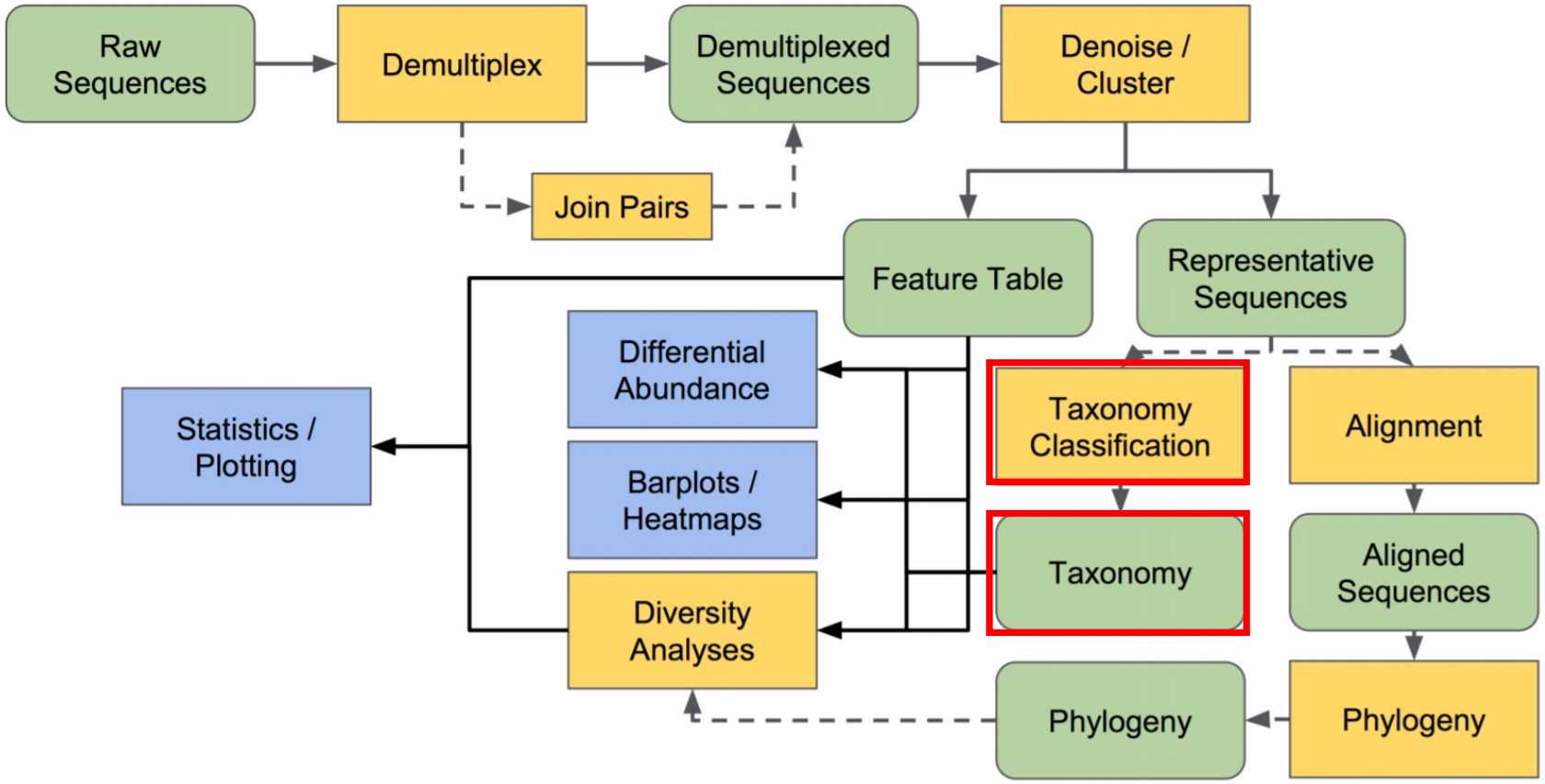
Module 7

Diversity Metrics

Module Outcomes

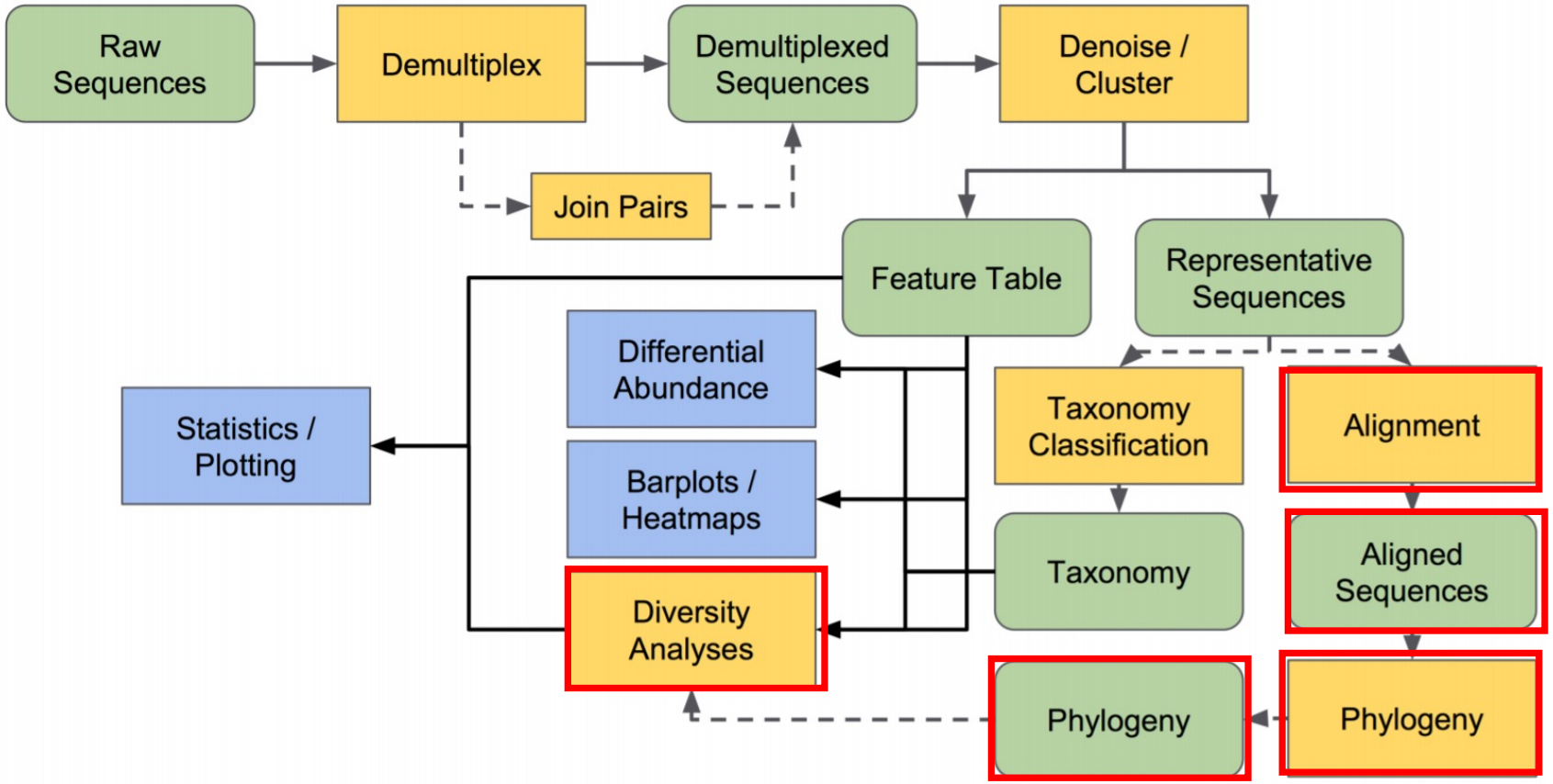
1. Define “microbial diversity” based on 3 key parameters
2. Rarefy your data before running your diversity metrics analysis
3. Interpret box plots and principle component analyses (PCA)

QIIME2 workflow



https://docs.qiime2.org/2018.8/tutorials/overview/#demultiplexing

QIIME2 workflow

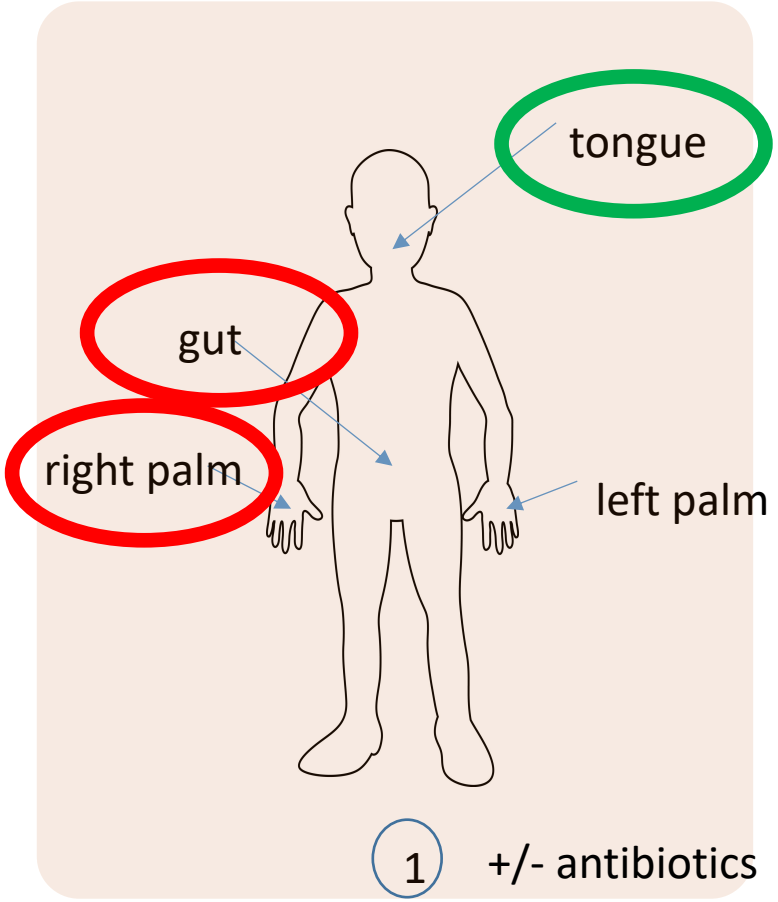


Green – input data, Orange – processes, Blue – statistical testing

Alpha versus Beta Diversity

- Alpha diversity – variation of microbes **<u>within</u>** a single environment
  - e.g. how many different microbes on the tongue?

- Beta diversity – comparison of microbial variation **<u>between</u>** environments
  - e.g. how different are the microbial communities sampled from the right palm vs the gut?

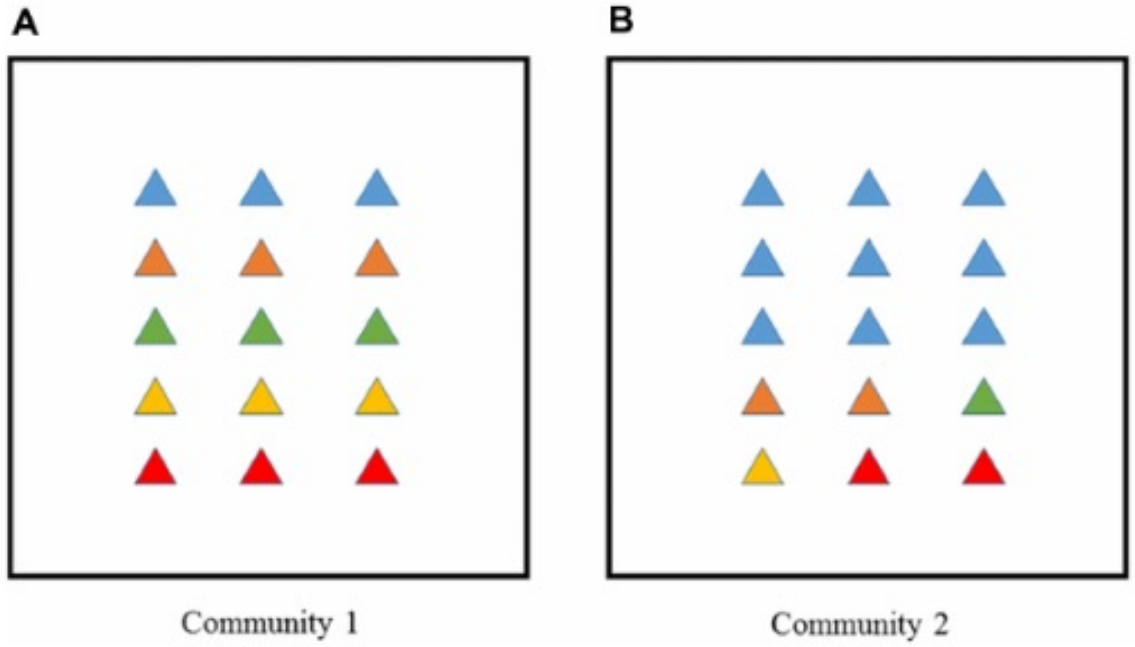


Diversity terms:

1. **Richness** – absolute number of organisms (specifically ASVs)
   - What is there?

2. **Abundance** – how much of an organism (amounts of that ASV) there is
   - How much of each?
   - **Evenness** – how uniform the species present are
     - Are the species evenly distributed?

Richness



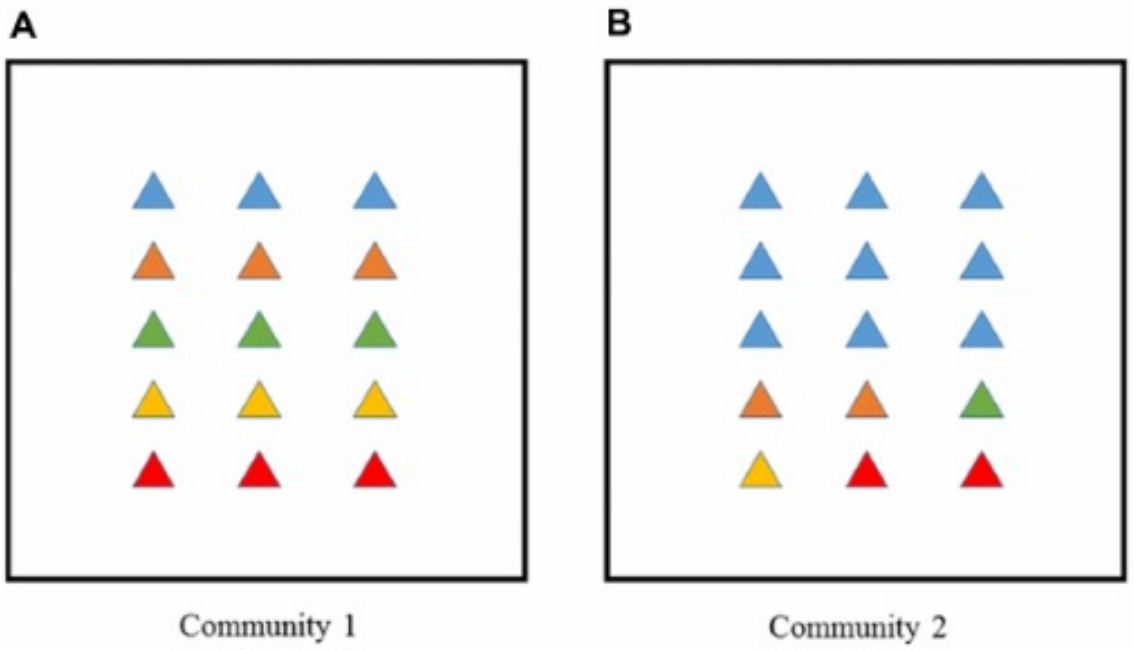
5 species

5 species

Community 1 and 2 are <u>equally</u> as diverse.

https://www.researchgate.net/publication/320452742_Deciphering_diversity_indices_for_better_understanding_of_the_microbial_communities

Abundance and Evenness



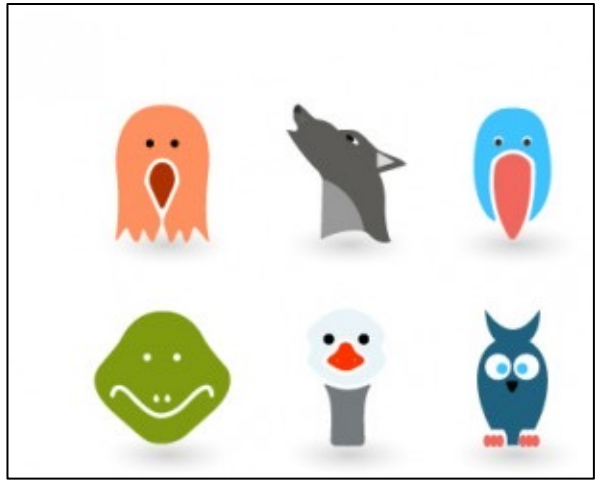
Community 1 is more diverse than Community 2

https://www.researchgate.net/publication/320452742_Deciphering_diversity_indices_for_better_understanding_of_the_microbial_communities

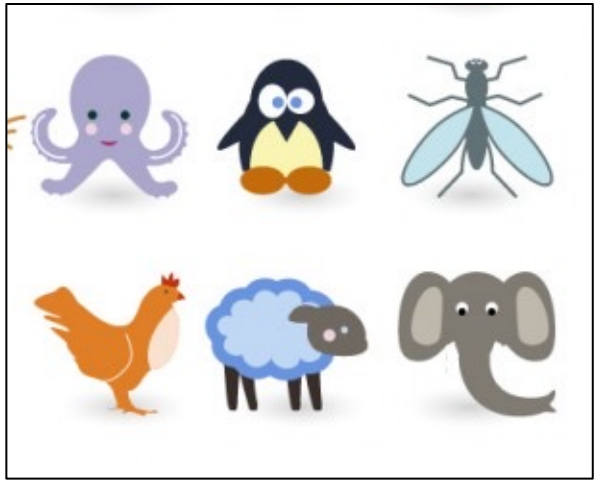
When “species” richness and evenness increase, diversity does too.

3. Phylogenetic Relatedness

- How closely or distally related the species are to each other
- The more distantly related species are, the more diverse



Community 1



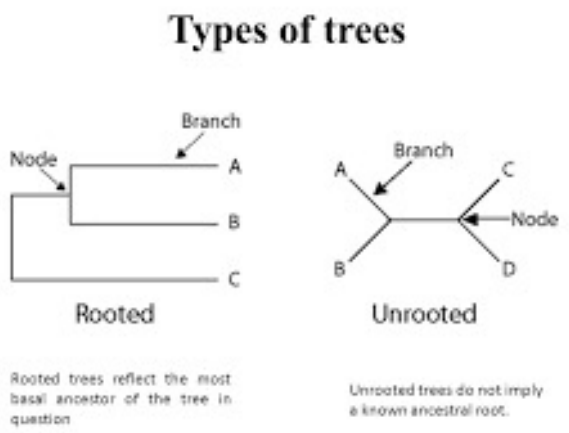
Community 2

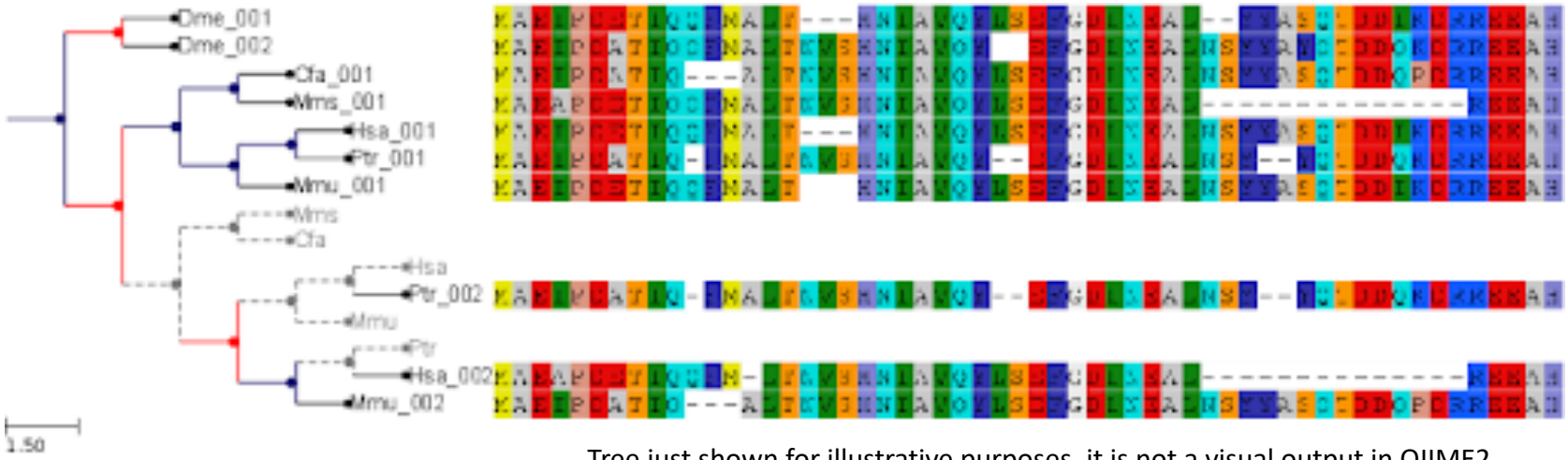
Community 1 is less diverse than 2

Generate a tree for phylogenetic diversity analyses

```
qiime phylogeny align-to-tree-mafft-fasttree \
--i-sequences rep-seqs.qza \
--o-alignment aligned-rep-seqs.qza \
--o-masked-alignment masked-aligned-rep-seqs.qza \
--o-tree unrooted-tree.qza \
--o-rooted-tree rooted-tree.qza
```

You do NOT actually visualize the tree





Tree just shown for illustrative purposes, it is not a visual output in QIIME2

QIIME2 - Alpha and beta diversity analysis summary

- Alpha diversity (within samples)
  - ✓ Shannon's diversity index (measure of community richness)
  - ✓ Observed Features (measure of community richness)
  - ✓ Faith's Phylogenetic Diversity (measure of community richness that incorporates phylogenetic relationships between the features)
  - ✓ Evenness (or Pielou's Evenness; measure of community evenness)

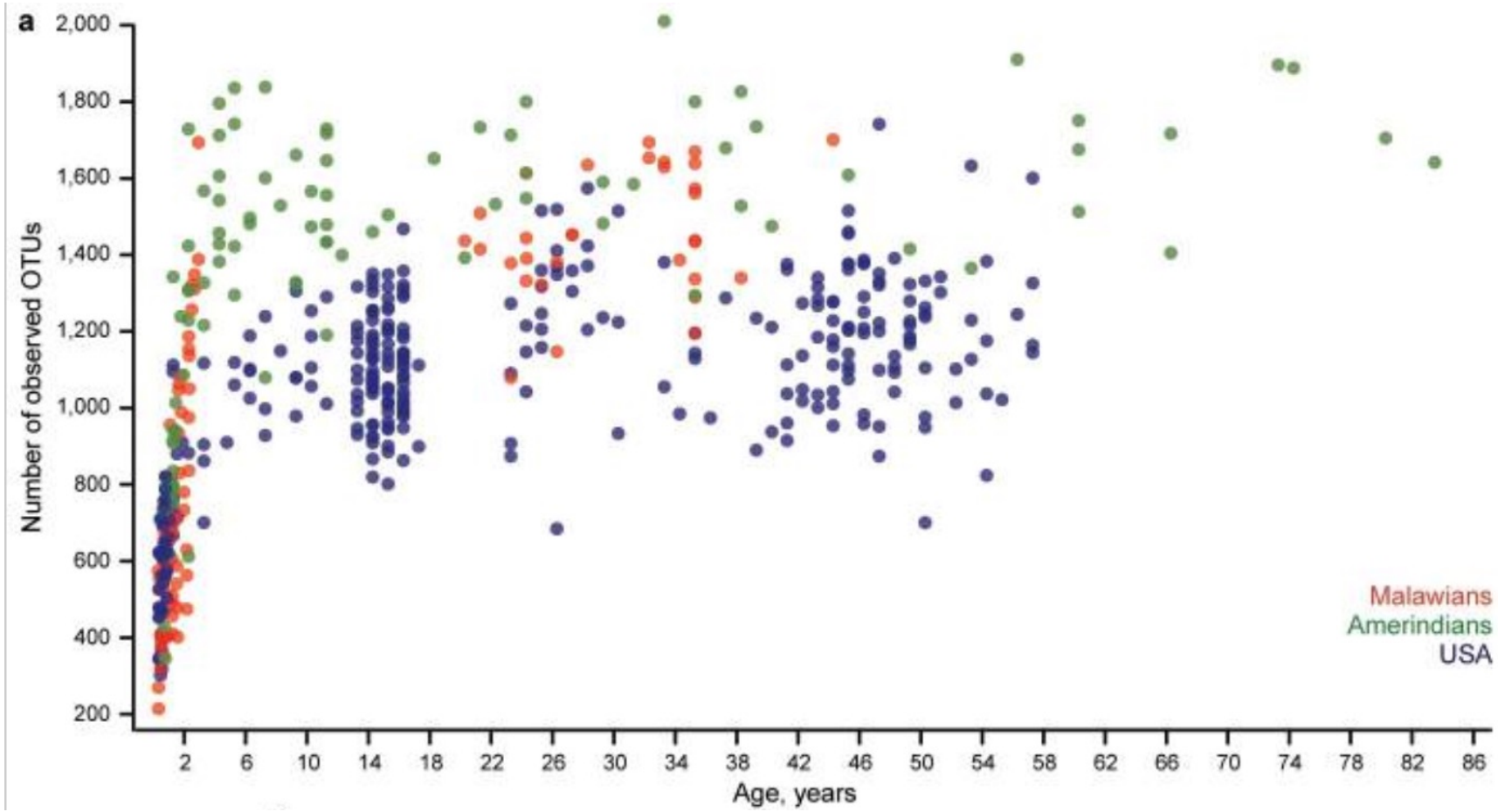
- Beta diversity (between samples)
  - ✓ Jaccard distance (measure of community dissimilarity)
  - ✓ Bray-Curtis distance (measure of community dissimilarity)
  - ✓ Unweighted UniFrac distance (measure of community dissimilarity that incorporates phylogenetic relationships between the features)
  - ✓ Weighted UniFrac distance (measure of community dissimilarity that incorporates phylogenetic relationships between the features)

Alpha Diversity

Alpha diversity

Considers phylogenetic distances	Considers abundance: No	Considers abundance: Yes
No	**Observed Features** (measure of community richness)	**Shannon Diversity** (measure of community richness and abundance)
Yes	**Faith's Phylogenetic Diversity** (measure of community richness that incorporates phylogenetic relationships between the features)	**n/a**

Alpha Diversity – Observed Features (richness)



(Note: remember that ASVs have now replaced OTUs)

Yatsunenko, Nature 2012

Alpha diversity

Considers phylogenetic distances	Considers abundance: No	Considers abundance: Yes
No	**Observed Features** (measure of community richness)	**Shannon Diversity** (measure of community richness and abundance)
Yes	**Faith's Phylogenetic Diversity** (measure of community richness that incorporates phylogenetic relationships between the features)	**n/a**

Observed Feature counts fail to capture phylogenetics (genetic relatedness)

- Sample A
  - *Pseudomonas aeruginosa*
  - *Pseudomonas argentinensis*
  - *Pseudomonas flavescens*
- Sample B
  - *Pseudomonas aeruginosa*
  - *Pseudomonas argentinensis*
  - *E. coli*
- Sample C
  - *Pseudomonas aeruginosa*
  - *Giardia lamblia*
  - *Methanobrevibacter smithii*

Observed Feature Counts
Sample A = 3
Sample B = 3
Sample C = 3

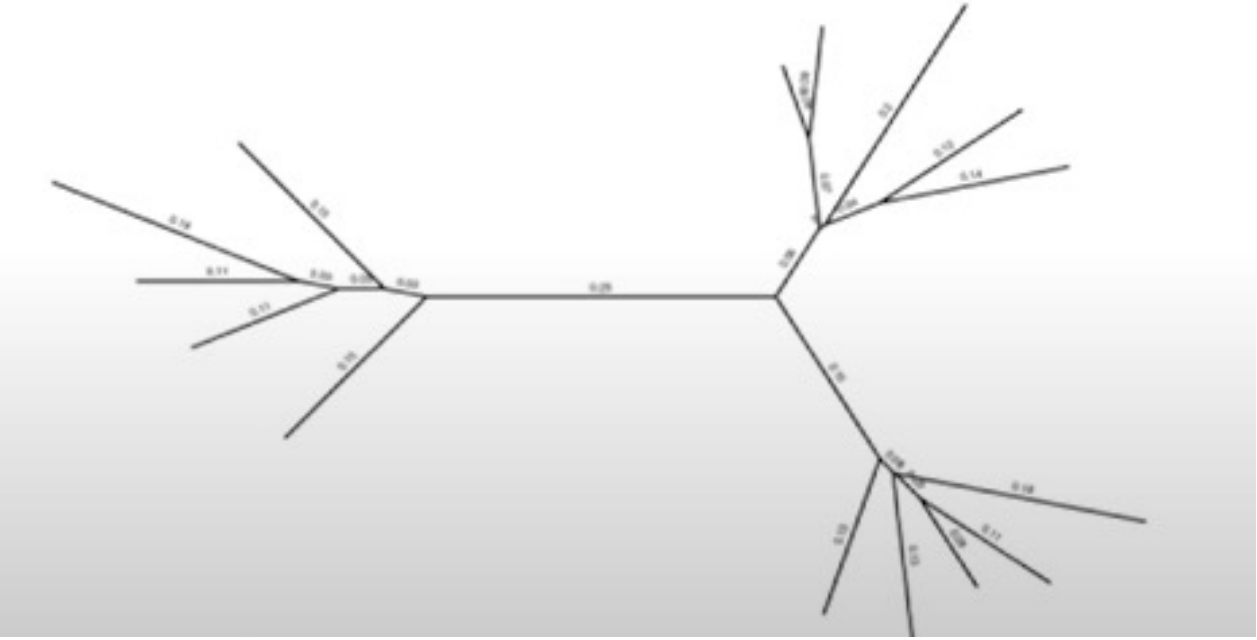
Conclusion
Sample A, B, and C are equally diverse.

Slide adapted from Rob Knight Lab

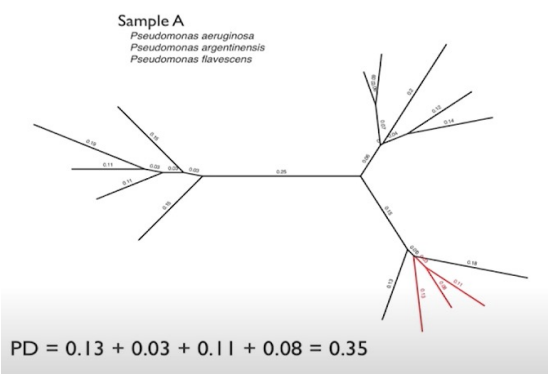
- Sample A
  - *Pseudomonas aeruginosa*
  - *Pseudomonas argentinensis*
  - *Pseudomonas flavescens*

- Sample B
  - *Pseudomonas aeruginosa*
  - *Pseudomonas argentinensis*
  - *E. coli*

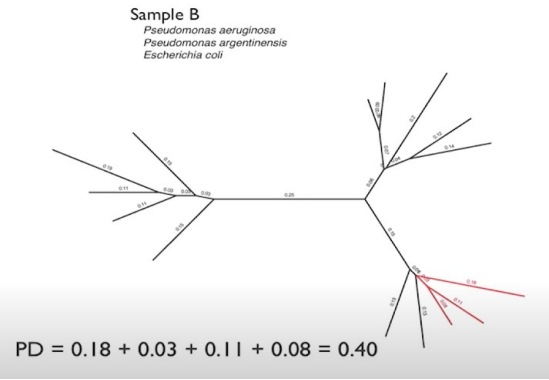
- Sample C
  - *Pseudomonas aeruginosa*
  - *Giardia lamblia*
  - *Methanobrevibacter smithii*



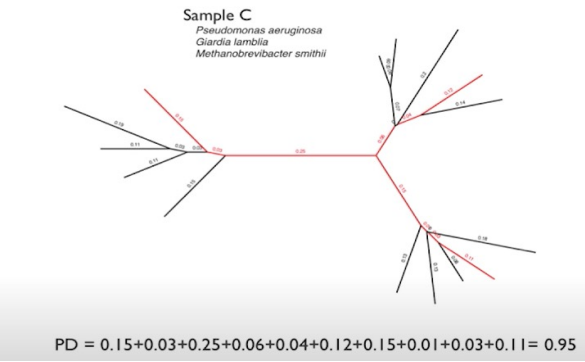
Illustrating phylogenetic differences between samples



Sample A



Sample B



Sample C

Add up the total branch length – gives a measure of phylogenetic diversity

Alpha diversity

Factors in phylogenetic distances	Factors in abundance: No	Factors in abundance: Yes
No	**Observed Features** #1. 4 different bugs #2. 4 different bugs	**Shannon Diversity** #1. Lower diversity, mostly bug A #2. Higher diversity, 4 bugs similar abundances
Yes	**Faith's Phylogenetic Diversity** #1. 4 different bugs #2. 4 different bugs All equally related	**n/a**

Pielou's Evenness: derived from measures in Shannon diversity to determine evenness

Consider this example:

Sample #1: **99% bug A + 1% (B + C + D)**

Sample #2: **25% bug A + 25% B + 25% C + 25% D**

Beta Diversity

Beta diversity

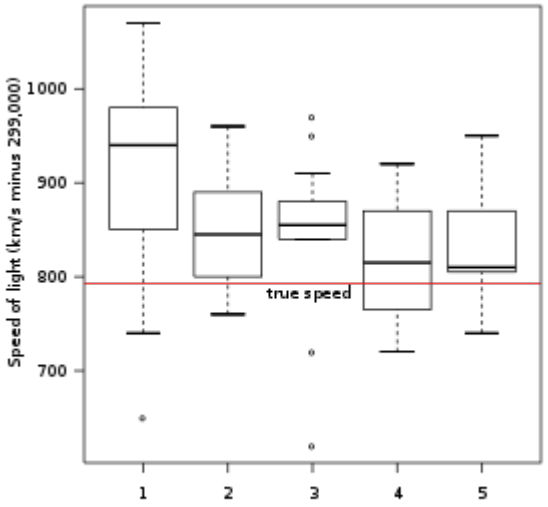
Cares about phylogenetic distance? \ Cares about abundance?	No	Yes
No	**Jaccard**	**Bray Curtis**
Yes	**Unweighted Unifrac**	**Weighted Unifrac**

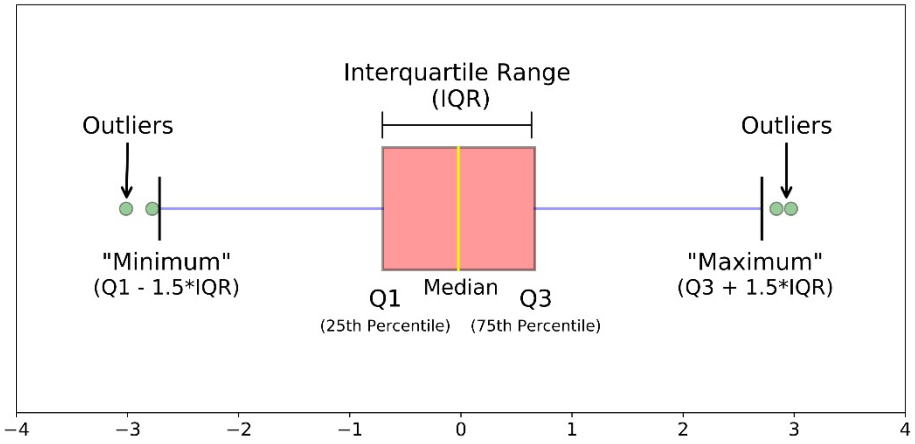
Beta Diversity Metrics

- Jaccard: only cares about presence/absence
  - [9 giraffes + 1 elephant] is the same as [5 giraffes + 5 elephants]
- Bray-Curtis: takes abundance (evenness) into account
  - [9 giraffes + 1 elephant] is different from, and LESS diverse than, [5 giraffes + 5 elephants]
- Unweighted Unifrac: cares about presence/absence, and also relatedness (phylogenic distance)
  - [9 giraffes + 1 elephant] is the same as [5 giraffes + 5 elephants]. However, [5 giraffes + 5 spiders] is different from, and MORE diverse than [5 giraffes + 5 elephants]
- Weighted Unifrac: cares about abundance AND relatedness (phylogenic distance)
  - [9 giraffes + 1 elephant] < [5 giraffes + 5 elephants] < [5 giraffes + 5 spiders]

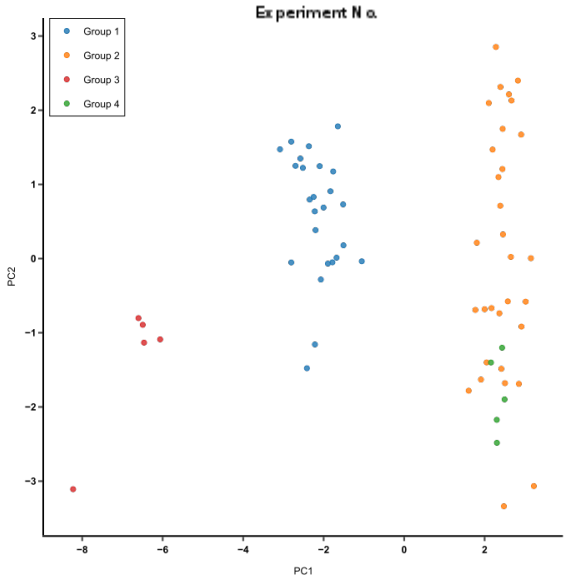
Beta and alpha diversity can be visualized as

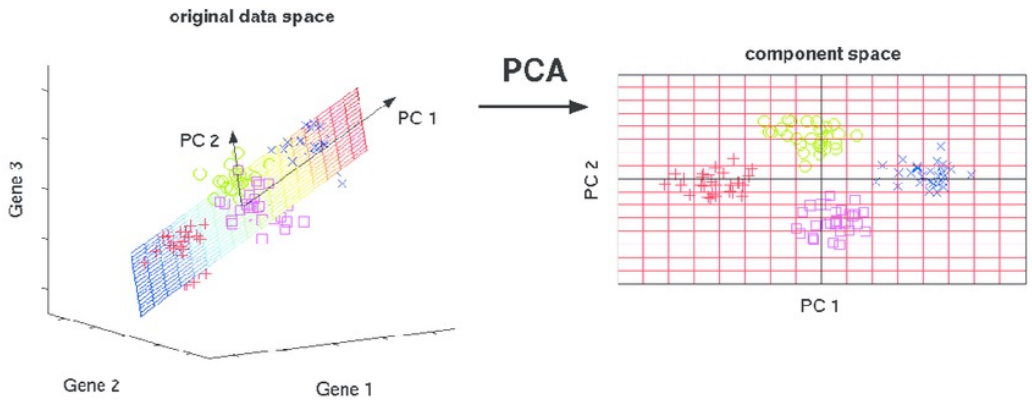
- Box plot





- PCA (PCoA) plot: principal coordinate analysis

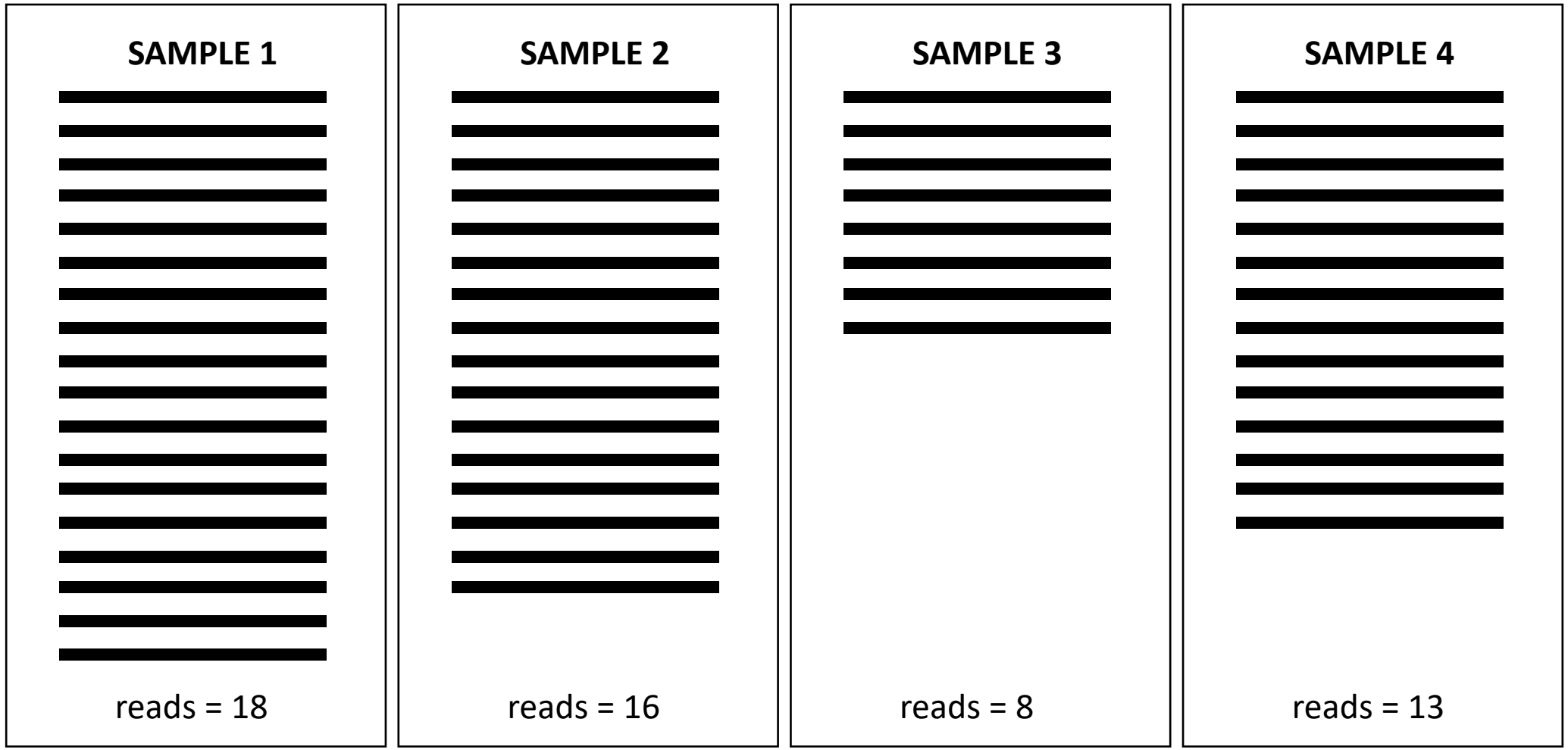




Rarefaction

Correcting for sequencing depth

Uneven sequencing depths

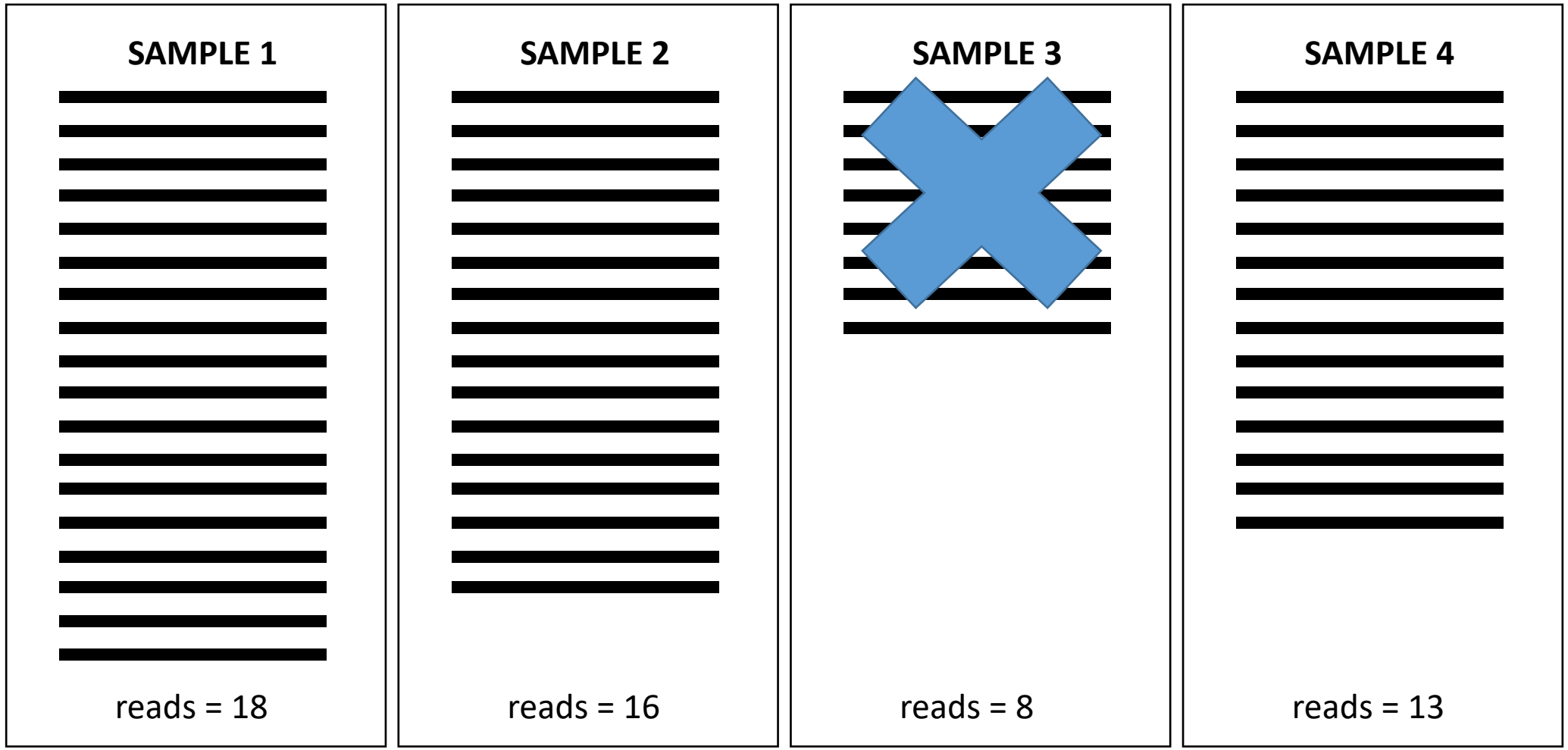


Can’t really compare diversity

Distinguishing Sequencing and Sampling Depth

- **Sequencing depth**: how many reads are in a sample (ie. how deeply did it get sequenced)
- **Sampling depth**: what we want to set as our rarefaction parameter (ie. how many reads we want to sample to normalize all the samples)

Set Sampling Depth of 10

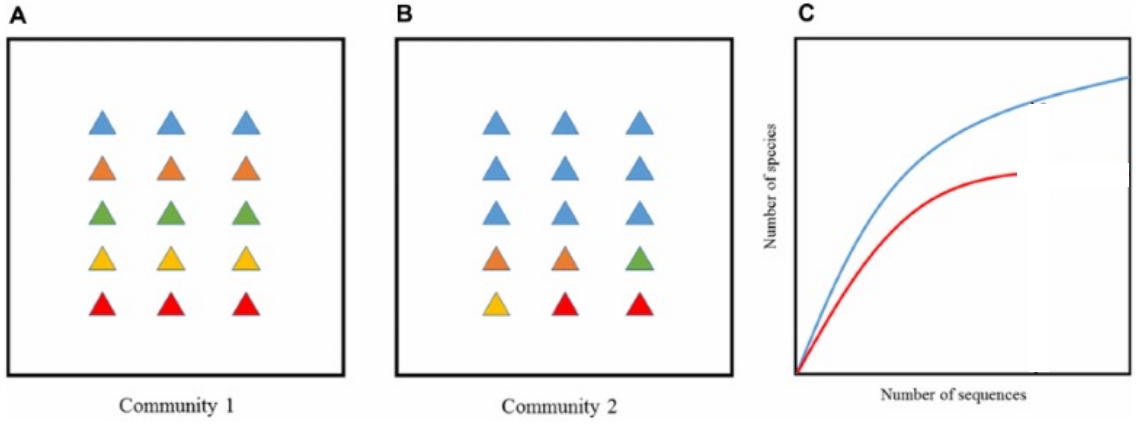


Set Sampling Depth of 10

Problem with sample representation

- A problem emerges from comparing samples of different library sizes.
- Richness measurements are affected by sequencing depth (library size).
- Therefore, it is difficult to determine immediately which community has higher species richness when we compare samples of different sizes.
- One way to overcome this problem is to standardize all samples from different communities to the same sampling depth

Community 1 and 2 show equal richness.



Because of uneven sampling depth, sample A (blue) would show higher richness

Rarefaction

- Rarefaction curves measure ASVs observed with a given depth of sequencing, and are used to compare observed richness among communities that have been unequally sampled.
- Rarefaction is a statistical technique to approximate the number of ASVs expected in a random sample of individuals taken from a sample collection.
- Rarefaction permits direct comparisons of samples of different sizes of their sample sizes.

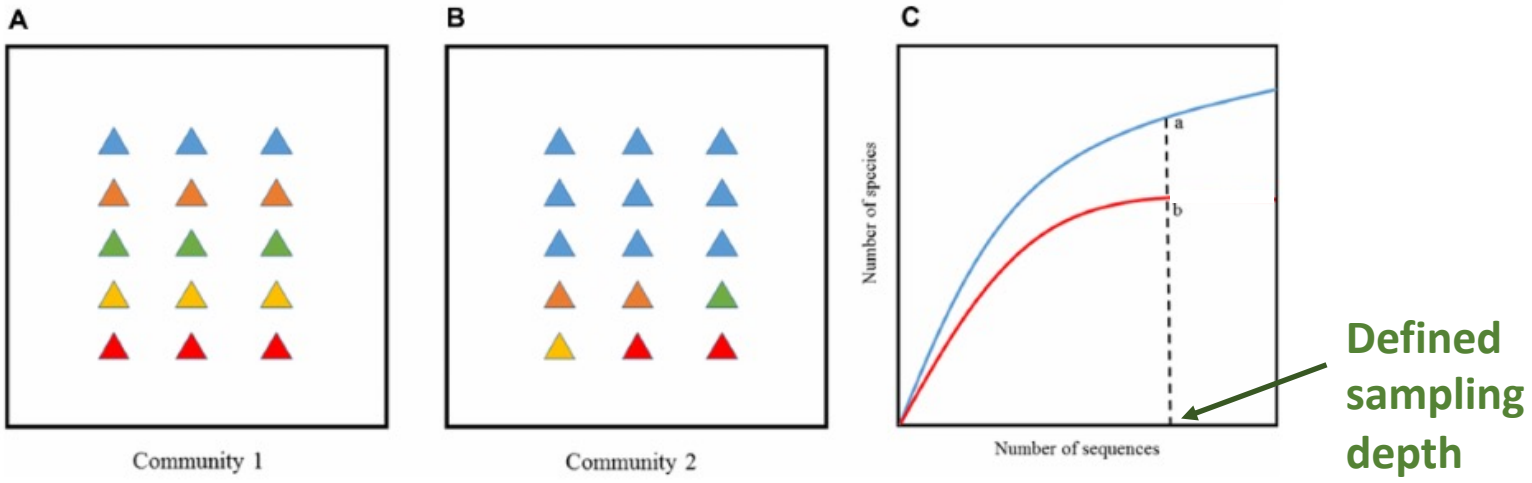
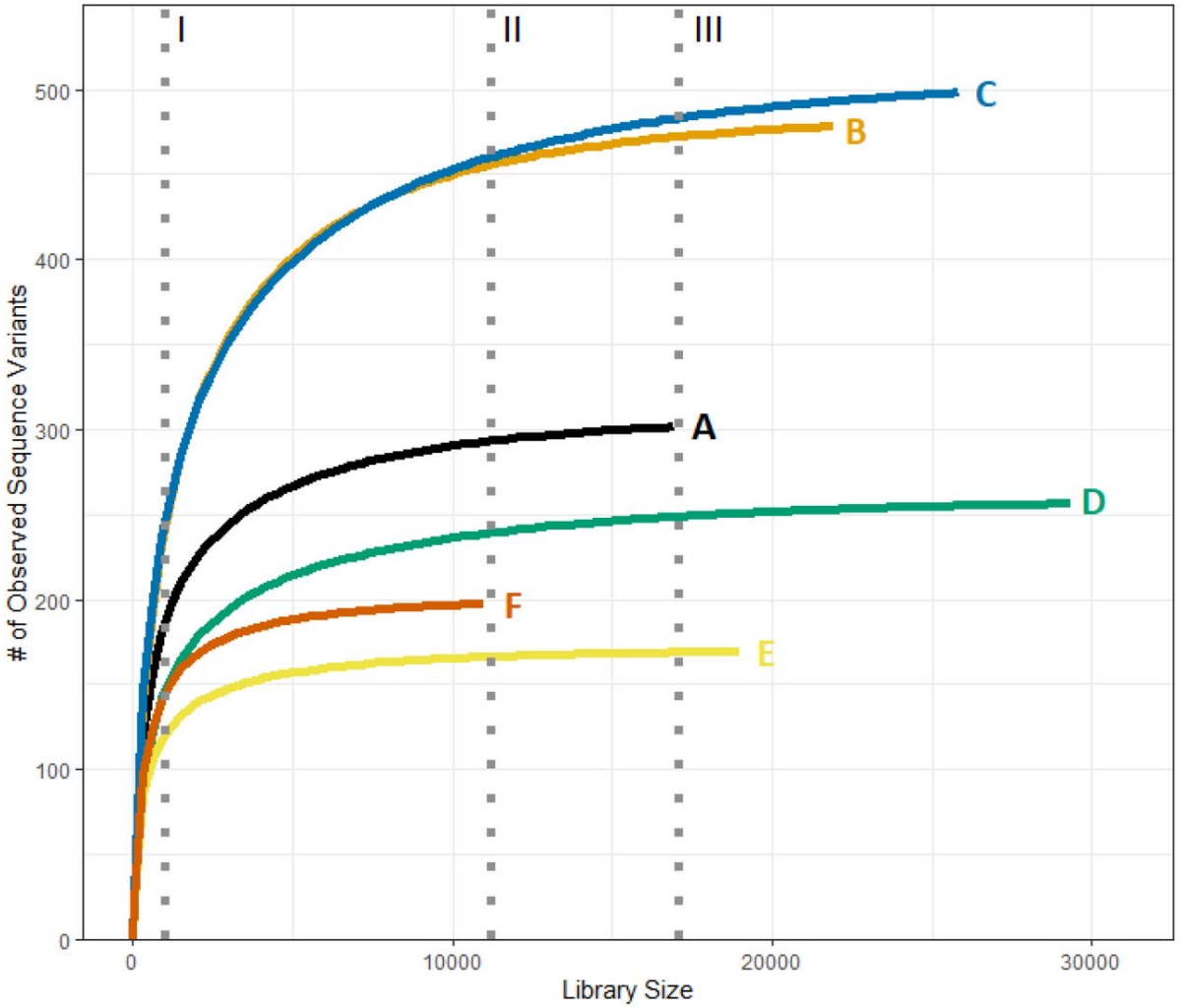


Fig. 1. Species richness, evenness, and rarefaction curve.
Both communities 1 (**A**) and 2 (**B**) have the same species richness, five species each. However, organisms in community 1 (**A**) are more evenly distributed than in community 2 (**B**). With the same sampling efforts, **A** is more diverse than **B** based on the rarefaction curve (**C**). The triangles represent bacterial species, and different species are presented in different colors.

Rarefaction

- Process of adjusting for differences in library sizes that will affect how you analyze alpha diversity by selecting a sampling depth to standardize across samples
- A very low sampling depth (I) means that you retain less variants
- An optimal sampling depth (II) you retain all samples and more variants
- At a high sampling depth (III) you retain the most features but can lose whole samples (eg. sample F would be omitted)



Major considerations for Rarefaction

1. Is dependent on your research question! You need to know which metadata categories to focus on and prioritize sample sizes using your table.qzv file
2. Use your alpha rarefaction curve to determine if the sampling depth you chose from the first step is at a saturation point where you have most all of your ASVs represented

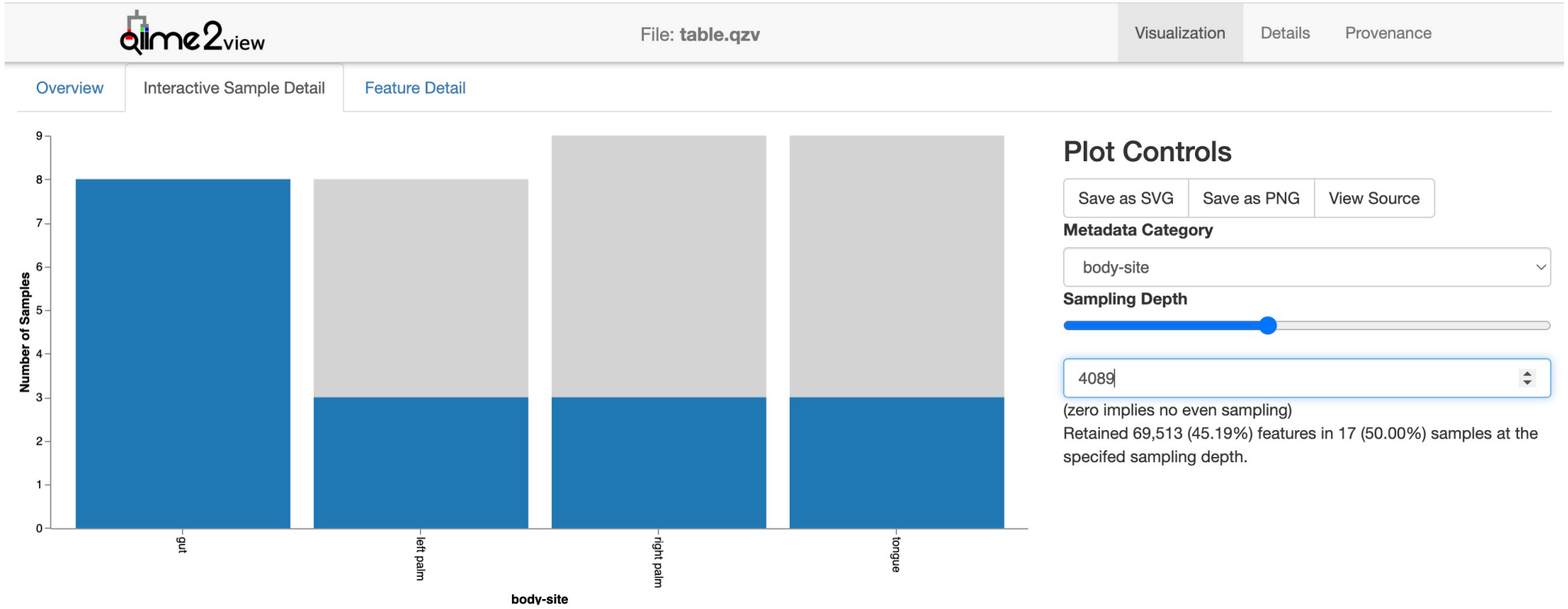
1. Alpha Rarefaction Curve

```
qiime diversity alpha-rarefaction \
--i-table table.qza \
--i-phylogeny rooted-tree.qza \
--p-max-depth 8000 \
--m-metadata-file /mnt/datasets/project_1/moving_pictures/sample-
metadata.tsv \
--o-visualization alpha-rarefaction.qzv
```

Goal is to resolve the saturation point

2. Table.qzv file

- Revisit the table.qzv file from the MPT



Application of rarefaction

-Running the alpha and beta diversity analysis in QIIME2

```
qiime diversity core-metrics-phylogenetic \
--i-phylogeny rooted-tree.qza \
--i-table table.qza \
--p-sampling-depth 4098 \
--m-metadata-file sample-metadata.tsv \
--output-dir core-metrics-results
```

Based on (1) and (2)

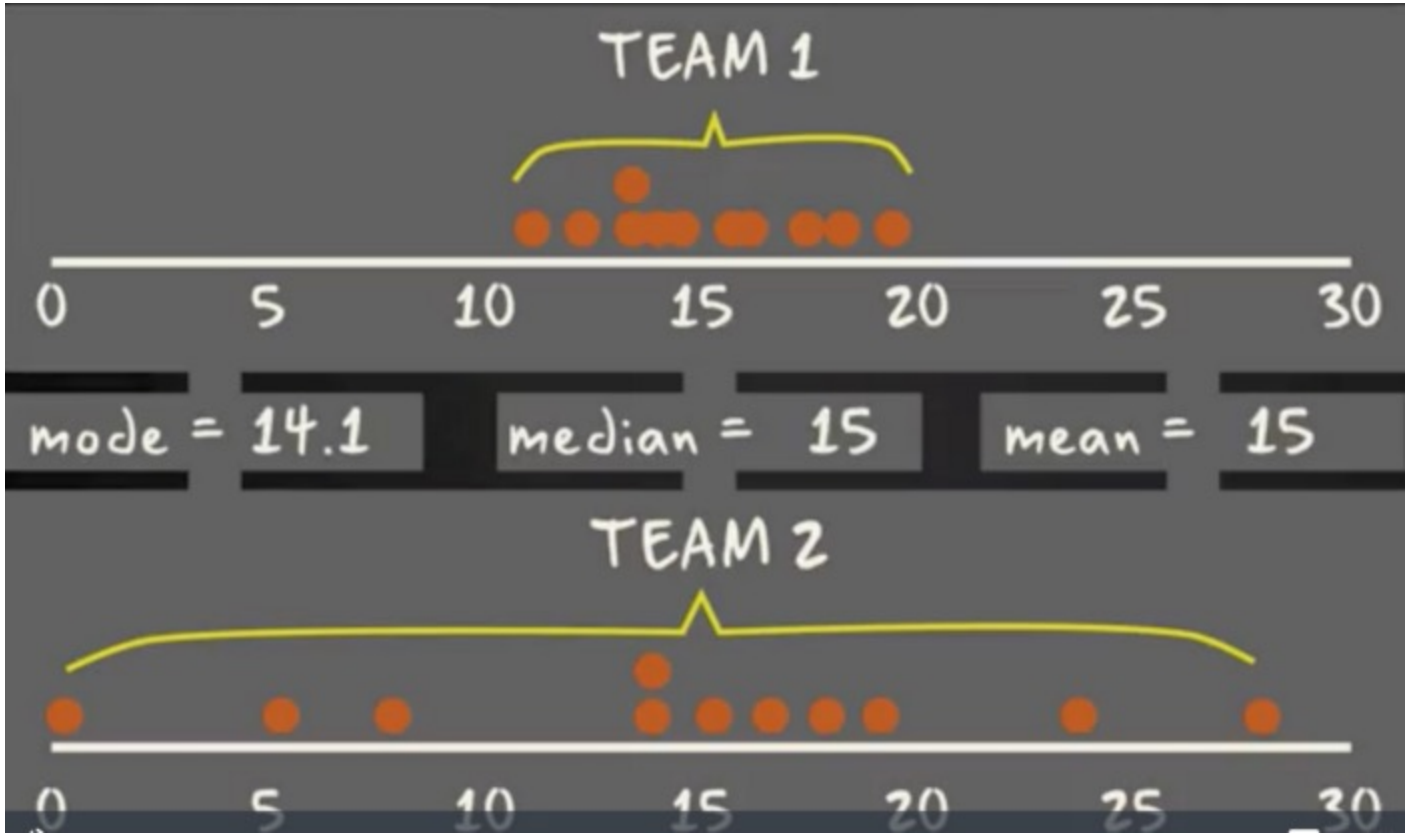
You generated a folder filled with the different metrics

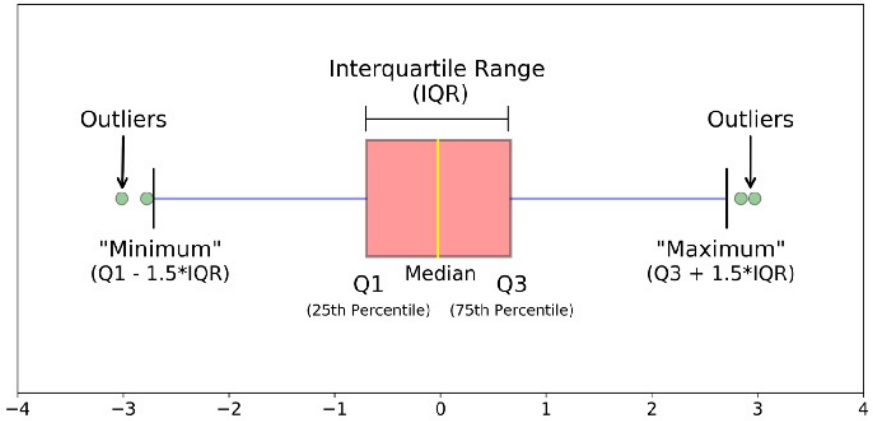
Optimizing Sampling Depth

1. Alpha rarefaction curve – look for the plateau where ASVs are satured
2. Use the table.qzv and focus on your metadata category of interest and <u>retain as many samples as possible</u>

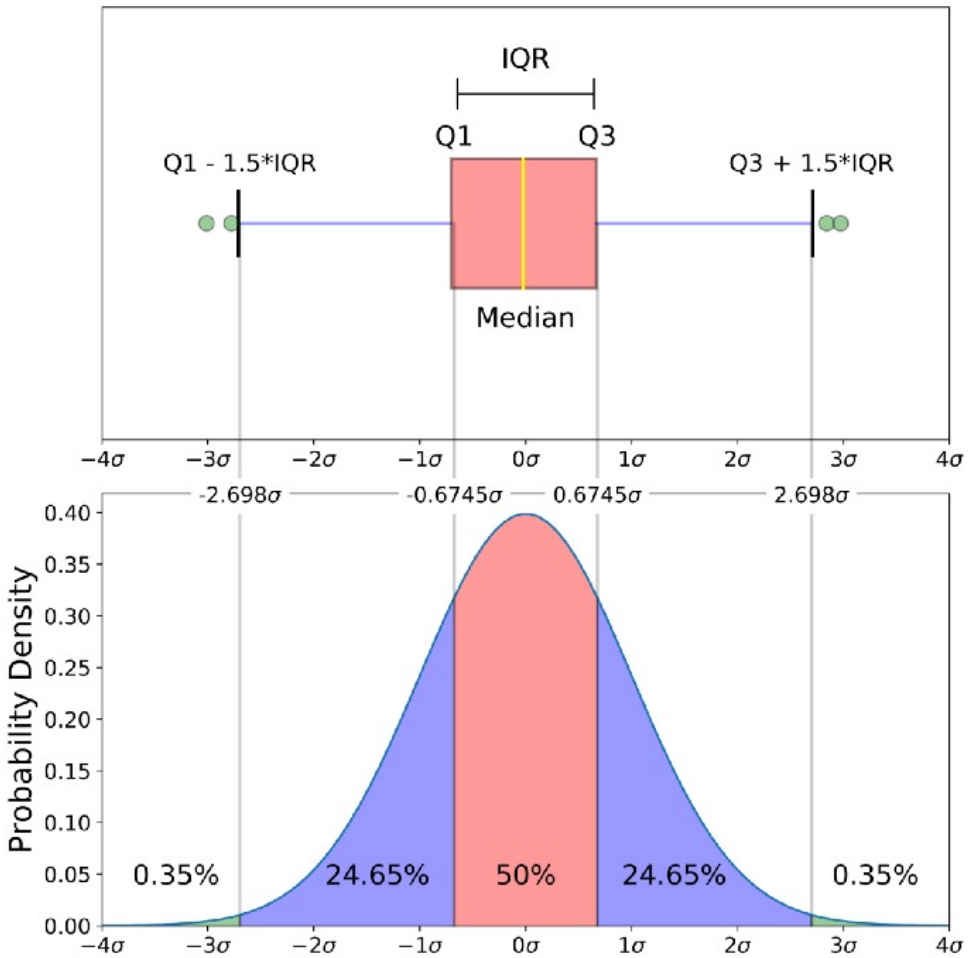
Boxplots

Boxplots


TEAM 1
0
5
10
15
20
25
30
mode = 14.1
median = 15
mean = 15
TEAM 2


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

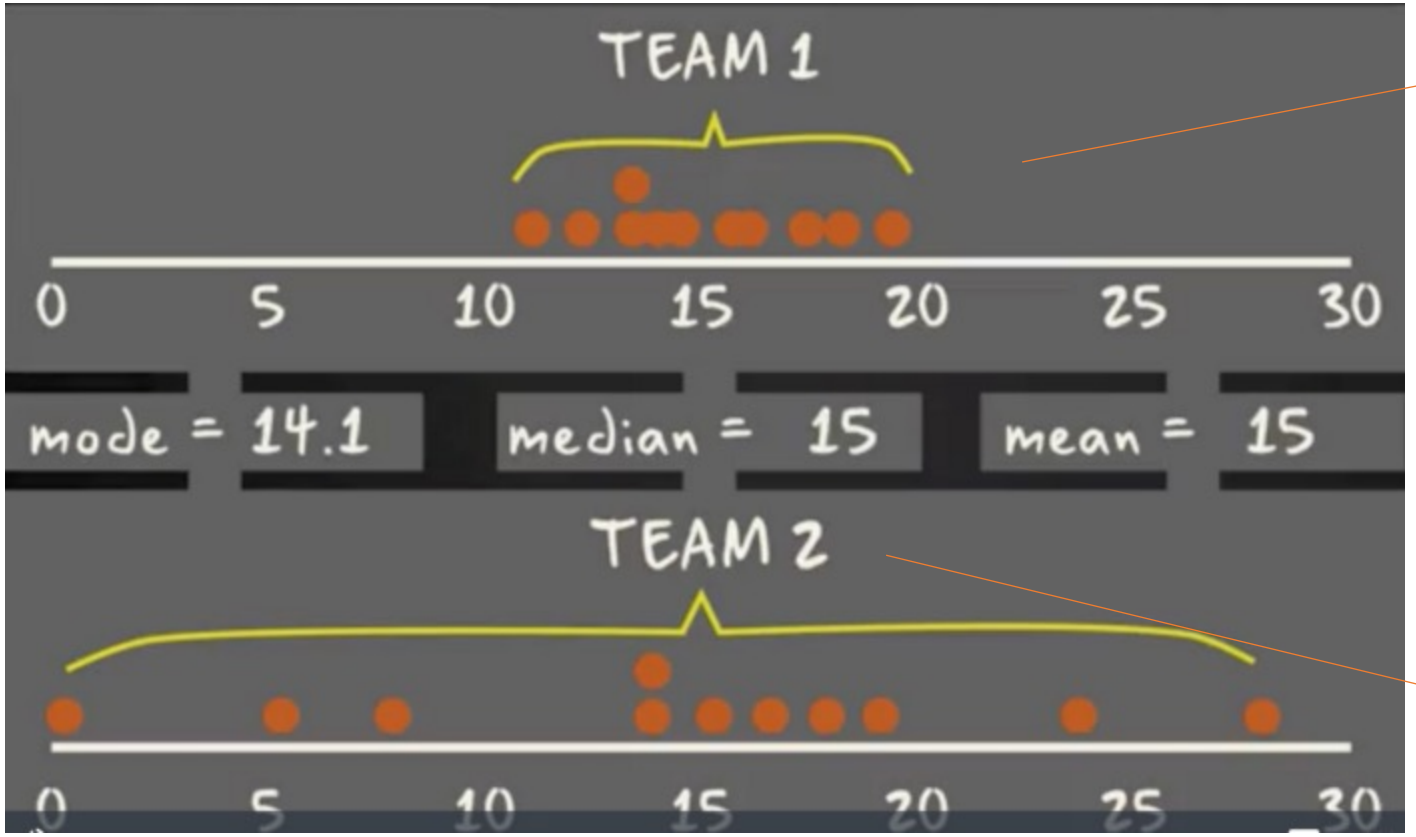
Interpreting boxplots

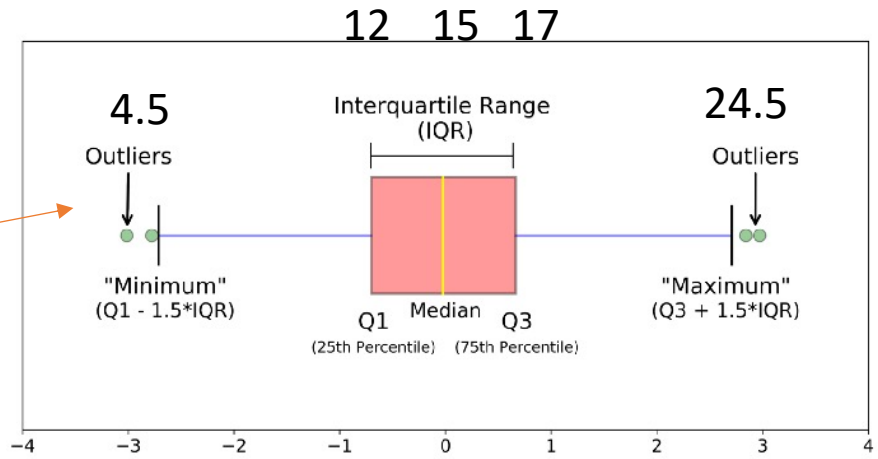


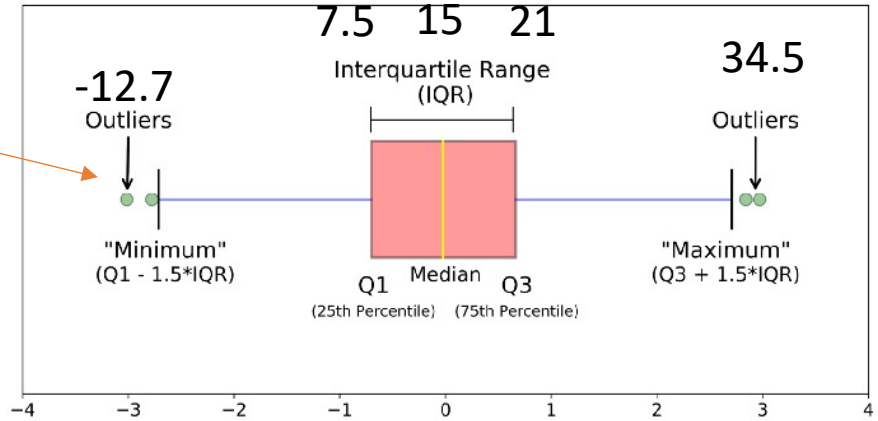
- **median (Q2/50th Percentile)**: the middle value of the dataset.
- **first quartile (Q1/25th Percentile)**: the middle number between the smallest number (not the "minimum") and the median of the dataset.
- **third quartile (Q3/75th Percentile)**: the middle value between the median and the highest value (not the "maximum") of the dataset.
- **interquartile range (IQR)**: 25th to the 75th percentile.
- **whiskers (shown in blue)**
- **outliers (shown as green circles)**
- **"maximum"**: Q3 + 1.5*IQR
- **"minimum"**: Q1 -1.5*IQR

https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Boxplots


TEAM 1
0
5
10
15
20
25
30
mode = 14.1
median = 15
mean = 15
TEAM 2

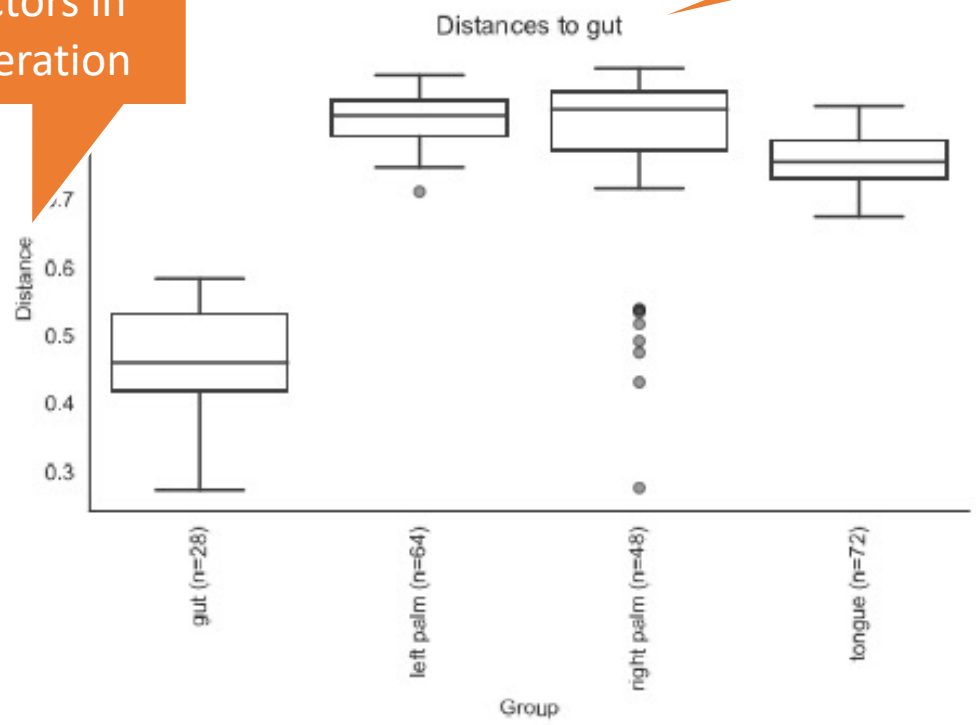

12 15 17
4.5
Outliers
Interquartile Range
(IQR)
24.5
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4 −3 −2 −1 0 1 2 3 4

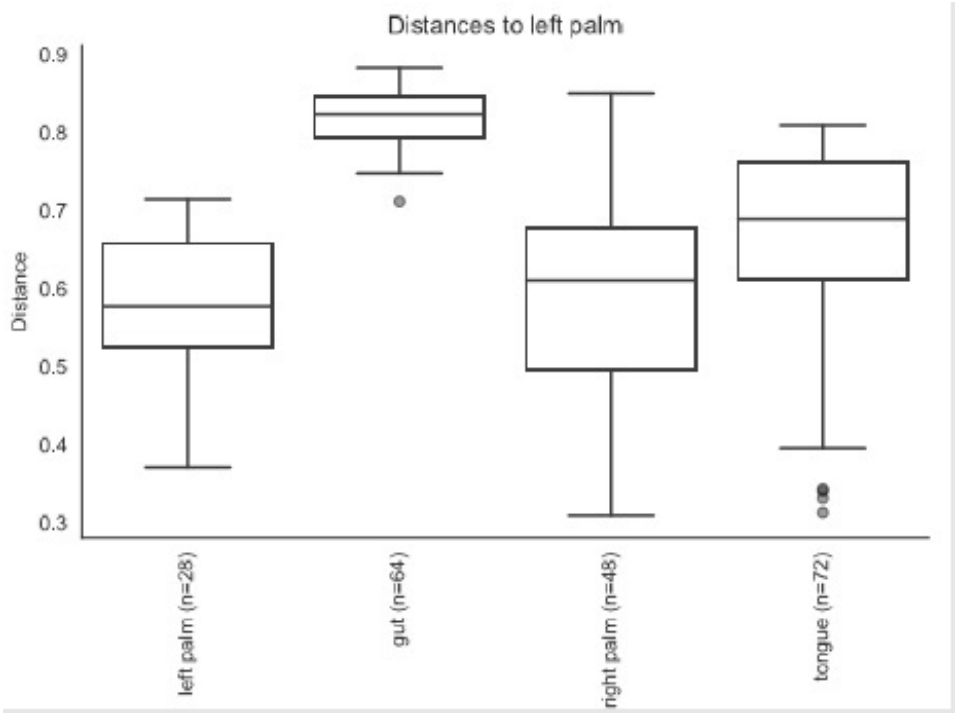

7.5 15 21
-12.7
Outliers
Interquartile Range
(IQR)
34.5
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4 −3 −2 −1 0 1 2 3 4

Beta diversity – measuring dissimilarity between communities
Unweighted UniFrac – considers presence/absence + genetic relatedness

Measure of all the factors in consideration

Reference point



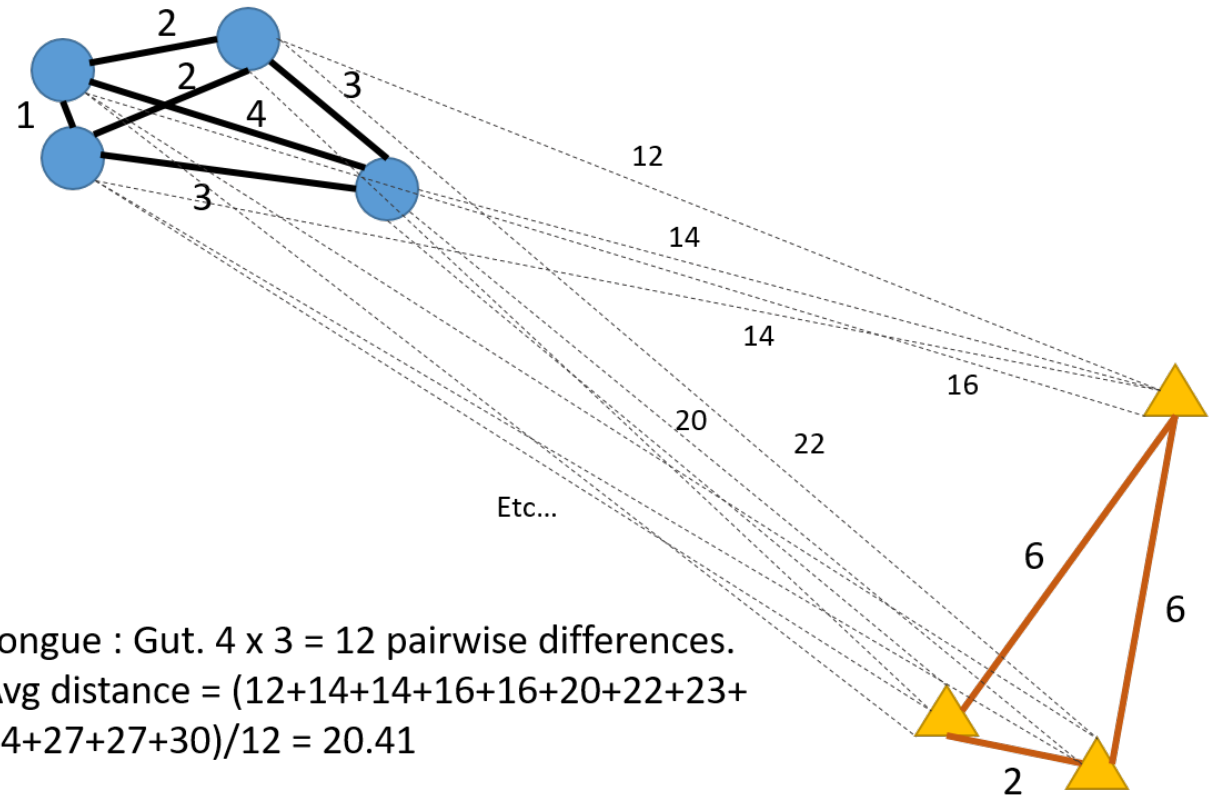


unweighted-unifrac-body-site-significance.qzv

https://view.qiime2.org/visualization/?type=html&src=https%3A%2F%2Fdocs.qiime2.org%2F2020.8%2Fdata%2Ftutorials%2Fmoving-pictures%2Fcore-metrics-results%2Funweighted-unifrac-body-site-significance.qzv

Measuring “distance”

“GUT” 4 gut samples. 6 pairwise distances. Avg distance = (1+2+2+3+3+4)/6 = 2.5

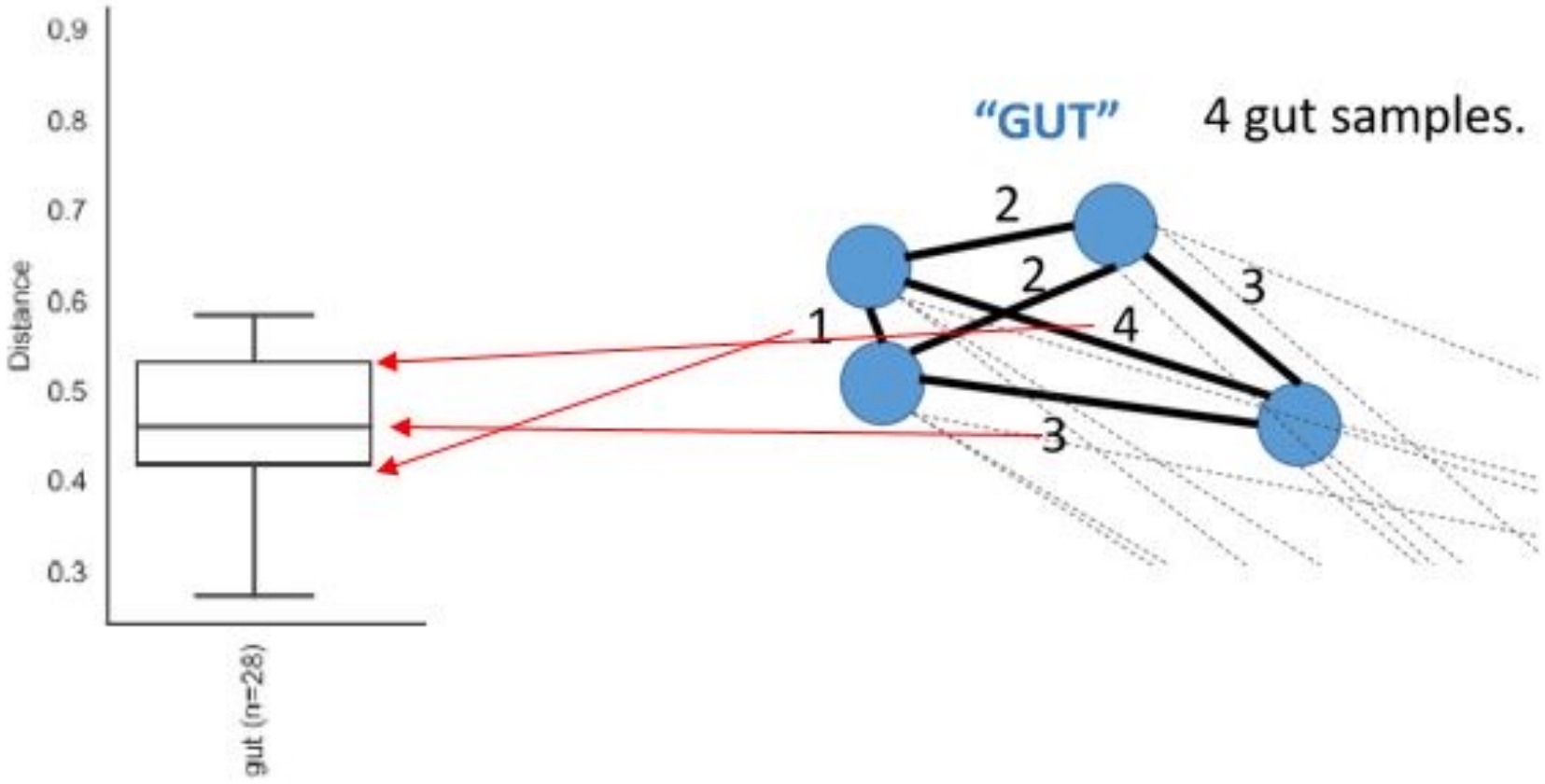


“TONGUE”

3 tongue samples. 3 pairwise distances.
Avg distance = (6+6+2)/3 = 4.67

Tongue : Gut. 4 x 3 = 12 pairwise differences.
Avg distance = (12+14+14+16+16+20+22+23+24+27+27+30)/12 = 20.41

Measuring “distance”

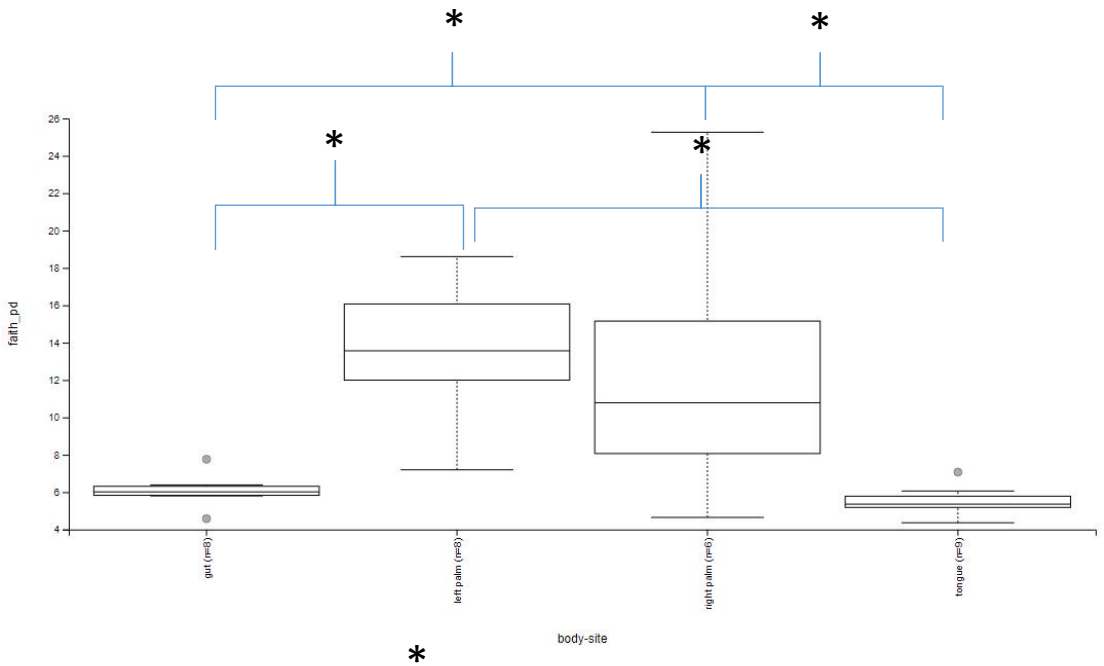


q-value

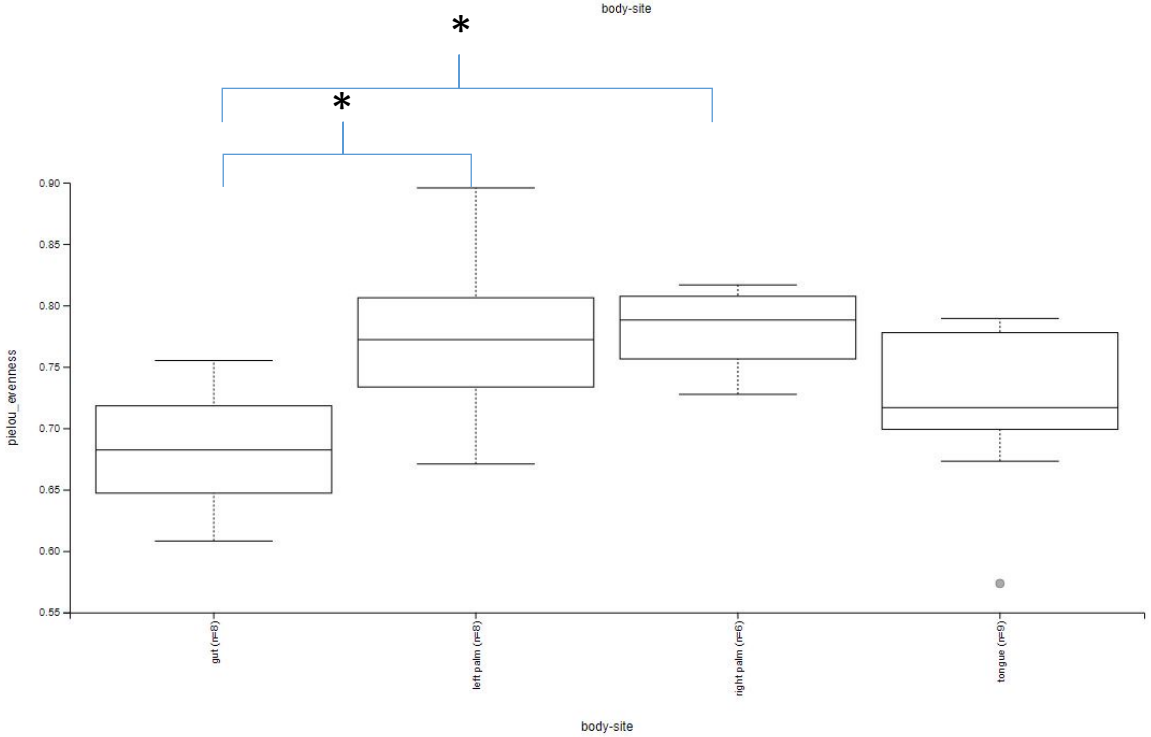
- q-value = adjusted p-value
- Read it the same way you do for a p-value
- Really important when making multiple comparisons
- Corrects for potential false positives when making multiple comparisons (false discovery rate)
- **Take this in consideration instead of the raw p-value for multiple comparisons**

Running all the metrics

- When doing alpha or beta diversity, you normally try all the types of analysis
- They all take richness in consideration
- See different patterns of significance when you run all the metrics.



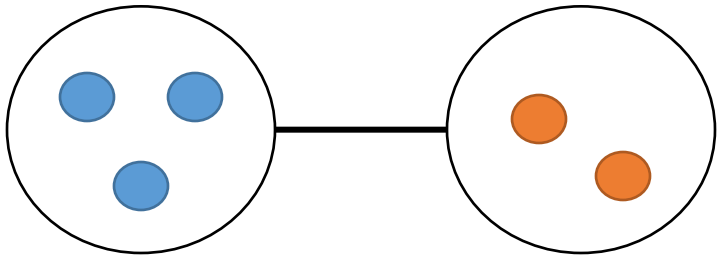
Faith's phylogenetic distance

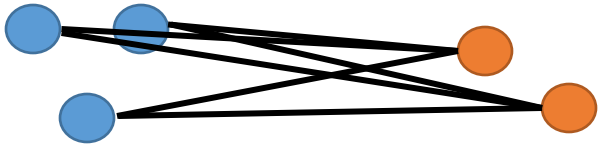


Evenness

Alpha versus Beta

- Alpha: look at the diversity of each sample/metadata category and you can compare them to each other but you will be comparing the average
- Beta: a more sophisticated comparison of multiple samples in one category to another as the reference point





PCoA

Principal Coordinate Analysis

PCoA Example: Olympics

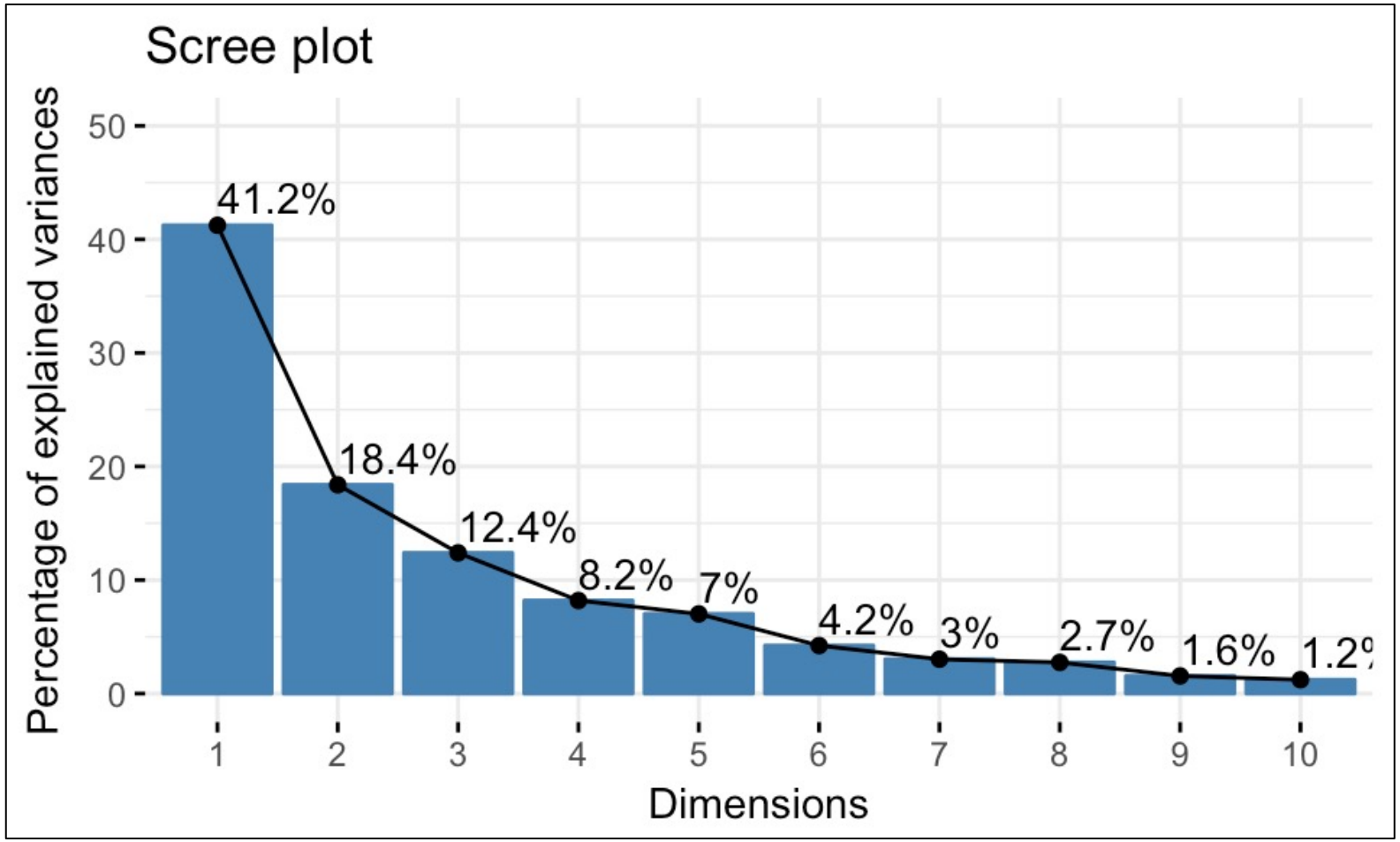
name	100m	Long.jump	//	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58		63.19	291.7	1	8217	Decastar
CLAY	10.76	7.4		60.15	301.5	2	8122	Decastar
Macey	10.89	7.47		58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74		55.39	278.05	5	8343	OlympicG
\\								
Zsivoczky	10.91	7.14		63.45	269.54	6	8287	OlympicG
Hernu	10.97	7.19		57.76	264.35	7	8237	OlympicG
Pogorelov	10.95	7.31		53.45	287.63	11	8084	OlympicG
Schoenbeck	10.9	7.3		60.89	278.82	12	8077	OlympicG
Barras	11.14	6.99		64.55	267.09	13	8067	OlympicG
KARPOV	11.02	7.3		50.31	300.2	3	8099	Decastar
WARNERS	11.11	7.6		51.77	278.1	6	8030	Decastar
Nool	10.8	7.53		61.33	276.33	8	8235	OlympicG
Drews	10.87	7.38		51.53	274.21	19	7926	OlympicG

Lots of data affected by many variables

We want to cluster the athletes based on how similar their data is to each other

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/

Not all variables can be represented

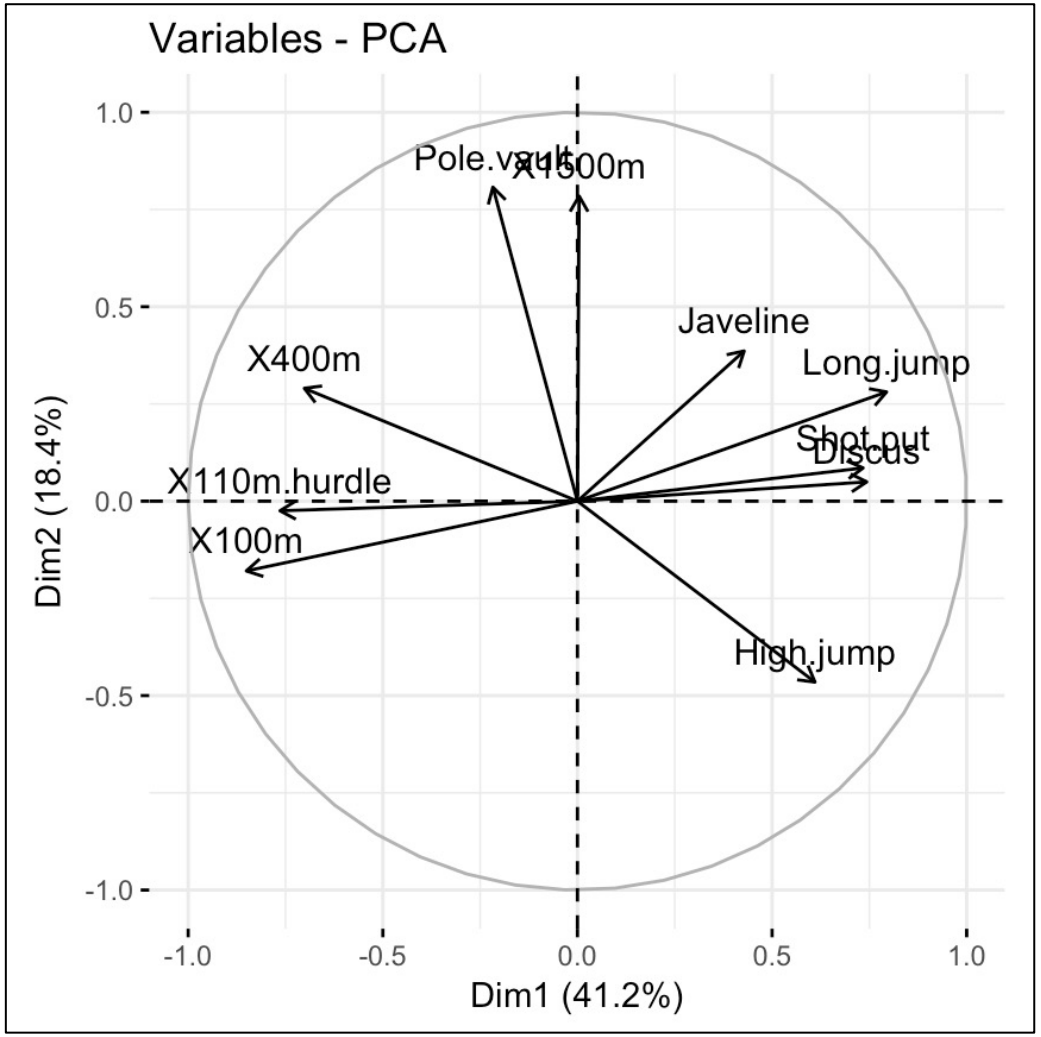


Can’t represent all the variables but there are some can be combined

You often see 2 dimensions that represent the most variables possible (dimension 1 and 2)

This plot is not one that you generate.

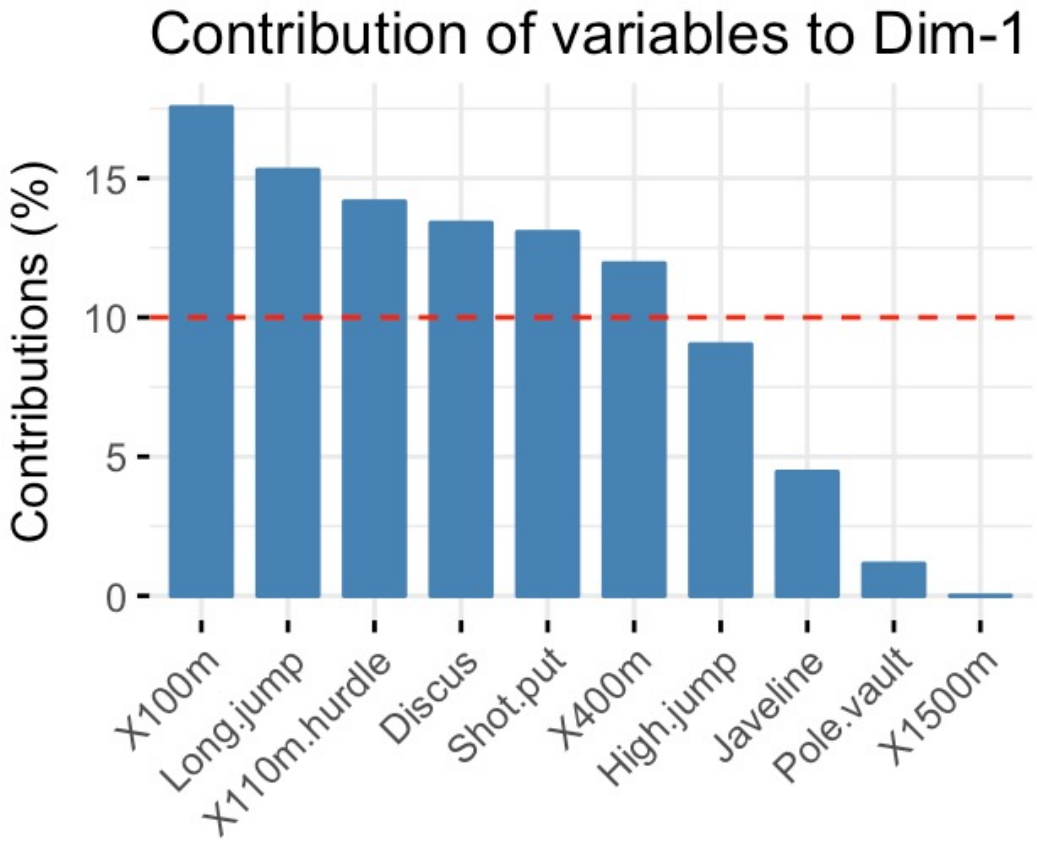
Loading Plot – Which variables affect data positioning

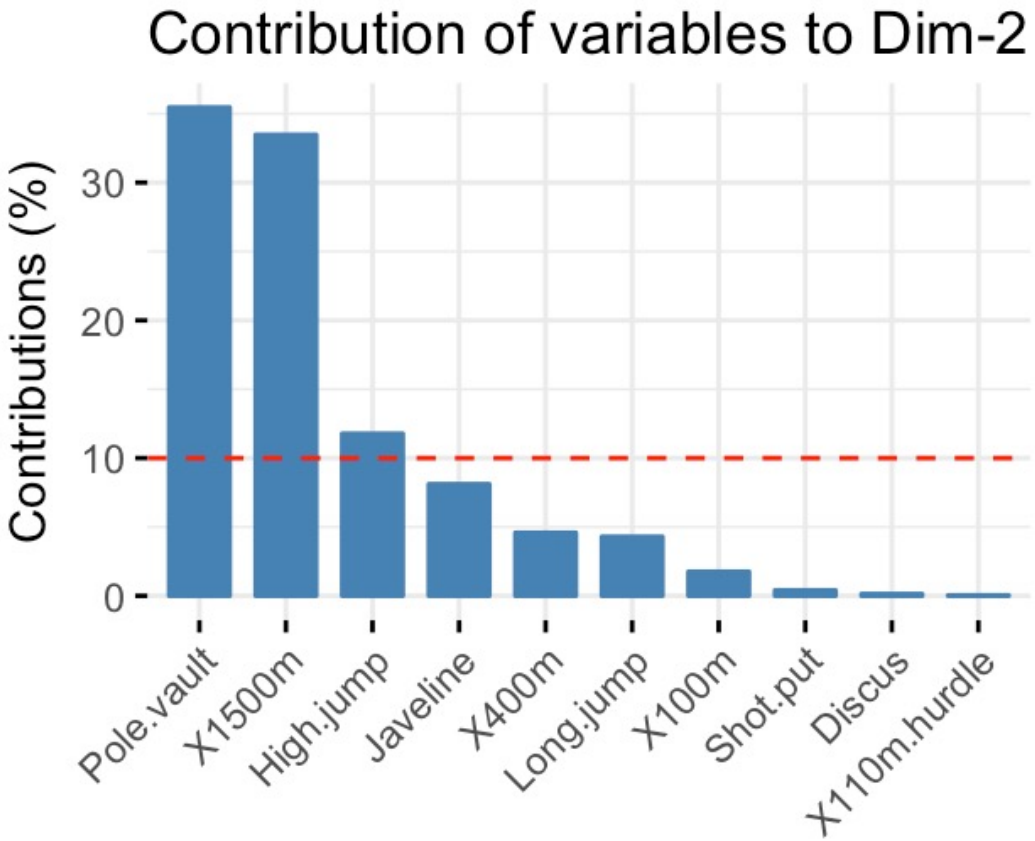


Different variables affect where the data points (ie athletes) are placed on the PCA plot based on how much impact they have on their overall comparison to other data points

This plot is not one that you generate.

Variables have different weights in how they affect the clustering of data



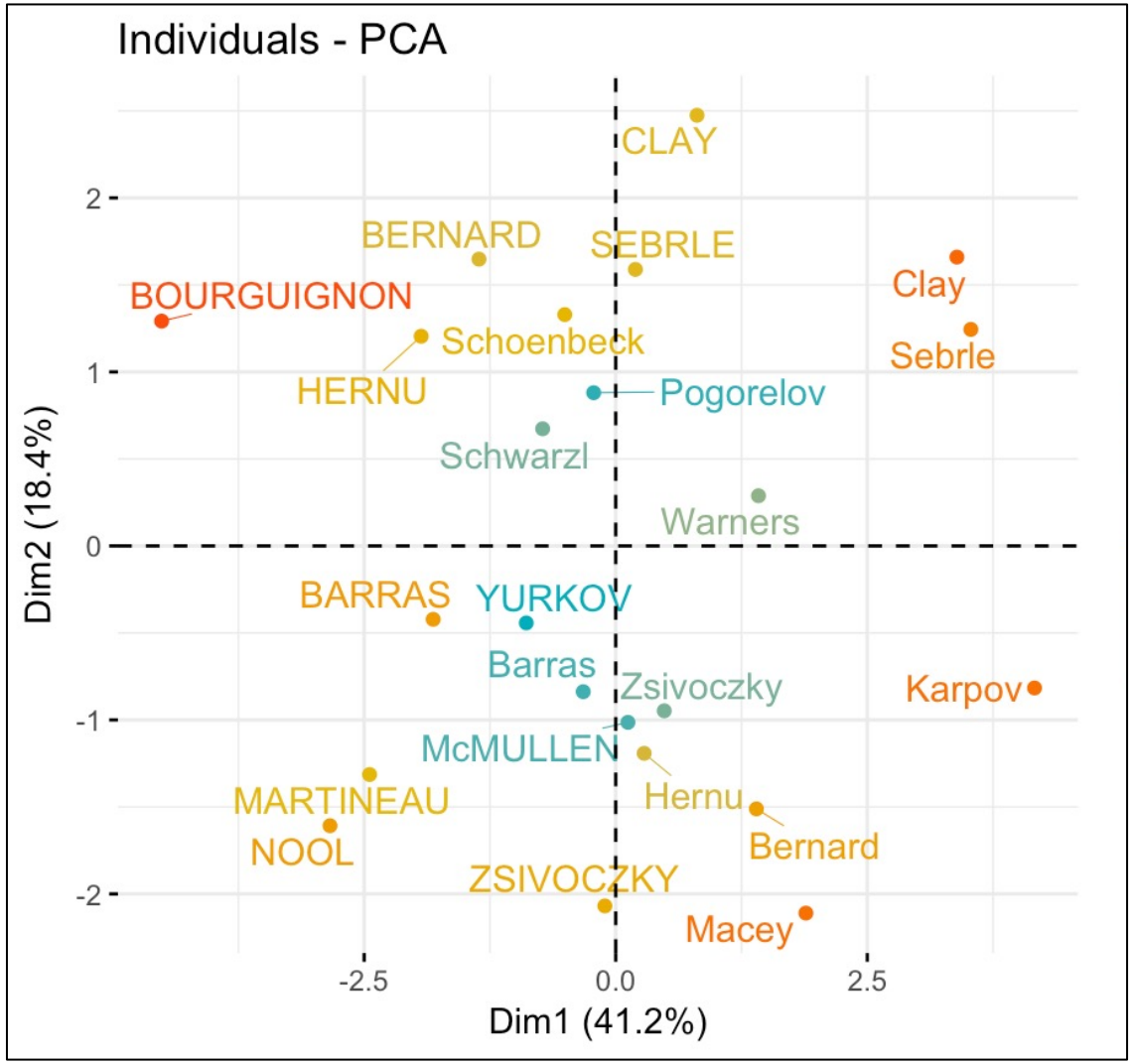


Within each dimension, different variable carry different weights on how they affect data distribution

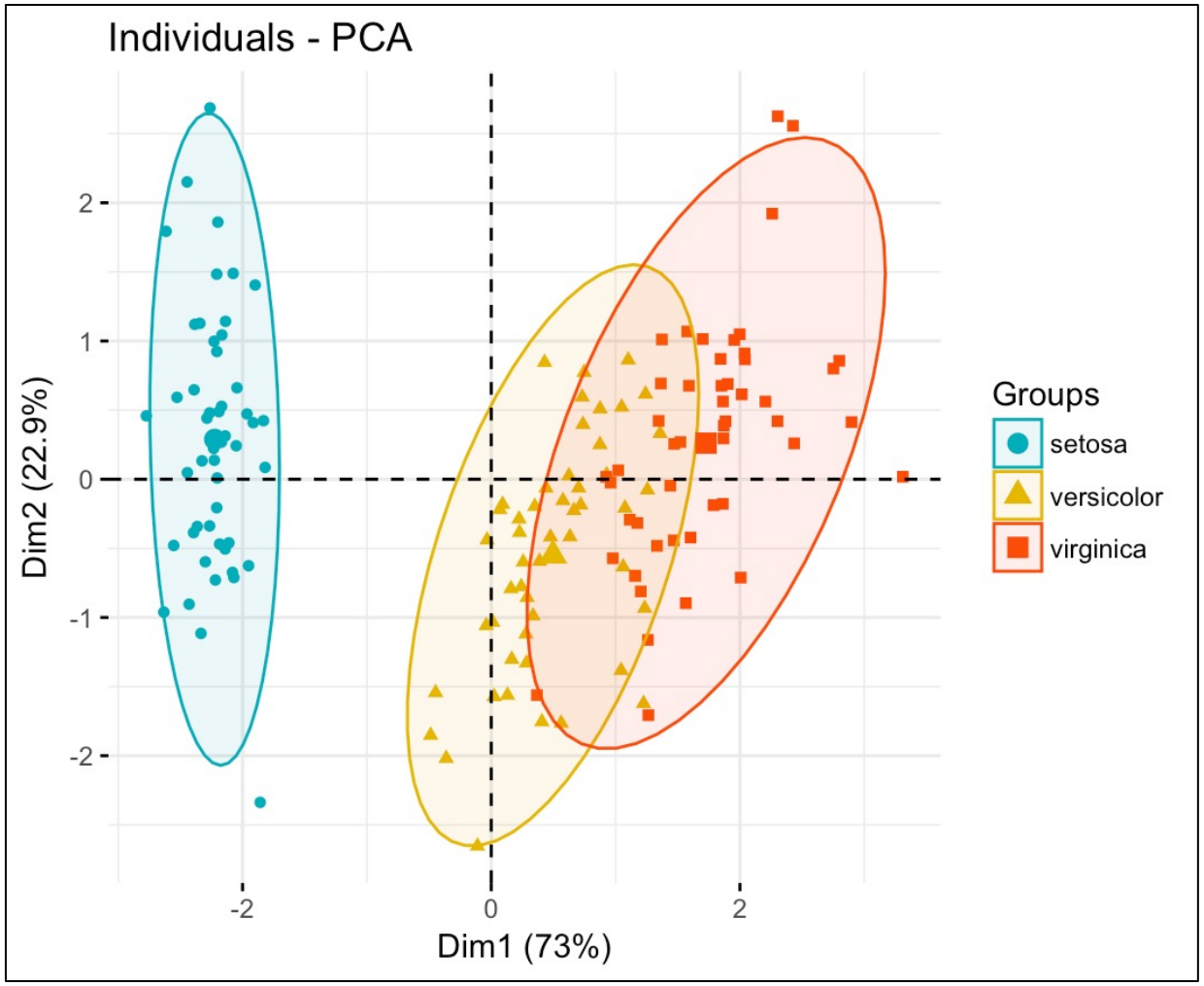
PCA Plot of Athletes based on the two dimensions

This is our final PCA plot!

We often try to look for clusters among samples.

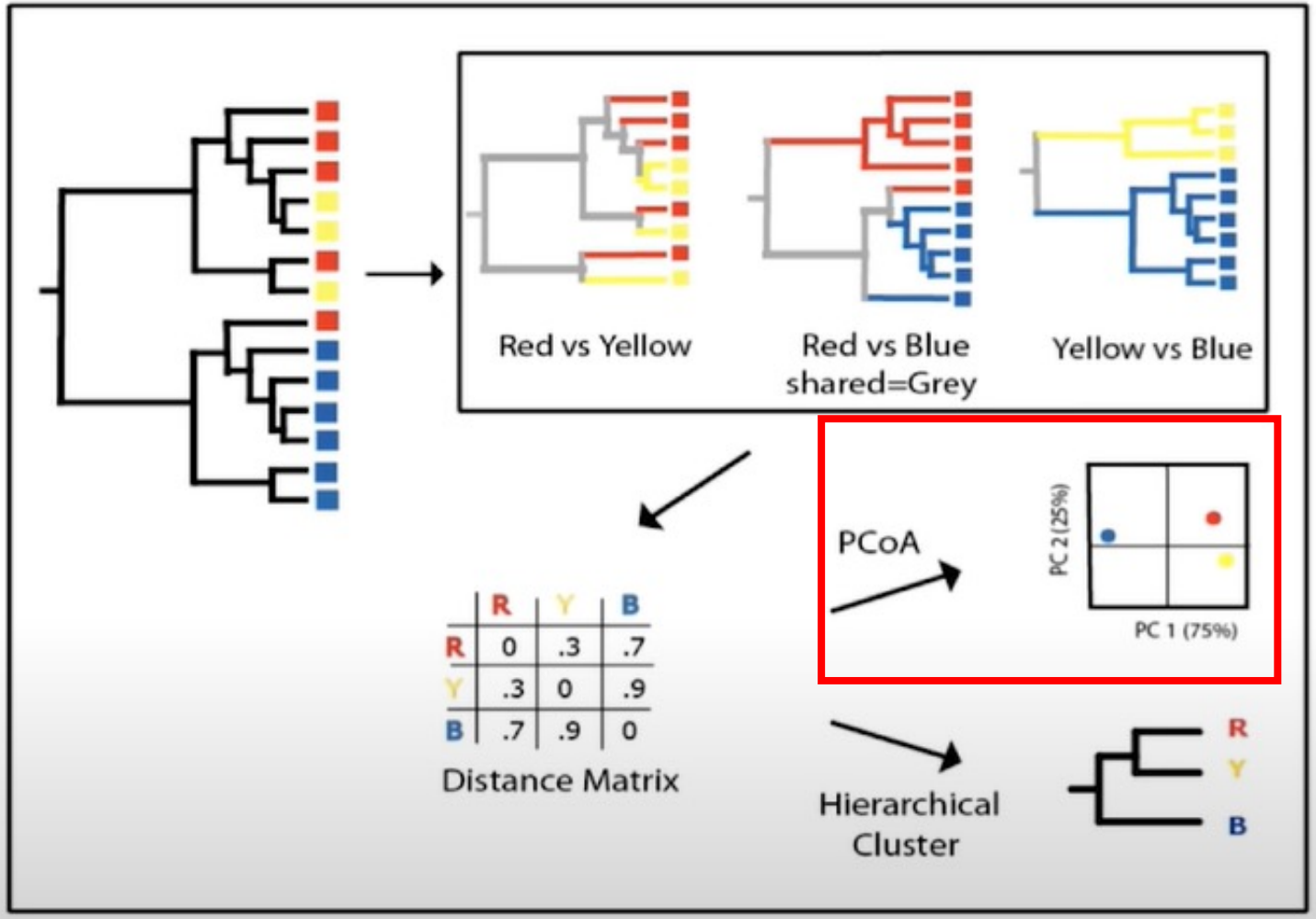


Finding clusters



You can use R to find clusters and draw the ellipses around them for better visualization

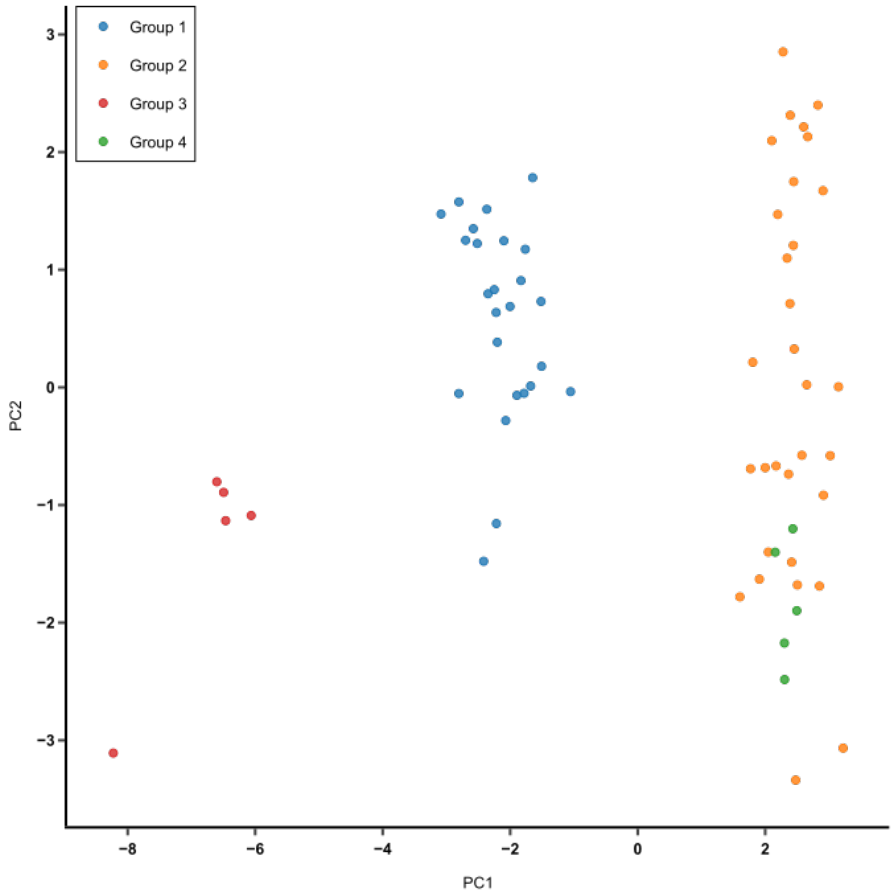
Beta diversity using UniFrac



https://www.youtube.com/watch?v=M8ylvsS0MHg

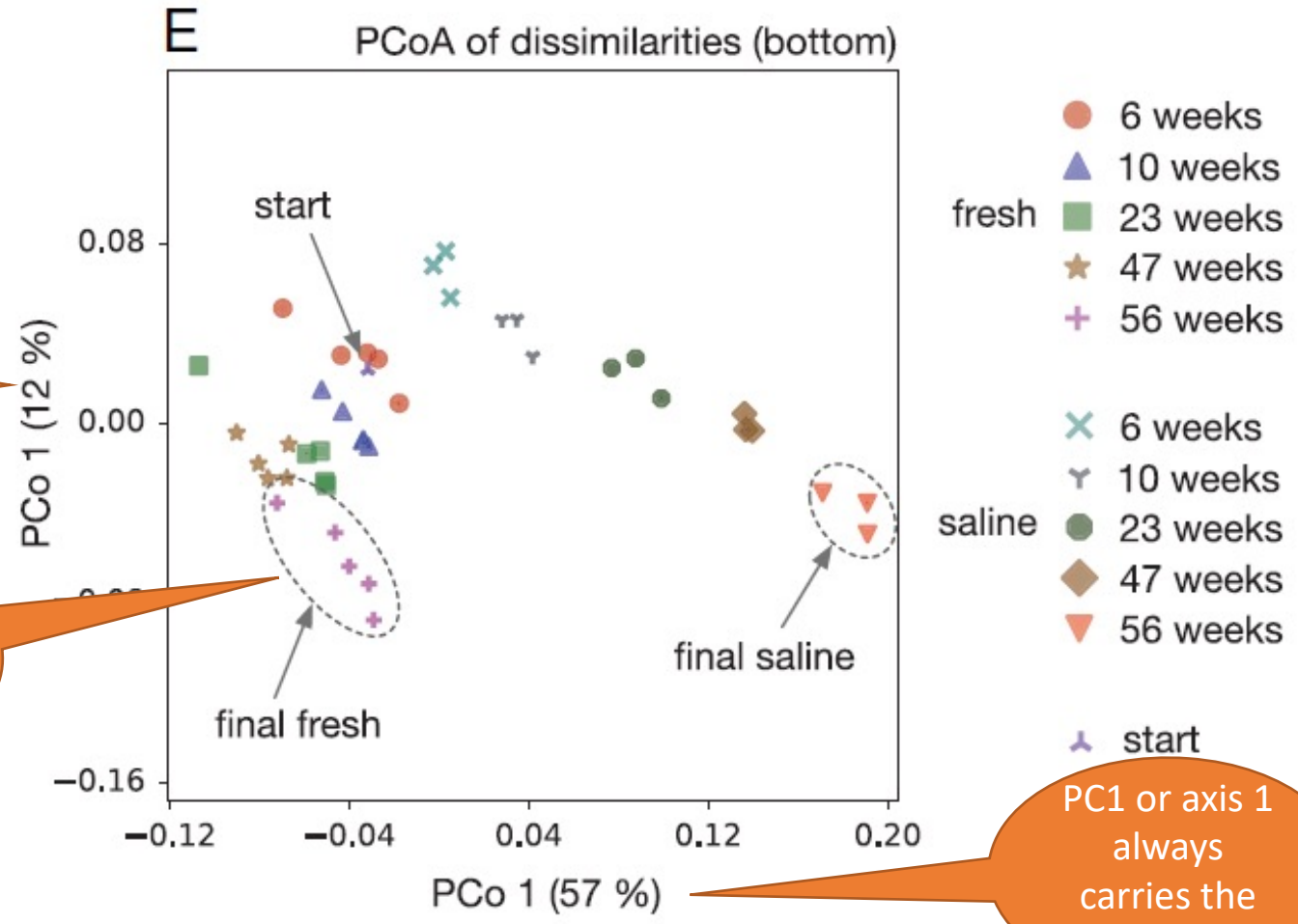
How are PCA plots generated?

- Summary of all the different variables within the data set
- Takes a multi-dimensional plot and flattens it into one dimension
- Principal component represents that set of variables = variance



https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/#:~:text=A%20PCA%20plot%20shows%20clusters%20of%20samples%20based%20on%20their%20similarity.&text=PCA%20does%20not%20discard%20any%20samples%20or%20characteristics%20(variables).&text=Such%20influences%2C%20or%20loadings%2C%20can,produces%20the%20differences%20among%20clusters.

Interpreting PCoA plots



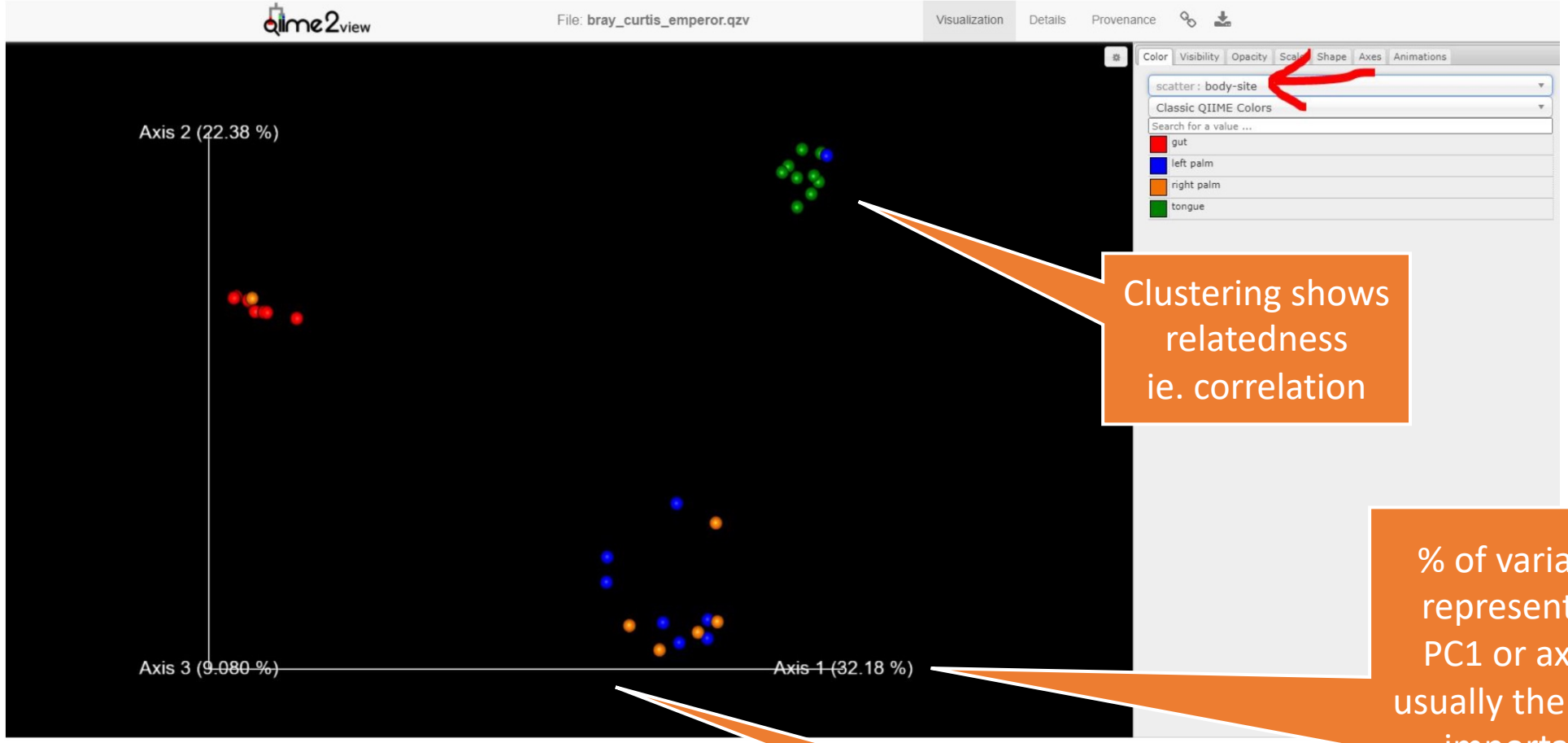
Variance represented

Clustering

PC1 or axis 1 always carries the most weight

Plots are courtesy of Ilan Rubin

Bray Curtis Metric – Principle Coordinate Analysis (PCoA) Plot



Clustering shows relatedness ie. correlation

% of variance represented; PC1 or axis 1 usually the most important

Axis take many variables in consideration