MGMA2024:

Beta diversity in microbiome analysis

Metrics and visualization

Duy Dao
khuongduying@gmail.com

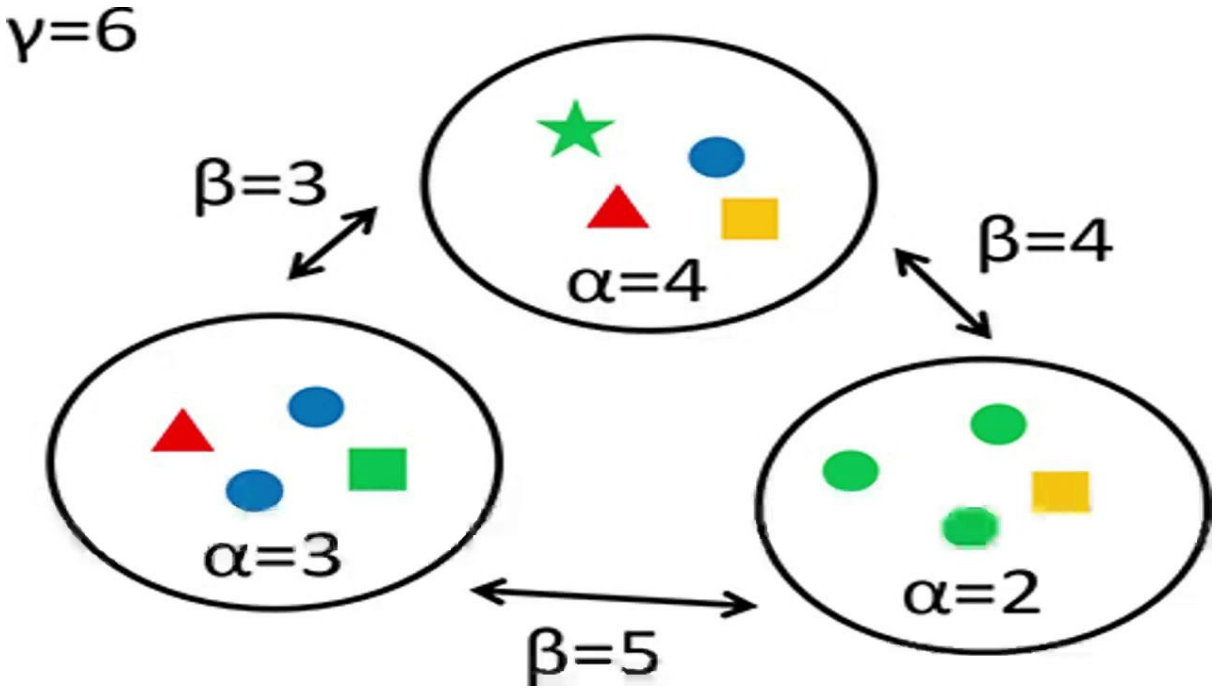
Content

- ❏ Introduction to Beta diversity
- ❏ Beta diversity metrics
- ❏ Ordination method
- ❏ Statistical testing of Beta diversity with PERMANOVA
- ❏ Practice

Introduction to Beta diversity

Introduction to Beta diversity

Compare between 3 ecosystems:



How different are they?

Introduction to Beta diversity

Hypothetical species	Woodland habitat	Hedgerow habitat	Open field habitat
A	X		
B	X		
C	X		
D	X		
E	X		
F	X	X	
G	X	X	
H	X	X	
I	X	X	
J	X	X	
K		X	
L		X	X
M			X
N			X
Alpha diversity	10	7	3
Beta diversity	Woodland vs. hedgerow: 7	Hedgerow vs. open field: 8	Woodland vs. open field: 13
Gamma diversity	14		

Source

Introduction to Beta diversity

Beta diversity

- Measures the **difference** between two samples or communities
- **Compare** the structure of the microbiome communities
- Requires a **distance** or **dissimilarity measure matrix** as input

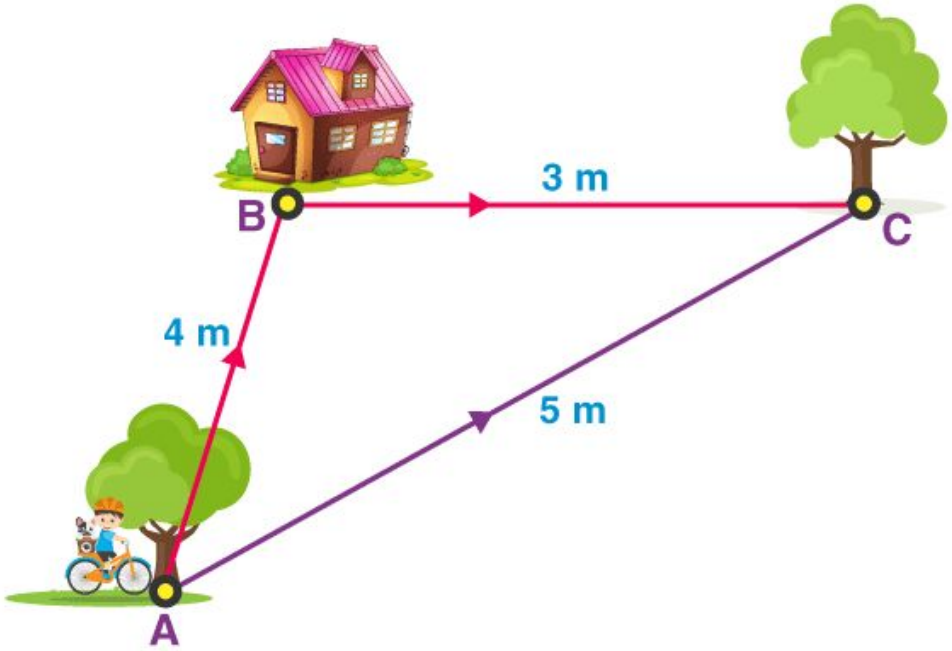
How different two or more communities are, either in their composition (richness) or in the abundance of the organisms that compose it (abundance).

Beta diversity metrics

distance / dissimilarity metrics

Beta diversity metrics

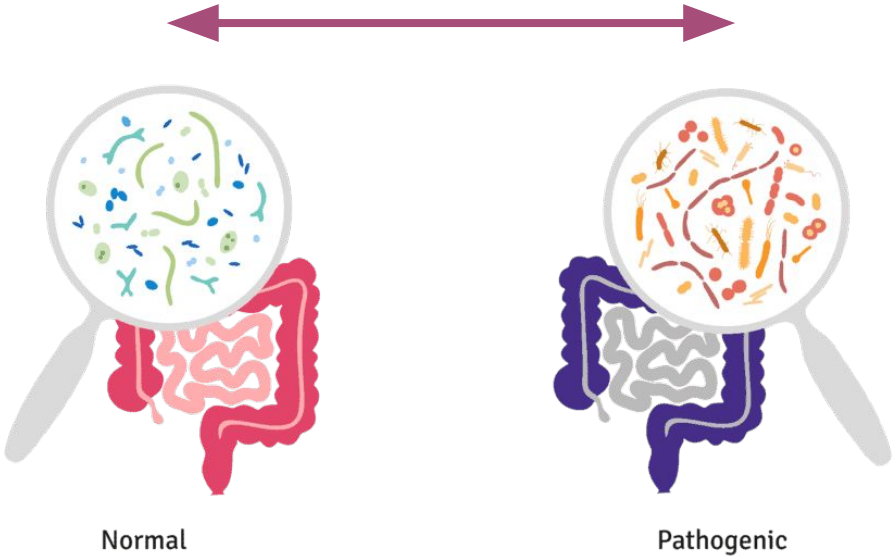
What is the distance in beta diversity?



Physical distance

Beta diversity metrics

What is the distance in beta diversity?



The distance measures…

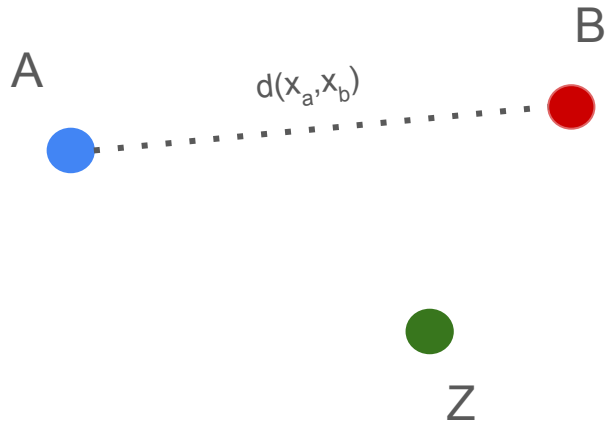
- How *“far”* they are from each other
- How *different* they are from each other

Image

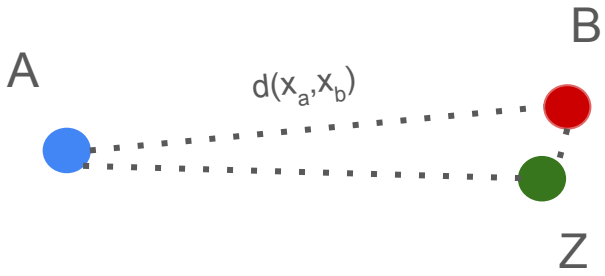
Beta diversity metrics

Some rules for distance

1. $d(x_a,x_b) \geq 0$
2. $d(x_a,x_b) = 0$, if x_a is equal to x_b
3. $d(x_a,x_b) = d(x_b,x_a)$
4. $d(x_a, x_b) \leq d(x_a,x_z) + d(x_z,x_b)$



Beta diversity metrics



- Each point is the entire diversity of a single sample
- The closer 2 points are to each other → The more similar the microbiomes of those samples are to each other.

Beta diversity metrics

★ **Non-phylogenetic dissimilarities:**
- Bray-Curtis index
- Jaccard index
- Sørensen index

★ **Phylogenetic-based beta diversity metrics:**
- Unweighted UniFrac distances
- Weighted UniFrac distances
- Generalized UniFrac distances

Beta diversity metrics

Jaccard index

- Non-phylogenetic
- Binary - Presence/absence observational data
- A 2 × 2 contingency table → calculation of the coefficients (or association)

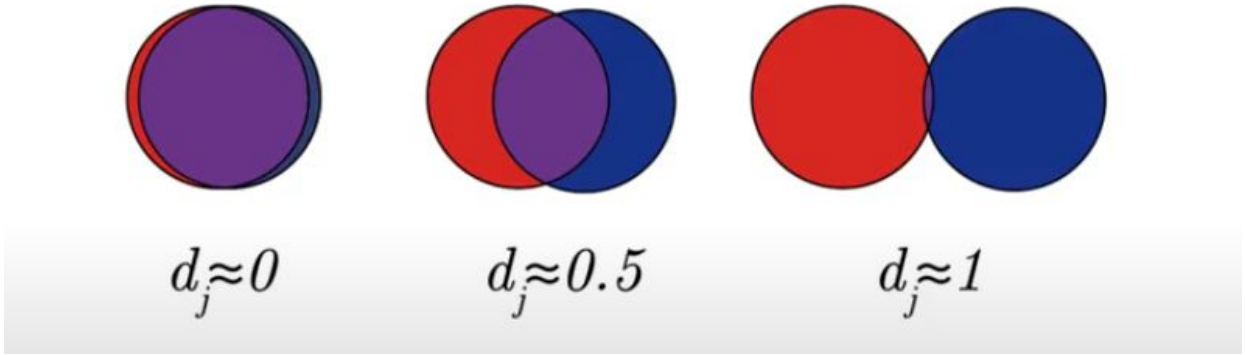
		Sample A	
		Presence	Absence
Sample B	Presence	a	b
	Absence	c	d

Jaccard similarity $S_j = \frac{a}{a+b+c}$, → Jaccard dissimilarity

dj = 1 - Sj

Beta diversity metrics

Jaccard index



Fraction of unique features, regardless of abundance

	Presence a	Absence a
Presence b	36	84
Absence b	13	0

dj = 0.73

	Presence a	Absence a
Presence b	113	2
Absence b	1	0

dj ≈ 0

https://www.youtube.com/watch?v=7H_LhmGafOc

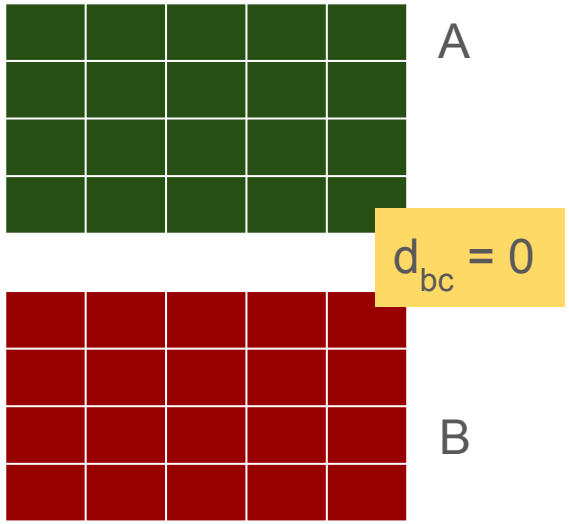
Beta diversity metrics

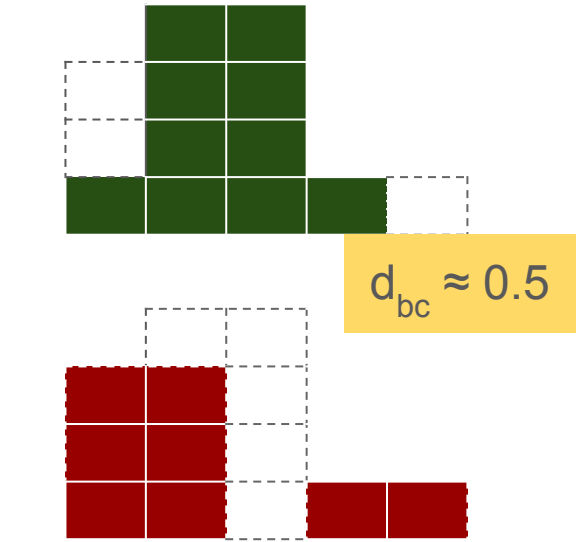
Bray-Curtis index

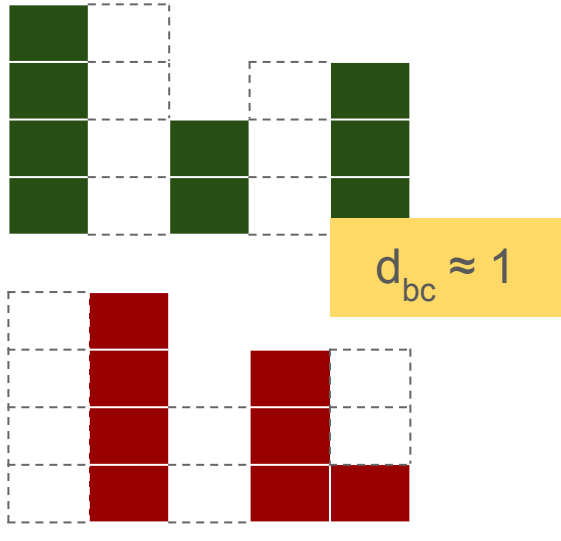
- Non-phylogenetic
- Quantitative
- Abundance (counts) + presence/absence
- Most widely used beta diversity in ecology and microbiome research fields

Beta diversity metrics

Bray-Curtis distance







Beta diversity metrics

Bray-Curtis distance

Feature table

	feature1	feature2	feature3	feature4	feature5
sampleA	42	0	37	99	1
sampleB	12	1	22	88	0
sampleC	25	3	23	86	0
sampleD	0	0	87	12	0

$$BC(A, B) = \frac{\Sigma_i \left| X_{iA} - X_{iB} \right|}{\Sigma_i (X_{iA} + X_{iB})}$$

X_{iA} : frequency of feature *i* in sample *A*

Distance matrix

	sampleA	sampleB	sampleC	sampleD
sampleA	0.0	0.19	0.15	0.65
sampleB	0.19	0.0	0.07	0.69
sampleC	0.15	0.07	0.0	0.70
sampleD	0.65	0.69	0.70	0.0

Useful if we see similar features but different in frequencies

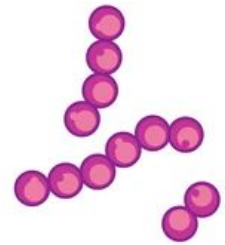
Phylogenetic Beta Diversity Metrics

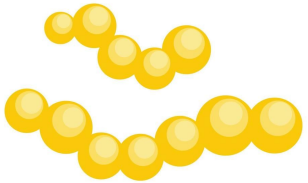
Why using phylogenetic-based beta diversity?

Situation:
2 samples have different species but related in taxonomy

- Our microbiome are different but closely related
- We cannot say that they are totally different

→ We want to have a more sensitive test





Phylogenetic Beta Diversity Metrics

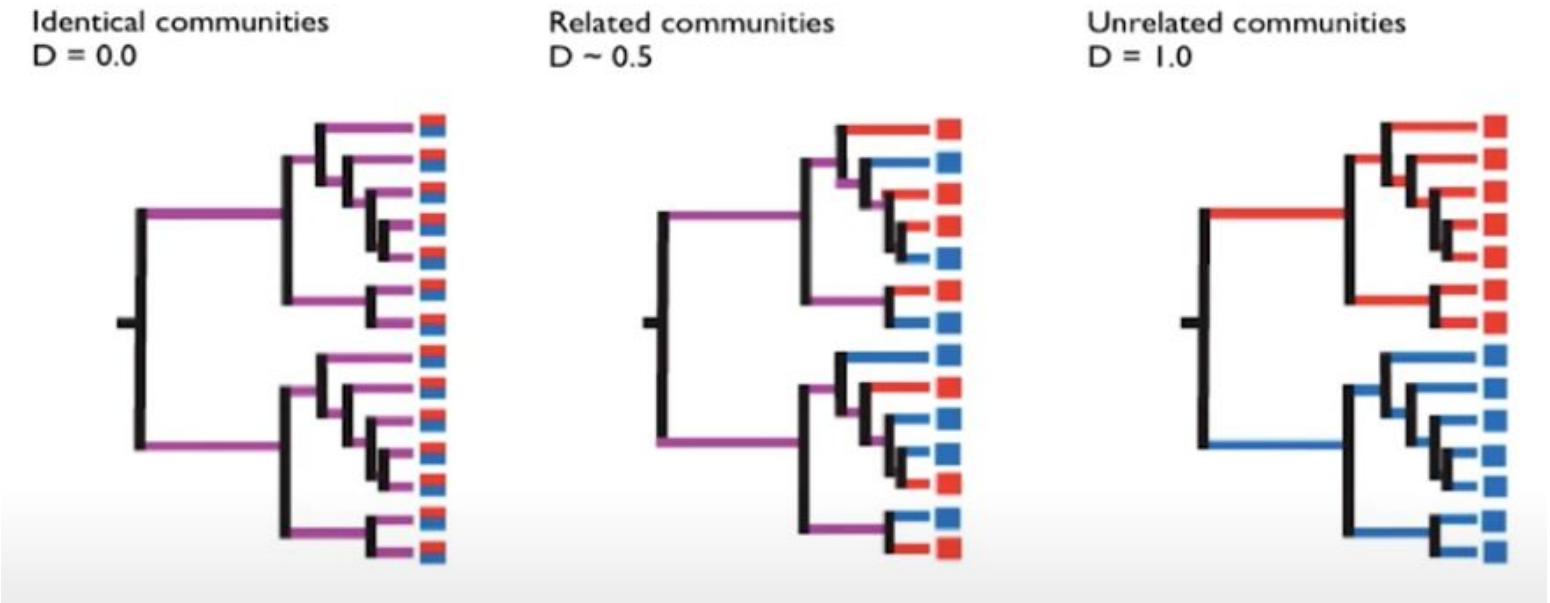
Unweighted UniFrac

- Phylogenetic-based
- Binary (presence/absence)

→ Measure the phylogenetic distance between sets of taxa in a phylogenetic tree

Phylogenetic Beta Diversity Metrics

Unweighted UniFrac



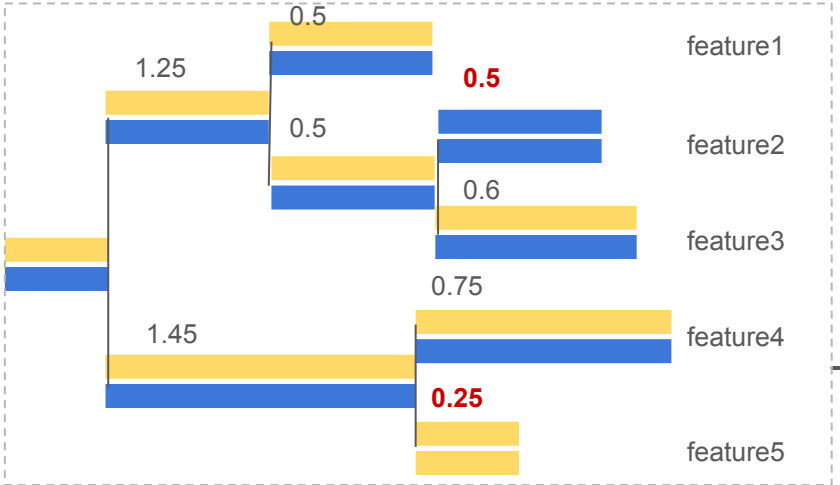
$$UU(A, B) = \frac{\textit{sum of unique branch length}}{\textit{sum of observed branch length}}$$

Phylogenetic Beta Diversity Metrics

Feature table

	feature1	feature2	feature3	feature4	feature5
sampleA	42	0	37	99	1
sampleB	12	1	22	88	0
sampleC	25	3	23	86	0
sampleD	0	0	87	12	0

Unweighted UniFrac



Distance matrix

	sampleA	sampleB	sampleC	sampleD
sampleA	0.0			
sampleB	0.13	0.0		
sampleC			0.0	
sampleD				0.0

$$\frac{0.5 + 0.25}{1.25 + 0.5 + 0.5 + 0.5 + 0.6 + 1.45 + 0.75 + 0.25}$$

Phylogenetic Beta Diversity Metrics

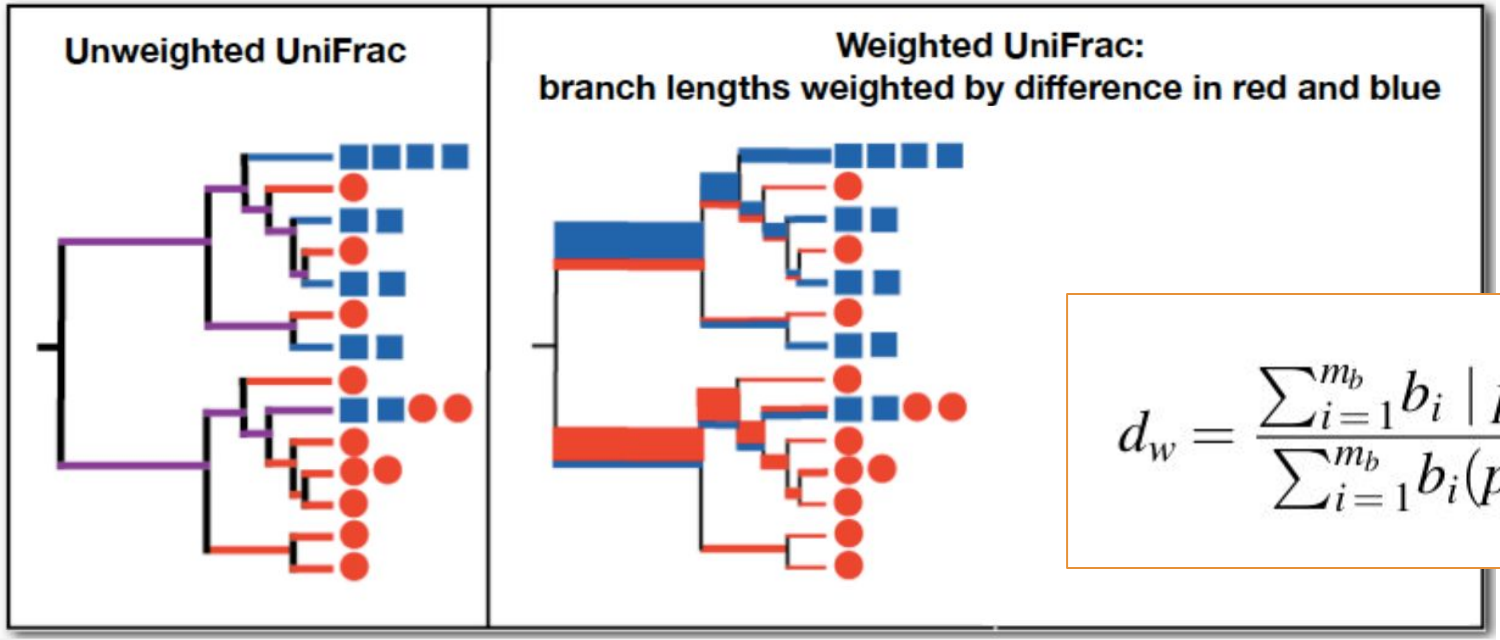
Unweighted UniFrac

Because of the probability that the rare taxa sequenced are directly related to the presence/absence of species

→ Unweighted UniFrac could most efficiently detect the variability in community membership or the abundance of rare lineages

Phylogenetic Beta Diversity Metrics

Weighted UniFrac distance



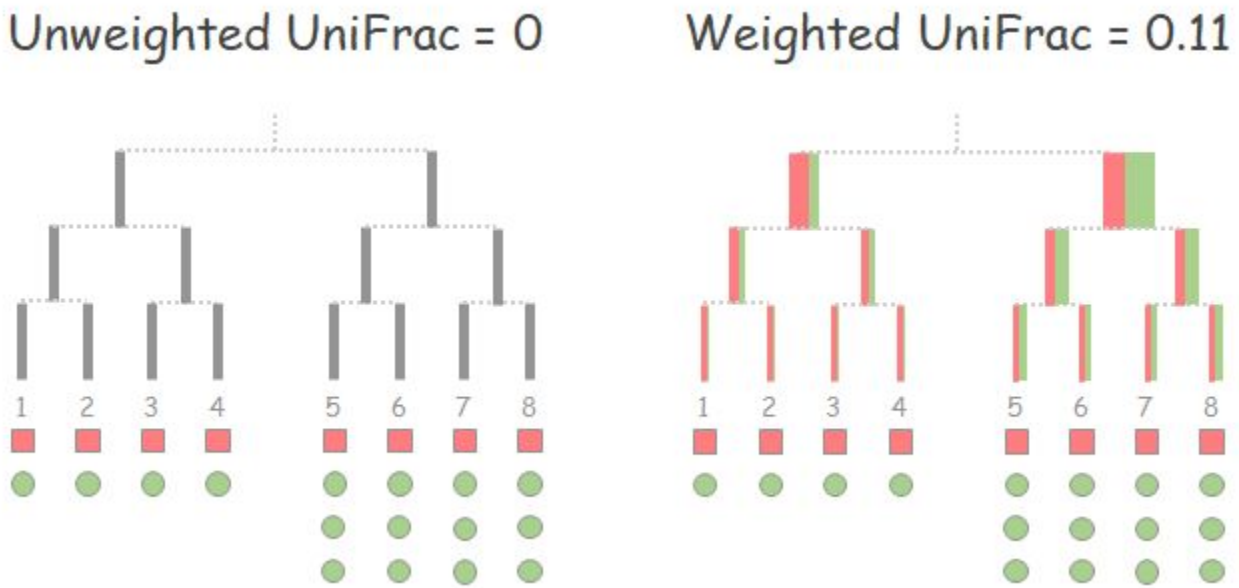
$$d_w = \frac{\sum_{i=1}^{m_b} b_i \mid p_i^A - p_i^B \mid}{\sum_{i=1}^{m_b} b_i (p_i^A + p_i^B)},$$

Uses species **abundance** information and **weights the branch length** with abundance difference.

https://sonnenburglab.github.io/MI221/lectures/distance_ordination.pdf

Phylogenetic Beta Diversity Metrics

Difference between Unweighted- and Weighted UniFrac



https://summer-inequality.blogspot.com/2020/02/robert-edgar-unifrac-distance_5.html

Phylogenetic Beta Diversity Metrics

GUniFrac

→ Capture the variability of taxa that have middle abundances.

$$d_G^{(\alpha)} = \frac{\sum_{i=1}^{m_b} b_i \left(p_i^A + p_i^B\right)^{\alpha} \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{m_b} b_i \left(p_i^A + p_i^B\right)^{\alpha}},$$

extra parameter α is used to control the weight on abundant lineages.

Beta diversity metrics - Summary

	Qualitative (presence/absence)	Quantitative
Non-phylogenetic	Jaccard	Bray-Curtis
Phylogenetic	Unweighted UniFrac	Weighted UniFrac

Beta diversity metrics - Summary

Different metrics tell different things…

Unweighted Metrics (“Qualitative”)	Weighted Metrics (“Quantitative”)
Presence/Absence of OTU	Considers relative abundance (composition)
More sensitive to rare OTUs	More sensitive to abundant taxa
Jaccard, Unweighted UniFrac	Bray-Curtis, Weighted UniFrac

Beta diversity metrics - Summary

Different metrics tell different things…

Non-Phylogenetic	Phylogenetic
Assume everything is equally dissimilar	Take into account similarity based on shared evolution
More likely to see differences based on close relatives	Better for scaling the differences which are seen
Shannon, Bray-Curtis, Jaccard,...	UniFrac

Beta diversity metrics - Summary

FAQ

What is the best distance metric?

- Different metrics show different properties of the data
- No one single metric is better than the rest

Can other metrics be used?

- Absolutely

Ordination method

Visualize the distances

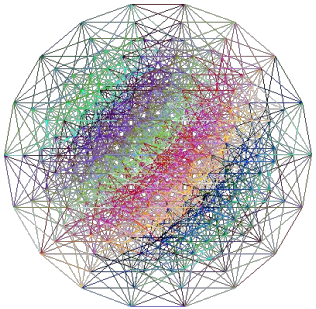
Ordination

- Microbiome sequencing datasets are high-dimensional,
- Large number of taxa - low number of samples
- Hard to explain the differences

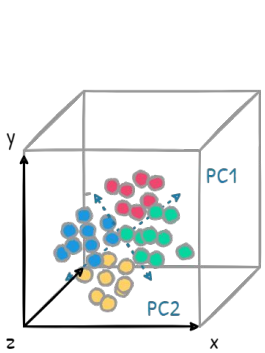
→ Reduce the high dimensionality of microbial taxa (PCA, PCoA, NMDS)

Ordination

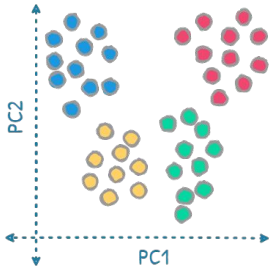
"Ordination" ~ setting in order



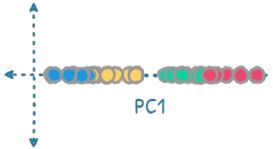
Multi-dimension



3D



2D



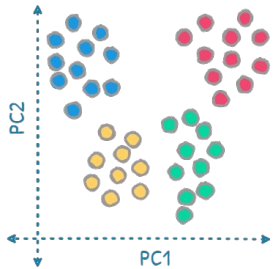
1D

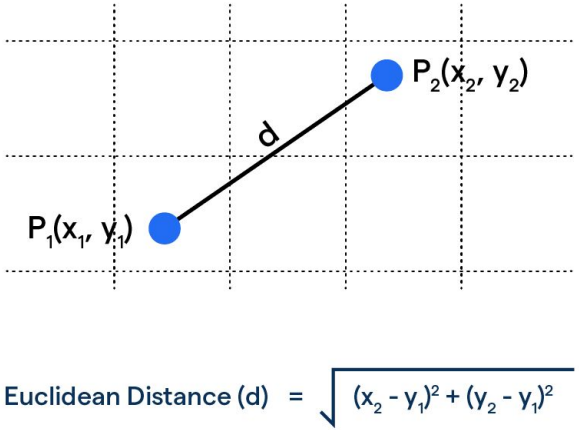
Ordination projects the multidimensional scatter diagram onto bivariate graphs whose axes are known to be of particular interest.

Ordination

Pricipal Component Analysis (PCA)

- Reduce the dimensionality of a dataset while preserving as much variance as possible
- Based on Euclidean distances

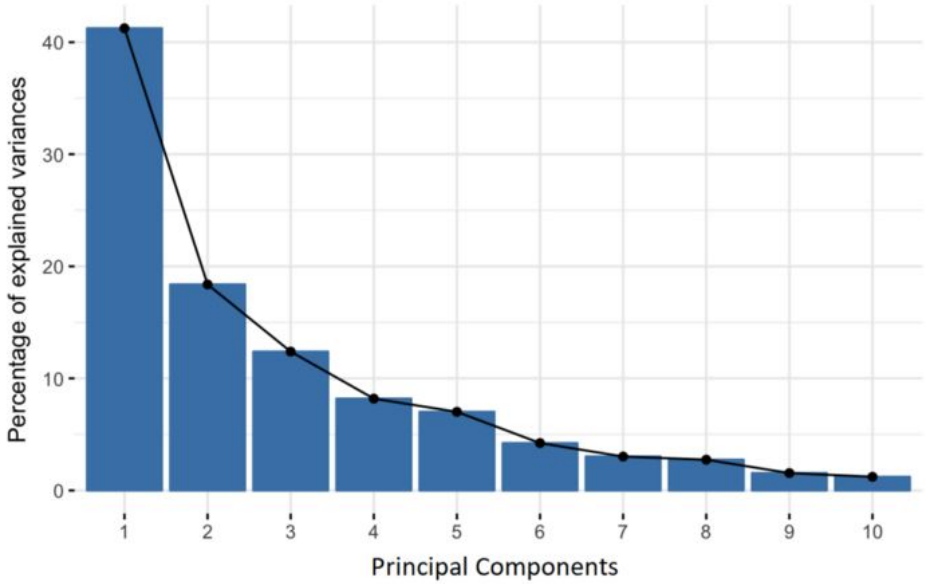




- Commonly used in data visualization, noise reduction, feature extraction, and exploratory data analysis.

Ordination

Pricipal Component Analysis (PCA)

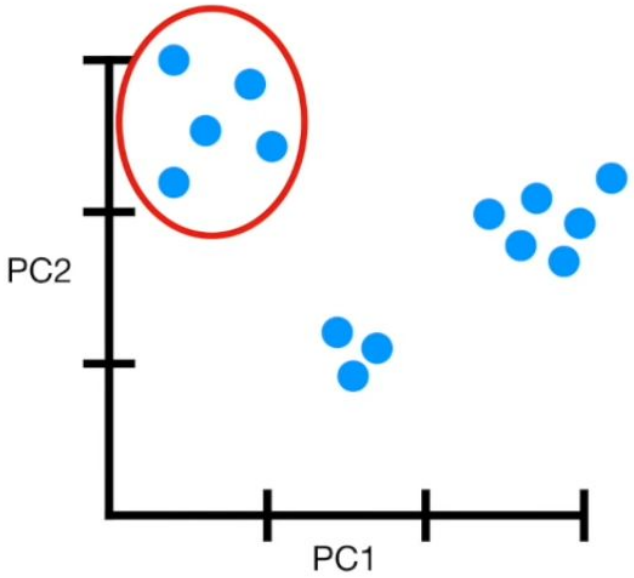


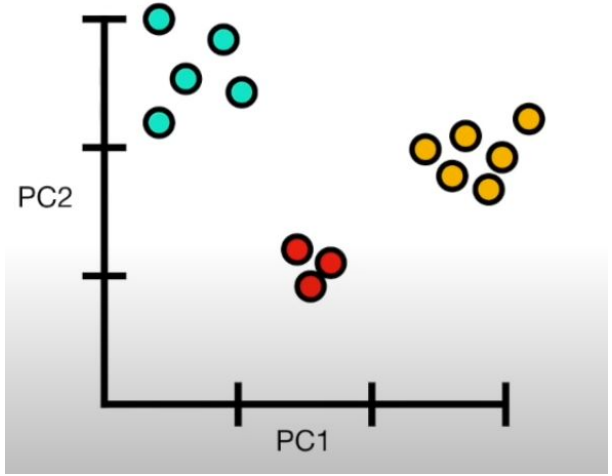
Choose the PCs which are best explained variances

Ordination

Pricipal Component Analysis (PCA)

A PCA plot converts the correlations among all of the samples into a 2-D graph





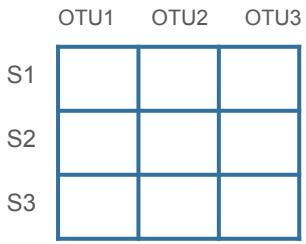
This cluster of samples are highly correlated to each other…

Ordination

Principal coordinate analysis (PCoA)

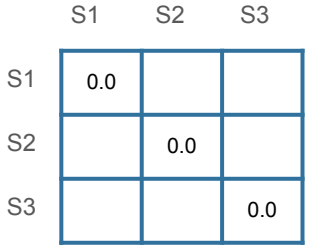
- Conceptual extension of PCA
- Metric (multidimensional) scaling method
- Very similar to PCA, except that instead of converting correlations into a 2-D graph, they convert distances among samples into a 2-D graph.

Workflow



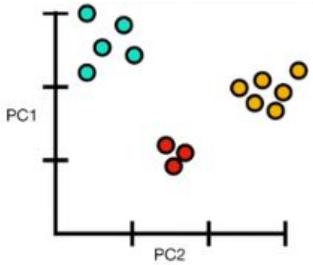
Feature table

- *OTU*
- *ASV*



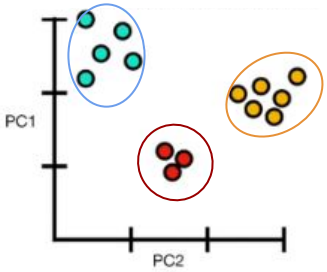
Distance matrix

- *Bray-Curtis*
- *Jaccard*
- *UniFrac*
- *…*



Ordination

- *PCoA*
- *NMDS*
- *…*

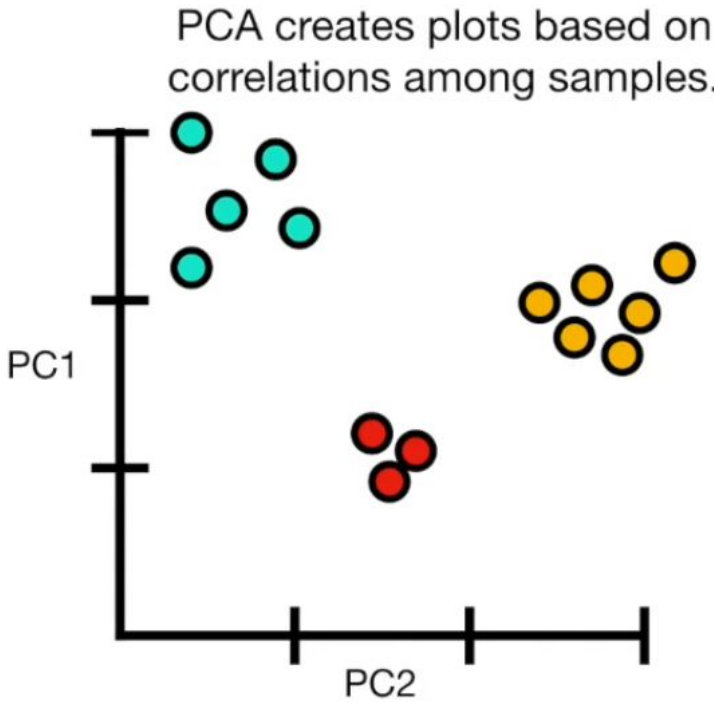


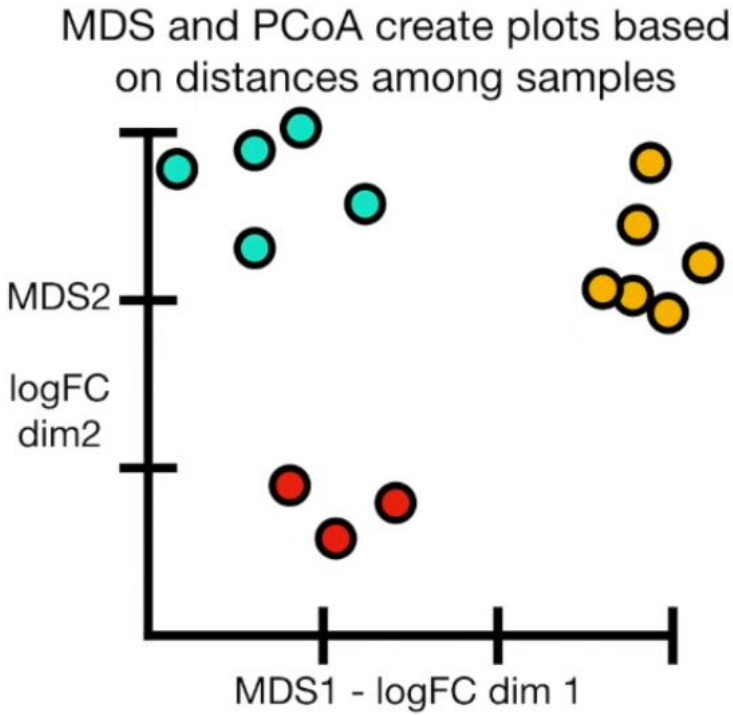
Statistical test

- *PERMANOVA*
- *ANOSIM*

Ordination

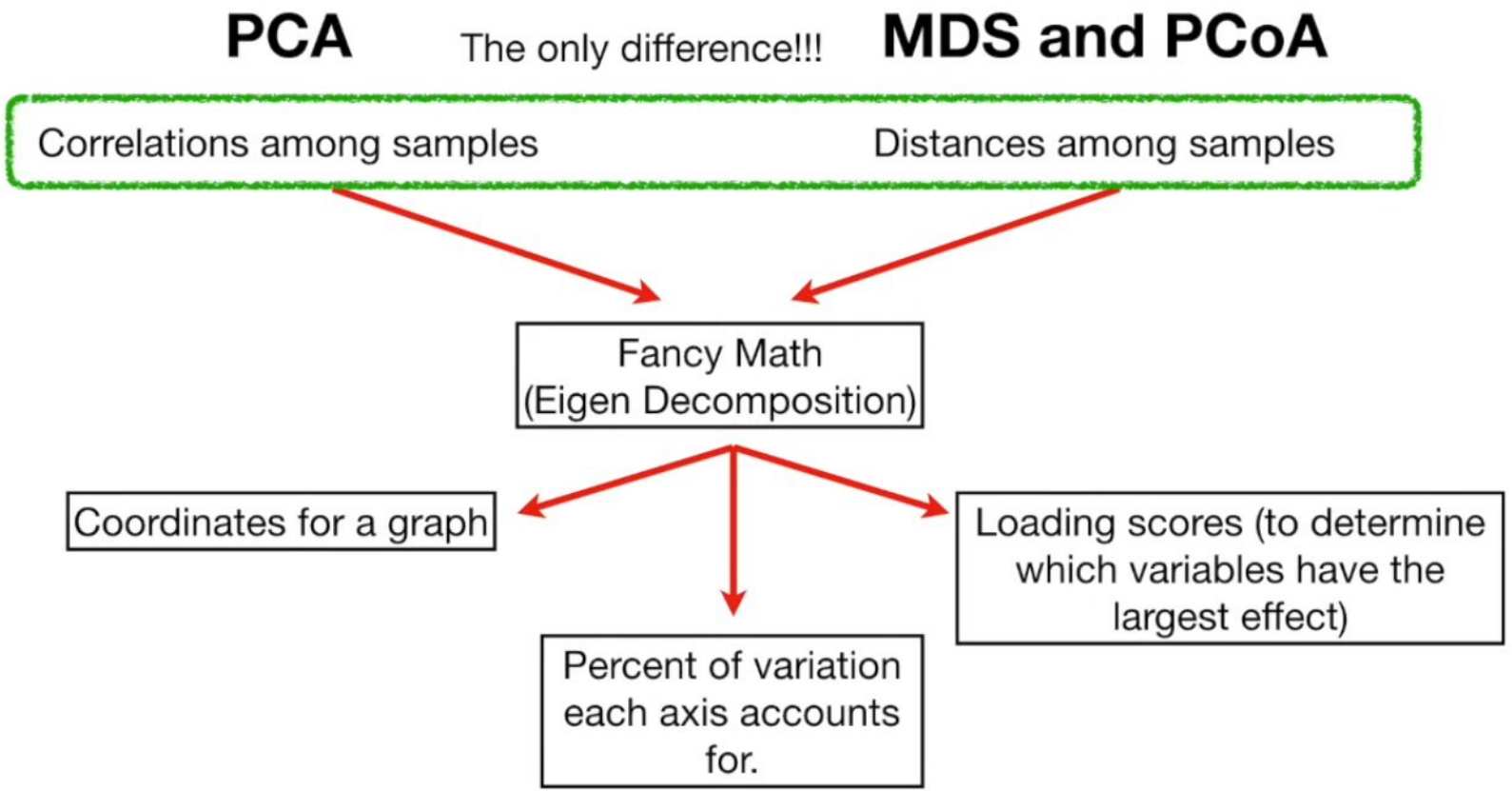
Principal coordinate analysis (PCoA)





https://www.youtube.com/watch?v=GEn-_dAyYME

Ordination



https://www.youtube.com/watch?v=GEn-_dAyYME

Practice 1

Beta diversity metrics calculation and ordination in QIIME2 with Parkinson's Disease Mouse dataset [Link]

Practice 1

Using QIIME2 to calculate and visualize the Beta diversity metrics of this dataset

sample_name #q2:types	barcode categorical	mouse_id categorical	genotype categorical	cage_id categorical	donor categorical	donor_status categorical	days_post_transplant numeric	genotype_and_donor_status categorical
recip.220.WT.OB1.D7	CCTCCGTCATGG	457	wild type	C35	hc_1	Healthy	49	wild type and Healthy
recip.290.ASO.OB2.D1	AACAGTAAACAA	456	susceptible	C35	hc_1	Healthy	49	susceptible and Healthy
recip.389.WT.HC2.D21	ATGTATCAATTA	435	susceptible	C31	hc_1	Healthy	21	susceptible and Healthy
recip.391.ASO.PD2.D14	GTCAGTATGGCT	435	susceptible	C31	hc_1	Healthy	14	susceptible and Healthy
recip.391.ASO.PD2.D21	AGACAGTAGGAG	437	susceptible	C31	hc_1	Healthy	21	susceptible and Healthy
recip.391.ASO.PD2.D7	GGTCTTAGCACC	435	susceptible	C31	hc_1	Healthy	7	susceptible and Healthy
recip.400.ASO.HC2.D14	CGTTCGCTAGCC	437	susceptible	C31	hc_1	Healthy	14	susceptible and Healthy
recip.401.ASO.HC2.D7	ATTTACAATTGA	437	susceptible	C31	hc_1	Healthy	7	susceptible and Healthy
recip.403.ASO.PD2.D21	CGCAGATTAGTA	456	susceptible	C35	hc_1	Healthy	21	susceptible and Healthy
recip.411.ASO.HC2.D14	ATGTTAGGGAAT	456	susceptible	C35	hc_1	Healthy	14	susceptible and Healthy
recip.411.ASO.HC2.D21	CTCATATGCTAT	457	wild type	C35	hc_1	Healthy	21	wild type and Healthy

Statistical testing of Beta diversity with PERMANOVA *

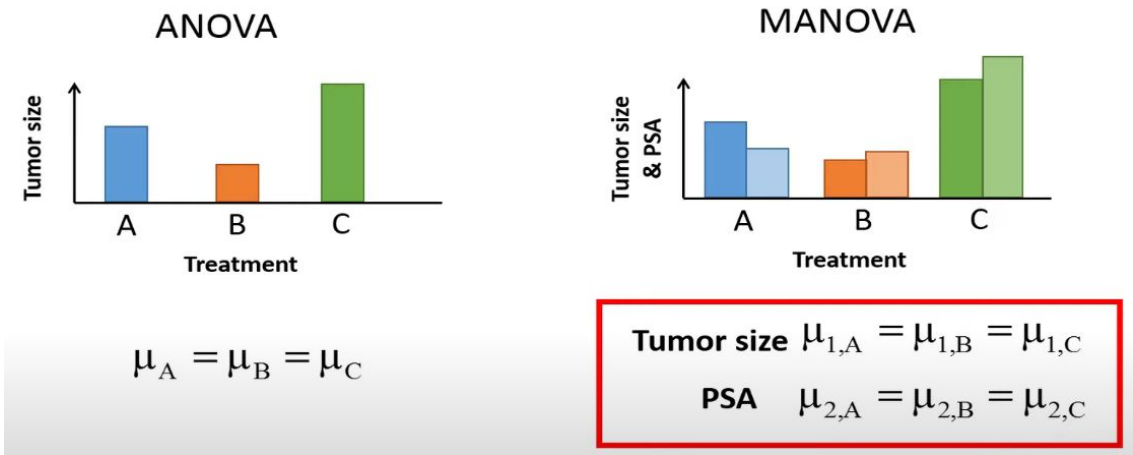
(*) Authorized by Minh Nguyen

In microbiome research, PERMANOVA is generally the preferred method for testing beta diversity differences due to its robustness and flexibility.

What is MANOVA ?

- Multivariate ANOVA (MANOVA)
- An extension to univariate ANOVA
- Includes at least ***two*** dependent variables to analyze the differences between ***multiple groups*** in the independent variable.
- MANOVA compares the *vectors* containing the group mean of each dependent variable.

- MANOVA maximises the discrimination in *between-groups* than within-groups.



https://www.r-bloggers.com/2021/11/manovamultivariate-analysis-of-variance-using-r/

Assumptions of MANOVA

MANOVA relies on the following assumptions:

1. Independent groups
2. Multivariate normality:
    - Multivariate normal distribution for each combination of independent and dependent variables.
3. Homogeneity of the variance-covariance matrices:
    - Equal variance-covariance matrices for each combination formed by each group in the independent variable (Box's M test).
4. No multicollinearity among dependent variables:
    - Not too strong linear correlation between the dependent variables.
5. Linear relationship of the dependent variables for each group

MANOVA Hypotheses

- Null hypothesis (H_0): group mean vectors are the same for all groups.
- Alternative hypothesis (H_a): group mean vectors are not same for all groups.

Permutational MANOVA (PERMANOVA)

Introduction to PERMANOVA

- Non parametric MANOVA: permutational multivariate analysis of variance (*do not assume any distribution*).
- *General multivariate* null hypothesis (H_0) of no differences in the composition and/or relative abundances of different species (taxa/variables) between *two or more groups*.
- Based on any measure of dissimilarity
- Test statistic: multivariate analogue to **Fisher's *F-ratio***.
- **P-values**: using *permutations*.

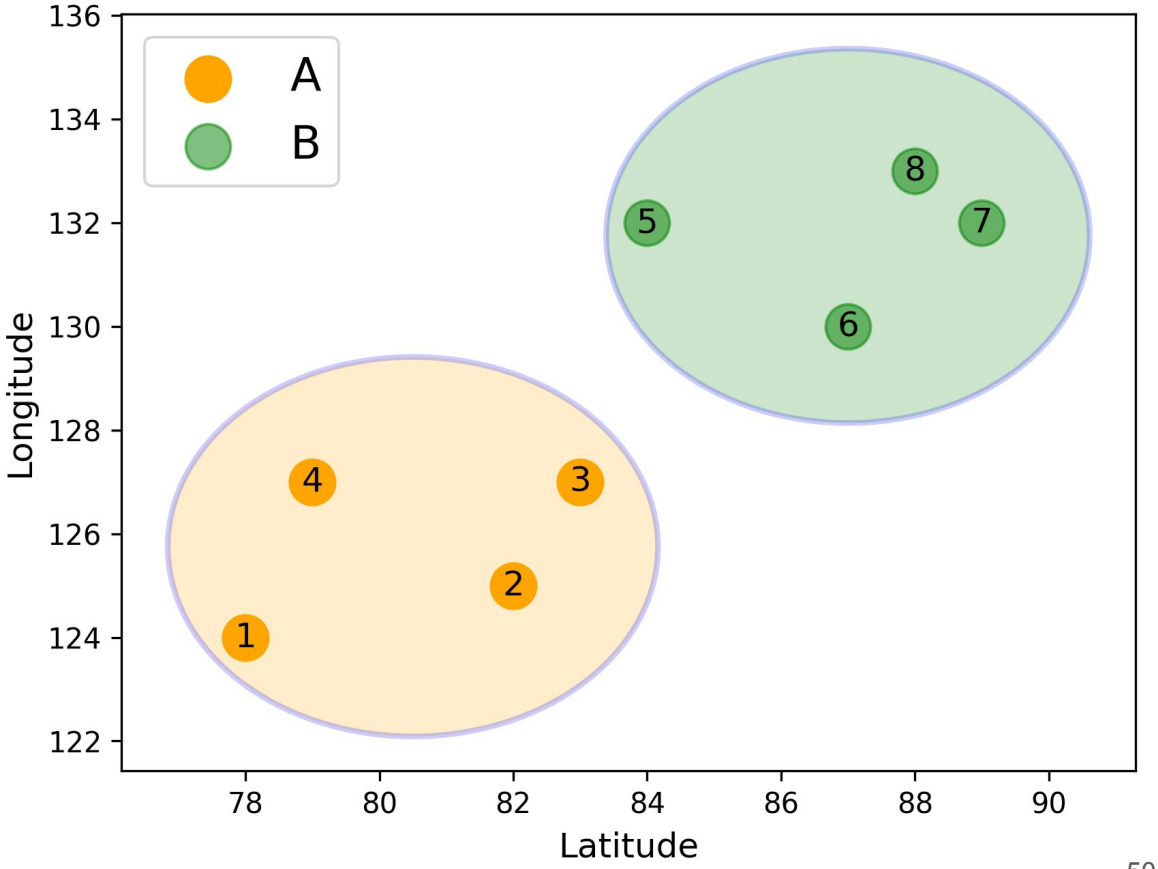
Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26: 32–46

Application of PERMANOVA

- PERMANOVA can be used if we do not fulfill the underlying assumptions in MANOVA.
- PERMANOVA still assumes that the different groups have about the same spread (dispersion) in the data, quite robust to differences in spread, especially if the groups have equal sample sizes.
- PERMANOVA is often used on count data, which is usually highly skewed and might contain lots of zeros.

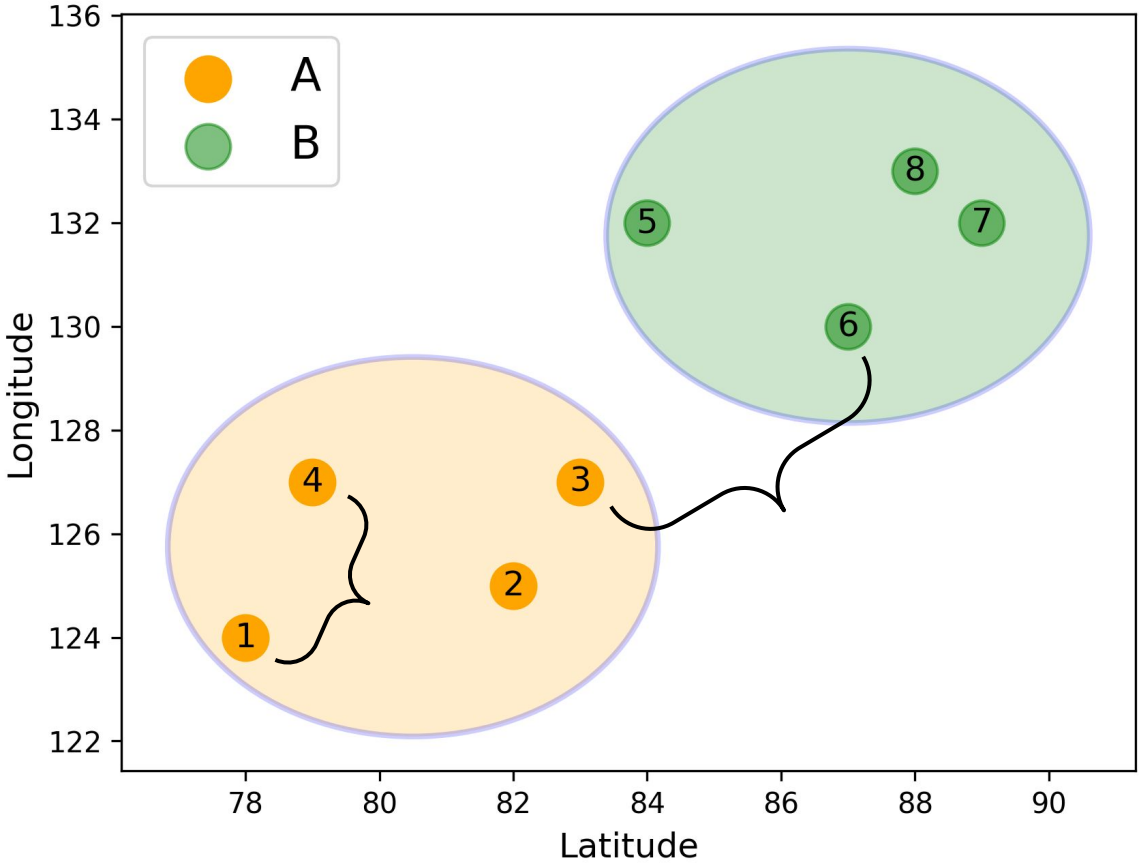
PERMANOVA

Longitude	Latitude	Group
124	78	A
125	82	A
127	83	A
127	79	A
132	84	B
130	87	B
132	89	B
133	88	B



PERMANOVA

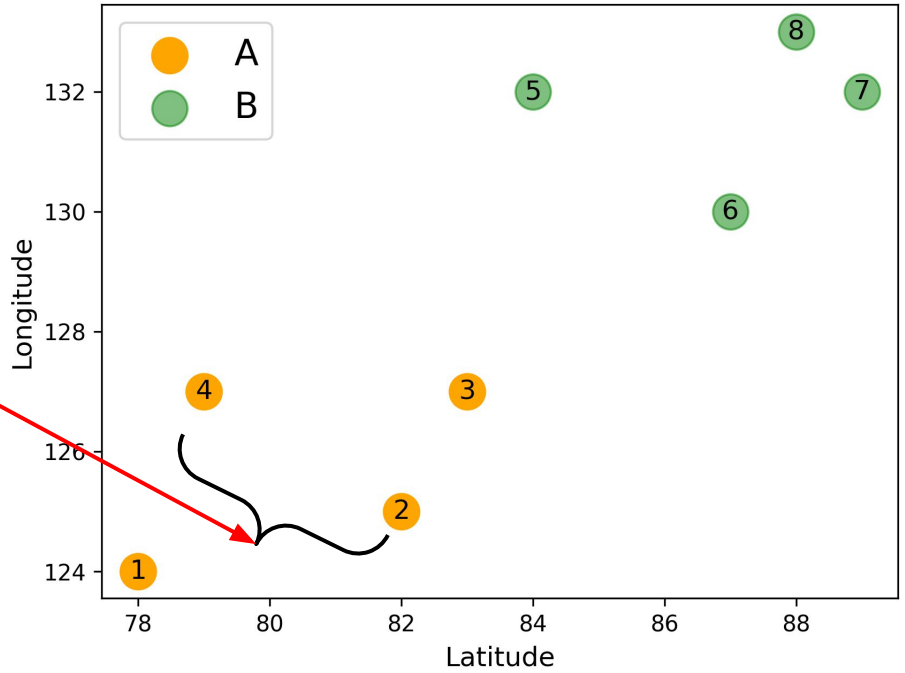
Longitude	Latitude	Group
124	78	A
125	82	A
127	83	A
127	79	A
132	84	B
130	87	B
132	89	B
133	88	B



$$d = \sqrt{(83-87)^2 + (127-130)^2} = 5$$

PERMANOVA: Distance matrix

	1	2	3	4	5	6	7	8
1	0	4.1	5.8	3.2	10.0	10.8	13.6	13.5
2		0	2.2	3.6	7.3	7.1	9.9	10.0
3			0	4.0	5.1	5.0	7.8	7.8
4				0	7.1	8.5	11.2	10.8
5					0	3.6	5.0	4.1
6						0	2.8	3.2
7							0	1.4
8								0



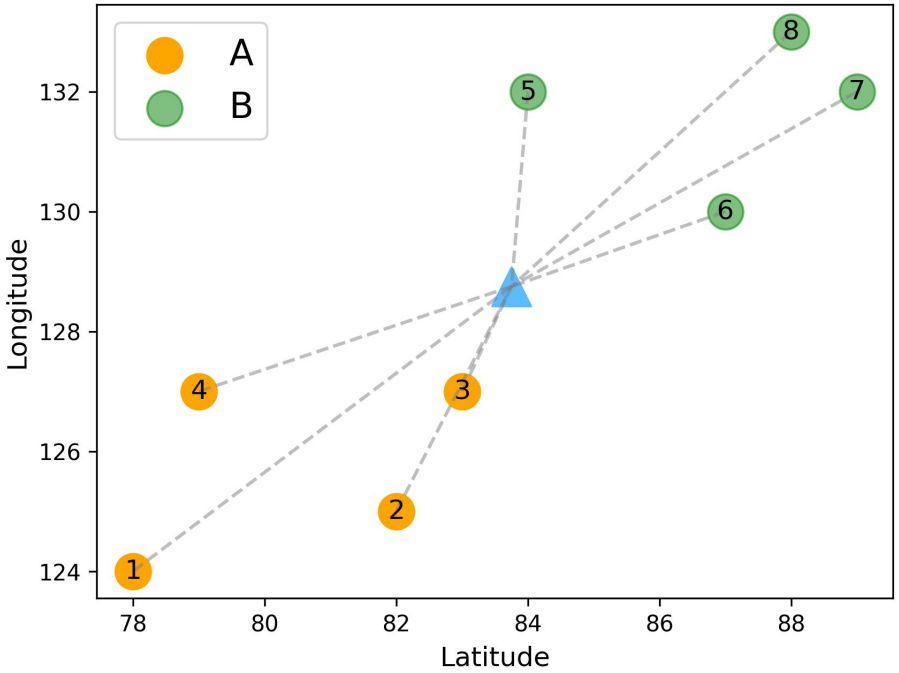
$$SS_T = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2$$

A matrix of distances between every pair of observations

- N: total number of observations (points)
- d_{ij}: the distance between observation; i = 1, …, N; j = 1, …, N
- SS_T: total sum of squares

PERMANOVA: Total sum of square (SST)

	1	2	3	4	5	6	7	8
1	0	4.1	5.8	3.2	10.0	10.8	13.6	13.5
2		0	2.2	3.6	7.3	7.1	9.9	10.0
3			0	4.0	5.1	5.0	7.8	7.8
4				0	7.1	8.5	11.2	10.8
5					0	3.6	5.0	4.1
6						0	2.8	3.2
7							0	1.4
8								0

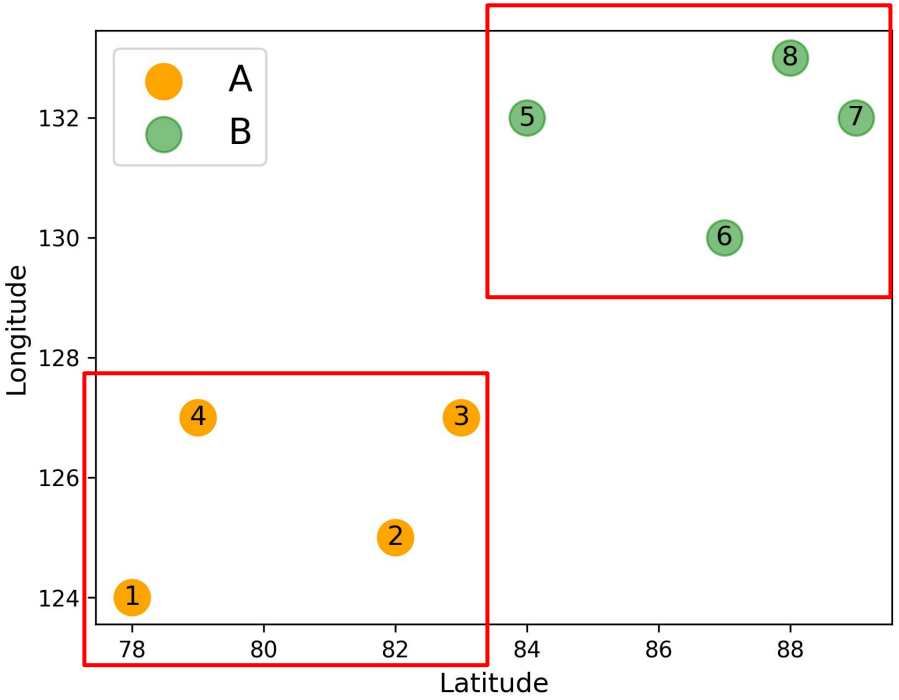


$$SST = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 = 199$$

$$SST = \sum_{i=1}^{N} d_{i,centroid}^2 = 199$$

PERMANOVA: Within-group/residual sum of square (SSW)

	1	2	3	4	5	6	7	8
1	0	4.1	5.8	3.2	10.0	10.8	13.6	13.5
2		0	2.2	3.6	7.3	7.1	9.9	10.0
3			0	4.0	5.1	5.0	7.8	7.8
4				0	7.1	8.5	11.2	10.8
5					0	3.6	5.0	4.1
6						0	2.8	3.2
7							0	1.4
8								0



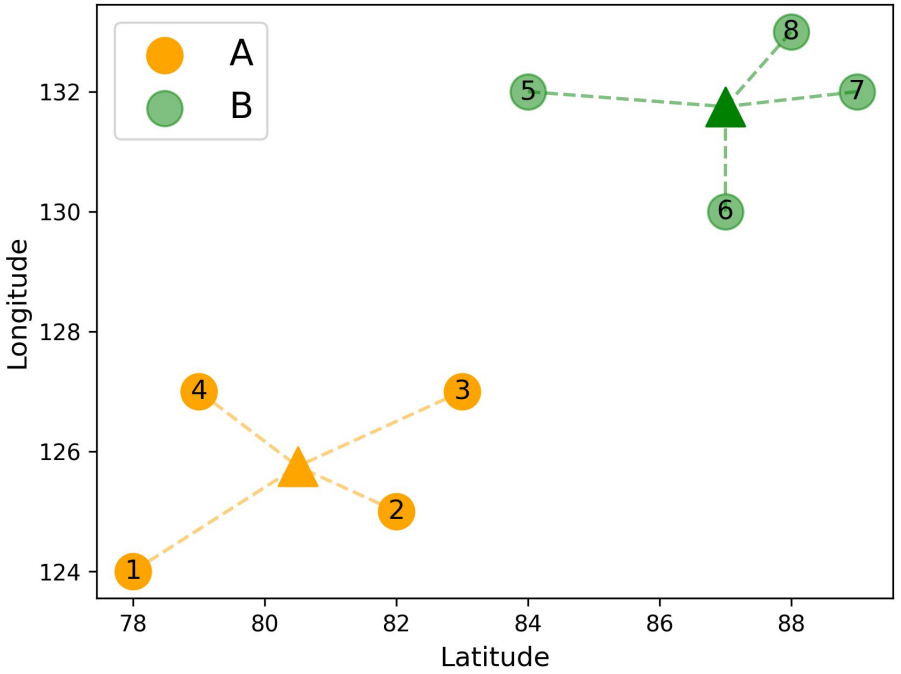
$$SST = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 = 199$$

$$SSW = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \delta_{ij} = 42.5$$

δ_{ij}: 1 if observation i and j are in the same group, otherwise 0

PERMANOVA: Within-group/residual sum of square (SSW)

	1	2	3	4	5	6	7	8
1	0	4.1	5.8	3.2	10.0	10.8	13.6	13.5
2		0	2.2	3.6	7.3	7.1	9.9	10.0
3			0	4.0	5.1	5.0	7.8	7.8
4				0	7.1	8.5	11.2	10.8
5					0	3.6	5.0	4.1
6						0	2.8	3.2
7							0	1.4
8								0



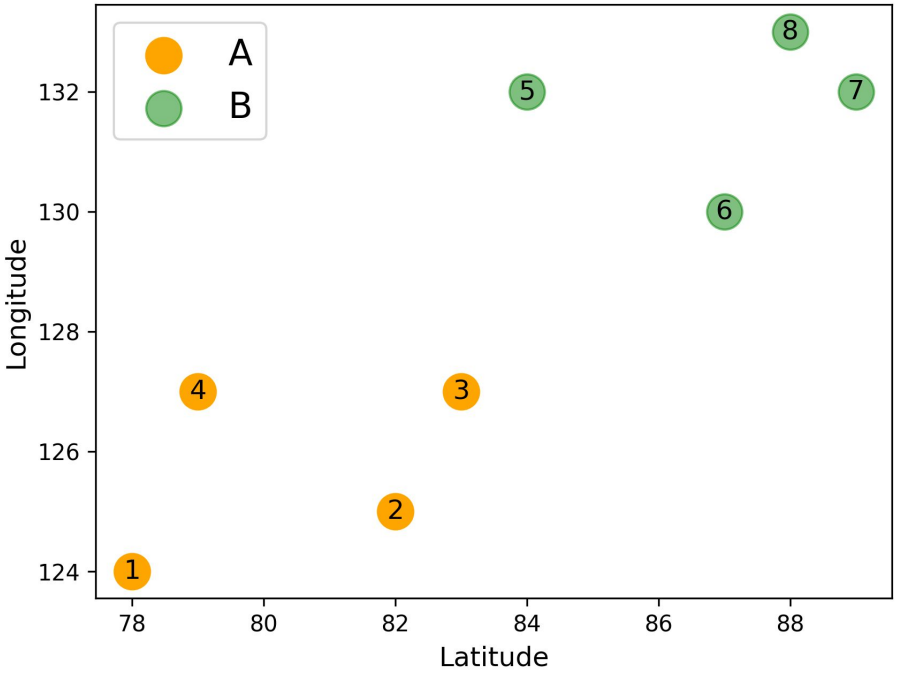
$$SST = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 = 199$$

$$SSW = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \delta_{ij} = 42.5$$

$$SSW = \sum_{k=1}^{p}\sum_{i=1}^{n} d_{i,centroid_k}^2 = 42.5$$

PERMANOVA: Between-group SS and F-ratio

	1	2	3	4	5	6	7	8
1	0	4.1	5.8	3.2	10.0	10.8	13.6	13.5
2		0	2.2	3.6	7.3	7.1	9.9	10.0
3			0	4.0	5.1	5.0	7.8	7.8
4				0	7.1	8.5	11.2	10.8
5					0	3.6	5.0	4.1
6						0	2.8	3.2
7							0	1.4
8								0



$$SST = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 = 199$$

$$SSW = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \delta_{ij} = 42.5$$

$$SS_A = SS_T - SS_W =$$

$$199 - 42.5 = 156.5$$

$$F = \frac{Variance\ between\ groups}{Variance\ within\ groups} = \frac{SS_A/(a-1)}{SS_W/(N-a)}$$

$$= \frac{156.5/(2-1)}{42.5/(8-2)} = 22.094$$

PERMANOVA: F-statistics based on the permutation

Longitude	Latitude	Group
124	78	A
125	82	A
127	83	A
127	79	A
132	84	B
130	87	B
132	89	B
133	88	B

Permu

Longitude	Latitude	Group
124	78	A
125	82	B
127	83	B
127	79	A
132	84	B
130	87	A
132	89	A
133	88	B

F-ratio

$F_{approx} = 22.094$

$F_{p1} = 3.47$

$F_{p2} = 5.65$

$F_{p3} = 22.094$

.

.

.

$F_{p1000} = 4.12$

PERMANOVA: P-value - indicating significance between groups

- Comparing the F-value obtained with the *original ordering* of the rows to the *empirical distribution* created for a true null by permuting the labeled rows
- The fraction of permuted F-ratios that are greater than the observed F-ratio

$$p = \frac{Number\ of |F_p| \geq |F_{approx}|}{Number\ of\ permutations} = 0.031$$

*Example p-value

Remarks on PERMANOVA

- Non-parametric framework, multifactorial analysis of variance of ecological and microbiome data.
- Non-parametric method without assumption of multivariate normality.
- Partitioning variation according to any ANOVA design.
- Any distance measure (or on ranks of distances) appropriate for the data and hypothesis being tested.
- Test statistic: ~ Fisher’s F-ratio, constructed from sums of squared distances (dissimilarities) within and between groups
- P-value: appropriate permutation methods.

Pairwise PERMANOVA

Introduction to Pairwise Permutation MANOVA

- In ANOVA and also PERMANOVA, if a null hypothesis of no difference among groups is rejected, then it suggests that there is a significant difference among the defined groups.
- However there is no way to know which groups are significantly separated.
- **Pairwise comparison** using an appropriate test and a statistical method to adjust the P-values for multiple comparisons:
  - Permutation test
  - Tukey’s method
  - ...

Pairwise PERMANOVA model

- RVAideMemoire Package
- Specifically designed statistical tools for pairwise comparisons of each group level with corrections for multiple testing after implementing permutational MANOVA.
- **pairwise.perm.manova()** function

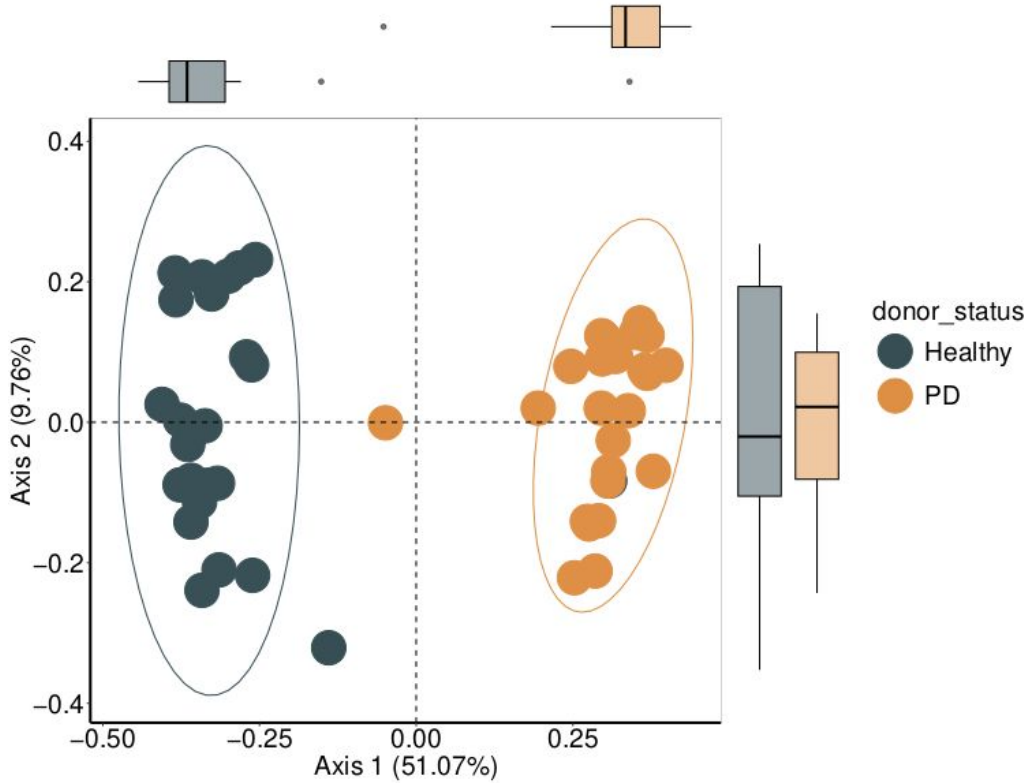
Hervé, Maxime. 2022. RVAideMemoire: Testing and plotting procedures for biostatistics. R package version 0.9-81-2

Practice 2

Ordination and Conducting statistical test for Beta diversity using R [link]

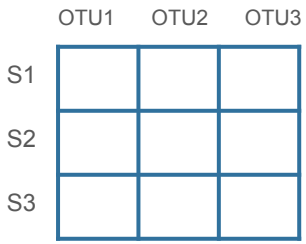
Practice 2

Use R to calculate the distance and make ordination plot for PD Mouse dataset



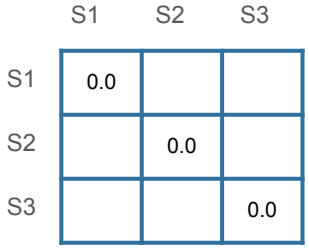
Is there any differences in microbiome community between 2 donor groups?

Workflow



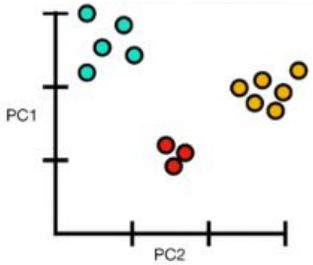
Feature table

- *OTU*
- *ASV*



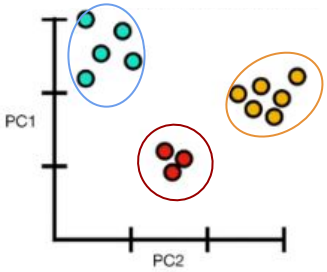
Distance matrix

- *Bray-Curtis*
- *Jaccard*
- *UniFrac*
- *…*



Ordination

- *PCoA*
- *NMDS*
- *…*



Statistical test

- *PERMANOVA*
- *ANOSIM*

THANK YOU