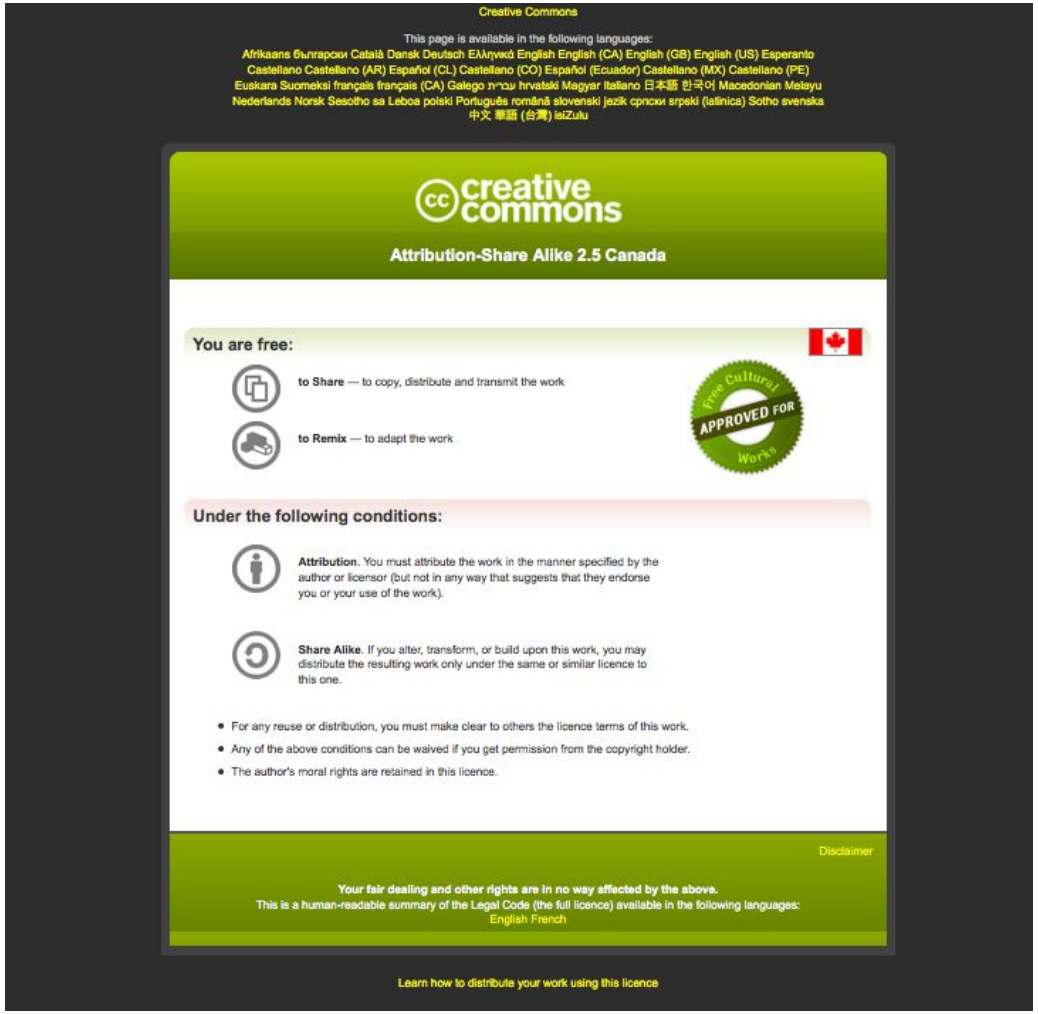
bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

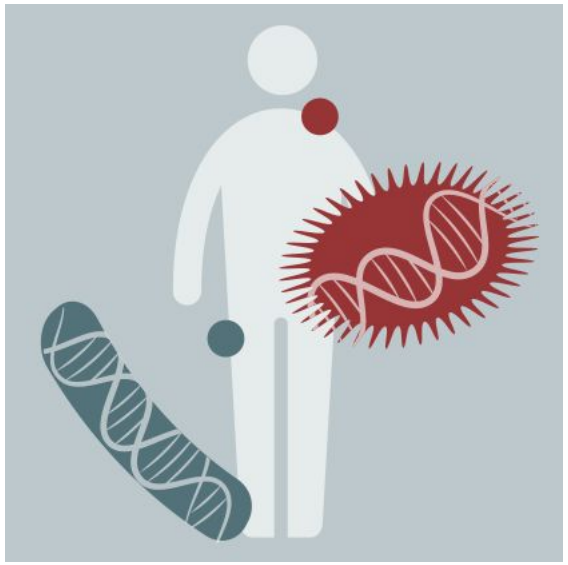

Creative Commons
This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu
creative commons
Attribution-Share Alike 2.5 Canada
You are free:
to Share — to copy, distribute and transmit the work
to Remix — to adapt the work
Free Cultural
APPROVED FOR
Works
Under the following conditions:
Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.
For any reuse or distribution, you must make clear to others the licence terms of this work.
Any of the above conditions can be waived if you get permission from the copyright holder.
The author's moral rights are retained in this licence.
Disclaimer
Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French
Learn how to distribute your work using this licence

Microbiome statistics and visualizations

Hena R. Ramay

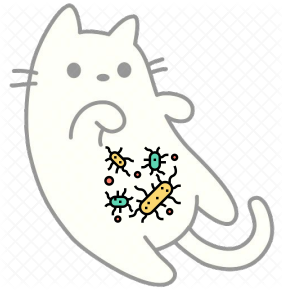
CBW-IMPACTT Microbiome Analysis

July 5-7, 2023



Learning Objectives

- By the end of this lecture, you will:
  - Statistical understanding of 16s data
  - Calculating alpha and beta-diversity
  - Basic statistics for diversity comparison of samples
  - Differential Abundance Analysis
  - Rarefy or not to rarefy



Definitions & Acronyms

Biosample:
Any type of sample that contains a microbial community from soil, water, animals

ASV or OTU table:
Amplicon Sequence Variant (ASV)
Operational Taxonomical Unit (OTU)

NGS: Next Generation Sequencing

Diversity:
Microbiome diversity refers to the variety and abundance of different microbial species or taxa within a given sample or ecosystem

Diversity measures: are quantitative metrics used to assess the richness and evenness of species or taxa within a given ecological community

Distance /dissimilarity metrics: The two terms are often used interchangeably when we talk about dissimilarity or similarity between pairs of objects. Dissimilarity matrix can be negative, non-symmetric etc.

Part1: Understanding 16s data

Is microbiome data count or compositional?

Compositional Data and Count Data are two distinct types of data commonly encountered in statistical analysis, each requiring different analytical approaches.

Compositional data

Represents relative proportions or percentages of different components within a whole.

Count data

Represents the absolute occurrences or events in each category or variable

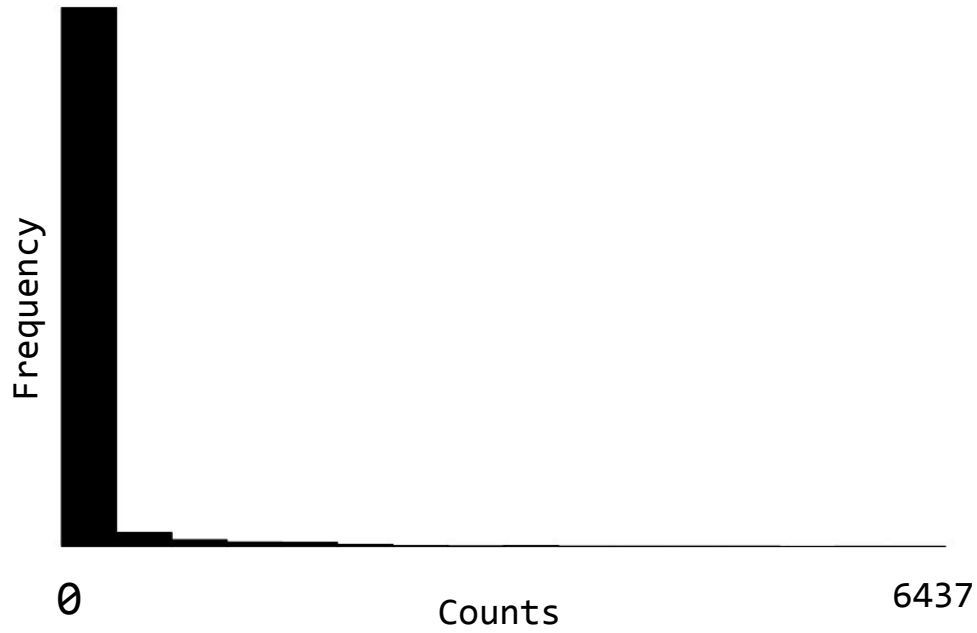
Is microbiome data count or compositional?

Sparse data

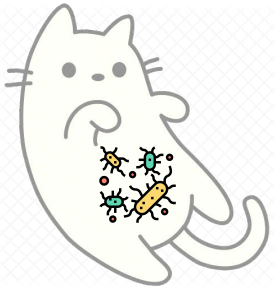
Samples

Taxa	S1	S2	S3	S4	S5	S6	S7
Taxa1	1240	0	33	56	0	77	45
Taxa2	0	0	5000	45	0	677	0
Taxa3	0	4563	0	1	0	0	0
Taxa4	534	34	0	0	563	45	0
Taxa5	99	84	0	6	856	643	6437
Taxa6	786	269	1	245	456	0	988

Distribution

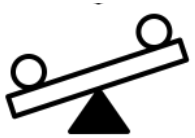


Are we seeing every species in this matrix that was supposed to be in the biosample?



Biosample

Amplify 16s region



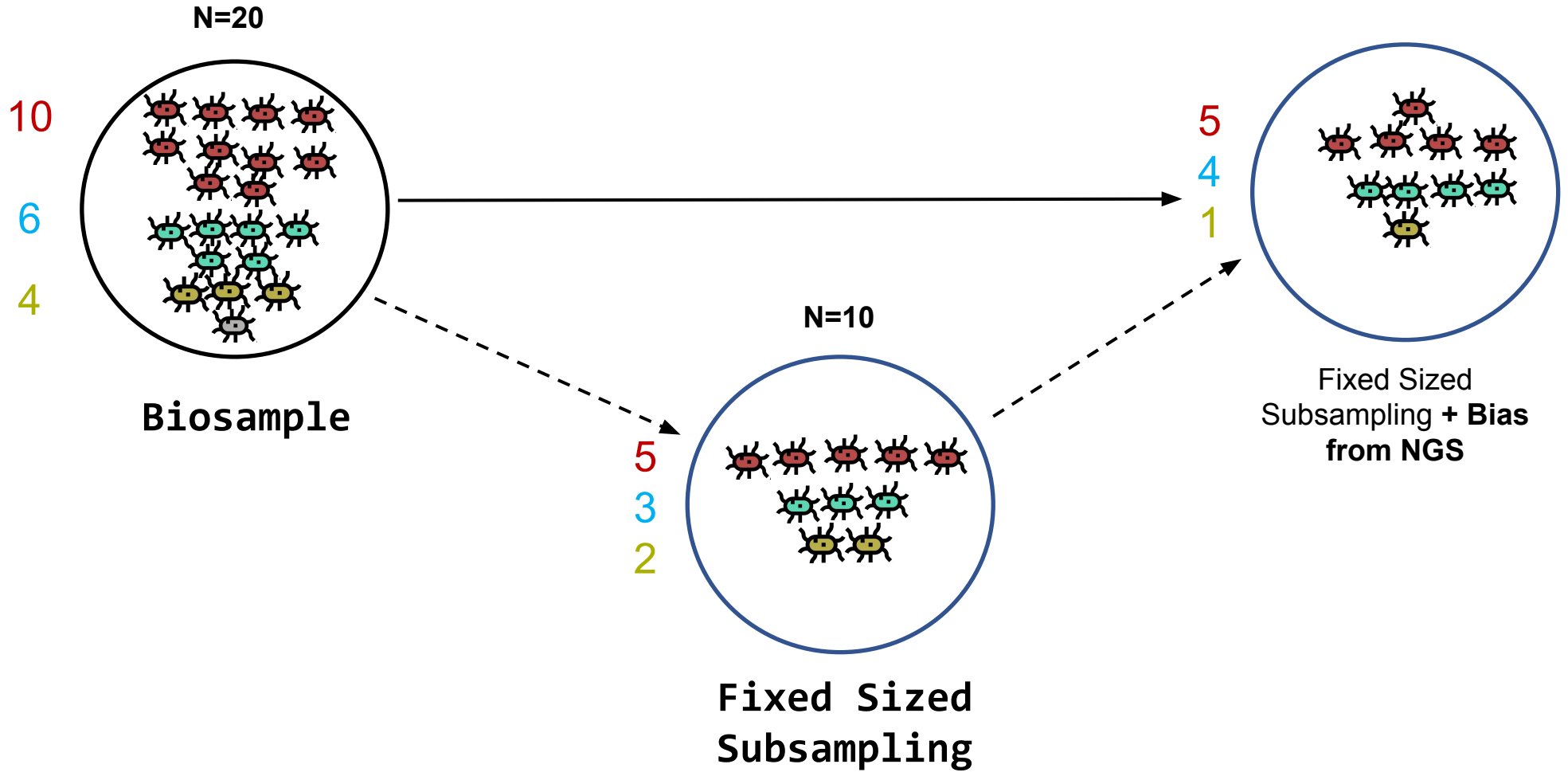
Next Generation Sequencing (NGS)

Maximum number of reads that a sample can be assigned is limited! (Fixed Sized)

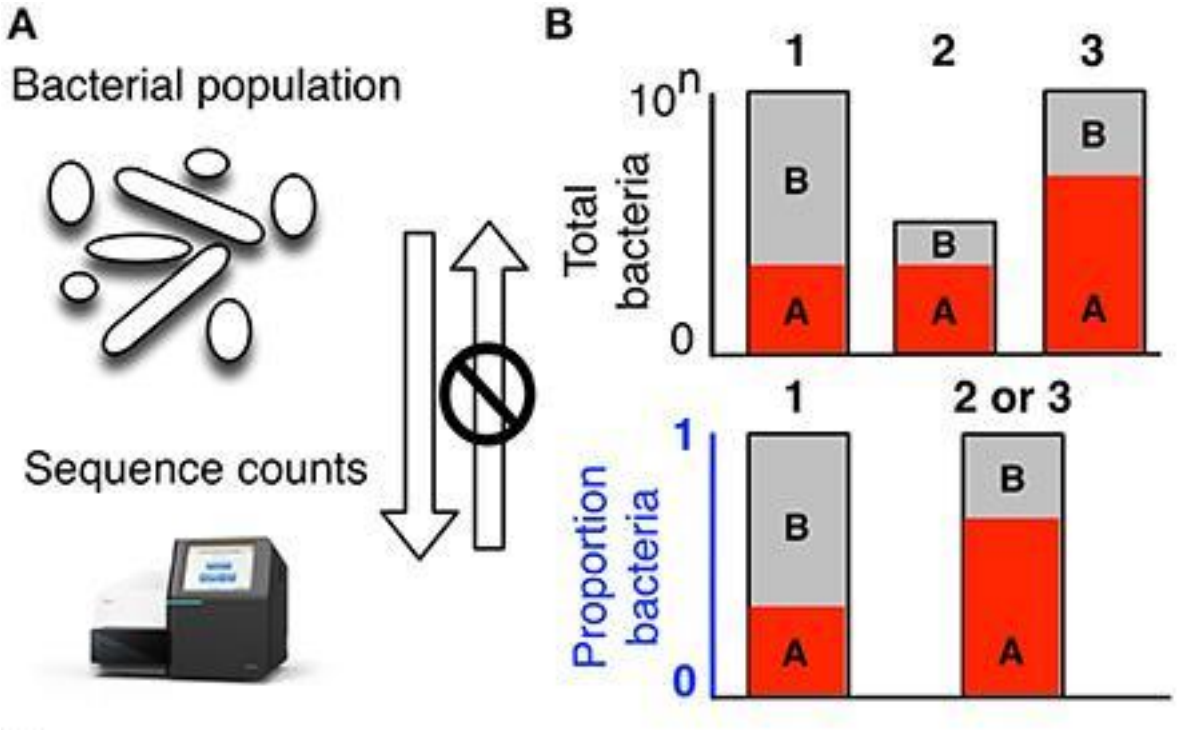
Samples

Taxa	S1	S2	S3	S4	S5	S6	S7
Taxa1	1240	0	33	56	0	77	45
Taxa2	0	0	5000	45	0	677	0
Taxa3	0	4563	0	1	0	0	0
Taxa4	534	34	0	0	563	45	0
Taxa5	99	84	0	6	856	643	6437
Taxa6	786	269	1	245	456	0	988

Who is left behind and why ?



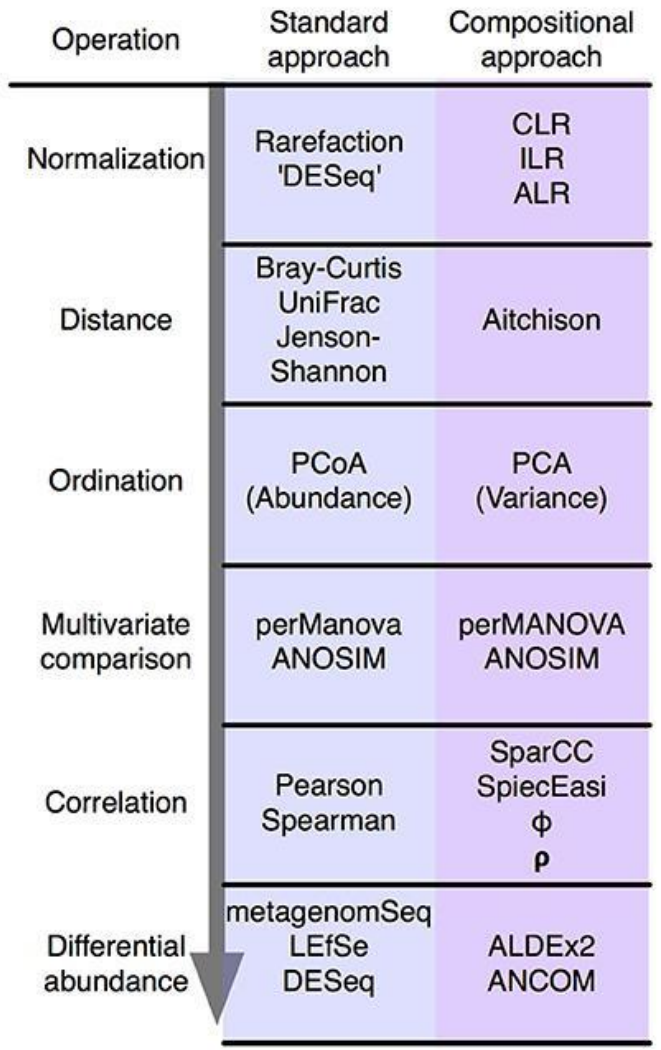
Compositional data



How does my choice of calling microbiome data compositional or count change things?

https://www.frontiersin.org/articles/10.3389/fmicb.2017.02224/full

The assumptions different methods make about the distribution of the data and the transformations that need to be made before a method can be used are different



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

As a user you must know these assumptions before using a method

https://www.frontiersin.org/articles/10.3389/fmicb.2017.02224/full

Part 2 :Diversity

Diversity

Diversity analysis plays a crucial role in understanding the microbial composition and dynamics within a microbiome dataset.

Alpha diversity

Within sample diversity
Who is in the sample

Number of Species ±
Abundance of species

Beta diversity

Between sample diversity
How similar are the samples

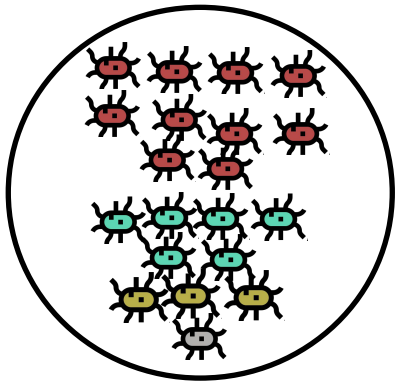
Distance/dissimilarity matrix

Alpha Diversity

10

6

4

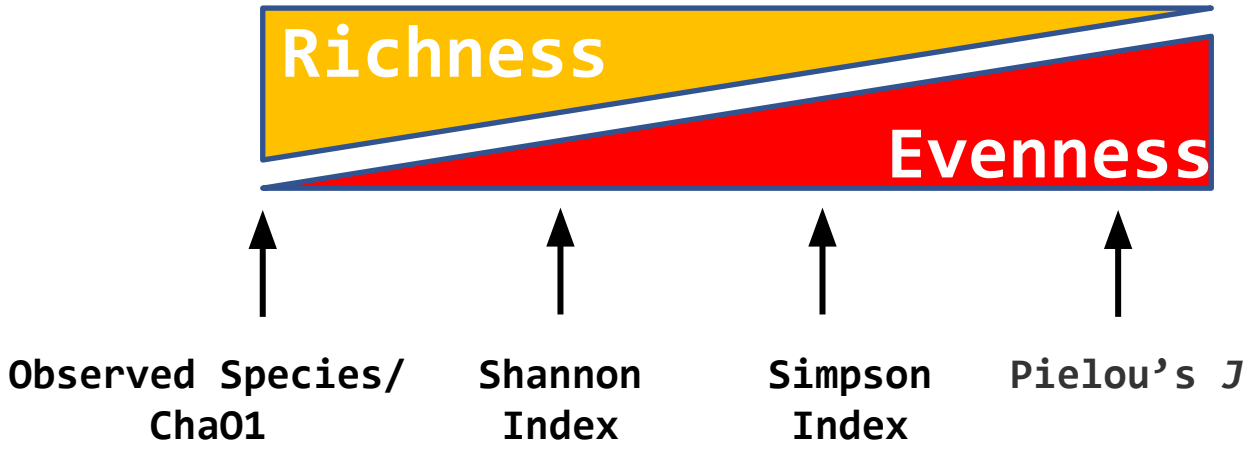


Biosample

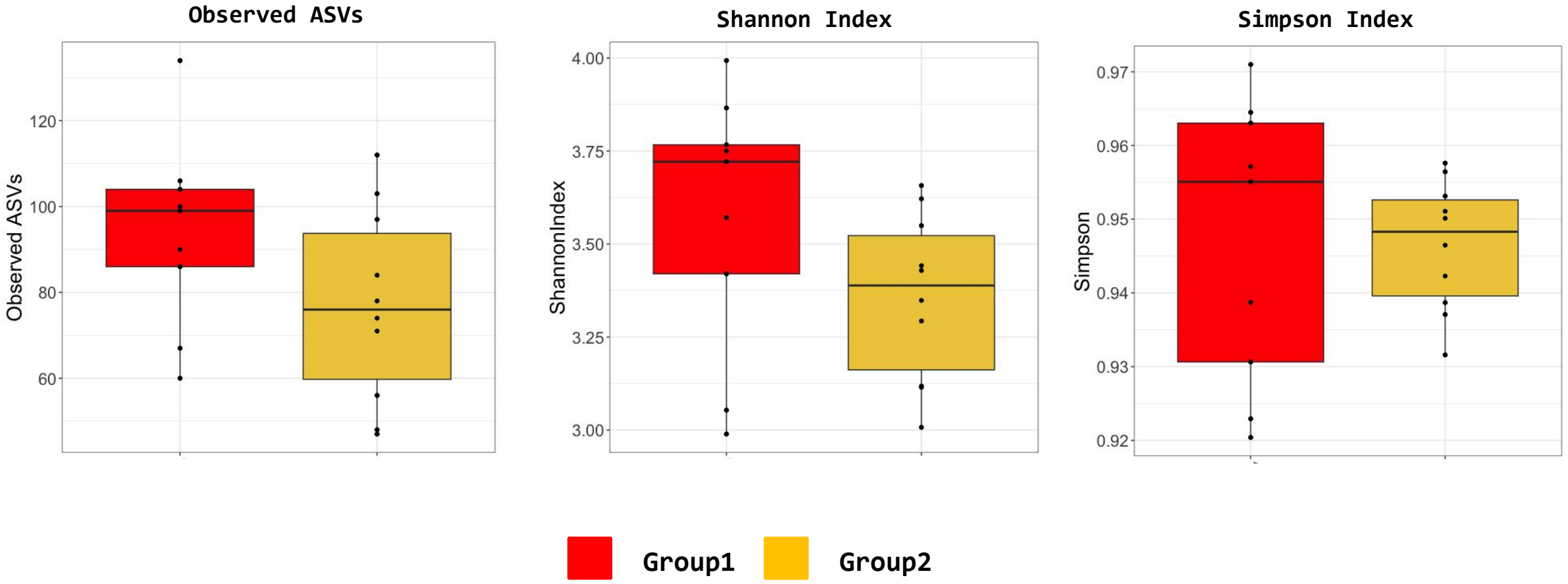
Richness
Number of species in a sample

Evenness
Evenness or relative abundance of species with a sample

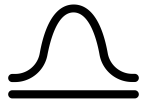
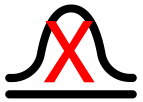
Quite a few alpha diversity indices exist which try to capture a **combination of richness and evenness** in a sample.



Alpha Diversity



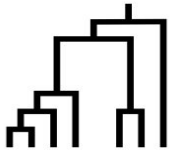
Alpha Diversity

ANOVA	Kruskal- Wallis
Parametric Test	Non-parametric Test
Requires the assumptions of **normality** and **equal variances** of groups 	No such requirement 
Compares **means** F-statistic: Measures the ratio of between-group variability to within-group variability.	Compares **median** H statistic: Measures the difference in the medians between groups.

If the p-value is significant, post-hoc tests can be conducted to identify which groups differ significantly from each other in case of more than two groups

Beta diversity

Beta diversity quantifies the differences in species composition between different locations or conditions



	Non-phylogenetic	Phylogenetic
Absence/Presence	Jaccard	Unifrac
Absence/Presence + Abundance	Bray-Curtis	Weighted- Unifrac

Beta diversity can be measured using various distance or dissimilarity metrics, such as Bray-Curtis dissimilarity, Jaccard index, and UniFrac distance.

Bray-Curtis dissimilarity

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

	S1	S2	Smaller value
Taxa1	1240	0	0
Taxa2	0	0	0
Taxa3	0	4563	0
Taxa4	534	34	34
Taxa5	99	84	84
Taxa6	786	269	269
Total Per Sample	**2659**	**4950**	**387**

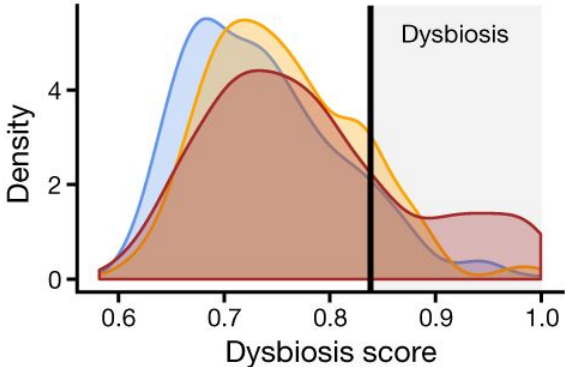
$$BC_{s1s2} = 1 - \frac{2(387)}{(2659 + 4950)}$$

$$BC_{s1s2} = \mathbf{0.89}$$

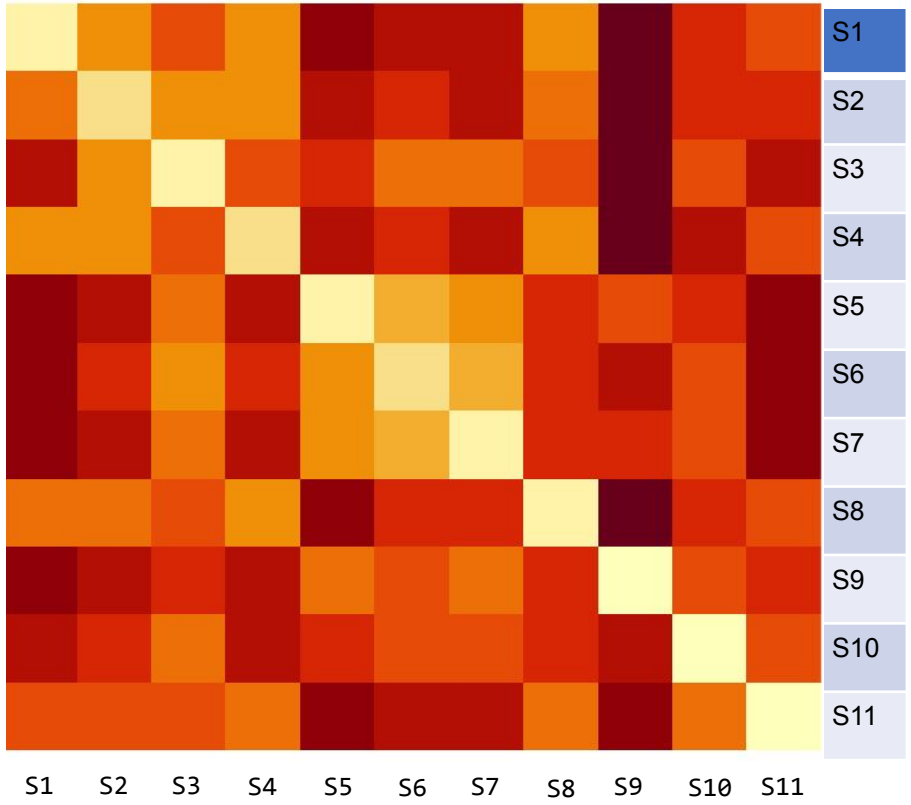
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
S1	0	**0.89**	0.47	0.28	0.69	0.61	0.63	0.32	0.79	0.56	0.45
S2	**0.89**	0	0.27	0.25	0.58	0.49	0.52	0.3	0.74	0.47	0.44
S3	0.47	0.27	0	0.37	0.4	0.28	0.33	0.36	0.62	0.38	0.5
S4	0.28	0.25	0.37	0	0.58	0.49	0.52	0.25	0.71	0.52	0.4
S5	0.69	0.58	0.4	0.58	0	0.22	0.26	0.57	0.5	0.51	0.66
S6	0.61	0.49	0.28	0.49	0.22	0	0.16	0.48	0.55	0.41	0.58
S7	0.63	0.52	0.33	0.52	0.26	0.16	0	0.45	0.51	0.41	0.59
S8	0.32	0.3	0.36	0.25	0.57	0.48	0.45	0	0.68	0.47	0.4
S9	0.79	0.74	0.62	0.71	0.5	0.55	0.51	0.68	0	0.54	0.65
S10	0.56	0.47	0.38	0.52	0.51	0.41	0.41	0.47	0.54	0	0.42
S11	0.45	0.44	0.5	0.4	0.66	0.58	0.59	0.4	0.65	0.42	0

Bray-Curtis dissimilarity

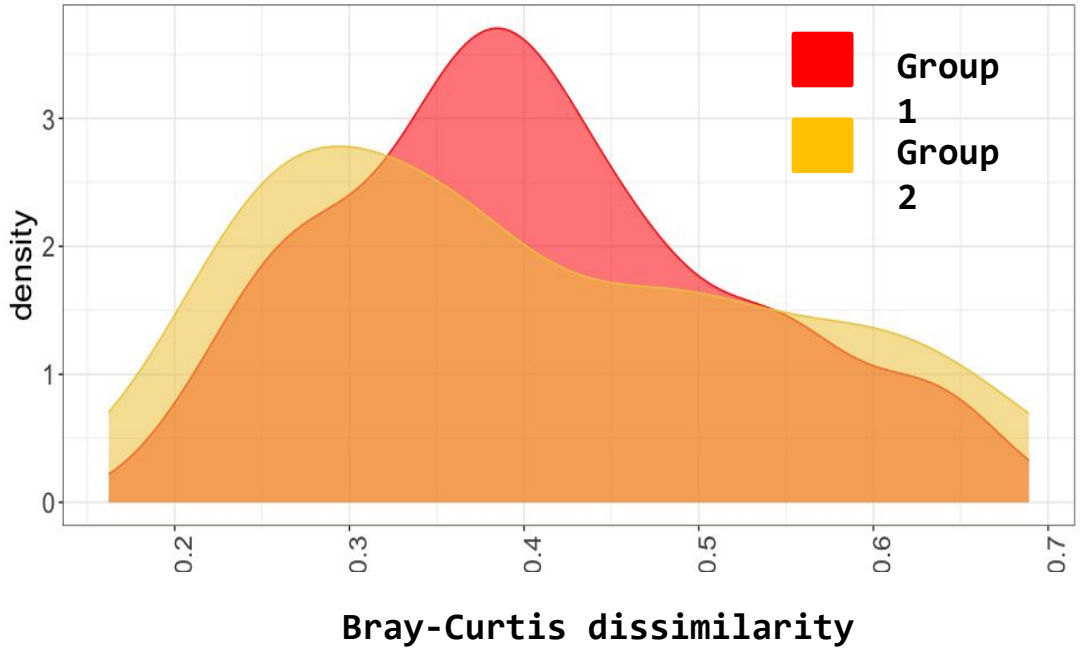
Example



Heatmap



Density Plots



https://www.nature.com/articles/s41586-019-1237-9

Ordination

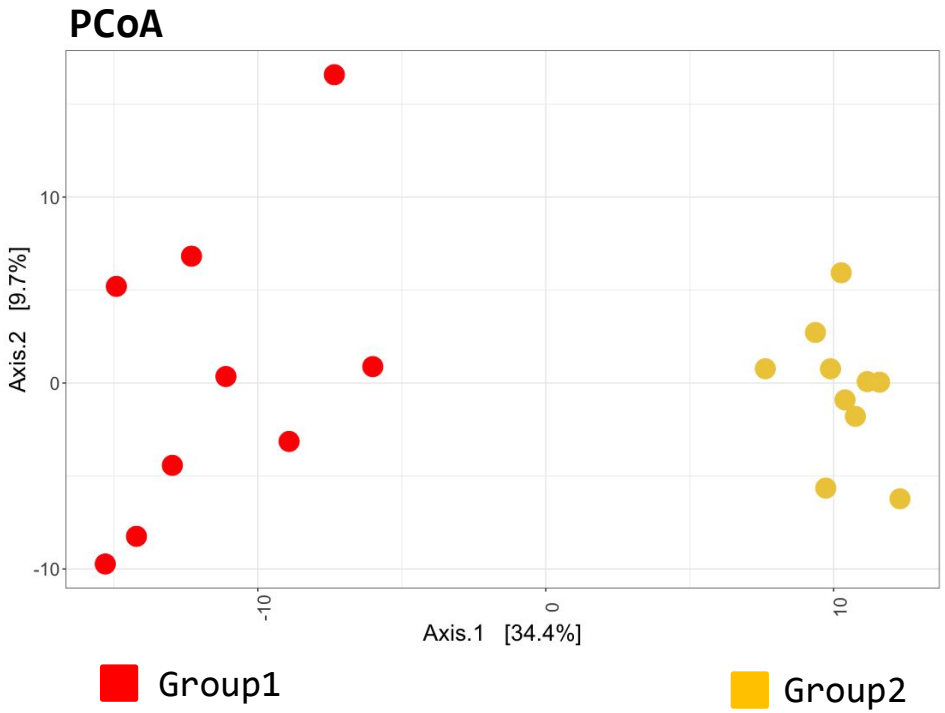
Ordination methods are statistical techniques used to visualize and analyze multivariate data. These methods aim to reduce the dimensionality of the data while preserving the relationships among samples or variables

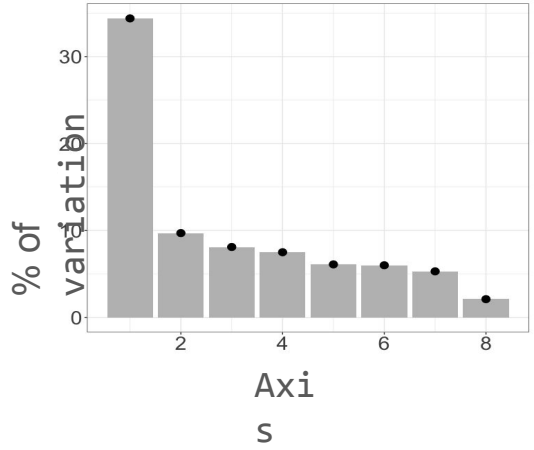
Methods:

Principal Component Analysis (**PCA**)
Principal Coordinates Analysis(**PCoA**)
Nonmetric Multidimensional Scaling (**NMD**
Correspondence Analysis (**CA**)

Example

In PCoA, Axis.1 captures the largest source of variation in the Bray-Curtis dissimilarities. It represents the most significant differences or patterns between the samples in terms of their community composition.





PERMANOVA: Permutational Multivariate Analysis of Variance

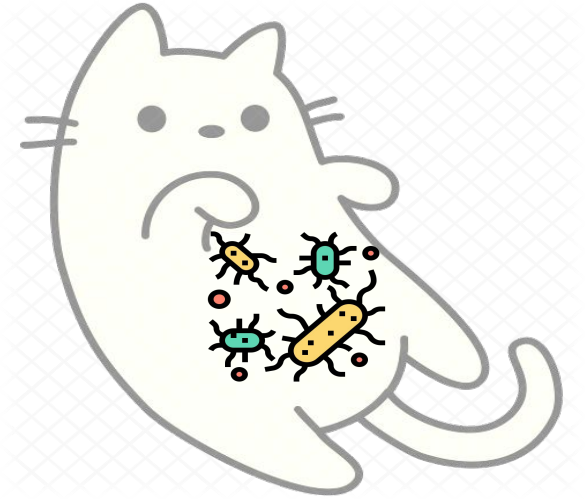
Non-parametric multivariate statistical test that assesses the significance of differences in community composition among groups.

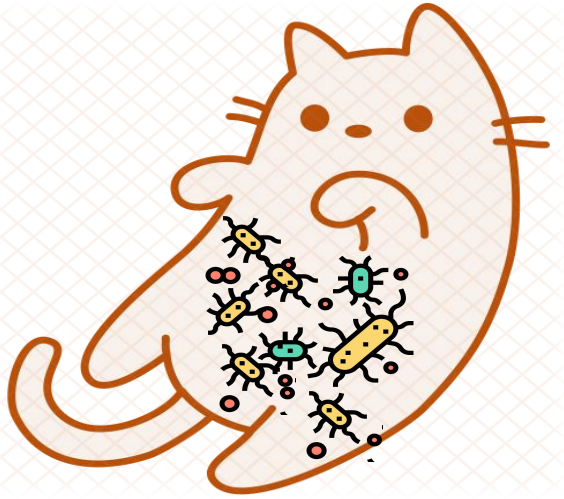
It uses permutation-based tests to determine if the dissimilarities between groups are significantly different from what would be expected by chance.

Calculates the F statistics like in ANOVA

Concerns:
Order in which terms are added matters
If the p-value is significant, it is recommended to check the **dispersion** of the groups

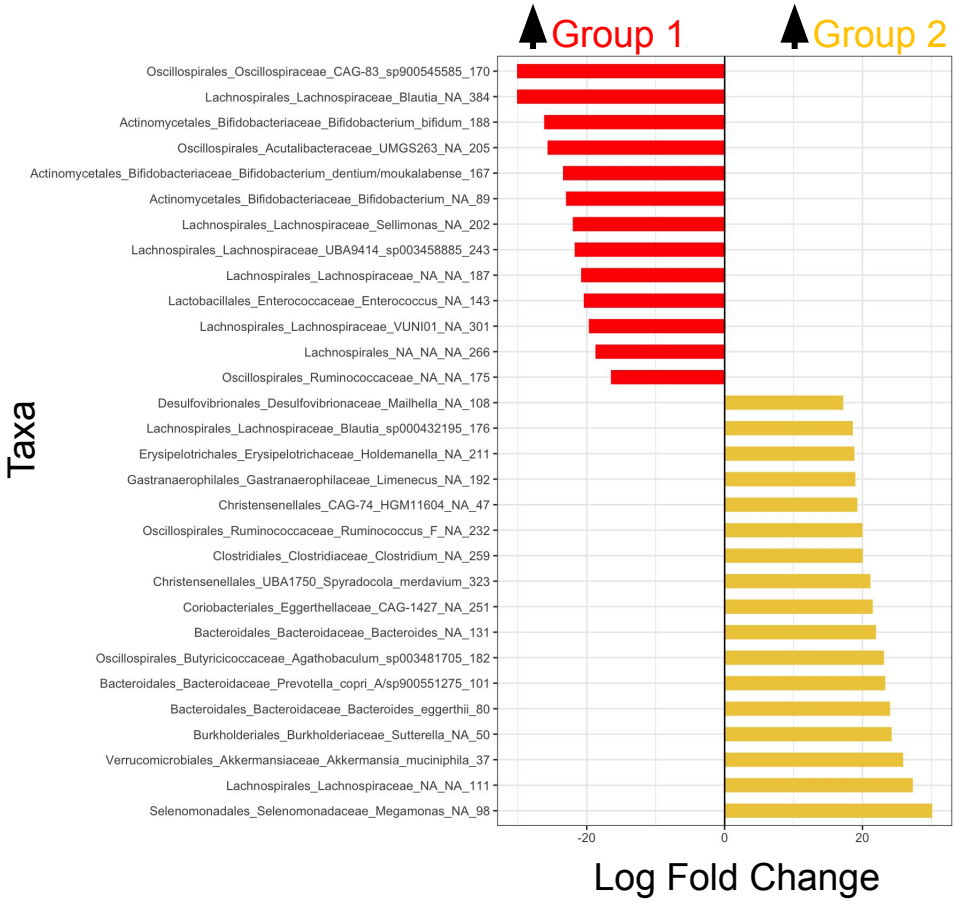
Differential Abundance Analysis





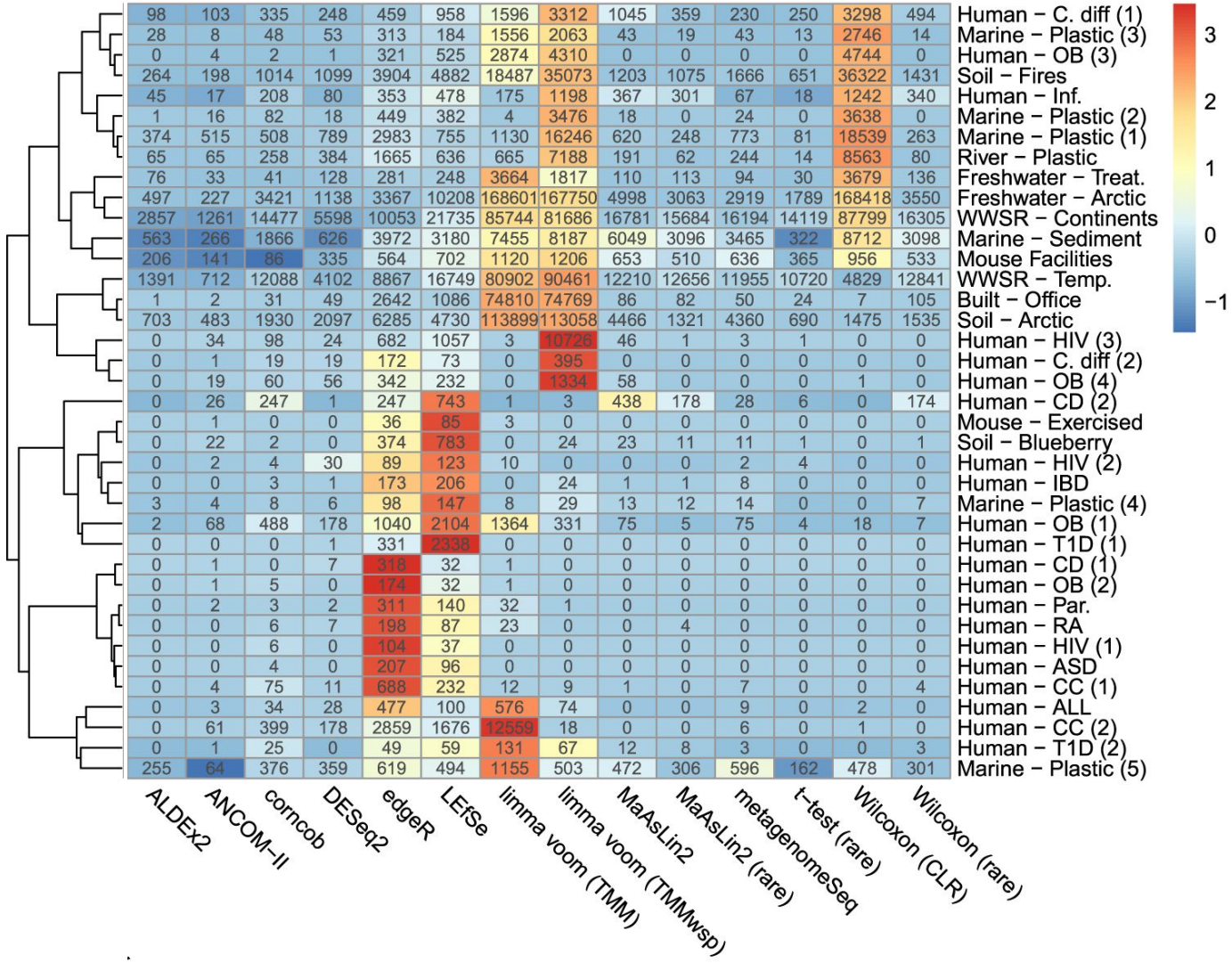
Are there any taxa that are present in different abundance in the two types of samples?

Differential Abundance Analysis



1. DESeq2
2. Maaslin2
3. edgeR
4. metagenomeSeq
5. ANCOM2
6. ANCOM-BC
7. corncob
8. ALDEx2

Differential Abundance Analysis



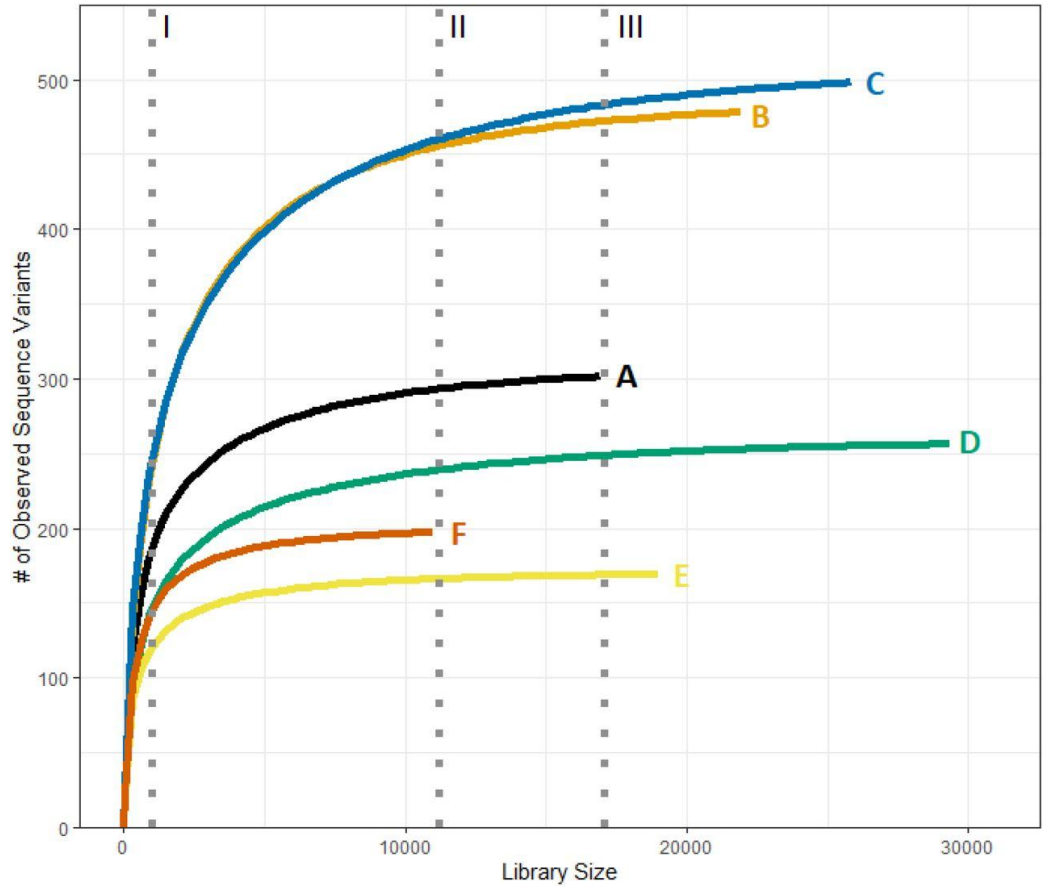
Authors' Recommendation: Use the same tools when comparing results between specific studies and otherwise use a consensus approach based on several DA tools to help ensure results are robust to DA choice!

https://www.nature.com/articles/s41467-022-28034-z

Variation in read depth

It is very difficult to get same number of reads assigned to all samples even on the same sequencing run.

Sample	Observed ASVs	Total reads
A	300	16500
B	458	21000
C	500	26000
D	252	29000
E	165	19000
F	200	11000



https://www.nature.com/articles/s41598-021-01636-1

Rarefy or not to Rarefy?

My rule: Instead of rarefying/subsampling use an appropriate transformation!Rarefaction can cause artifacts.

No	Check
Rare Taxa Might be missed by subsampling	**Sample Coverage** 10x > variation in sample depth Uneven sampling depth can mitigate biases in diversity measures.
Statistical Power decreased statistical power for detecting significant differences between groups	

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003531

We are on a Coffee Break & Networking Session

Workshop Sponsors: