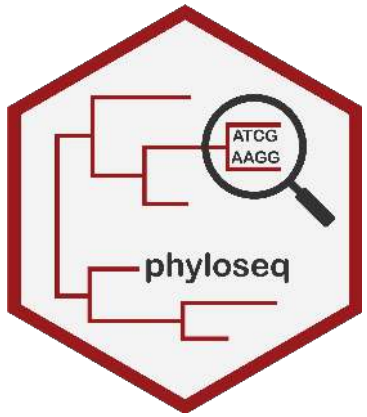


Convergence: All errors are plausible

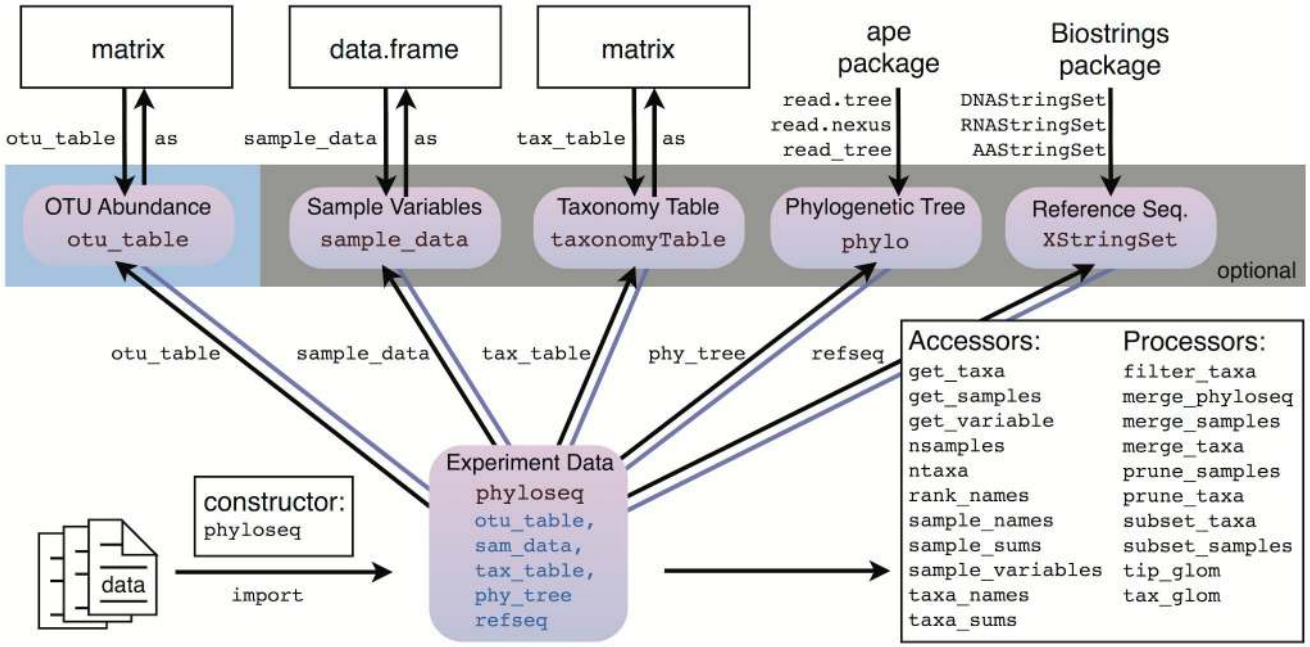
Callahan et al. 2016. Nature Methods

DADA2

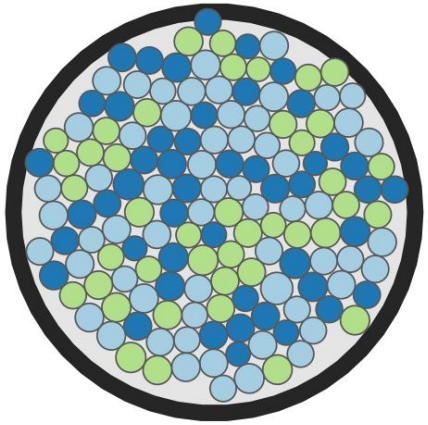
Amplicon Sequencing. **Exactly.**



https://benjjneb.github.io/dada2/tutorial.html



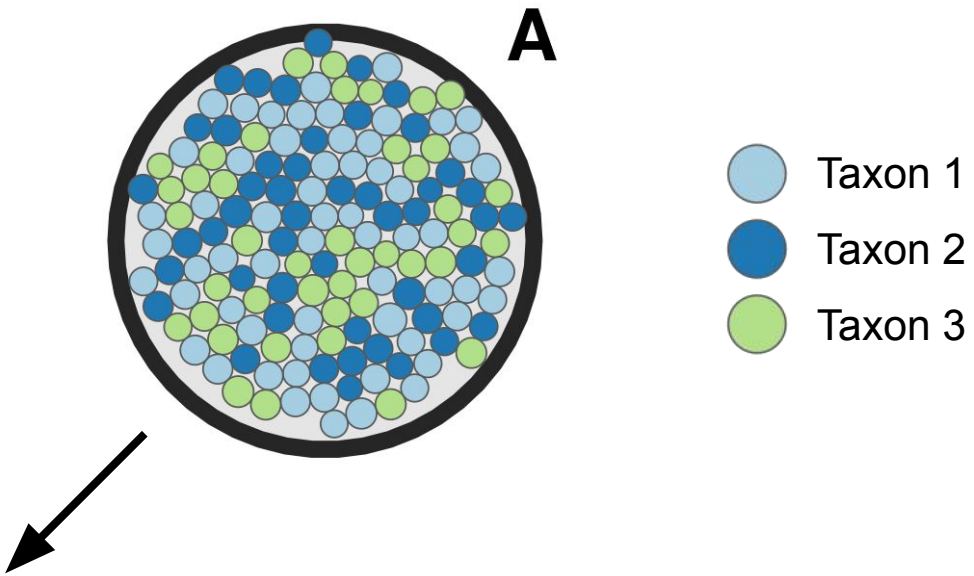
Analysing taxon abundance data



?

1. Diversity
2. Dissimilarity
3. Differential Abundance

Ecosystem **Diversity**

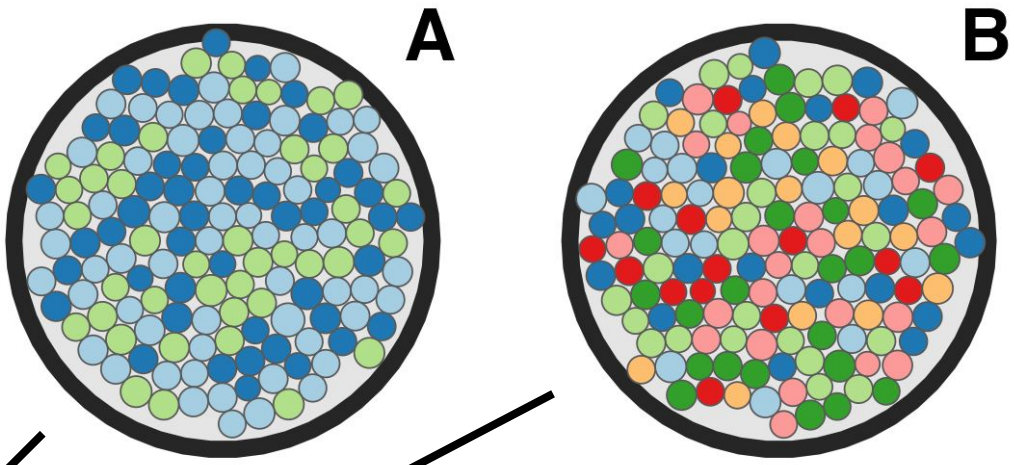


“Observed”

“Pielou’s”

“Effective Shannon”

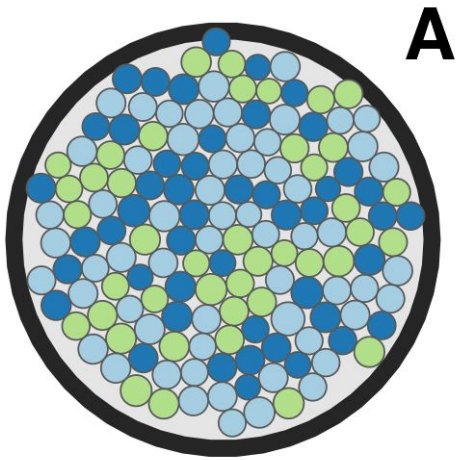
Ecosystem **Diversity**

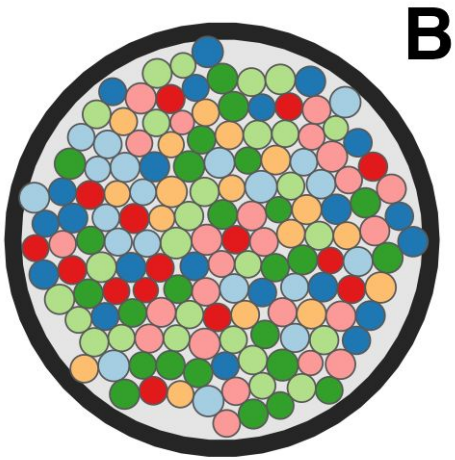


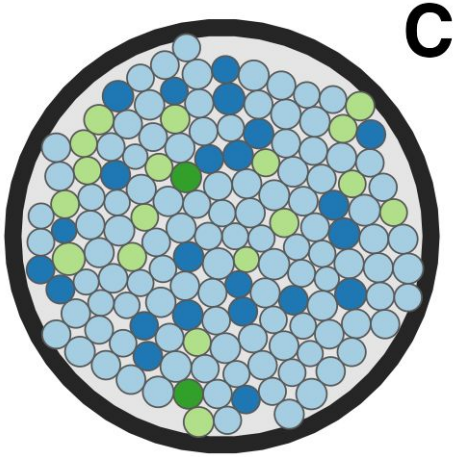
A

Richness	3.0
Evenness	1.0
Diversity	3.0

Ecosystem **Diversity**

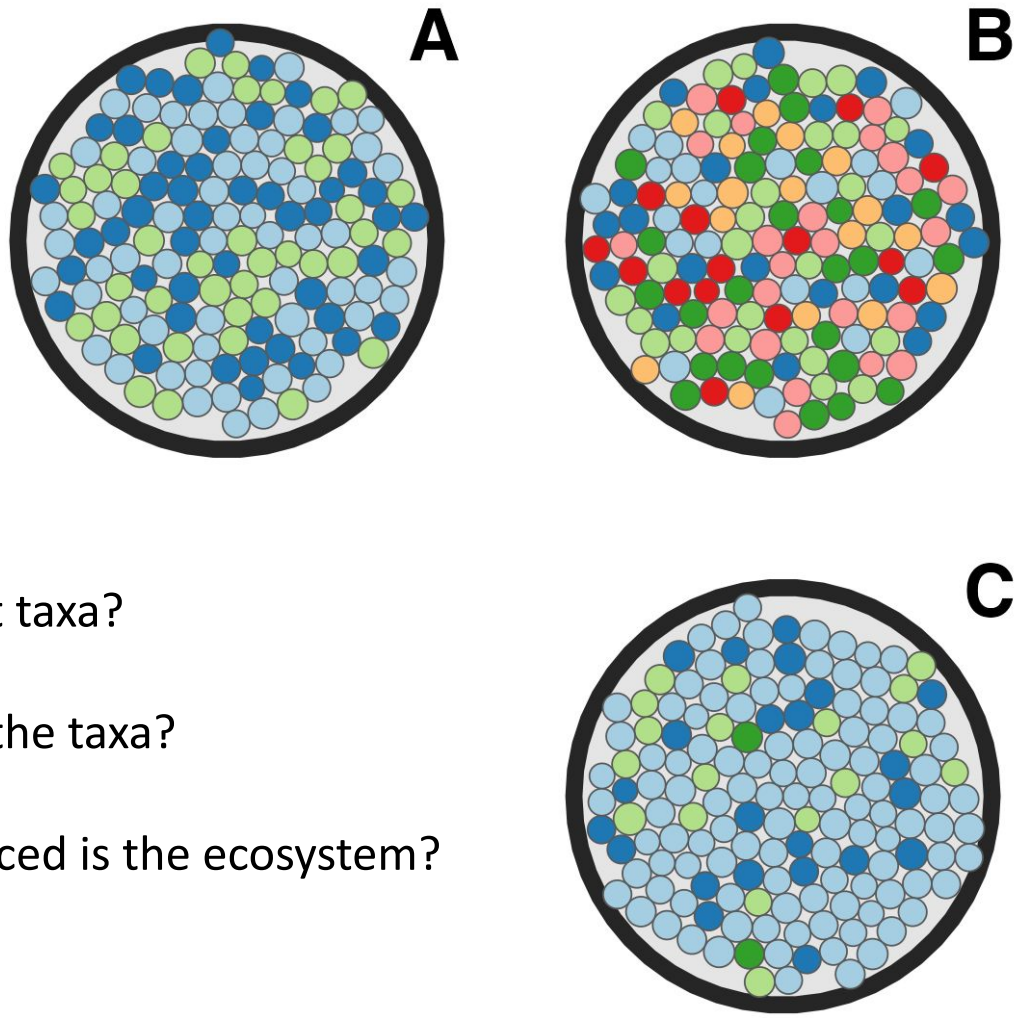






	A	B	C
Richness	3.0	7.0	4.0
Evenness	1.0	1.0	0.6
Diversity	3.0	7.0	2.3

Ecosystem **Diversity**



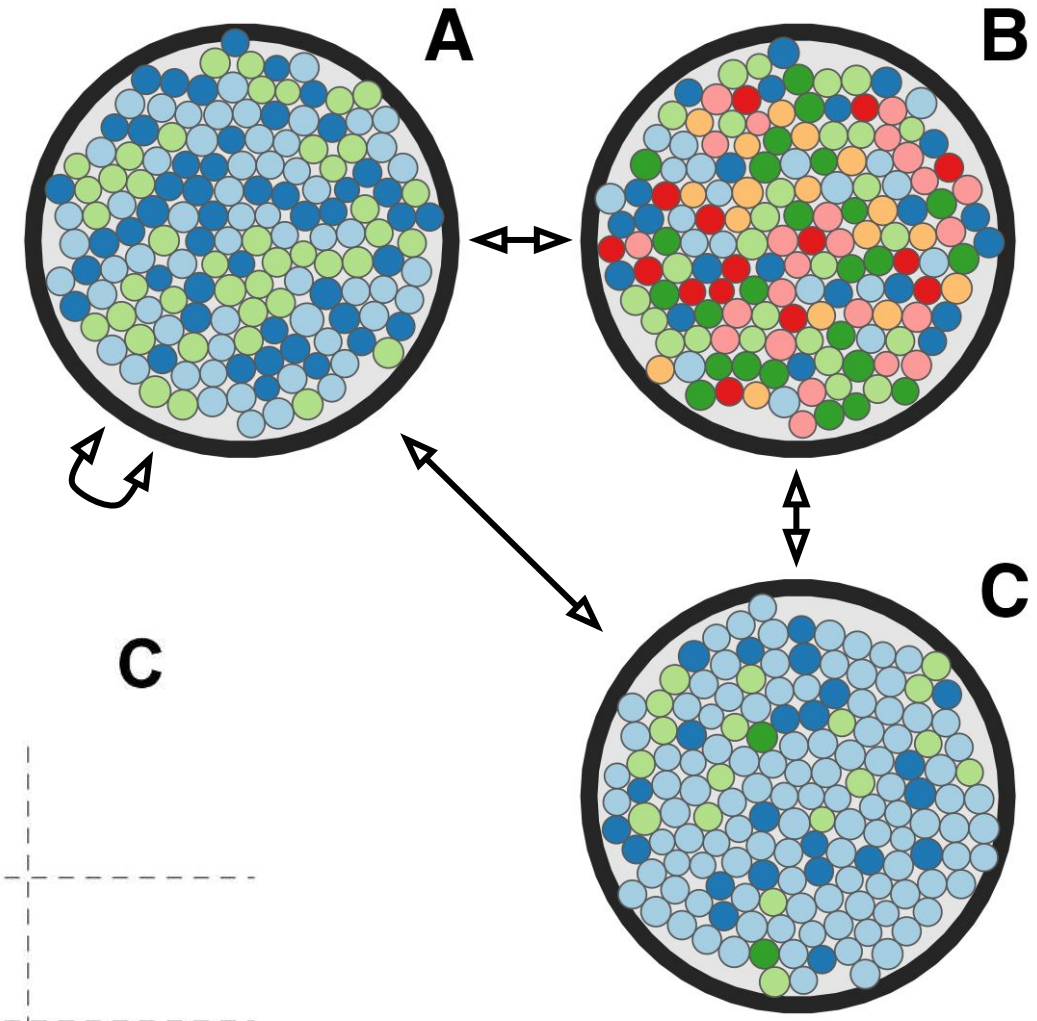
- **Richness** - how many different taxa?
- **Evenness** - how balanced are the taxa?
- **Diversity** - how rich and balanced is the ecosystem?

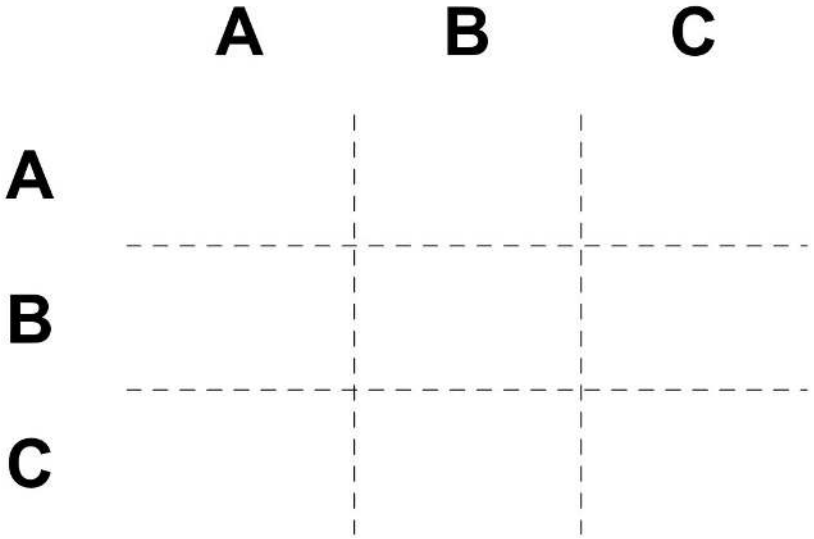
There are many different ways to calculate these things!

More about diversity indices -> https://www.davidzeleny.net/anadat-r/doku.php/en:div-ind

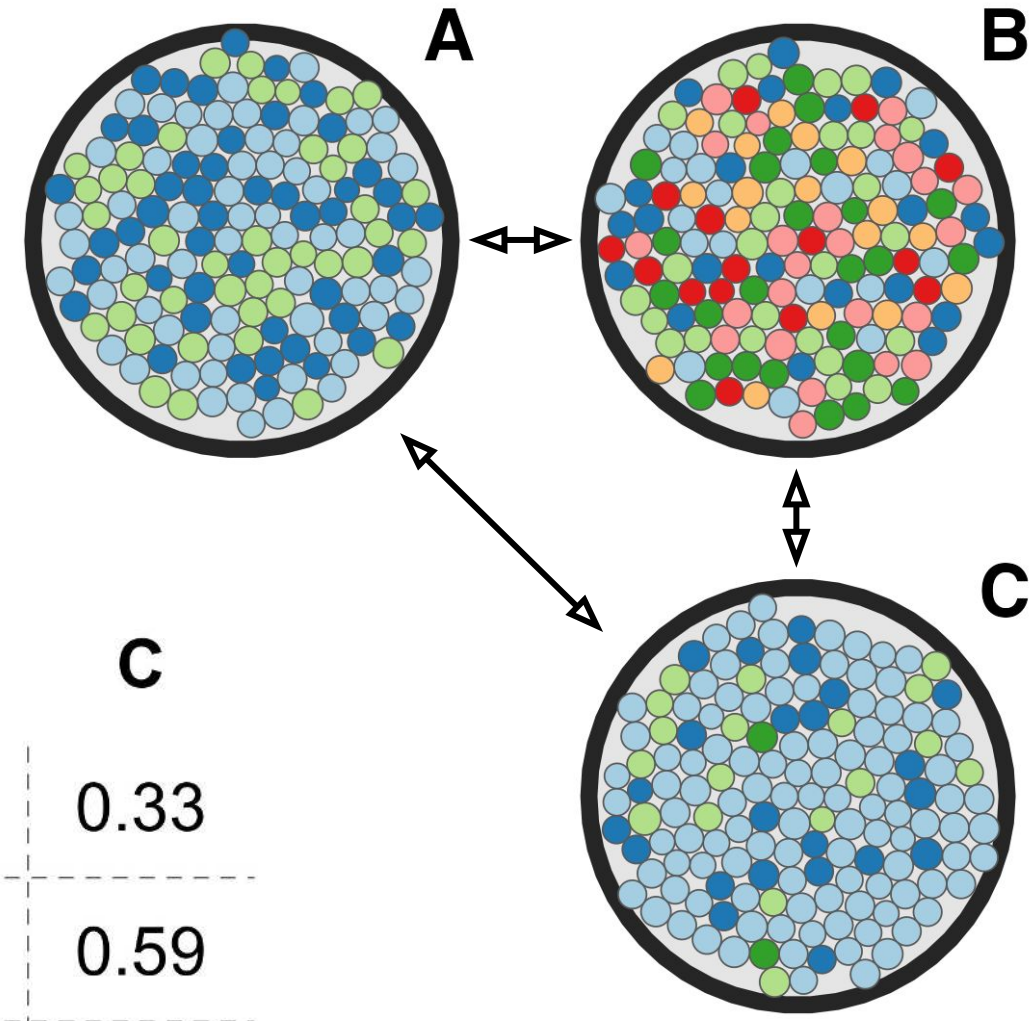
Dissimilarity

between ecosystems





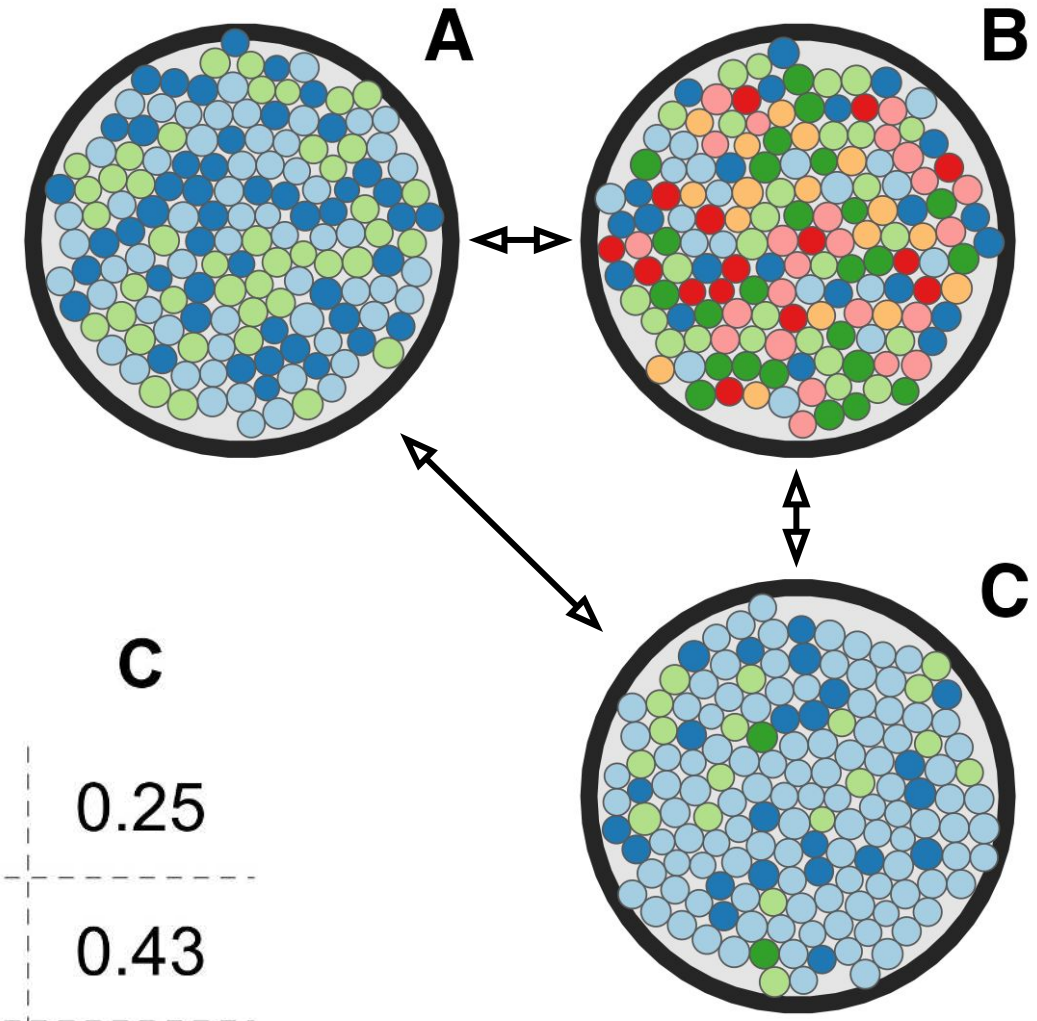
Dissimilarity between ecosystems



	A	B	C
A	0.00	0.54	0.33
B		0.00	0.59
C			0.00

Bray-Curtis Dissimilarity

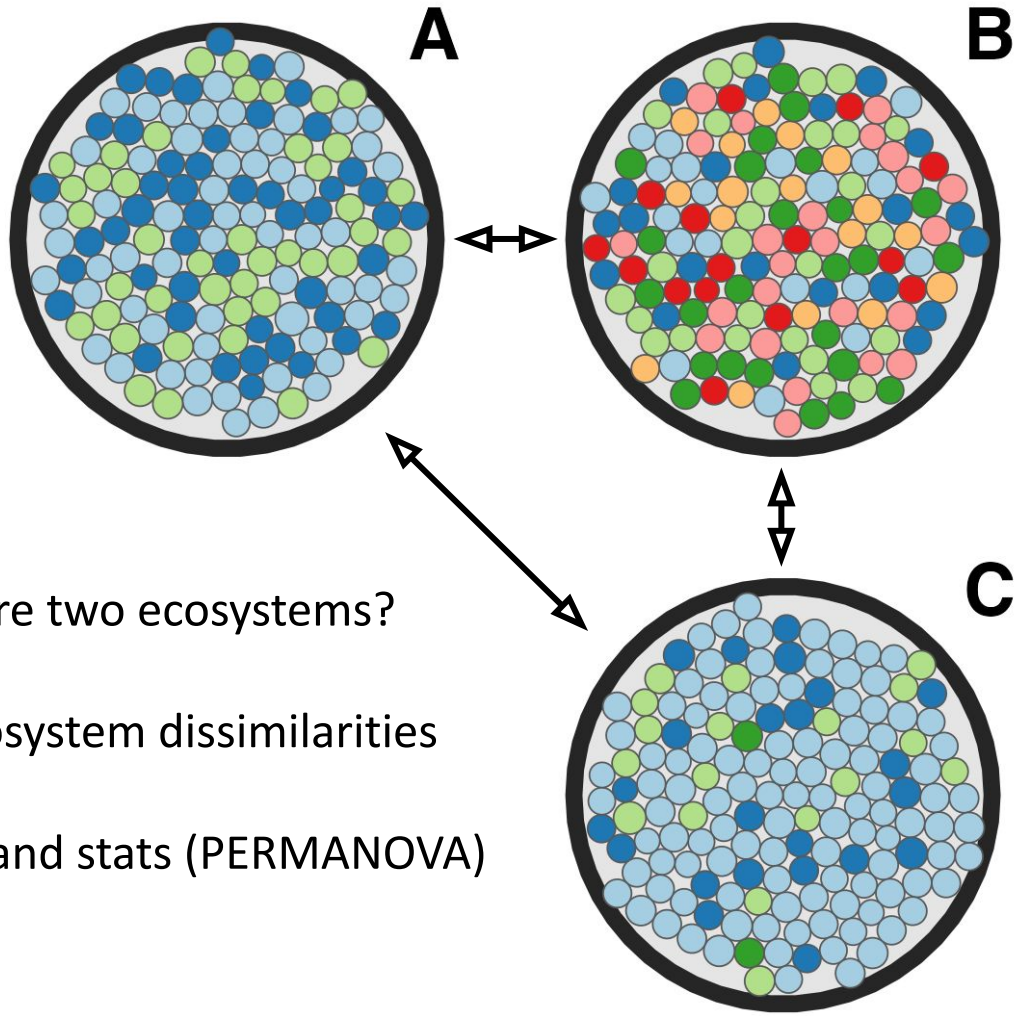
Dissimilarity between ecosystems



	A	B	C
A	0.00	0.57	0.25
B		0.00	0.43
C			0.00

Binary Jaccard Distance

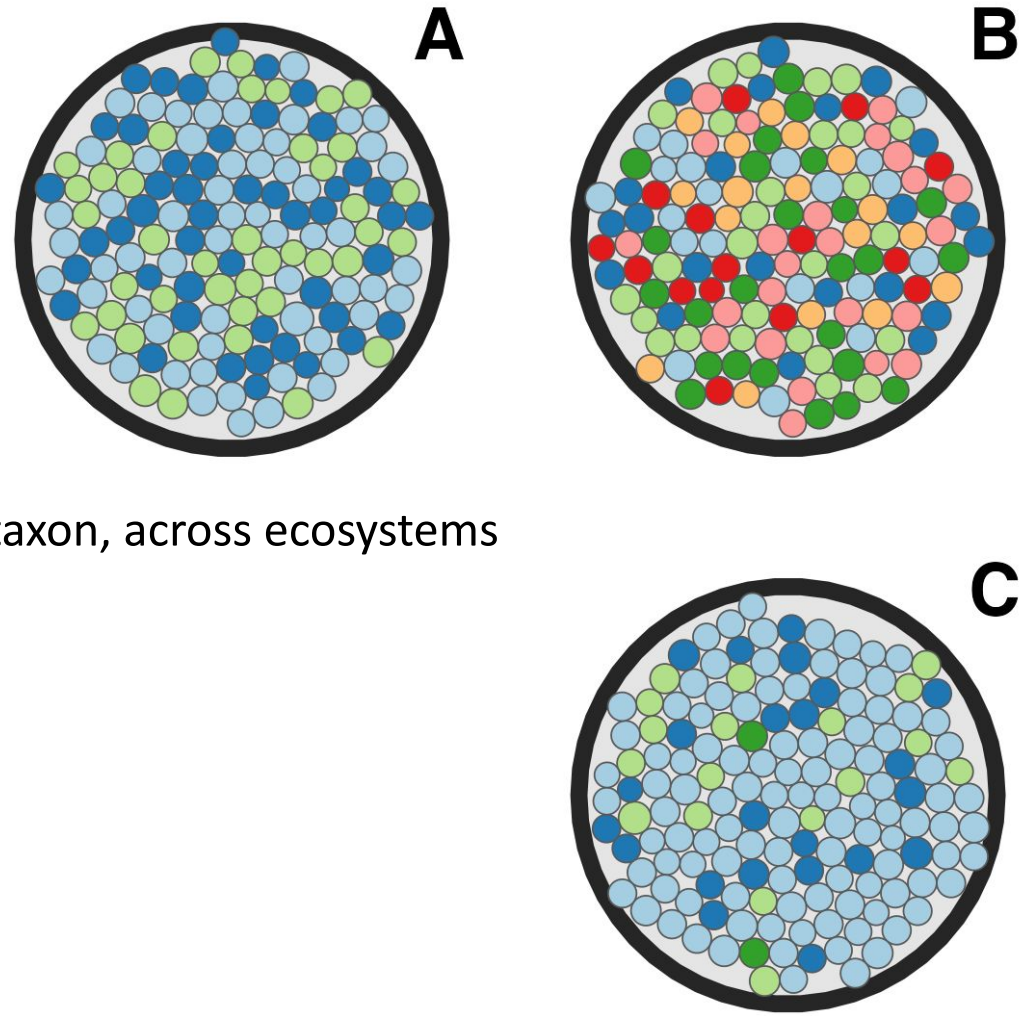
Dissimilarity between ecosystems



- **Dissimilarity** - how different are two ecosystems?
- **Distance matrix** - pairwise ecosystem dissimilarities
- **Very useful** - for plots (PCoA) and stats (PERMANOVA)

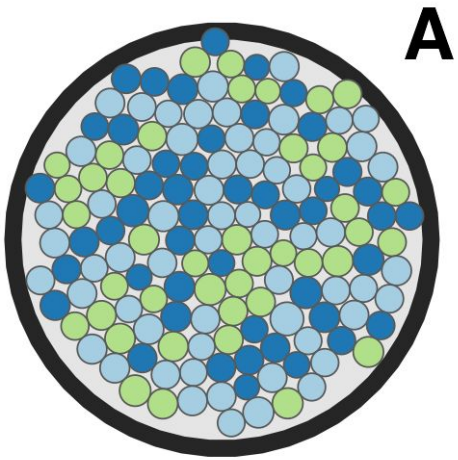
There are many different dissimilarity measures!

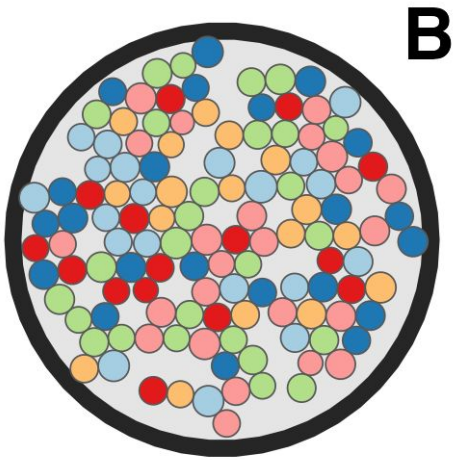
Differential Abundance of each taxon



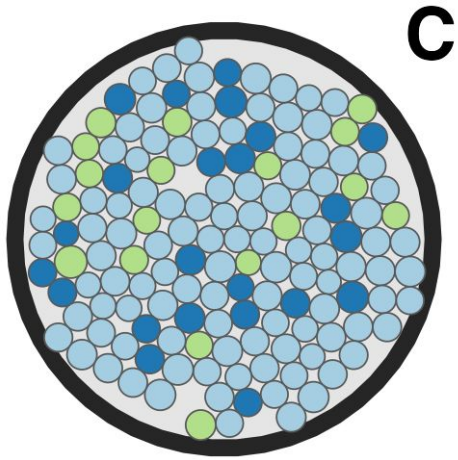
- Compare abundance of each taxon, across ecosystems

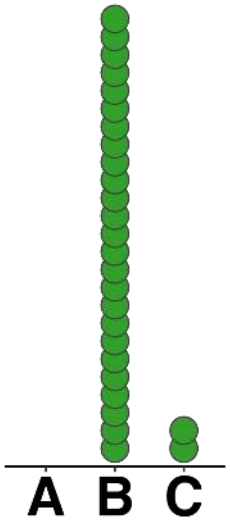
Differential Abundance of each taxon



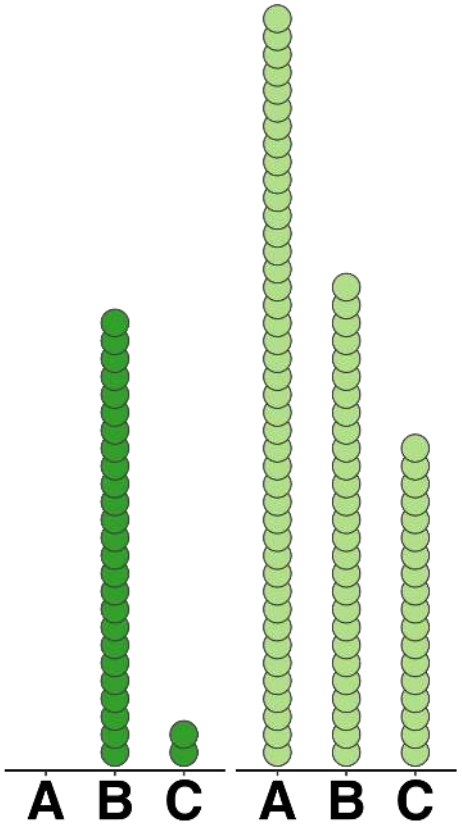


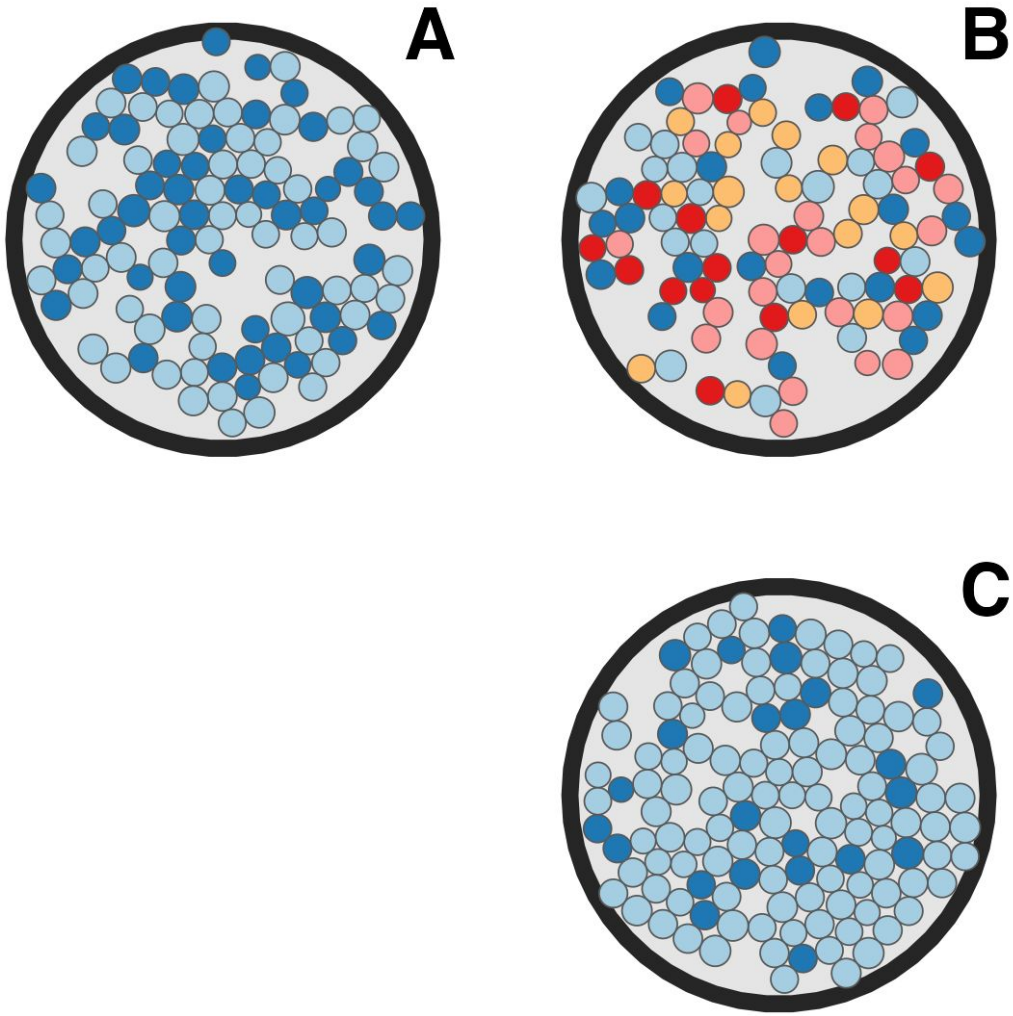
- Compare abundance of each taxon, across ecosystems
- One taxon at a time





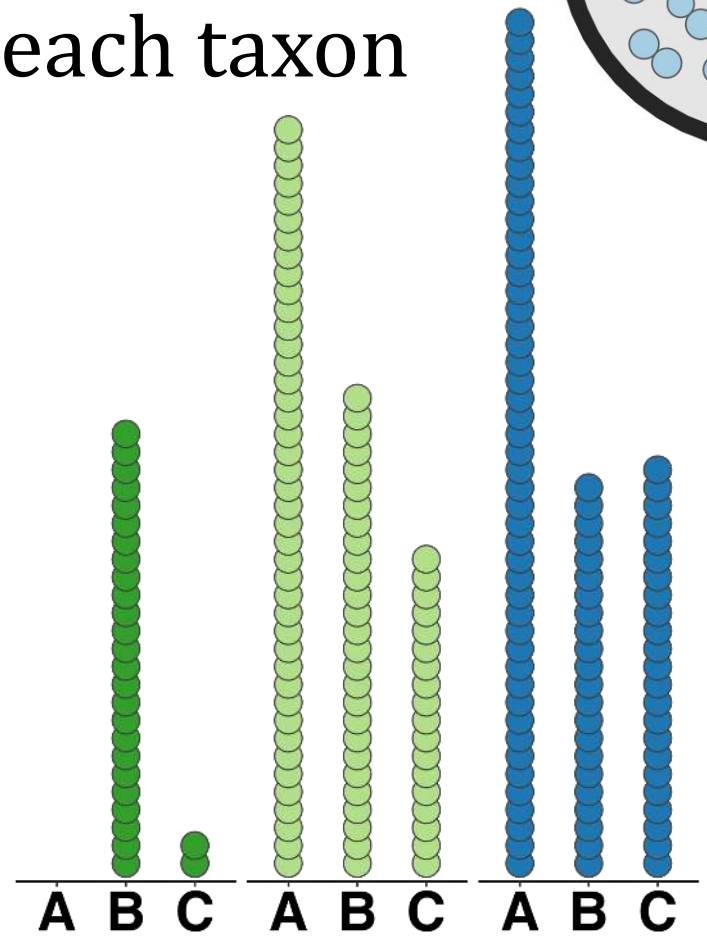
Differential Abundance of each taxon

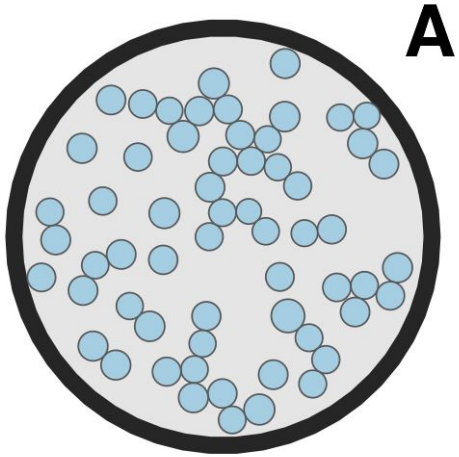


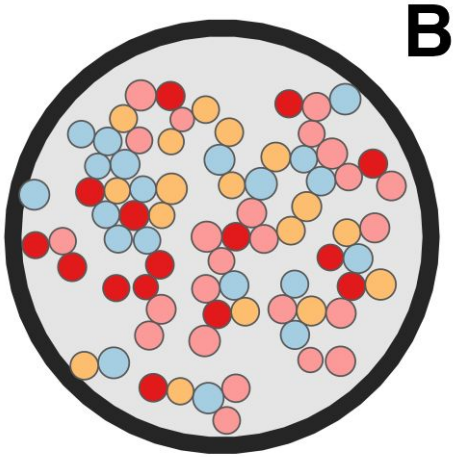


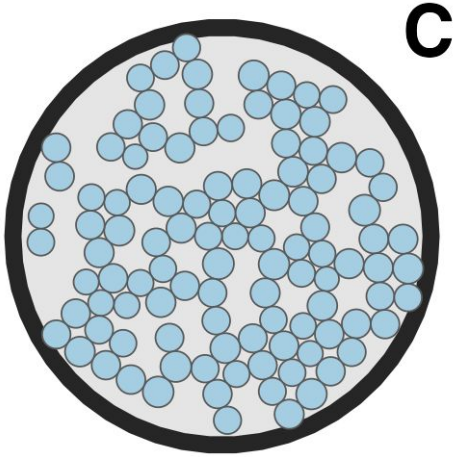
Differential Abundance

of each taxon







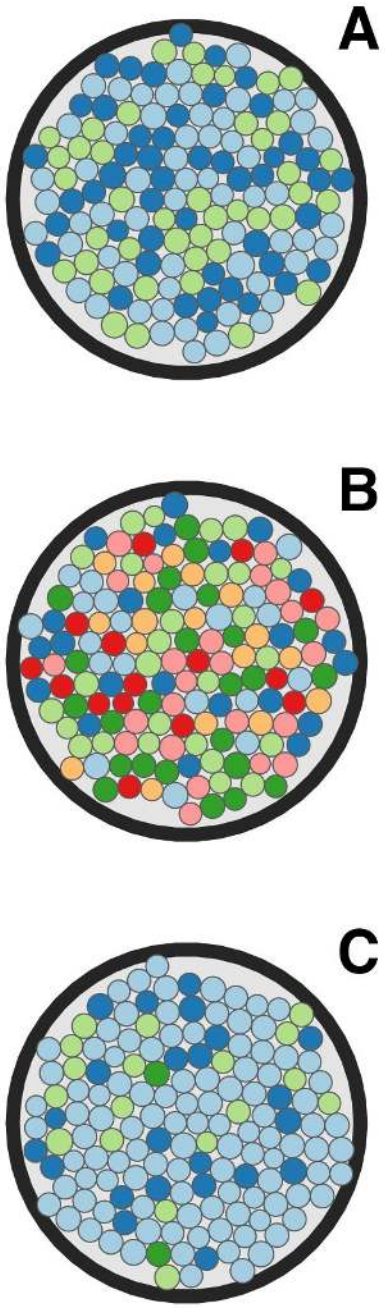


etc… etc… etc…

Differential Abundance of each taxon

- **Compare across groups of samples**

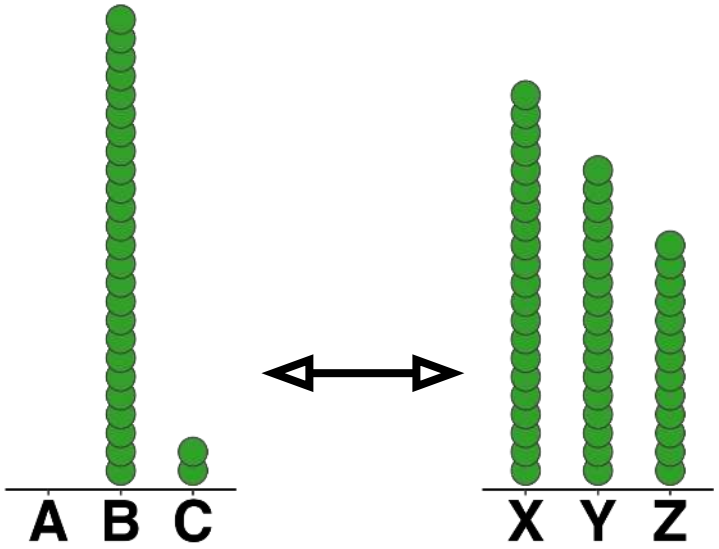
  e.g. - group ABC vs. group XYZ

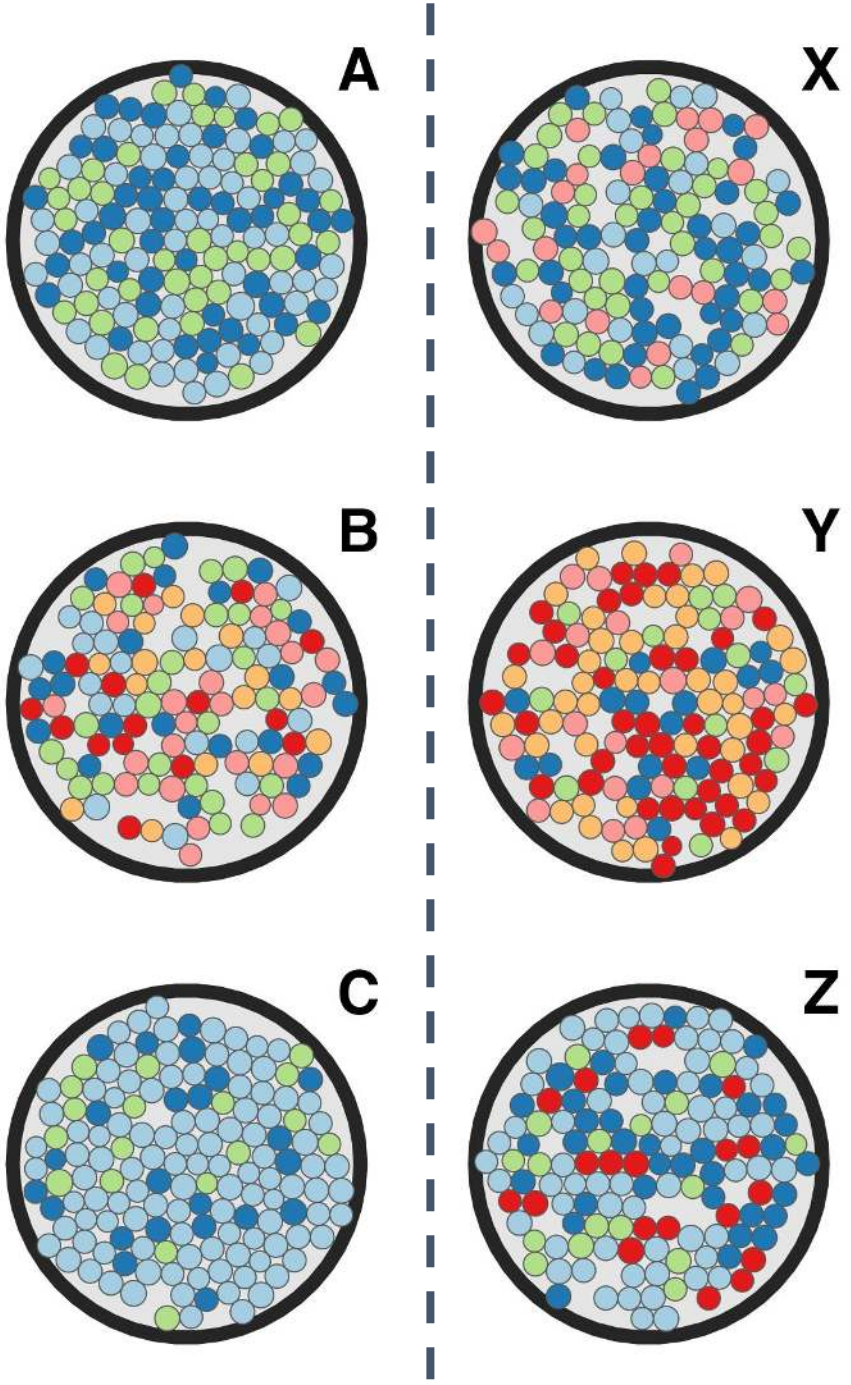


Differential Abundance of each taxon

- **Compare across groups of samples**

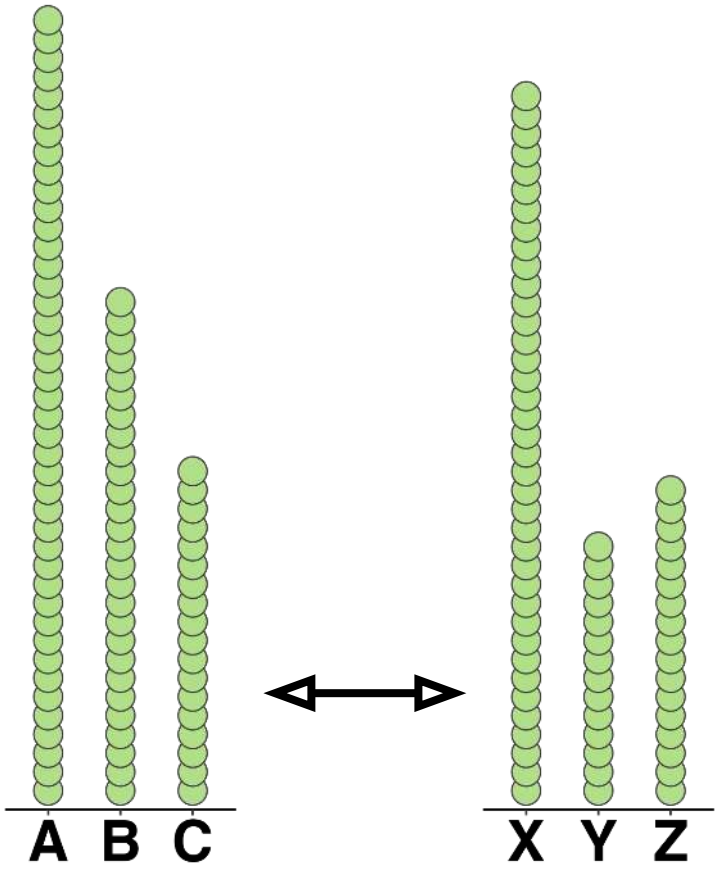
  e.g. - group ABC vs. group XYZ

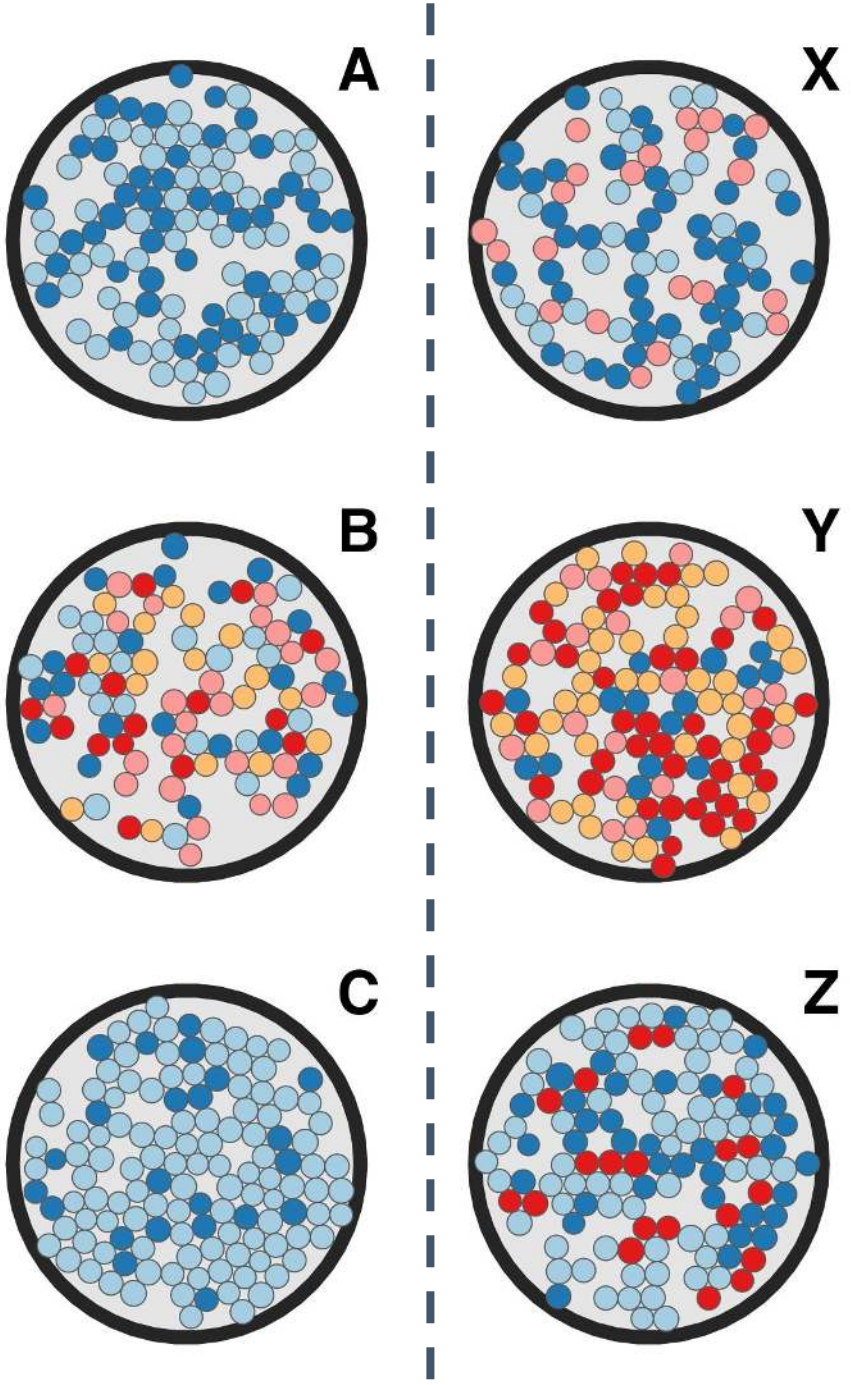




Differential Abundance

of each taxon

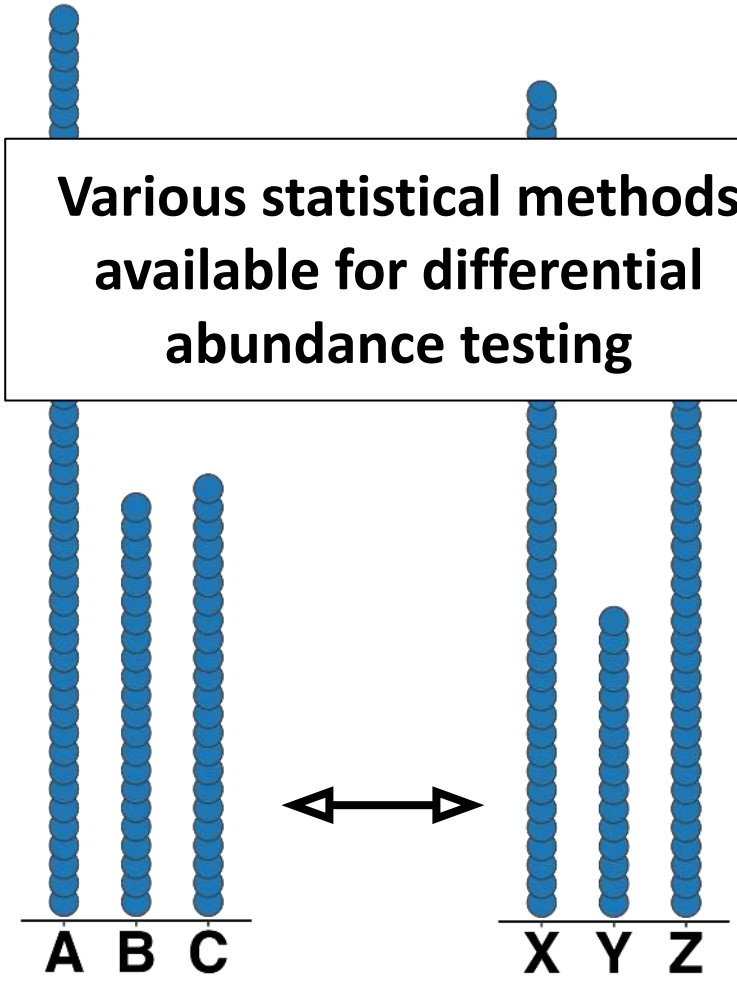


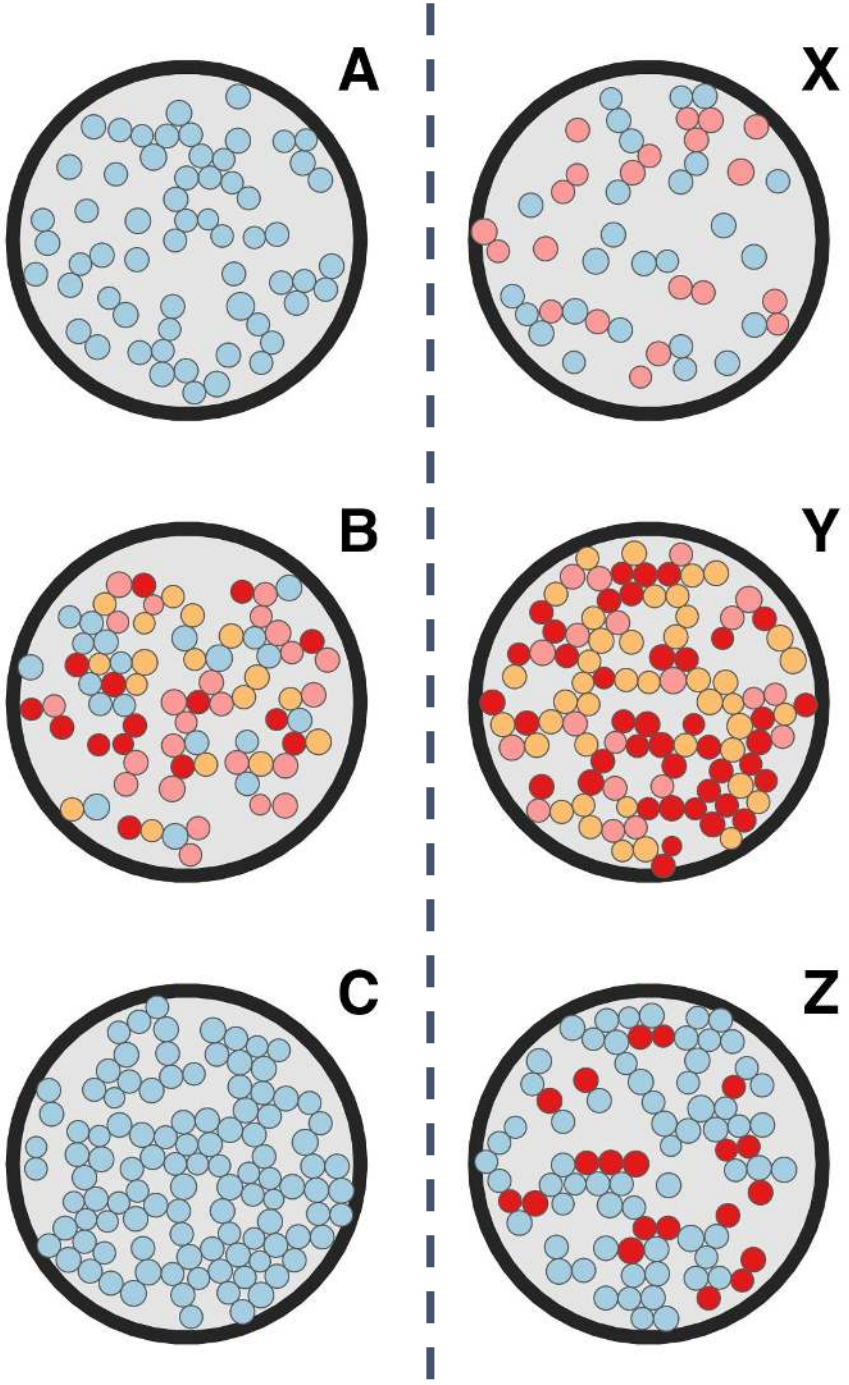


Differential Abundance

of each taxon

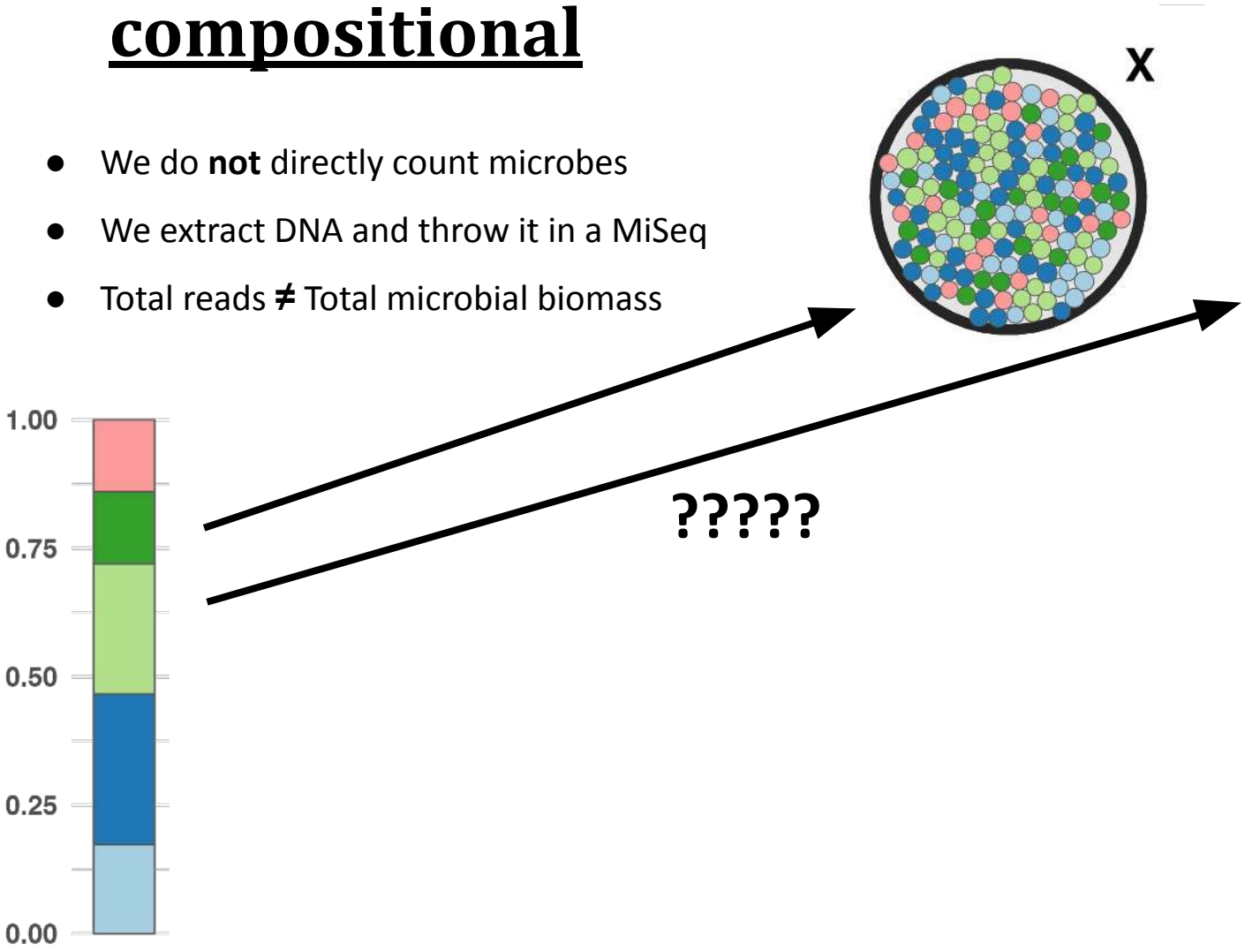
Various statistical methods available for differential abundance testing





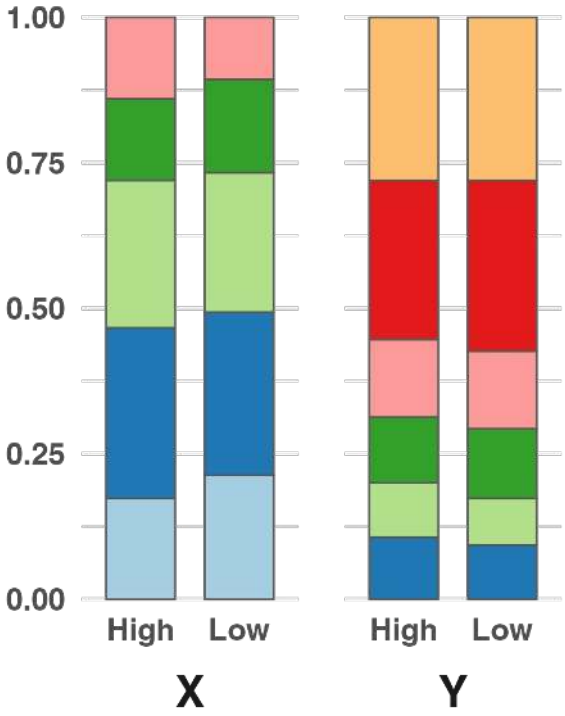
Microbiome data are <u>compositional</u>

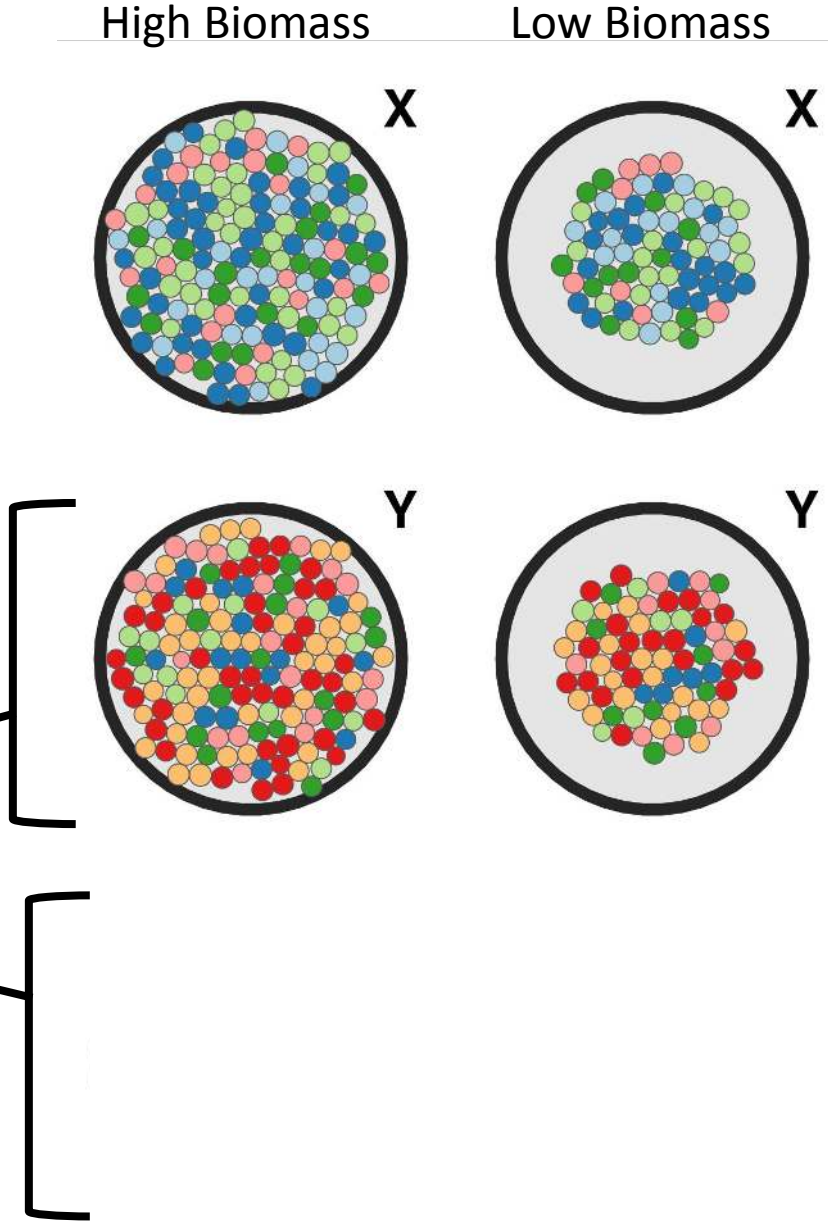
- We do **not** directly count microbes
- We extract DNA and throw it in a MiSeq
- Total reads ≠ Total microbial biomass



Microbiome data are <u>compositional</u>

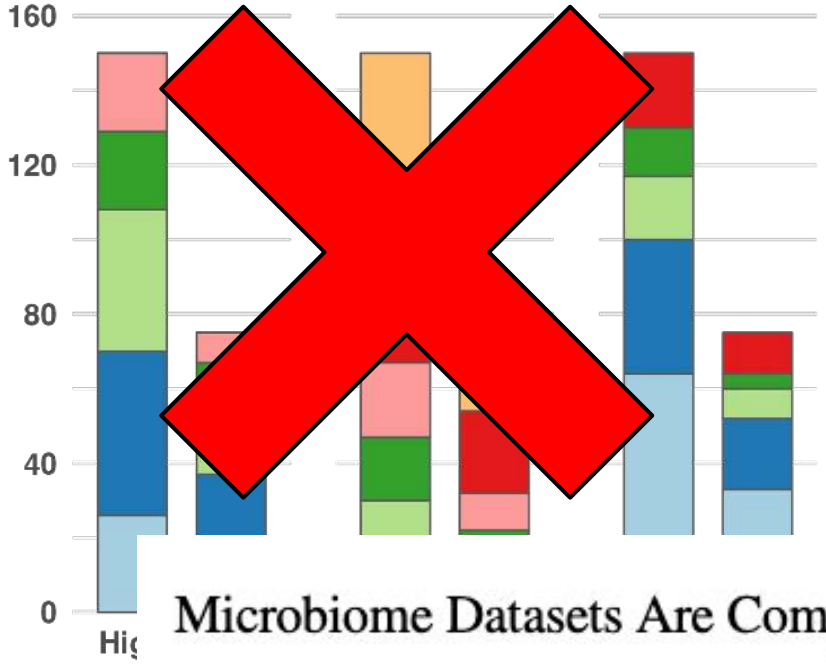
- We do **not** directly count microbes
- We extract DNA and throw it in a MiSeq
- Total reads ≠ Total microbial biomass

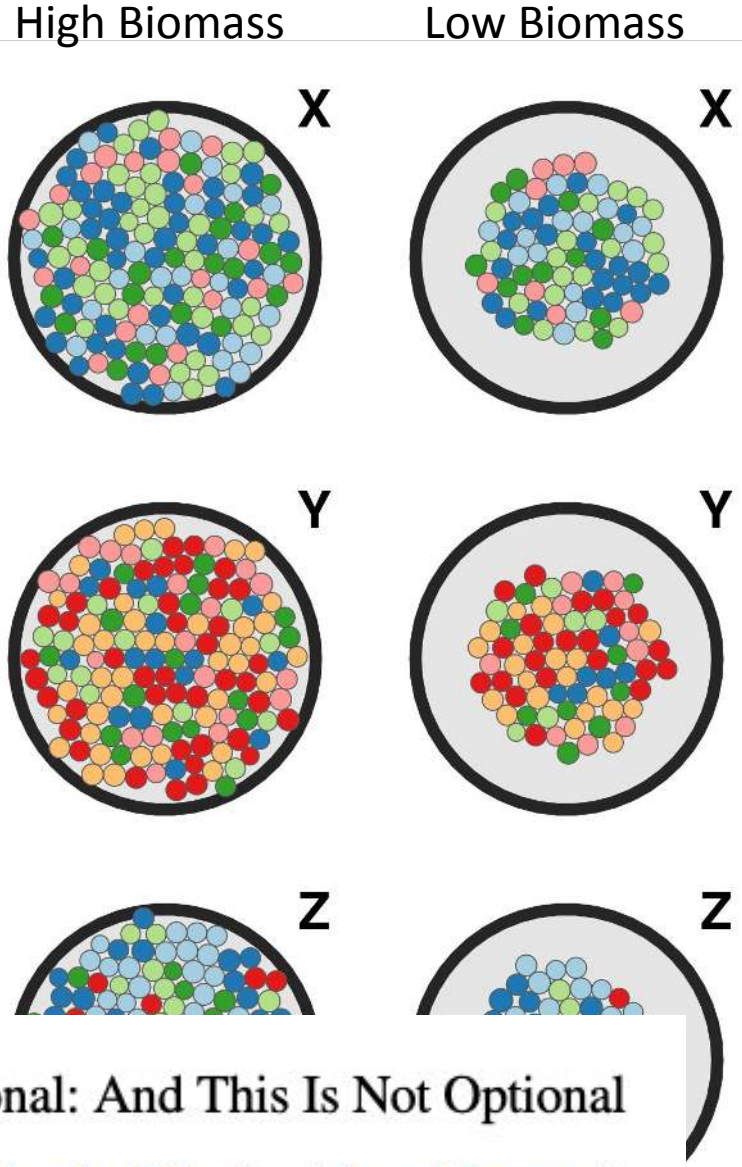




Microbiome data are compositional

- We do **not** directly count microbes
- We extract DNA and throw it in a MiSeq
- Total reads ≠ Total microbial biomass





Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor,[1,*] Jean M. Macklaim,[1] Vera Pawlowsky-Glahn,[2] and Juan J. Egozcue[3]

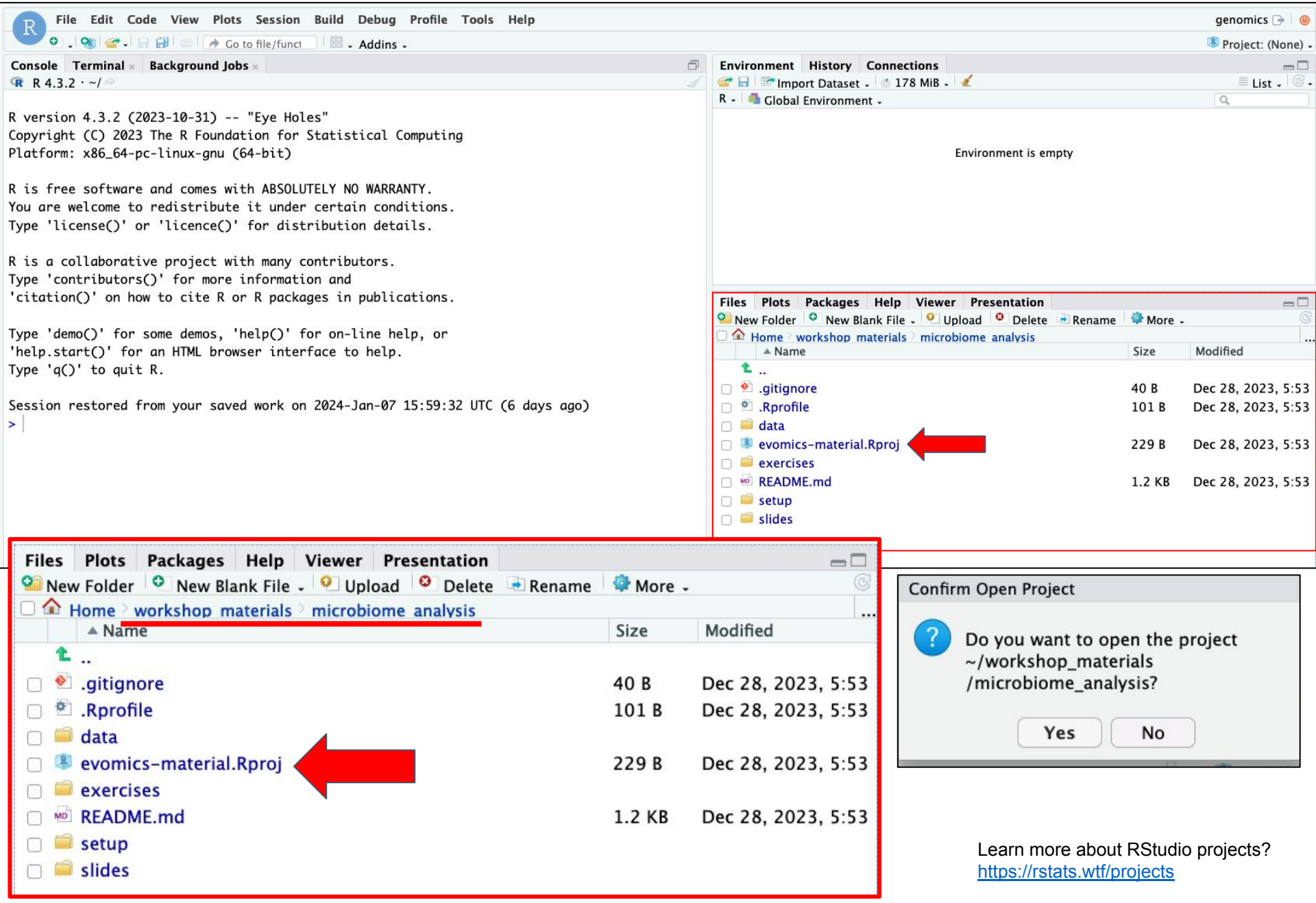
NOW: Barcharts and Diversity - getting started in R

david-barnett.github.io/evomics-material/exercises/exercises_1.html

Next lecture at 20:30

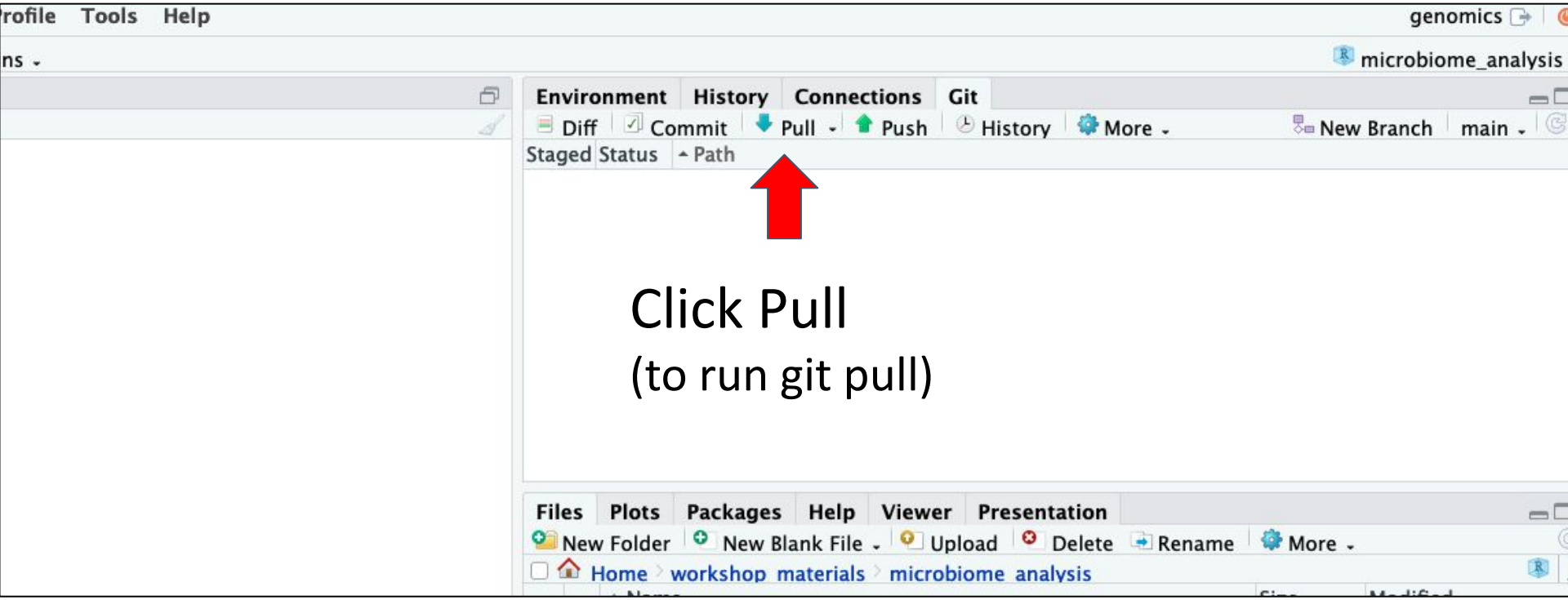
Remember to take a
break before then!

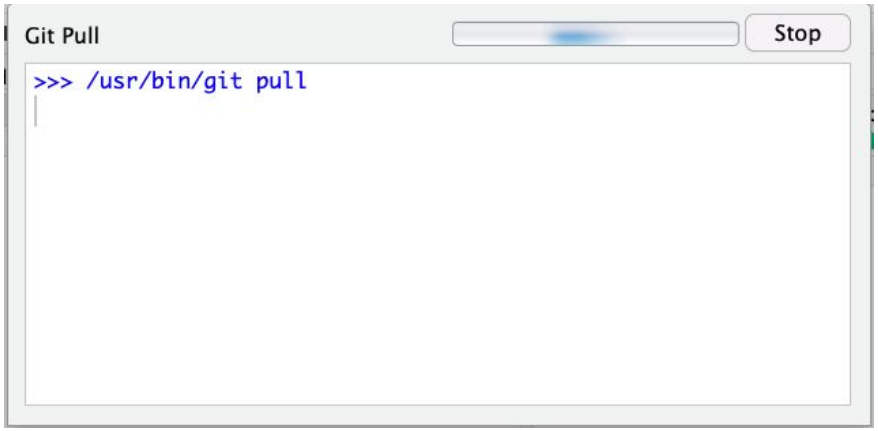
Activate RStudio Project in microbiome_analysis directory

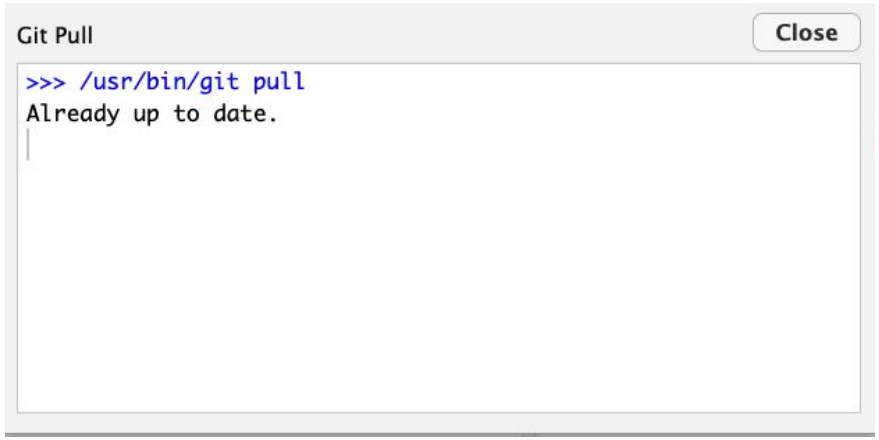


Learn more about RStudio projects?
https://rstats.wtf/projects

Ensure you have the latest version of the project git repo







More about using git with R & RStudio?
https://happygitwithr.com/

NOW: Barcharts and Diversity - getting started in R

david-barnett.github.io/evomics-material/exercises/exercises_1.html

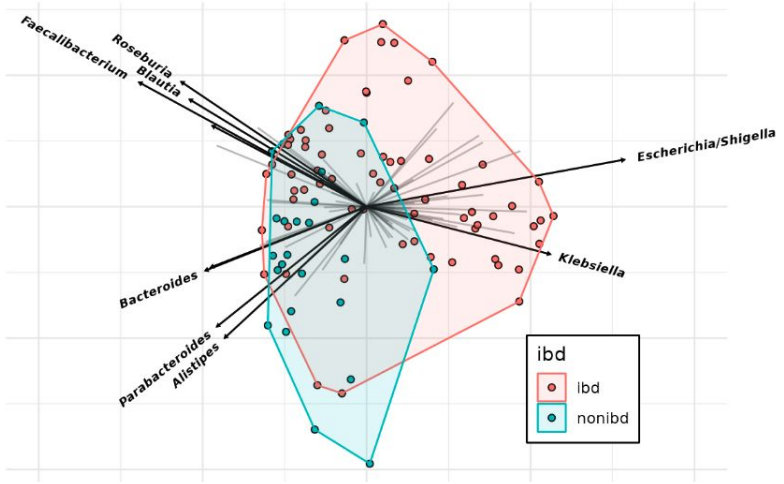
Next lecture at 20:30

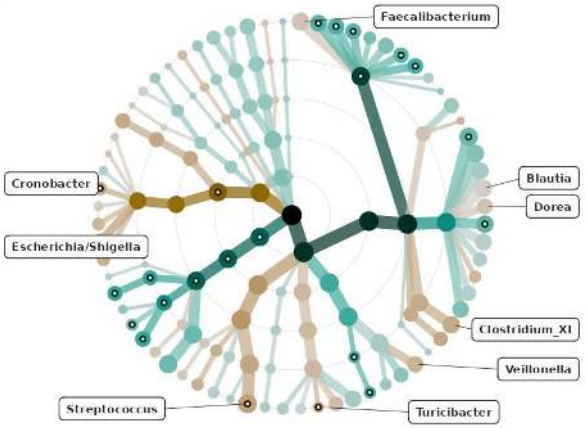
Remember to take a
break before then!

Dissimilarity, Ordination, and Differential Abundance

3. From Dissimilarity to Ordination

- Common dissimilarity measures
- PCoA, PERMANOVA, and PCA

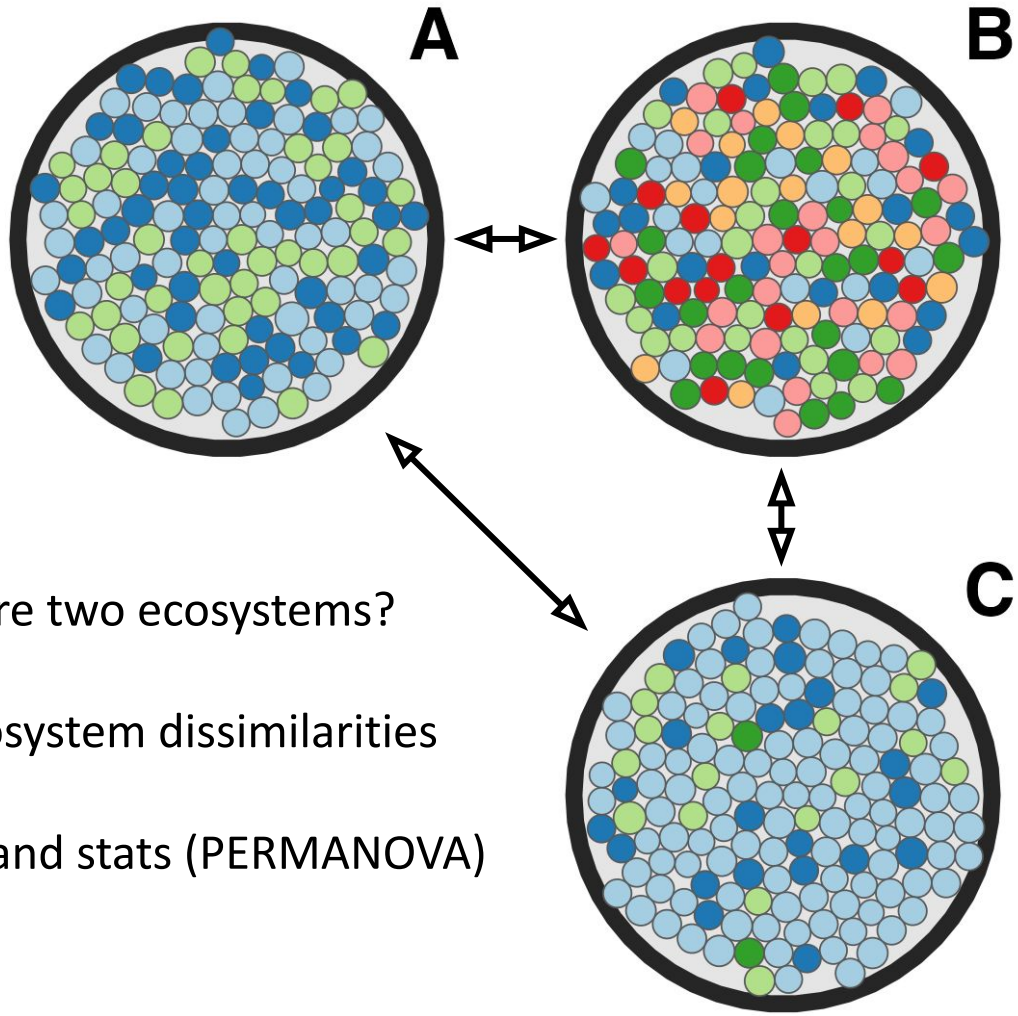




4. Differential Abundance testing

- A gentle intro to modelling individual taxa

Dissimilarity between ecosystems

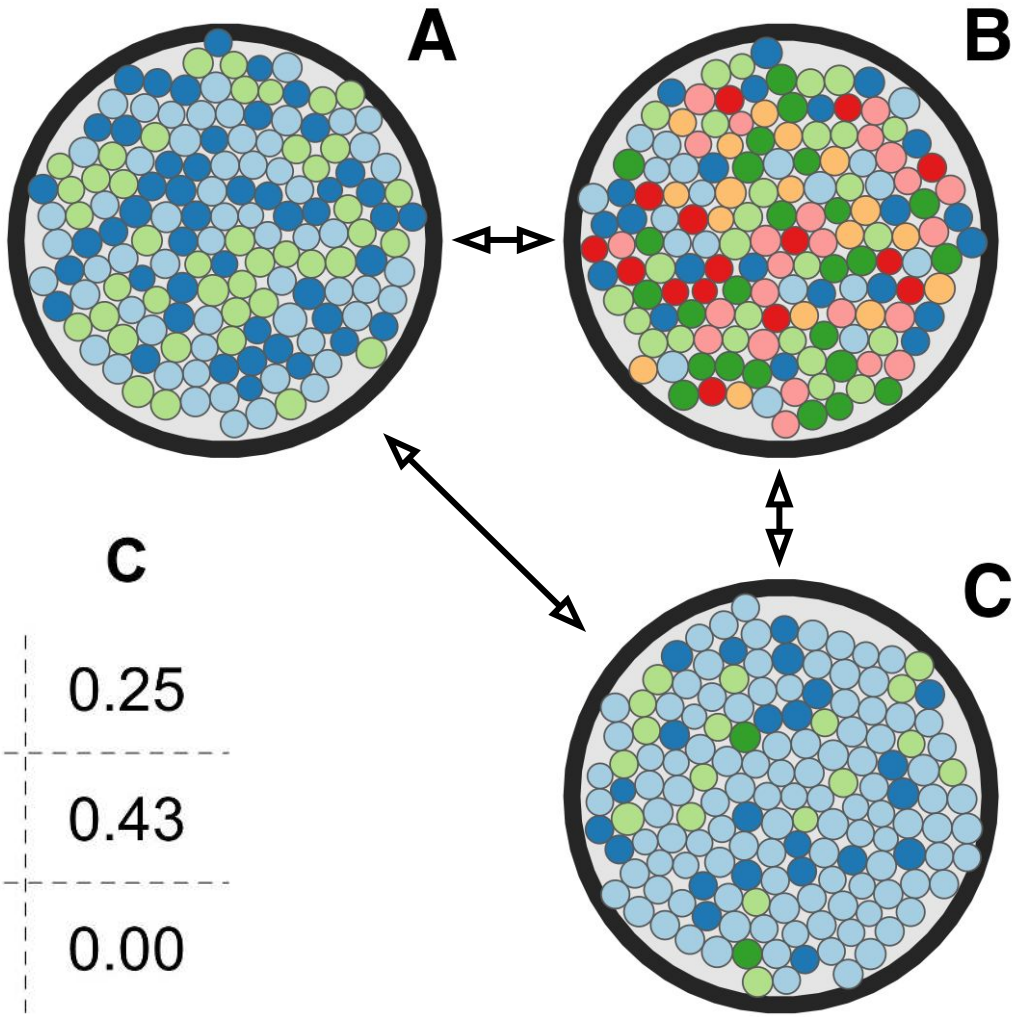


- **Dissimilarity** - how different are two ecosystems?
- **Distance matrix** - pairwise ecosystem dissimilarities
- **Very useful** - for plots (PCoA) and stats (PERMANOVA)

There are many different dissimilarity measures!

Dissimilarity Measures

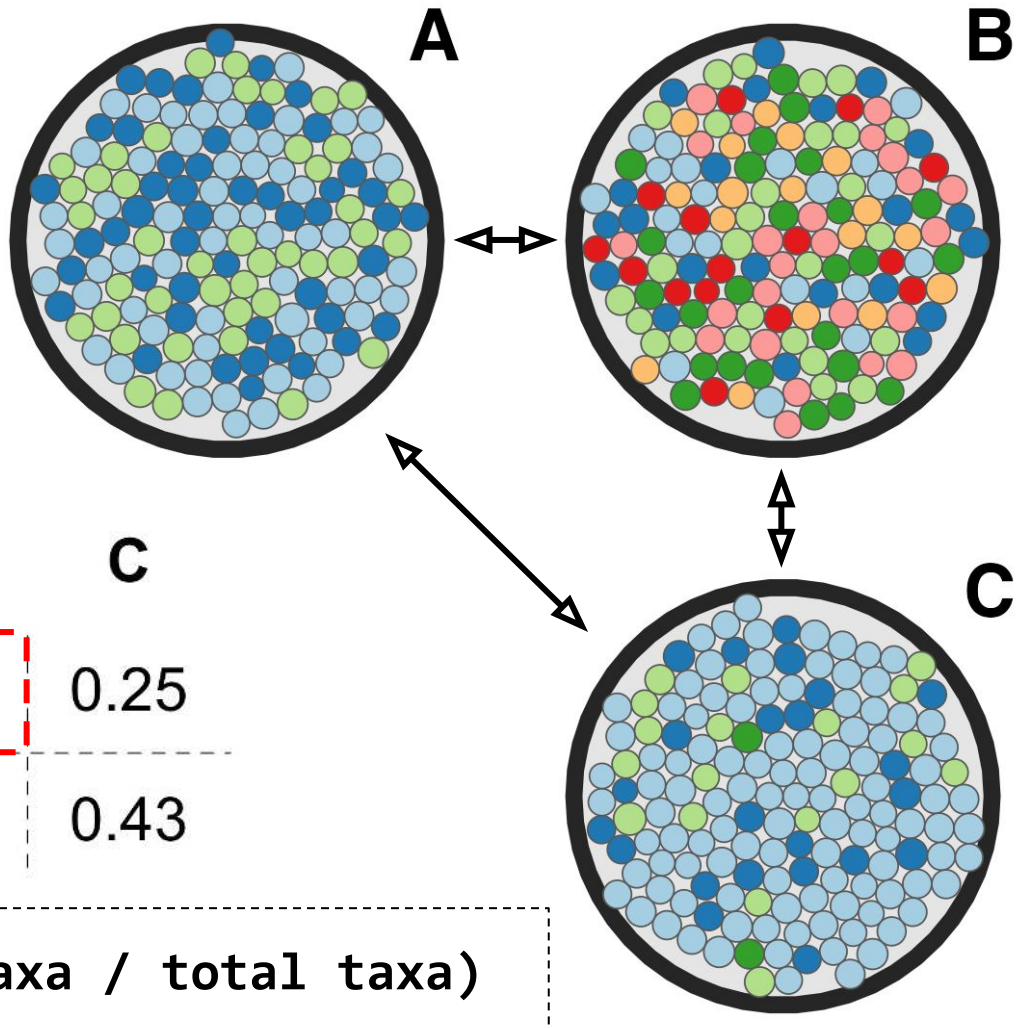
1. **Binary Jaccard Distance**
   - an “unweighted” measure



	A	B	C
A	0.00	0.57	0.25
B		0.00	0.43
C			0.00

Dissimilarity Measures

1. **Binary Jaccard Distance**
   - an “unweighted” measure



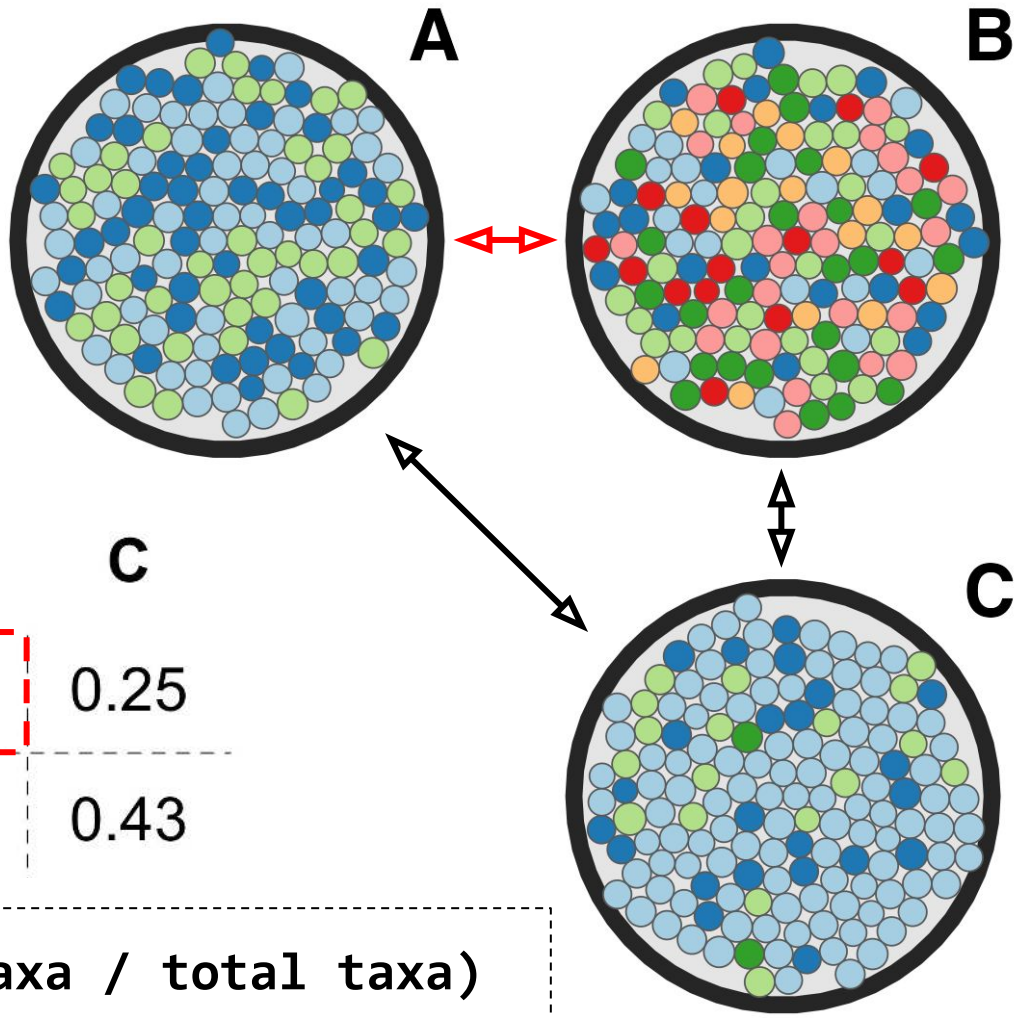
	B	C
A	0.57	0.25
B		0.43

d_{ij} = 1 - (shared taxa / total taxa)

➡ AB = ???

Dissimilarity Measures

1. **Binary Jaccard Distance**
   - an “unweighted” measure



	B	C
A	0.57	0.25
B		0.43

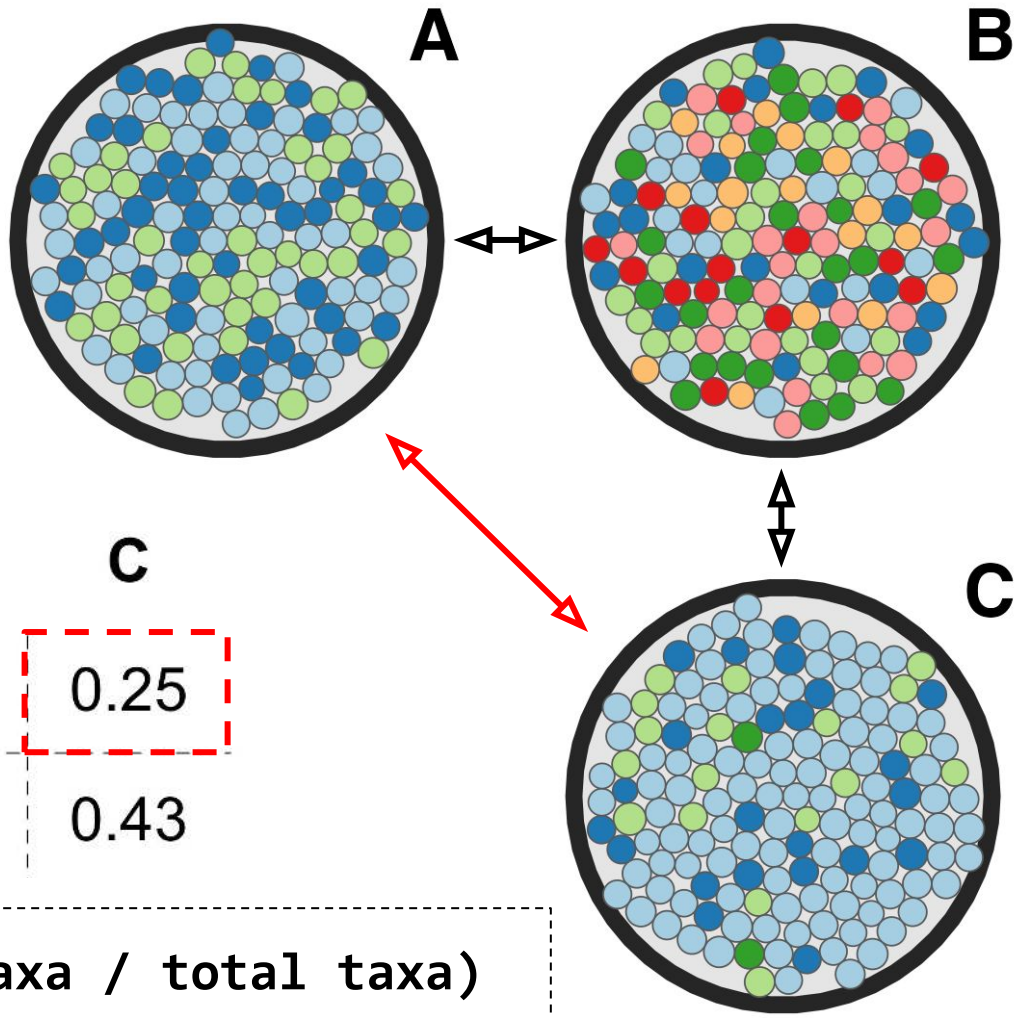
d_{ij} = 1 - (shared taxa / total taxa)

➡ AB = 1 - (● ● ● / ● ● ● ● ● ● ●)

➡ AB = 1 - (3 / 7) = 4 / 7 = **0.57**

Dissimilarity Measures

1. **Binary Jaccard Distance**
   - an “unweighted” measure



	B	C
A	0.57	0.25
B		0.43

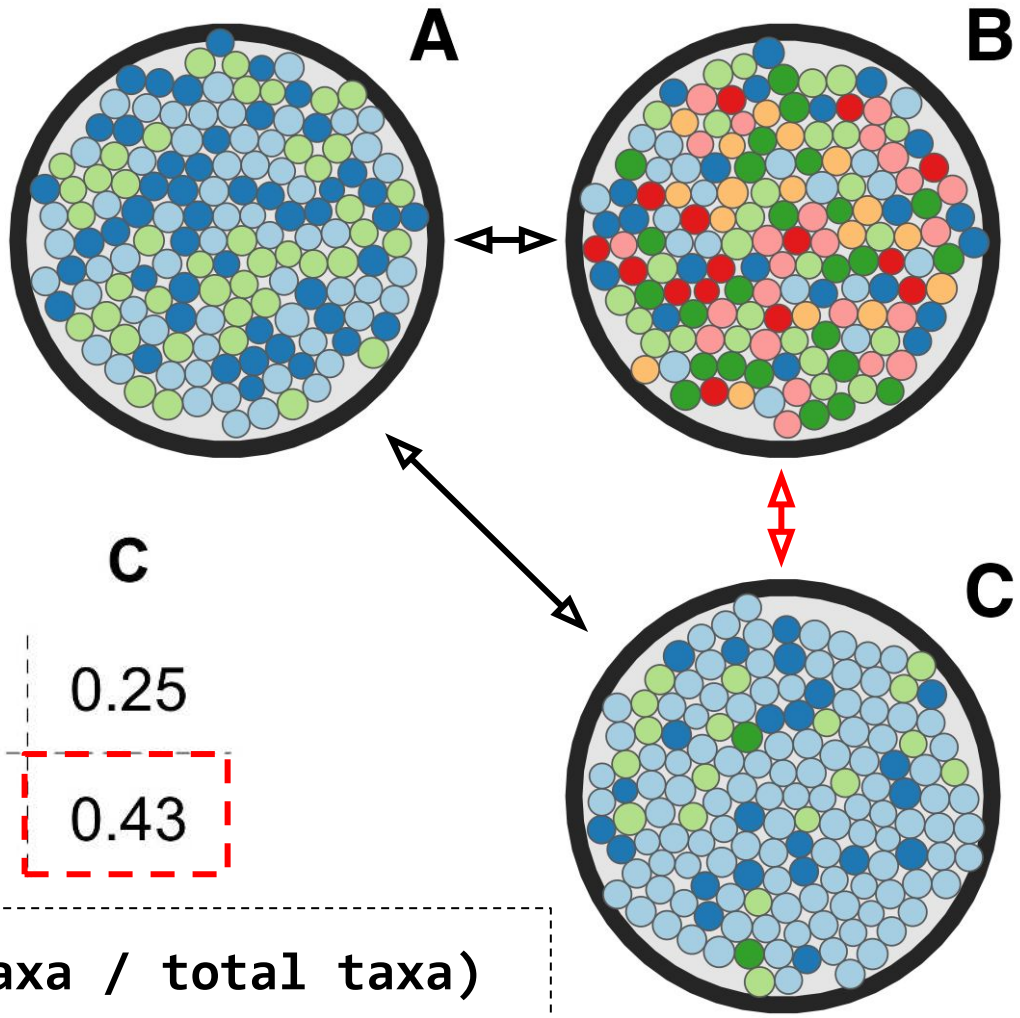
d_{ij} = 1 - (shared taxa / total taxa)

➡ AC = 1 - (●●● / ●●●●)

➡ AC = 1 - (3 / 4) = 1 / 4 = **0.25**

Dissimilarity Measures

1. **Binary Jaccard Distance**
   - an “unweighted” measure



	B	C
A	0.57	0.25
B		0.43

d_{ij} = 1 - (shared taxa / total taxa)

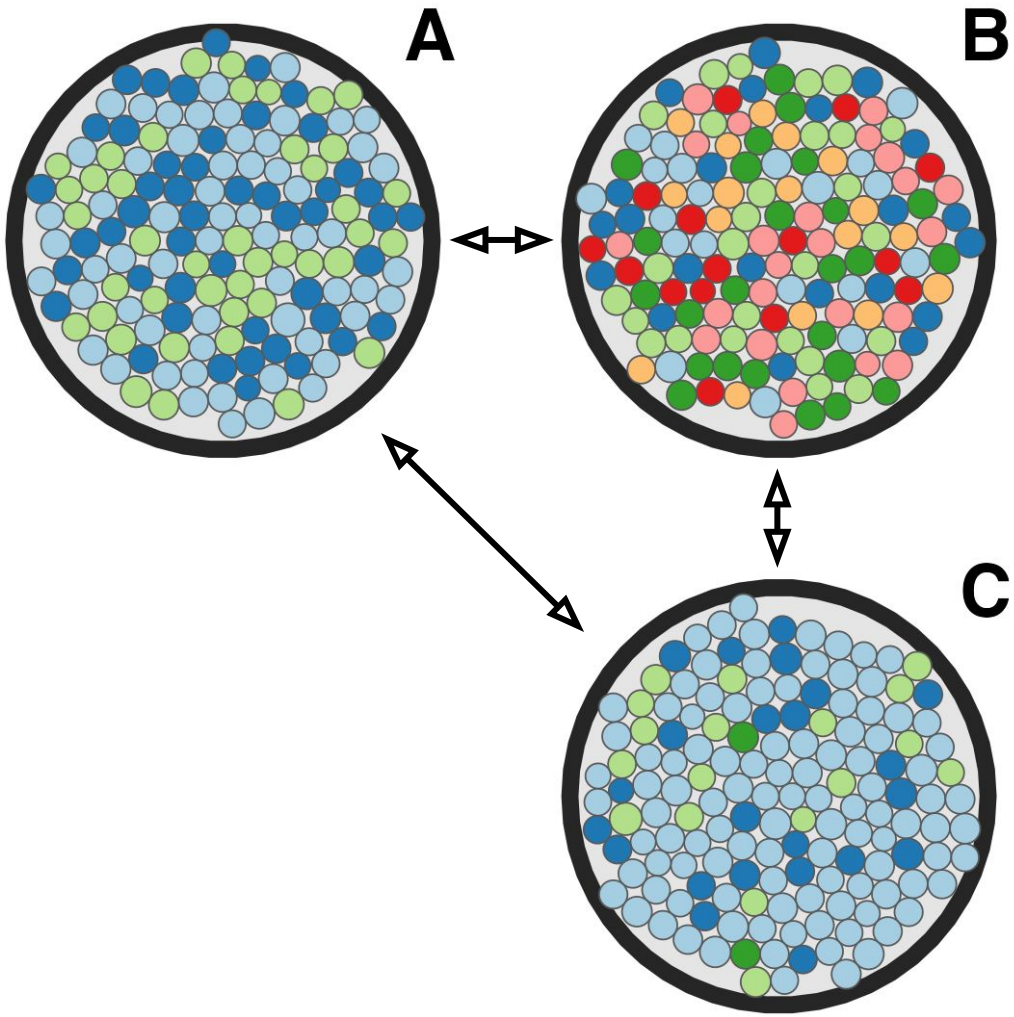
➡ BC = 1 - (● ● ● ● / ● ● ● ● ● ● ●)

➡ BC = 1 - (4 / 7) = 3 / 7 = **0.43**

Dissimilarity Measures

2. Bray-Curtis Dissimilarity

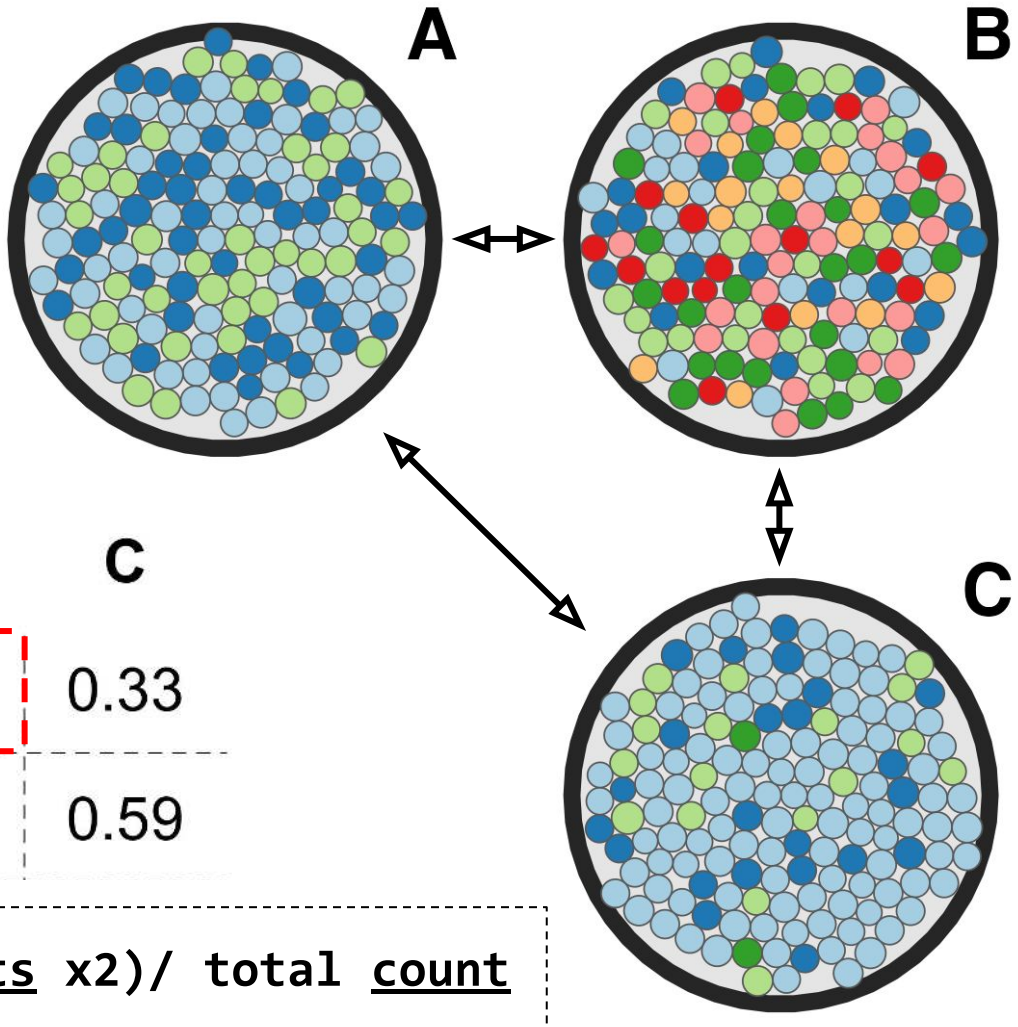
- “abundance-weighted”



Dissimilarity Measures

2. Bray-Curtis Dissimilarity

- "abundance-weighted"



	B	C
A	0.54	0.33
B		0.59

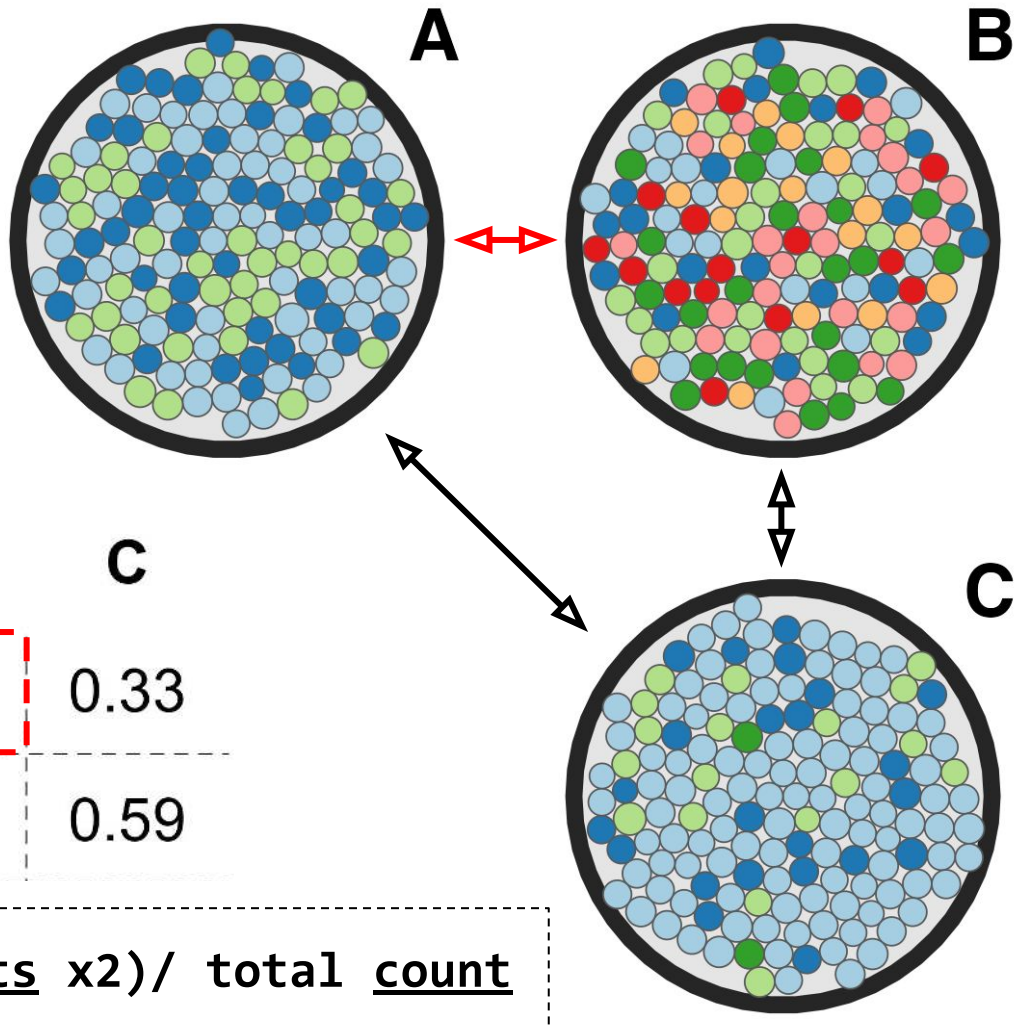
d_{ij}= 1 - (shared counts x2)/ total count

➡ AB= ???

Dissimilarity Measures

2. Bray-Curtis Dissimilarity

- "abundance-weighted"



	B	C
A	0.54	0.33
B		0.59

d_{ij}= 1 - (shared counts x2)/ total count

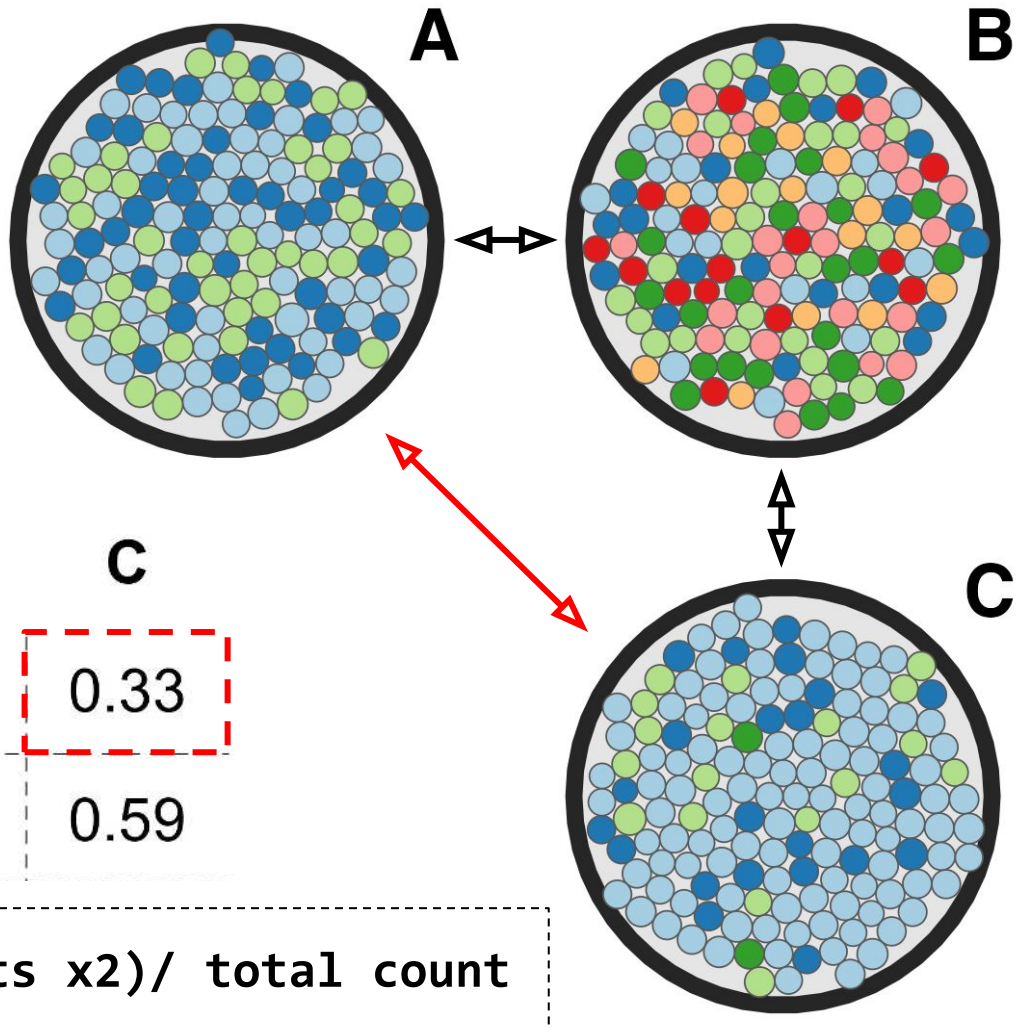
➡ AB= 1 - (20+22+27)x2 / (A + B)

➡ AB= 1 - 138/300 = 1 - 0.46 = **0.54**

Dissimilarity Measures

2. Bray-Curtis Dissimilarity

- "abundance-weighted"



	B	C
A	0.54	0.33
B		0.59

d_{ij}= 1 - (shared counts x2)/ total count

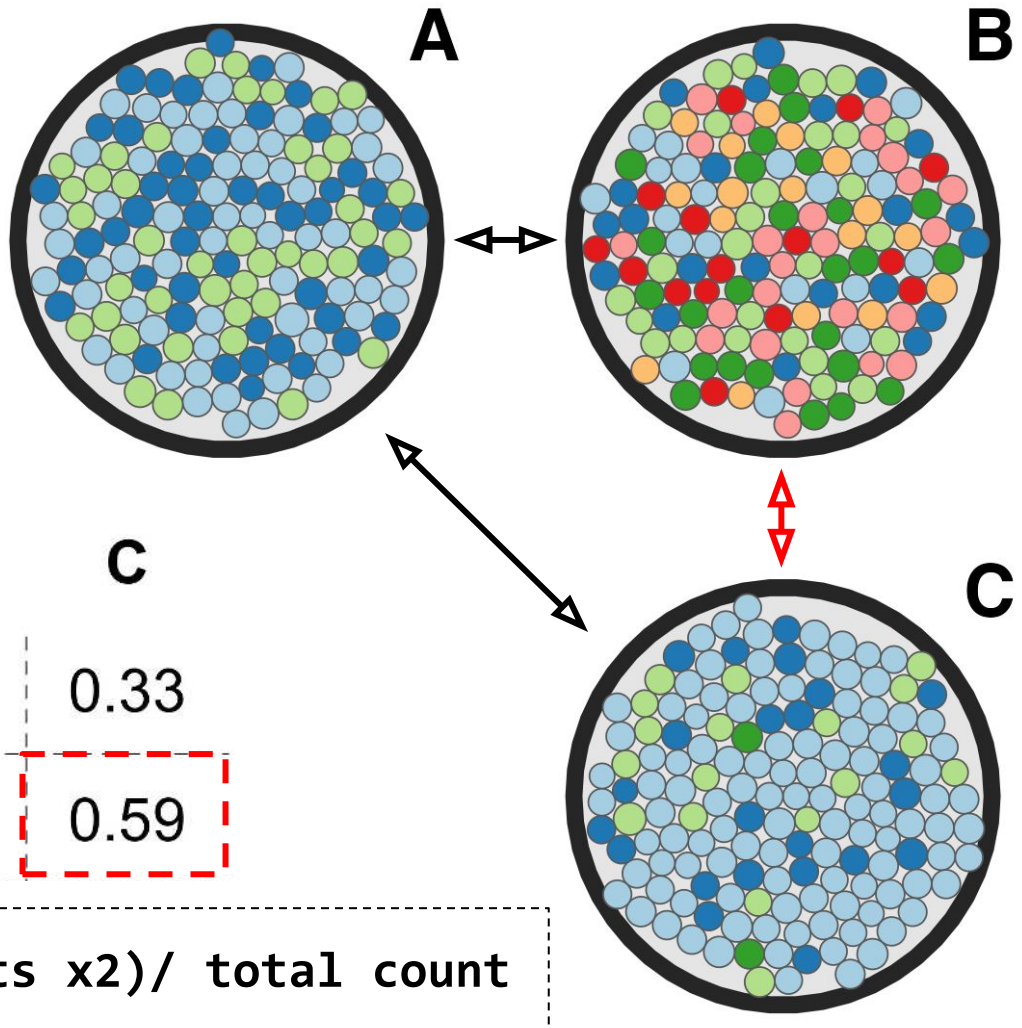
➡ AC= 1 - (60 +23 +18)x2 / (+)

➡ AC= 1 - 202/300 = 1 - 0.67 = **0.33**

Dissimilarity Measures

2. Bray-Curtis Dissimilarity

- "abundance-weighted"



	B	C
A	0.54	0.33
B		0.59

d_{ij}= 1 - (shared counts x2)/ total count

➡ BC= 1 -(20+22+18+2)x2/(+)

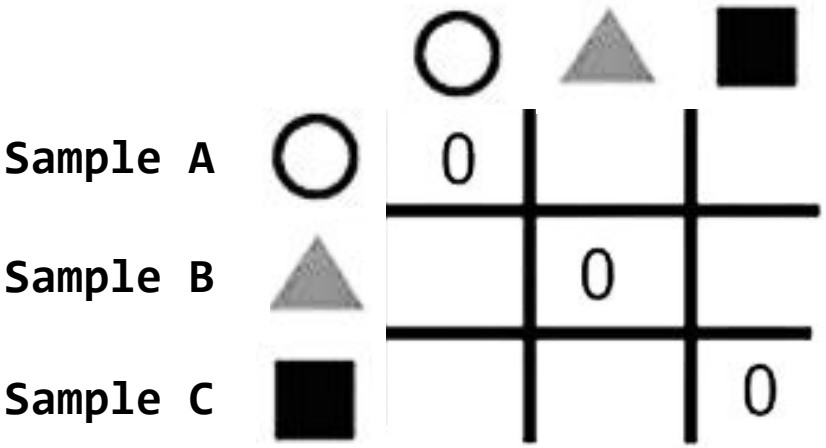
➡ BC= 1 - 124/300 = 1 - 0.41 = **0.59**

Dissimilarity Measures

3. UniFrac distance family

- "phylogenetic" distances

- Samples share the same taxa ➡ UniFrac is very low (or zero)
- Samples contain distinct but **related** taxa ➡ UniFrac is low
- Samples contain **unrelated** taxa ➡ UniFrac is higher



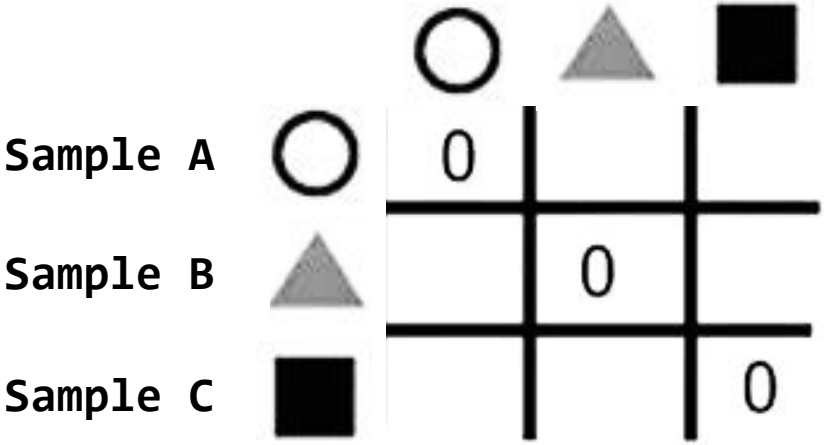
Example adapted from Lozupone and Knight 2005

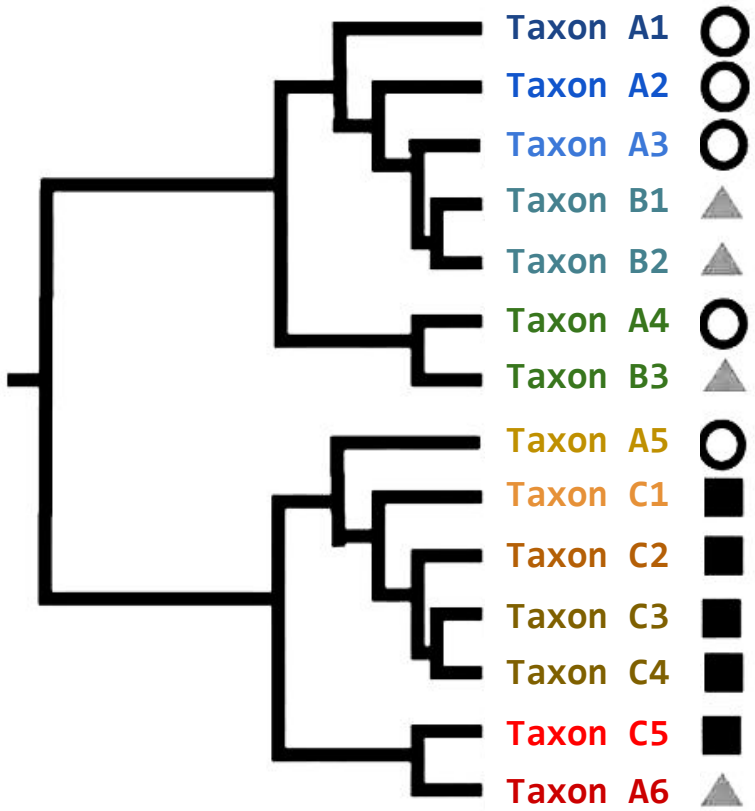
Dissimilarity Measures

3. UniFrac distance family

- "phylogenetic" distances

- share the same taxa ➡ UniFrac is very low (or zero)
- … distinct but **related** taxa ➡ UniFrac is low
- … **unrelated** taxa ➡ UniFrac is higher





Tree tips are **taxa**!
Shape shows their **source**

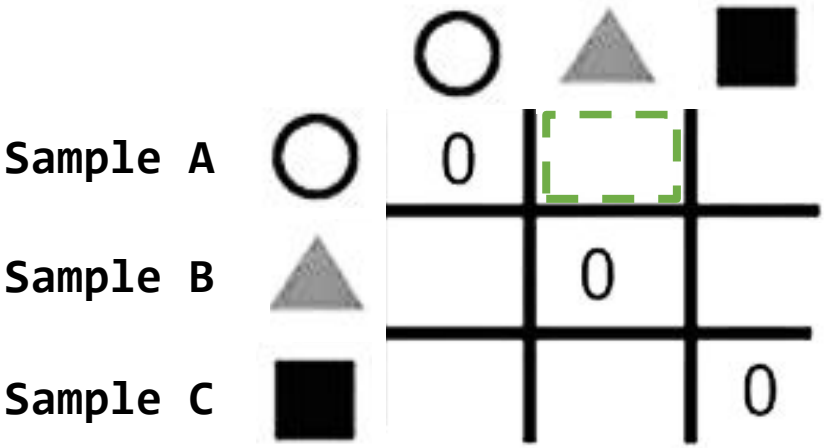
UniFrac = "**Uni**que **Frac**tion"
of branch length leading to
taxa from only one sample

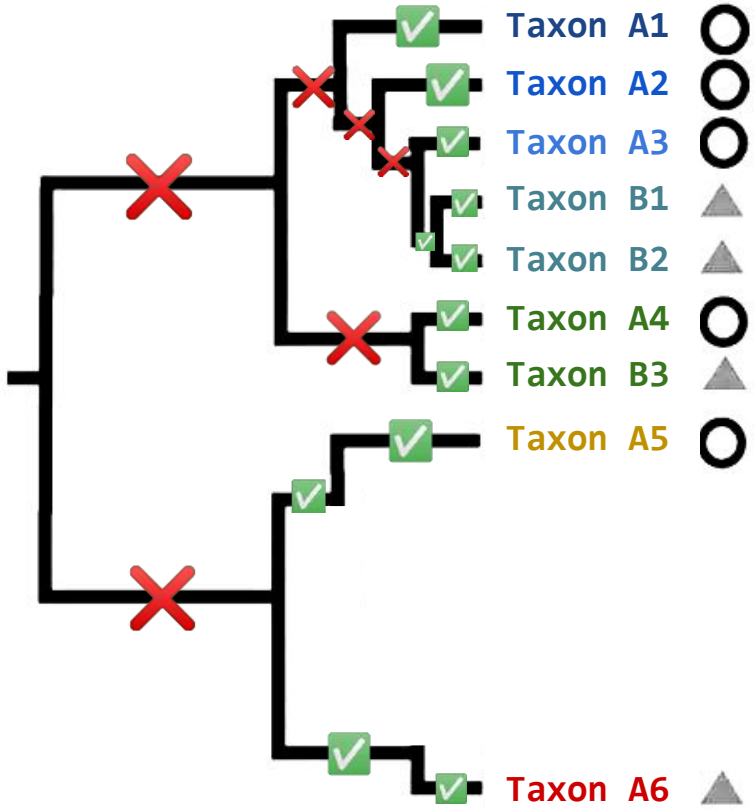
Dissimilarity Measures

3. UniFrac distance family

- "phylogenetic" distances

- share the same taxa ➡ UniFrac is very low (or zero)
- … distinct but **related** taxa ➡ UniFrac is low
- … **unrelated** taxa ➡ UniFrac is higher





Tree tips are **taxa**!
Shape shows their **source**

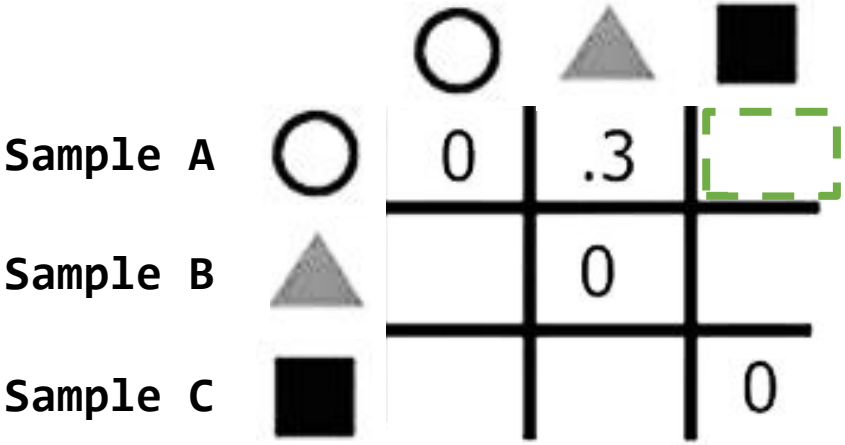
UniFrac = "**Uni**que **Frac**tion"
of branch length leading to
taxa from only one sample

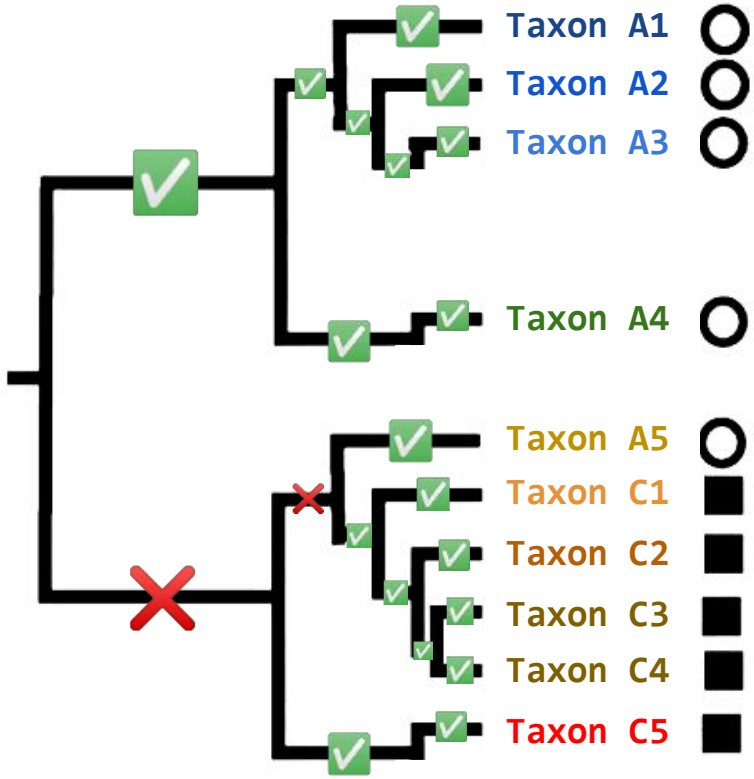
Dissimilarity Measures

3. UniFrac distance family

- "phylogenetic" distances

- share the same taxa ➡ UniFrac is very low (or zero)
- … distinct but **related** taxa ➡ UniFrac is low
- … **unrelated** taxa ➡ UniFrac is higher





Tree tips are **taxa**!
Shape shows their **source**

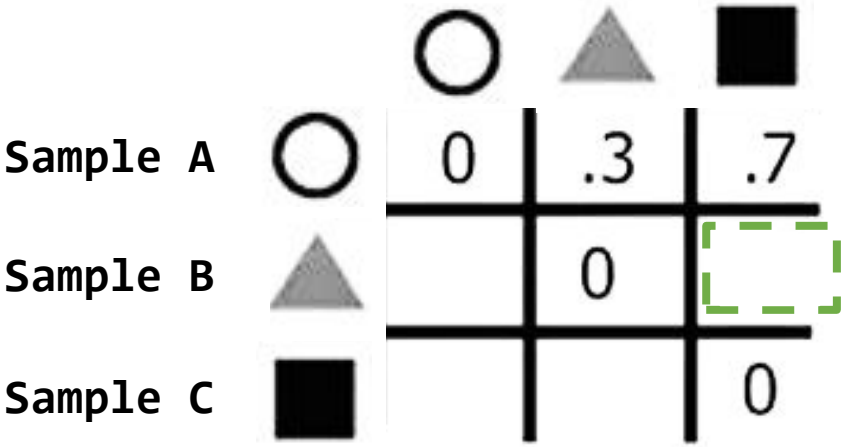
UniFrac = "**Uni**que **Frac**tion"
of branch length leading to
taxa from only one sample

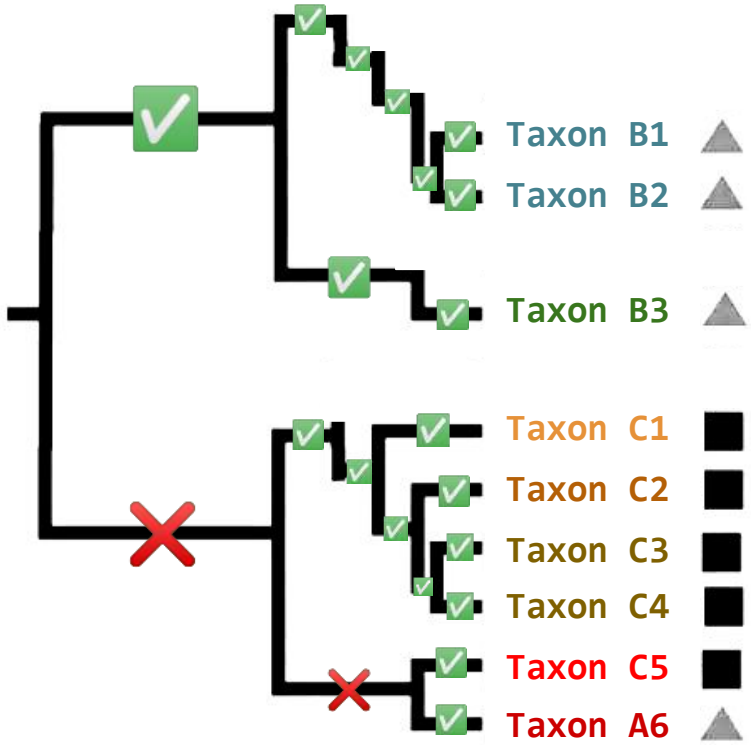
Dissimilarity Measures

3. UniFrac distance family

- "phylogenetic" distances

- share the same taxa ➡ UniFrac is very low (or zero)
- … distinct but **related** taxa ➡ UniFrac is low
- … **unrelated** taxa ➡ UniFrac is higher





Tree tips are **taxa**!
Shape shows their **source**

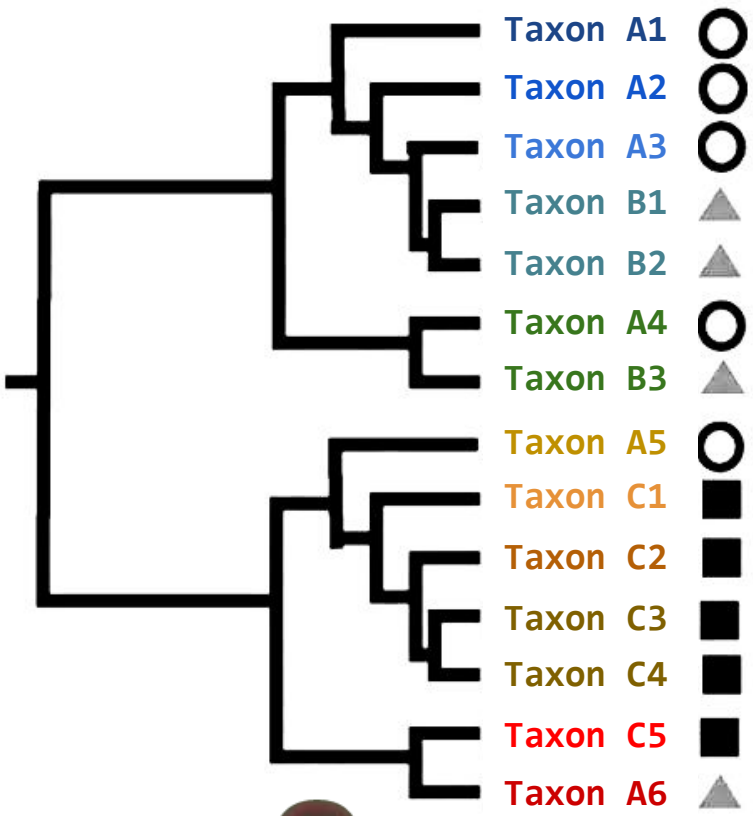
UniFrac = "**Uni**que **Frac**tion"
of branch length leading to
taxa from only one sample

Dissimilarity Measures

3. UniFrac distance family

- "phylogenetic" distances

- share the same taxa ➡ UniFrac is very low (or zero)
- … distinct but **related** taxa ➡ UniFrac is low
- … **unrelated** taxa ➡ UniFrac is higher



The UniFrac family:

1. UniFrac (unweighted)
2. Abundance-weighted UniFrac
3. Generalised UniFrac (balanced)

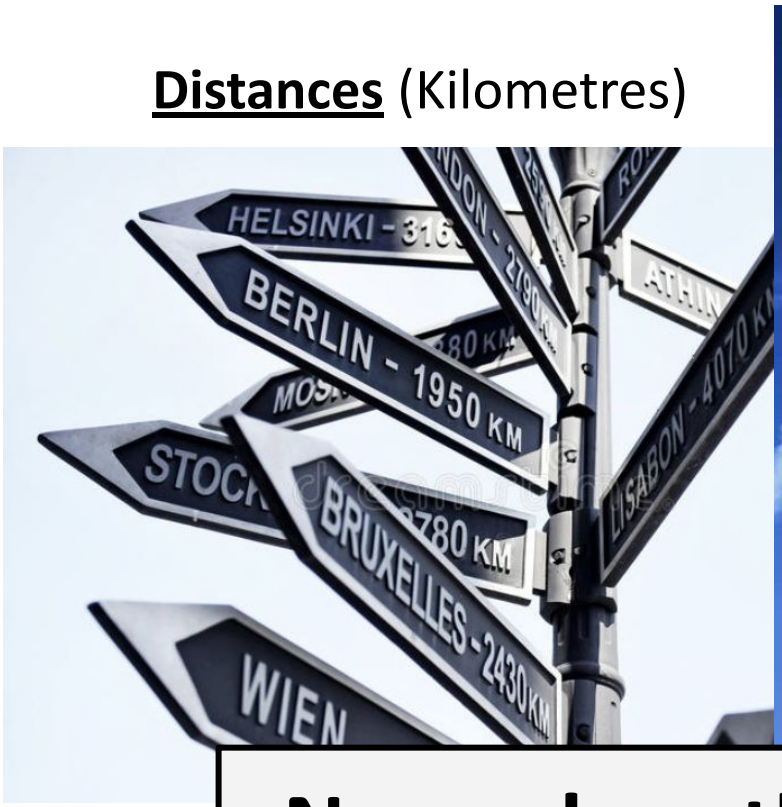
Tree tips are **taxa**!
Shape shifts their **source**

UniFrac = "**Uni**que **Frac**tion"
of branch length leading to
taxa from only one sample

Principal Coordinates Analysis

Dissimilarities ➡ Ordination

Distances (Kilometres)



Now, place the cities on map …?

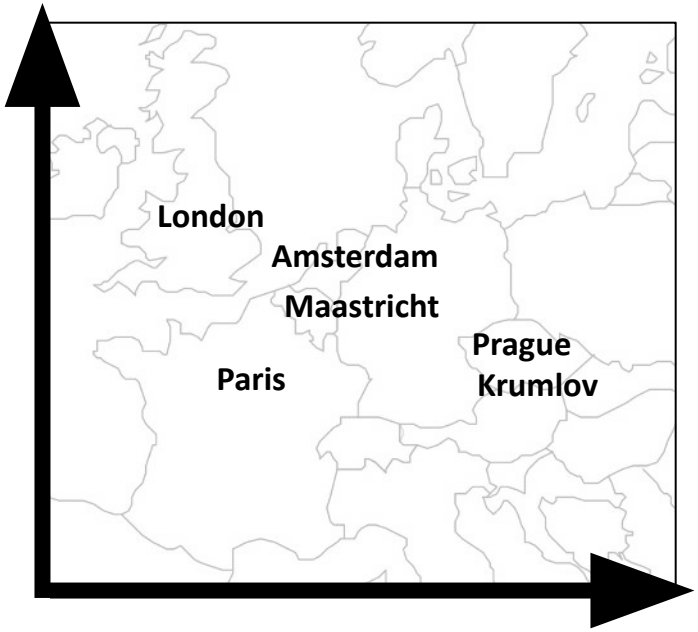
Principal Coordinates Analysis

Dissimilarities ➡ Ordination

Distances (Kilometres)

	Lon.	Par.	Ams.	Maa.	Pra.	Kru.
London	0	…	…	…	…	…
Paris	…	0	…	…	…	…
Amsterdam	…	…	0	…	…	…
Maastricht	…	…	…	0	…	…
Prague	…	…	…	…	0	…
Krumlov	…	…	…	…	…	0

Map (Coordinates)



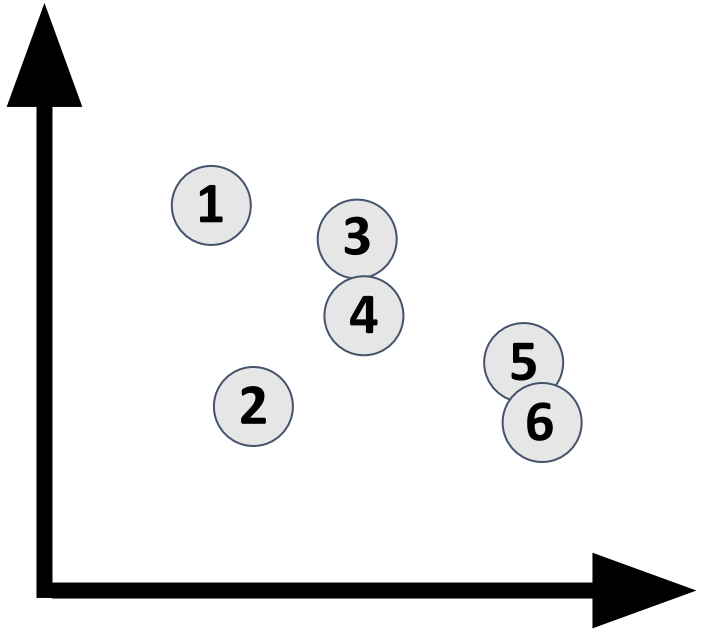
Principal Coordinates Analysis

Dissimilarities ➡ Ordination

Dissimilarities (Bray-Curtis)

	S1	S2	S3	S4	S5	S6
Sample 1	0	…	…	…	…	…
Sample 2	…	0	…	…	…	…
Sample 3	…	…	0	…	…	…
Sample 4	…	…	…	0	…	…
Sample 5	…	…	…	…	0	…
Sample 6	…	…	…	…	…	0

Plot (Principal Coordinates)

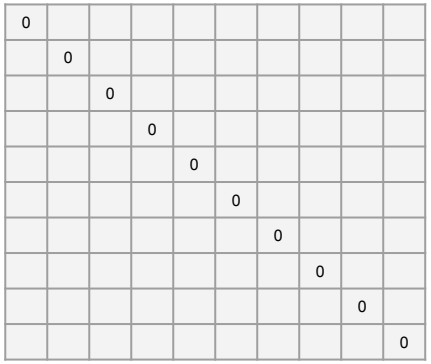


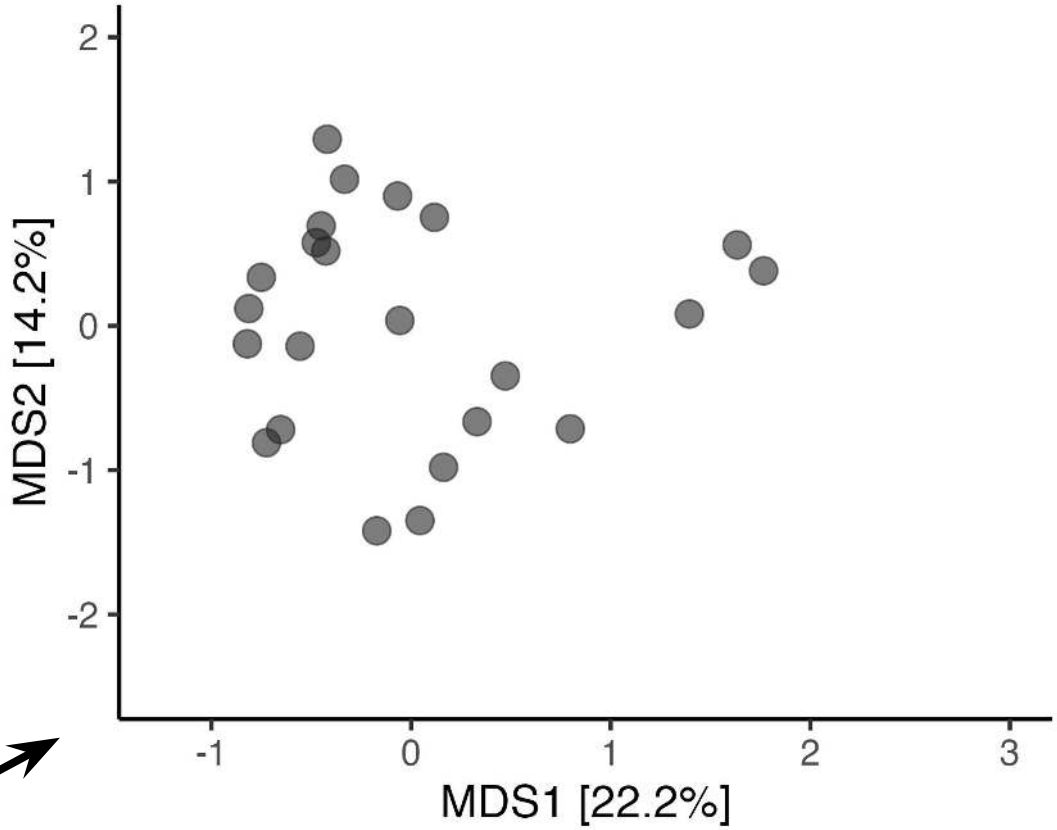
PCoA
on real data

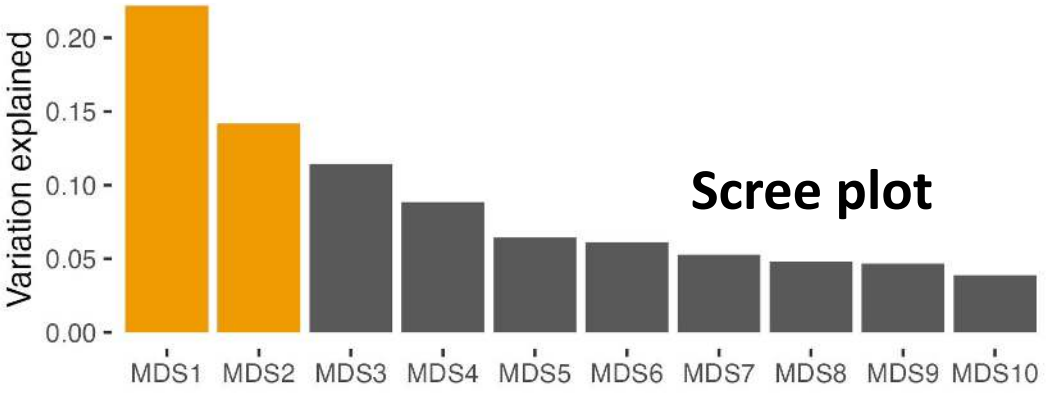
Ulcerative Colitis patients and Healthy Controls

Stool samples --> 16S microbiota abundances

Compute Bray-Curtis dissimilarities with genera







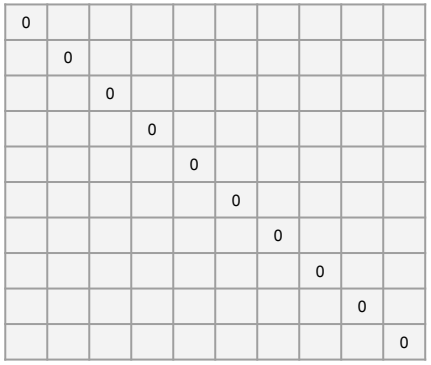
Scree plot

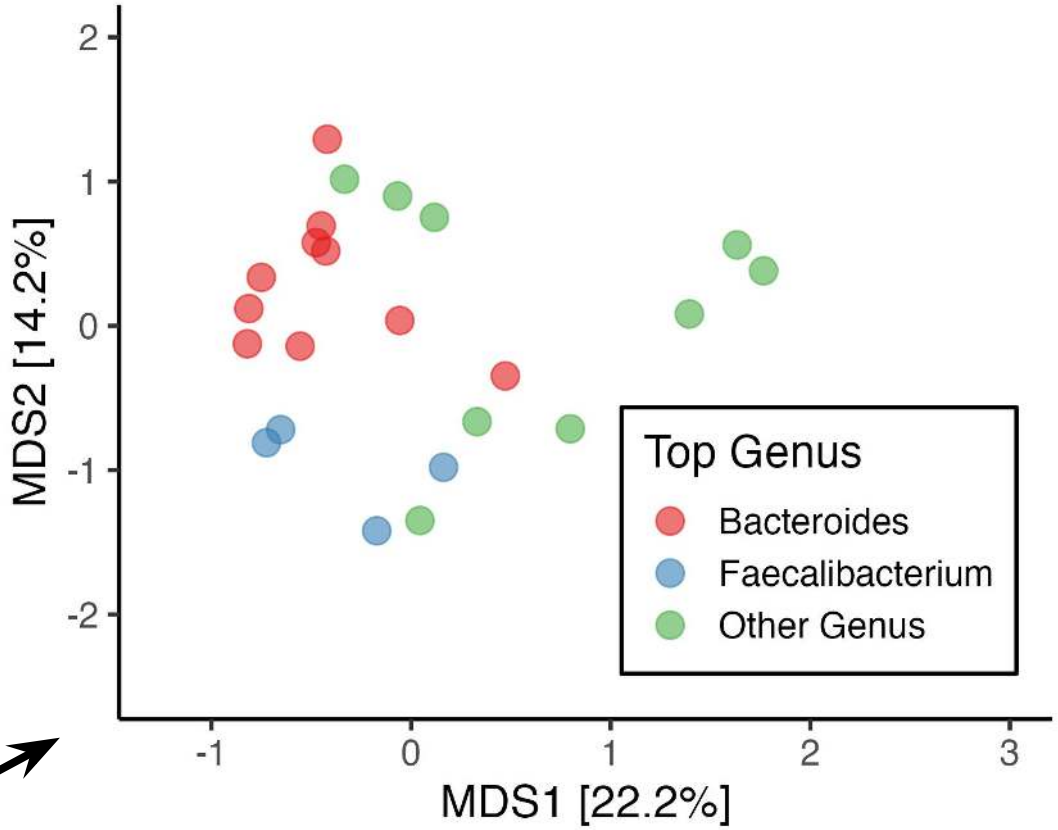
PCoA
on real data

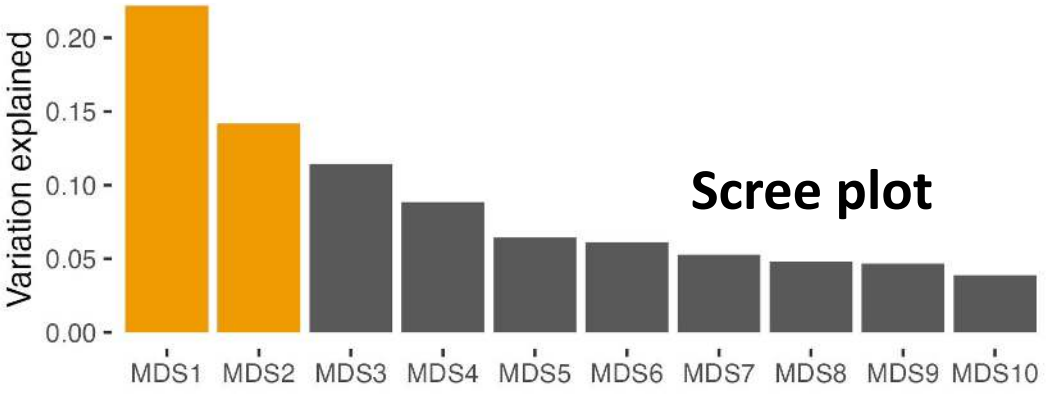
Ulcerative Colitis patients and Healthy Controls

Stool samples --> 16S microbiota abundances

Compute Bray-Curtis dissimilarities with genera







Scree plot

PCoA
on real data

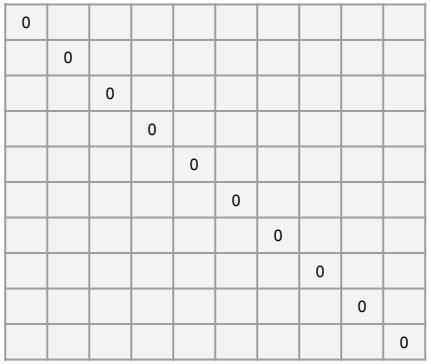
Ulcerative Colitis patients and Healthy Controls

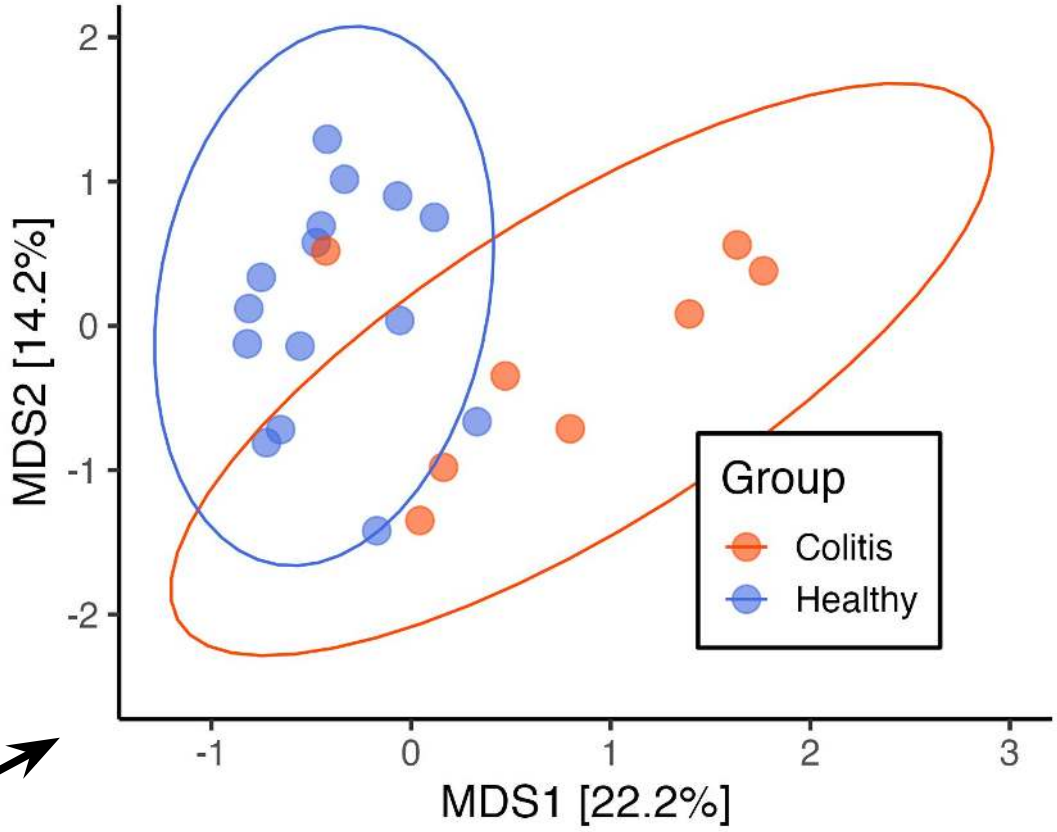
↓

Stool samples --> 16S microbiota abundances

↓

Compute Bray-Curtis dissimilarities with genera

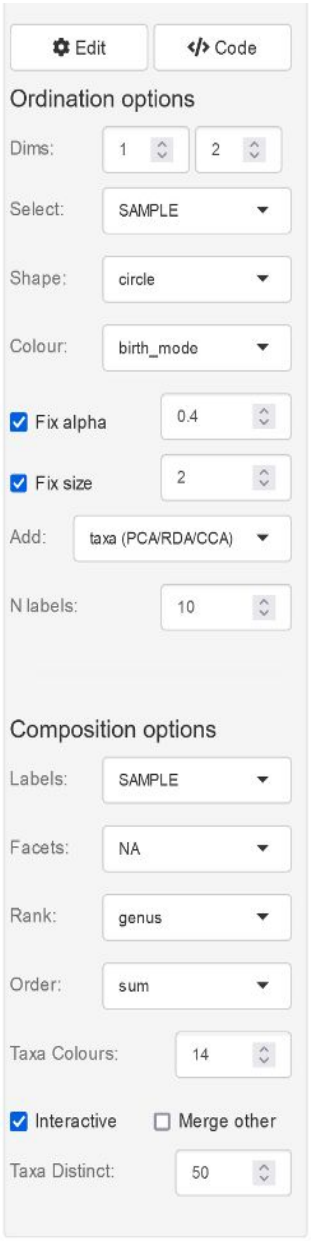


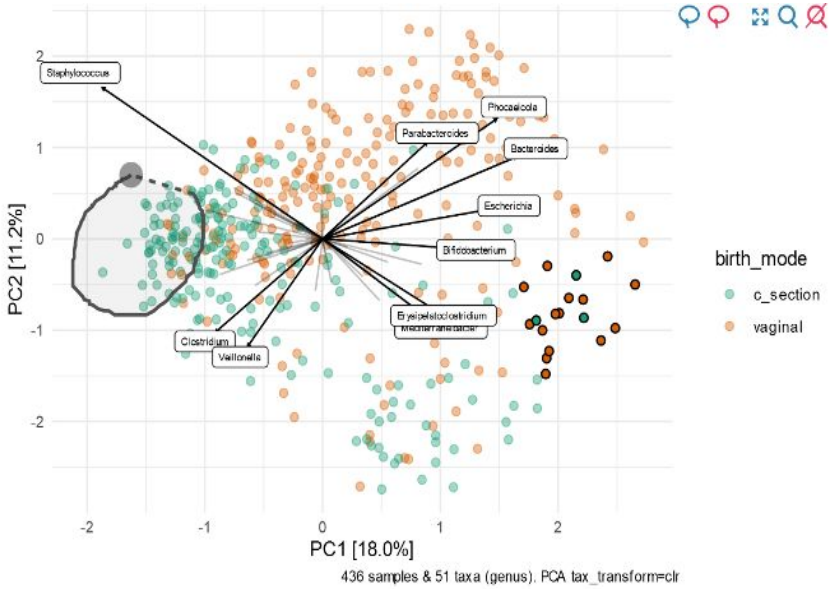


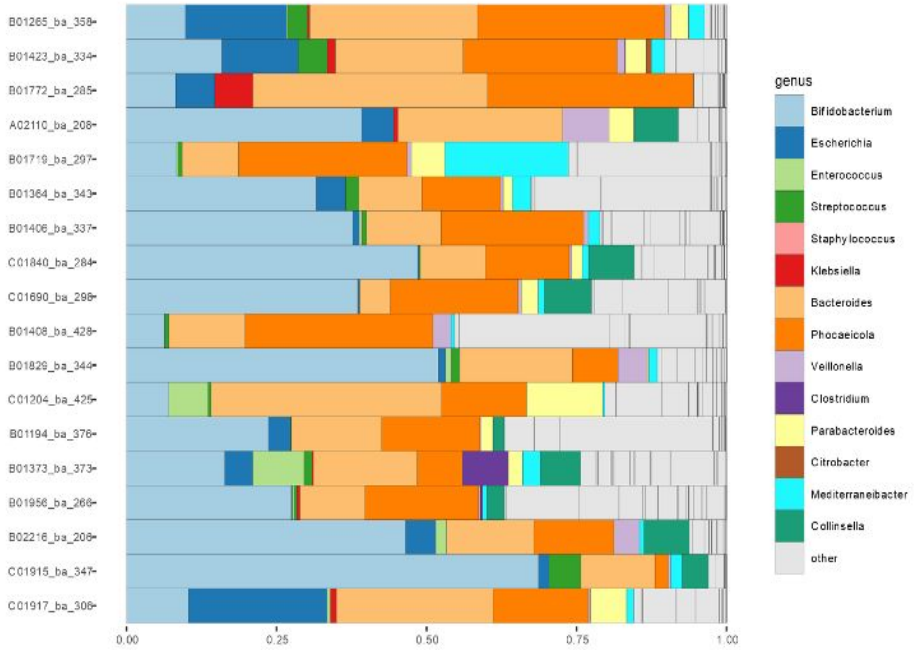
→ **PERMANOVA:** $p < 0.01$

- Permutational Multivariate ANOVA
- Does average composition differ by Group?

Interactive Ordination!







PCoA

Principal Coordinates Analysis

VS.

PCA

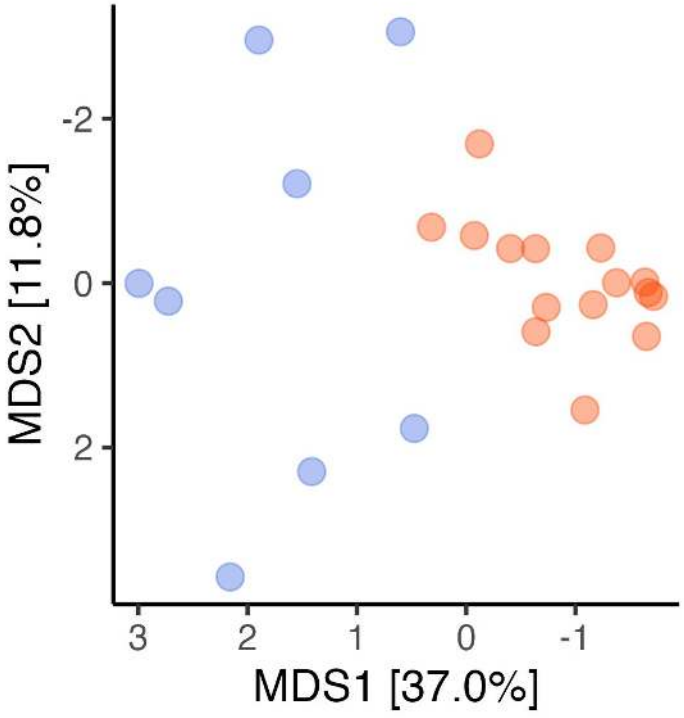
Principal Components Analysis

PCoA		PCA
Taxon abundances		Taxon abundances
Calculate Dissimilarities	↓	**Transform abundances**
Distance Matrix	↓	Transformed abundances
PCoA / MDS ordination	↓	**PCA ordination** (rotate input)
New dimensions (Coordinates)	↓	New dimensions (Components)
Plot first 2 or 3 dims	↓	**Plot first 2 or 3 dims**

PCoA

Coordinates

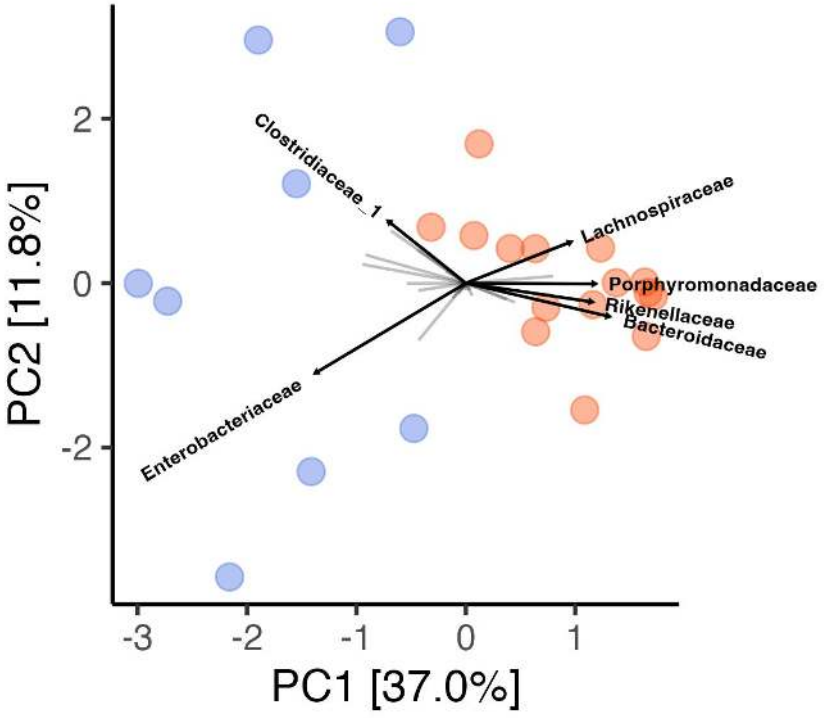
on Aitchison distances



PCA

Components

on CLR-transformed taxa



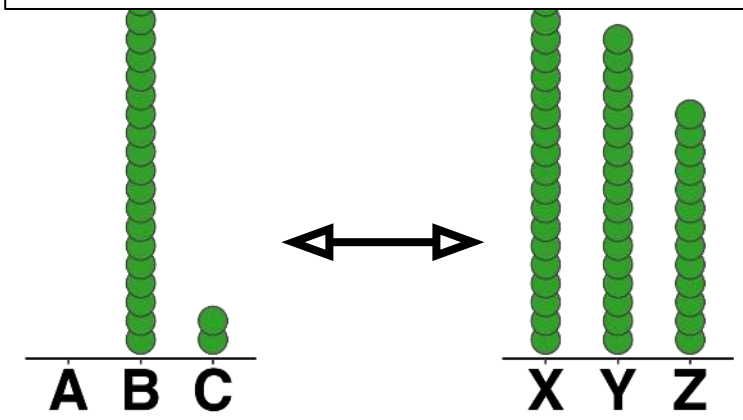
Differential Abundance

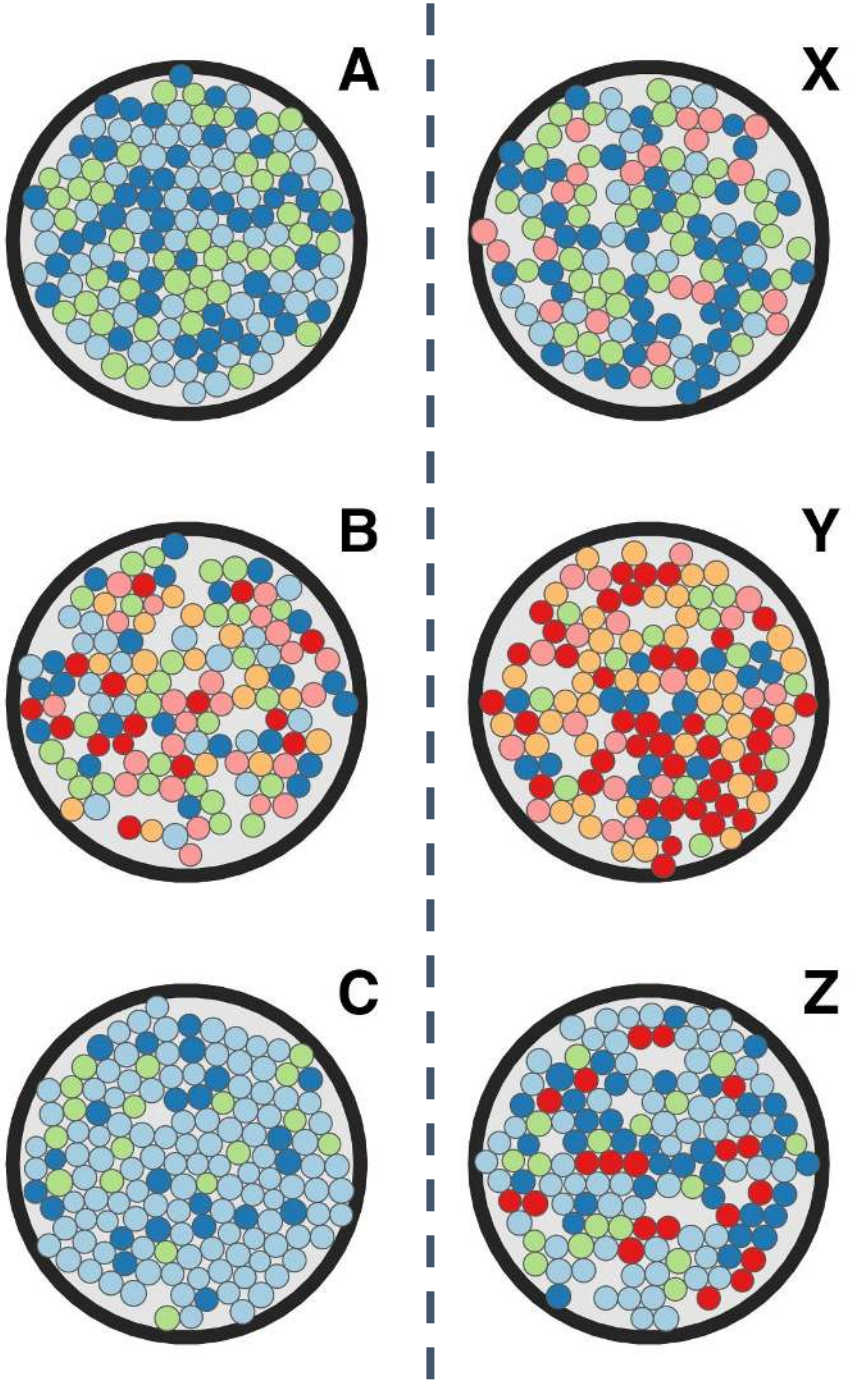
of each taxon

- **Compare across groups of samples**

  e.g. - group ABC vs. group XYZ

Various statistical methods available for differential abundance testing





Differential Abundance

- **Just comparing two groups of numbers, how hard can it be?**
- **But, we have compositional, noisy, and zero-inflated abundance counts…**
- **And about 15 different specialist methods to choose from…**

Microbiome differential abundance methods produce disturbingly different results across 38 datasets

Microbiome differential abundance methods produce different results across 38 datasets

Jacob T. Nearing[1,7✉], Gavin M. Douglas[1,7], Molly G. Hayes[2], Jocelyn MacDonald[3], Dhwani K. Desai[4], Nicole Allward[5], Casey M. A. Jones[6], Robyn J. Wright[6], Akhilesh S. Dhanani[4], André M. Comeau[4] & Morgan G. I. Langille[4,6]

Microbiome differential abundance methods produce different results… 😬 😱 😵

So what can you do?

1. **Filter out the noise**

   - Most methods perform poorly on rare taxa (which are also often less relevant)

2. **Keep it simple, and visualise your data!**

   - Check for visible patterns and start with non-parametric measures (Spearman correlations or Wilcox tests)

3. **Try a couple of common DA methods** (not DESeq2)

   - Pick methods that suit your dataset (covariates? repeated samples?)
   - Check where the methods agree on overlapping results

NOW: PCoA, PCA, and DA exercises in R

david-barnett.github.io/evomics-material/exercises/exercises_2.html

Continue until 10

Or until the need for beer or bed
becomes too strong…

Thank you & good luck

Microbiomes, imagined by Lexica