

# Microbiome diversity analysis

Analysis of alpha and beta diversity using phyloseq R package

2024.04.24. Phytobiome symposium

**Session 2**

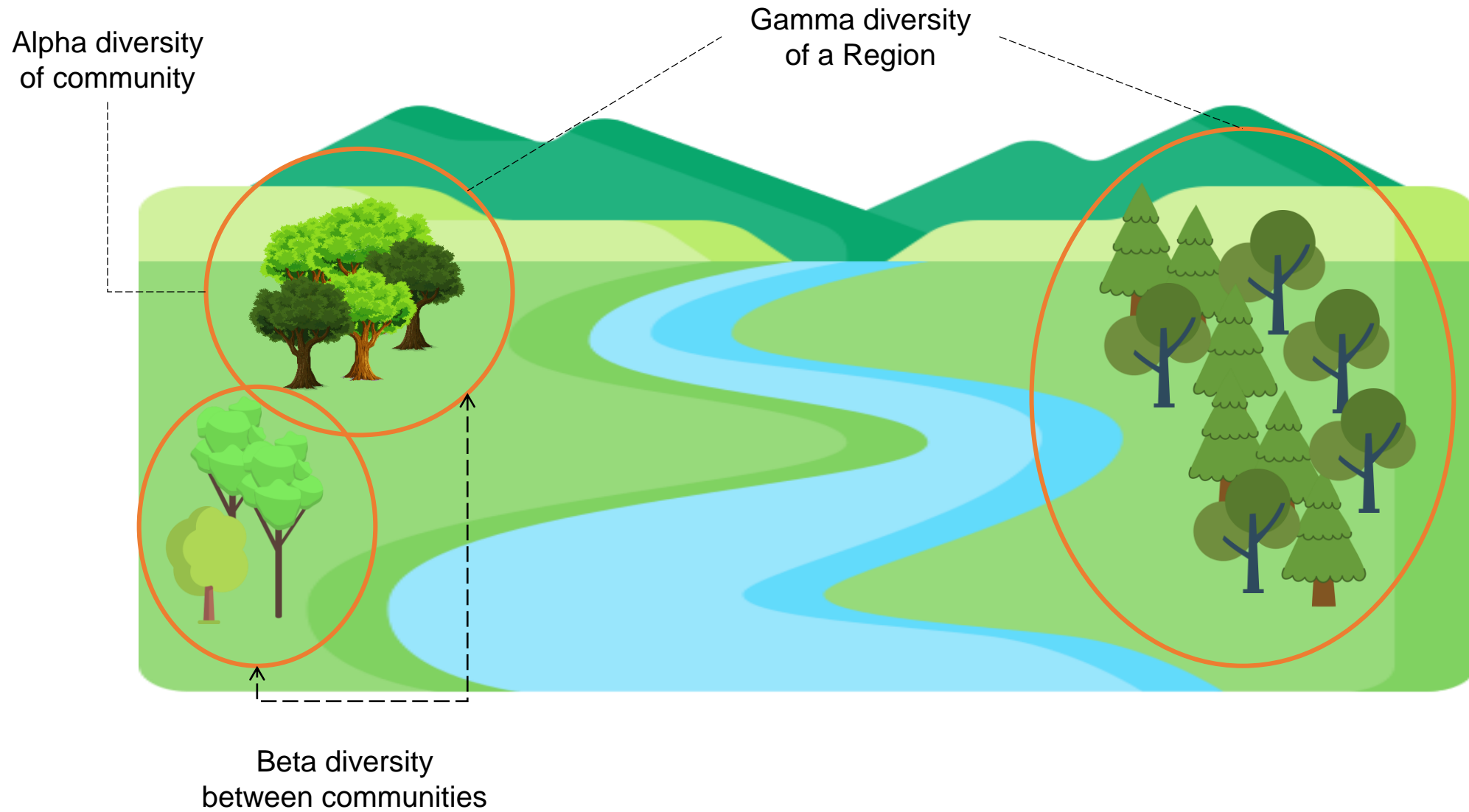
**Dong-A University Master course**

**Dongmin Lee**

# **Table of Contents**

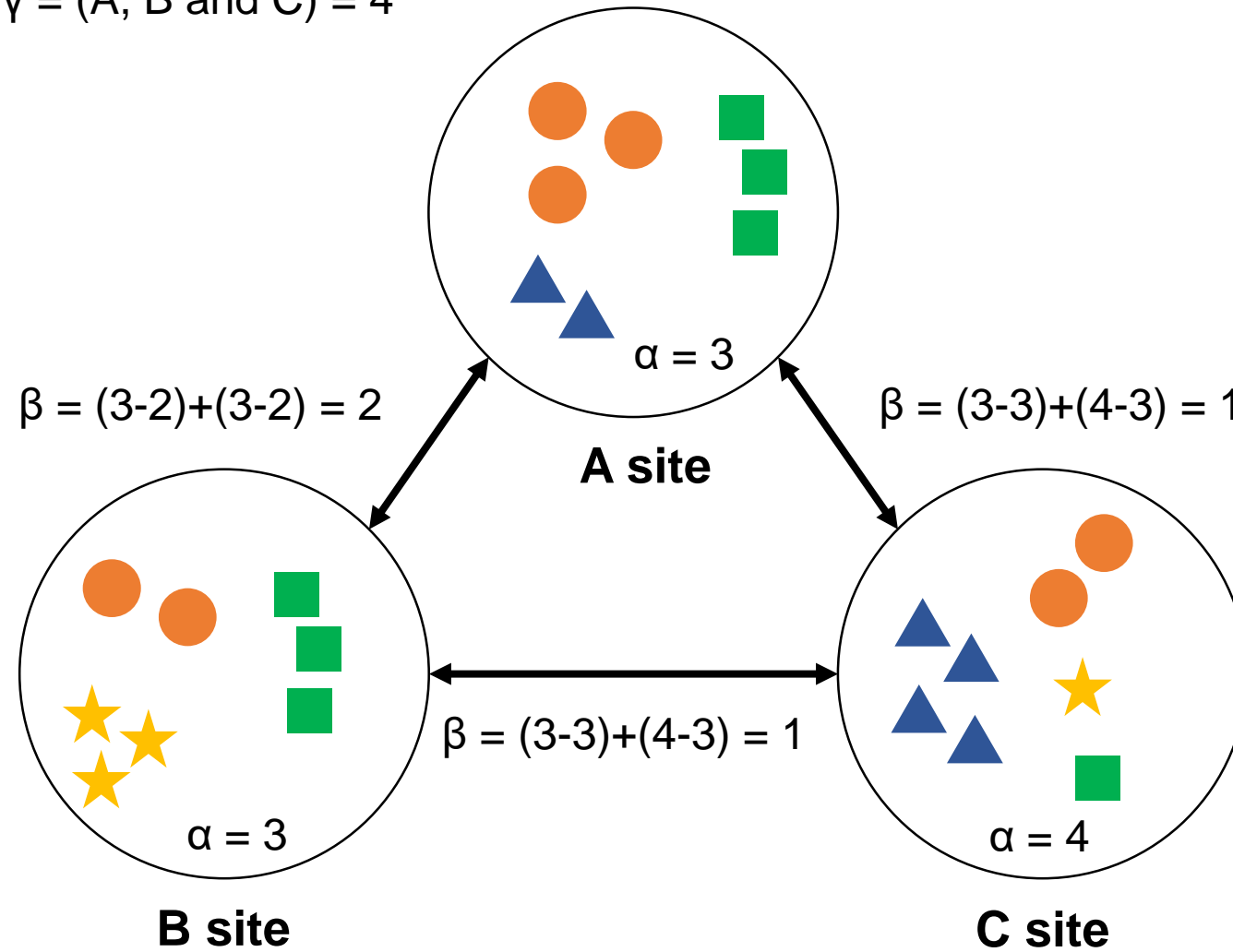
1. Diversity
2. Alpha diversity
3. Rarefaction
4. Beta diversity
5. Ordination
6. Analysis using phyloseq R package

# 1. Diversity



# 1. Diversity

$$\gamma = (A, B \text{ and } C) = 4$$



## 2. Alpha diversity

---

- **Within-sample diversity**
- **Two components of alpha diversity**
  - Richness (number of taxonomic groups)
  - Evenness (distribution of abundances of the groups)

### 2-1. Observed OTU or Taxa (Richness)

- Count of different species/OTUs

### 2-2. Shannon's diversity Index

- $p_i$ : ratio of  $i$  species,  $S$ : number of speices in a sample
- Consider both species richness and evenness
- Associated with entropy concept (Shannon diversity)
- Quantify the uncertainty in predicting the taxa identity of an individual selected at random from the sample

$$H = - \sum_{i=1}^S p_i \ln p_i$$

## 2. Alpha diversity

---

### 2-3. Simpson's index

$$D = \sum_{i=1}^s p_i^2$$

$$\textit{Gini-Simpson} = 1 - D$$

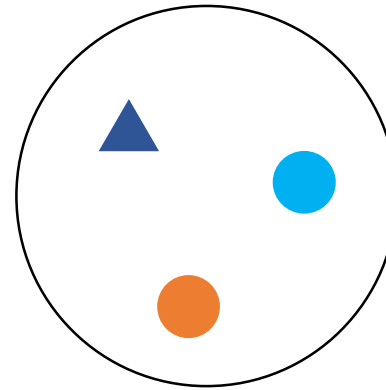
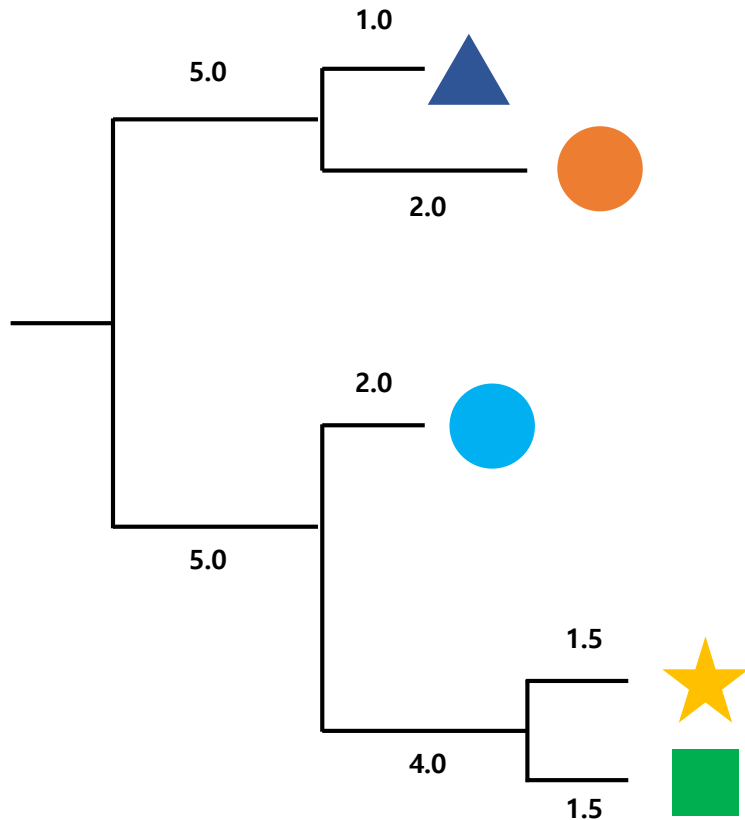
$$\textit{Inverse Simpson} = \frac{1}{D}$$

- $p_i$ : ratio of  $i$  species
- Probability that two individuals belong to the same taxa randomly chosen
- Higher  $D$  value suggest that community has low diversity
- More sensitive to evenness than richness
- So called, Simpson's evenness index.

## 2. Alpha diversity

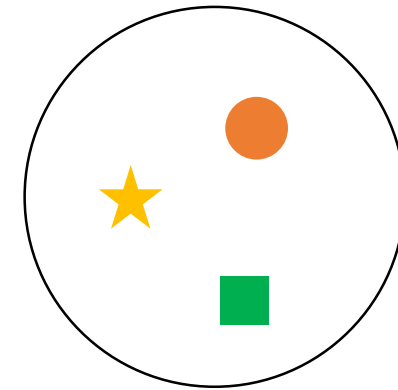
### 2-4. Faith's phylogenetic diversity

- Sum of branch length between observed species in the phylogenetic tree



**Faith's PD**

$$= 5 + 1 + 2 + 7$$
$$= 15$$



**Faith's PD**

$$= 5 + 2 + 5 + 4 + 1.5 + 1.5$$
$$= 19$$

### 3. Rarefaction

- Low sampling depth vs High sampling depth
- Briefly, it is unfair
- Amplicon sequencing data is not total DNA sequence data
- Randomly choose observations without replacement from our samples up to a specific depth

	Sample 1	Sample 2			Sample 1	Sample 2
<b>Species 1</b>	148	36	→ Rarefaction	<b>Species 1</b>	34	36
<b>Species 2</b>	53	10		<b>Species 2</b>	23	10
<b>Species 3</b>	2	4		<b>Species 3</b>	1	4
<b>Species 4</b>	48	24		<b>Species 4</b>	18	24
<b>Species 5</b>	6	3		<b>Species 5</b>	1	3
<b>Total</b>	257	77		<b>Total</b>	77	77

Is it fair to compare samples with different sequencing depth?

Can say that the Species 3 exists more in Sample 2 than in Sample 1?



## 4. Beta diversity

---

- **Between-sample diversity**
- **How different is it between different communities**

### 4-1. Bray-Curtis Dissimilarity

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

- Measure the compositional dissimilarity between the communities of two samples
- $C_{ij}$  is sum of the lesser values for given taxa in common between sample i and j
- $S_i$  and  $S_j$  are the total number of taxa counted in i and j, respectively
- Ranges between 0 (two samples share all taxa) and 1 (two samples do not share any taxa)
- Computed pairwise between all samples

## 4. Beta diversity

---

### 4-2. Jaccard distance

- Based on presence or absence of species
- Ratio between the number of members that are common between the two samples and the number of members that are distinct
- Ranges between 0 (the communities are identical) and 1 (the two communities are different)

$$\textit{Jaccard coefficient} \quad J(i, j) = \frac{|i \cap j|}{|i \cup j|}$$

$$\textit{Jaccard distance} \quad Jd(i, j) = 1 - J(i, j)$$

## 4. Beta diversity

### 4-3. UniFrac distance

- Based on phylogenetic tree, UniFrac distance is calculated as sum of the branch length
- Measure community dissimilarity based on the presence or absence of branch

#### -> Unweighted UniFrac

$$UU(A, B) = \frac{\text{sum of unique branch length}}{\text{sum of observed branch length}} = \text{fraction of total unshared branch lengths}$$

- Consider the abundance of sequences

#### -> Weighted UniFrac

$$WU(A, B) = \sum_i^n b_i * \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$

- n is the total number of branches in the tree
- $b_i$  is the length of branch i
- $A_i$  and  $B_i$  are the numbers of sequences that descend from branch i in A and B
- $A_T$  and  $B_T$  are the total numbers of sequences in A and B

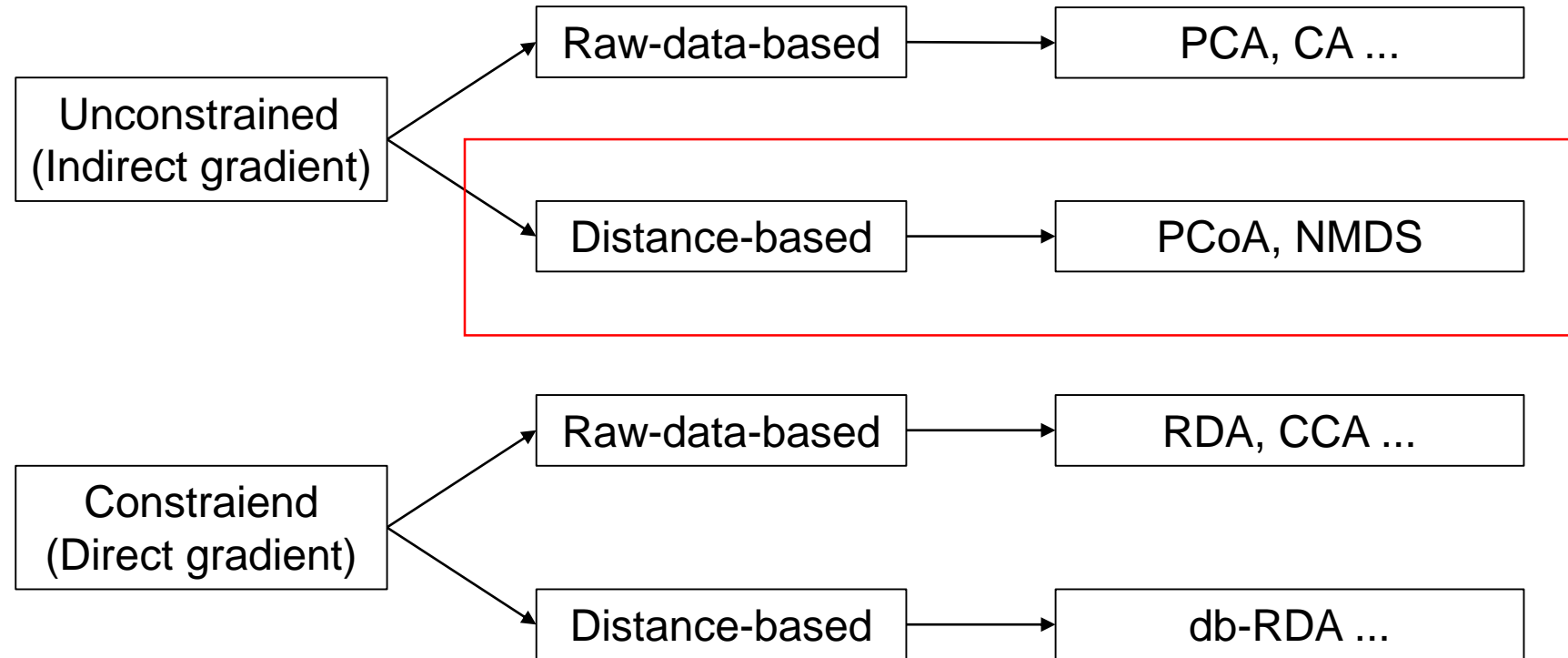
## 4. Beta diversity

### 4-4. Summary

	Consider abundance	Not consider abundance (Presence/Absence)
Not consider phylogenetic relationship	Bray-Curtis	Jaccard
Consider phylogenetic relationship	Weighted UniFrac	Unweighted UniFrac

## 5. Ordination

- Statistical method for transforming multidimensional data into a concise and interpretable form for visualization and analysis purposes
- Represent sample and species relationships in a low-dimensional space



# 5. Ordination

---

## Distance-based approach

- Rely on a square, symmetric distance matrix or similarity matrix

- **PCoA (Principal Coordinates Analysis)**

- To describe the data by reducing the dimensions of a distant matrix among objects
- Maximize linear correlation distance measures and distance in the ordination
- Cannot indicate combinations of variables

- **NMDS (Nonmetric Multidimensional Scaling)**

- To describe data by reducing the number of dimensions
- To discover nonlinear relationships
- Maximizing the rank-order correlation between distance measures and distance in ordination space
- Points are moved to minimize: stress (a measure of the mismatch between the two kinds of distance)
- To increase the likelihood of finding the correct solution

**Thank you !**

Department of Applied Biology

Dong-A University, Busan, Republic of Korea

237182@donga.ac.kr