Intro to Compositional Data

Alise Ponsero

DS meeting – January 2025

This workshop

Why microbiome data is compositional data?

What is the issue with dealing with compositional data?

What are log-ratio transform approaches ?

How to use compositional data analysis methods with microbiome data?

This workshop

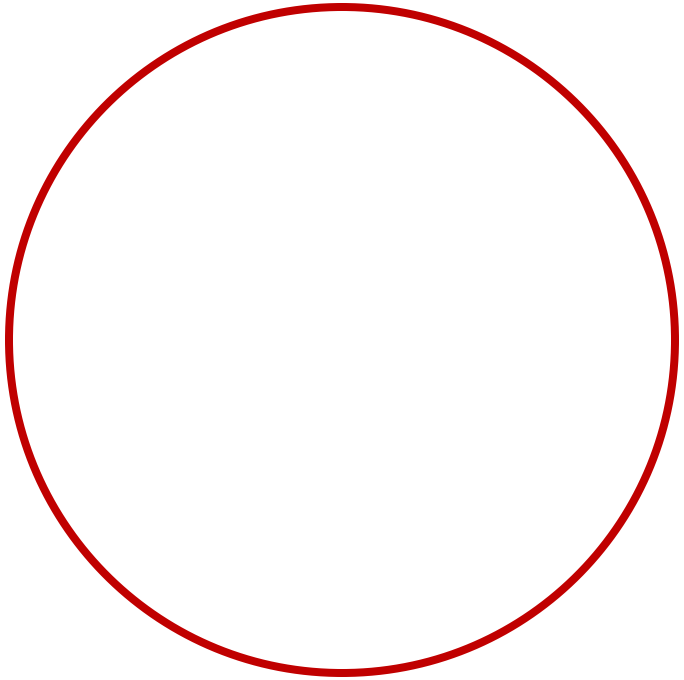
Why microbiome data is compositional data?

What is the issue with dealing with compositional data?

What are log-ratio transform approaches ?

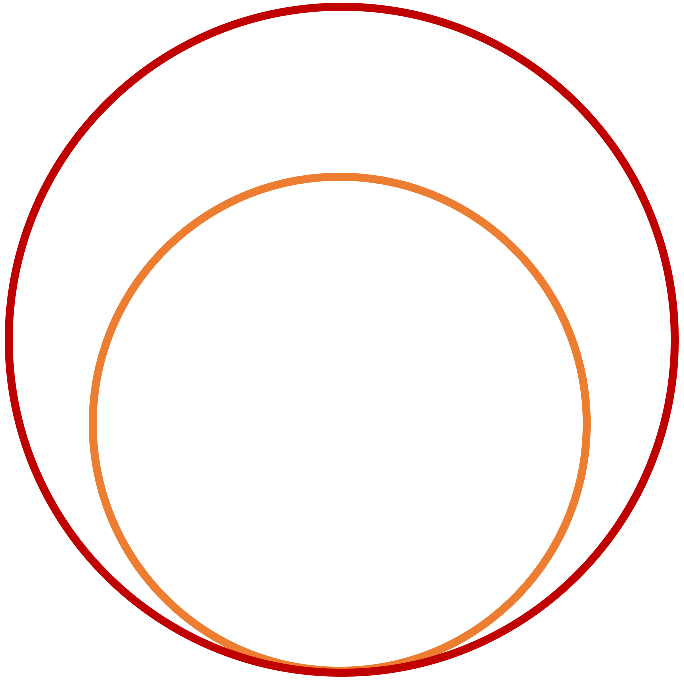
How to use compositional data analysis methods with microbiome data?

Microbiome data characteristics



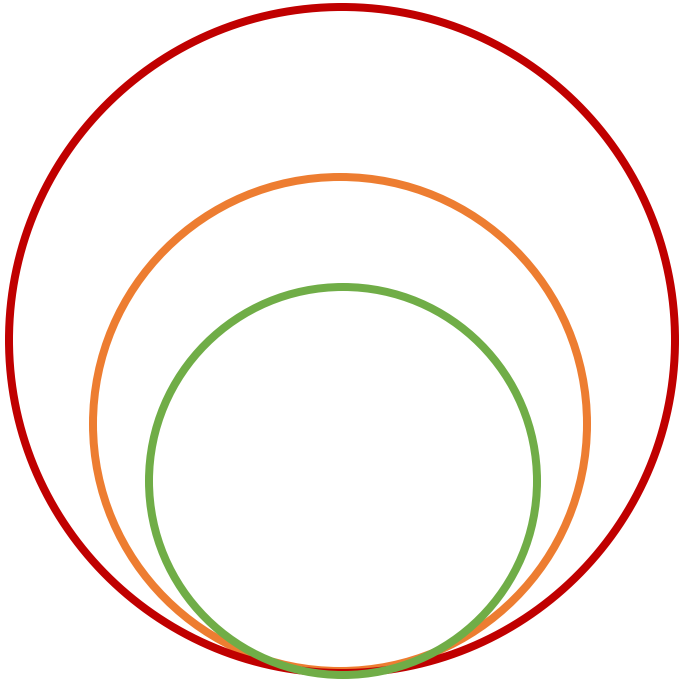
- Microbial population in the ecosystem

Microbiome data characteristics



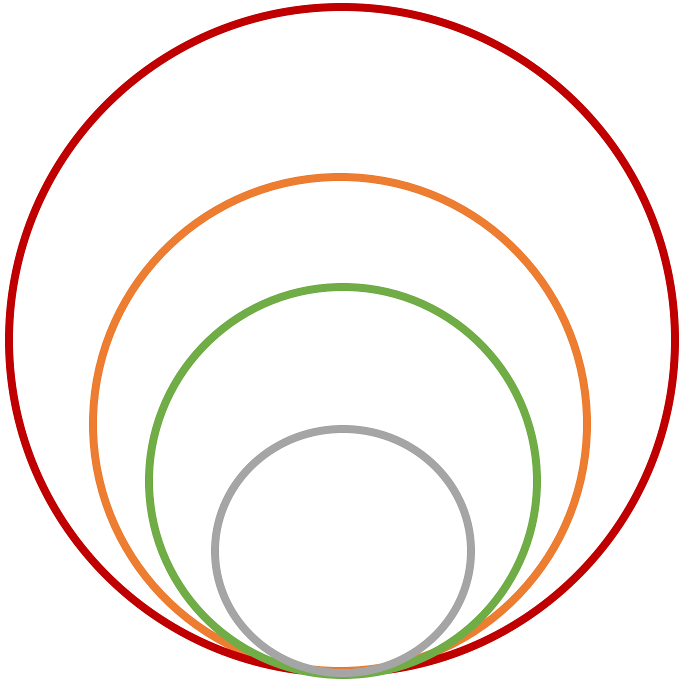
- Microbial population in the ecosystem
- Sample

Microbiome data characteristics



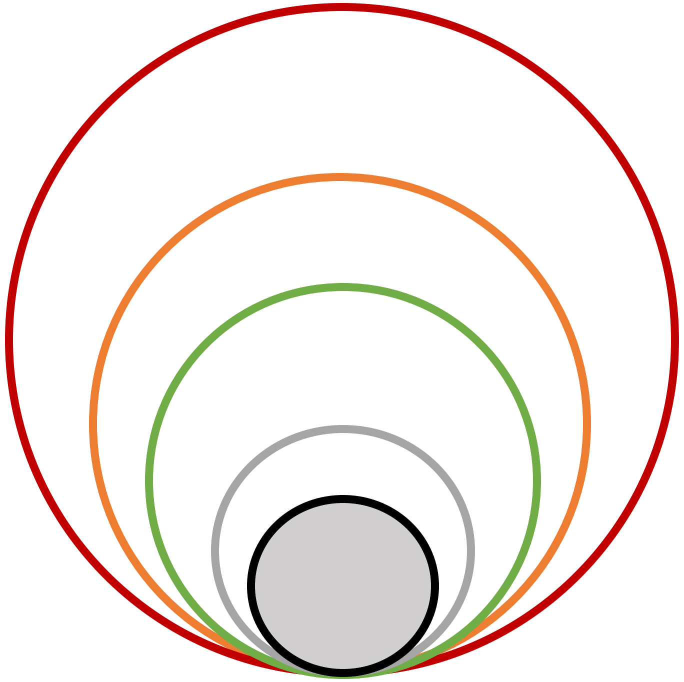
- Microbial population in the ecosystem
- Sample
- DNA/RNA extraction

Microbiome data characteristics



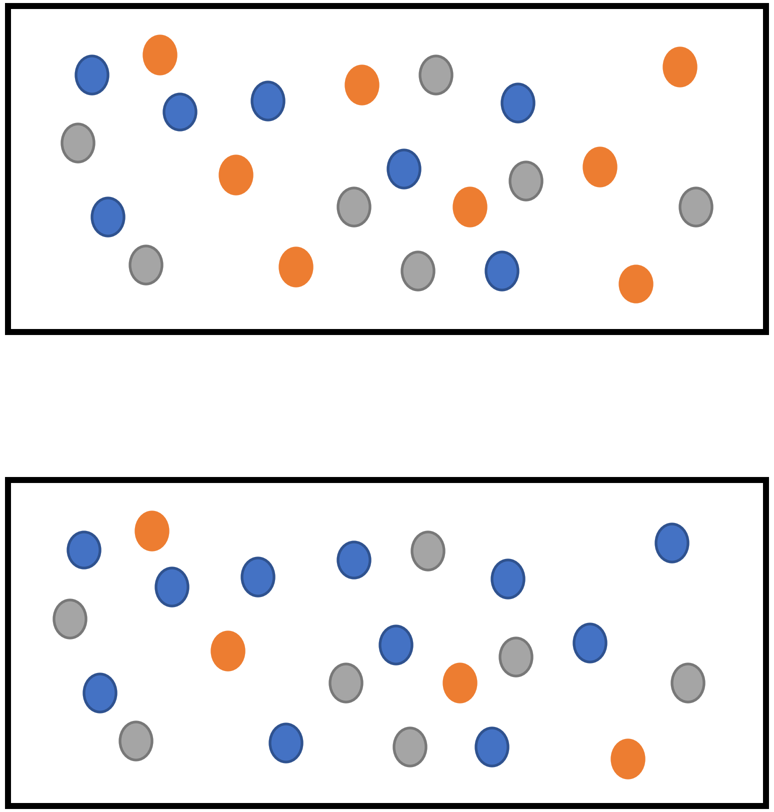
- Microbial population in the ecosystem
- Sample
- DNA/RNA extraction
- Sequencing library

Microbiome data characteristics

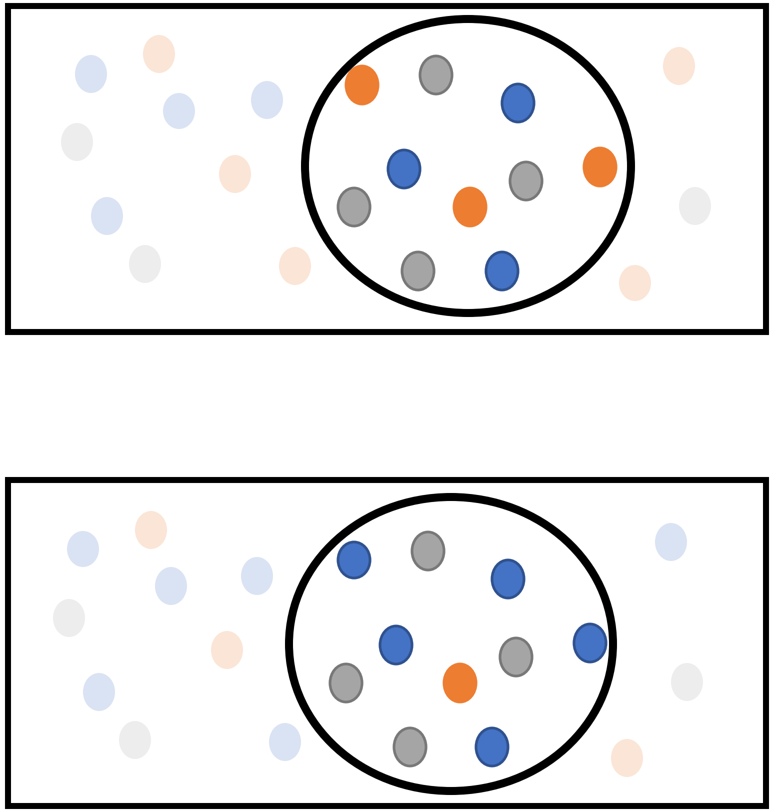


- Microbial population in the ecosystem
- Sample
- DNA/RNA extraction
- Sequencing library
- Reads from sequencer

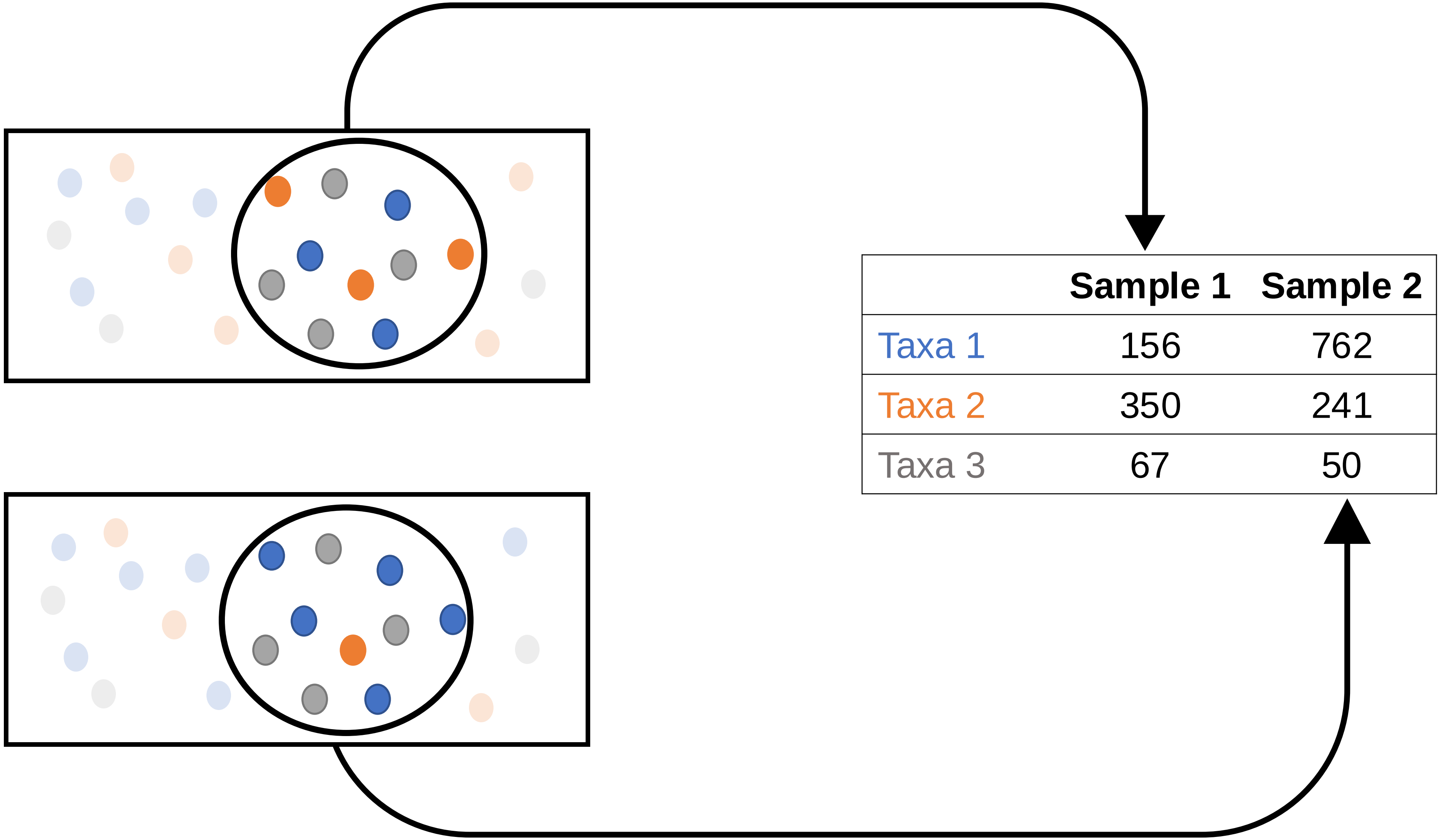
Microbiome data characteristics



Microbiome data characteristics



Microbiome data characteristics



	Sample 1	Sample 2
Taxa 1	156	762
Taxa 2	350	241
Taxa 3	67	50

Microbiome data characteristics

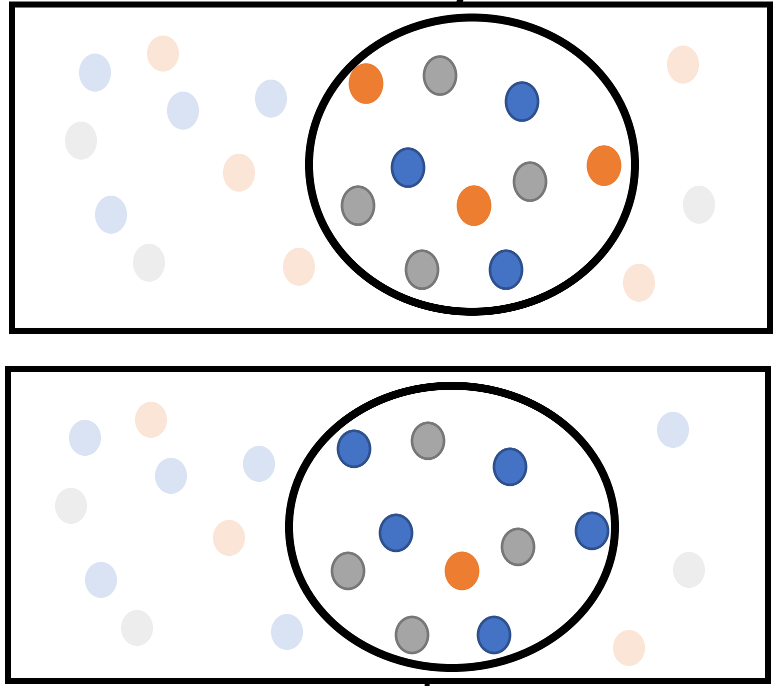
What does it mean to have a “zero” count?

	Sample 1	Sample 2
Taxa 1	156	762
Taxa 2	350	241
Taxa 3	67	**0**

Essential zero = Taxa 3 is absent from the ecosystem

Count Zero = Taxa 3 is present in the ecosystem but was missed by the sampling or sequencing

Microbiome data is inherently compositional



	Sample 1	Sample 2
Taxa 1	156	762
Taxa 2	350	241
Taxa 3	67	0
sum	573	1003

Absolute counts are not biologically informative

Microbiome data is inherently compositional

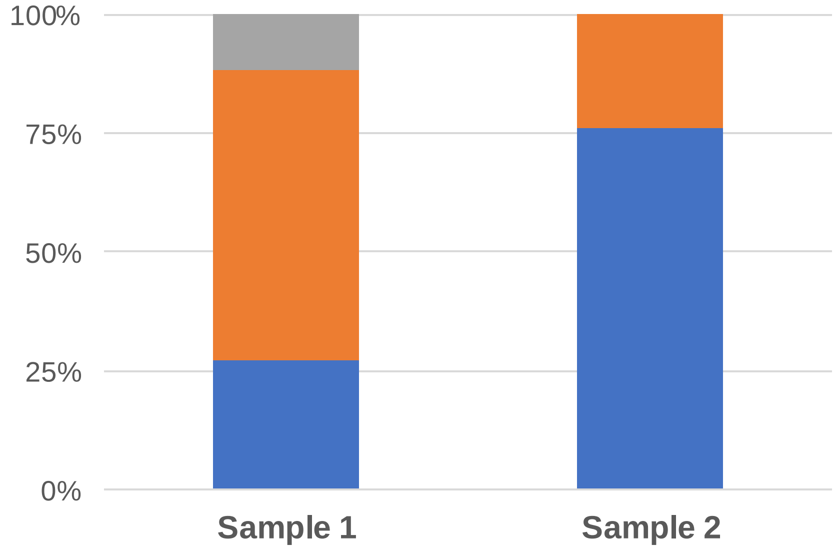
	Sample 1	Sample 2
Taxa 1	156	762
Taxa 2	350	241
Taxa 3	67	0
sum	573	1003

$$\xrightarrow[\frac{\text{Count Taxa}}{\text{Total counts}}]{\textbf{TSS}}$$

	Sample 1	Sample 2
Taxa 1	0.27	0.76
Taxa 2	0.61	0.24
Taxa 3	0.12	**0**
sum	1	1

Microbiome data is inherently compositional

	Sample 1	Sample 2
Taxa 1	156	762
Taxa 2	350	241
Taxa 3	67	0
sum	573	1003



This workshop

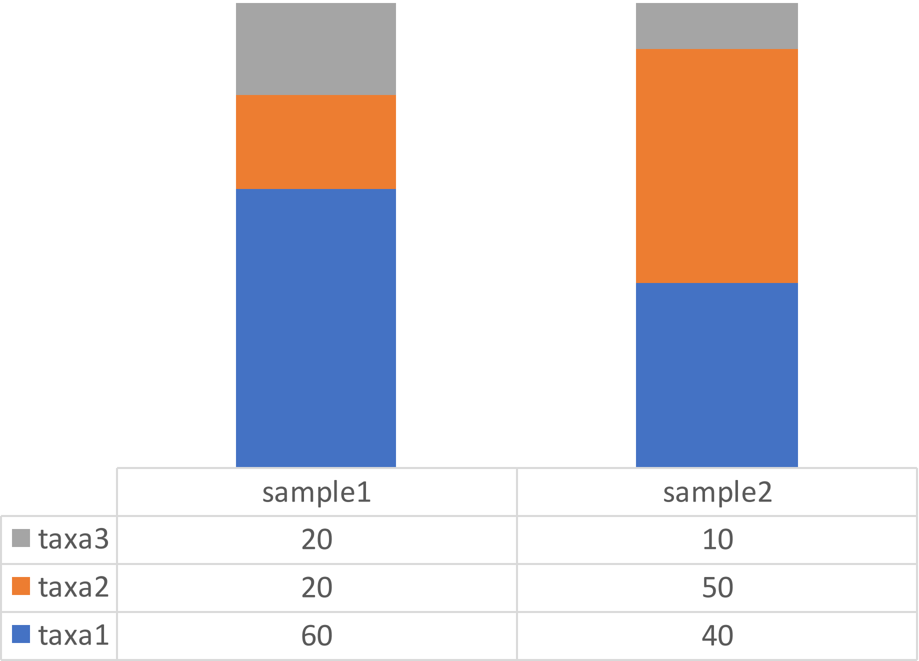
Why microbiome data is compositional data?

What is the issue with dealing with compositional data?

What are log-ratio transform approaches ?

How to use compositional data analysis methods with microbiome data?

The issue with compositional data

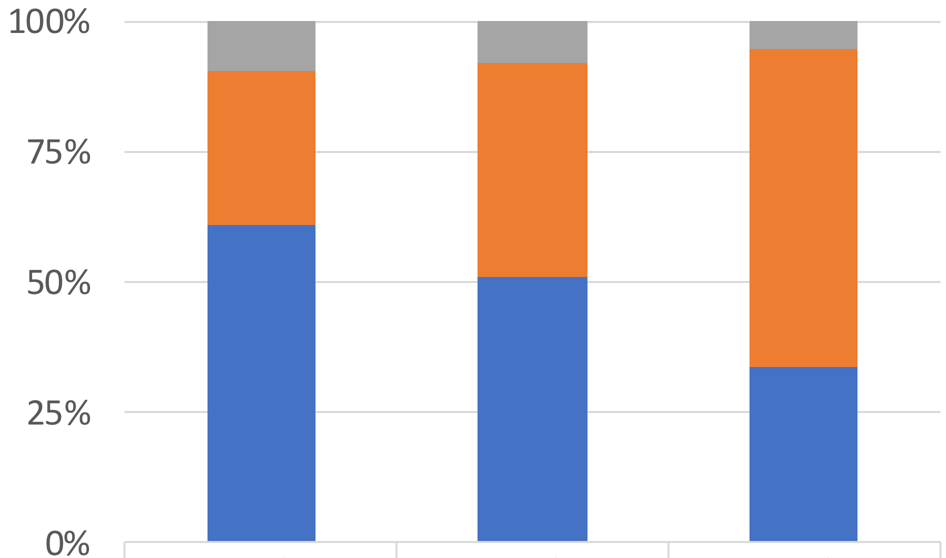


	sample1	sample2
■ taxa3	20	10
■ taxa2	20	50
■ taxa1	60	40

What is the true population composition?

Relative abundances carry no meaning for the absolute abundance of a specific component

The issue with compositional data



	Sample1	Sample2	Sample3
■ Taxa3	4	4	4
🟧 Taxa2	12	20	45
🟦 Taxa1	25	25	25

One change in abundance will drive abundance changes in another species violates assumptions of independence

The issue with compositional data

”Beware of attempts to interpret correlations between ratios whose numerators and denominators contain common parts.”

Pearson *1896*

Microbiome data is compositional

Analyzing relative data as if they were absolute can yield erroneous results for several common techniques

- Statistical models that assume independence between features are flawed because of the mutual dependency between components

- Distances between samples are misleading and erratically sensitive to the arbitrary inclusion or exclusion of components

- Components can appear definitively correlated even when they are statistically independent

This workshop

Why microbiome data is compositional data?

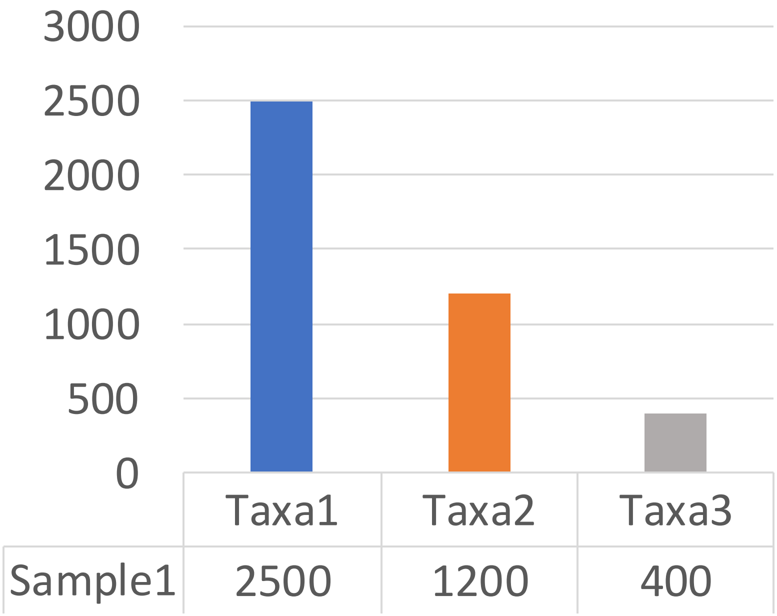
What is the issue with dealing with compositional data?

What are log-ratio transform approaches ?

How to use compositional data analysis methods with microbiome data?

CoDa

The starting point for any **CO**mpositional **DA**ta analyses is a ratio transformation of the data...



	Taxa1	Taxa2	Taxa3
Sample1	2500	1200	400

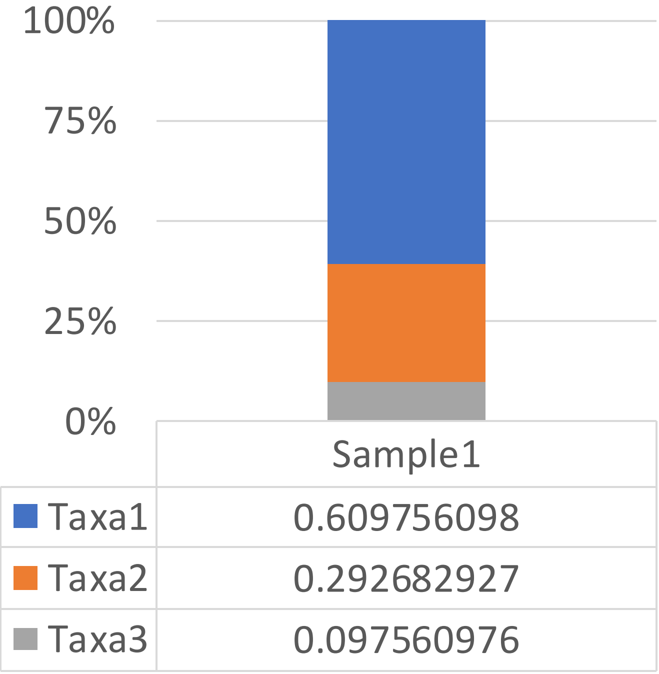
Ratio transformations

- capture the relationships between the features in the dataset

$$\frac{Taxa1}{Taxa2} = \frac{2500}{1200} = 2.083$$

CoDa

The starting point for any COmpositional DAta analyses is a ratio transformation of the data...



	Sample1
🟦 Taxa1	0.609756098
🟧 Taxa2	0.292682927
⬜ Taxa3	0.097560976

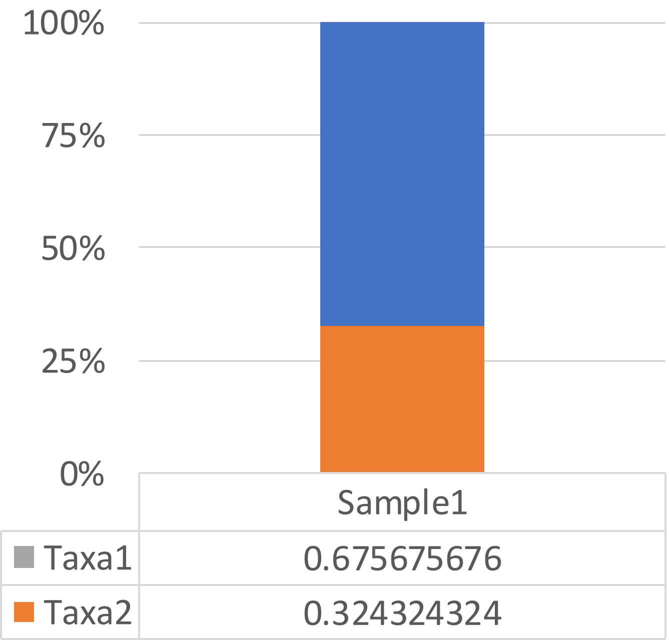
Ratio transformations

- capture the relationships between the features in the dataset
- ratios are the same whether the data are counts or proportions.

$$\frac{Taxa1}{Taxa2} = \frac{0.6098}{0.29268} = 2.083$$

CoDa

The starting point for any COmpositional DAta analyses is a ratio transformation of the data...



	Sample1
■ Taxa1	0.675675676
🟧 Taxa2	0.324324324

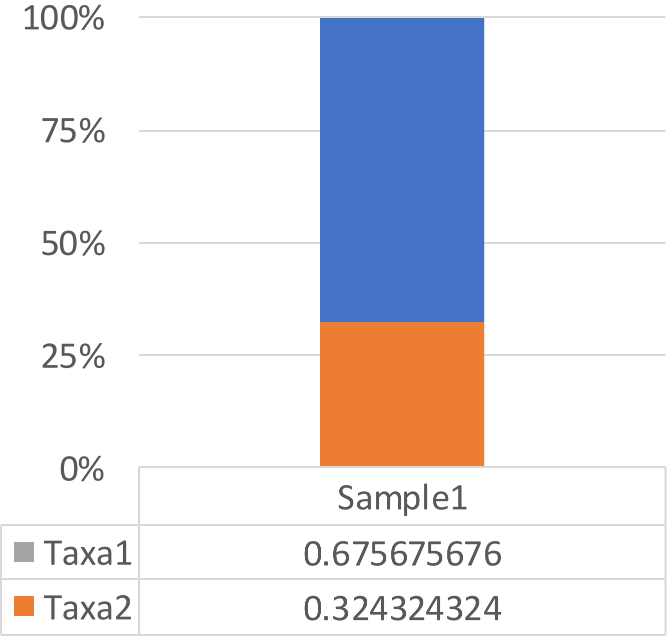
Ratio transformations

- capture the relationships between the features in the dataset
- ratios are the same whether the data are counts or proportions

$$\frac{Taxa1}{Taxa2} = \frac{0.6757}{0.3243} = 2.083$$

CoDa

The starting point for any COmpositional DAta analyses is a ratio transformation of the data...



	Sample1
■ Taxa1	0.675675676
🟧 Taxa2	0.324324324

Ratio transformations

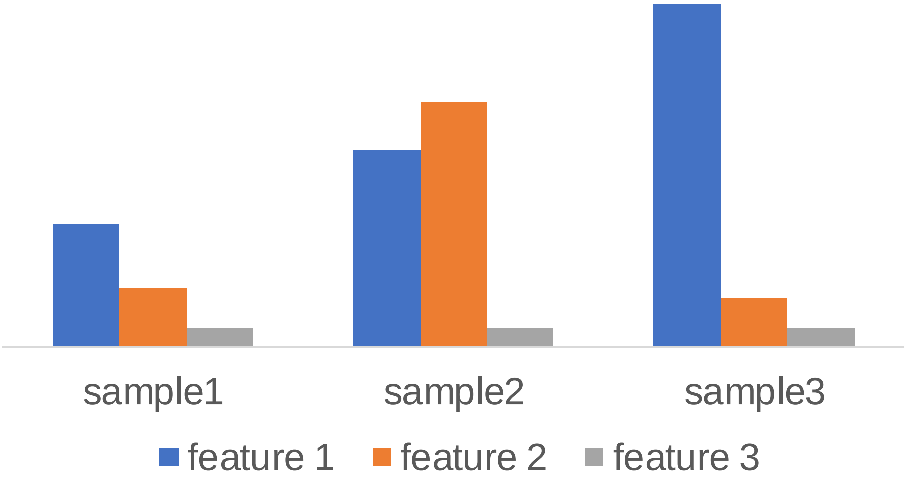
- capture the relationships between the features in the dataset
- ratios are the same whether the data are counts or proportions

logarithm of ratios (log-ratios)

- makes the data symmetric and linearly related
- places the data in a log-ratio coordinate space

Log-ratio transform

Additive Log Ratio

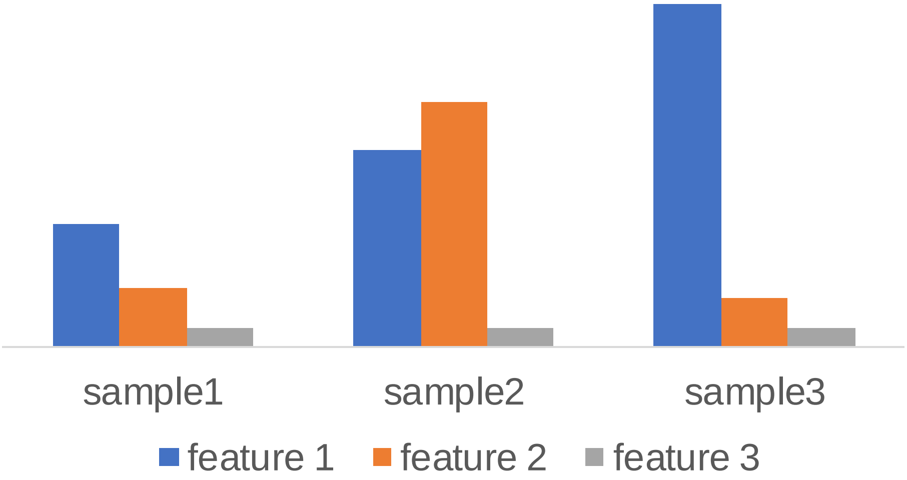


Feature 3 is a reference that can be "sacrificed" to transform the other counts.

alr (sample) = $[\ \log\frac{feature\ 1}{feature\ 3},\ \log\frac{feature\ 2}{feature\ 3}\ ]$

Log-ratio transform

Centered Log Ratio

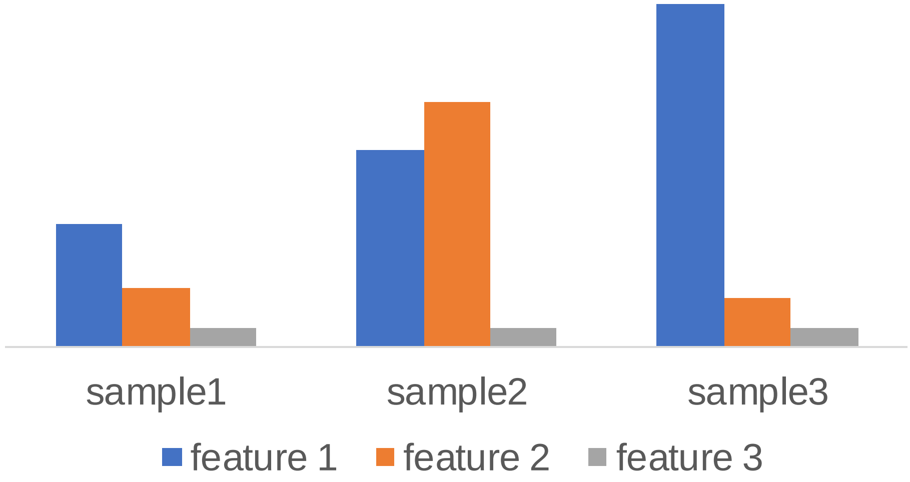


Instead of using a taxa as reference, we can use the **geometric mean** of the counts from the sample

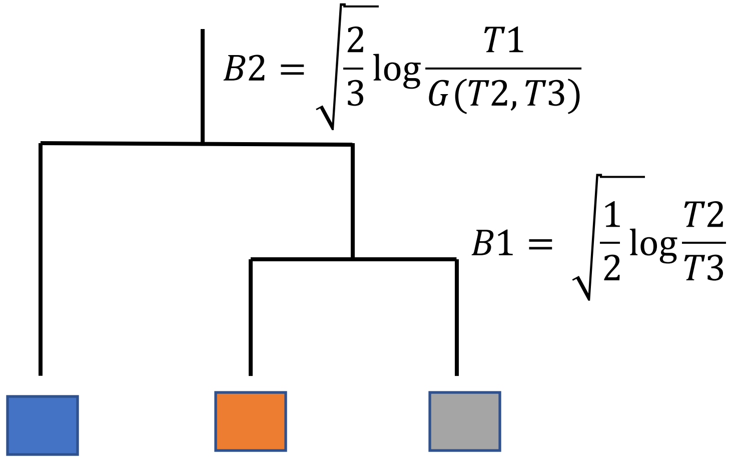
clr (sample) = $[\ \log \frac{feature\ 1}{G(sample)},\ \log \frac{feature\ 2}{G(sample)}\ ,\ \log \frac{feature\ 3}{G(sample)}\ ]$

Log-ratio transform

Isometric Log Ratio



We can also use ratio of sub-groups.
e.g: phylogenetic relationship can be leveraged (phILR)



$$B2 = \sqrt{\frac{2}{3}} \log \frac{T1}{G(T2, T3)}$$

$$B1 = \sqrt{\frac{1}{2}} \log \frac{T2}{T3}$$

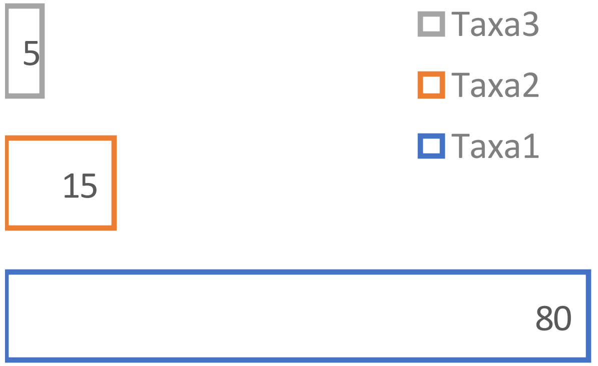
ilr (sample) = [B1, B2]

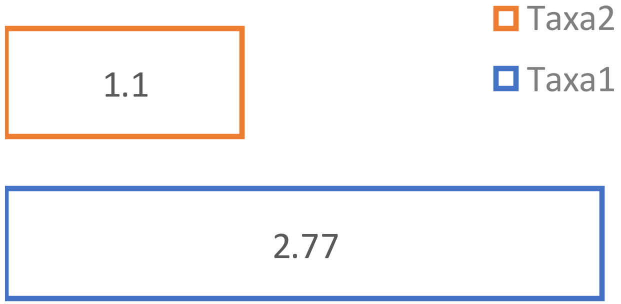
Log-ratio transform

Log-ratio transformations are NOT normalizations

- Normalization : recast data in absolute terms
- Transformation : must be interpreted with respect of a chosen reference

Log-ratio transform



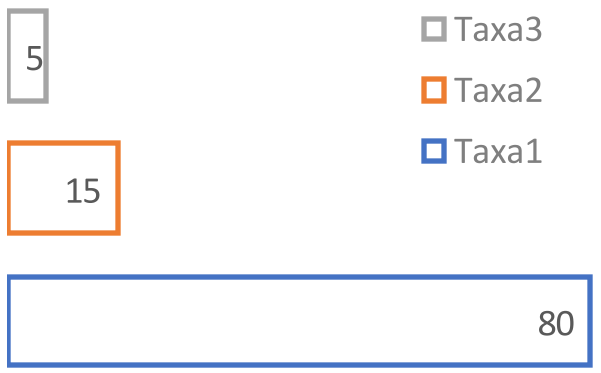


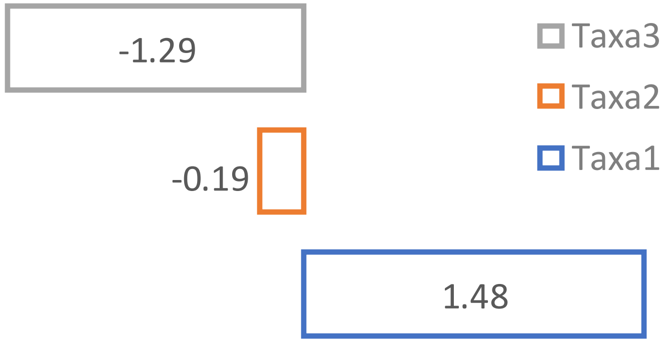
Additive Log Ratio : log ratio of the counts and a component of reference

Centered Log Ratio : log ratio of the counts and the sample geometric mean

Isometric Log Ratio : series of sequential log-ratios between subgroups of features

Log-ratio transform



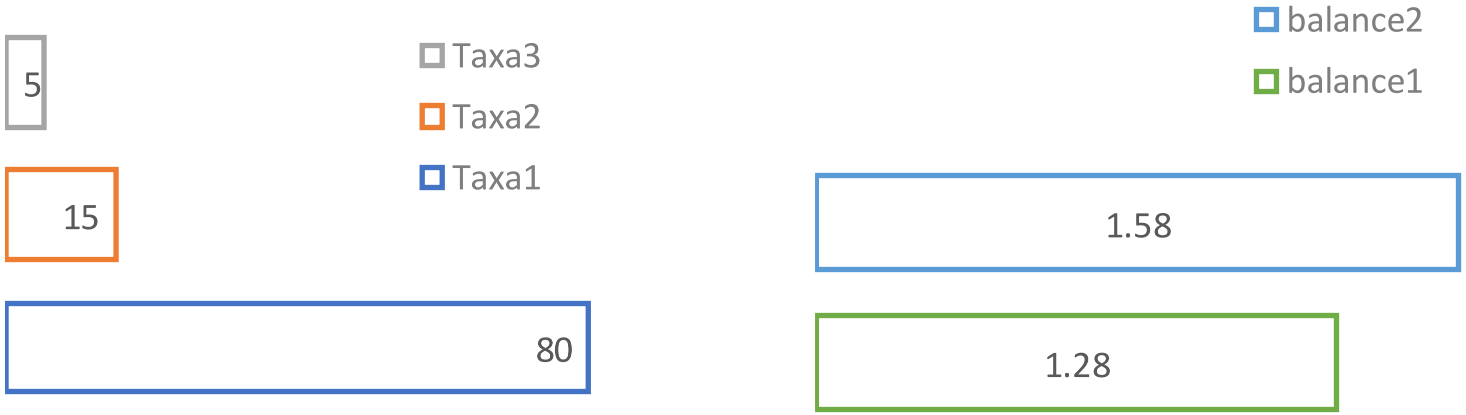


Additive Log Ratio : log ratio of the counts and a reference

Centered Log Ratio : log ratio of the counts and the sample geometric mean

Isometric Log Ratio : series of sequential log-ratios between subgroups of features

Log-ratio transform



Additive Log Ratio : log ratio of the counts and a reference

Centered Log Ratio : log ratio of the counts and the sample geometric mean

Isometric Log Ratio : series of sequential log-ratios between subgroups of features

Handling Zeros with log ratio transform

How to handle Zero counts before a log-transform?

	Sample 1	Sample 2
Taxa 1	156	762
Taxa 2	350	241
Taxa 3	67	0

Handling Zeros with log ratio transform

How to handle Zero counts before a log-transform?

	Sample 1	Sample 2
Taxa 1	156	762
Taxa 2	350	241
Taxa 3	67	0

Solution 1: Remove all component with 1 or more zero counts

Handling Zeros with log ratio transform

How to handle Zero counts before a log-transform?

	Sample 1	Sample 2
Taxa 1	156	762
Taxa 2	350	241
Taxa 3	67	1

Solution 1: Remove all component with 1 or more zero counts

Solution 2: Add a pseudo-count of 1 to all zero counts

Handling Zeros with log ratio transform

How to handle Zero counts before a log-transform?

	Sample 1	Sample 2
Taxa 1	156	762
Taxa 2	350	241
Taxa 3	67	P(x)

Solution 1: Remove all component with 1 or more zero counts

Solution 2: Add a pseudo-count of 1 to all zero counts

Solution 3: Dealing with 0 count values as point estimates or as a probability distribution

This workshop

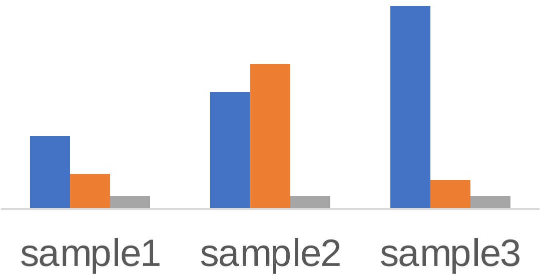
Why microbiome data is compositional data?

What is the issue with dealing with compositional data?

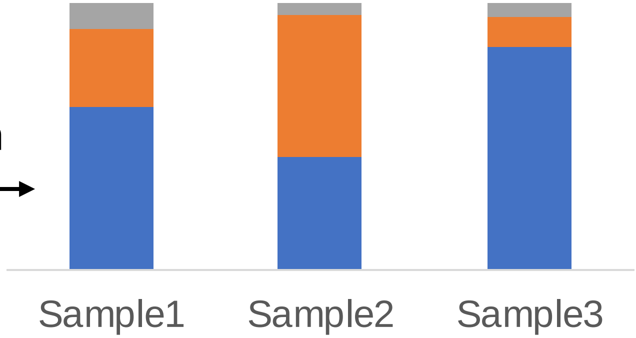
What are log-ratio transform approaches ?

How to use compositional data analysis methods with microbiome data?

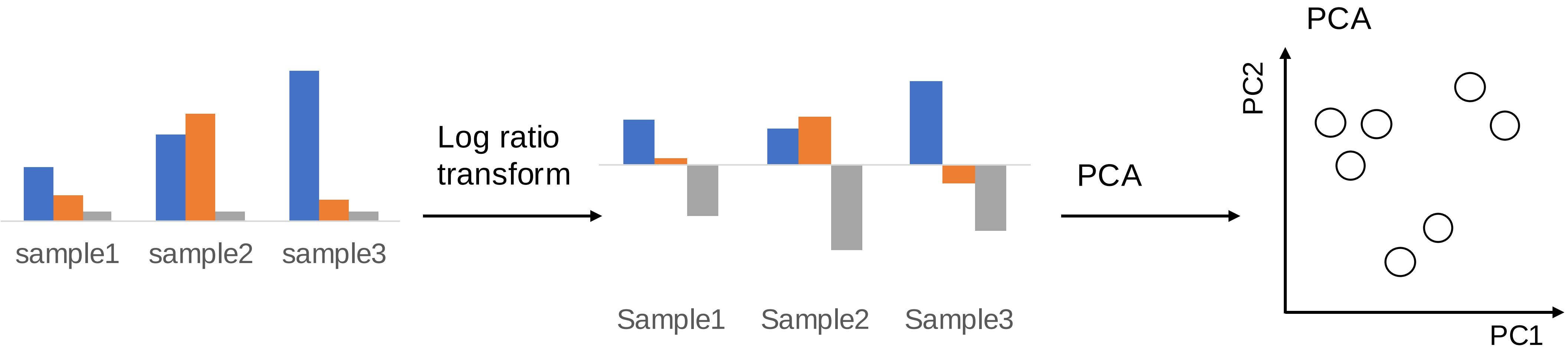
Classic vs CoDa



Count normalization



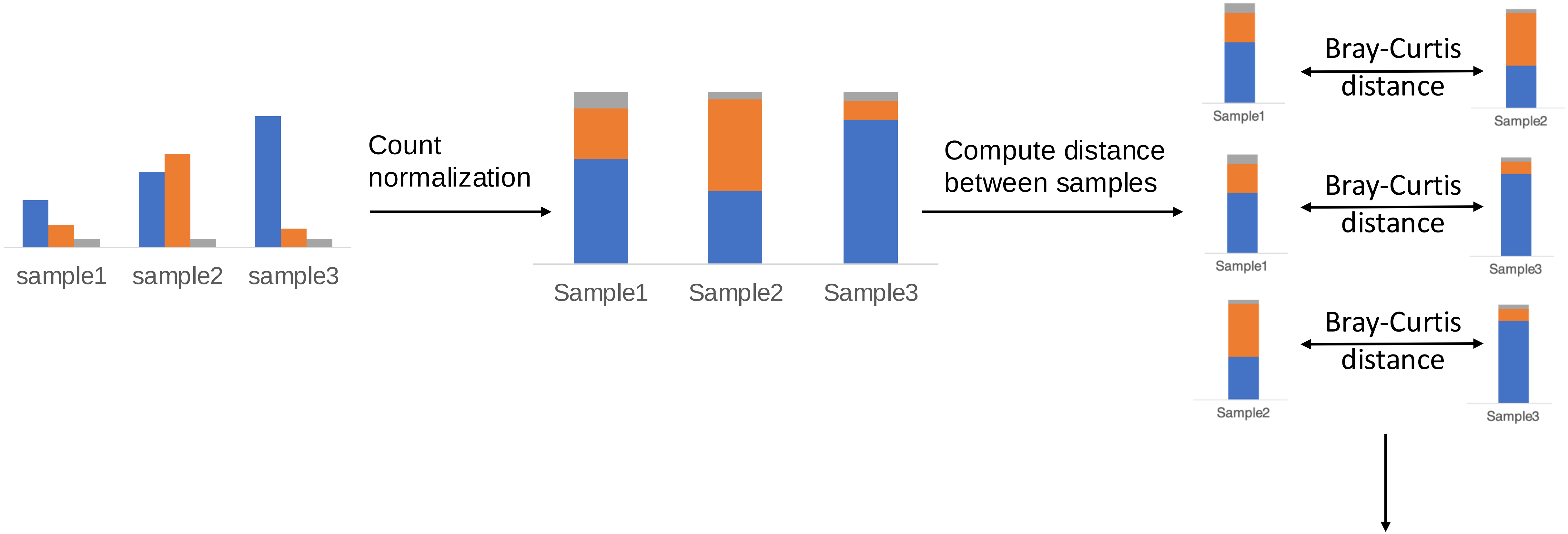
Classic vs CoDa



Log ratio transform

PCA

Classic vs CoDa



Count normalization

Compute distance between samples

Bray-Curtis distance

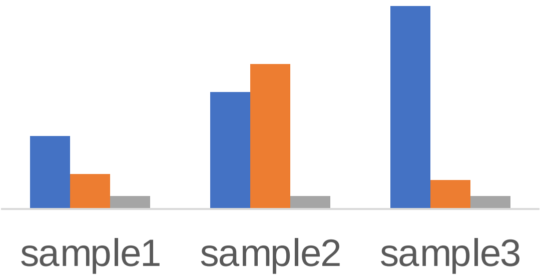
Bray-Curtis distance

Bray-Curtis distance

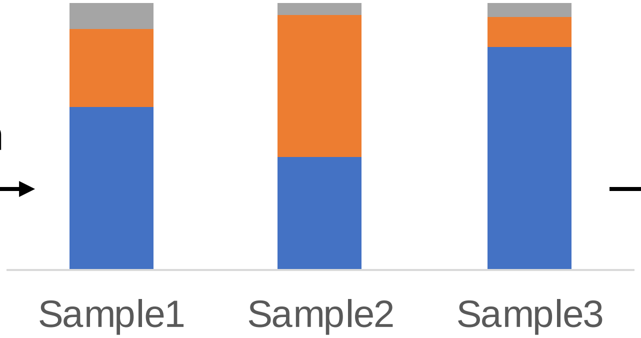
Distance matrix

	S1	S2	S3
S1	0		
S2	0.4	0	
S3	0.2	0.6	0

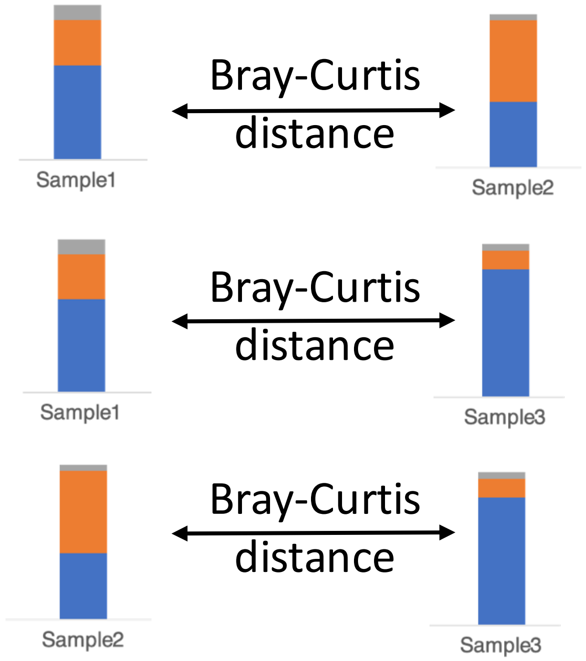
Classic vs CoDa



Count normalization



Compute distance between samples



Bray-Curtis distance

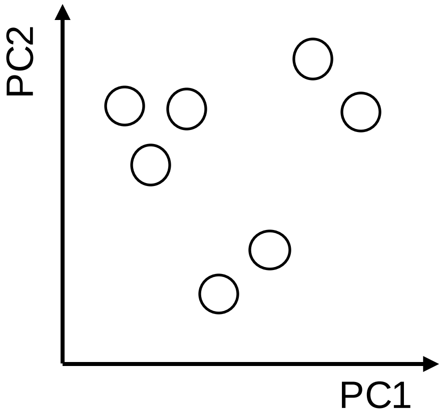
Bray-Curtis distance

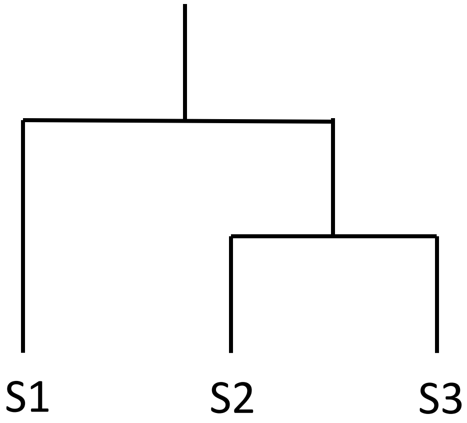
Bray-Curtis distance

Distance matrix

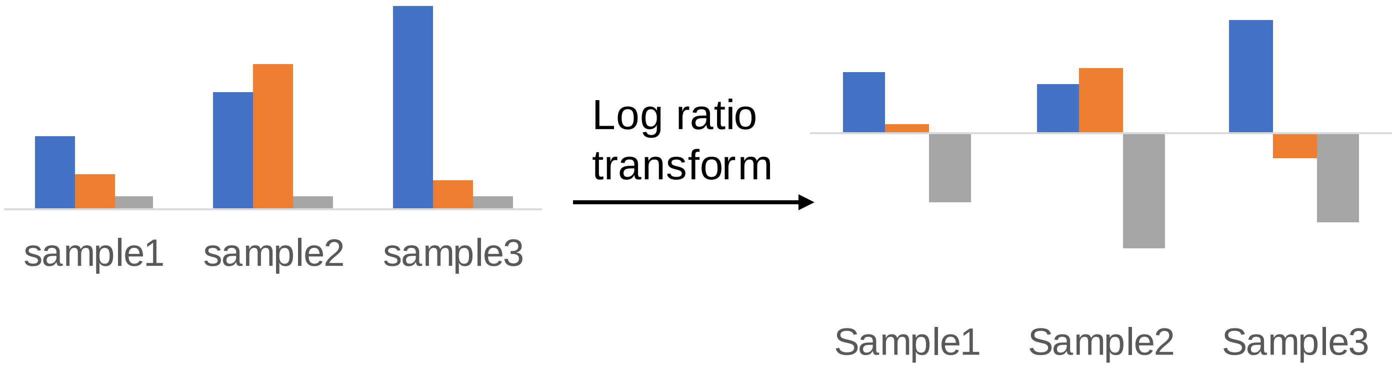
	S1	S2	S3
S1	0		
S2	0.4	0	
S3	0.2	0.6	0

PCoA





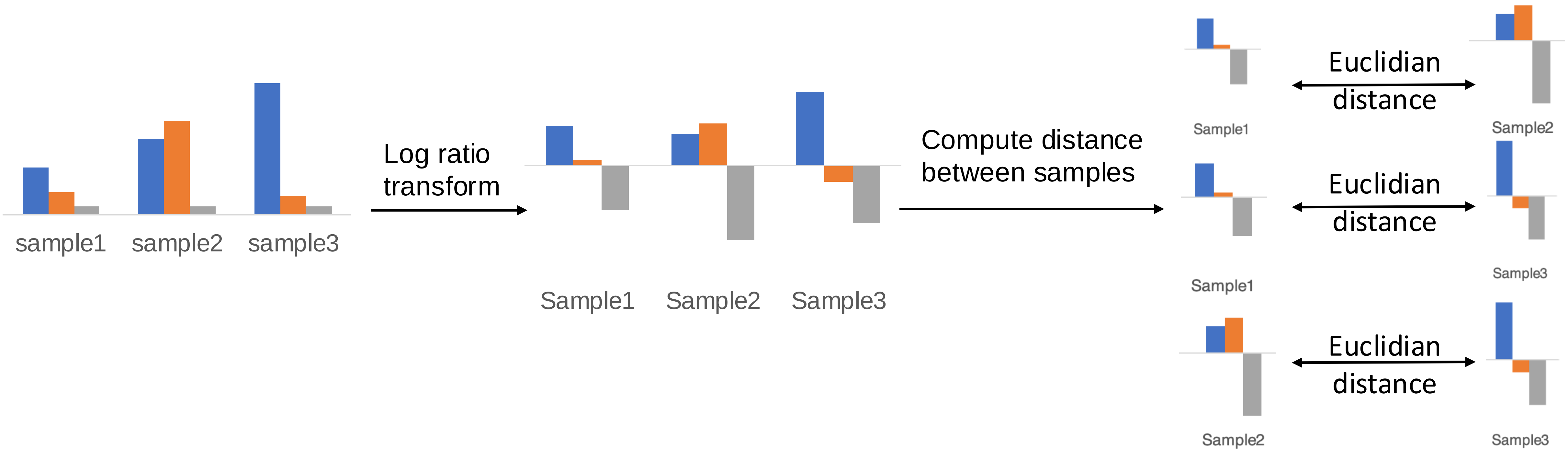
After a log-transfom, visualize samples similarities



Log ratio transform

What methods should we apply after a log ratio transform?

Classic vs CoDa



Log ratio transform

Compute distance between samples

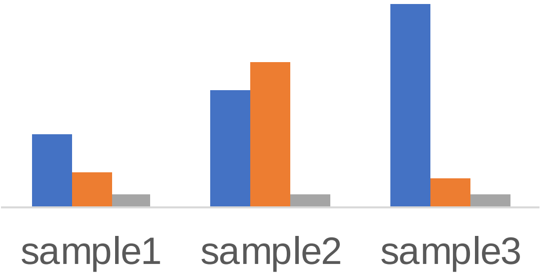
Euclidian distance

Euclidian distance

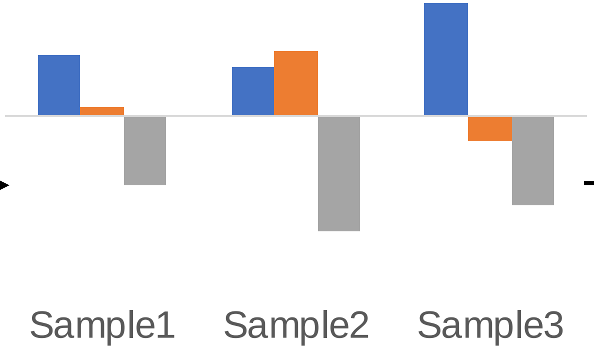
Euclidian distance

Euclidian
distance between log-transform
is called
“Aitchison distance”

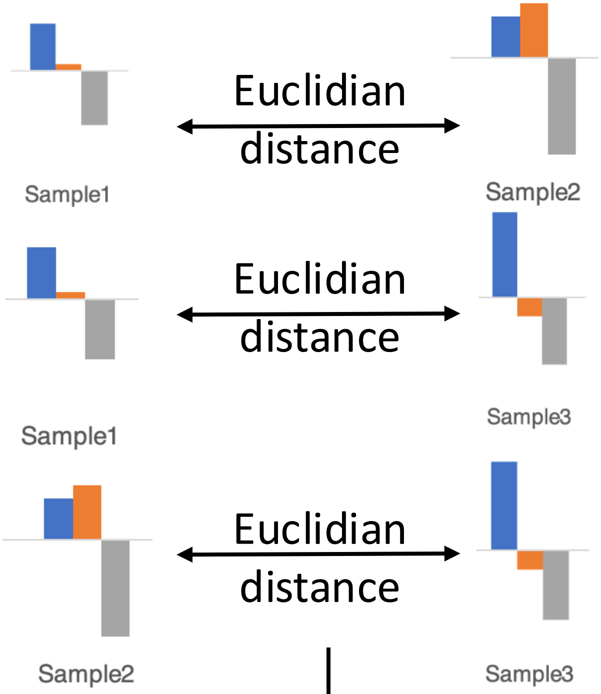
Classic vs CoDa



Log ratio transform

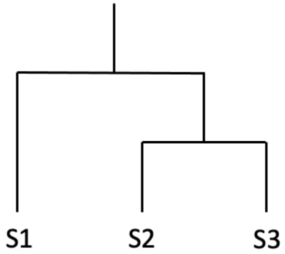


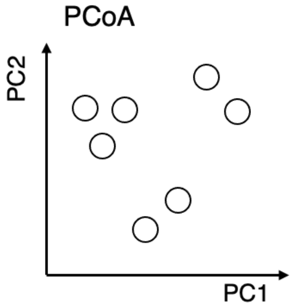
Compute distance between samples



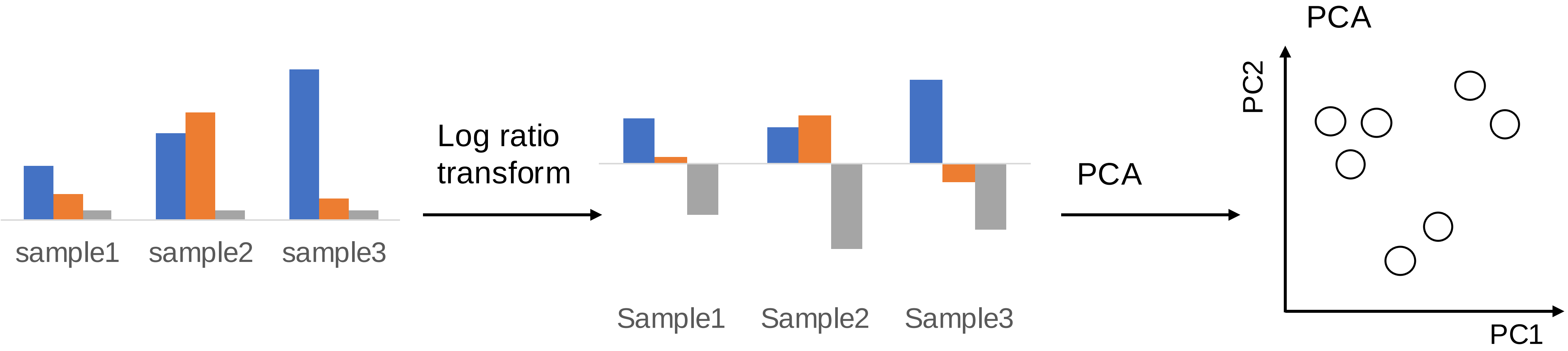
Distance matrix

	S1	S2	S3
S1	0		
S2	0.4	0	
S3	0.2	0.6	0





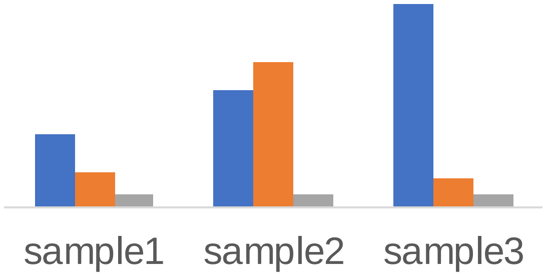
Classic vs CoDa



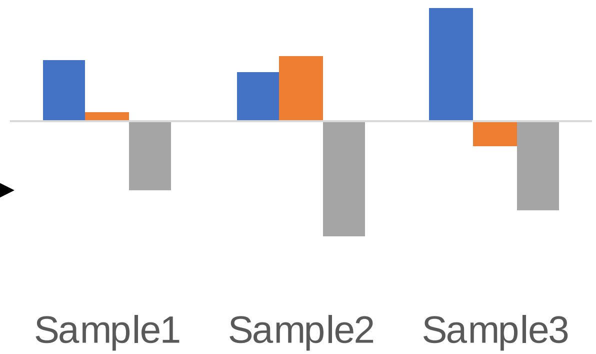
Log ratio transform

PCA

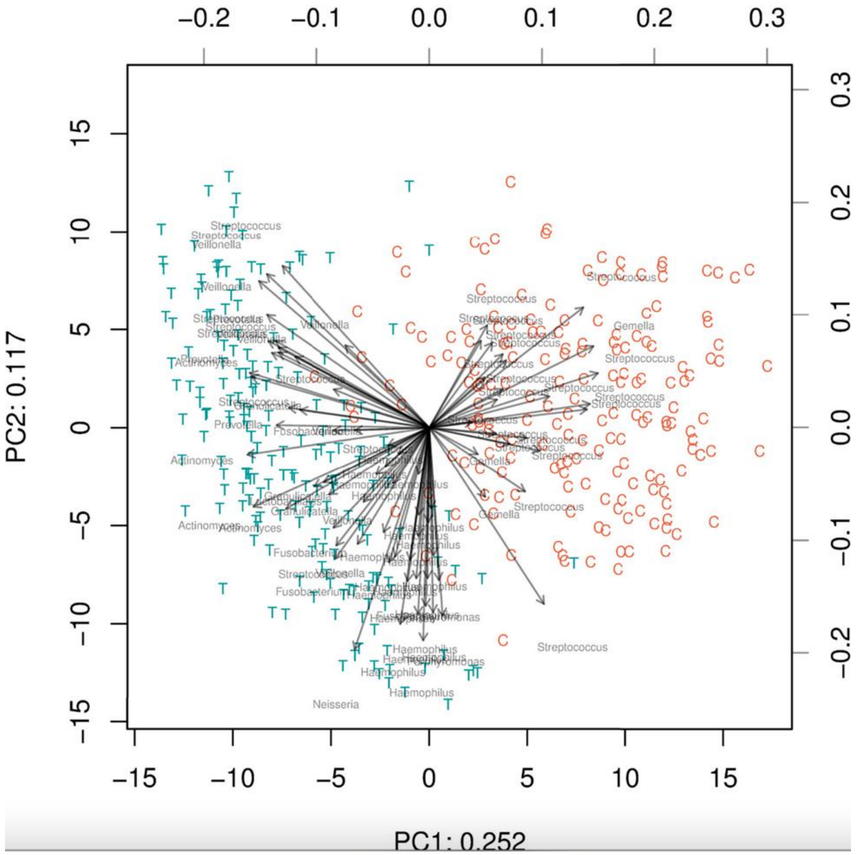
Classic vs CoDa



Log ratio transform



PCA



From Gloor *et al.* 2016

After a log-transform…

Operation	Standard method	Compositional method
Normalization & transformation	Rarefaction	ALR/CLR/ILR
Distance Ordination	BC, Unifrac, Jenson PCoA	Aitchison PCA
Multivariate comparison	PerMANOVA ANOSIM	PerMANOVA ANOSIM
Correlation	Pearson Spearman	SparCC SpiecEasi
Differential abundance	metagenomSeq DESeq	ALDEx2 ANCOM

Tools & useful links

Review articles

Review articles by Quinn et al. 2019 , Lin and Peddada 2020, and Luz Calle 2019 give you a good recap of the reasons and methods behind compositional data analysis for microbiome data

Tutorial

This excellent tutorial by Nicholas Ollberding introduce statistical analysis of microbiome data and in particular the use of CLR transform

Books

If during this talk you developed a true passion for CoDa approaches, I can recommend these amazing books to go deeper:

- A Concise Guide to Compositional Data Analysis by J. Aitchison
- Analyzing Compositional Data with R by Boogaart and Tolosana-Delgado (2013)
- Applied Compositional Data Analysis by Filzmoser, Hron, and Templ (2018)

Tools & useful links

Compositions R package

Book outlining how to use the compositions R package by Van den Boogaart and Tolosana-Delgado (2013) is particularly helpful, although none of the examples are drawn from the biological literature.

Tools to deal with 0 counts

Methods of dealing with 0 count values as point estimates using the zCompositions R package (Palarea-Albaladejo and Martín-Fernández, 2015), and as a probability distribution using ALDEx2 available on Bioconductor.

PhILR

This introduction paper will give you an overview of what the PhILR R package allows and what it can reveal in your microbiome dataset