Alpha diversity

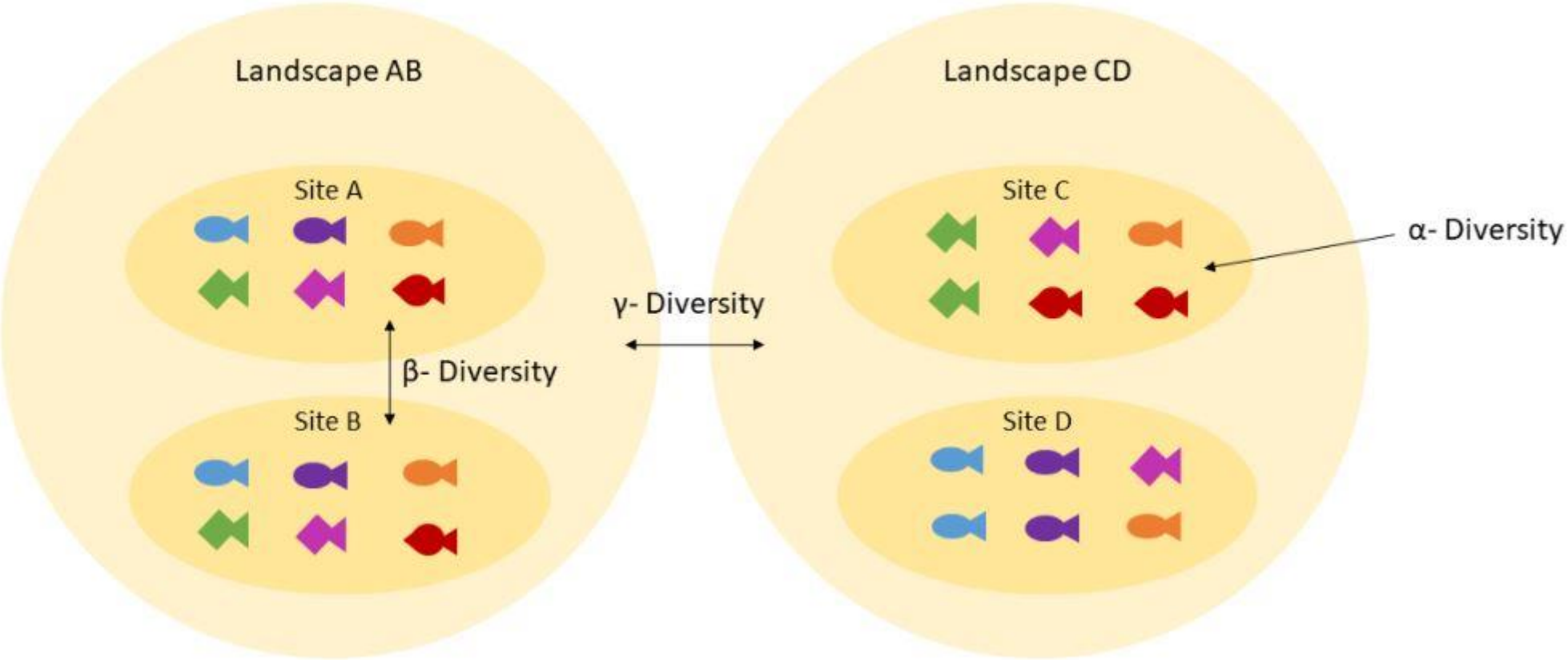
Ho Phu Quy

Part 1: introduction about alpha diversity

Image source:
http://miriadna.com/desctopwalls/images/max/Field-of-yellow-tulips.jpg

Image source:
https://imgflip.com/memetemplate/62338435/Flower-garden

https://www.youtube.com/watch?v=VqcLbKKa7j8&list=PLbVDKwGpb3XnwBACQ5w8L8NzjZ-6D0M8u&index=18



https://www.youtube.com/watch?v=VqcLbKKa7j8&list=PLbVDKwGpb3XnwBACQ5w8L8NzjZ-6D0M8u&index=18

	Sample 1	Sample 2
Species A	10	50
Species B	10	0
Species C	10	3
Species D	10	5
Species E	10	1
Species D	10	25
Species H	10	11

What different between 2 sample?

https://www.youtube.com/watch?v=VqcLbKKa7j8&list=PLbVDKwGpb3XnwBACQ5w8L8NzjZ-6D0M8u&index=18

	Sample 1	Sample 2
Species A	10	50
Species B	10	0
Species C	10	3
Species D	10	5
Species E	10	1
Species D	10	25
Species H	10	11

There are 2 things different between 2 sample here. And i ‘m using 2 question to make it clear:

- First, how many species in each sample?
- Second, how species in each sample distribution?

https://www.youtube.com/watch?v=VqcLbKKa7j8&list=PLbVDKwGpb3XnwBACQ5w8L8NzjZ-6D0M8u&index=18

Richness index

	Sample 1	Sample 2
Species A	1	1
Species B	1	0
Species C	1	1
Species D	1	1
Species E	1	1
Species D	1	1
Species H	1	1

So, we can see in sample 1 the number of species is more than sample 2
In alpha diversity, this calculate is richness, which reflects both the number of species present (species richness)

https://www.youtube.com/watch?v=VqcLbKKa7j8&list=PLbVDKwGpb3XnwBACQ5w8L8NzjZ-6D0M8u&index=18

	Sample 1	Sample 2
Species A	10	50
Species B	10	0
Species C	10	3
Species D	10	5
Species E	10	1
Species D	10	25
Species H	10	11

About distribution, we can see in the sample 1 the have the same number of species, but in the sample 2 we can see the number of each species seem have big different. In here, we call the different like that is the species of evenness, which distribute the number of organisms per species.

https://www.youtube.com/watch?v=VqcLbKKa7j8&list=PLbVDKwGpb3XnwBACQ5w8L8NzjZ-6D0M8u&index=18

	Sample 1	Sample 2
Species A	10	50
Species B	10	0
Species C	10	3
Species D	10	5
Species E	10	1
Species D	10	25
Species H	10	11

With both richness and evenness in each sample, we can say that the sample 1's evenness and richness is larger than sample 2.

https://www.youtube.com/watch?v=VqcLbKKa7j8&list=PLbVDKwGpb3XnwBACQ5w8L8NzjZ-6D0M8u&index=18

	Sample 1	Sample 2
Species A	10	50
Species B	10	0
Species C	10	3
Species D	10	5
Species E	10	1
Species D	10	25
Species H	10	11

Beside of richness and eveness, we have another index.

Diversity index which is combine of richness and evenness

https://www.youtube.com/watch?v=VqcLbKKa7j8&list=PLbVDKwGpb3XnwBACQ5w8L8NzjZ-6D0M8u&index=18

Why incorporate phylogeny in a diversity metric?

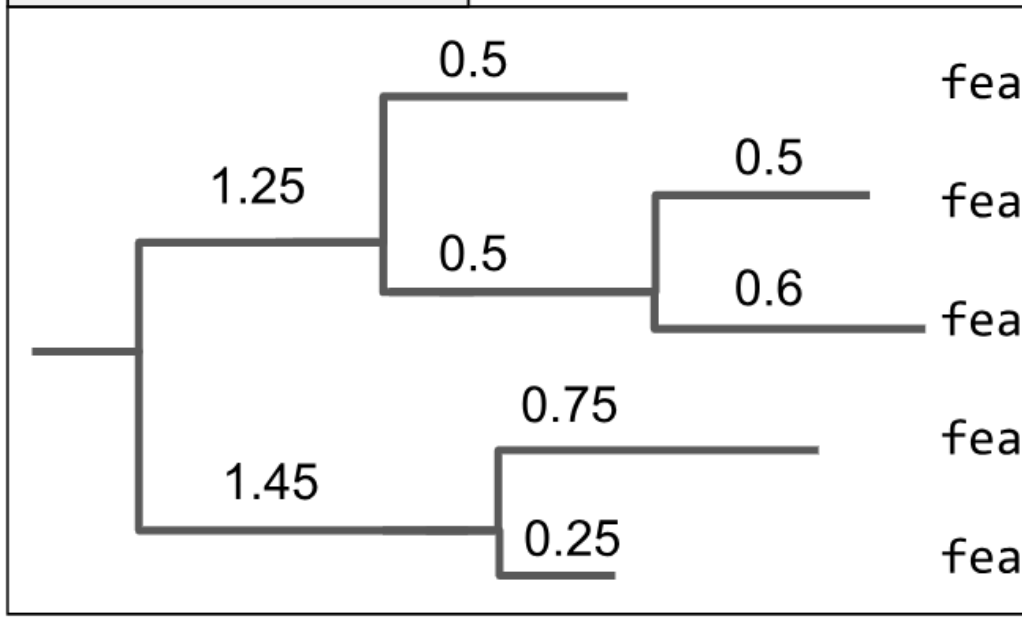
FeatureTable[PresenceAbsence]

	feature1	feature2	feature3	feature4	feature 5
4ac2	1	1	1	0	0
e375	0	1	1	1	0

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature 5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

Phylogeny[Rooted]



https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

Why incorporate phylogeny in a diversity metric?

FeatureTable[PresenceAbsence]

	feature1	feature2	feature3	feature4	feature 5
4ac2	1	1	1	0	0
e375	0	1	1	1	0

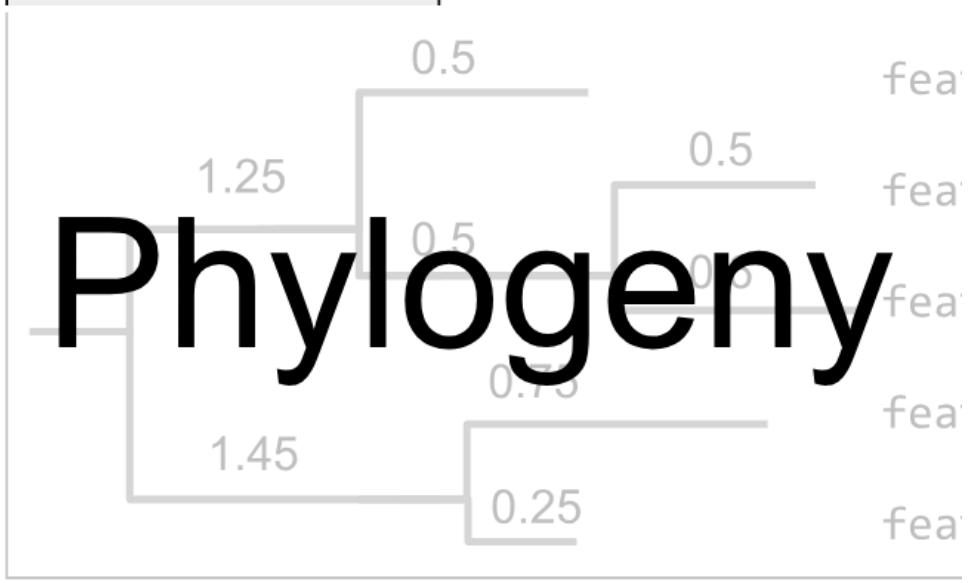
Richness

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature 5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

Evenness

Phylogeny[Rooted]



Phylogeny

https://www.youtube.com/watch?v=tucmby3xl88&list=PLbVDKwGpb3XmkQmoBy1wh3QrWWwdn_pT&index=20

Part 2: Rarefraction

So let make another feature table.

	Feature 1	Feature 2	Feature3	Feature 4	Feature 5
Sample 1	11	1	44	75	2
Sample 2	23	32	1	60	11
Sample 3	14	0	33	11	7
Sample 4	0	2	22	2	8

So let make another feature table.

	Feature 1	Feature 2	Feature3	Feature 4	Feature 5
Sample 1	11	1	44	75	2
Sample 2	23	32	1	60	11
Sample 3	14	0	33	11	7
Sample 4	0	2	12	2	8

Can we compare each sample using this Observed feature ?

Here we can see the observed feature but not only whether that feature exists in the sample but also the frequency of that feature.

	Observed OTUs
Sample 1	133
Sample 2	127
Sample 3	65
Sample 4	24

The answer is we can't do this beacause it's simply not the equal. Beacause what we care is Feature in each sample so if total observed feature is not the equal we can't not conpare them together.

Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total:	10	10	10

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total:	4.29	4.5	4.25

https://www.youtube.com/watch?v=TTUrtCY2k-w

So let make another feature table.

	Feature 1	Feature 2	Feature3	Feature 4	Feature 5
Sample 1	7	1	22	35	0
Sample 2	12	15	0	29	9
Sample 3	14	0	33	11	7
Sample 4	0	2	12	2	8

This is benchmark

In here we sampling the feature sequencing in each sample with the total larger or equal with the sample chosen as the benchmark .

	Observed OTUs
Sample 1	65
Sample 2	65
Sample 3	65
Sample 4	Null

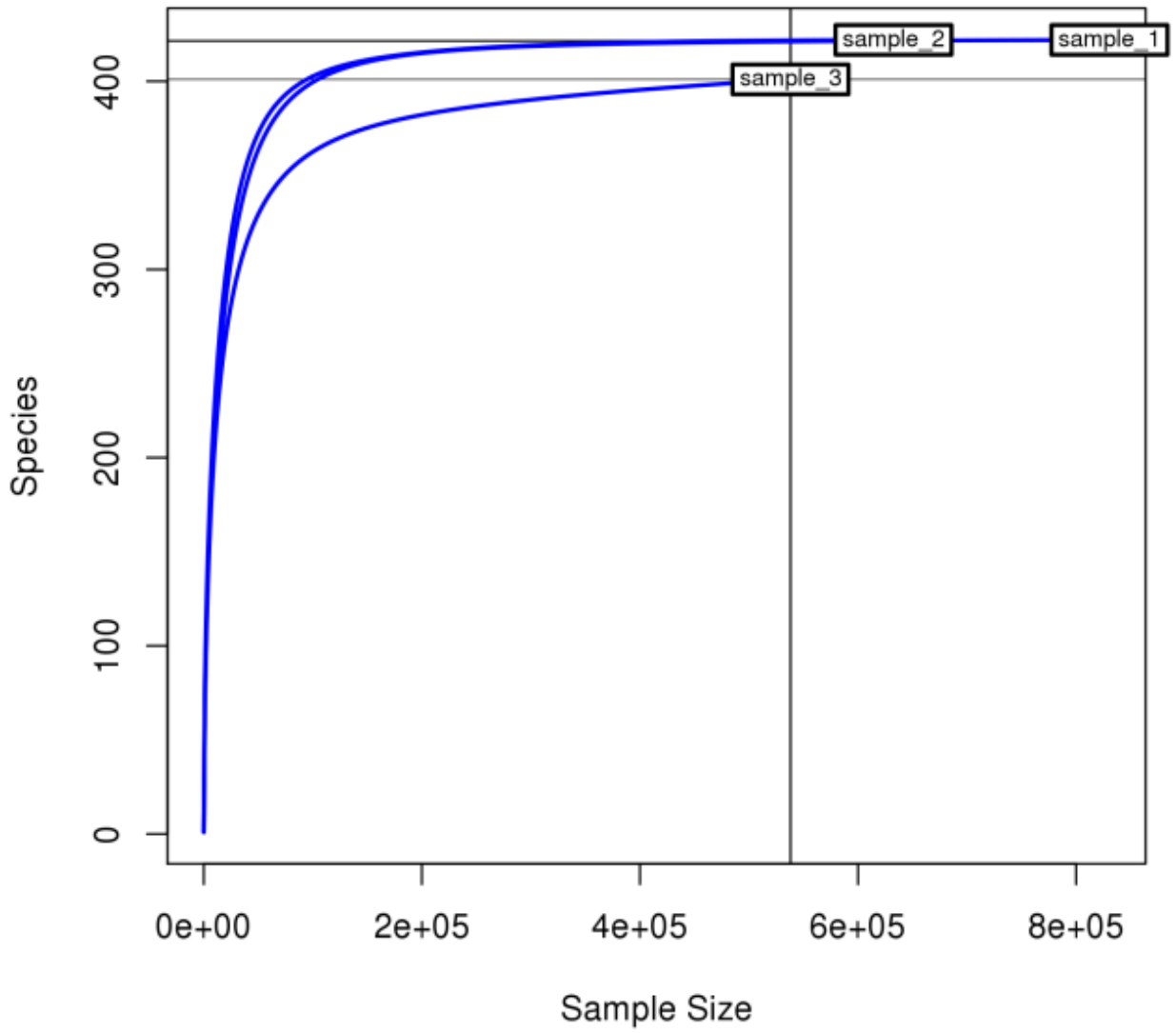
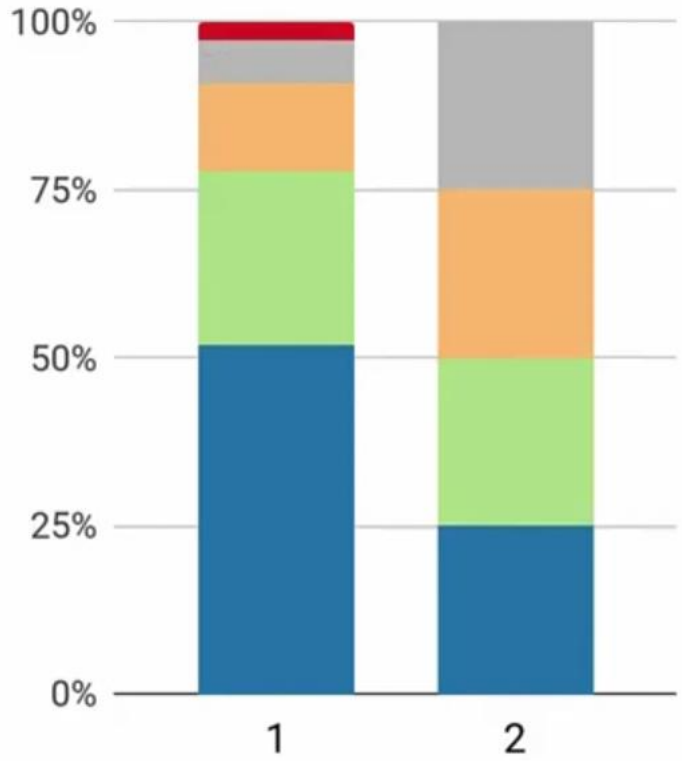


Figure 5: Rarefaction curve of annotated species richness.

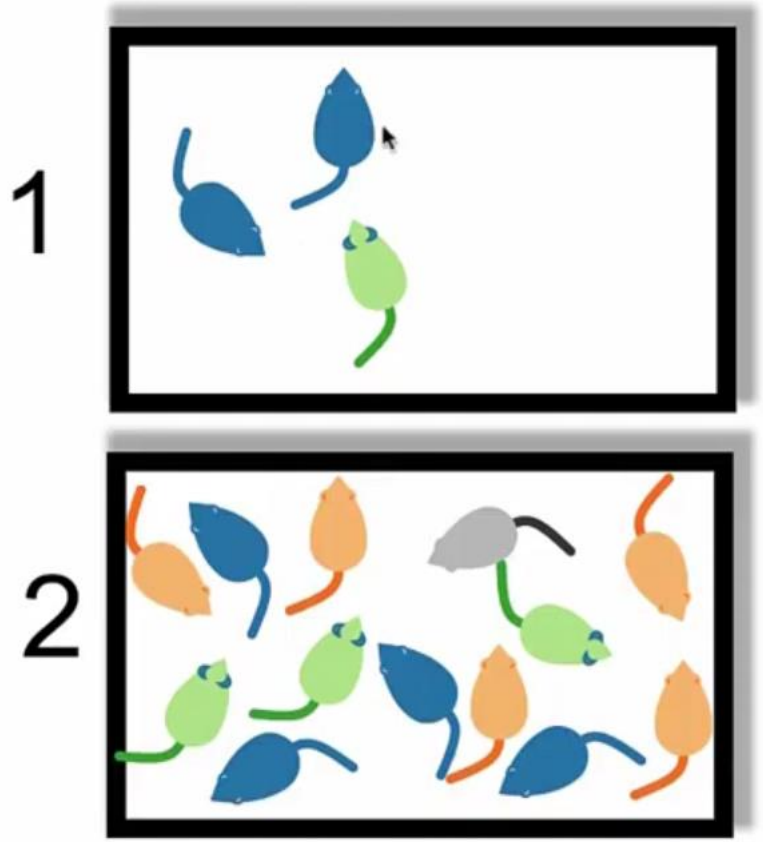
https://eurofinsgenomics.eu/media/1610468/inview_metagenome_explore_demo_report.pdf

A theoretical sampling problem

Actual distribution



Observed distribution



https://www.youtube.com/watch?v=g5BdGP4V5YA

Part 3: index of alpha diversity

The Measures of Chao 1 Richnes

$$S_{Chao1} = S_{obs} + \frac{n_1^2}{2n_2},$$

Where:

- Chao1 is the estimated number of species.
- Sobs is the observed number of species in a sample in total.
- n1 and n2 are the number of singletons and doubletons.

The Measures of Chao 1 Richnes

$$S_{Chao1} = S_{obs} + \frac{n_1^2}{2n_2},$$

Sample C:

- Species A: 10 individuals
- Species B: 1 individual (singleton)
- Species C: 2 individuals (doubleton)
- Species D: 5 individuals
- Species E: 1 individual (singleton)
- Species F: 3 individuals

The Measures of Chao 1 Richnes

$$S_{Chao1} = S_{obs} + \frac{n_1^2}{2n_2},$$

Sample C:

- Species A: 10 individuals
- Species B: 1 individual (singleton)
- Species C: 2 individuals (doubleton)
- Species D: 5 individuals
- Species E: 1 individual (singleton)
- Species F: 3 individuals

From this data:

- Sobs = 6 (there are 6 observed species)
- n1 = 2 (Species B and E are singletons)
- N2 = 1 (Species C is a doubleton)

The Measures of Chao 1 Richnes

$$S_{Chao1} = S_{obs} + \frac{n_1^2}{2n_2},$$

Sample C:

•Species A: 10 individuals
•Species B: 1 individual (singleton)
•Species C: 2 individuals (doubleton)
•Species D: 5 individuals
•Species E: 1 individual (singleton)
•Species F: 3 individuals

From this data:

•Sobs = 6 (there are 6 observed species)
•n1 = 2 (Species B and E are singletons)
•n2 = 1 (Species C is a doubleton)

Using the Chao1 formula:

$$\text{Chao1} = 6 + \frac{2^2}{2 \times 1}$$

$$\text{Chao1} = 6 + \frac{4}{2}$$

$$\text{Chao1} = 6 + 2 = 8$$

The Measures of Chao 1 Richnes

$$S_{Chao1} = S_{obs} + \frac{n_1^2}{2n_2},$$

Sample C:

•Species A: 10 individuals
•Species B: 1 individual (singleton)
•Species C: 2 individuals (doubleton)
•Species D: 5 individuals
•Species E: 1 individual (singleton)
•Species F: 3 individuals

From this data:

•Sobs = 6 (there are 6 observed species)
•n1 = 2 (Species B and E are singletons)
•n2 = 1 (Species C is a doubleton)

Using the Chao1 formula:

$$\text{Chao1} = 6 + \frac{2^2}{2\times 1}$$

$$\text{Chao1} = 6 + \frac{4}{2}$$

$$\boxed{\text{Chao1} = 6 + 2 = 8}$$

The Measures of Shannon Diversity

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

Where:

- H is the Shannon Diversity Index.
- S is the total number of species in the community (species richness).
- pi is the proportion of individuals belonging to species iii (i.e., the number of individuals of species iii divided by the total number of individuals in the community).

Shannon's index (labeled as H) was originally developed by Claude Shannon in 1948 to quantify the entropy (uncertainty or information content) in strings of text.

The Measures of Shannon Diversity

Shannon Diversity Index:

non-phylogenetic, alpha diversity metric measuring richness and evenness

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

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature 5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

SampleData[AlphaDiversity]

	Shannon
4ac2	
e375	

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

The Measures of Shannon Diversity

Shannon Diversity Index:

non-phylogenetic, alpha diversity metric measuring richness and evenness

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

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature 5
4ac2	25	30	15	0	0
e375	0	17	32	25	0

SampleData[AlphaDiversity]

	Shannon
4ac2	1.061
e375	

$$H' = -\left(0.375(-1.030)+0.429(-0.847)+0.214\ (-1.540)\right)$$

25 /(25 + 30 + 15)

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

The Measures of Shannon Diversity

Shannon Diversity Index:

non-phylogenetic, alpha diversity metric measuring richness and evenness

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature 5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

SampleData[AlphaDiversity]

	Shannon
4ac2	1.061
e375	1.064

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

The Measures of Simpson Diversity

$$D_0 = \frac{\sum n(n-1)}{N(N-1)}. \quad \Leftrightarrow \quad D_O = \sum_{i=1}^{S} p_i^2,$$

Where:

- S is the total number of species.
- pi is the proportion of individuals belonging to species i.

"the probability that two individuals chosen at random and independently from the population will be found to belong to the same group"

Simpson's Index focuses more on the dominance of species within a community

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

The Measures of Simpson Diversity

	Feature 1	Feature 2	Feature 3	Feature 4
Sample 1	4	55	2	0
Sample 2	0	12	34	11

D1 = (4*3 + 55*51 + 2)/(61*60) = 0.82

D2 =(12*11 + 34*33 + 11*10)/(57*56) = 0.43

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

The Measures of Simpson of Diversity index

Simpson's Diversity Index = 1 - D

Where:

D is simpson index.

	Feature 1	Feature 2	Feature 3	Feature 4
Sample 1	4	55	2	0
Sample 2	0	12	34	11

	Simpson’s Diversity Index
Sample 1	0.18
Sample 2	0.57

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWIWdn_pTT&index=20

The Measures of Simpson of Diversity index

Simpson's Diversity Index = 1 - D

Where:

D is simpson index.

	Feature 1	Feature 2	Feature 3	Feature 4
Sample 1	4	55	2	0
Sample 2	0	12	34	11

	Simpson’s Diversity Index
Sample 1	0.18
Sample 2	0.57

Beside of that we have:

Simpson’s Reciprocal Index = 1/D

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWIWdn_pTT&index=20

The Measures of Pielou's Evenness

$$J' = \frac{H}{ln(S)}$$

Where:

- H is Shannon's diversity index
- S is the total number of species observed in a sample

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

The Measures of Pielou's Evenness

$$J' = \frac{H}{ln(S)}$$

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature 5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

SampleData[AlphaDiversity]

	Shannon
4ac2	1.061
e375	1.064

Calculate Pielou's Evenness (J'):

- For sample 4ac2:

$$J'_{4ac2} = \frac{1.061}{1.0986} \approx 0.966$$

- For sample e375:

$$J'_{e375} = \frac{1.064}{1.0986} \approx 0.969$$

	Pielou's Evenness
4ac2	0.966
E375	0.969

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

Phylogenetic Diversity: faiths_pd index

How many feature we can observe in each sample ?

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Sample1	11	40	11	0	0
Sampl2	0	22	33	2	0

	Observed OTUs
Sample1	?
Sample 2	?

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1whpt0fWlWdn_pTT&index=20

Phylogenetic Diversity: faiths_pd index

We can only see 3 observed Features in each sample and with the right panel we cannot compare the 2 samples because they have the same number of observed features.

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Sample1	1	1	1	0	0
Sampl2	0	1	1	1	0

	Observed OTUs
Sample1	3
Sample 2	3

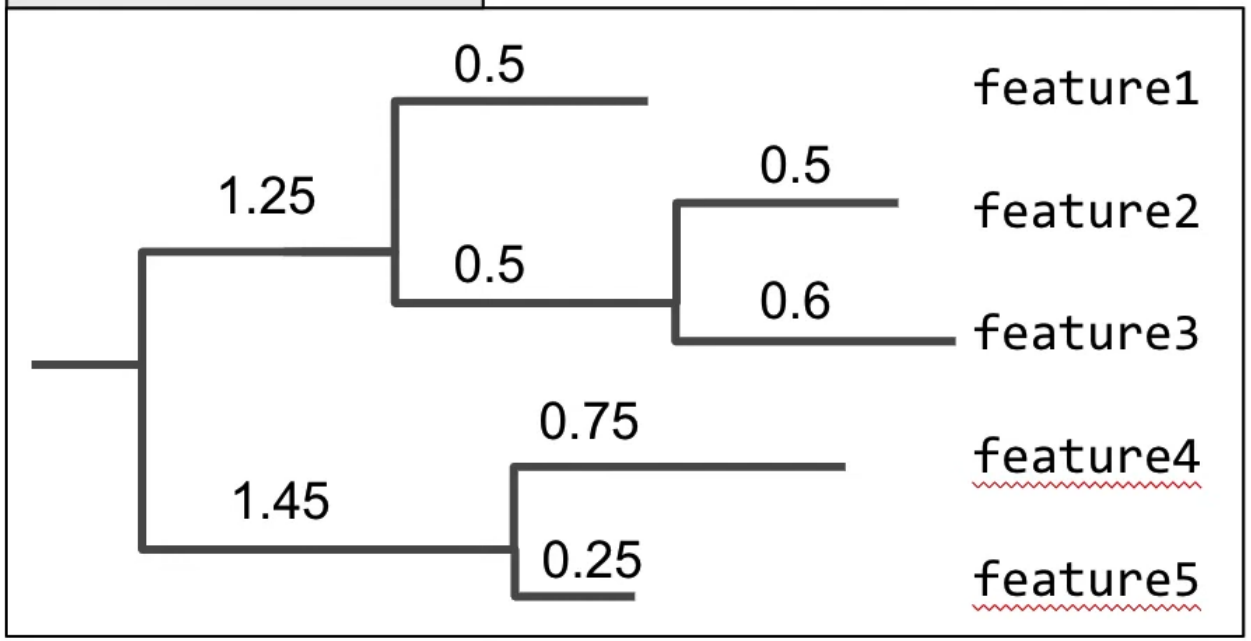
https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

Phylogenetic Diversity: faiths_pd index

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Sample1	1	1	1	0	0
Sampl2	0	1	1	1	0

	Observed OTUs
Sample1	3
Sample 2	3

Phylogeny[Rooted]



FeatureData[Taxonomy]

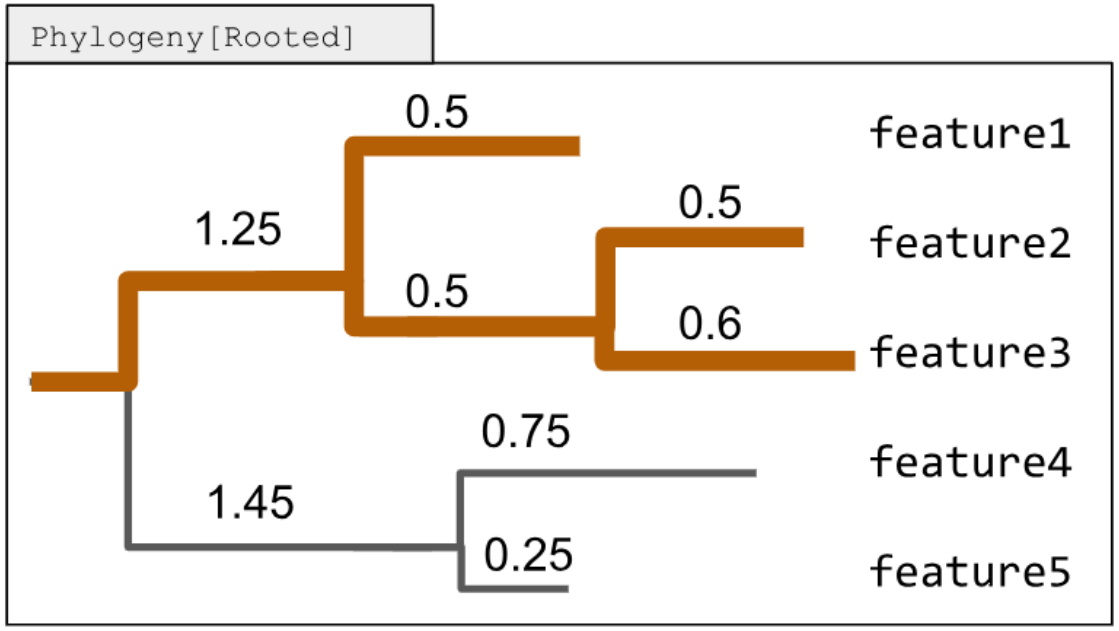
	Domain
feature1	Bacteria
feature2	Bacteria
feature3	Bacteria
feature4	Archaea
feature5	Archaea

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

Phylogenetic Diversity: faiths_pd index

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Sample1	1	1	1	0	0
Sampl2	0	1	1	1	0

	Observed OTUs
Sample1	3.35
Sample 2	3



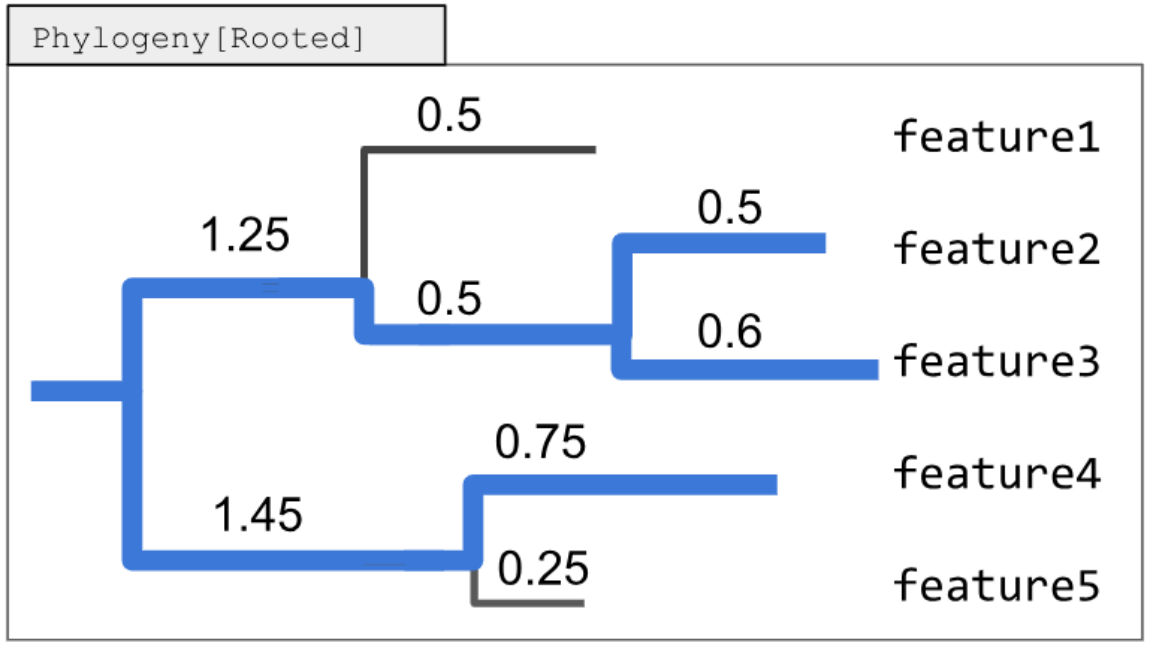
Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

Phylogenetic Diversity: faiths_pd index

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Sample1	1	1	1	0	0
Sampl2	0	1	1	1	0

	Observed OTUs
Sample1	3.35
Sample 2	5.05



Sum of branch length covered by a sample.

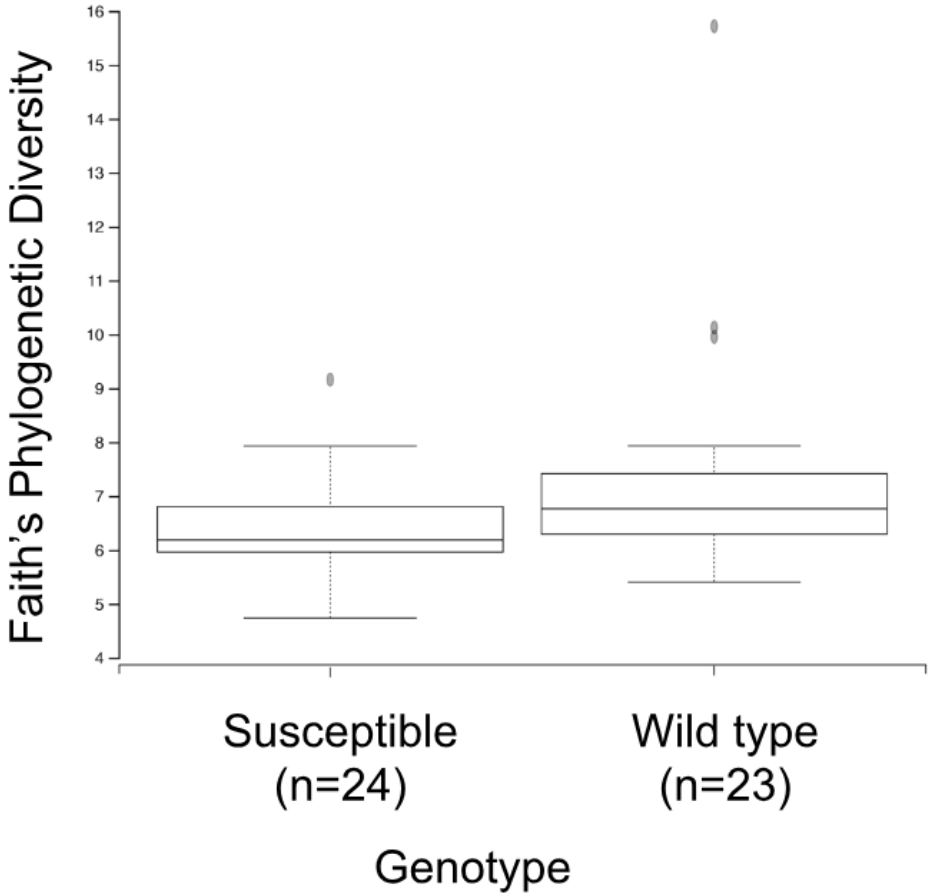
Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

Alpha diversity comparison

- visually
  - distribution comparison plots (discrete)
  - scatter plots (continuous)

- statistically
  - Kruskal-Wallis (discrete data)
  - Spearman correlation (continuous)
  - Regression (when asymptotically normal)



https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

Summary:

- Difinetion of alpha diversity: richness, evenness and diversity.
- What is rarefration and why we use it.
- Some index of alpha diversity: Chao 1, Shannon diversity index, Simpson diversity index, Pielou's Evenness and phylogeny diversity.

https://www.youtube.com/watch?v=tucm8y5xi88&list=PLbVDKwGpb3XmkQmoBy1wh3QfWlWdn_pTT&index=20

```python
for u in range(0, 1000):
    print('Thank you!')
```