Exploring the phylogenetic composition of microbiotas

F. Mazel & J. Sanders

ECOSCOPE
routing the microcosmos

UBC

—

June, 20, 2018

Organisation

Who are we?

Florent Mazel
flo.mazel@gmail.com

Jon Sanders
jonsan@gmail.com

Question

How can we (better) describe the composition of microbiomes using phylogenetic trees?

Targeted Skills

Reconstruct a microbial phylogenetic tree

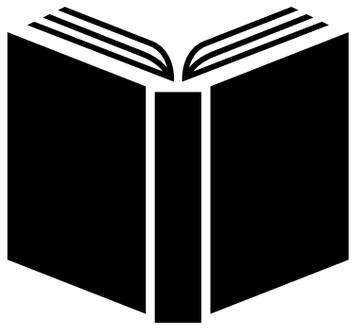
Document microbiota compositions using this phylogenetic tree

Statistically test their relationship with metadata

Visualize these compositions

Organisation of the workshop

Theory



Practice

[These Slides]

```
> read.table()
```

Important point
Try NOT to be shy!

Do NOT hesitate to ask questions

Important point
Try NOT to be shy!

Do NOT hesitate to ask questions

Use red/green cards to inform us about your coding situation

Introduction

Why Exploring the phylogenetic composition of microbiomes?

Why Phylogeny?

What is a phylogeny?

What is a phylogeny?

Phylogeny:
“a diagrammatic hypothesis
about the history of the evolutionary relationships of a group of organisms”

What is a phylogeny?

A hypothesis about the relationship between organisms

What is a phylogeny?

A hypothesis about the relationship between organisms

An estimation of something that *actually happened*

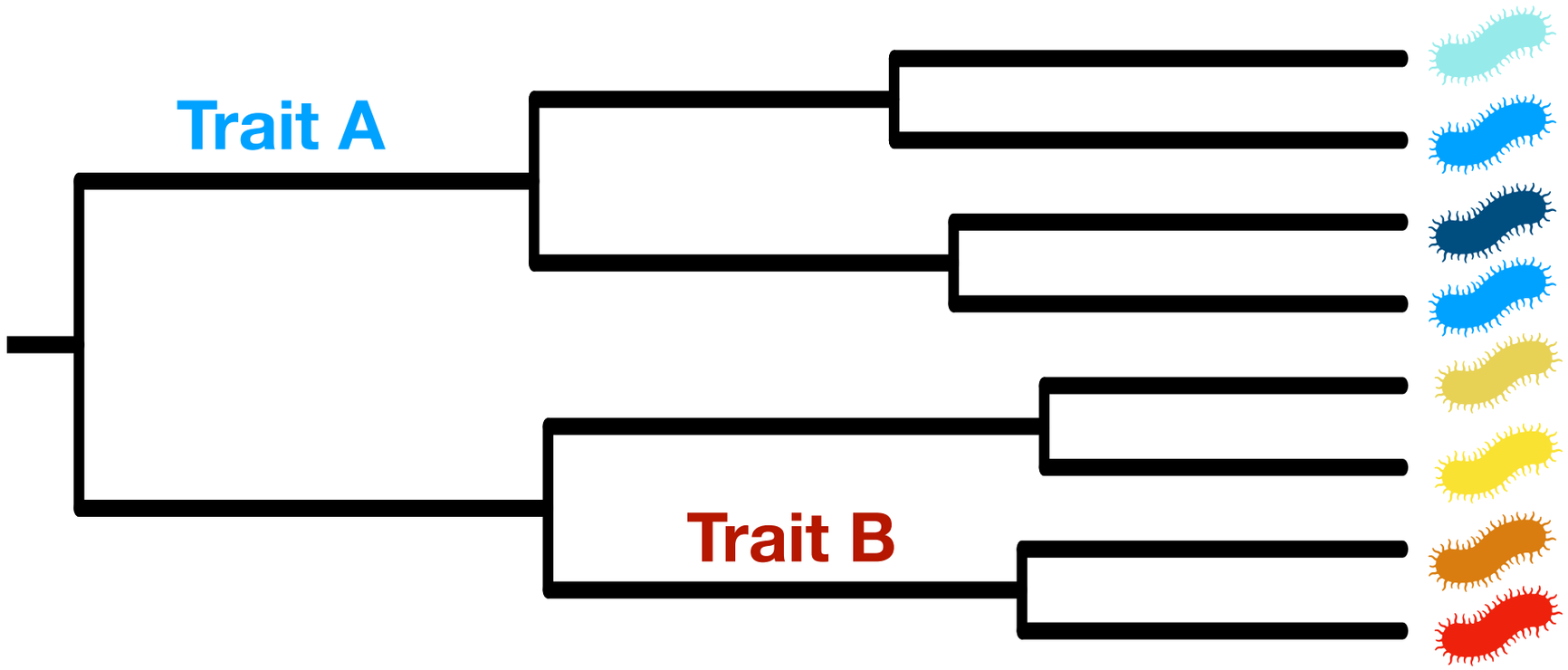
What is a phylogeny?

A hypothesis about the relationship between organisms

An estimation of something that *actually happened*

…and hence, a proxy for other things that happened

The phylogeny as a proxy for functions

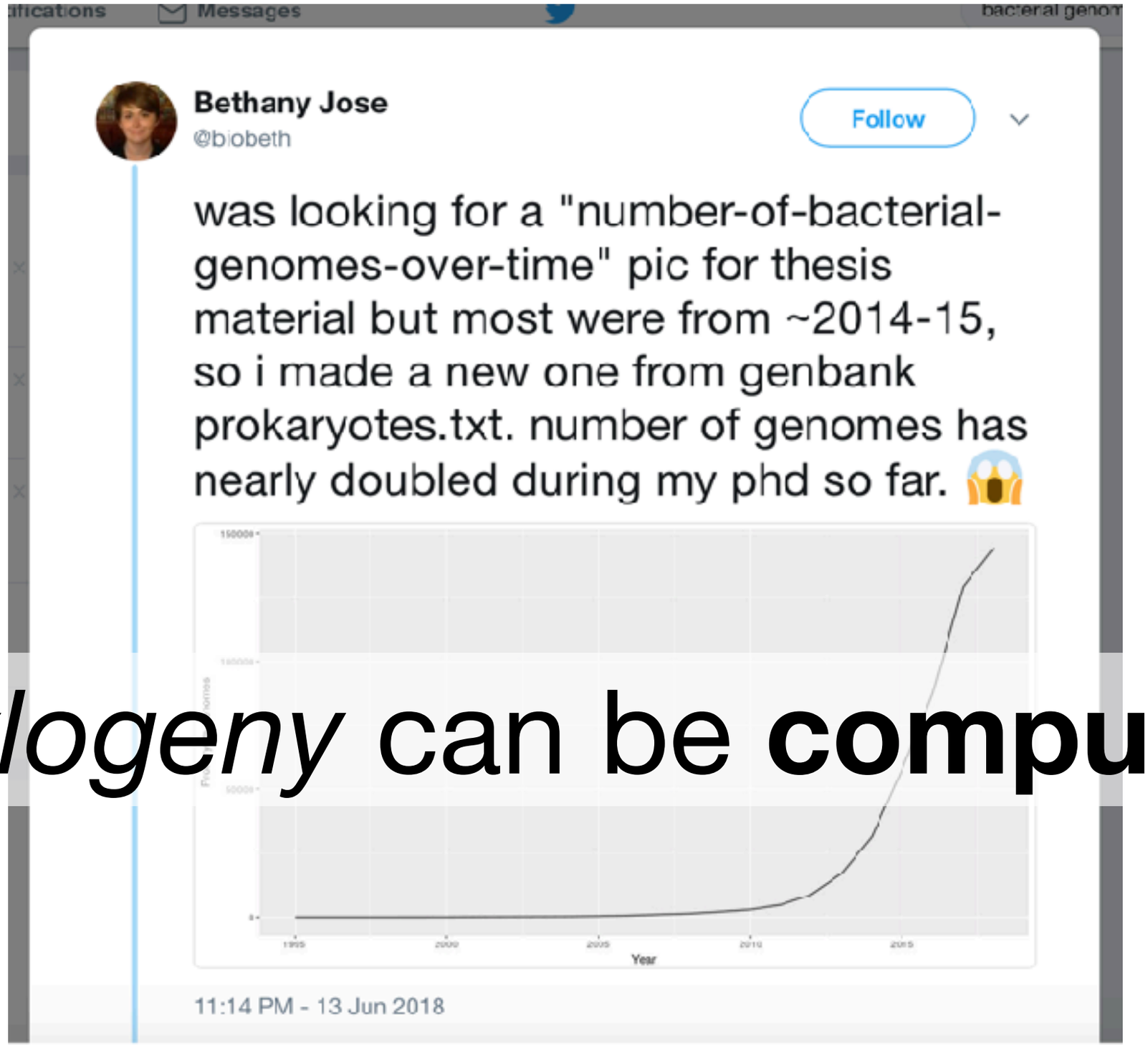


Phylogeny is a data structure

Taxonomy is also a data structure: systematic categorization

…but most microbes **don’t have names**

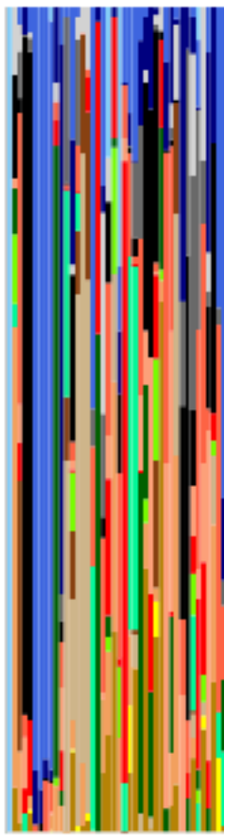
Phylogeny is a data structure

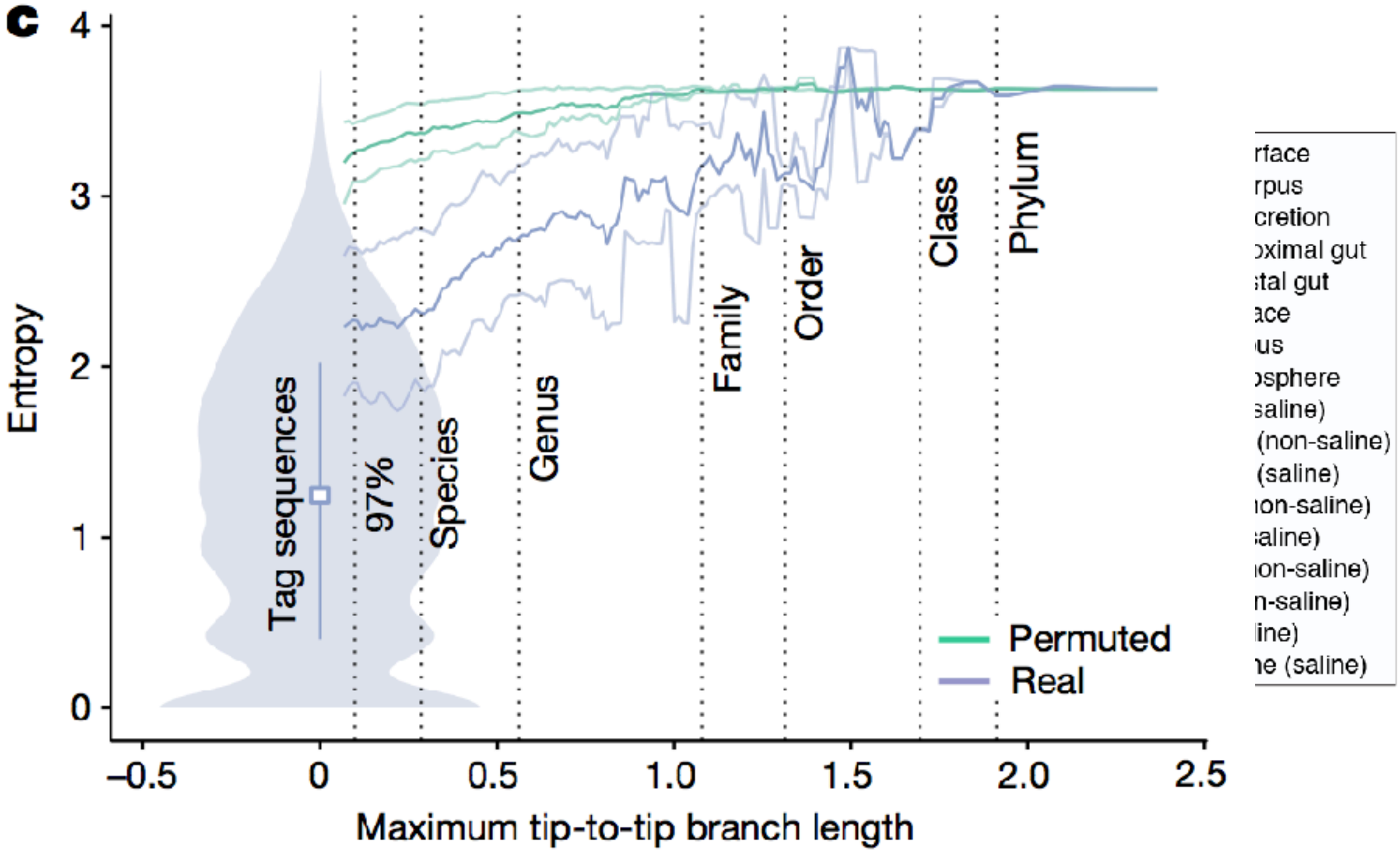


Phylogeny can be **computed**

Phylogeny correlates well with nature

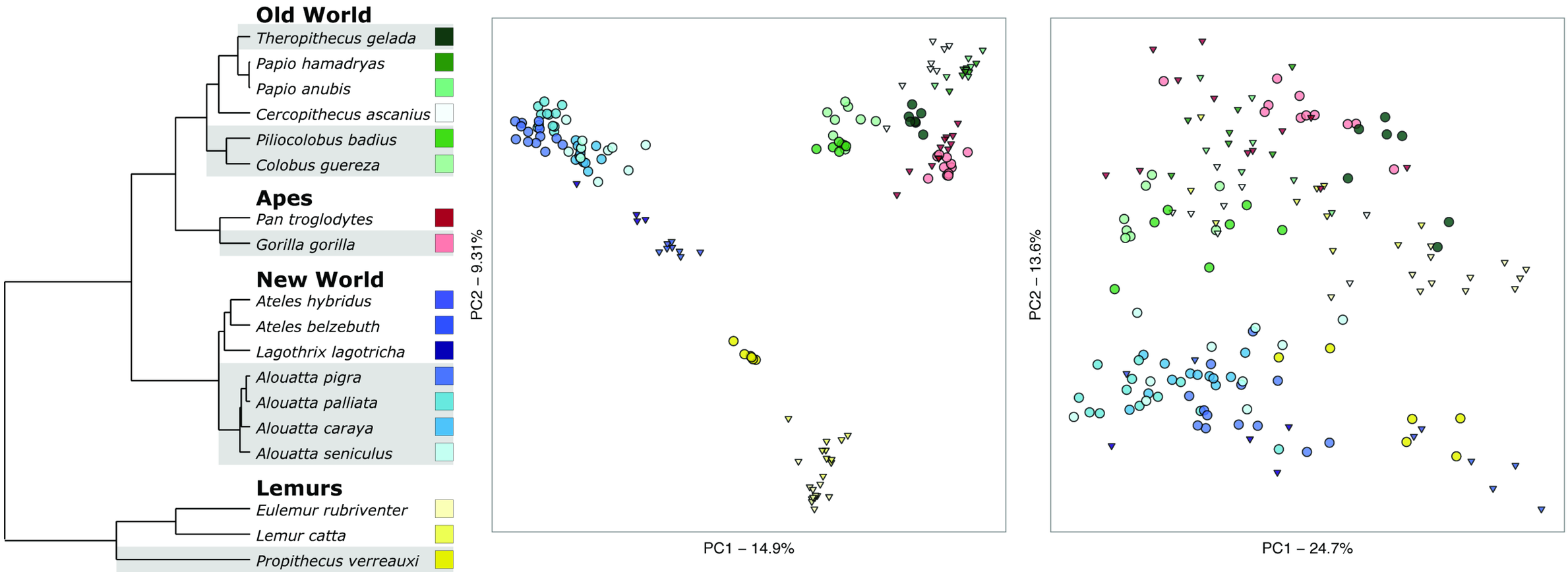
Example: which environments are microbes found in?





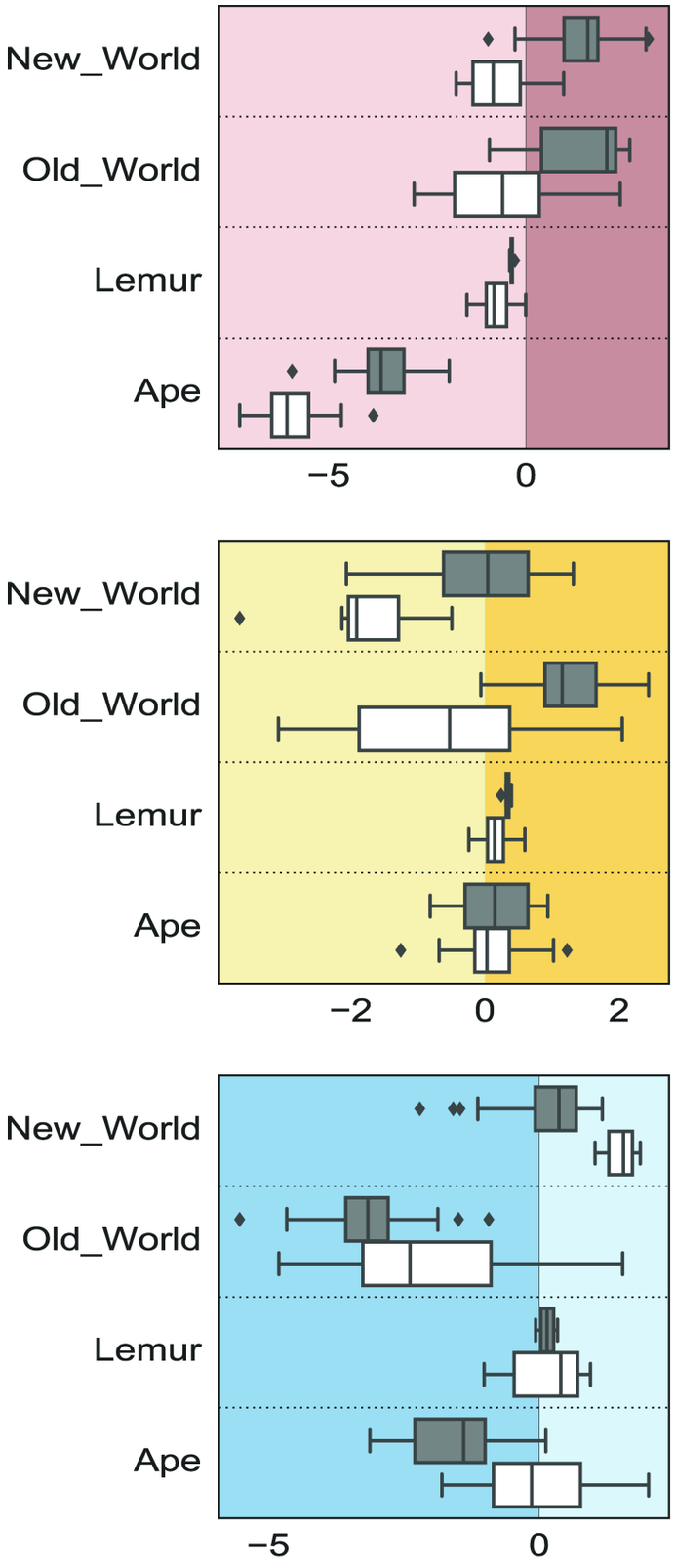
Thompson *et al.* **Nature** 2017

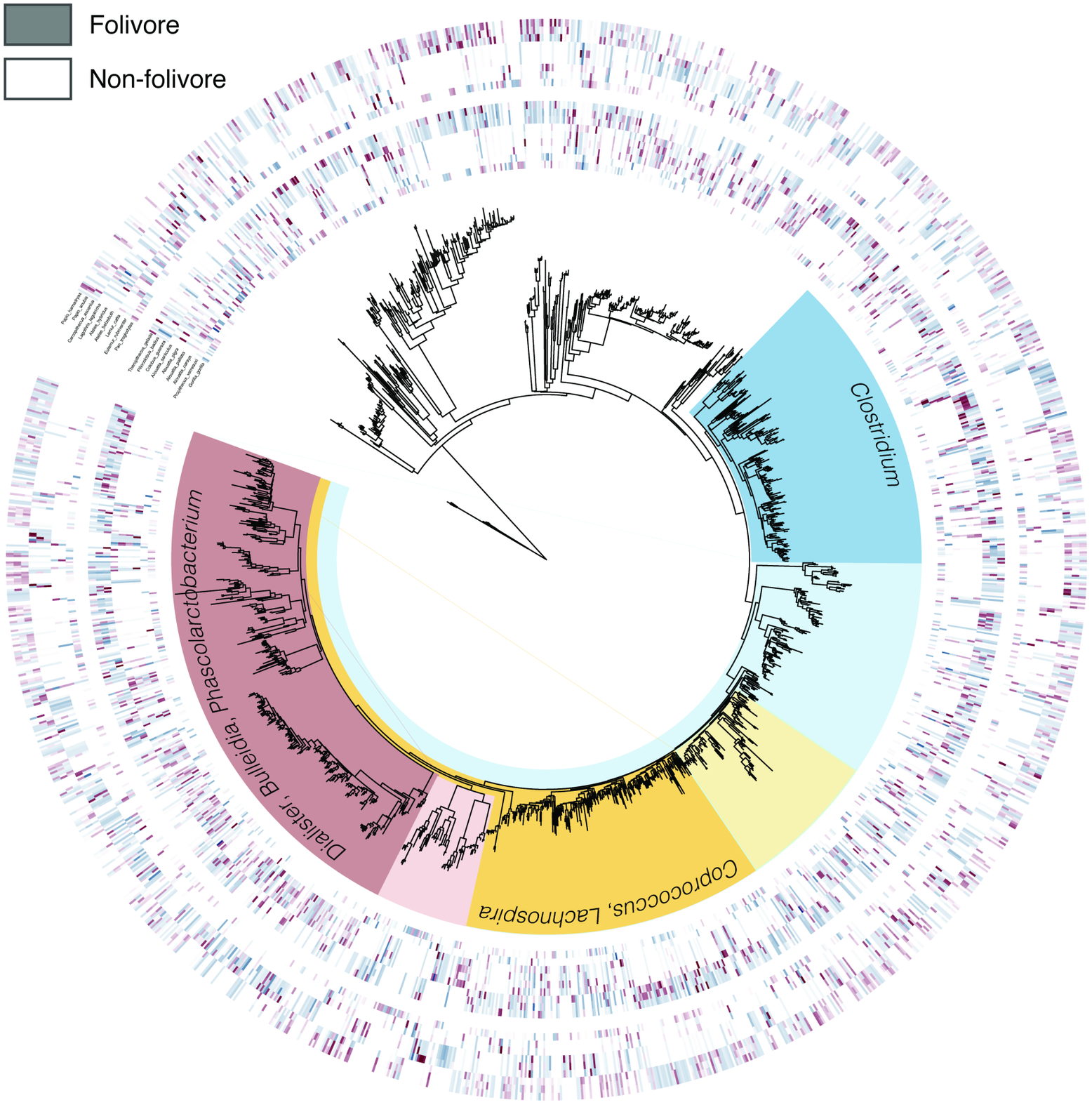
Phylogeny allows scale-independent analysis



Amato *et al*., **ISMEJ** *In press*

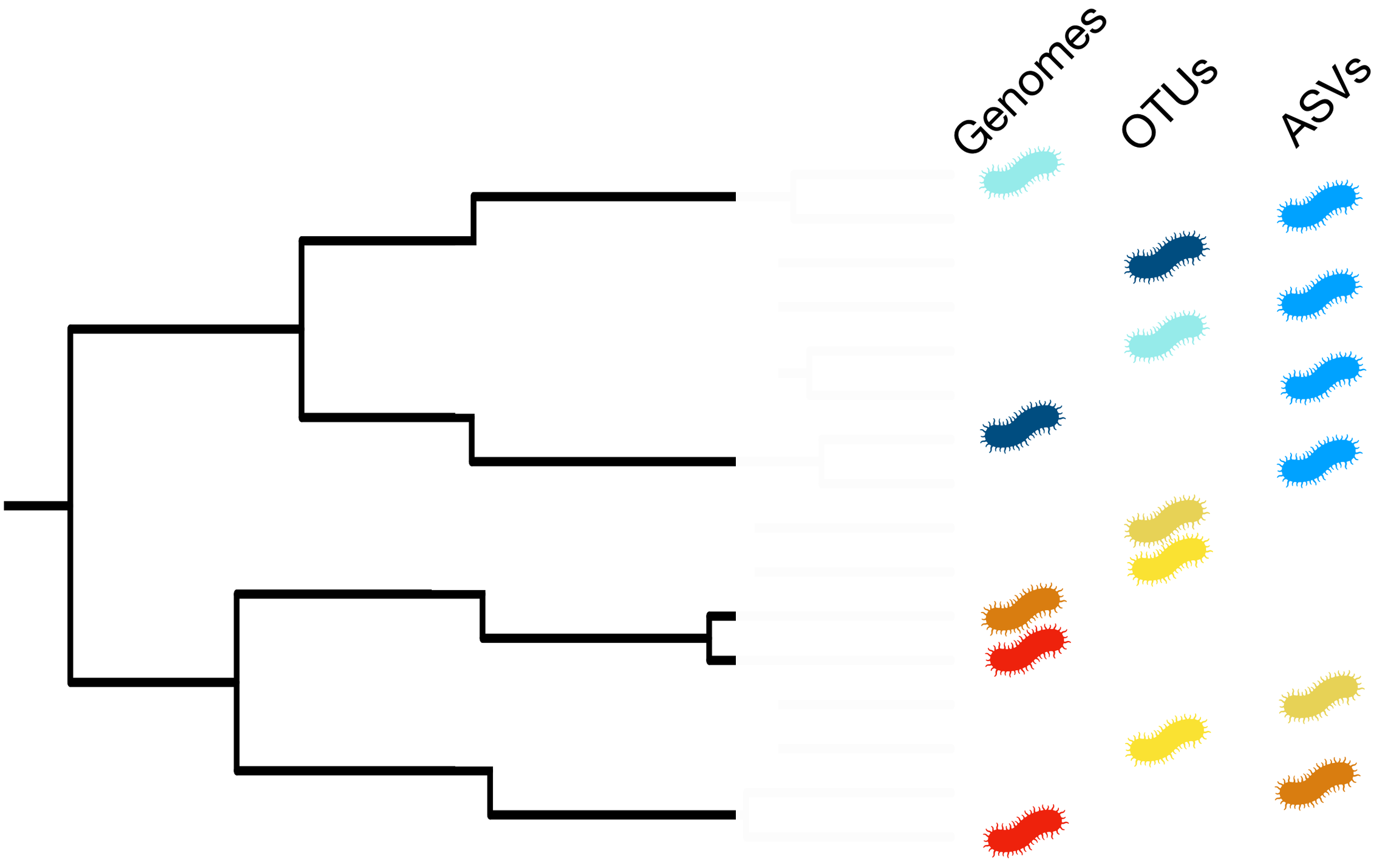
Phylogeny allows scale-independent analysis





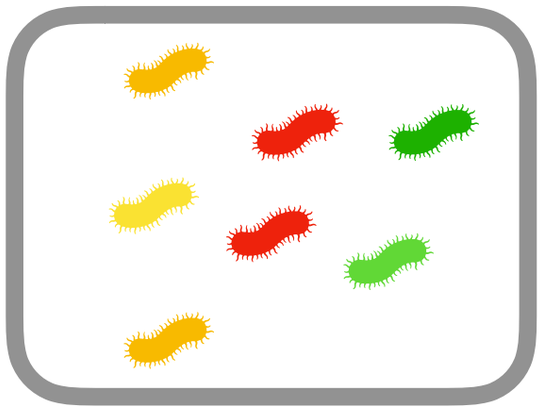
Amato *et al*., **ISMEJ** *In press*

Phylogeny allows integration of data types

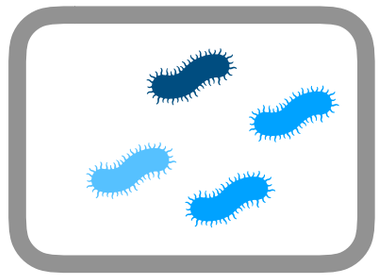


The composition of microbiomes

Microbiome 1

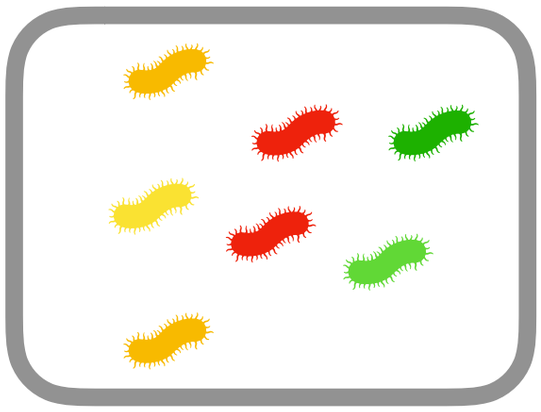


Microbiome 2

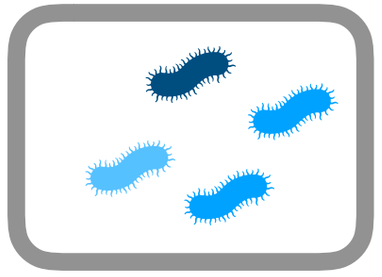


The composition of microbiomes

Microbiome 1

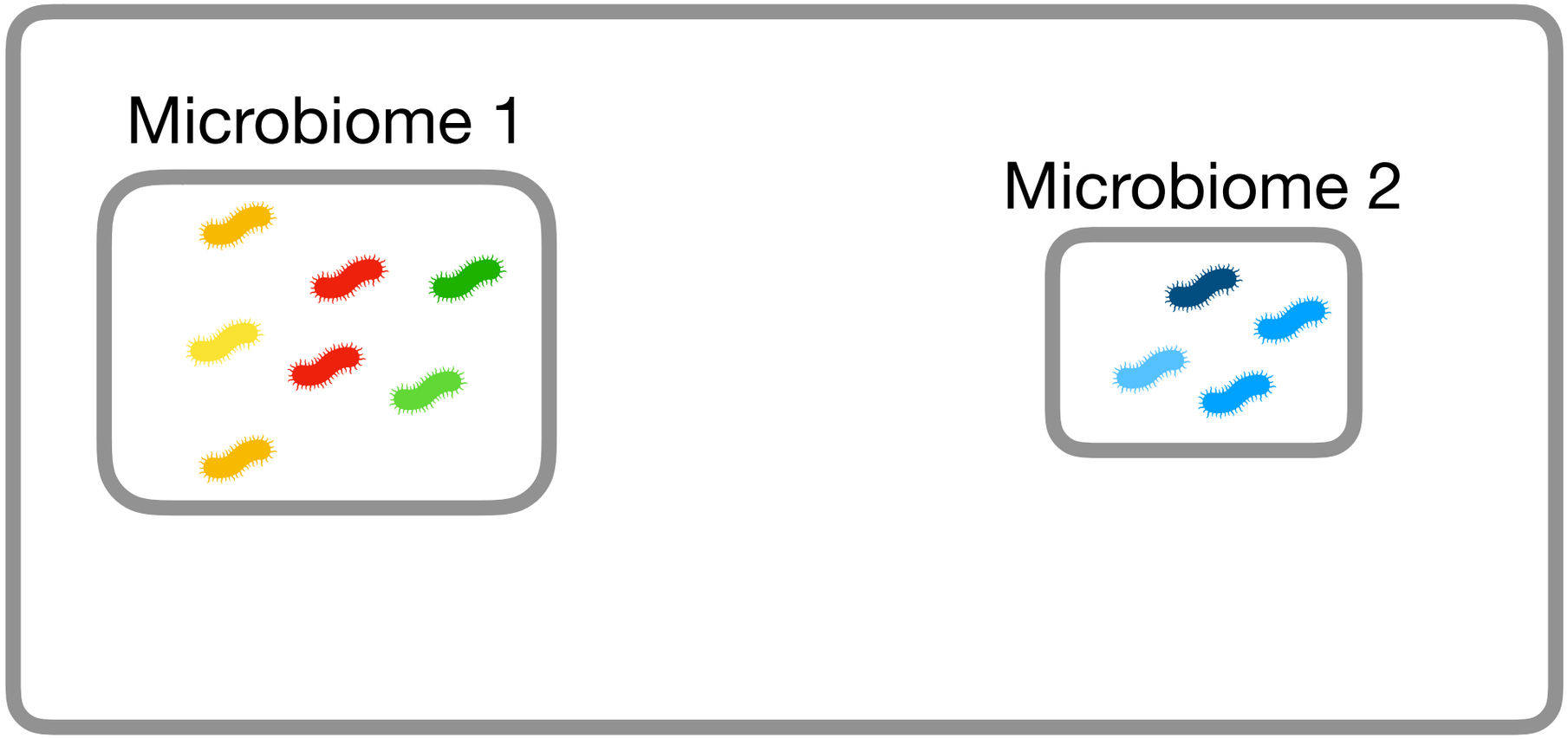


Microbiome 2



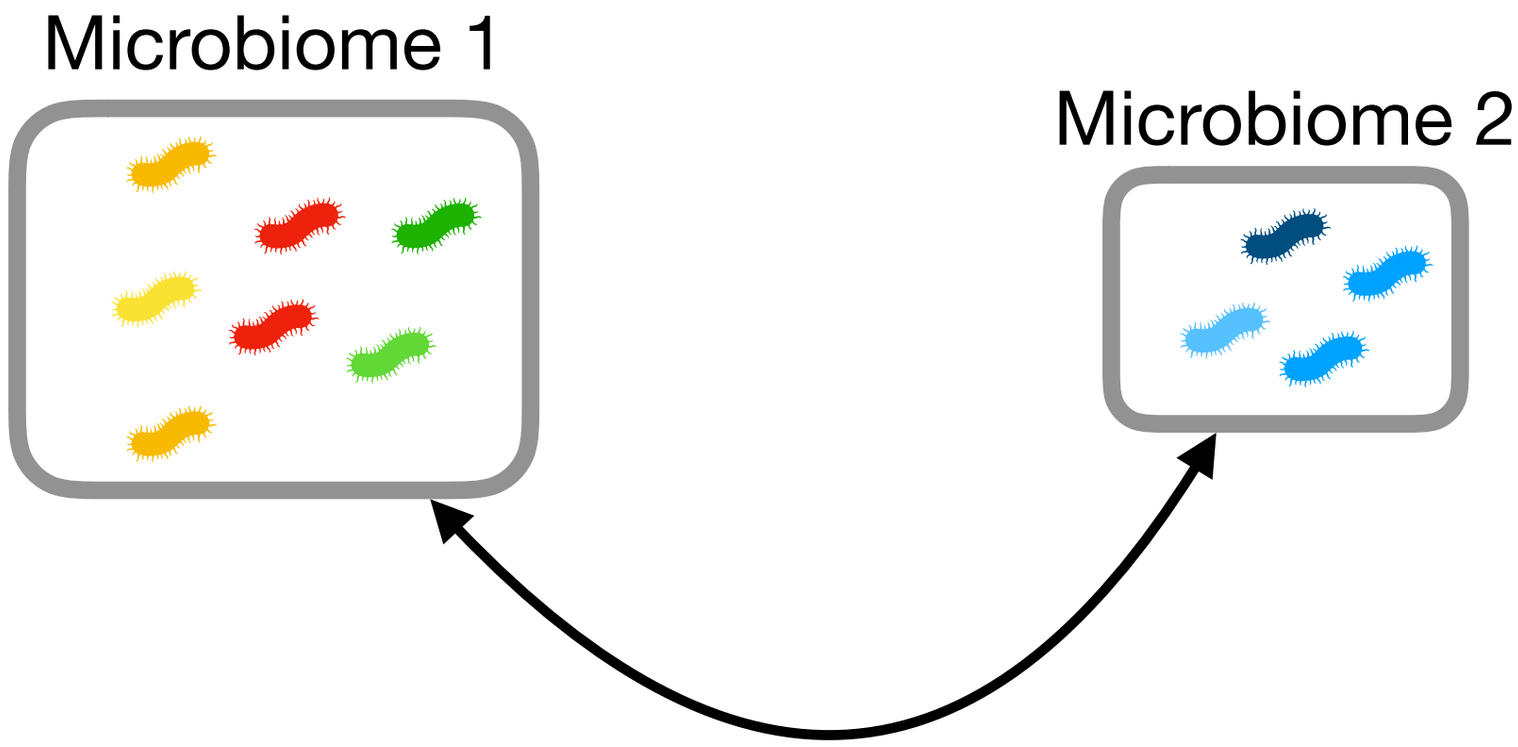
Alpha Diversity = “quantity” of “species” of **ONE** microbiome

The composition of microbiomes



Gamma Diversity = “quantity” of “species” of **ALL** microbiome

The composition of microbiomes



Beta Diversity = “quality” of “species” in the microbiomes

The composition of microbiomes

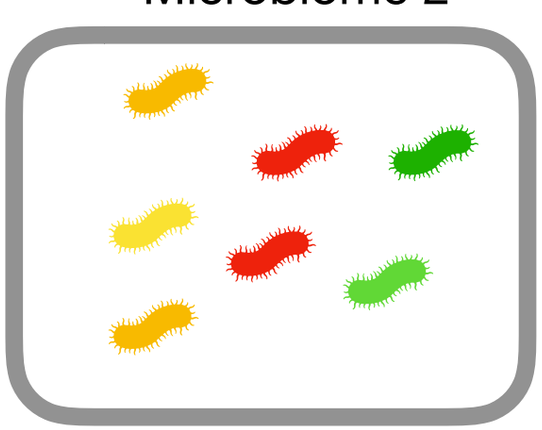
$$\textbf{Beta Diversity} = \frac{\text{Gamma}}{\text{Alpha}}$$

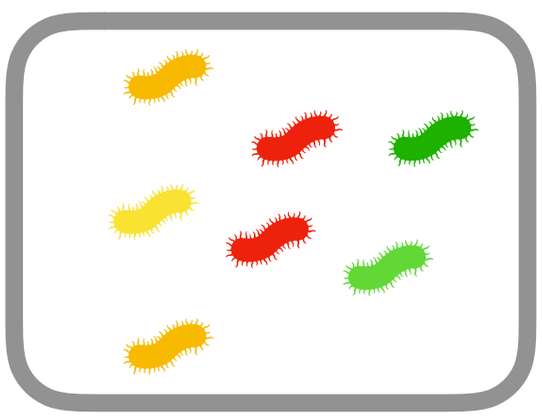
The composition of microbiomes

$$\textbf{Beta Diversity} = \frac{\text{Gamma}}{\text{Alpha}}$$

Example

Microbiome 2





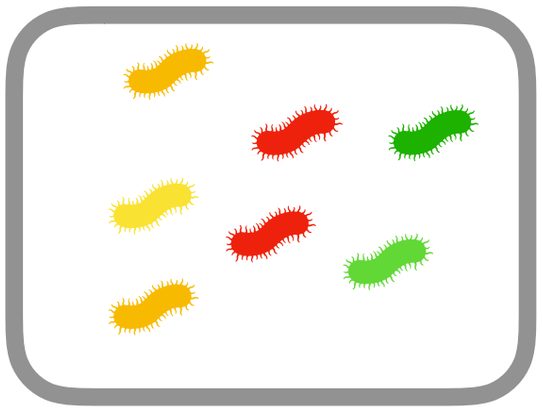
Microbiome 1

The composition of microbiomes

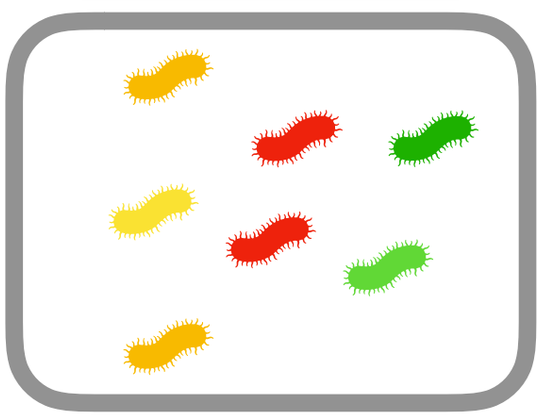
$$\textbf{Beta Diversity} = \frac{\text{Gamma}}{\text{Alpha}}$$

Example

Microbiome 2



Alpha = 5



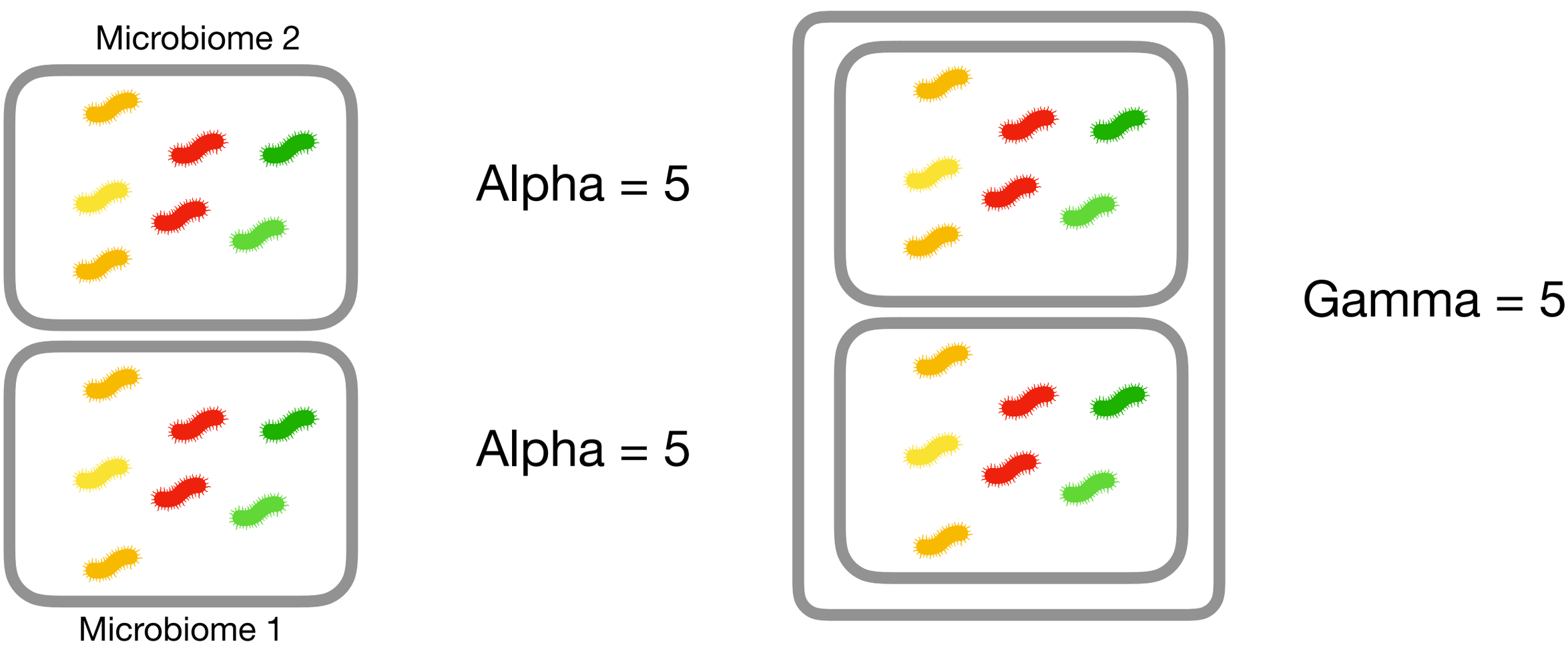
Alpha = 5

Microbiome 1

The composition of microbiomes

$$\textbf{Beta Diversity} = \frac{\text{Gamma}}{\text{Alpha}}$$

Example

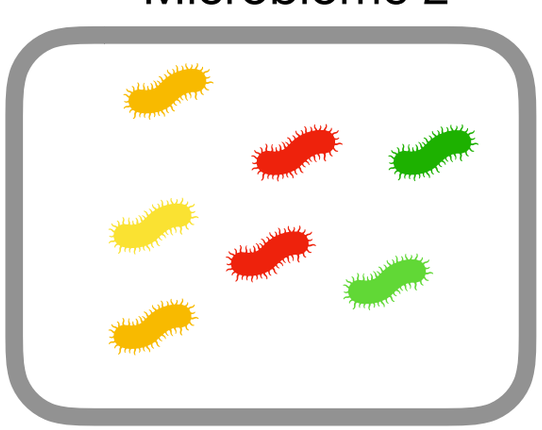


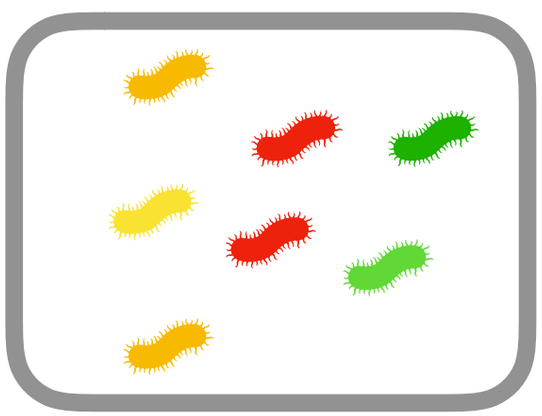
The composition of microbiomes

$$\textbf{Beta Diversity} = \frac{\text{Gamma}}{\text{Alpha}}$$

Example

Microbiome 2



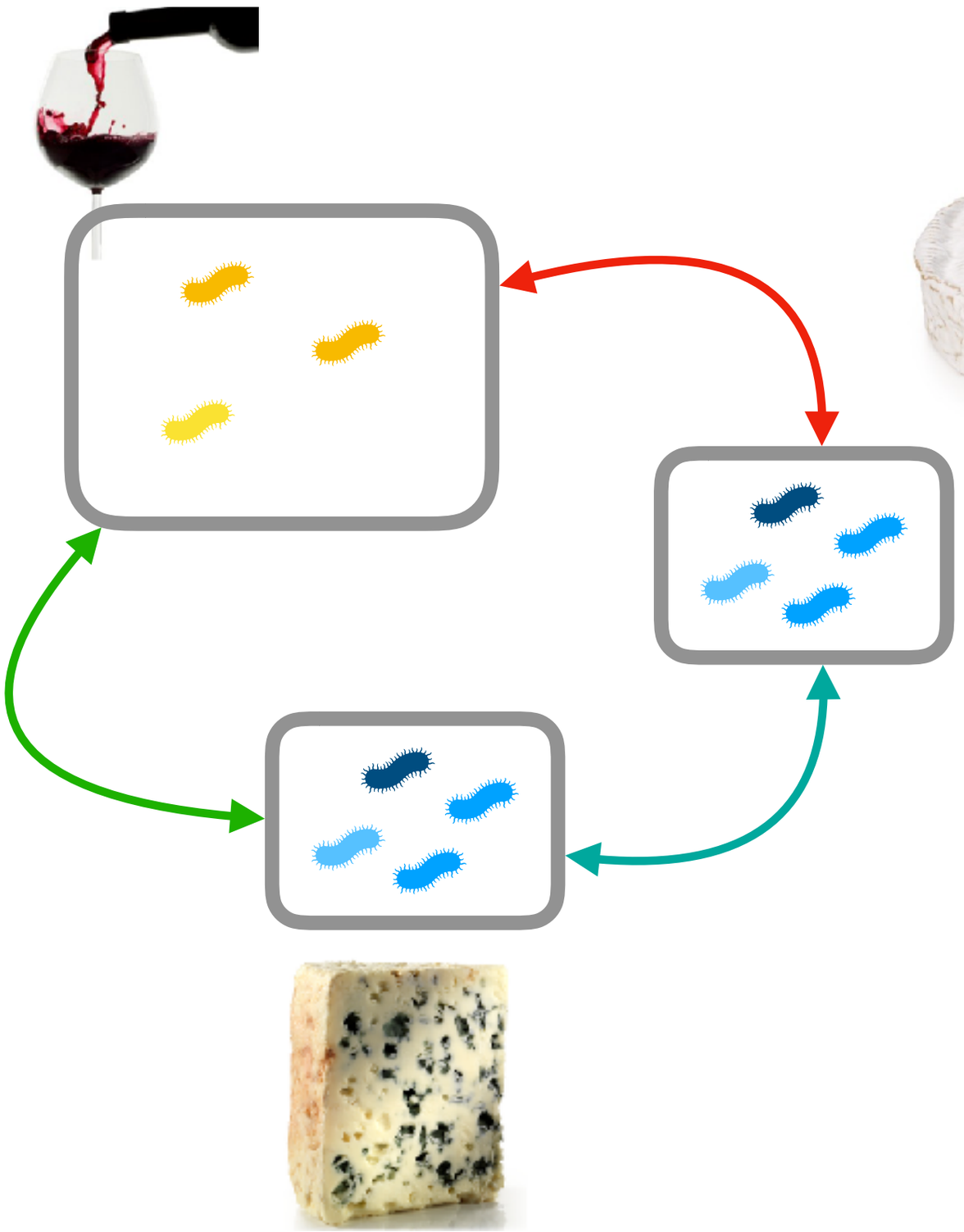


Microbiome 1

Alpha = Gamma

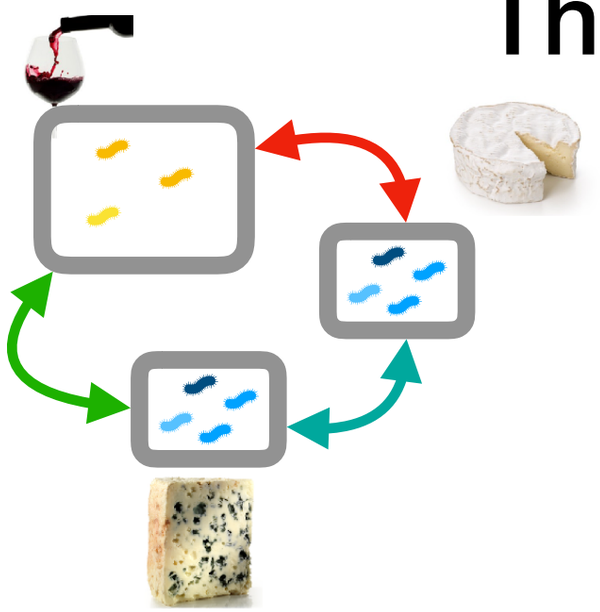
Beta is minimal

The composition of microbiomes

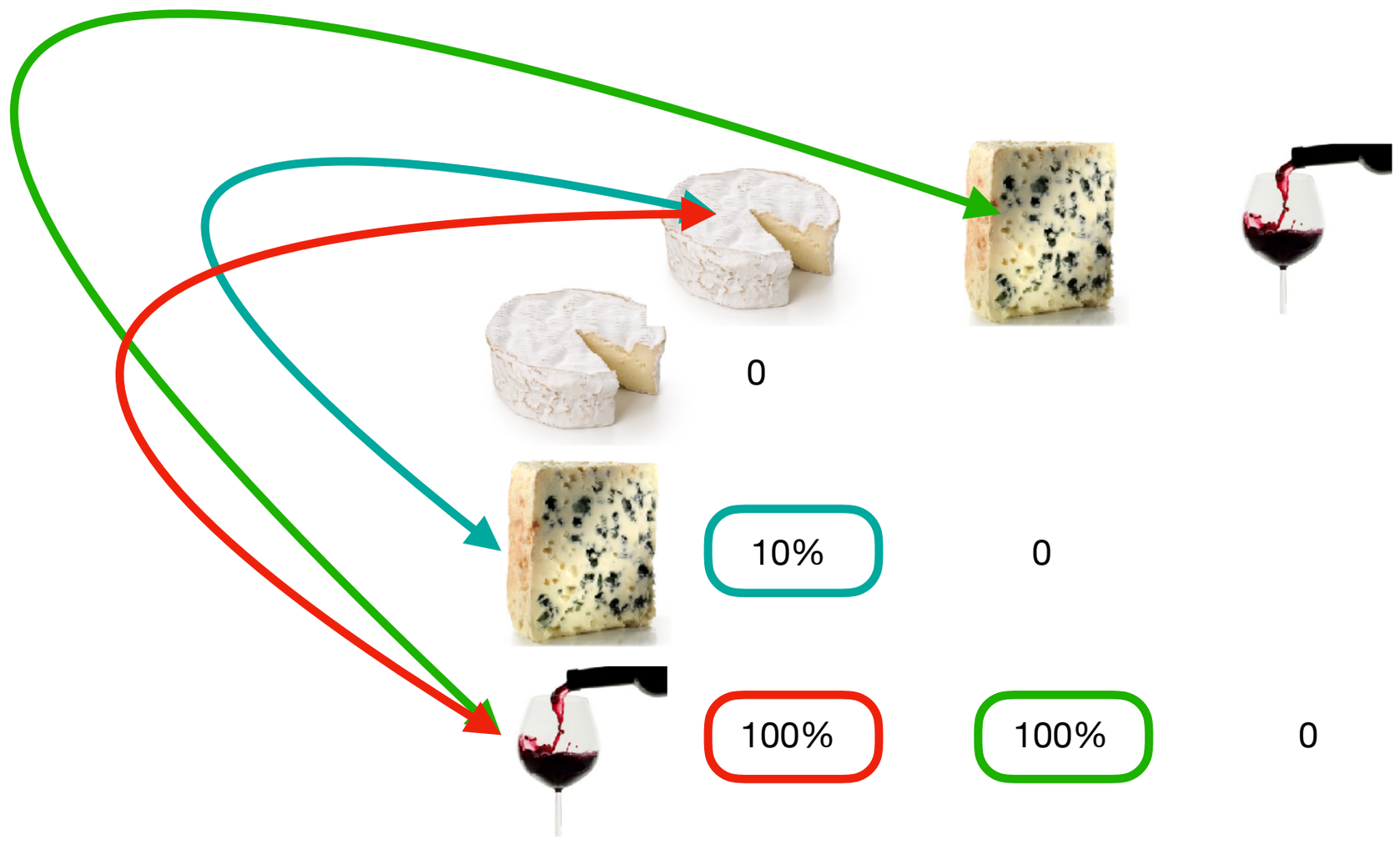


= pairwise measures
of distance or dissimilarities
between microbiota

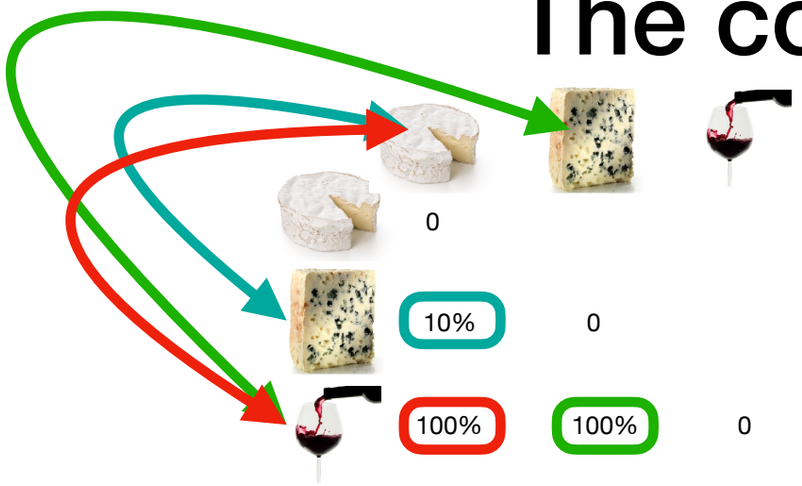
The composition of microbiomes

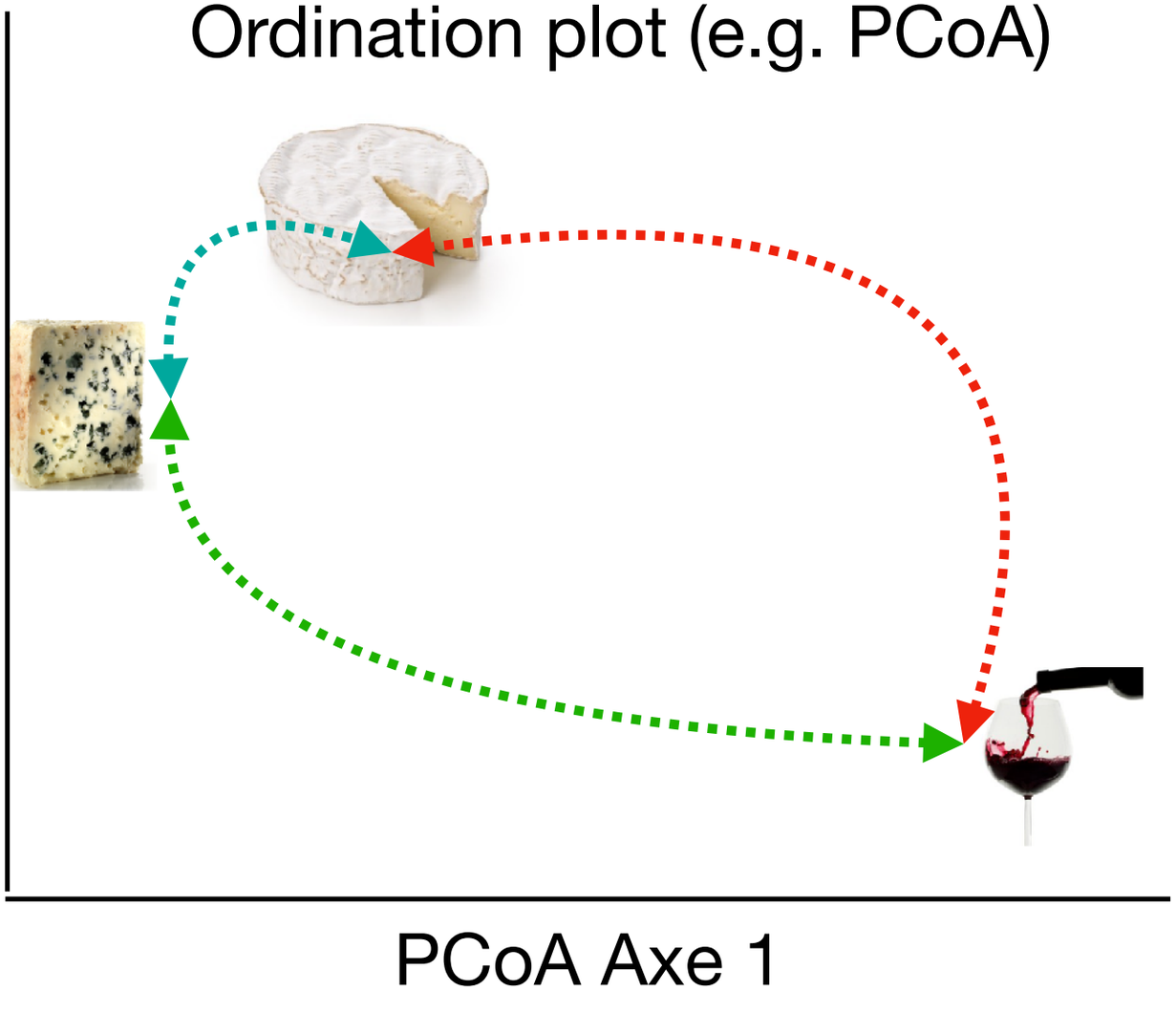


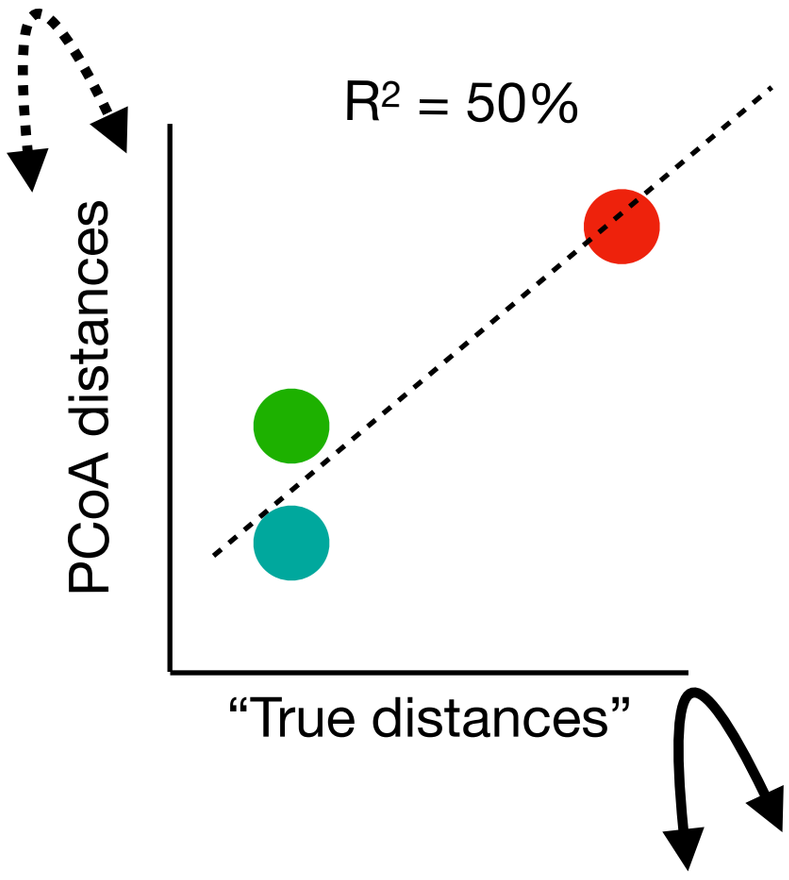
Distance matrix



The composition of microbiomes

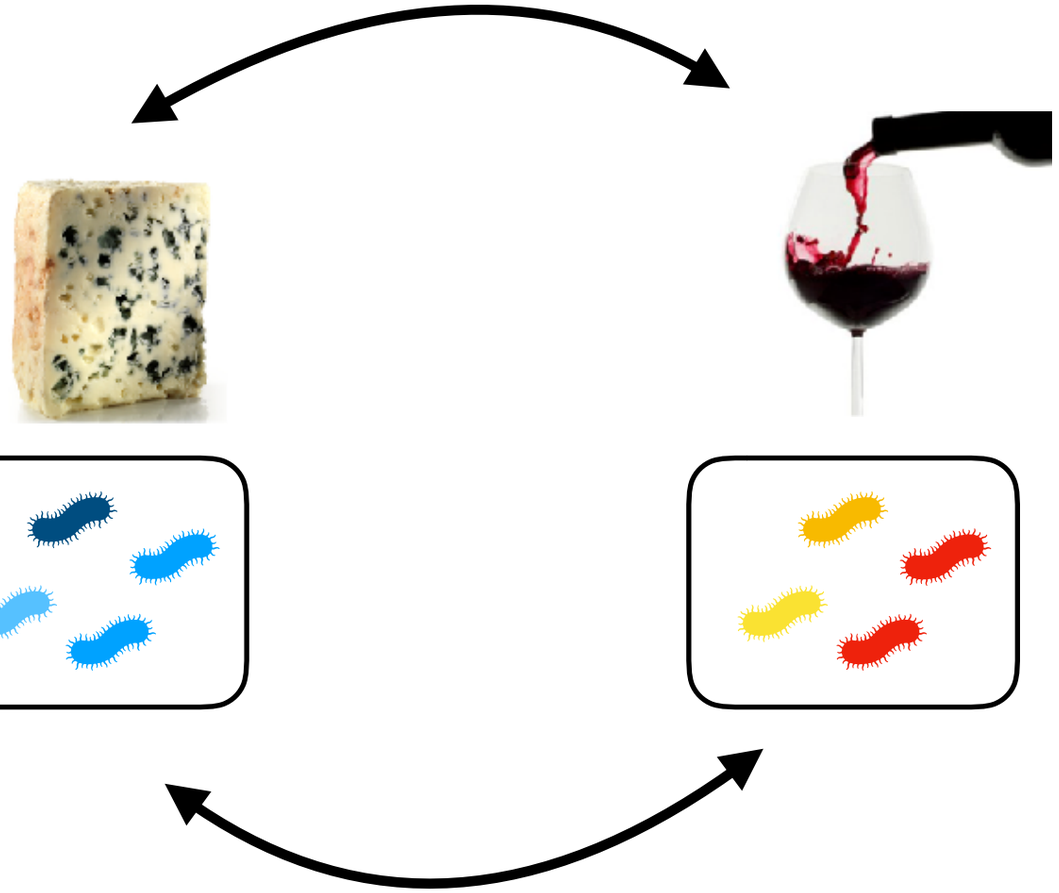






The composition of microbiomes

Change in environment

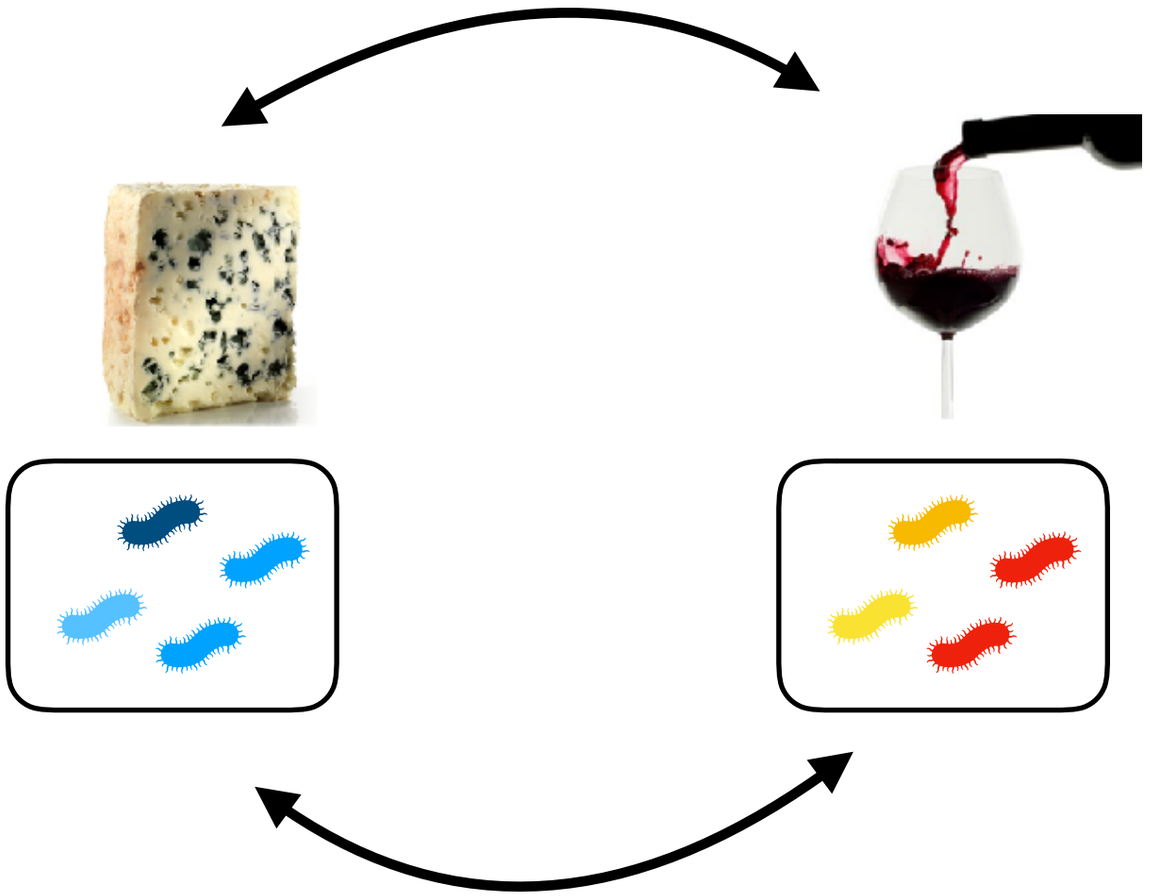


Change in microbiome

<u>OTUs composition</u>

The composition of microbiomes

Change in environment

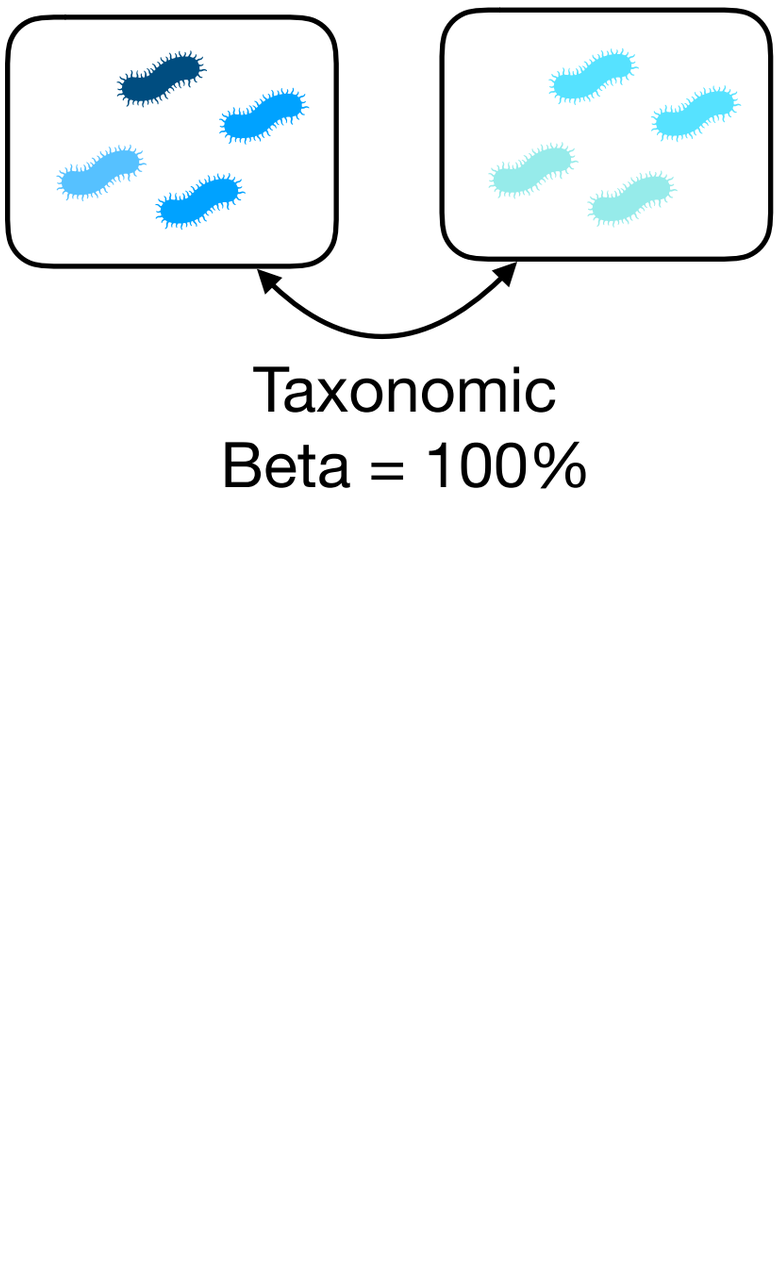


Change in microbiome

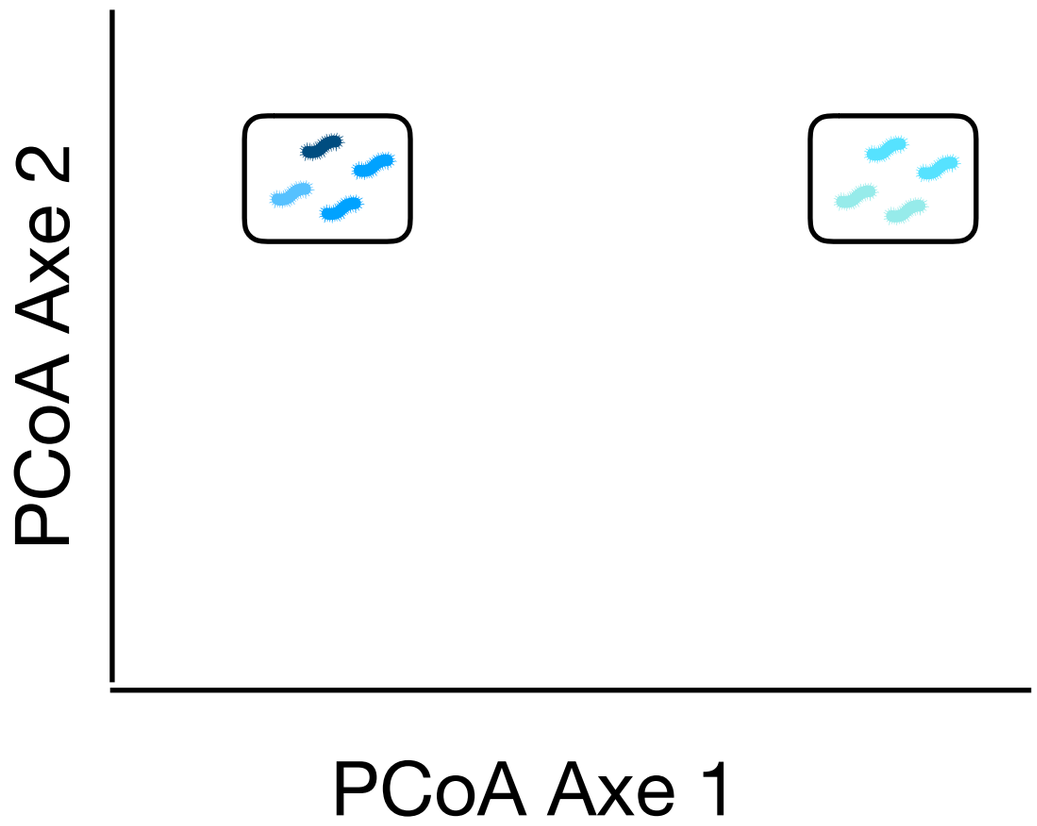
OTUs composition

Change in microbiome **functional composition**

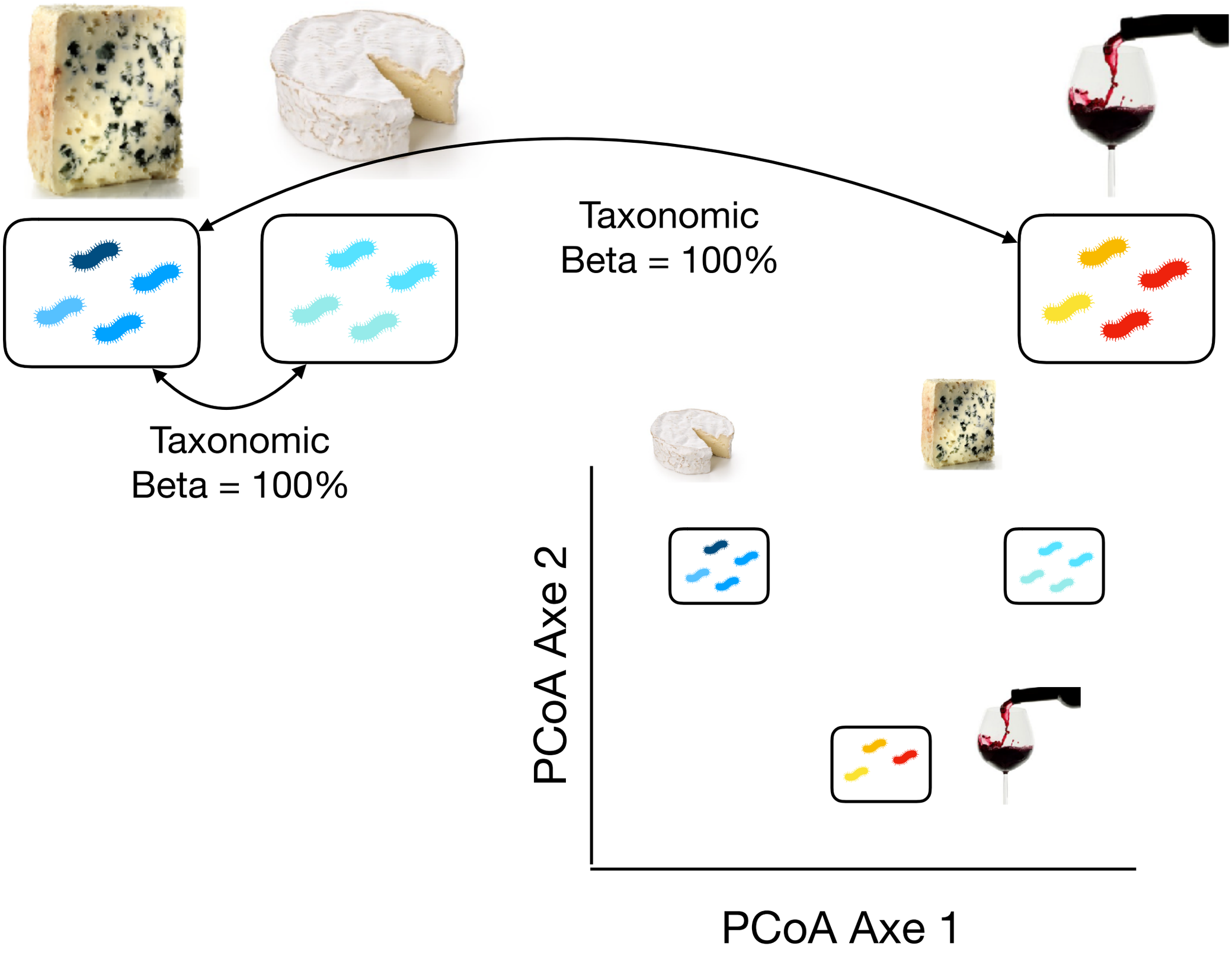
The limits of the OTU approach



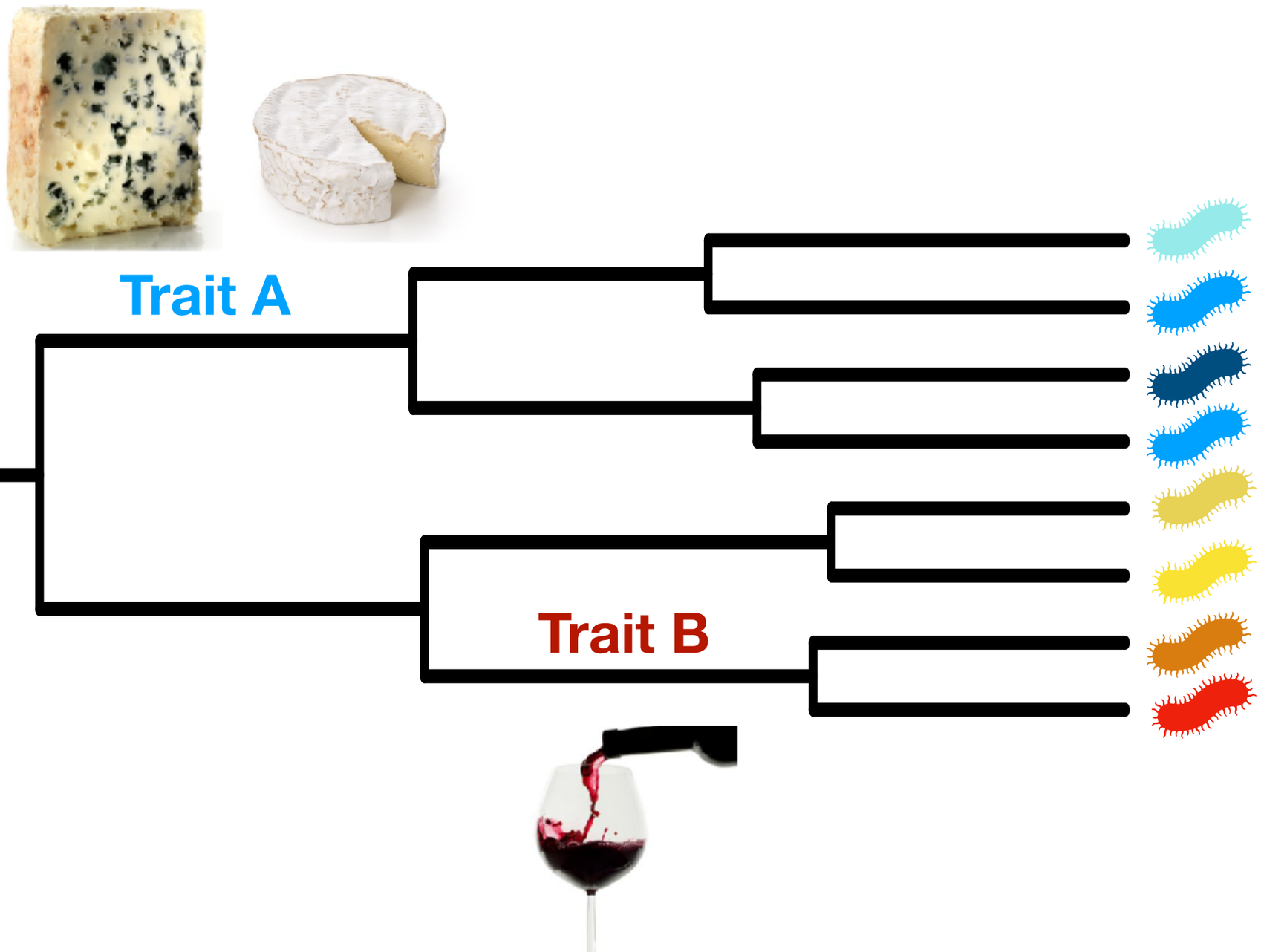
Taxonomic
Beta = 100%



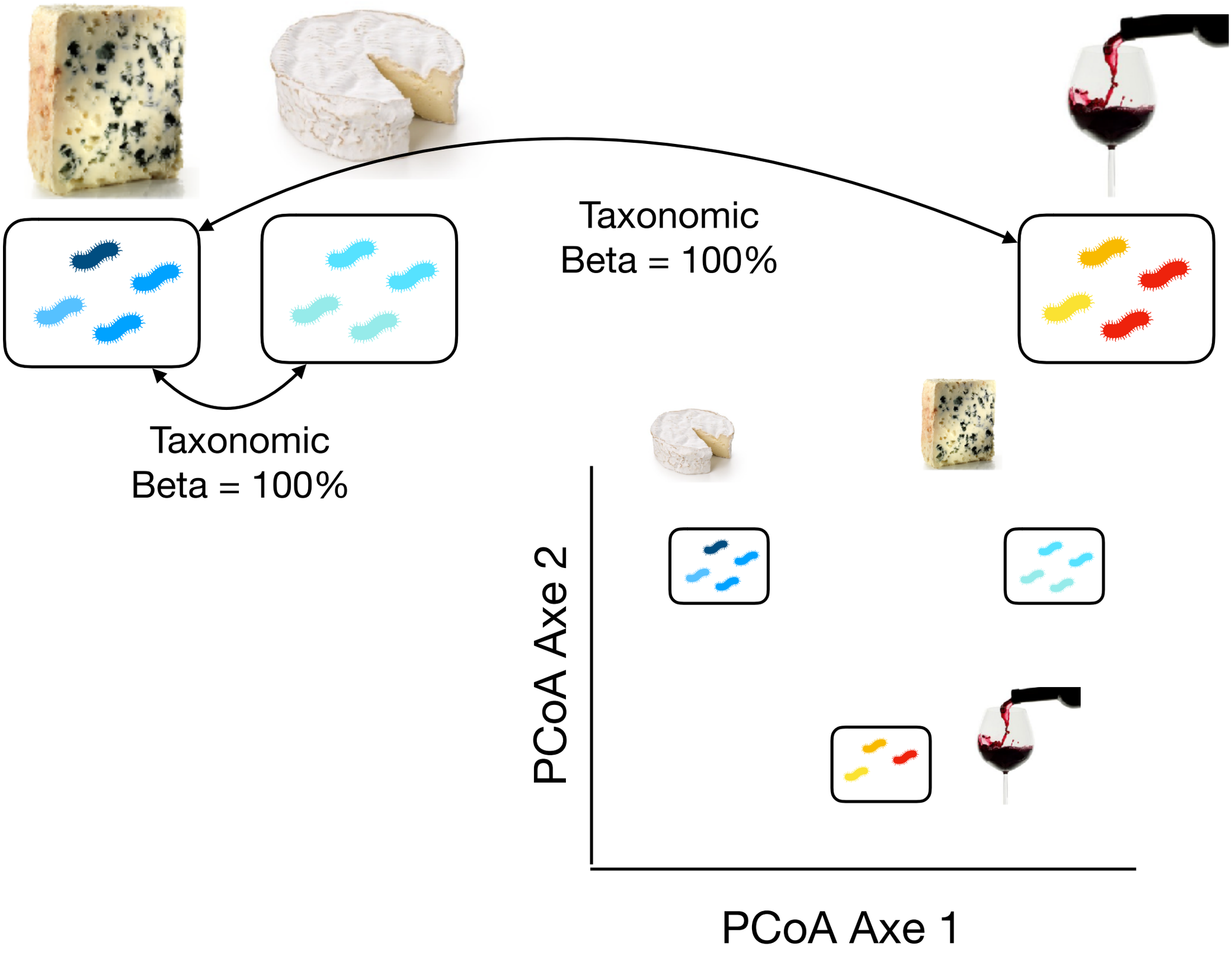
The limits of the OTU approach



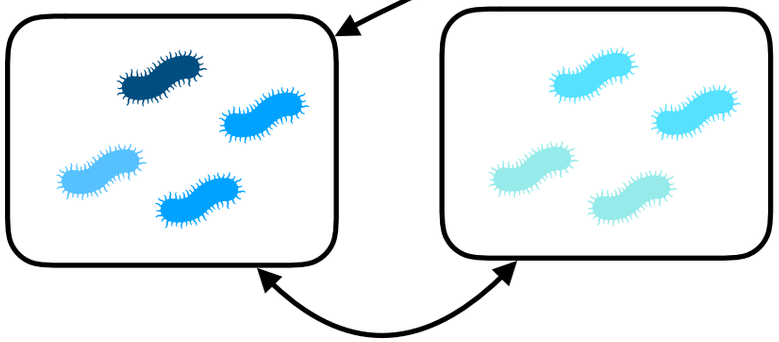
The phylogeny as a proxy for functions



The limits of the OTU approach

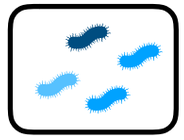


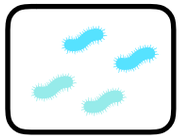
The limits of the OTU approach

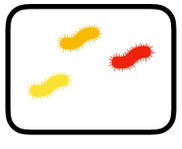


Phylogenetic Beta = 70%

Phylogenetic Beta = 40%







PCoA Axe 2

PCoA Axe 1

Objective

To better describe
the composition of microbiomes
using phylogenetic trees

Overall structure of the workshop

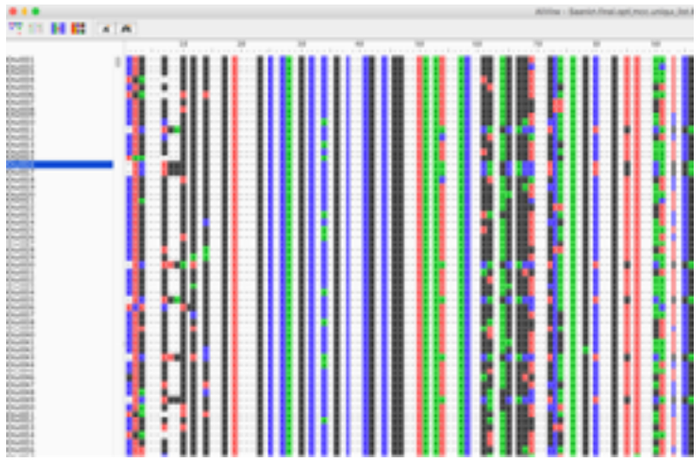
Overall structure of the workshop

0 — The dataset and the softwares

Overall structure of the workshop

0 — The dataset and the softwares

1 — Building a phylogenetic tree



Overall structure of the workshop

0 — The dataset and the softwares

1 — Building a phylogenetic tree

2 — Classical analysis of microbiome

Classical metrics	**Classical visualization**	**Classical test**
Bray-Curtis	PCoA	PERMANOVA
Unifrac	NMDS	

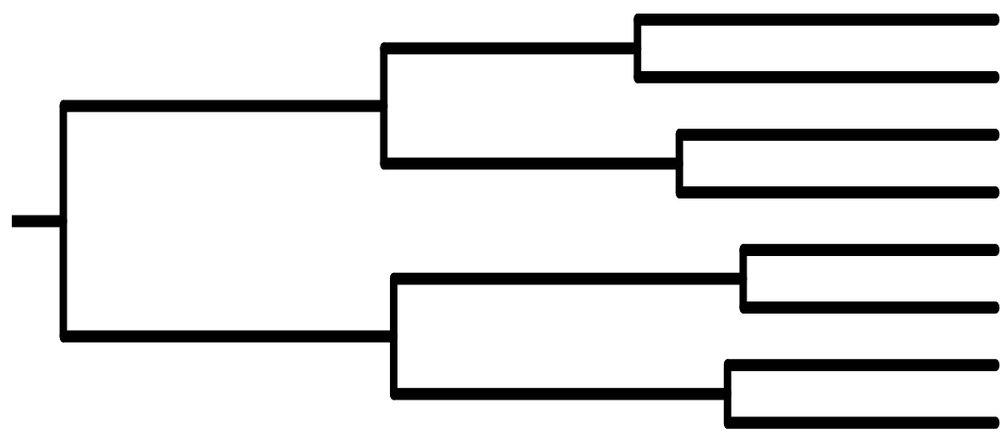
Overall structure of the workshop

0 — The dataset and the softwares

1 — Building a phylogenetic tree

2 — Classical analysis of microbiome

3 — Varying the phylogenetic resolution



Overall structure of the workshop

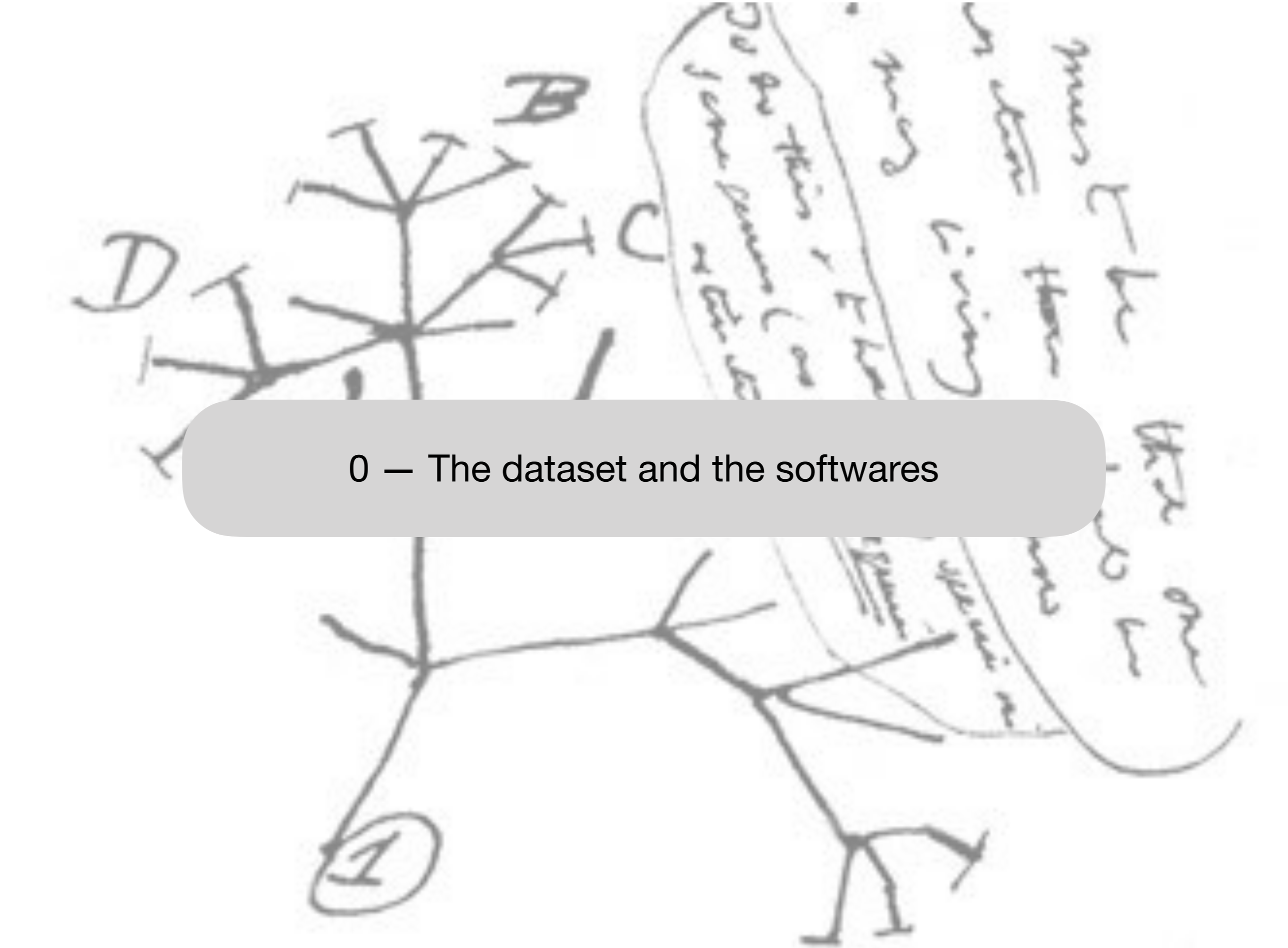
0 — The dataset and the softwares

1 — Building a phylogenetic tree

2 — Classical analysis of microbiome

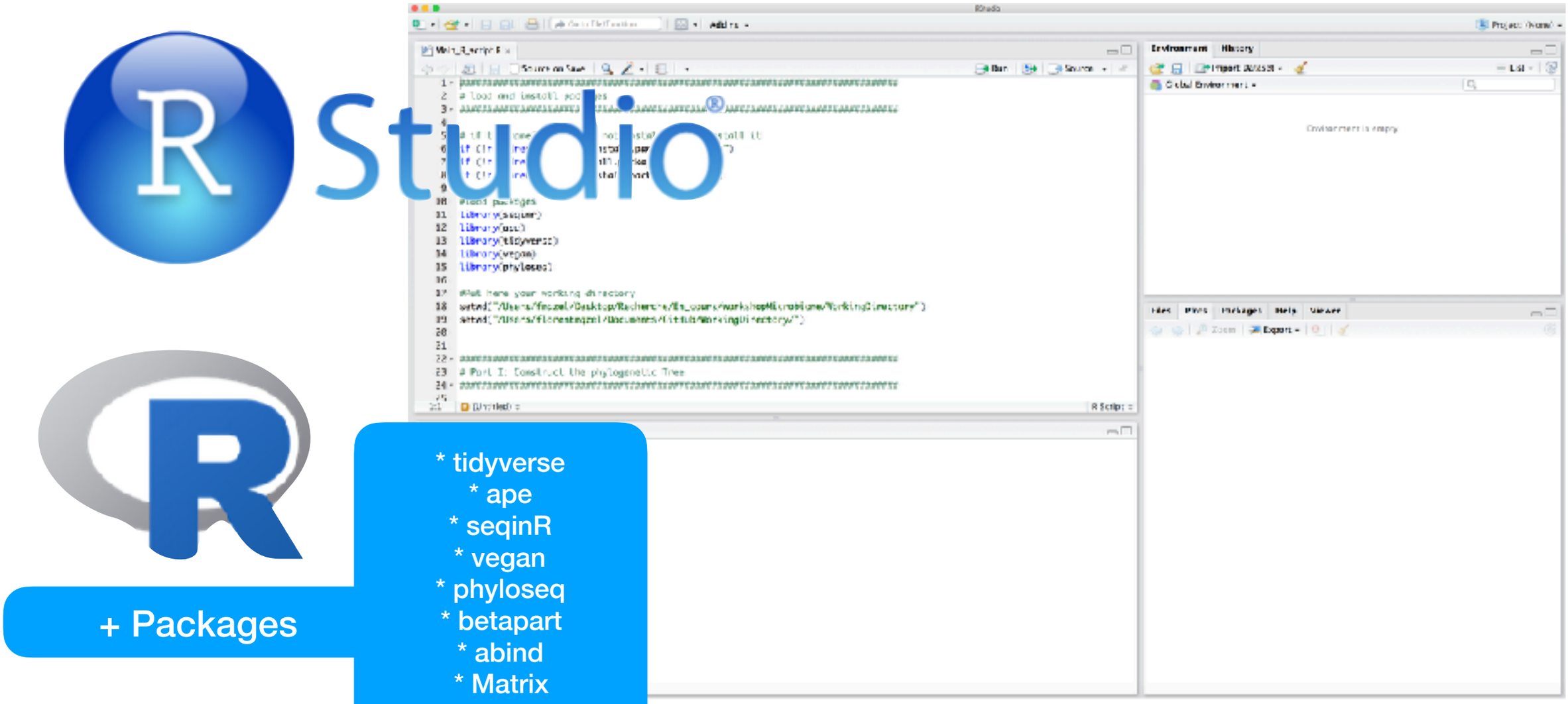
3 — Varying the phylogenetic resolution

4 — Exploring the branches of the phylogenetic tree



0 — The dataset and the softwares

Getting started with softwares



+ Packages

FastTree

Where to find the code / the data ?

All material available here:

https://github.com/FloMazel/Microbiome_Phylo_Diversity_Workshop

Organisation of the folders

Practice

39 commits | 2 branches | 0 releases | 2 contributors

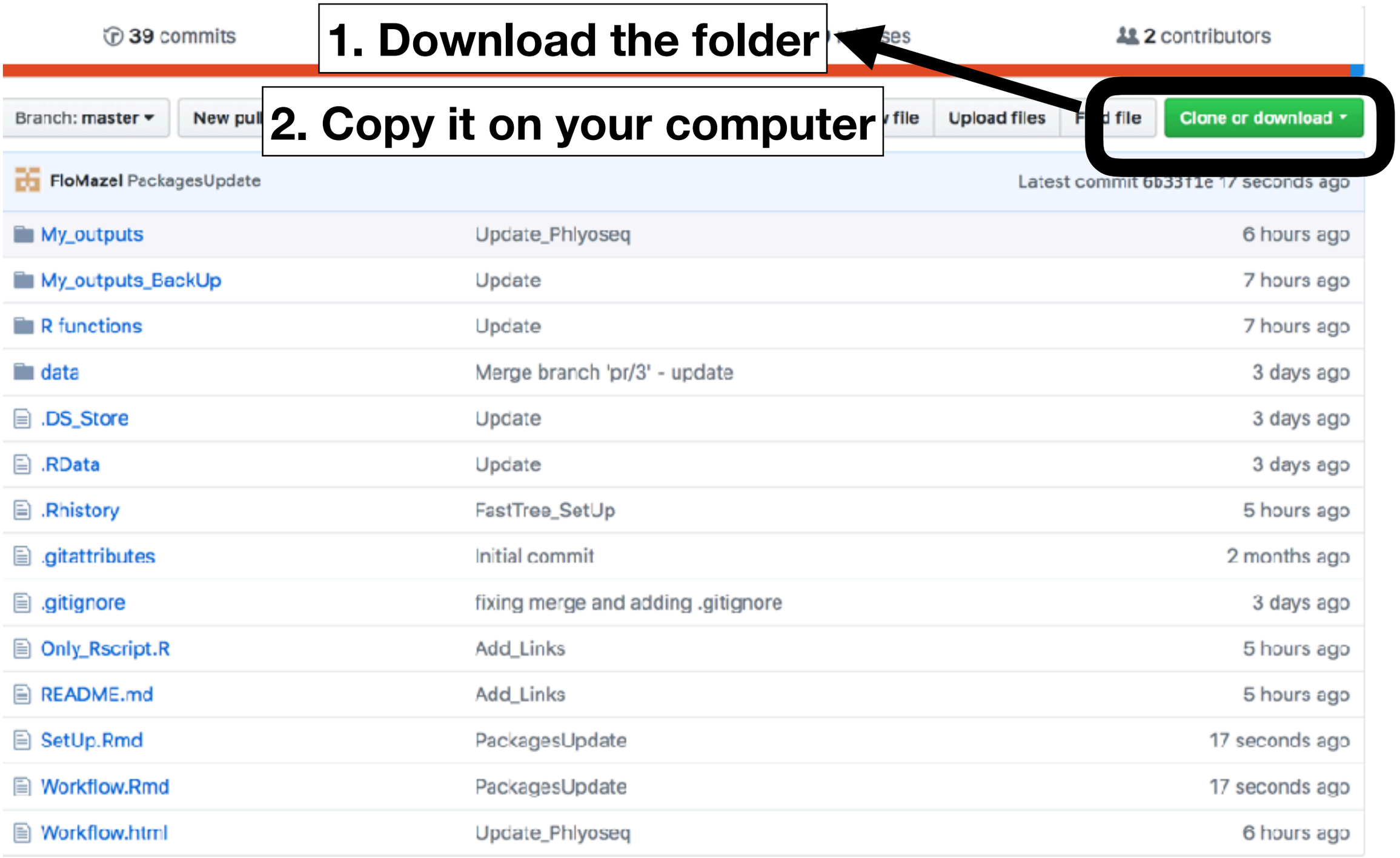
Branch: master ▾ | New pull request | Create new file | Upload files | Find file | Clone or download ▾

FloMazel PackagesUpdate — Latest commit 6b33f1e 17 seconds ago

My_outputs	Update_Phlyoseq	6 hours ago
My_outputs_BackUp	Update	7 hours ago
R functions	Update	7 hours ago
data	Merge branch 'pr/3' - update	3 days ago
.DS_Store	Update	3 days ago
.RData	Update	3 days ago
.Rhistory	FastTree_SetUp	5 hours ago
.gitattributes	Initial commit	2 months ago
.gitignore	fixing merge and adding .gitignore	3 days ago
Only_Rscript.R	Add_Links	5 hours ago
README.md	Add_Links	5 hours ago
SetUp.Rmd	PackagesUpdate	17 seconds ago
Workflow.Rmd	PackagesUpdate	17 seconds ago
Workflow.html	Update_Phlyoseq	6 hours ago

Organisation of the folders

Practice



Organisation of the folders

Practice

1. Download the folder

2. Copy it on your computer

3. The different files

39 commits | 2 contributors

Branch: master | New pull | file | Upload files | file | Clone or download

FloMazel PackagesUpdate — Latest commit 6b33f1e 17 seconds ago

My_outputs		6 hours ago
My_outputs_BackUp	Update	7 hours ago
R functions	Update	7 hours ago
data	Merge branch 'pr/3' - update	3 days ago
.DS_Store		3 days ago
.RData		3 days ago
.Rhistory		5 hours ago
.gitattributes		2 months ago
.gitignore	fixing merge and adding .gitignore	3 days ago
Only_Rscript.R	Add_Links	5 hours ago
README.md	Add_Links	5 hours ago
SetUp.Rmd	PackagesUpdate	17 seconds ago
Workflow.Rmd	PackagesUpdate	17 seconds ago
Workflow.html	Update_Phlyoseq	6 hours ago

(i) Data
(ii) R-scripts and workflow
(iii) Outputs (and backup)

Organisation of the folders

Practice

Exploring the phylogenetic composition of microbiomes

Some of this code as been adapted from the ECOSCOPE GitHub account

Table of contents

1. Prior to the workshop

Setup

Please come to the workshop with your laptop set up with the required software and data files as described in our setup instructions.

Background

Please read Hallam SJ *et al.* 2017. Sci Data 4: 170156 "Monitoring microbial responses to ocean deoxygenation in a model oxygen minimum zone" to learn more about the data used in this workshop. You can also check out this short video showing how the sampling was done!

Data description

The data in this workshop were collected as part of an on-going oceanographic time series program in Saanich Inlet, a seasonally anoxic fjord on the East coast of Vancouver Island, British Columbia

2. Getting started with R

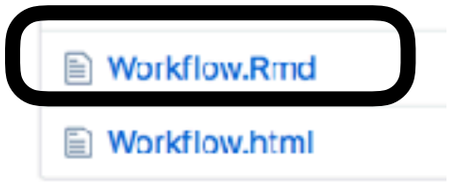
Installing and loading packages

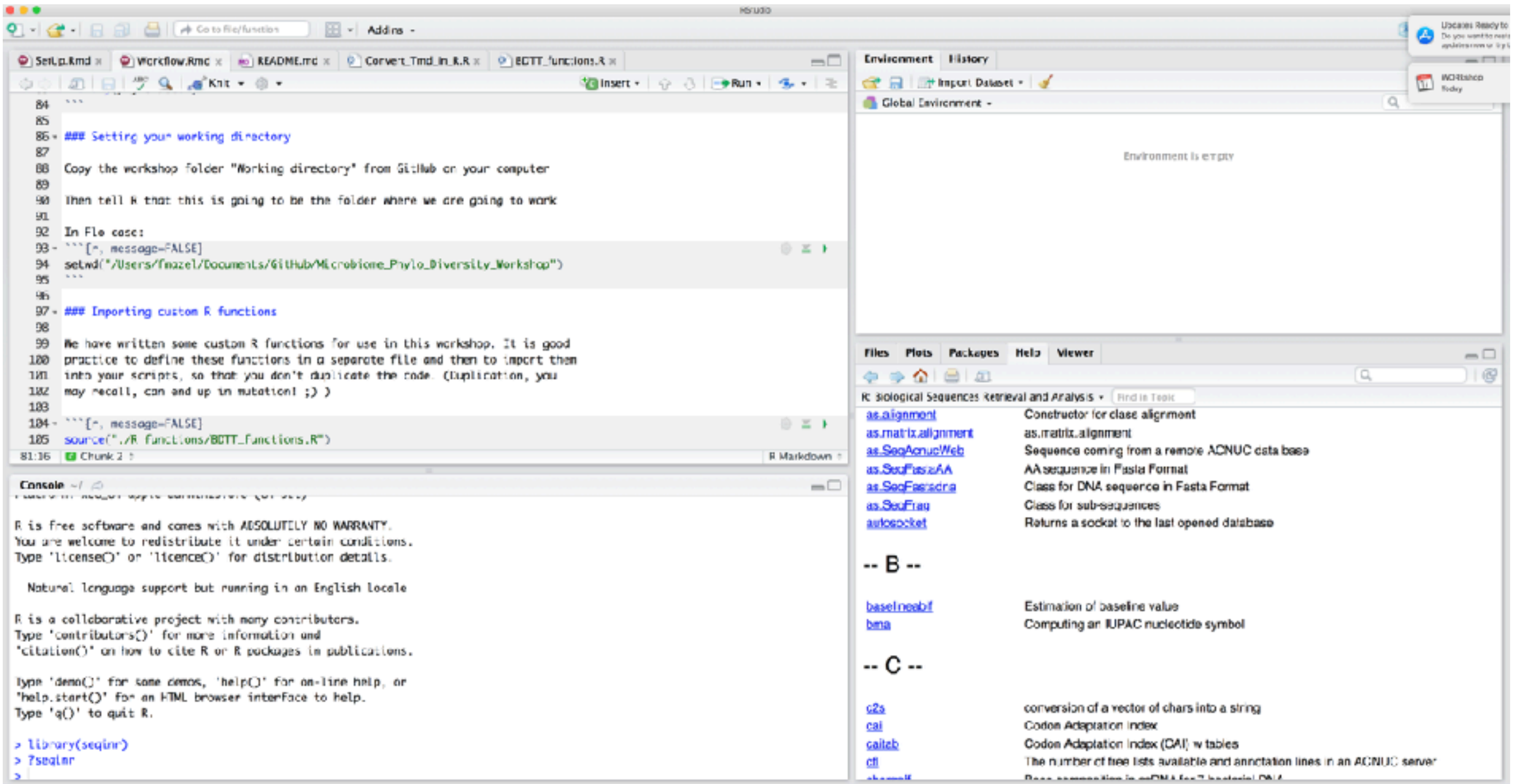
At the beginning of every R script, you should have a dedicated space for loading R packages. R packages allow any R user to code reproducible functions and share them with the R community. Packages exist for anything ranging from microbial ecology to complex graphics to multivariate modelling and beyond.

In this workshop, we will use several packages:

Organisation of the folders

Practice

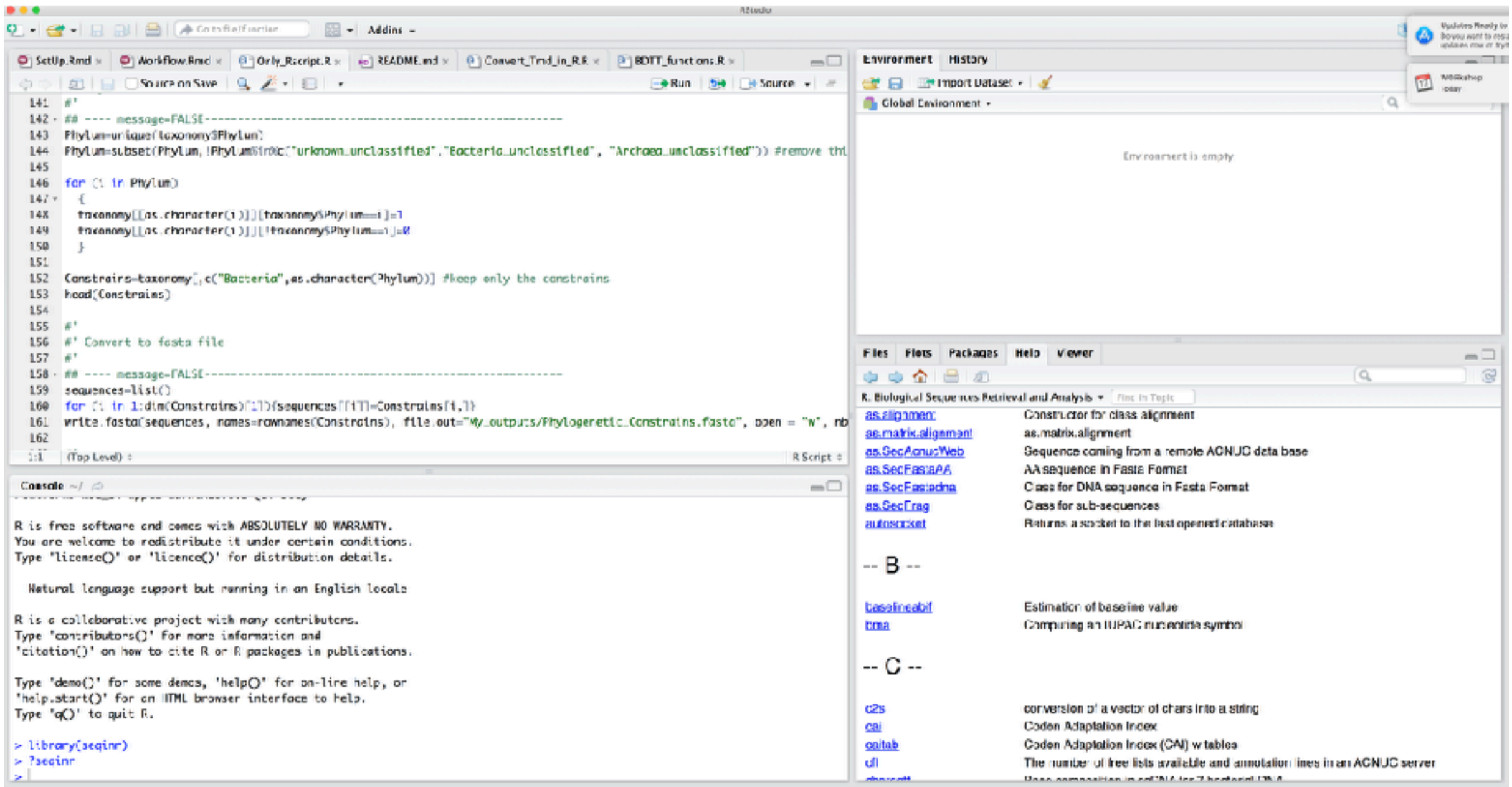




Organisation of the folders

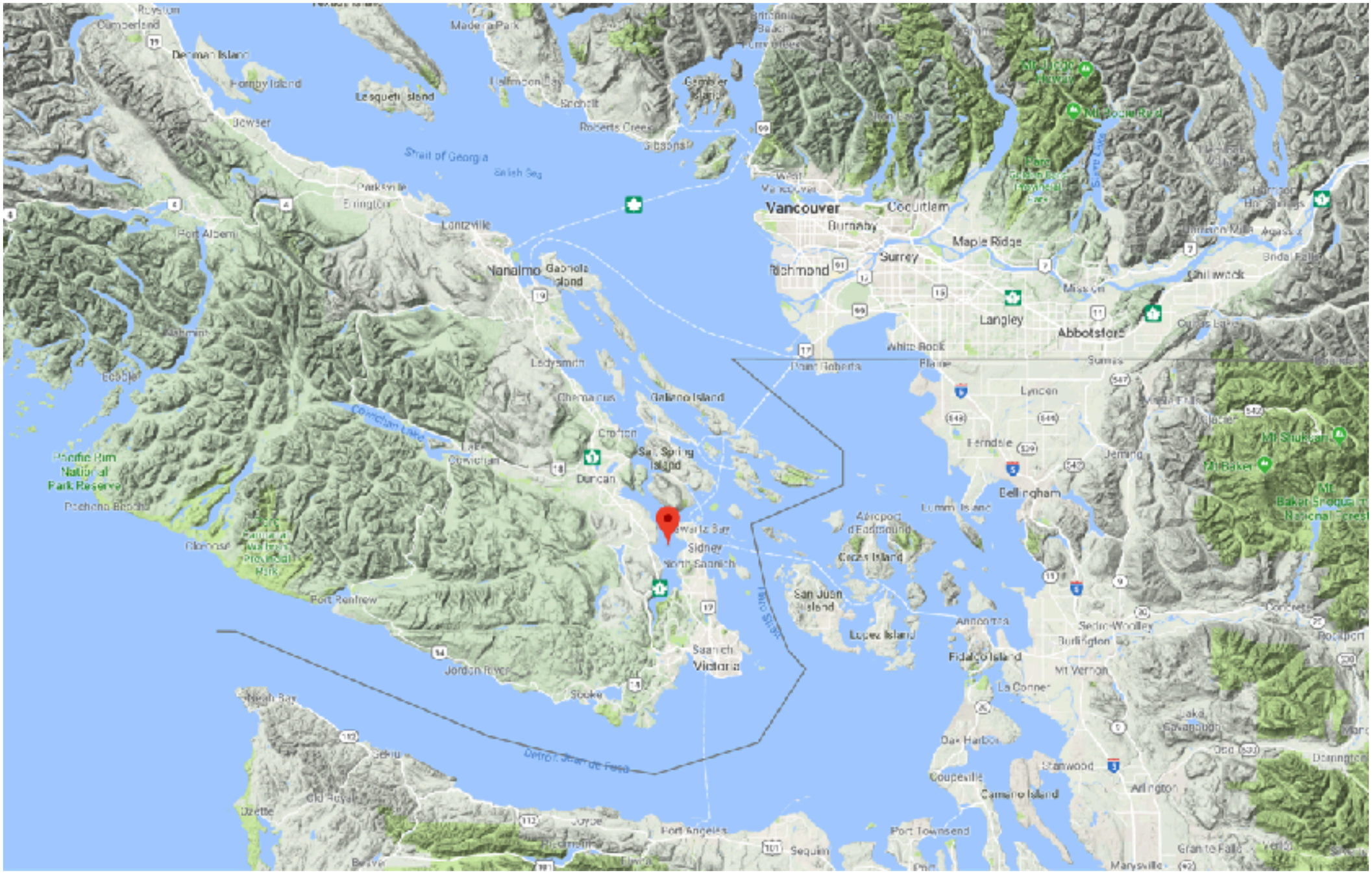
Practice

Only_Rscript.R

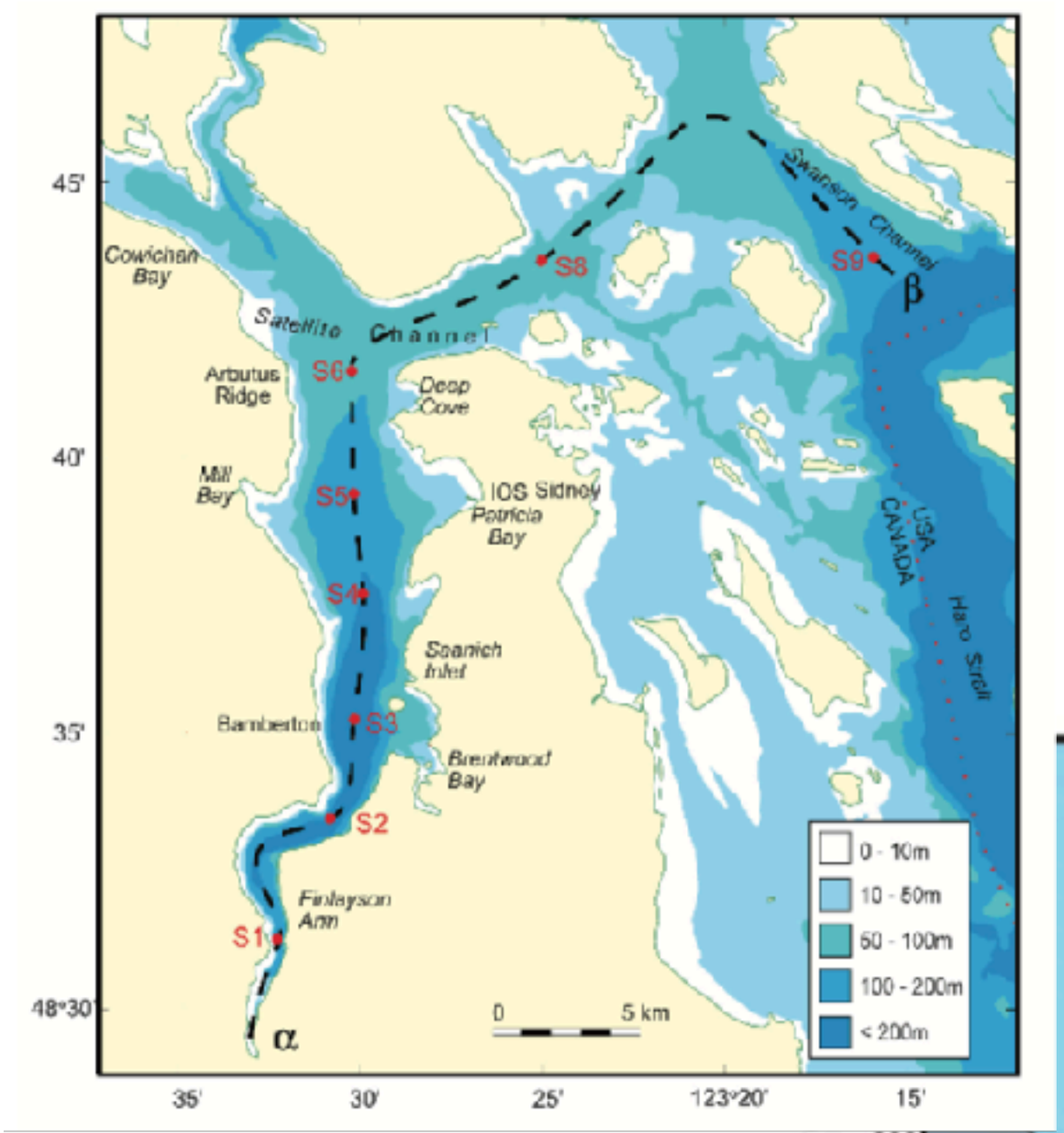


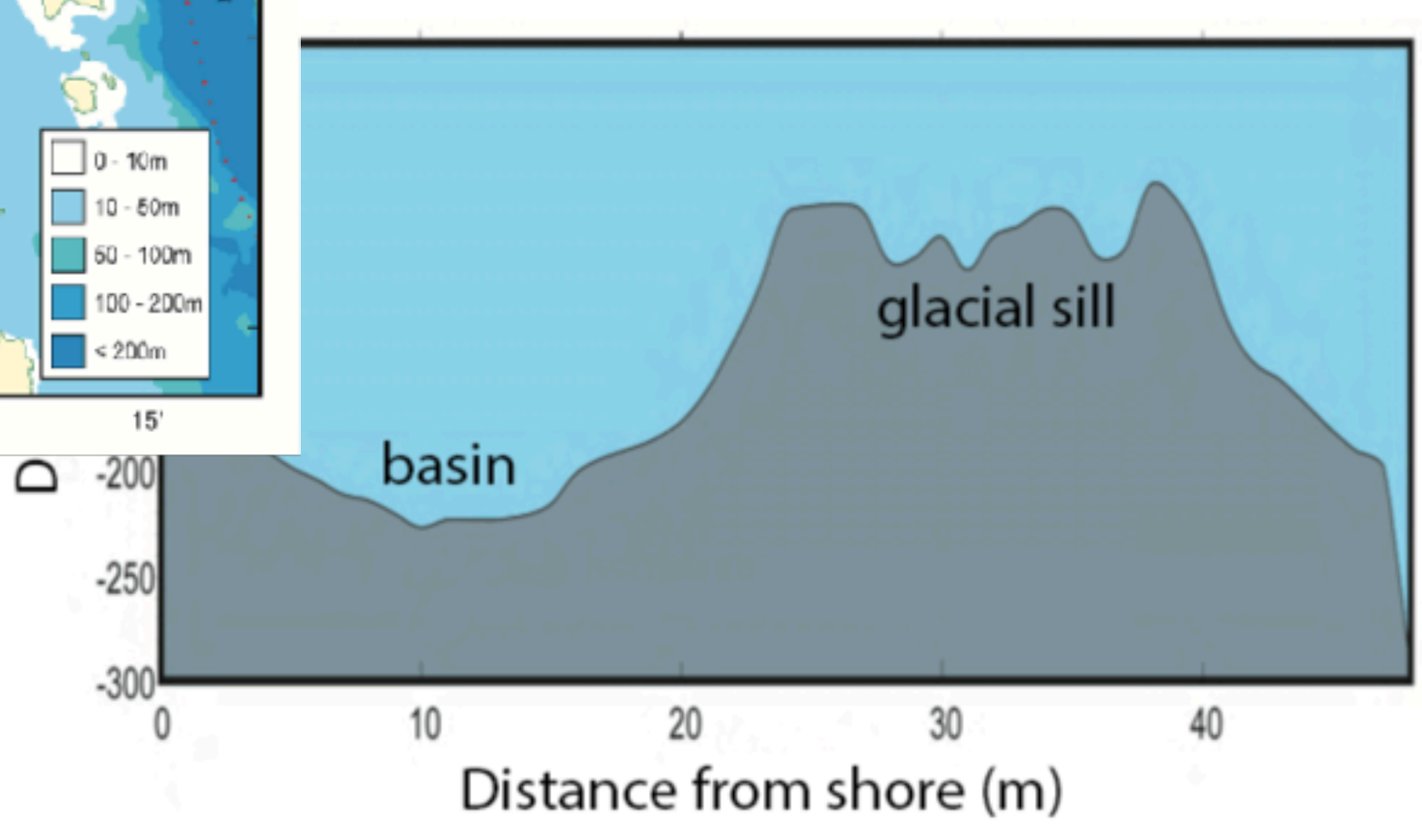
The data

Water columns samples from Saanish inlet



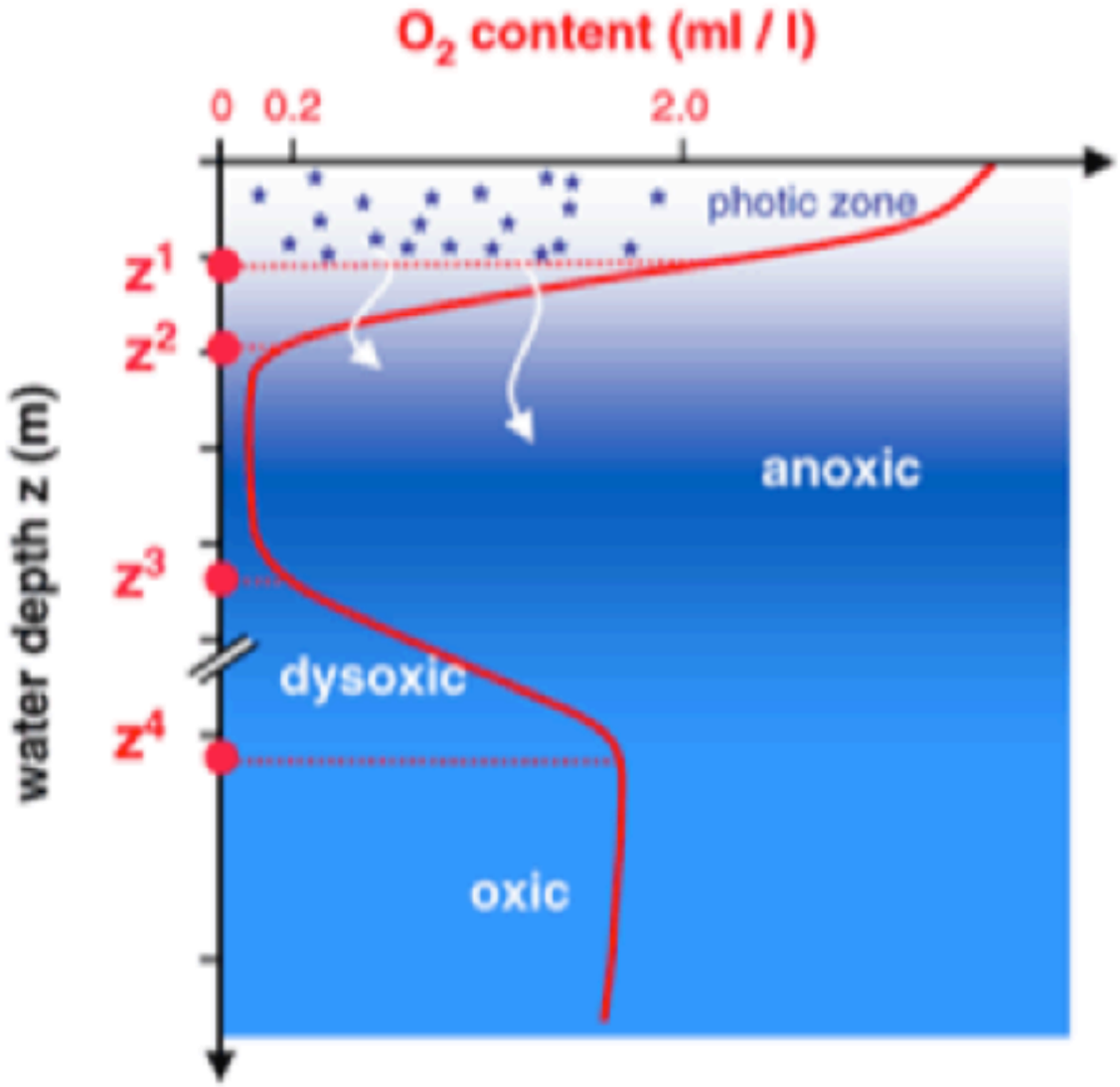
The data


45'
40'
35'
48°30'
35'
30'
25'
123°20'
15'
Cowichan Bay
Satellite Channel
Arbutus Ridge
Deep Cove
Mill Bay
IOS Sidney
Patricia Bay
Saanich Inlet
Brentwood Bay
Finlayson Arm
Swanson Channel
USA
CANADA
Haro Strait
S1
S2
S3
S4
S5
S6
S8
S9
α
β
0 5 km
0 - 10m
10 - 50m
50 - 100m
100 - 200m
< 200m


glacial sill
basin
-200
-250
-300
0
10
20
30
40
Distance from shore (m)

The data

A model for
Oxygen Minimum Zones
(OMZ)



The data

Branch: master ▾ Microbiome_Phylo_Diversity_Workshop / data / | Create new file | Upload files | Find file | History

FloMazel a — Latest commit da2db6e 16 days ago

..		
mothur_intermediate_files		16 days ago
.DS_Store		16 days ago
Saanich_cruise72_metadata.txt		2 months ago
Saanich_cruise72_mothur_NJ_tree.RData	Initial commit	2 months ago
Saanich_cruise72_mothur_OTU_table.shared		2 months ago
Saanich_cruise72_mothur_OTU_taxonomy.taxonomy		2 months ago
Saanich_cruise72_mothur_phyloseq.RData		2 months ago
Saanich_timeseries_geochemical_data.csv	Initial commit	2 months ago
mothur_pipeline.html	Initial commit	2 months ago

Fasta files

Metadata

OTU table

Taxonomy table

1 — Building a phylogenetic tree

Theory

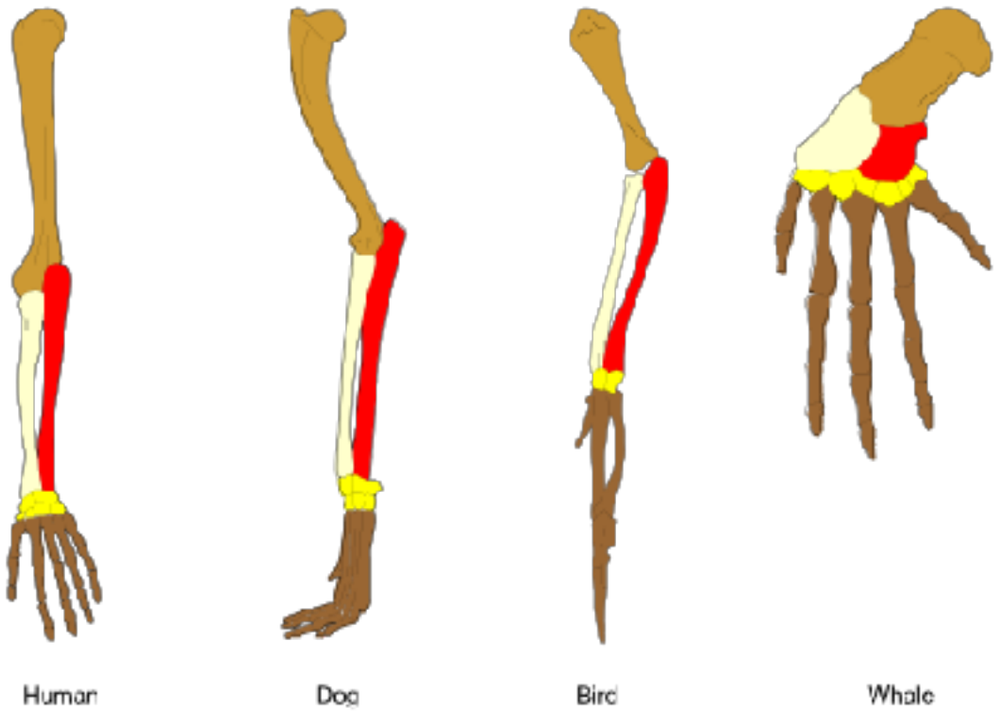
(1) Define homologous characters

(2) Construct a tree

Theory

(1) Define homologous characters

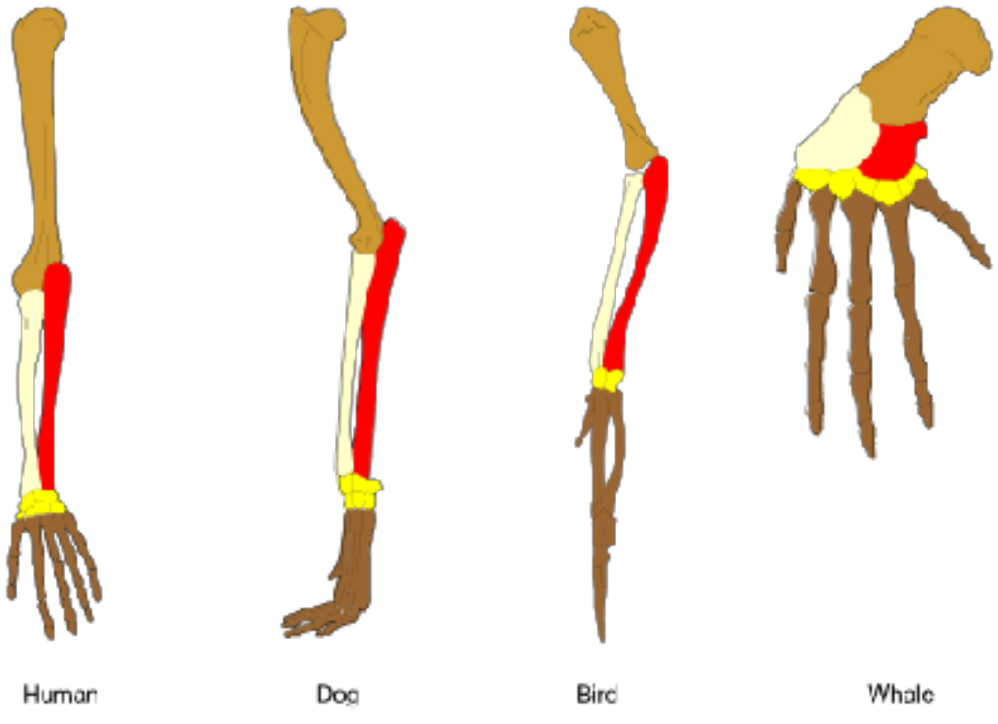
Morphological data



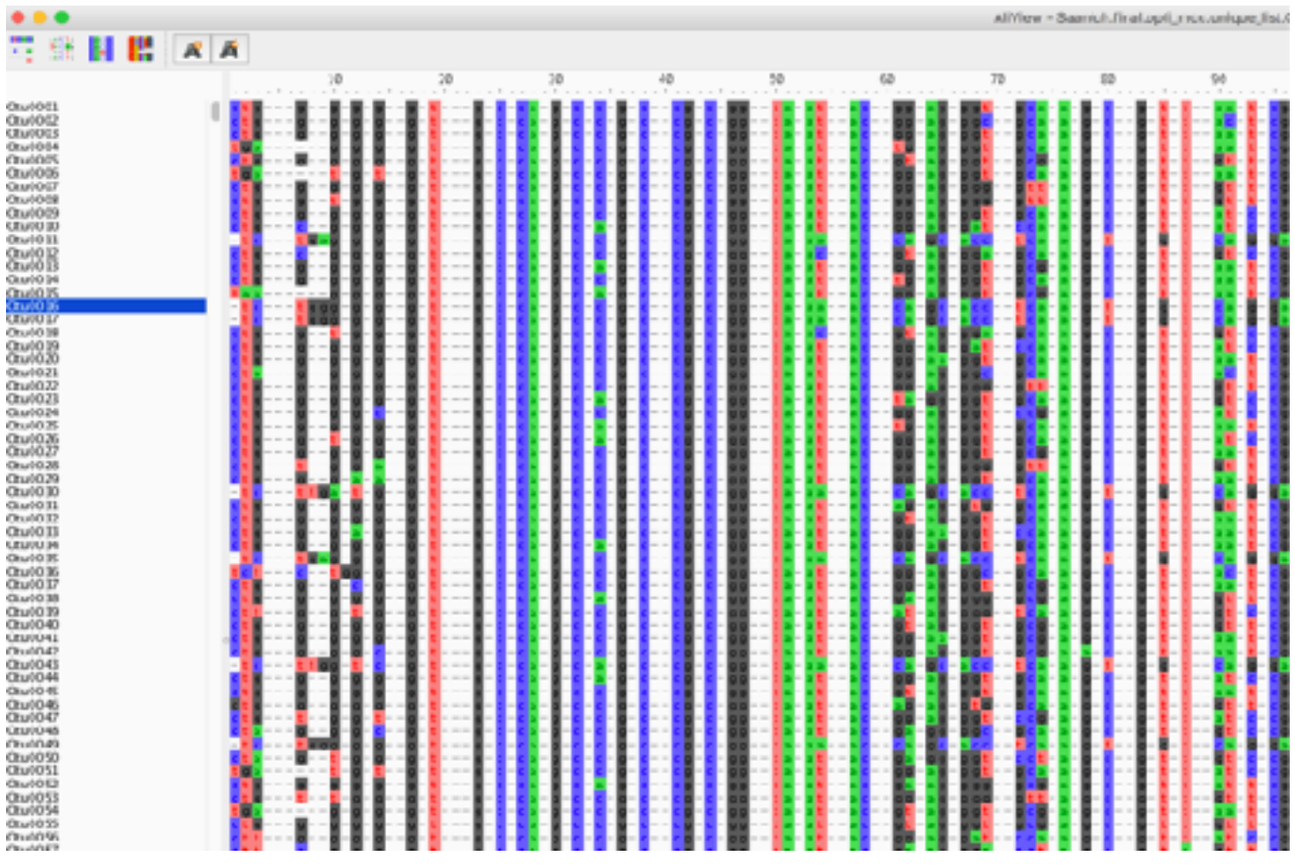
Theory

(1) Define homologous characters

Morphological data



Molecular data

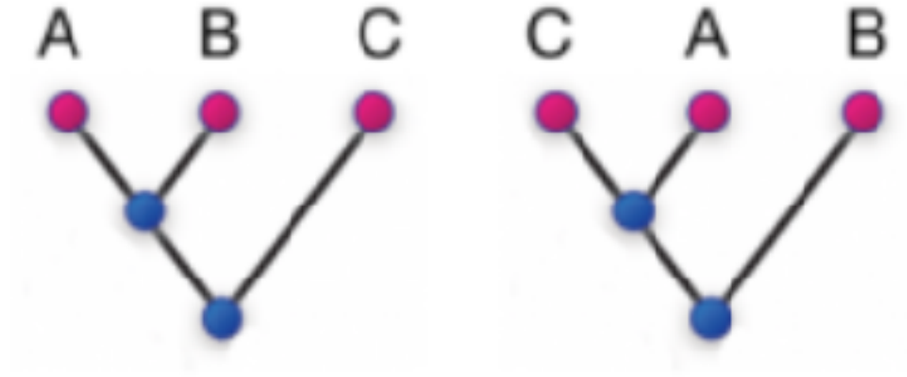


Theory

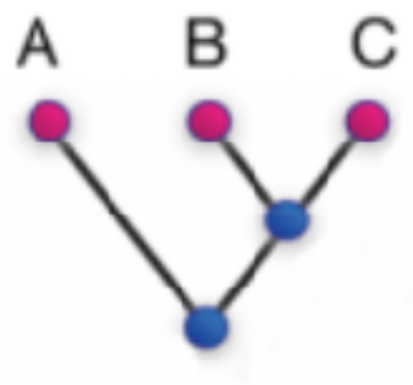
(2) Construct a tree

Idea: *Define a **criteria** to rank trees & **explore tree space***

Parsimony approach



Maximum likelihood



Theory

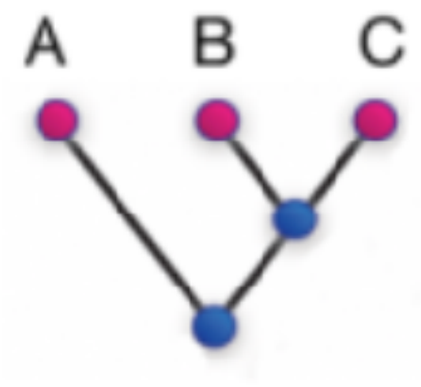
(2) Construct a tree

Idea: *Define a **criteria** to rank trees & **explore tree space***

Parsimony approach

#ParsimonyGate

Maximum likelihood



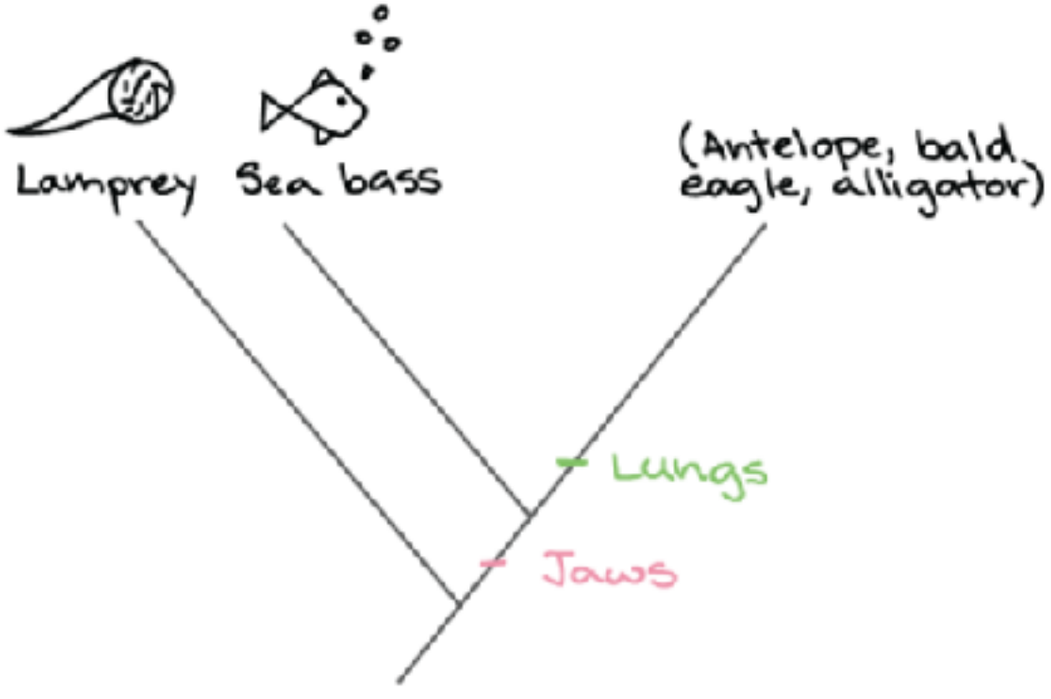
Theory

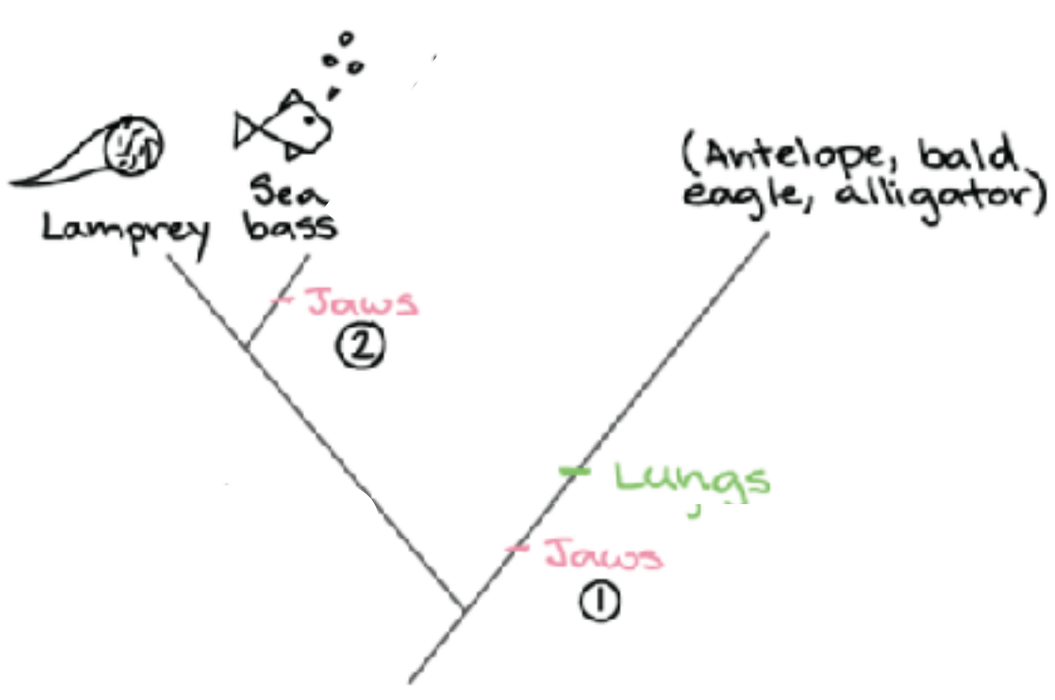
(2) Construct a tree

Idea: *Define a **criteria** to rank trees & **explore tree space***

Parsimony approach

The scenario with the minimal number of changes is the best

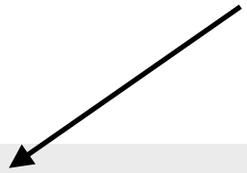




Theory

(2) Construct a tree

Idea: *Define a **criteria** to rank trees & **explore tree space***



Maximum likelihood

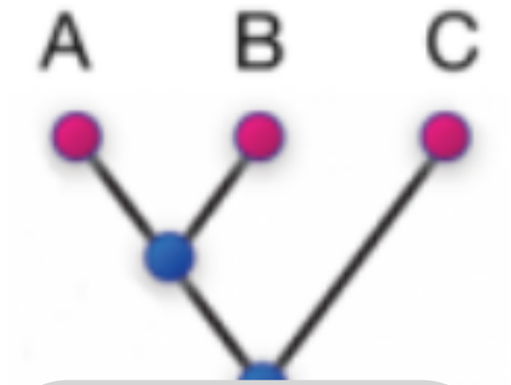
Theory

(2) Construct a tree

Idea: *Define a **criteria** to rank trees & **explore tree space***

Maximum likelihood

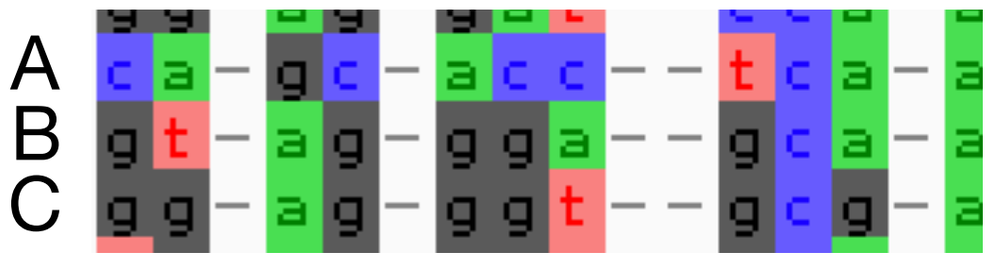
Model



Tree

substitution model

Data



Fit: Estimate parameters + compute likelihood

Theory

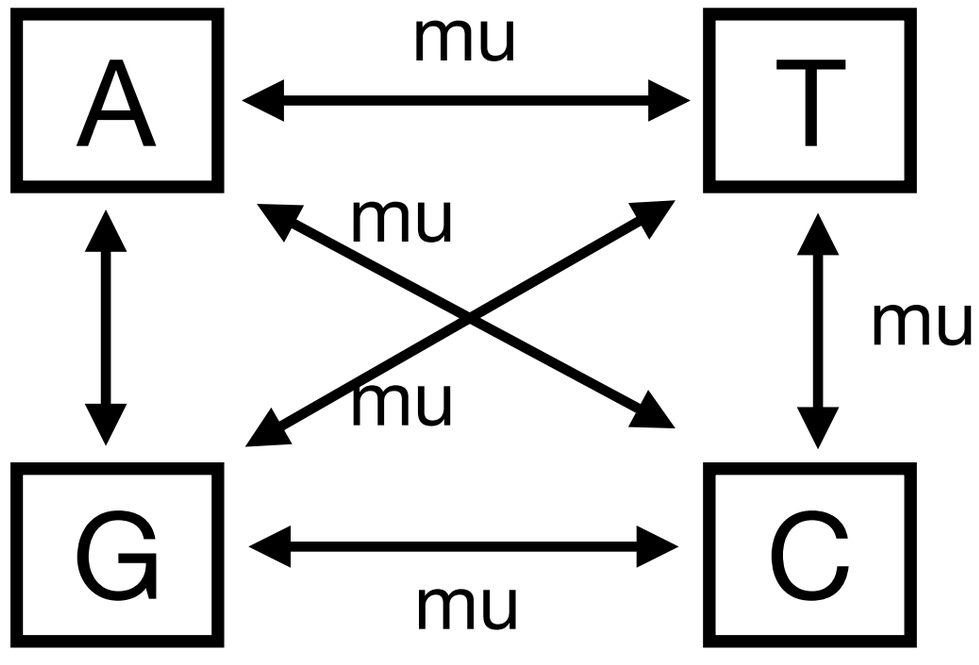
(2) Construct a tree

Idea: *Define a **criteria** to rank trees & **explore tree space***

Maximum likelihood

Use of a *substitution model*

A lot of different models exist
(they vary in parameters richness)



mu = transition rate

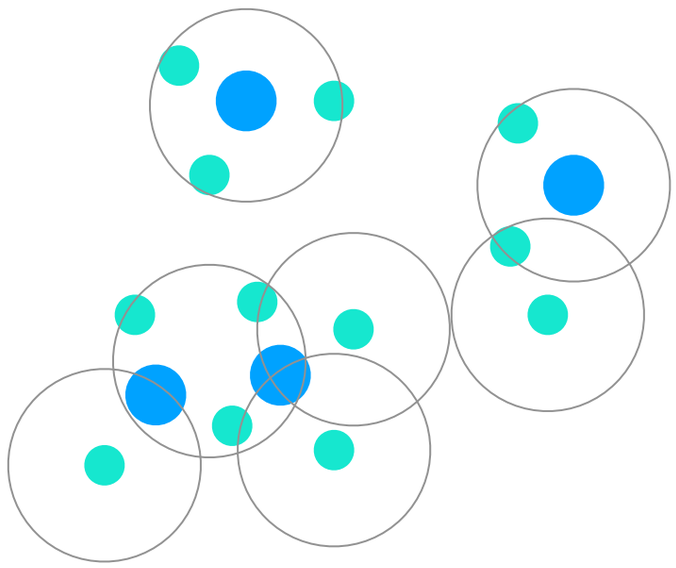
Theory

New methods for microbiomes:

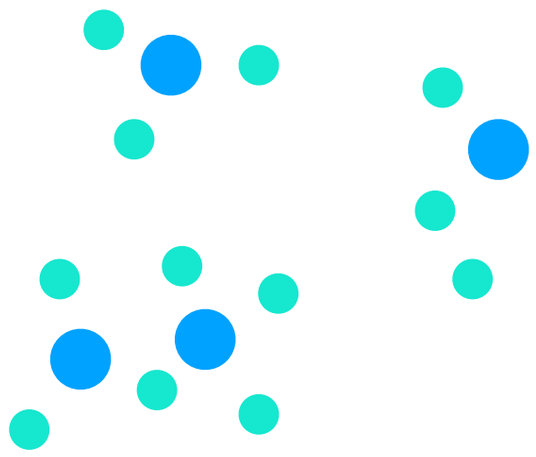
Exact Sequence Variants

- DADA2, Deblur
- Stable IDs
- Less inflation

OTUs



ESVs

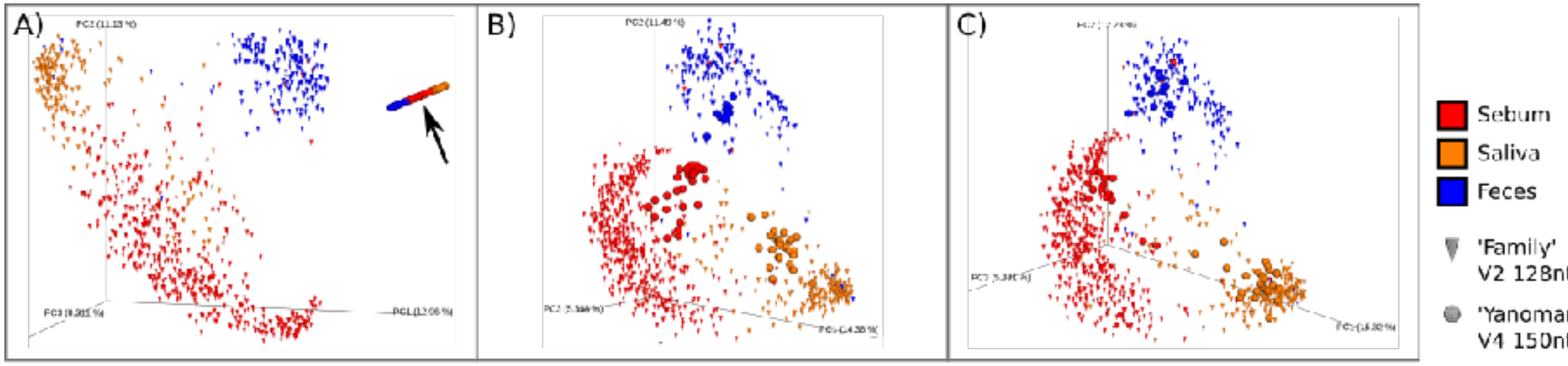


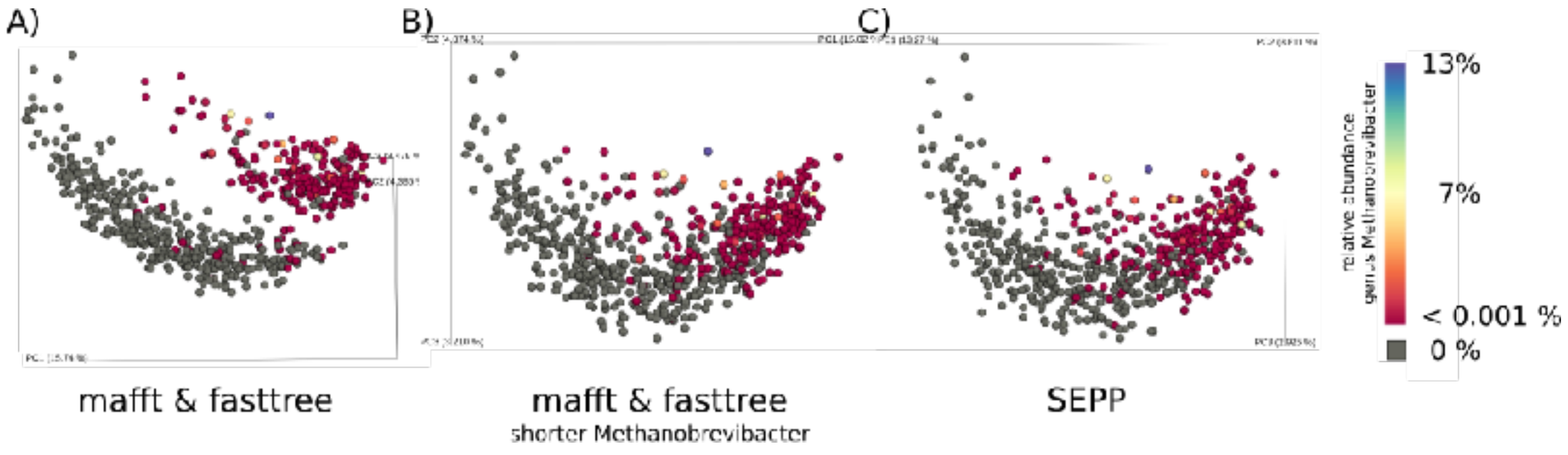
Theory

New methods for microbiomes:

Phylogenetic Insertion

- q2-fragment-insertion
- fewer artifacts
- more stable meta-analysis





Theory

New methods for microbiomes:

Phylogenetic Insertion

Exact Sequence Variants

- both available as Qiime2 plugins
- not covered further here :)

Practice

Build the tree

Part 1 — Constrains

Part 2 — Tree search

Build the tree — Part 1: Constrains

Build the tree — Part 1: Constrains

Rationale

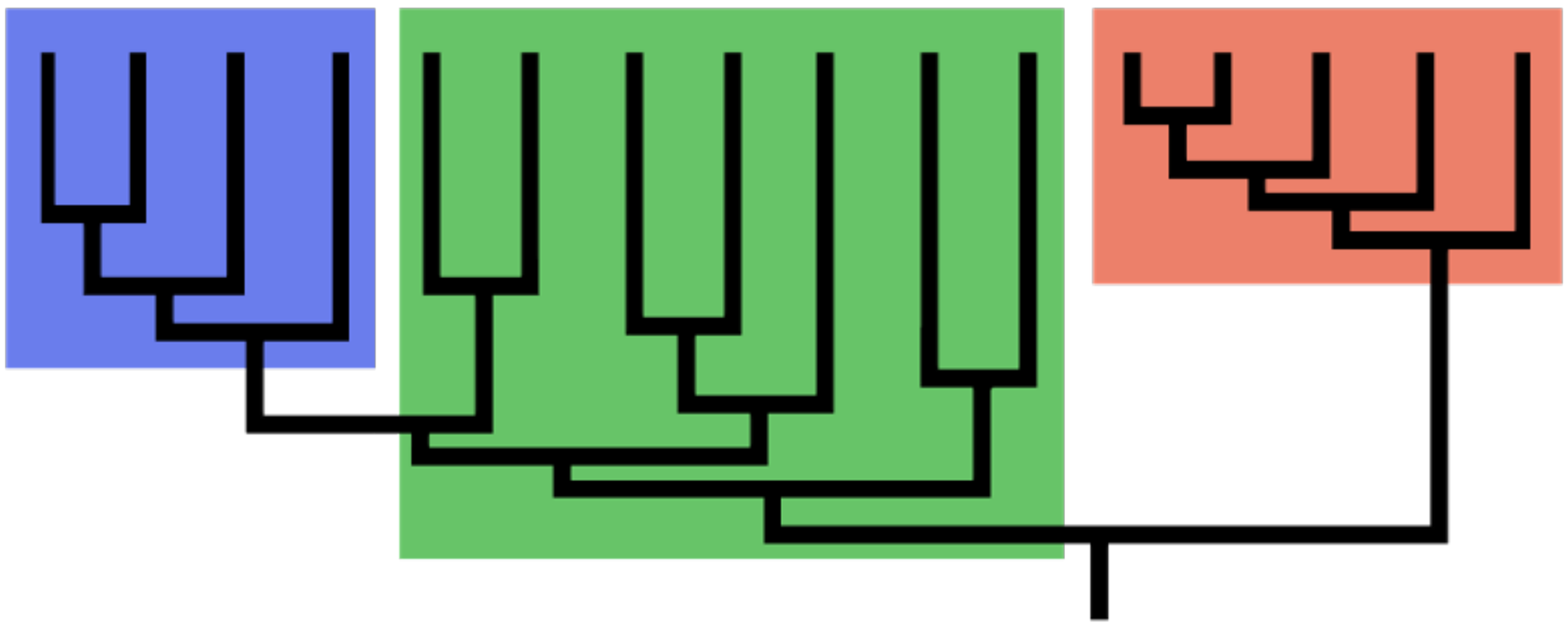
High taxonomic classes (e.g. phylum) have been validated as monophyletic based on longer sequences

Build the tree — Part 1: Constrains

Rationale

High taxonomic classes (e.g. phylum) have been validated as monophyletic based on longer sequences

A **monophyletic** group, or clade, is a group of organisms that consists of all the descendants of a common ancestor



Wikipedia

Build the tree — Part 1: Constrains

Rationale

High taxonomic classes (e.g. phylum) have been validated as monophyletic based on longer sequences

Idea

Constrain the construction of the tree so that OTUs belonging to the same taxonomic group are grouped in a monophyletic group in the tree

Build the tree — Part 1: Constrains

FastTree -help

Constrains Input for FastTree

```
fmazels-iMac:~ fmazel$ Desktop/Programmes_Unix/FastTree -help
FastTree 2.1.10 SSE3:
  FastTree protein_alignment > tree
  FastTree < protein_alignment > tree
  FastTree -out tree protein_alignment
  FastTree -nt nucleotide_alignment > tree
  FastTree -nt -gtr < nucleotide_alignment > tree
  FastTree < nucleotide_alignment > tree
FastTree accepts alignments in fasta or phylip interleaved formats

Common options (must be before the alignment file):
  -quiet to suppress reporting information
  -nopr to suppress progress indicator
  -log logfile -- save intermediate trees, settings, and model details
  -fastest -- speed up the neighbor joining phase & reduce memory usage
       (recommended for >50,000 sequences)
  -n <number> to analyze multiple alignments (phylip format only)
       (use for global bootstrap, with seqboot and CompareToBootstrap.pl)
  -nosupport to not compute support values
  -intree newick_file to set the starting tree(s)
  -intree1 newick_file to use this starting tree for all the alignments
       (for faster global bootstrap on huge alignments)
  -pseudo to use pseudocounts (recommended for highly gapped sequences)
  -gtr -- generalized time-reversible model (nucleotide alignments only)
  -lg -- Le-Gascuel 2008 model (amino acid alignments only)
  -wag -- Whelan-And-Goldman 2001 model (amino acid alignments only)
  -quote -- allow spaces and other restricted characters (but not ' ) in
          sequence names and quote names in the output tree (fasta input only;
          FastTree will not be able to read these trees back in)
  -noml to turn off maximum-likelihood
  -nome to turn off minimum-evolution NNIs and SPRs
       (recommended if running additional ML NNIs with -intree)
  -nome -mllen with -intree to optimize branch lengths for a fixed topology
  -cat # to specify the number of rate categories of sites (default 20)
      or -nocat to use constant rates
  -gamma -- after optimizing the tree under the CAT approximation,
      rescale the lengths to optimize the Gamma20 likelihood
  -constraints constraintAlignment to constrain the topology search
       constraintAlignment should have 1s or 0s to indicate splits
  -expert -- see more options
For more information, see http://www.microbesonline.org/fasttree/
```

http://www.microbesonline.org/fasttree/constrained.html

Build the tree — Part 1: Constrains

Constrains Input for FastTree

A table OTU*constrained clades in a fasta file

```
> head(Constrains)
        Bacteria Proteobacteria Bacteroidetes Thaumarchaeota Actinobacteria Marinimicrobia_(SAR406_clade)
Otu0001        1              1             0              0              0                             0
Otu0002        1              1             0              0              0                             0
Otu0003        1              1             0              0              0                             0
Otu0004        1              1             0              0              0                             0
Otu0005        1              1             0              0              0                             0
Otu0006        1              1             0              0              0                             0
```

To construct with the taxonomy file:

```
> head(taxonomy)
          Domain         Phylum                  Class                      Order
Otu0001 Bacteria Proteobacteria   Gammaproteobacteria            Oceanospirillales
Otu0002 Bacteria Proteobacteria Epsilonproteobacteria            Campylobacterales
Otu0003 Bacteria Proteobacteria   Gammaproteobacteria            Oceanospirillales
Otu0004 Bacteria Proteobacteria   Gammaproteobacteria                 Chromatiales
Otu0005 Bacteria Proteobacteria   Deltaproteobacteria SAR324_clade(Marine_group_B)
Otu0006 Bacteria Proteobacteria   Gammaproteobacteria            Oceanospirillales
```

Build the tree — Part 1: Constrains

Constrains Input for FastTree

Code lines 122 — 185

SetUp.Rmd | Workflow.Rmd | README.md | Additional Data Table14 R-file_rev... | Convert_Tmd_in_R.R | BDTT_

Knit | Insert | Run

```
115
116 ### 3.2. Tree Building <a name="Tree-Building"></a>
117
118 Typical read length in microbiome studies are relatively short and have thus contains limited information to
    reconstruct phylogeentic trees, especially to reconstruct deep branches. To avoid biased phylogenies, we thus
    constrains deep branches to follow taxonomic classification as it is admitted that large taxonomic clades are
    monophyletic (informations based on longer sequences). The choice of the constrains is not easy. Here we will
    constain tree reconstruction by Domain (Bacteria/Archea) and phylums.
119
120 We will use FastTree to reconstruct the phylogeneic hypotheses. While fastTree is not the best software to
    reconstruct phylogenies (because it makes a lot of approximations), it has the huge avantage to be fast, which is
    often critical in microbial species, as there are a very high number of "species".
121
122 #### Topological constrains
123
124 Load the taxonomic file
125
126 ```{r, message=FALSE}
127 taxonomy.raw = read.table("data/Saanich_cruise72_mothur_OTU_taxonomy.taxonomy", sep="\t", header=TRUE, row.names=1)
128 taxonomy= taxonomy.raw %>%
129   select(-Size) %>%
130   separate(Taxonomy, c("Domain", "Phylum", "Class", "Order", "Family", "Genus", "Species"), sep=";")
```

Build the tree — Part 2 : Run FasTree

On the Terminal

Function model Constrain fasta file

FastTree -gtr -cat 20 -constraints [Path] -nt [Path] > [Path]

Alignment

Output

Build the tree — Part 2 : Run FasTree

On the Terminal

Code lines 185 — 211

Build the tree — Part 3 : Plot the tree

Code lines 211 — 260

Load the trees in R

```
> tree=read.tree()
```

Explore the tree file

Colors of the tips

```
> colors=
```

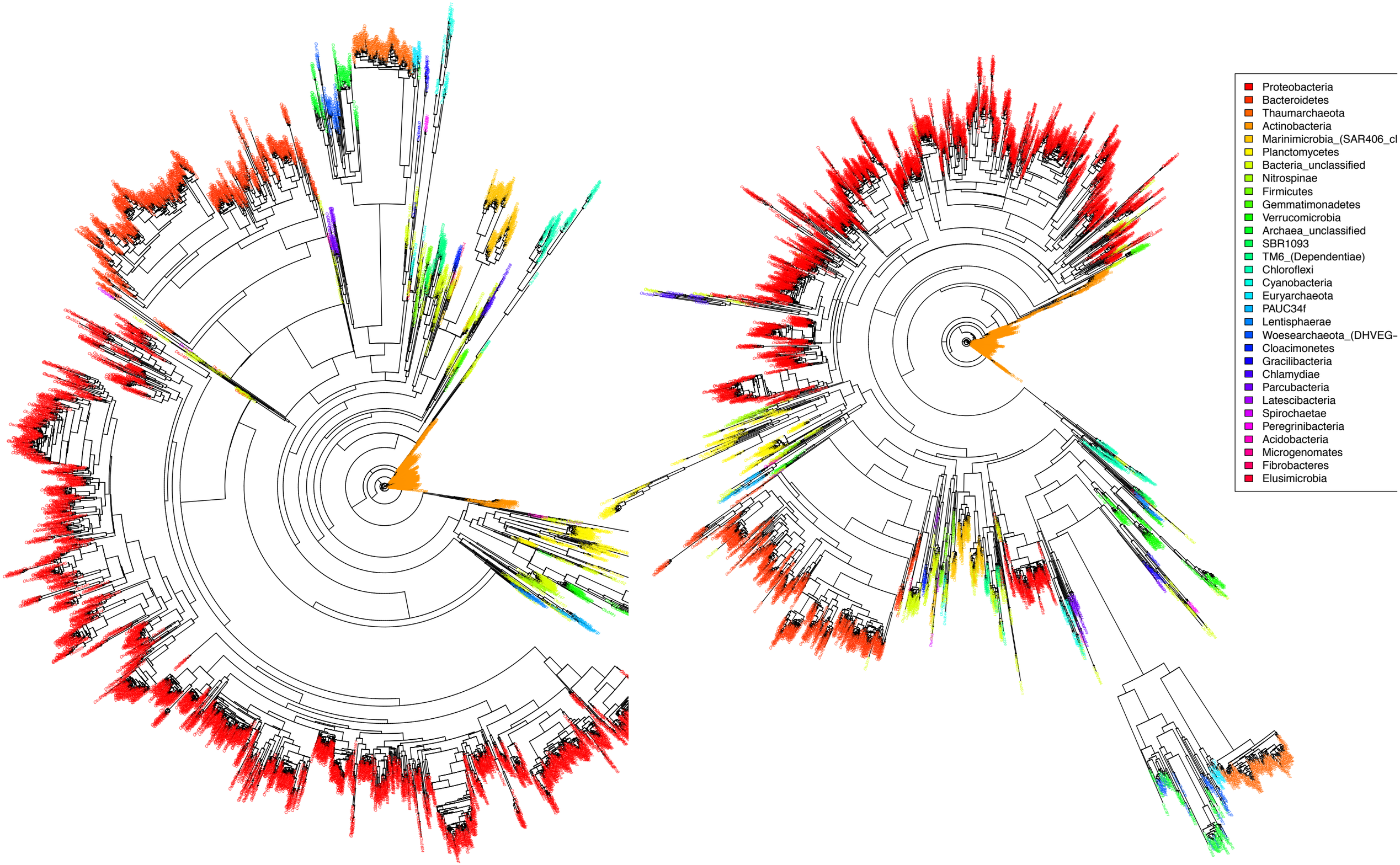
Plot the trees in R

```
> pdf()
> plot(tree, col=colors)
> dev.off()
```

Build the tree — Part 3 : Plot the tree

With constrains

Without constrains

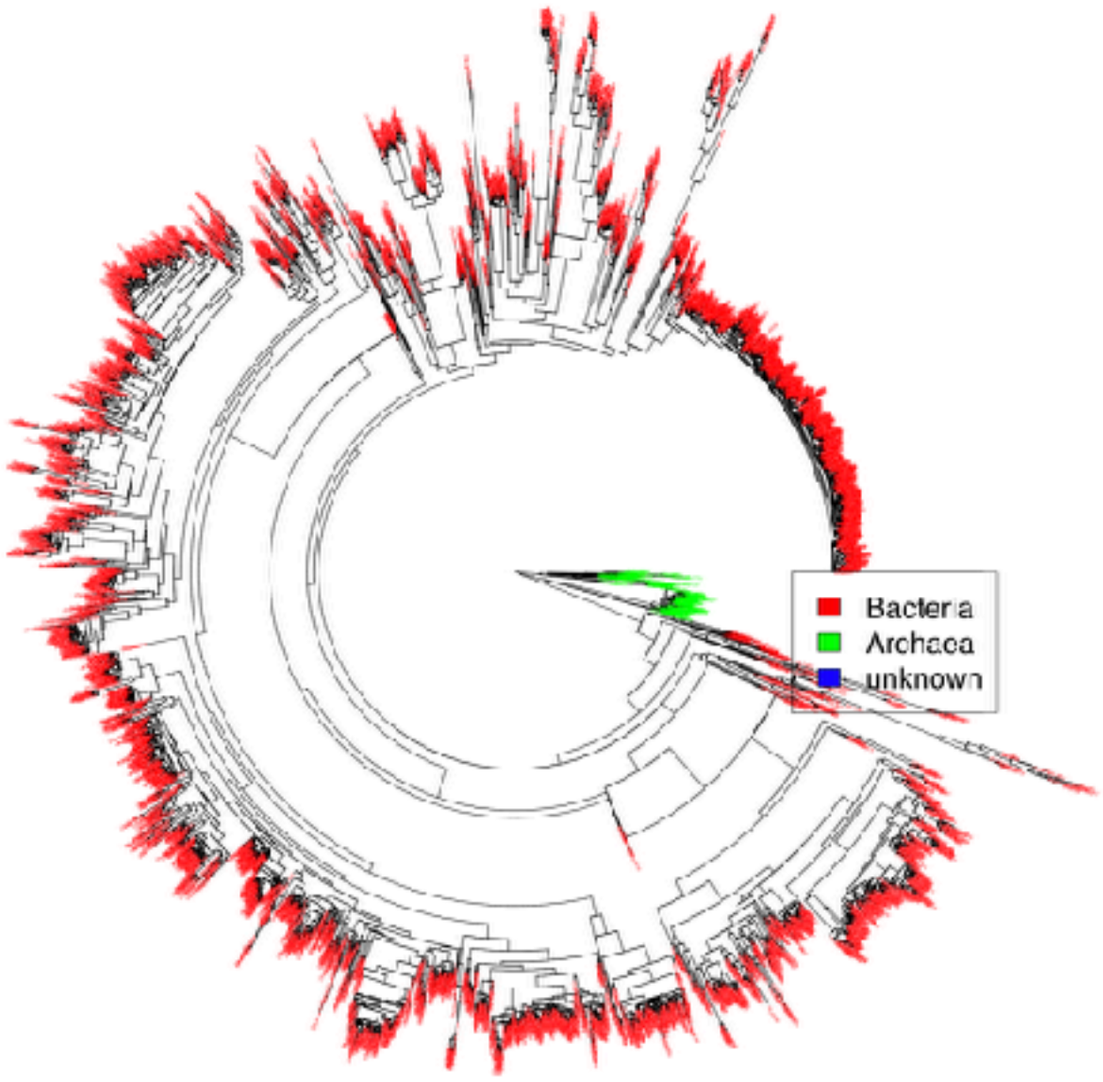


Build the tree — Part 3 : Root the tree

Code lines 260 — 295

Root the tree in R

```
> tree=root()
```



Overall structure of the workshop

1 — Building a phylogenetic tree

2 — Classical analysis of microbiome

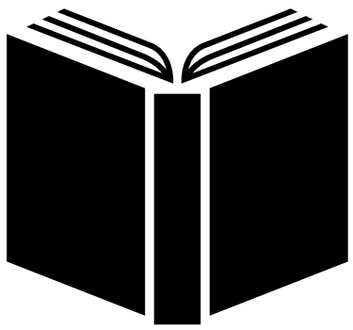
3 — Varying the phylogenetic resolution

4 — Exploring the branches of the phylogenetic tree

2 — Classical analysis of microbiome compositions

“Classical analysis” of the microbiota composition

Theory



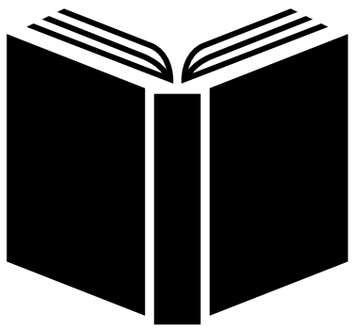
Beta-Diversity metric

Visualization

Statistical test

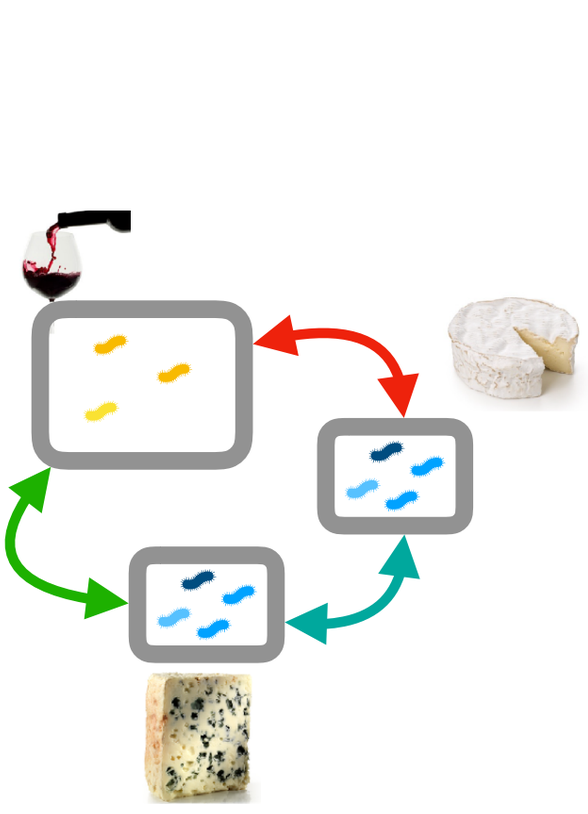
“Classical analysis” of the microbiota composition

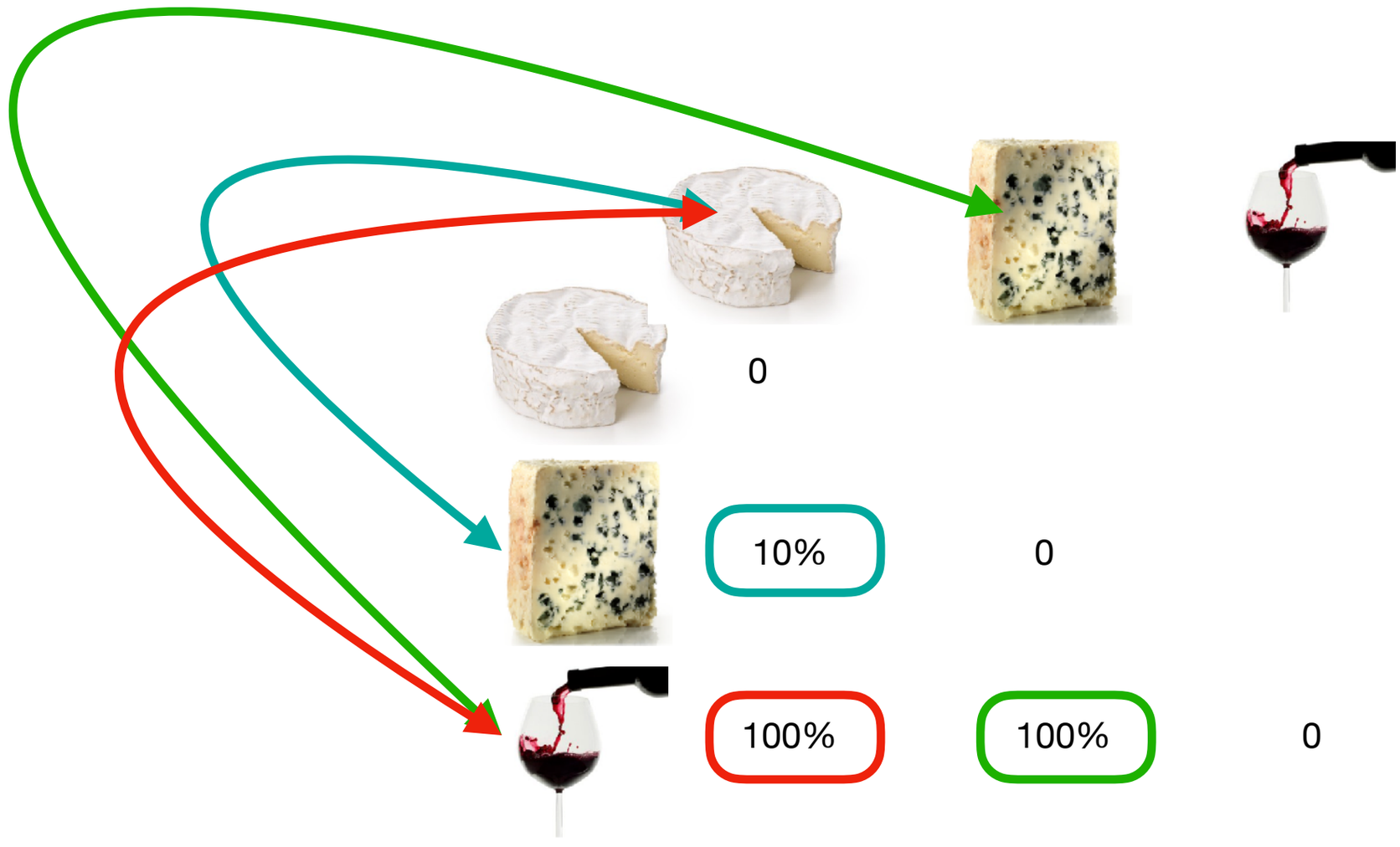
Theory



Beta-Diversity metric

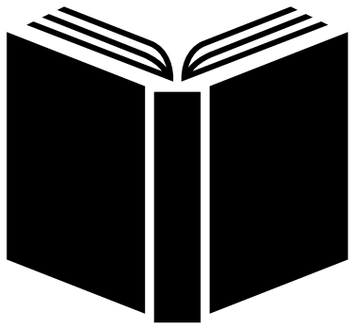
Distance matrix





“Classical analysis” of the microbiota composition

Theory



Beta-Diversity metric

Non — Phylogenetic

Presence/Absence

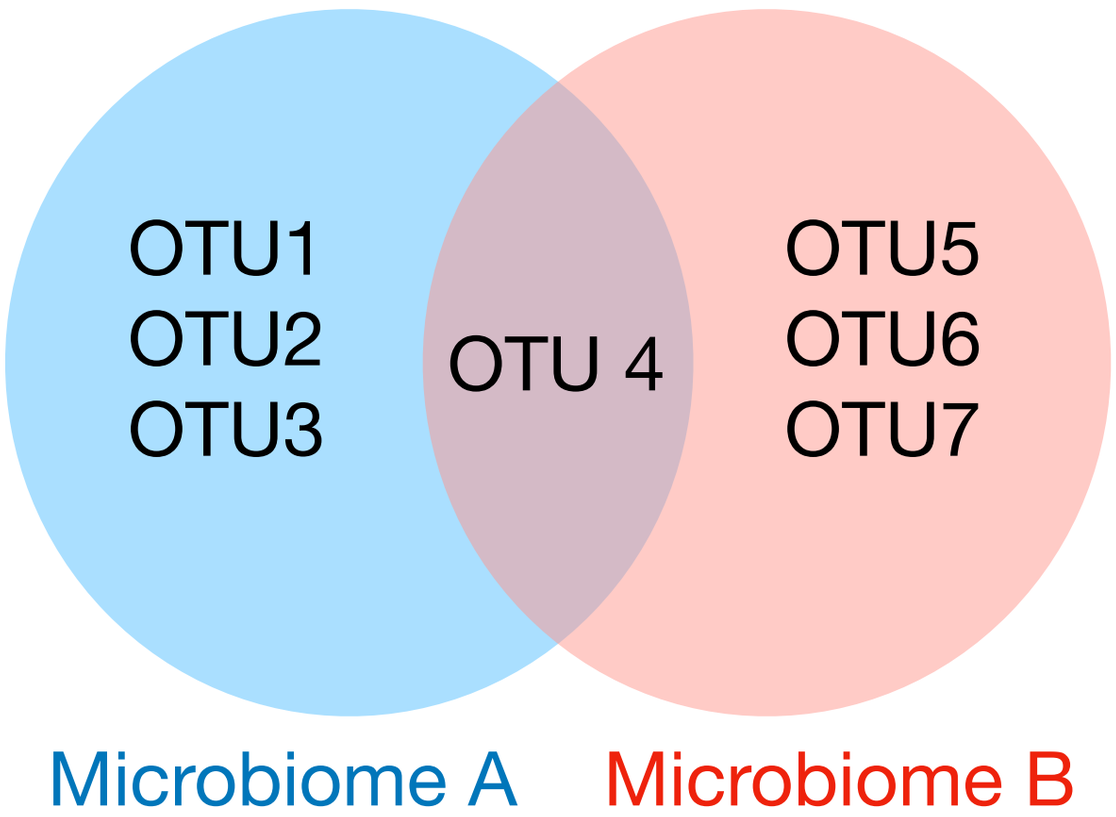
“Classical analysis” of the microbiota composition

Theory

Beta-Diversity metric

How to quantify similarity?
Example of Jaccard index

$$J(A, B) = 1 - \frac{A \cap B}{A \cup B}$$



$A \cap B$ = 1 (OTU4)

$A \cup B$ = 7 (OTU1-7)

Jaccard, Paul. *Nouvelles recherches sur la distribution florale*. 1908.

“Classical analysis” of the microbiota composition

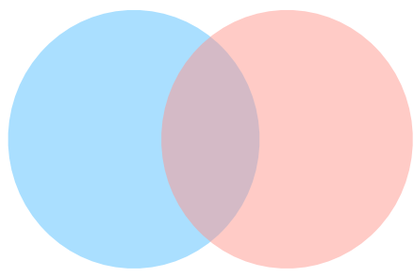
Theory

Beta-Diversity metric

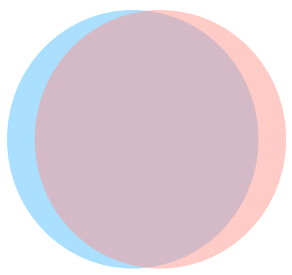
How to quantify similarity?
Example of Jaccard index

$$J(A, B) = 1 - \frac{A \cap B}{A \cup B}$$

J=1



0<J<1



J=0

Jaccard, Paul. *Nouvelles recherches sur la distribution florale*. 1908.

“Classical analysis” of the microbiota composition

Theory

Beta-Diversity metric

Non — Phylogenetic

Presence/Absence

Jaccard
Sorensen

“Classical analysis” of the microbiota composition

Theory

Beta-Diversity metric

Non — Phylogenetic

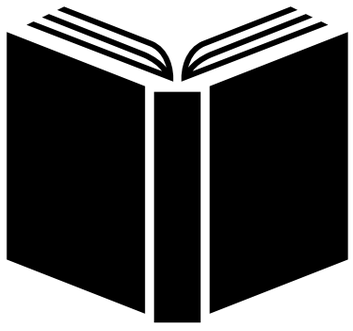
Presence/Absence

Jaccard
Sorensen

Abundances

“Classical analysis” of the microbiota composition

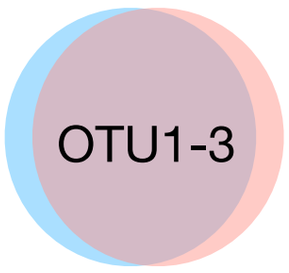
Theory



Beta-Diversity metric

Example of Bray Curtis

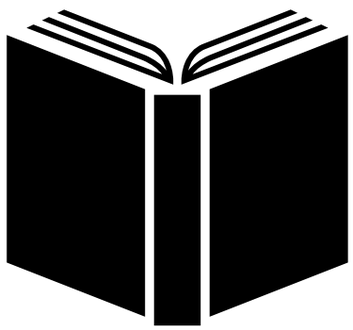
Same idea but considering reads, not only OTUs



J=0

“Classical analysis” of the microbiota composition

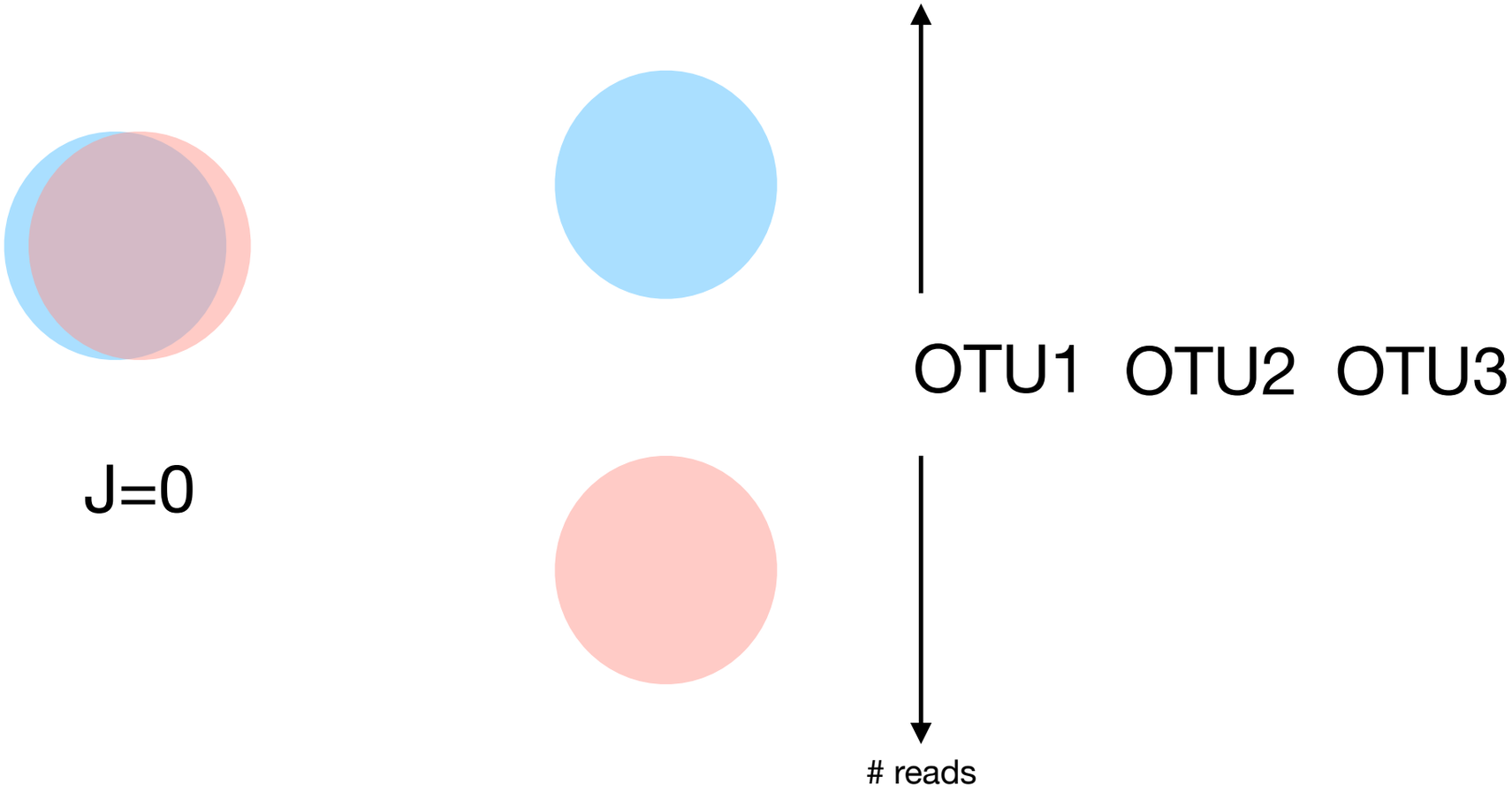
Theory



Beta-Diversity metric

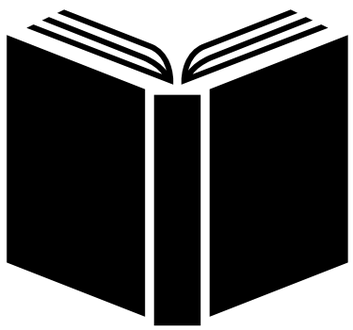
Example of Bray Curtis

Same idea but considering reads, not only OTUs



“Classical analysis” of the microbiota composition

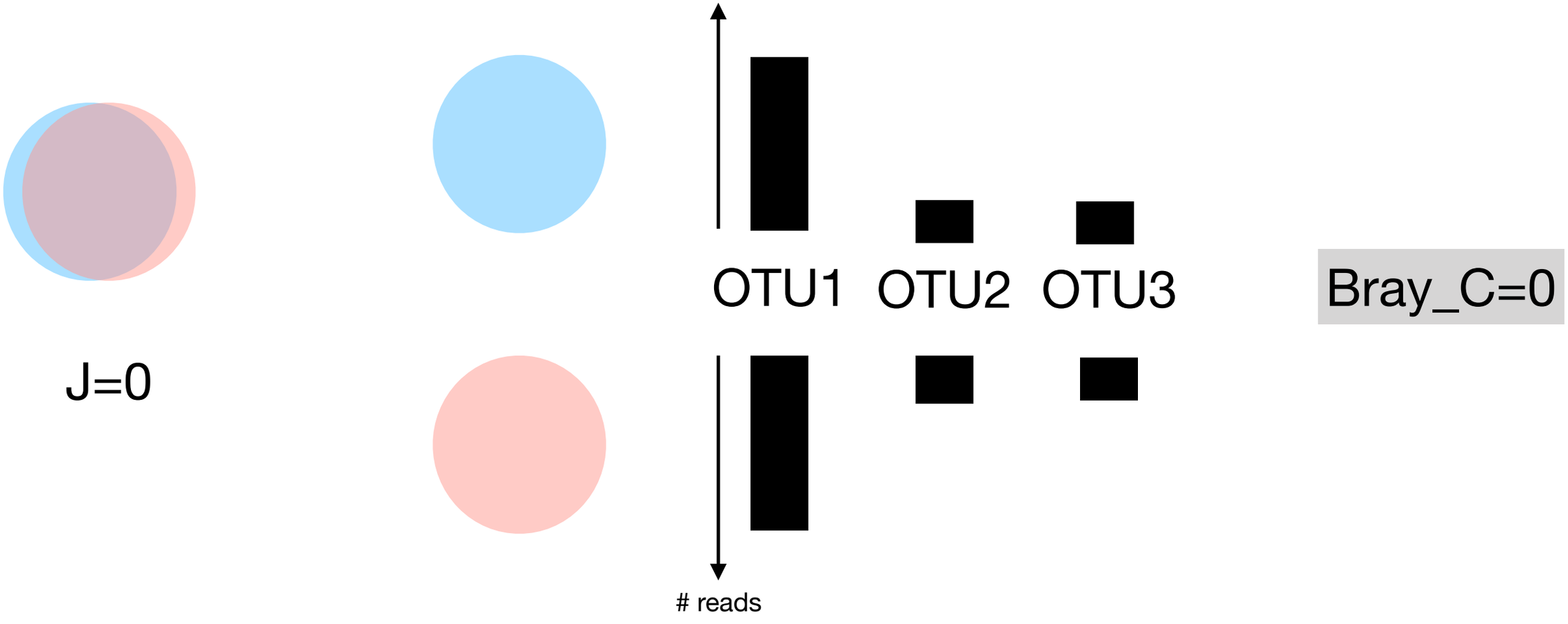
Theory



Beta-Diversity metric

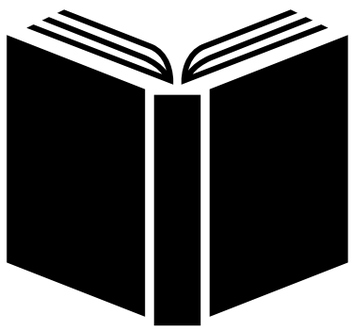
Example of Bray Curtis

Same idea but considering reads, not only OTUs



“Classical analysis” of the microbiota composition

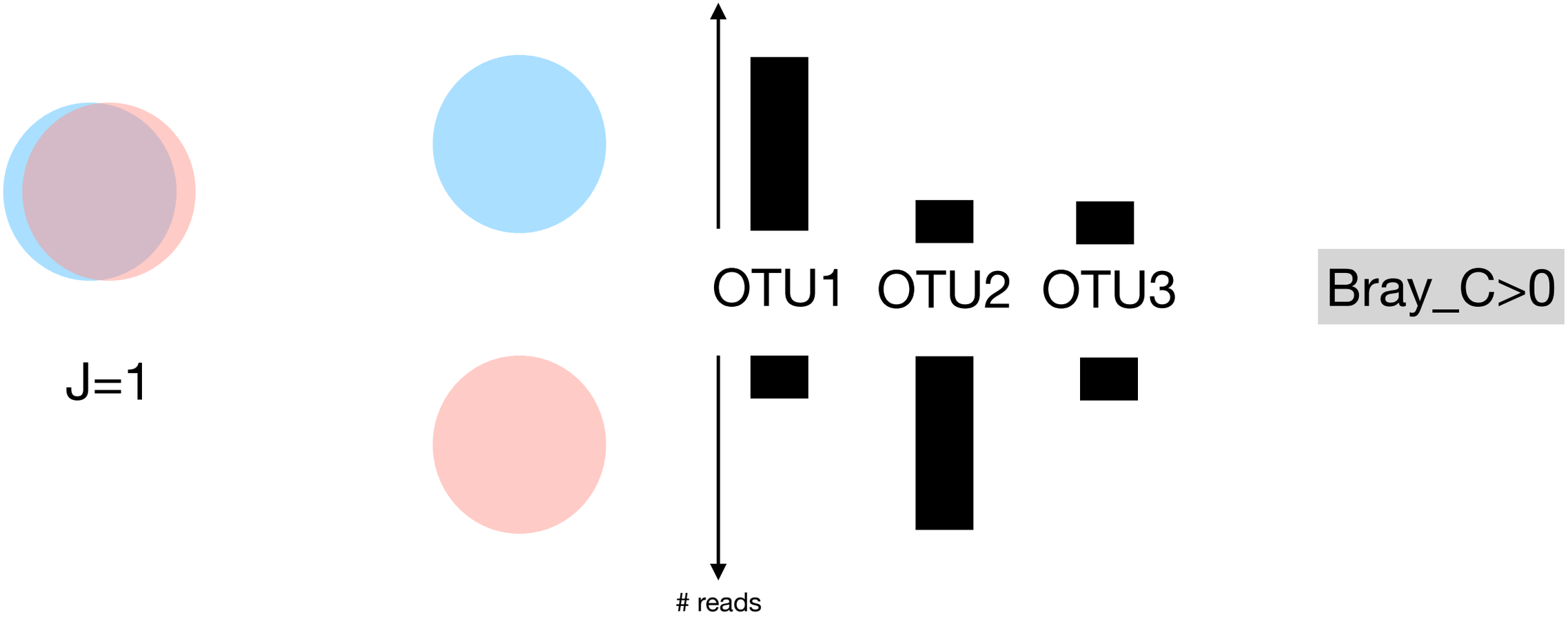
Theory



Beta-Diversity metric

Example of Bray Curtis

Same idea but considering reads, not only OTUs



“Classical analysis” of the microbiota composition

Theory

Beta-Diversity metric

Non — Phylogenetic

Presence/Absence	**Jaccard** **Sorensen**
Abundances	**BrayCurtis**

“Classical analysis” of the microbiota composition

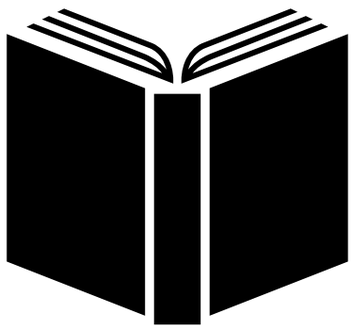
Theory

Beta-Diversity metric

	Non — Phylogenetic	Phylogenetic
Presence/Absence	**Jaccard** **Sorensen**	
Abundances	**BrayCurtis**	

“Classical analysis” of the microbiota composition

Theory



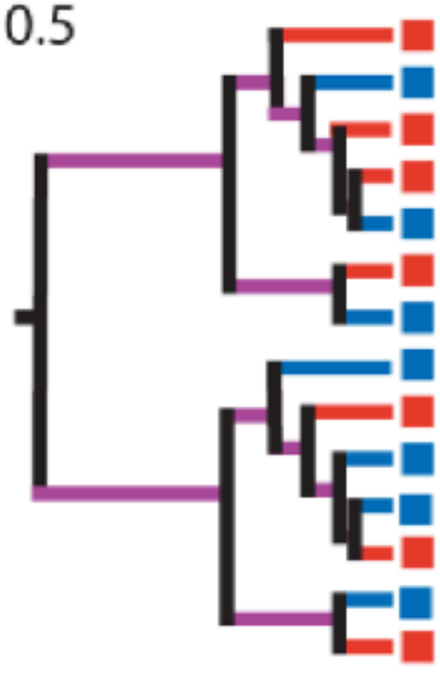
Beta-Diversity metric

With Jaccard
(Non phylogenetic)

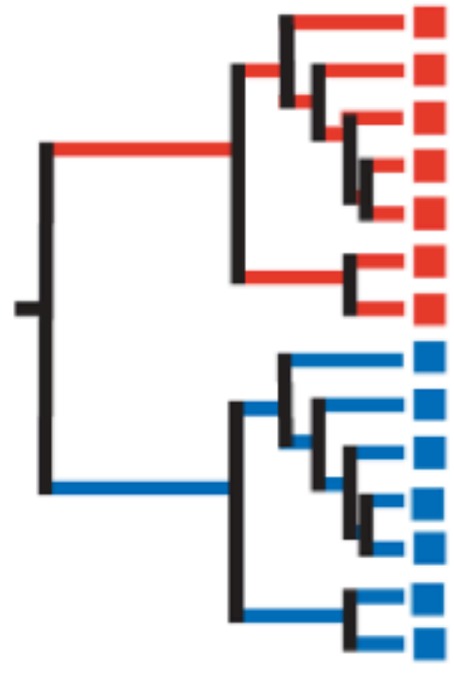
J=1

Example of Unifrac
(Phylogenetic-based metric)

Related communities:
D ~ 0.5



Unrelated communities:
D = 1



“Classical analysis” of the microbiota composition

Theory

Beta-Diversity metric

	Non — Phylogenetic	Phylogenetic
Presence/Absence	**Jaccard** **Sorensen**	Unifrac Phylosor
Abundances	**BrayCurtis**	**Weighted Unifrac**

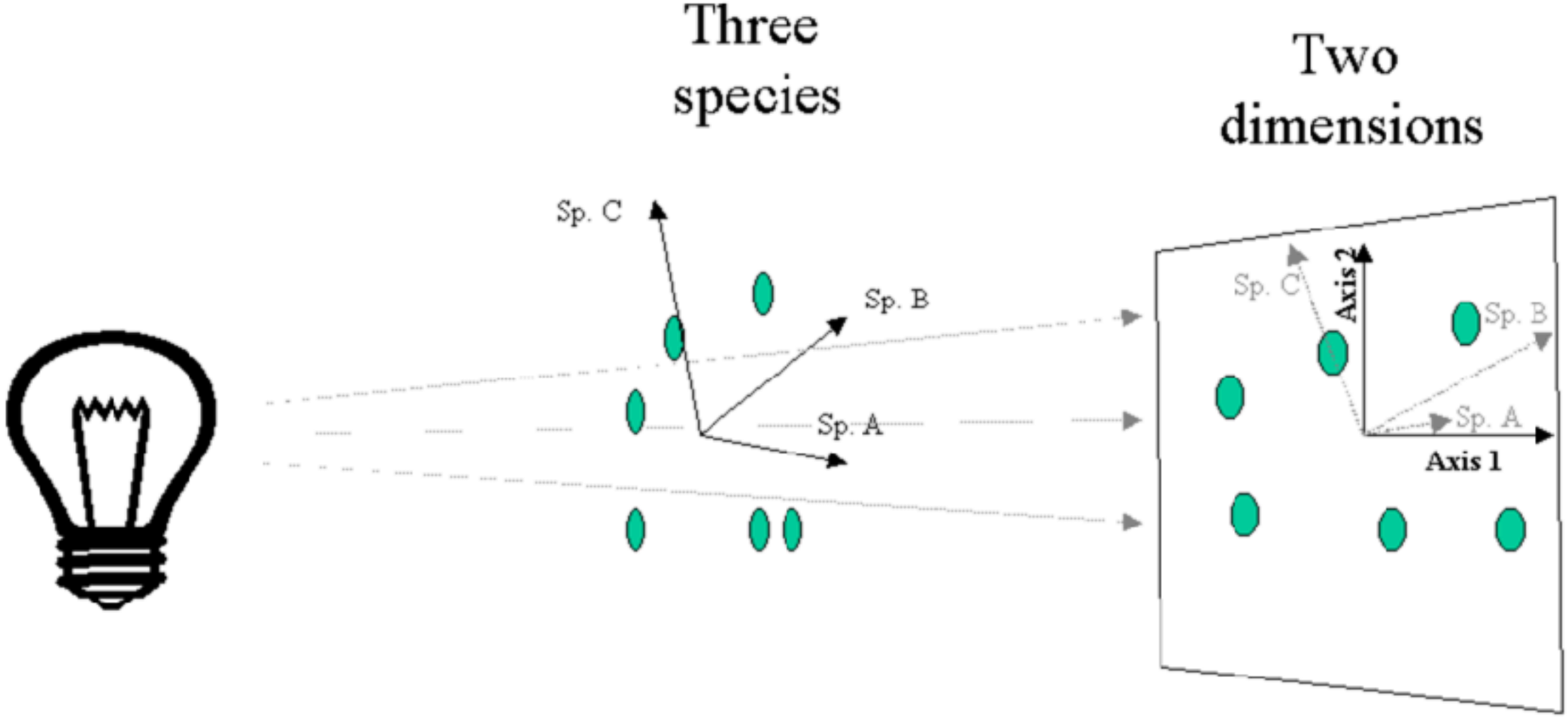
“Classical analysis” of the microbiota composition

Theory

Visualization

Try to best represent there distances on a limited number of axes

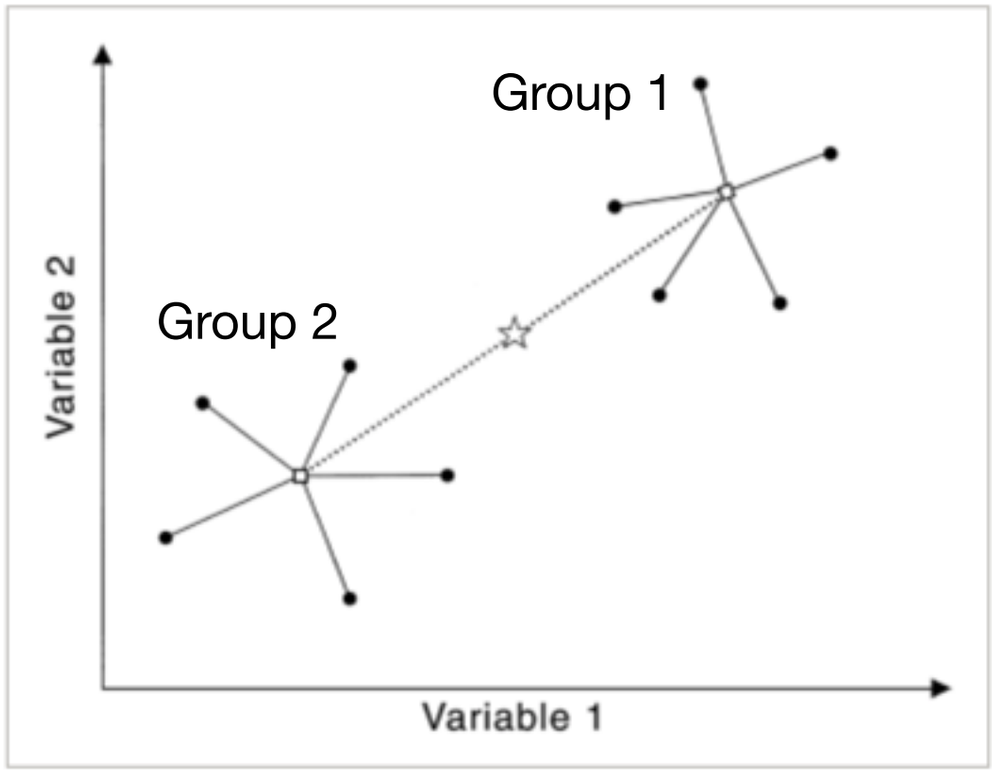
for n (number of OTUs) dimensions to 2 (or 3) dimensions



“Classical analysis” of the microbiota composition

Statistical test

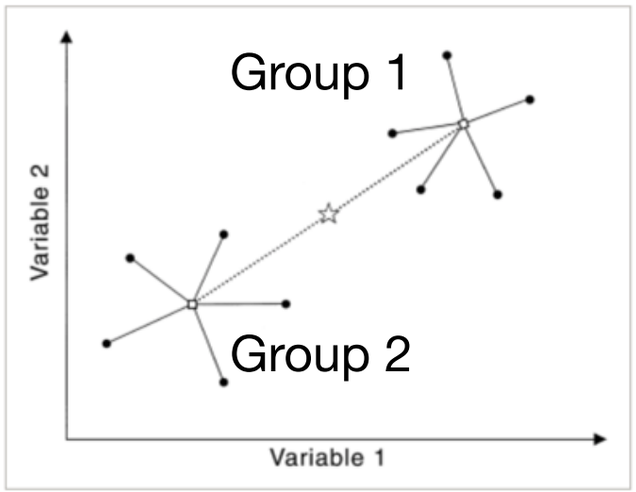
Example of the PERMANOVA test



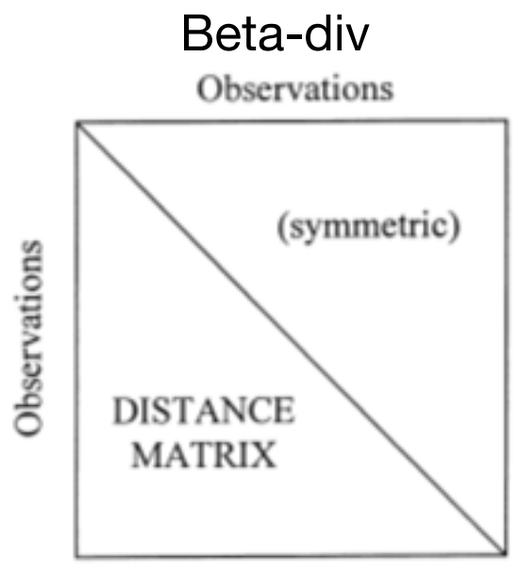
“Classical analysis” of the microbiota composition

Statistical test

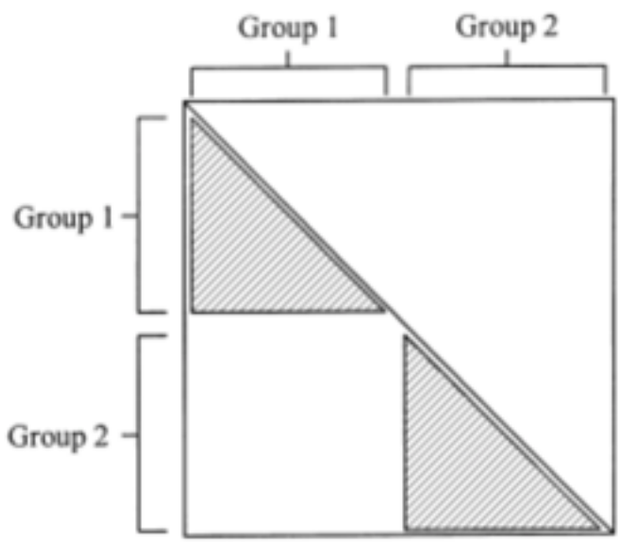
Example of the PERMANOVA test



Beta-div

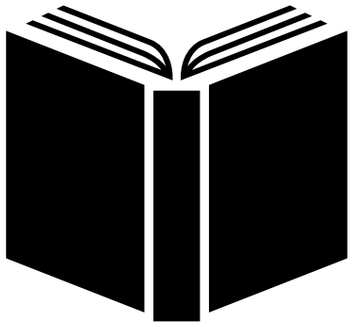


Beta-div



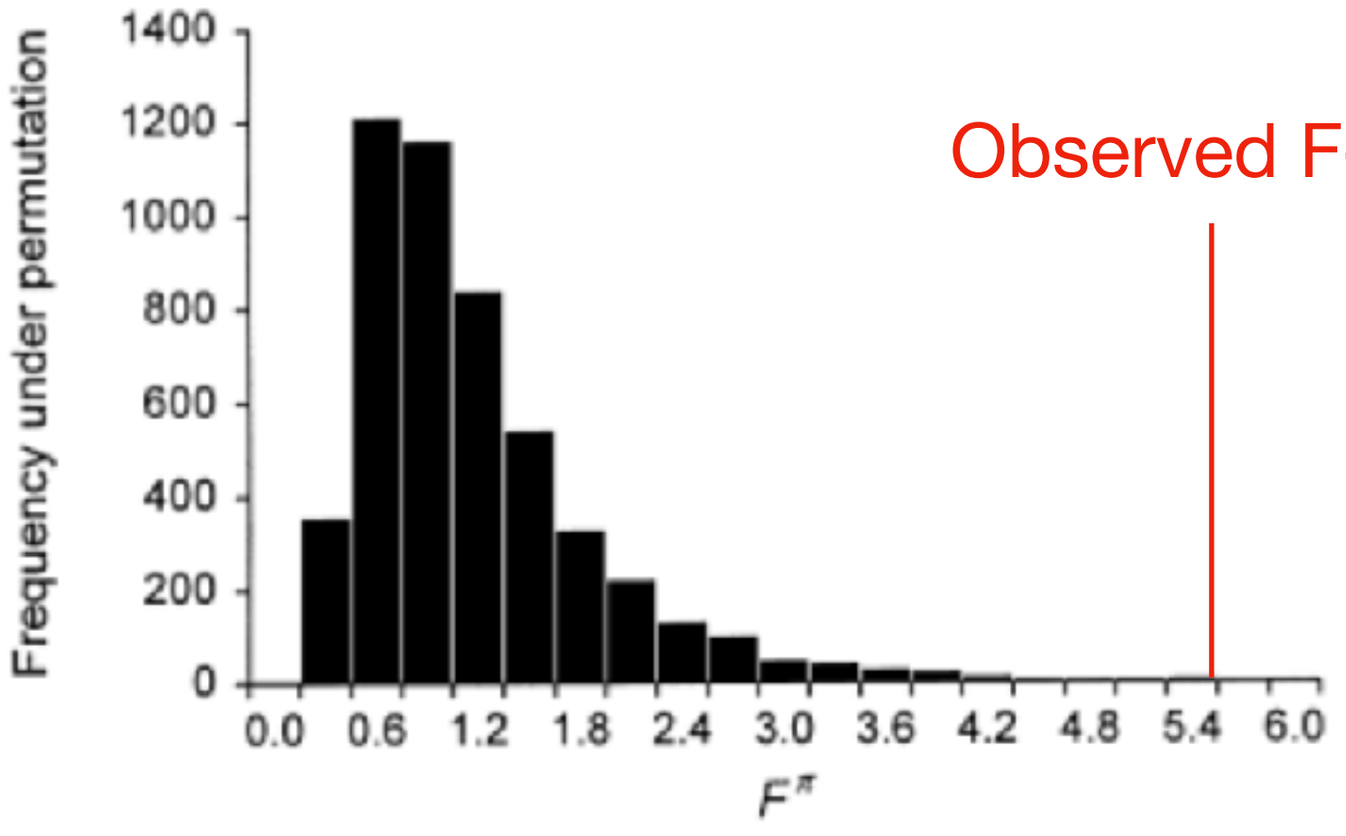
F ratio= [between] / [within]

“Classical analysis” of the microbiota composition



Statistical test

Example of the PERMANOVA test



p-value

“Classical analysis” of the microbiota composition

Statistical test

Example of the PERMANOVA test

```
Call:
adonis(formula = UniFracBeta ~ Depth_m, data = data.frame(sample_data(mothur)))

Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

          Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
Depth_m    1   0.34826 0.34826  2.5771 0.34012  0.001 ***
Residuals  5   0.67569 0.13514         0.65988
Total      6   1.02395                 1.00000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. Analysis the phylogenetic composition of microbiomes

Practice

Code lines 295 — 391

Import data

```
> read.table()
> read.tree()
```

Create a unified object

```
> phyloseq()
```

Compute “traditional” Beta-diversity metrics

```
> vegdist()
> UniFrac()
```

Plot their relationships

```
> plot()
```

Visualize Beta-diversities

```
> ordinate()
```

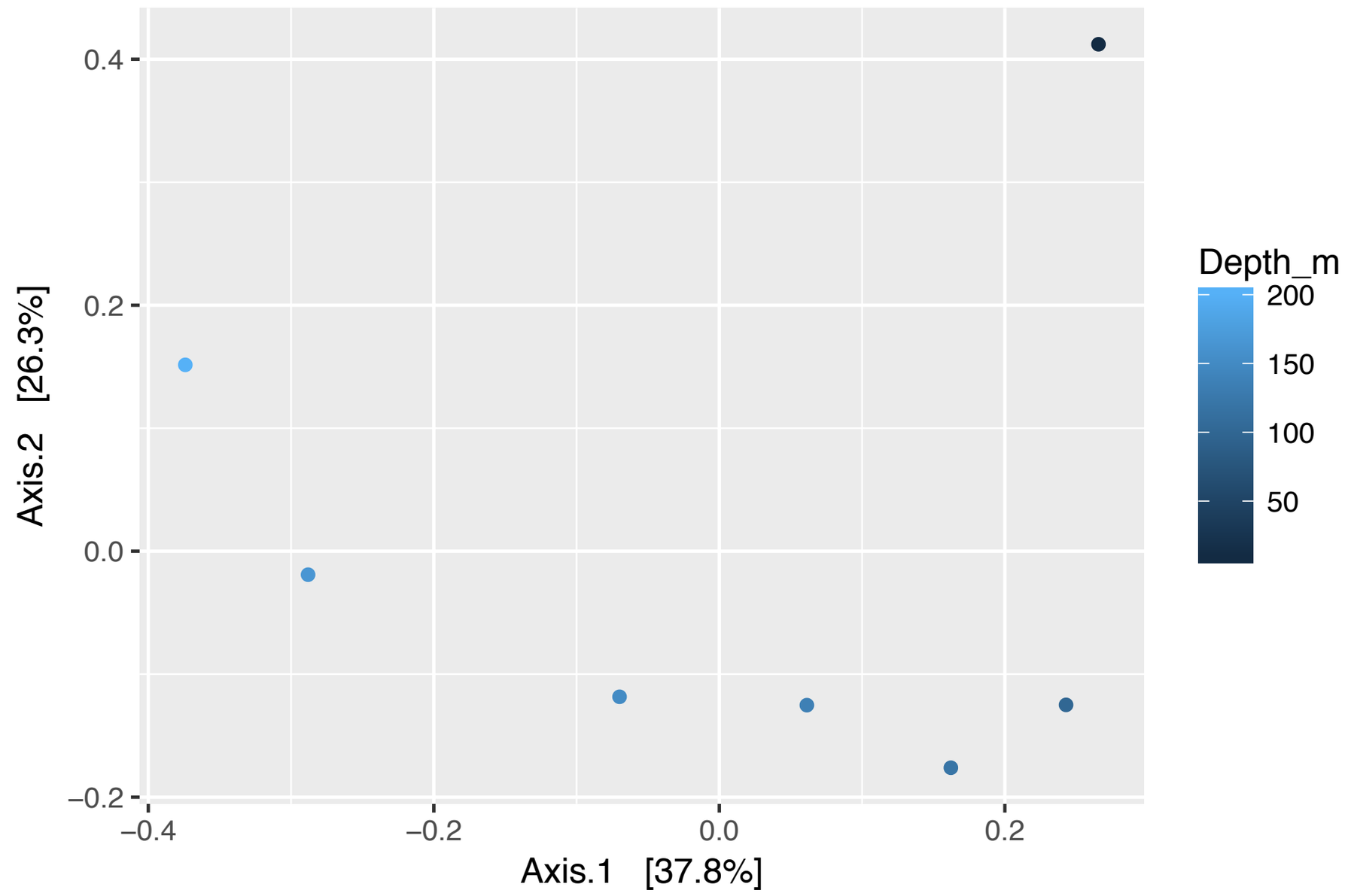
Statistical test

```
> adonis()
```

“Classical analysis” of the microbiota composition

Practice

Visualization



“Classical analysis” of the microbiota composition

Practice

Statistical test

R function

```
> adonis()
```

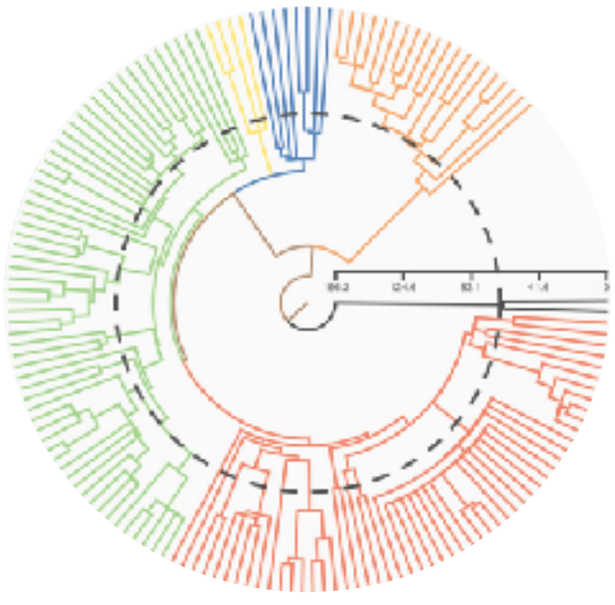
Test several beta-diversity metrics: difference between phylogenetic / non phylogenetic?

Test one or more explanatory variables

3 — Varying the phylogenetic resolution

Example 1: Mammalian guts

Gut Microbiota Composition ~ Host Phylogeny or Host Diet ?



Microb Ecol
DOI 10.1007/s00248-017-1041-8

CrossMark

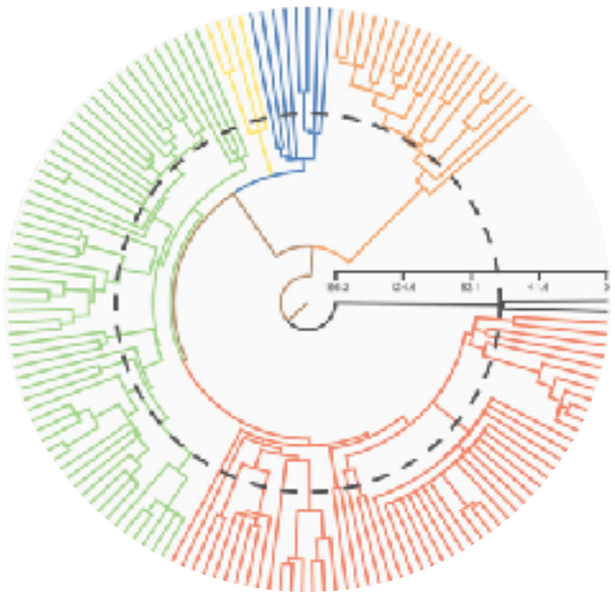
HOST MICROBE INTERACTIONS

Diet Versus Phylogeny: a Comparison of Gut Microbiota in Captive Colobine Monkey Species

3 — Varying the phylogenetic resolution

Example 1: Mammalian guts

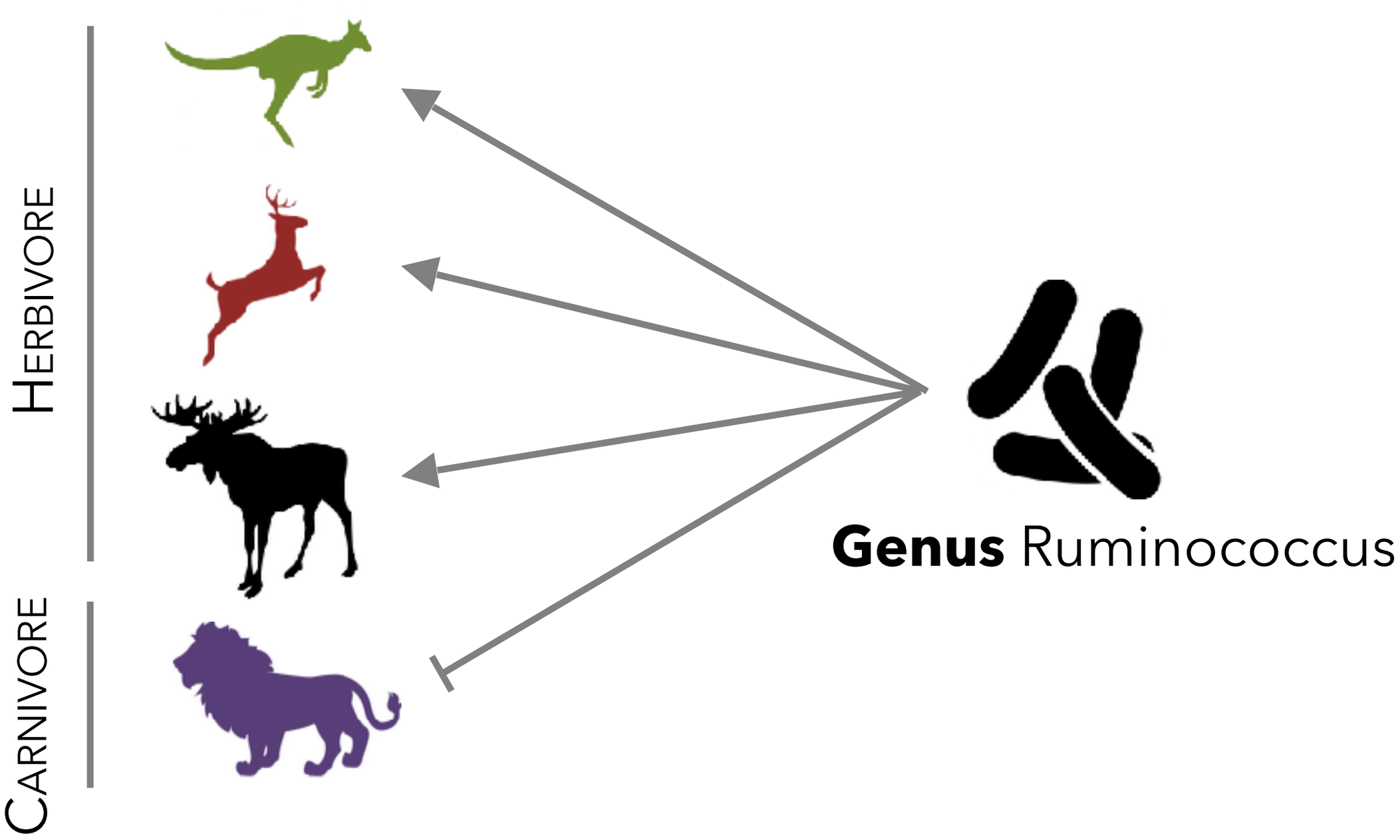
Gut Microbiota Composition ~ Host Phylogeny or Host Diet ?



***Idea**: Different factors might play at different phylogenetic resolutions*

Example 1: Mammalian guts

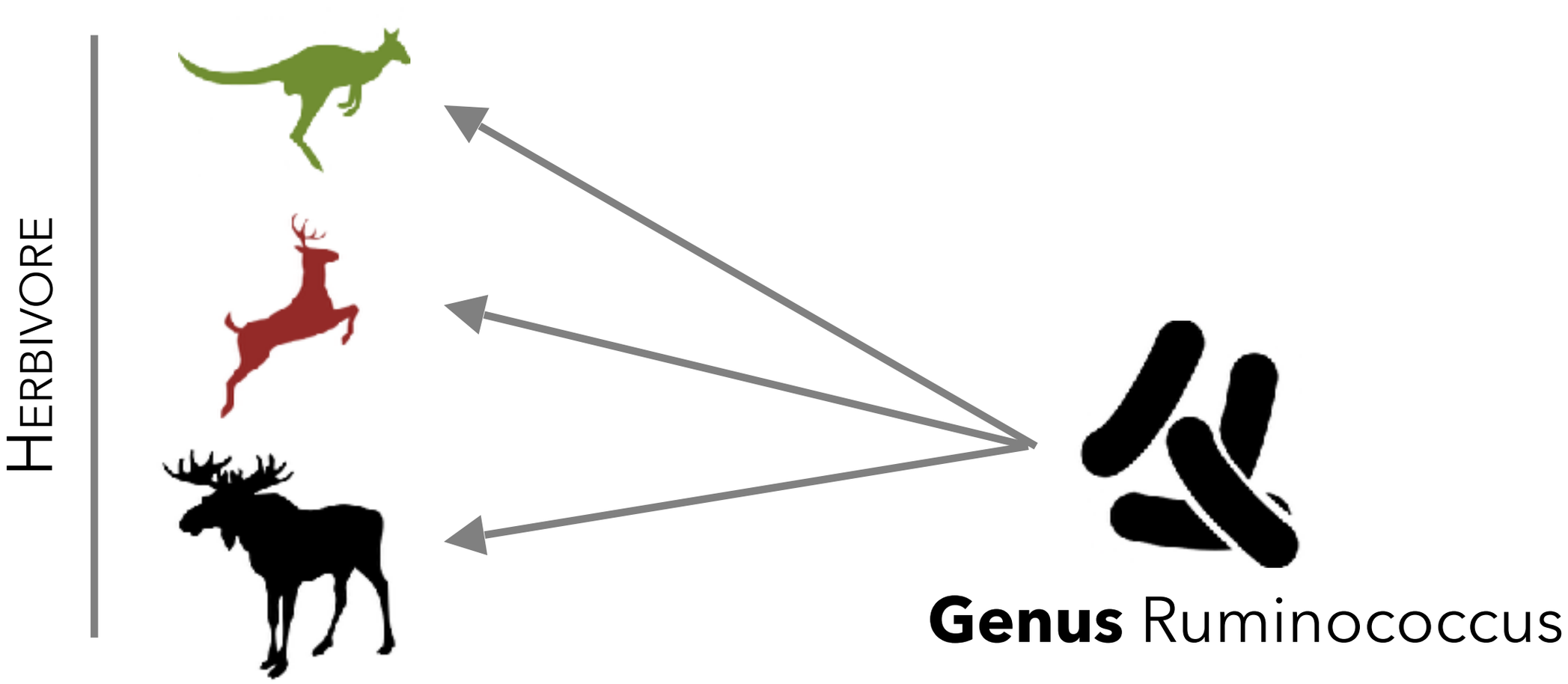
Coarse Bacterial Phylogenetic Scale



3 — Varying the phylogenetic resolution

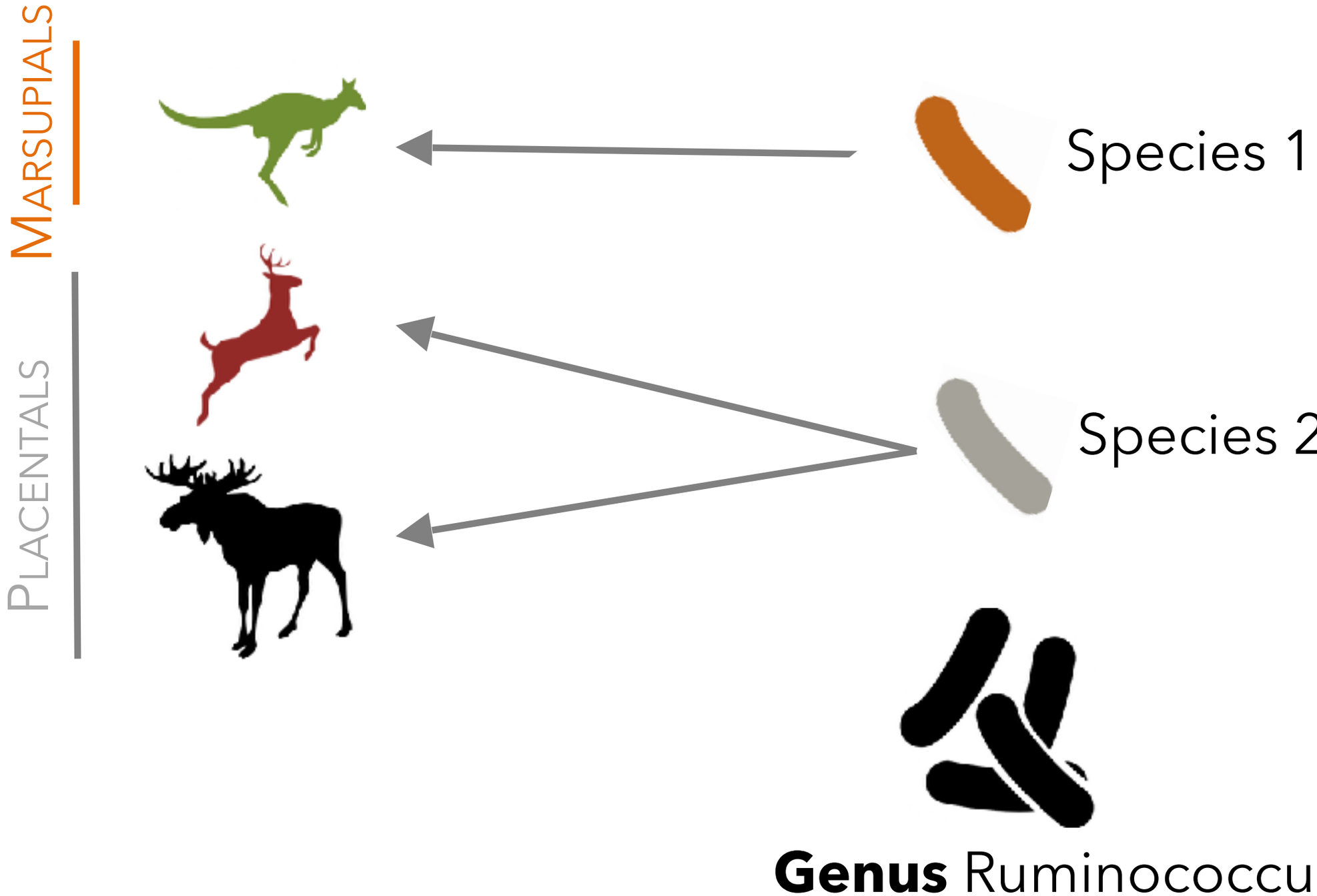
Example 1: Mammalian guts

Coarse Bacterial Phylogenetic Scale



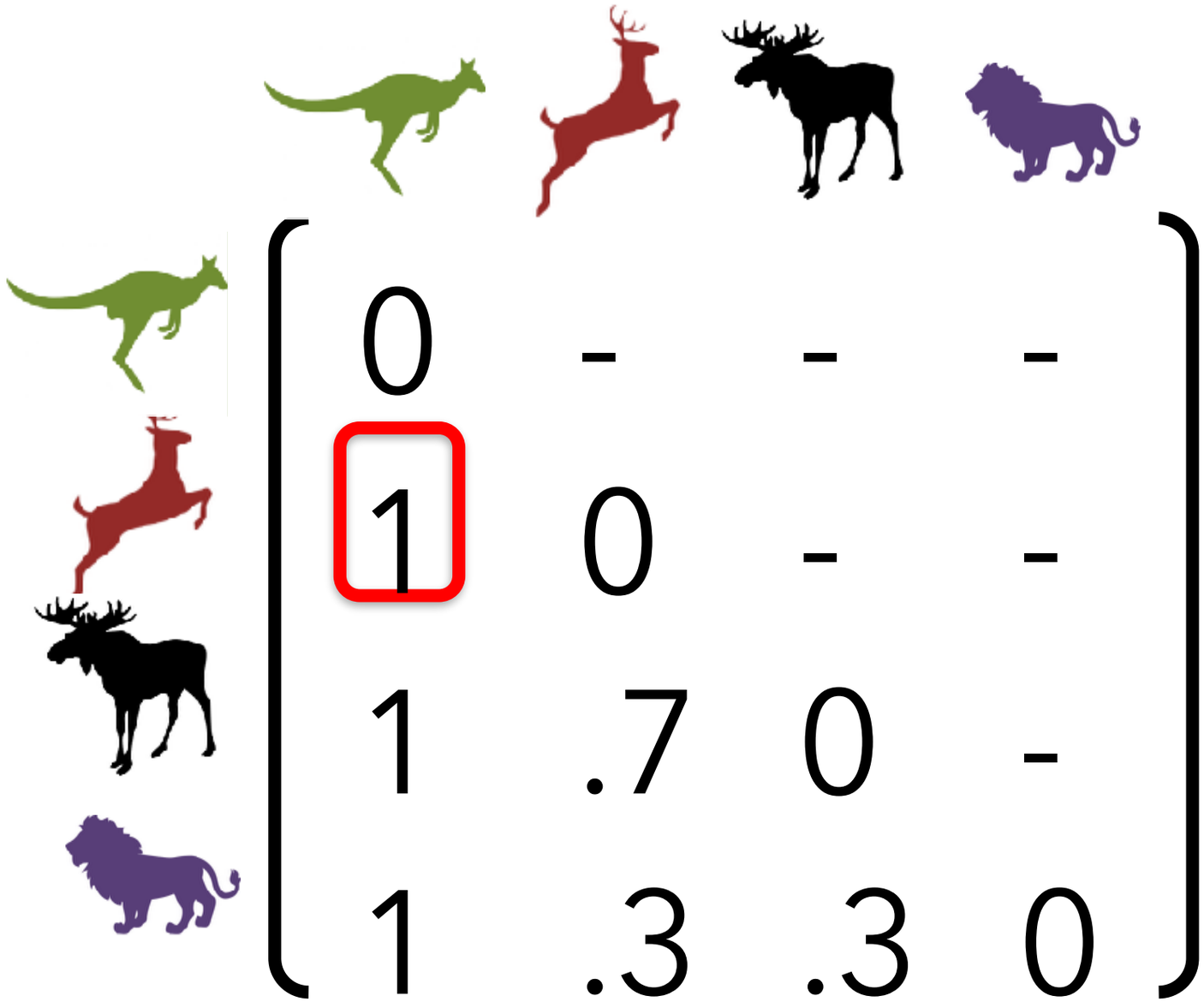
Example 1: Mammalian guts

Fine Bacterial Phylogenetic Scale



3 — Varying the phylogenetic resolution

Theory

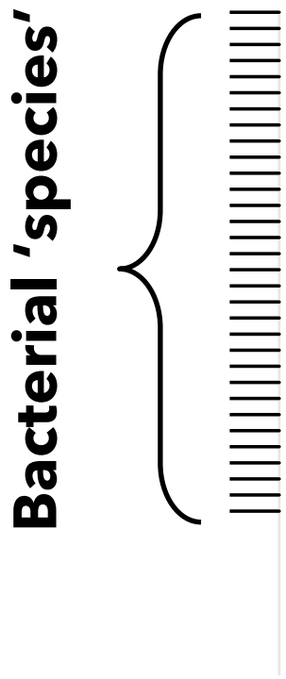


% of shared **‘species’** between hosts

3 — Varying the phylogenetic resolution

Theory

Method: “Slicing the tree of life”

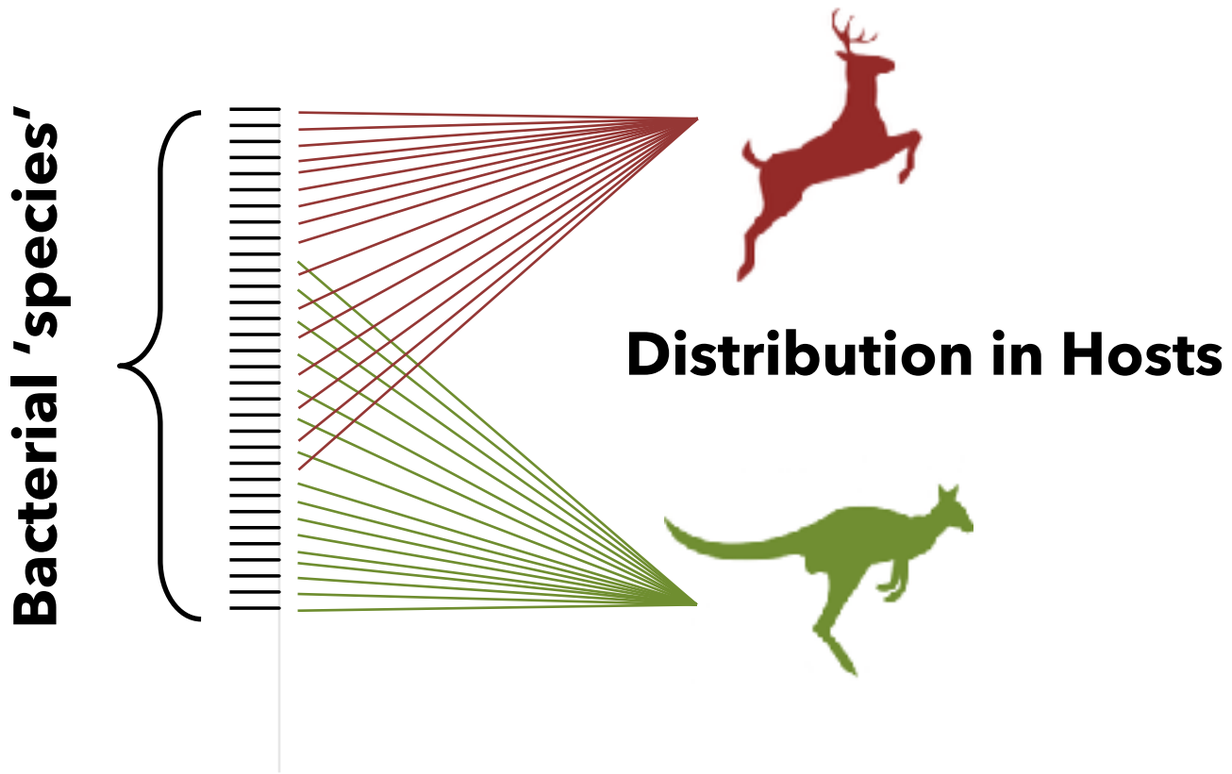


Distribution in Hosts

3 — Varying the phylogenetic resolution

Theory

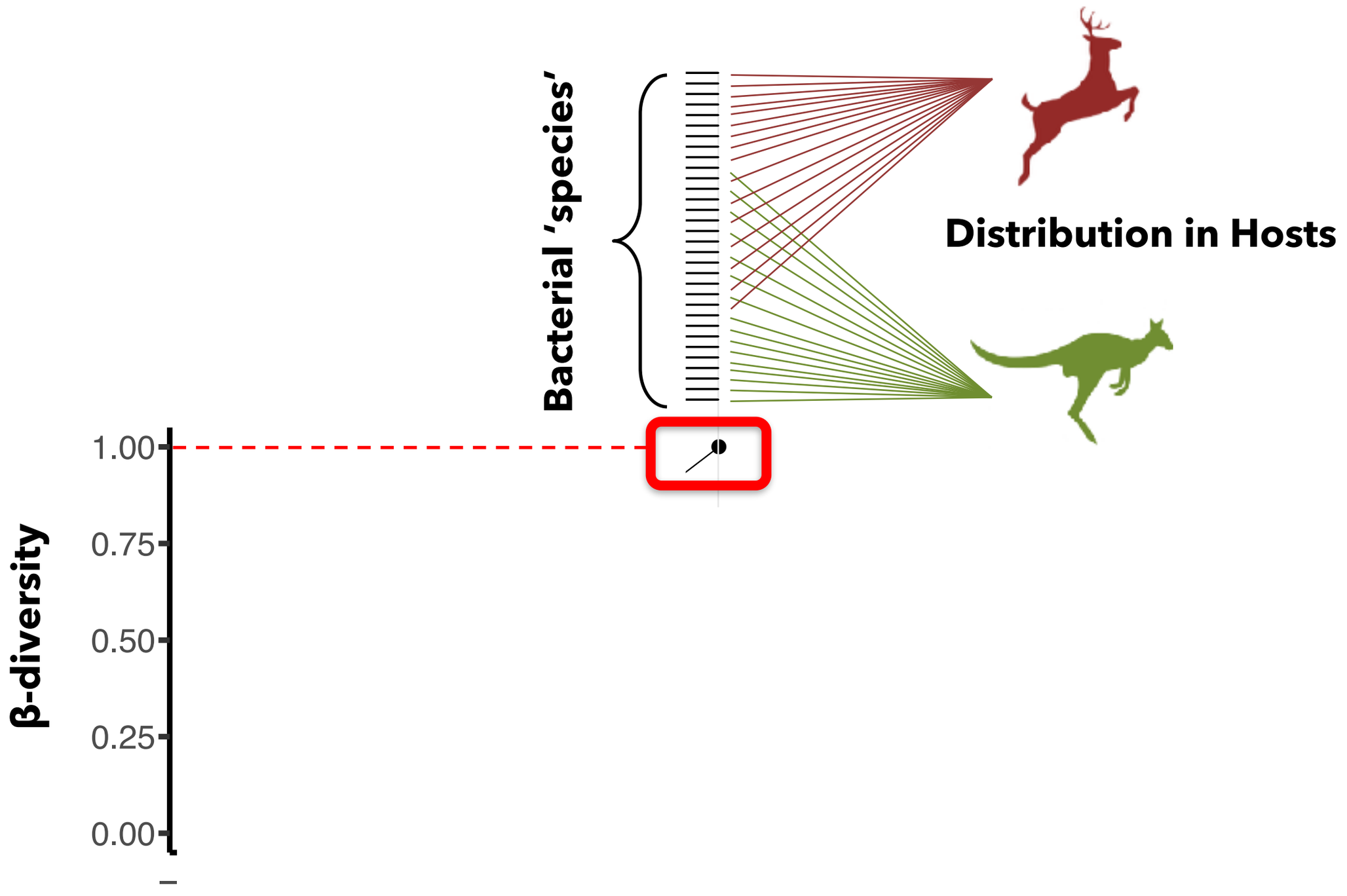
Method: “Slicing the tree of life”



3 — Varying the phylogenetic resolution

Theory

Method: "Slicing the tree of life"

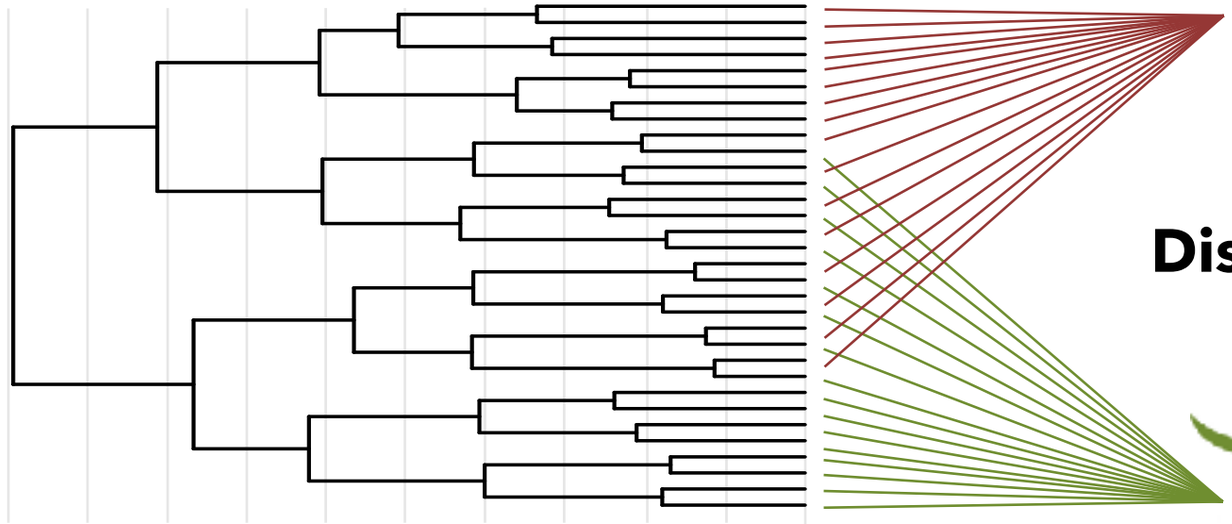


3 — Varying the phylogenetic resolution

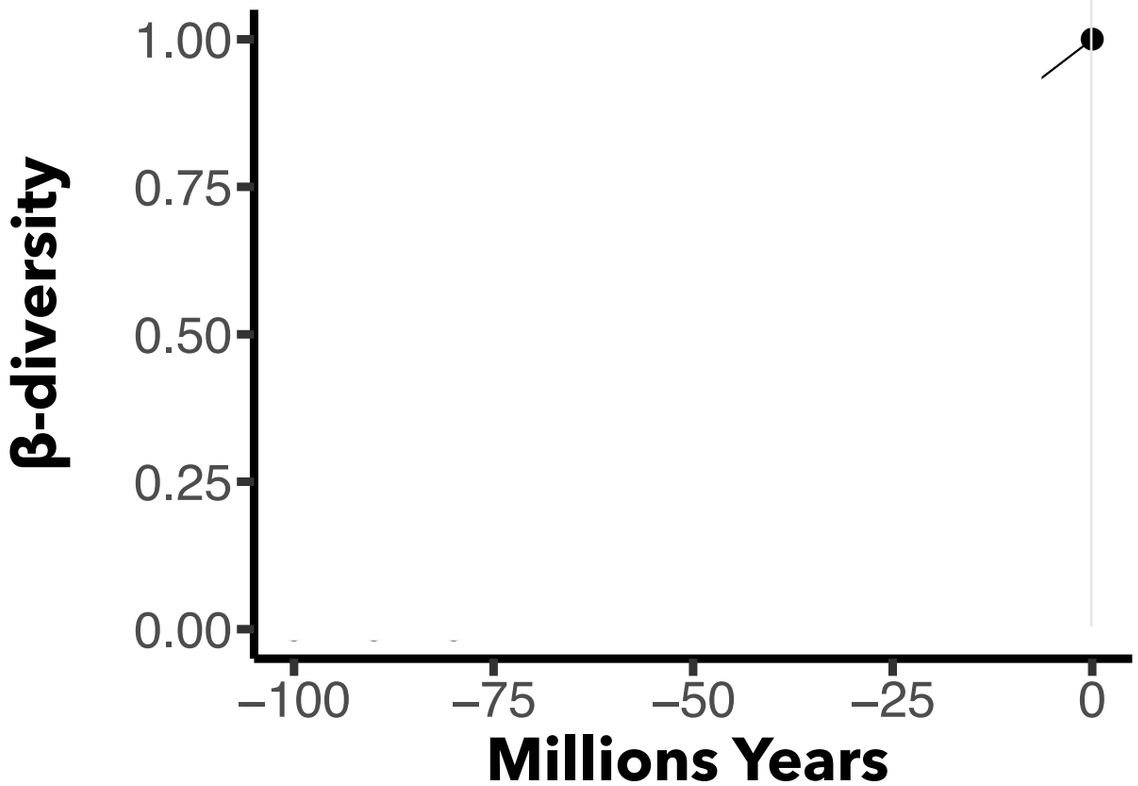
Theory

Method: "Slicing the tree of life"

Bacterial Phylogenetic tree



Distribution in Hosts

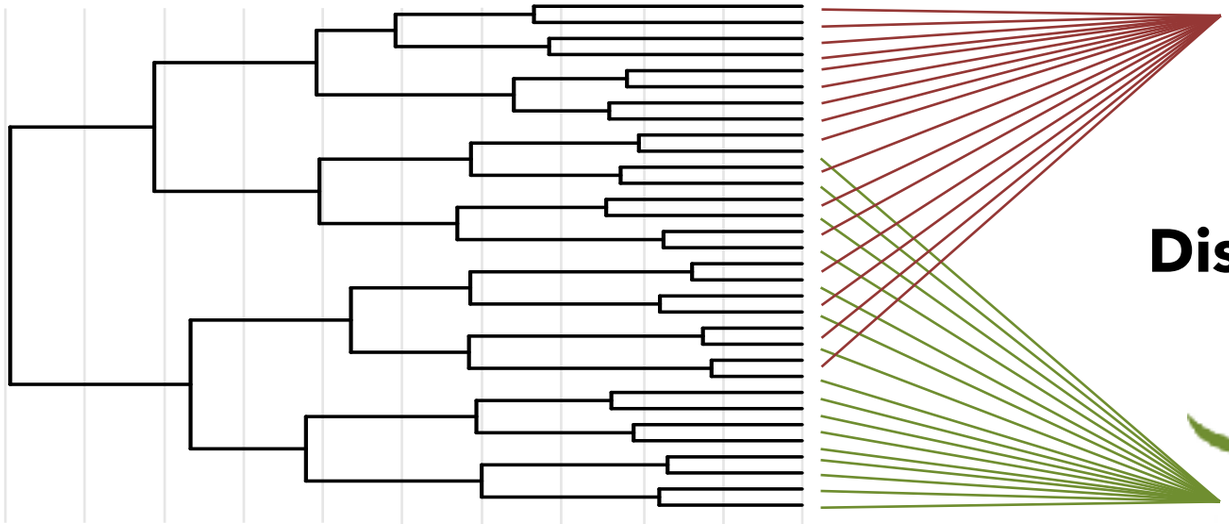


3 — Varying the phylogenetic resolution

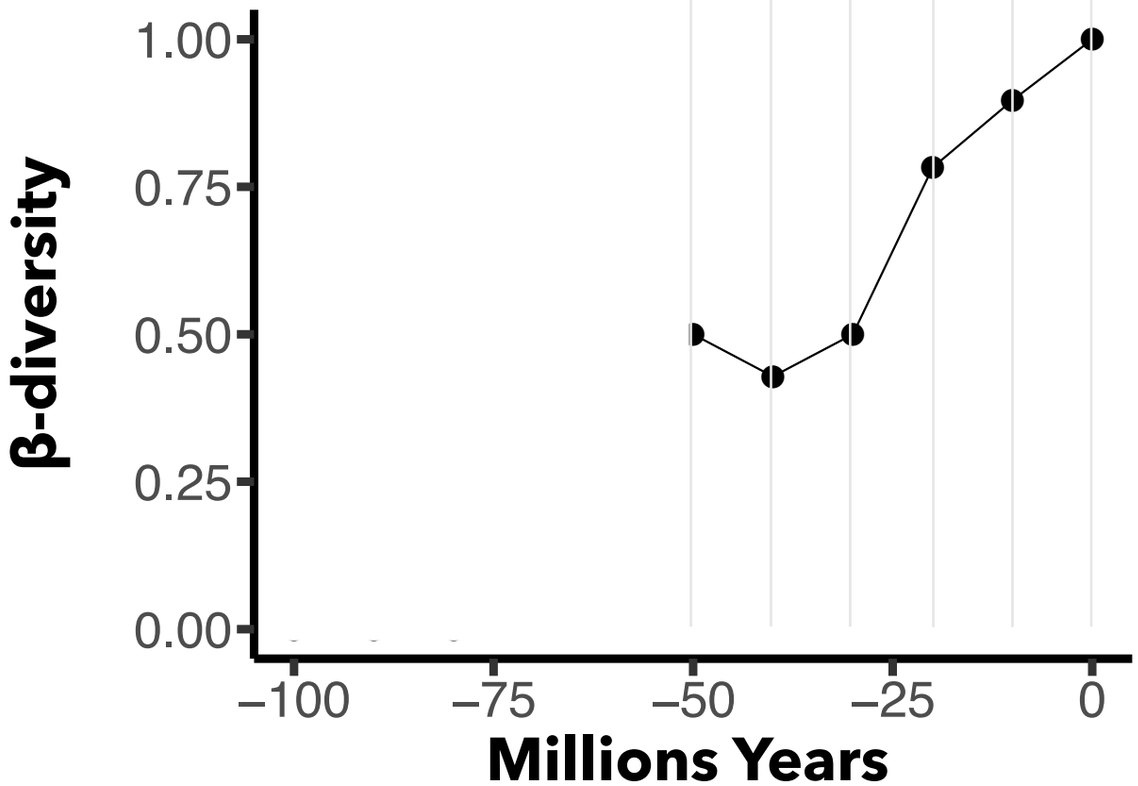
Theory

Method: "Slicing the tree of life"

Bacterial Phylogenetic tree



Distribution in Hosts

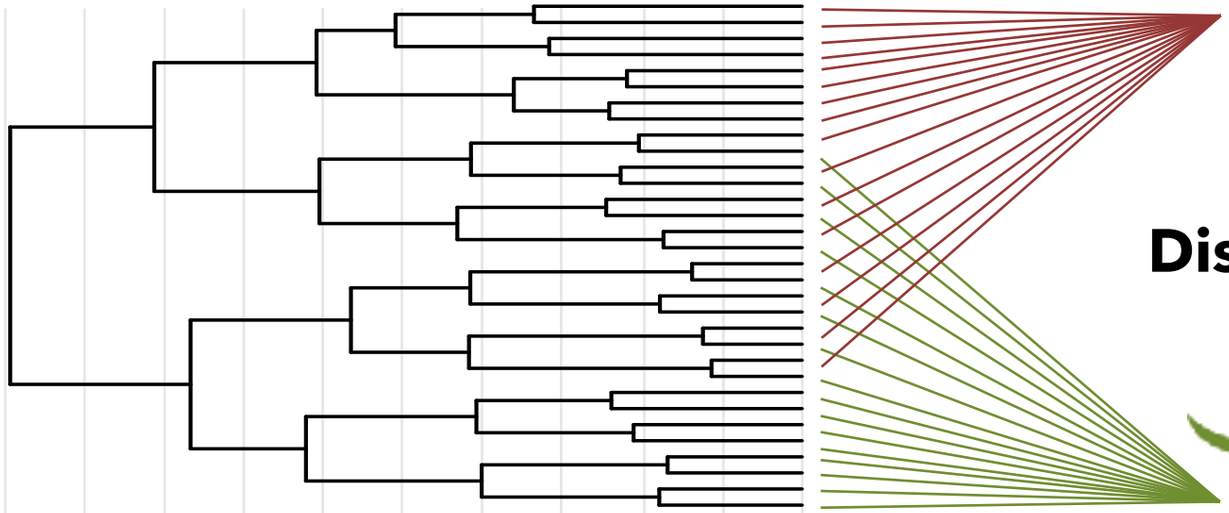


3 — Varying the phylogenetic resolution

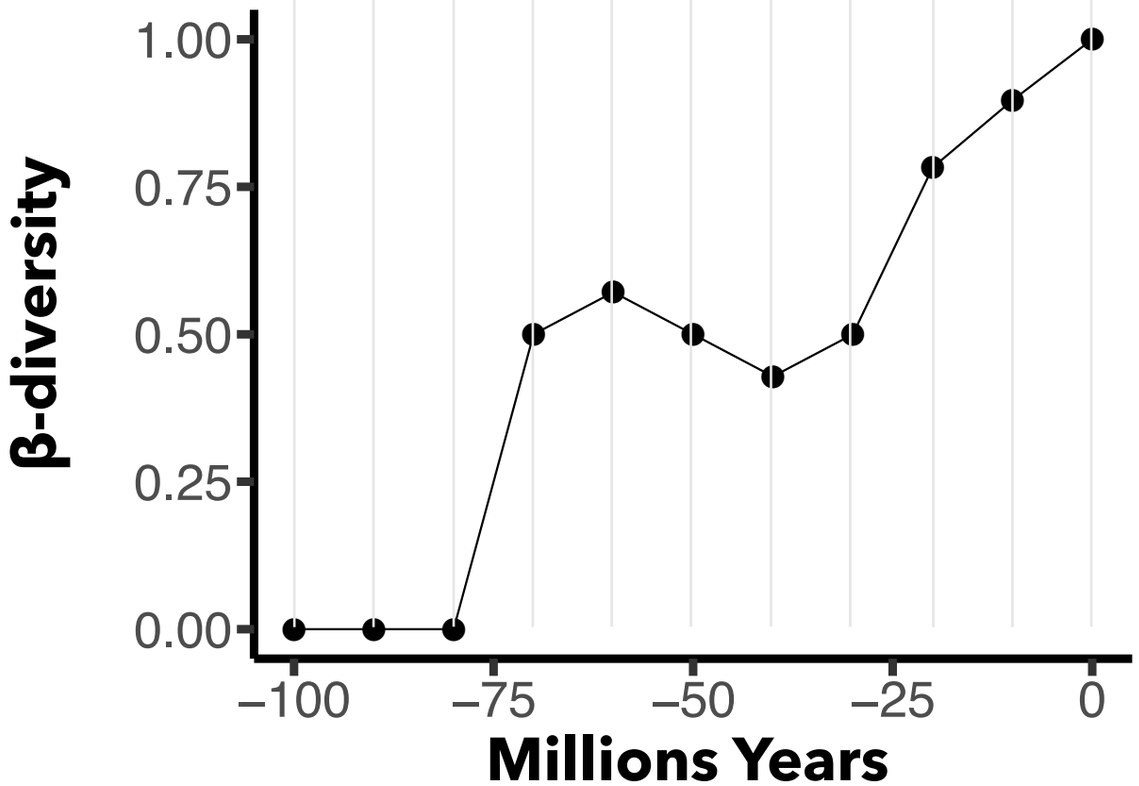
Theory

Method: "Slicing the tree of life"

Bacterial Phylogenetic tree



Distribution in Hosts



3 — Varying the phylogenetic resolution

Theory

Method: "Slicing the tree of life"

β-Diversity

Host Phylo Distances
Host Diet Distances

3 — Varying the phylogenetic resolution

Theory

Method: "Slicing the tree of life"

β-Diversity

Host Phylo Distances
Host Diet Distances

3 — Varying the phylogenetic resolution

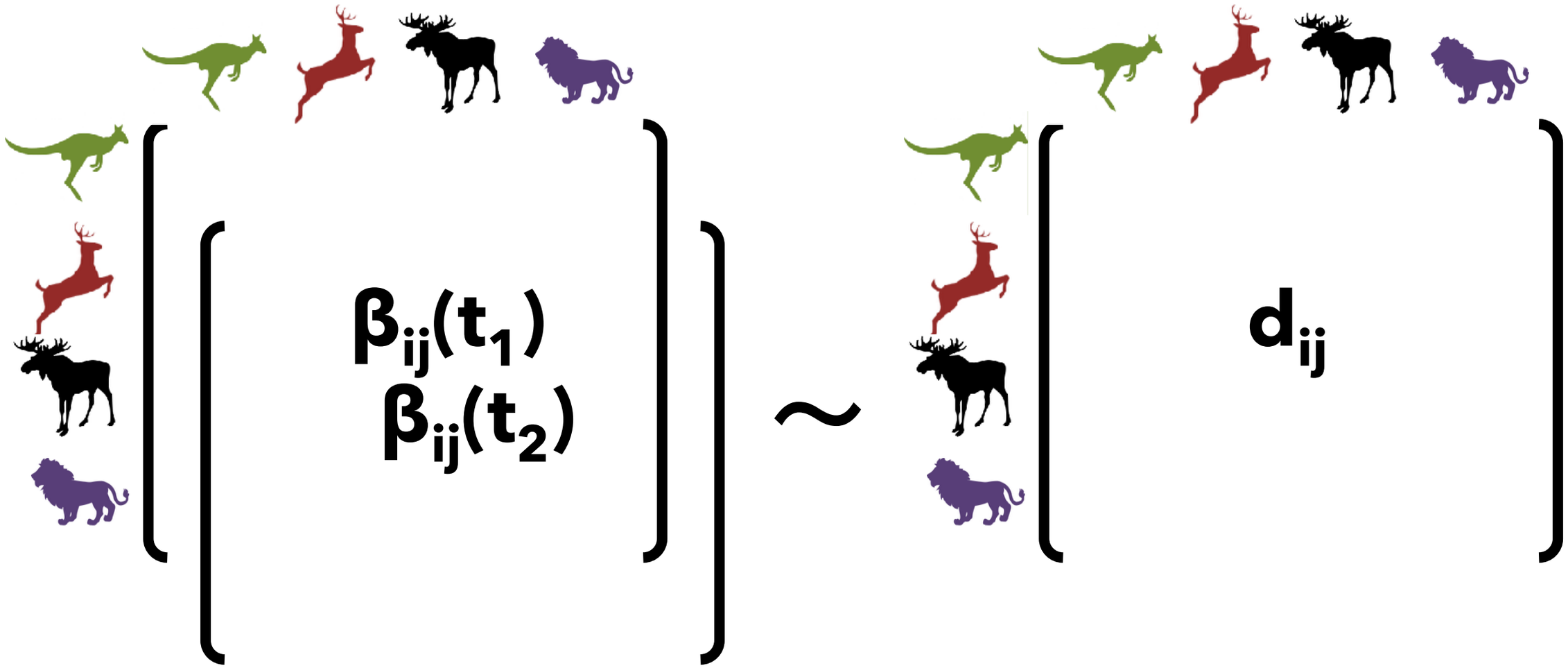
Theory

Method: "Slicing the tree of life"

β-Diversity

Host Phylo Distances

Host Diet Distances

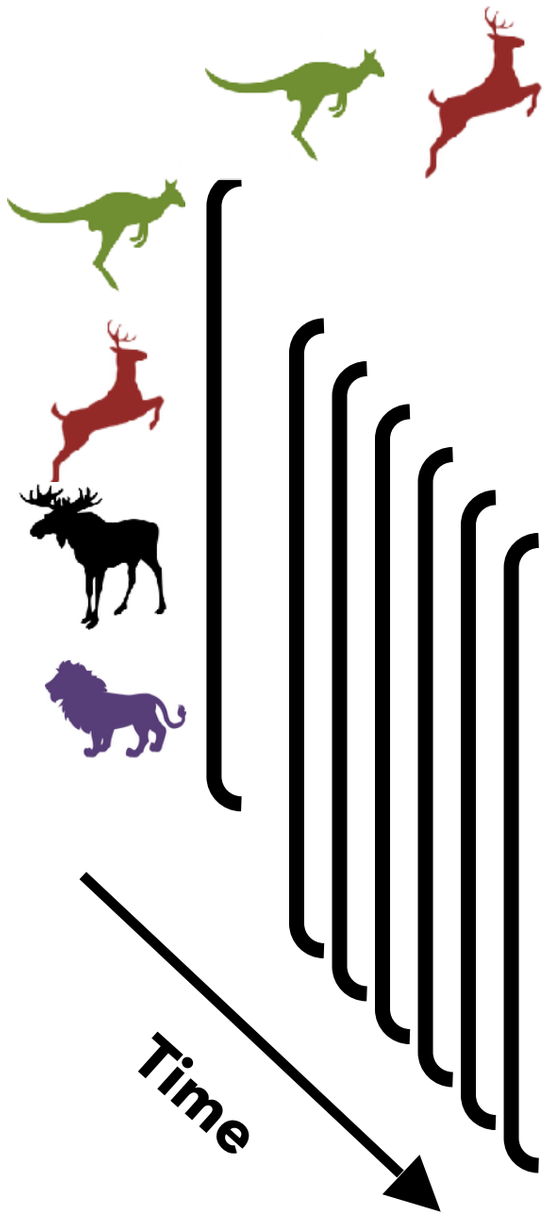


3 — Varying the phylogenetic resolution

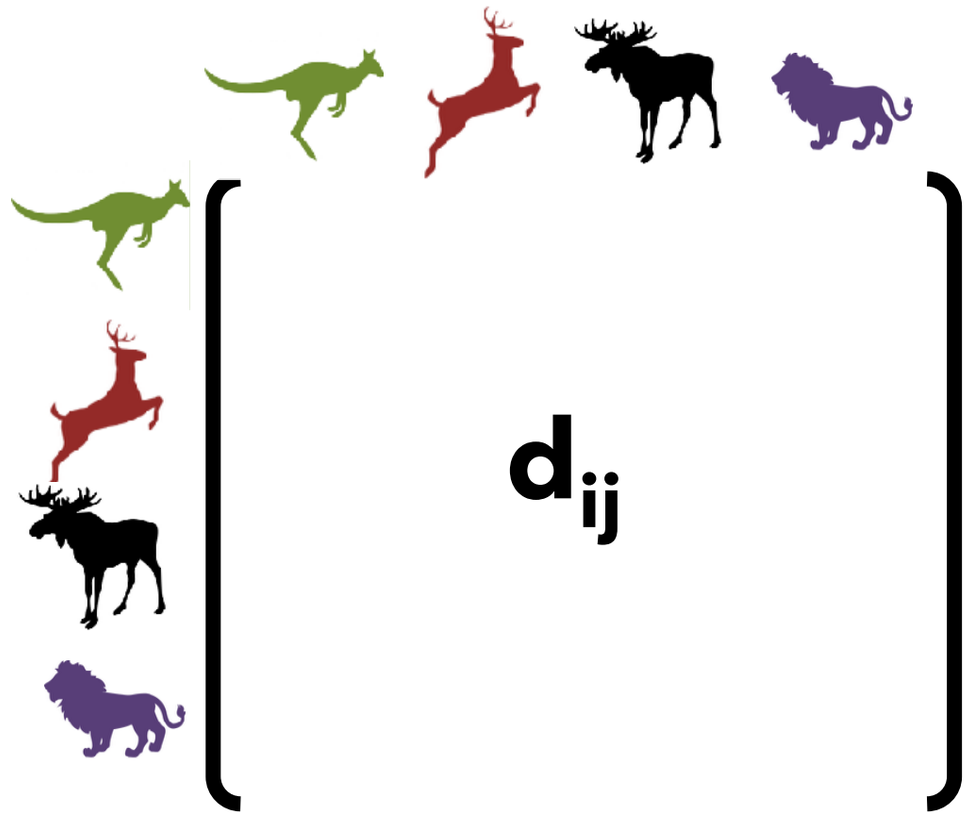
Theory

Method: "Slicing the tree of life"

β-Diversity



Host Phylo Distances
Host Diet Distances

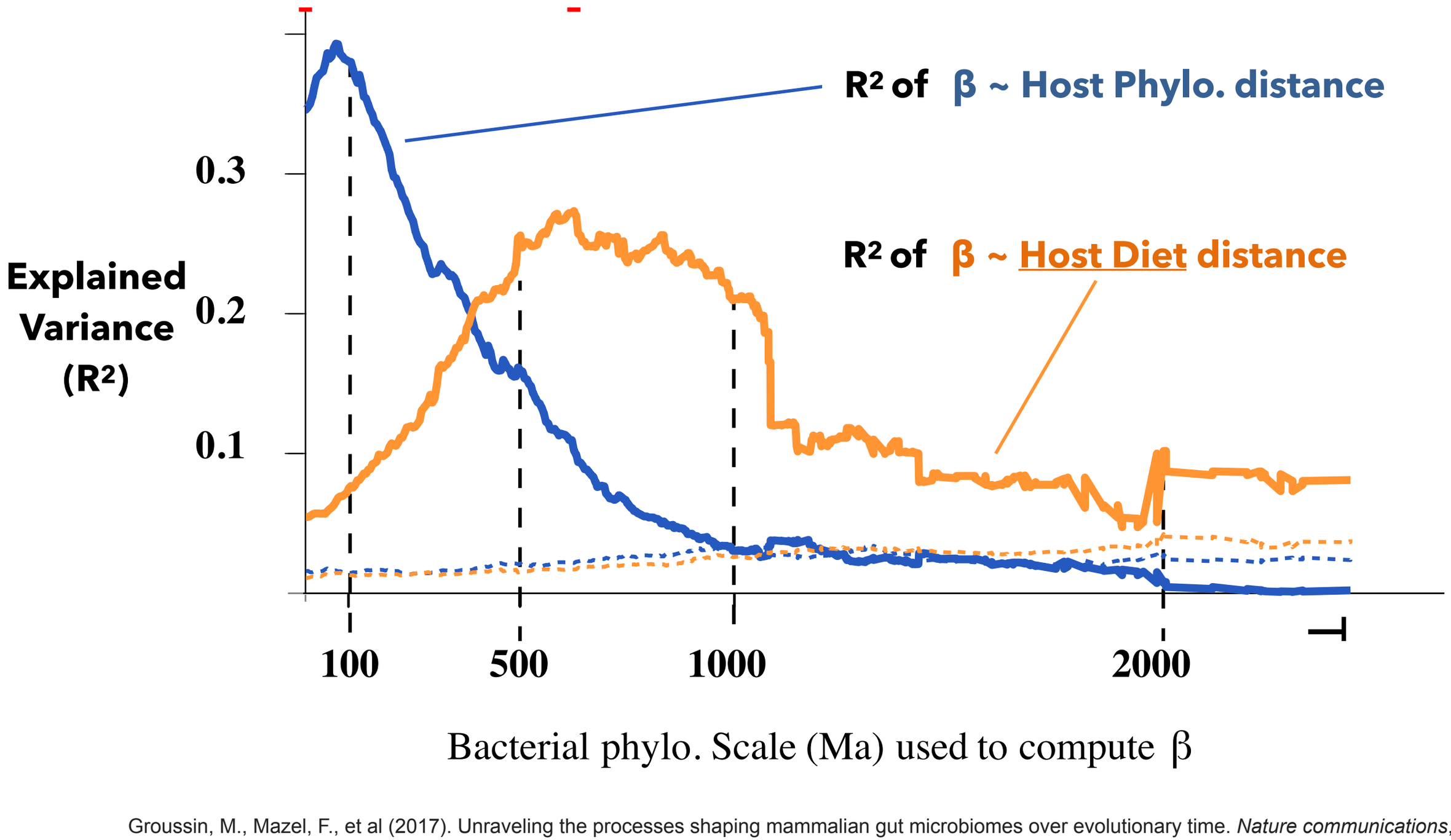


3 — Varying the phylogenetic resolution

Theory

***Results**: Does that really work ?*

Example 1: Mammalian guts



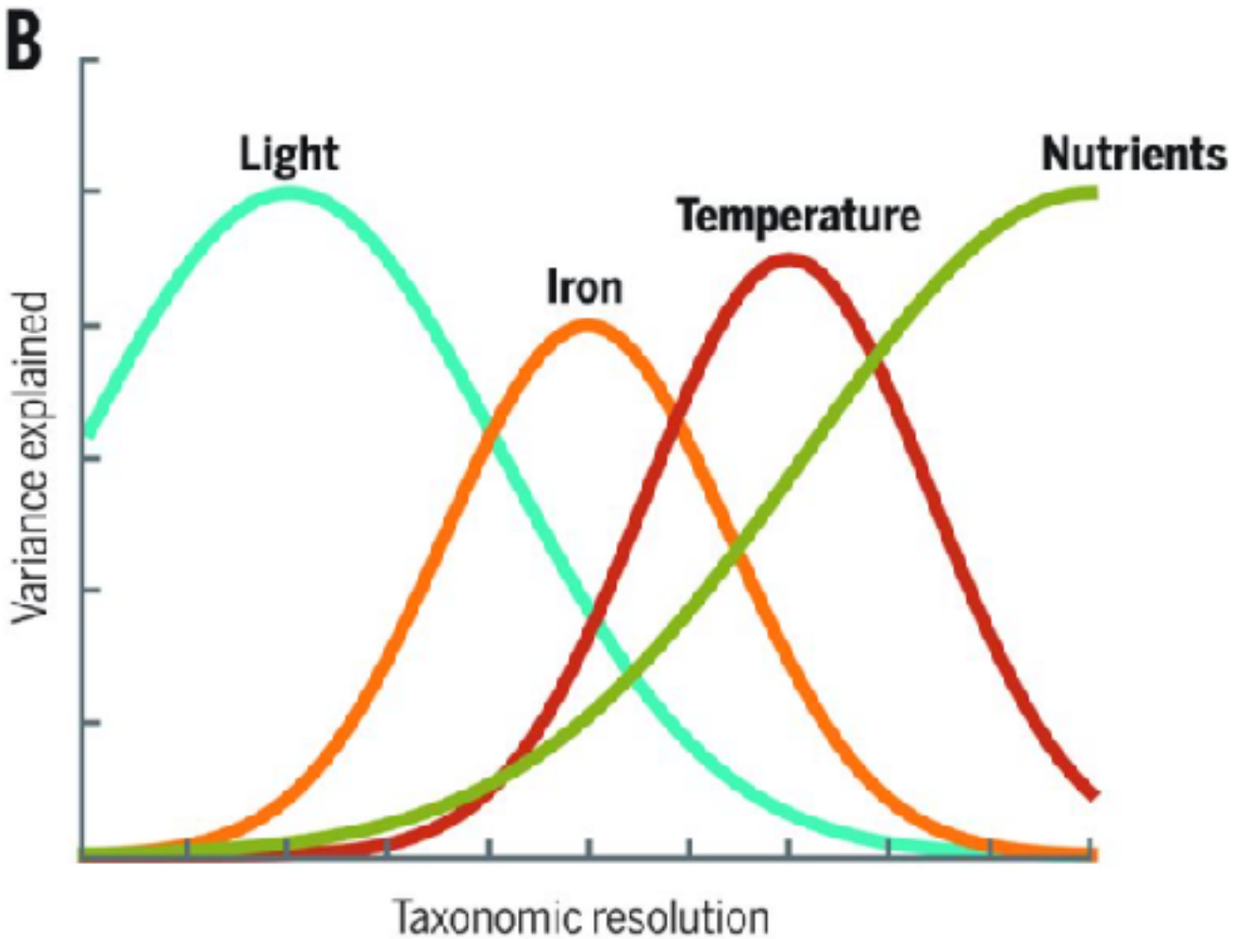
Groussin, M., Mazel, F., et al (2017). Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nature communications*, *8*, 14319.

3 — Varying the phylogenetic resolution

Theory

***Results**: Does that really work ?*

Example 2: Prochlorococcus sp.



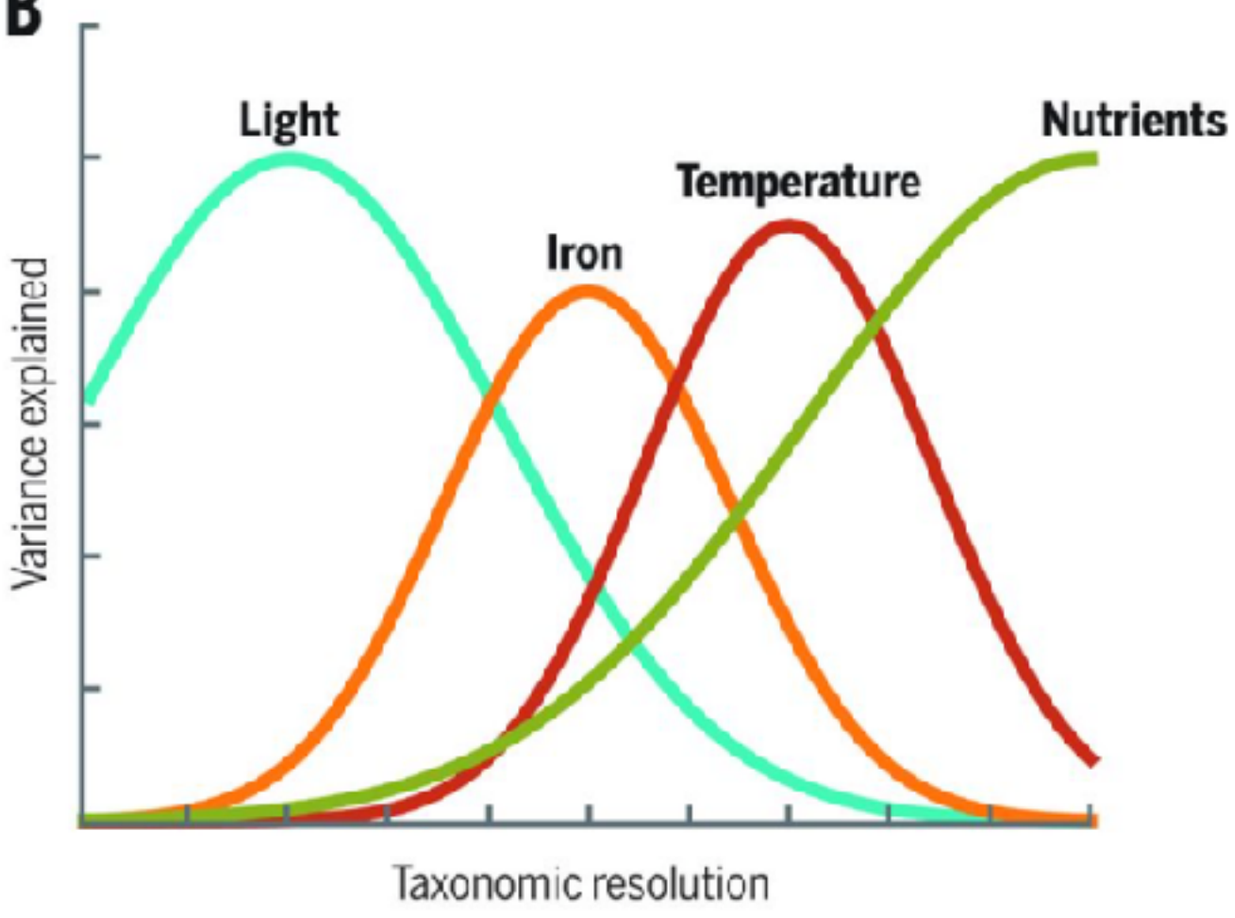
Martiny, J. B., Jones, S. E., Lennon, J. T., & Martiny, A. C. (2015). Microbiomes in light of traits: a phylogenetic perspective. *Science*, *350*(6261), aac9323.

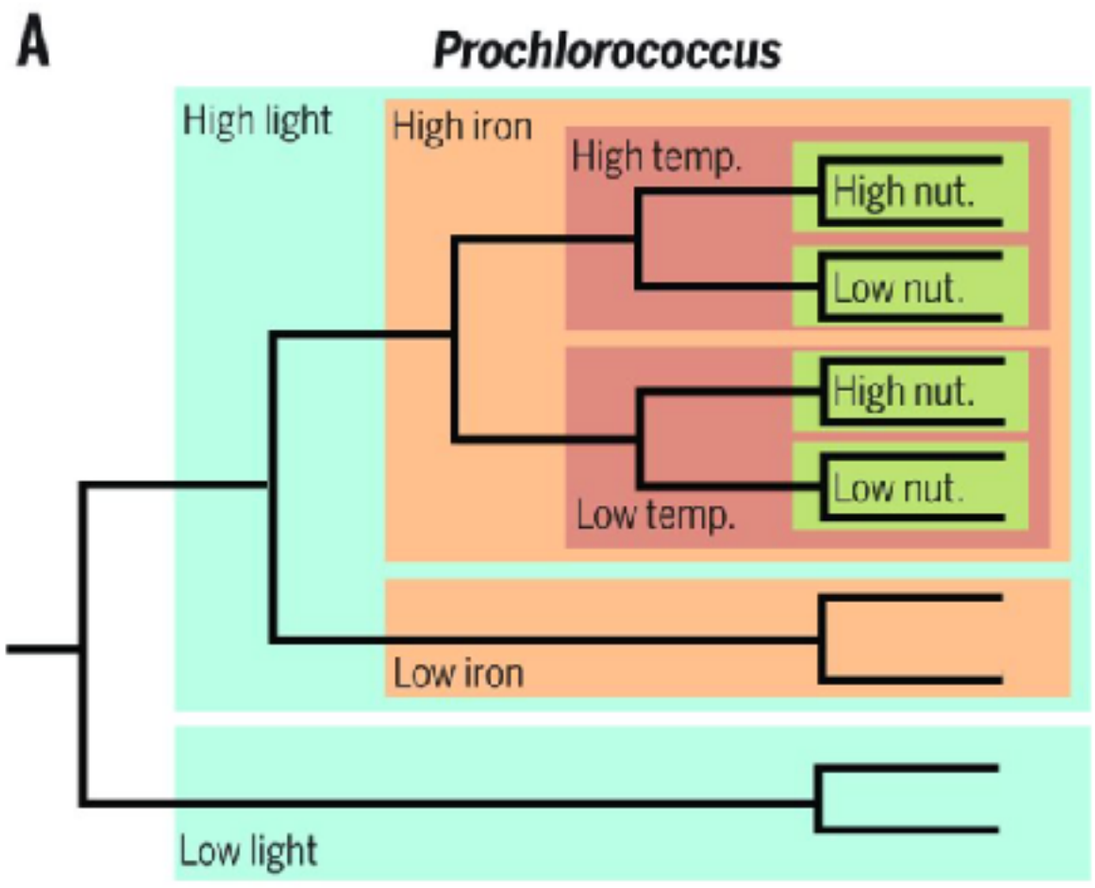
3 — Varying the phylogenetic resolution

Theory

***Results**: Does that really work ?*

Example 2: Prochlorococcus sp.





Martiny, J. B., Jones, S. E., Lennon, J. T., & Martiny, A. C. (2015). Microbiomes in light of traits: a phylogenetic perspective. *Science*, *350*(6261), aac9323.

Practice — 3 — Varying the phylogenetic resolution

Get a sense of the phylogenetic scale

```
> Hnodes()
```

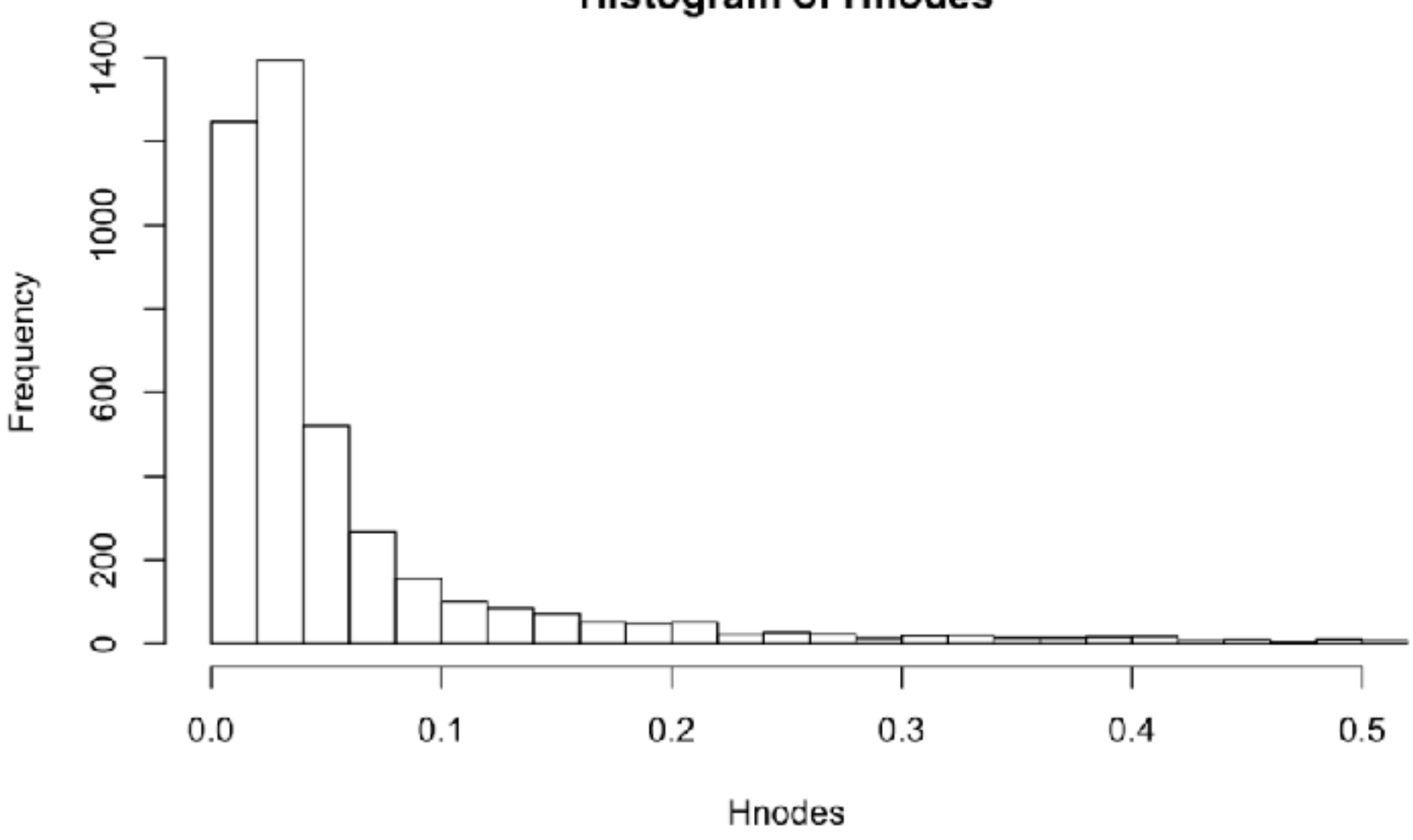
Compute Beta diversity through time (BDTT)

```
> BDTT()
```

3 — Varying the phylogenetic resolution

Get a sense of the phylogenetic scale

```
> Hnodes()
```



Practice

3 — Varying the phylogenetic resolution

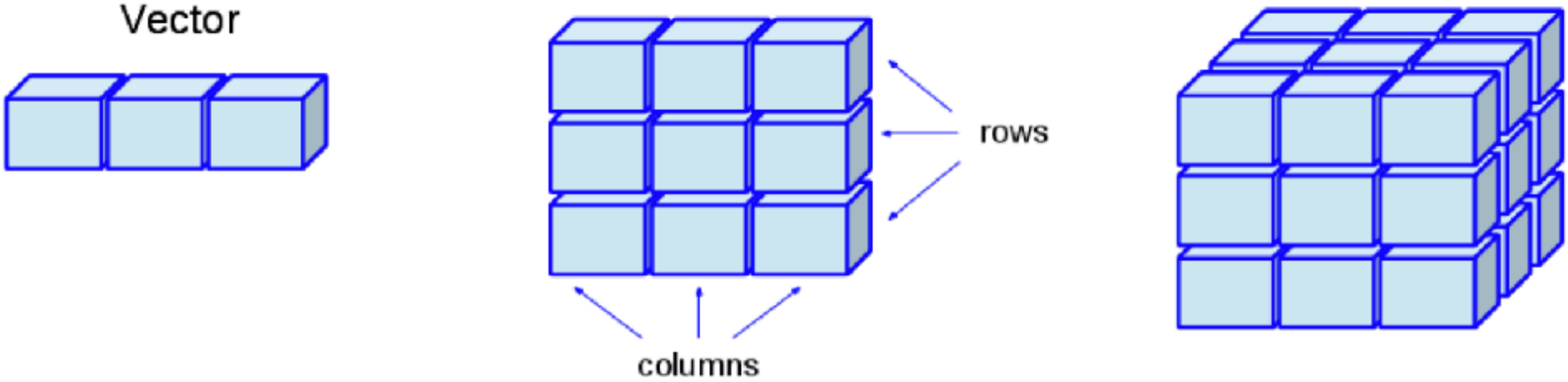
Run the BDTT

```
> BDTT()
```

```
> MultipleBetaJac=BDTT(similarity_slices=slices,tree=Tree,sampleOTUs=mat,onlyBeta=T,metric="ja
[1] "0 similarity provides 4315 total new OTUs"
[1] "0.025 similarity provides 2591 total new OTUs"
[1] "0.05 similarity provides 1365 total new OTUs"
[1] "0.075 similarity provides 932 total new OTUs"
[1] "0.1 similarity provides 733 total new OTUs"
[1] "0.125 similarity provides 606 total new OTUs"
[1] "0.15 similarity provides 515 total new OTUs"
[1] "0.175 similarity provides 435 total new OTUs"
[1] "0.2 similarity provides 378 total new OTUs"
[1] "0.225 similarity provides 322 total new OTUs"
```

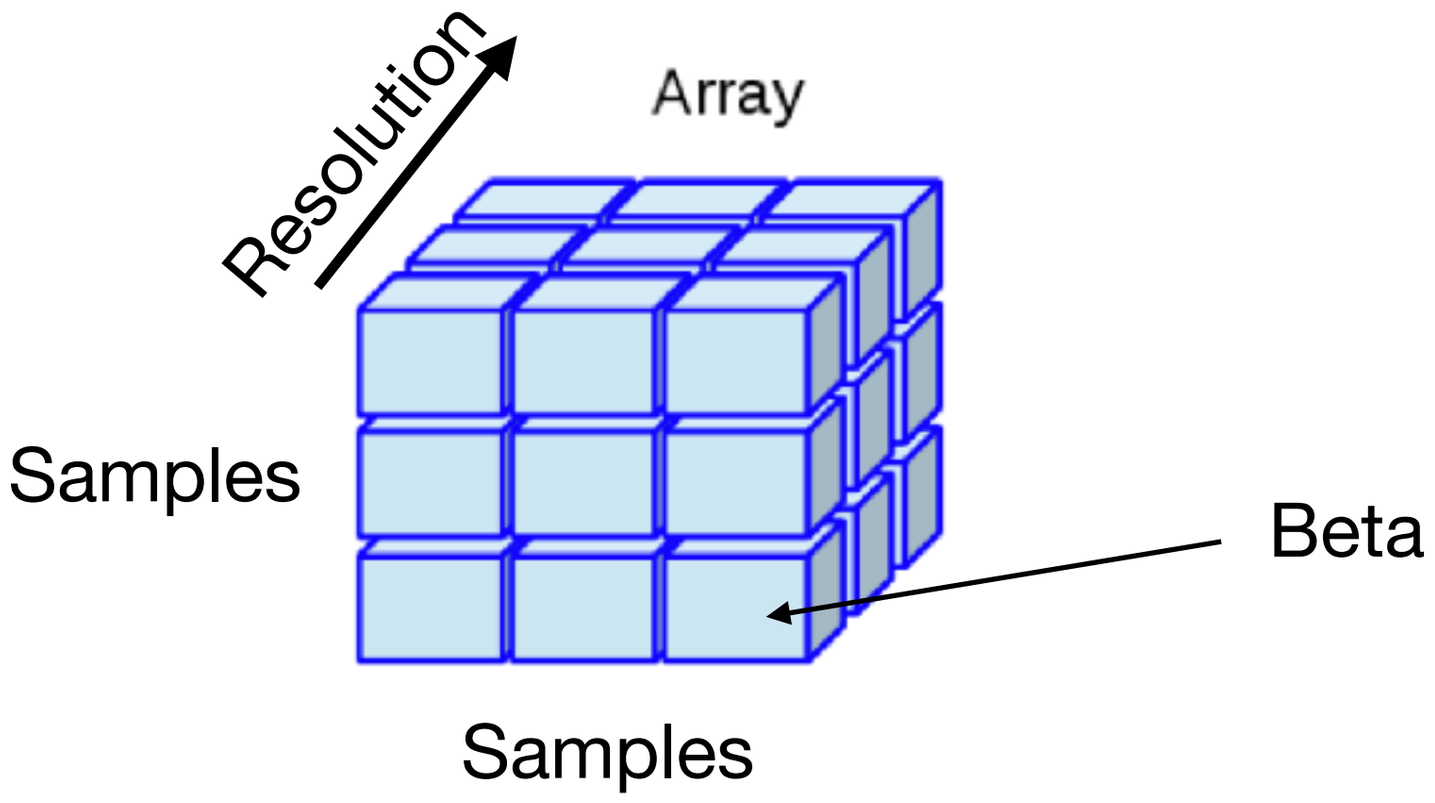
3 — Varying the phylogenetic resolution

Explore the output



3 — Varying the phylogenetic resolution

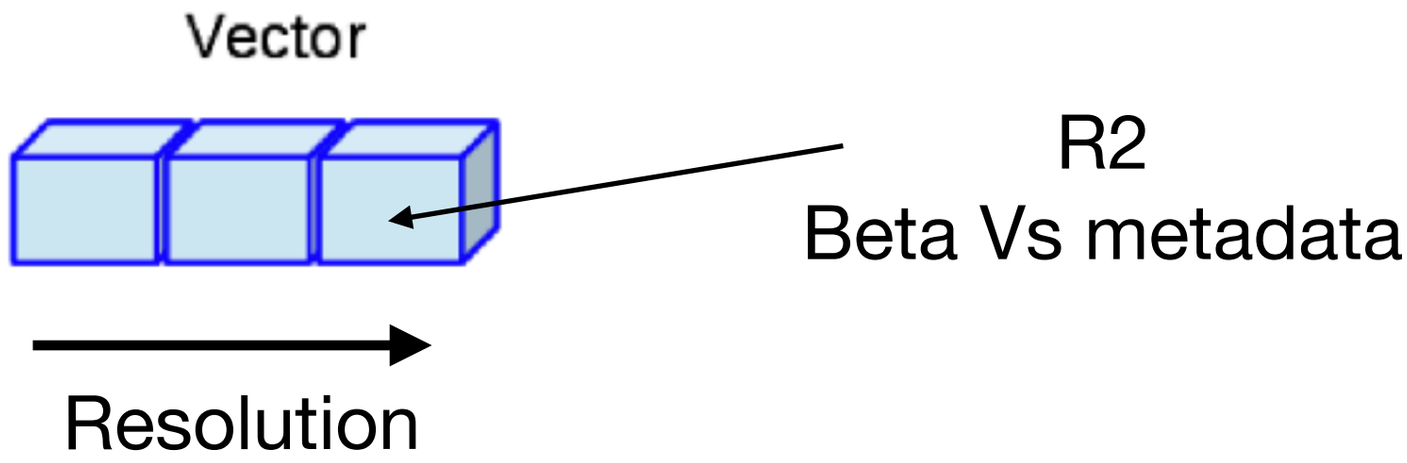
Explore the output



Practice

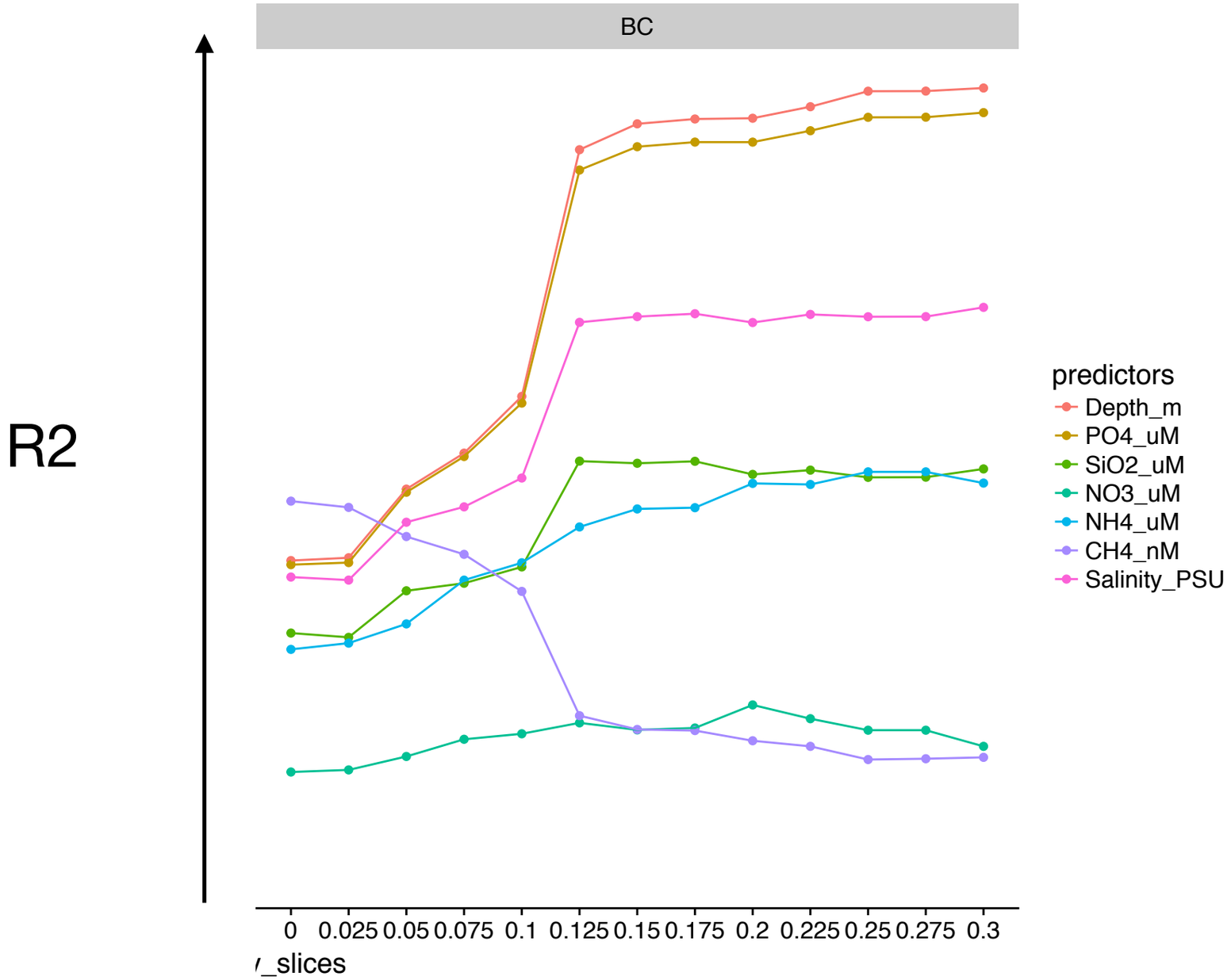
3 — Varying the phylogenetic resolution

Link with metadata



Practice

3 — Varying the phylogenetic resolution



Overall structure of the workshop

1 — Building a phylogenetic tree

2 — Classical analysis of microbiome

3 — Varying the phylogenetic resolution

4 — Exploring the branches of the phylogenetic tree

4 — Phylogenetic compositional factor analysis

Theory

Two problems:

1. *How do you identify changing microbes at unknown levels?*
2. *How do you deal with compositionality?*

4 — PhyloFactor

Theory

Jamie Morton

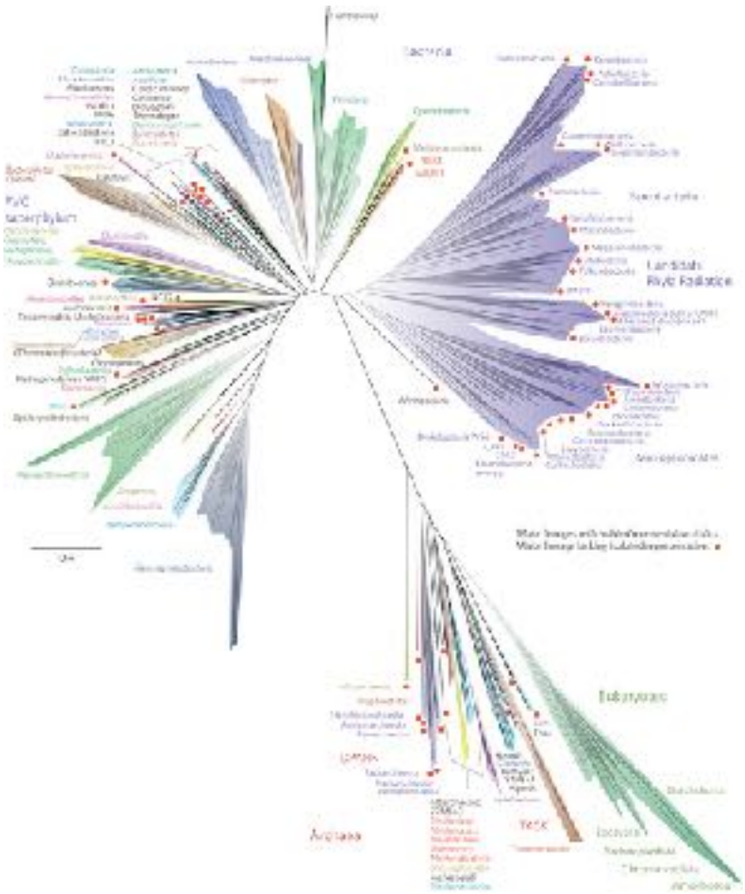
Alex Washburne

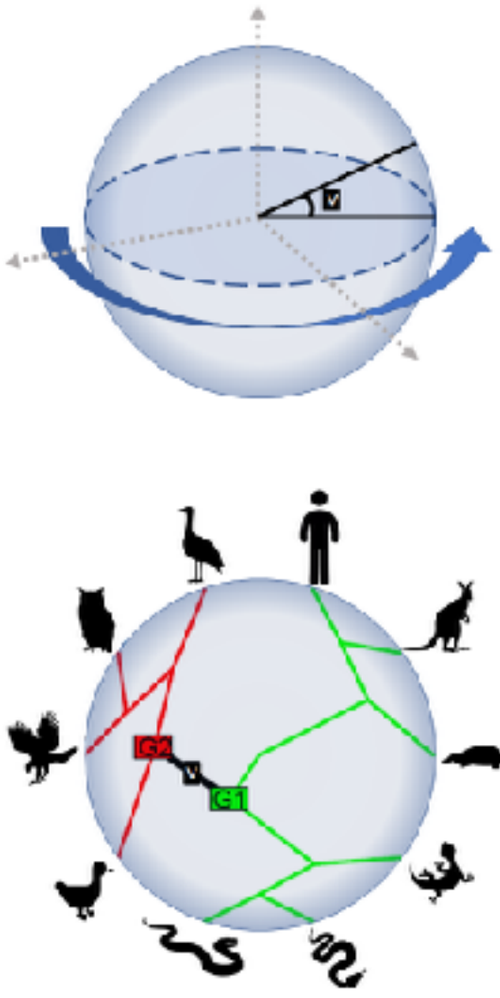
Washburne, Morton *et al*. **2018.** Methods for phylogenetic analysis of microbiome data. *Nature Microbiology*

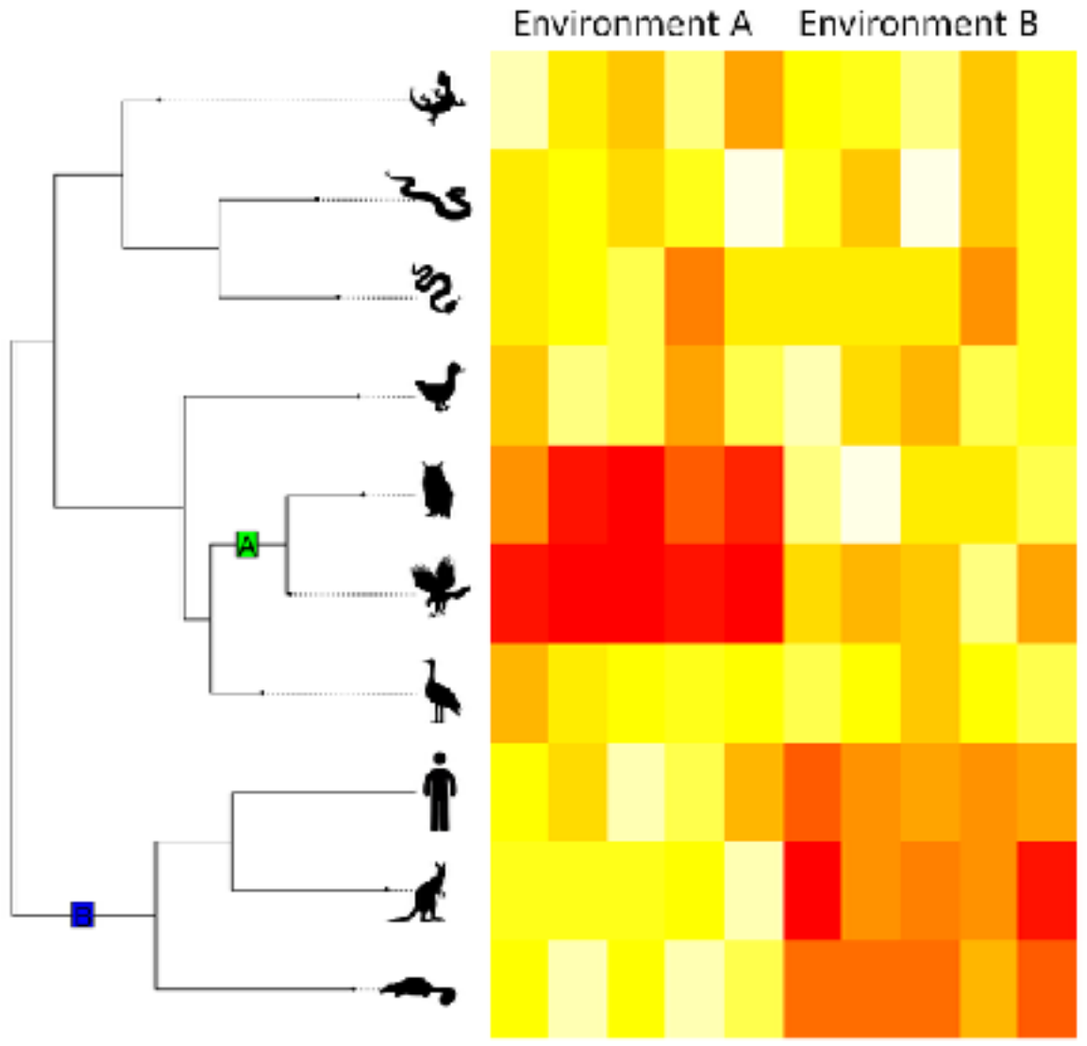
4 — PhyloFactor

Theory

Phylogenetic Variables





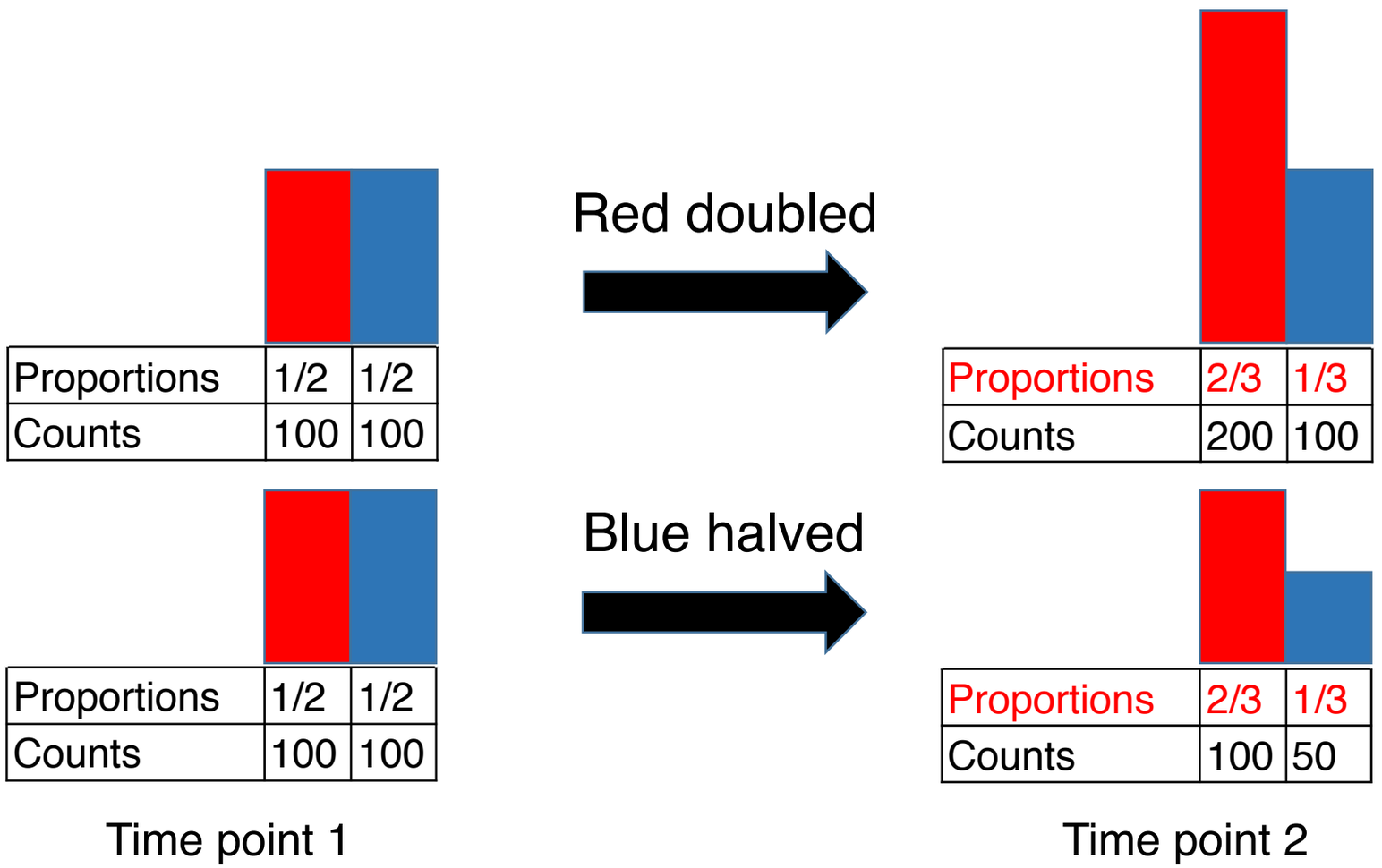


Washburne, Morton et al. (2018) Methods for phylogenetic analysis of microbiome data. *Nature Microbiology*

Theory

Compositional data

For one element to increase, remaining elements must decrease



Theory

Compositional data

Impossible to know from sequence data alone whether one population is increasing, or another population is decreasing!

Compositional analyses ask instead: how does ***share of community*** *change?*

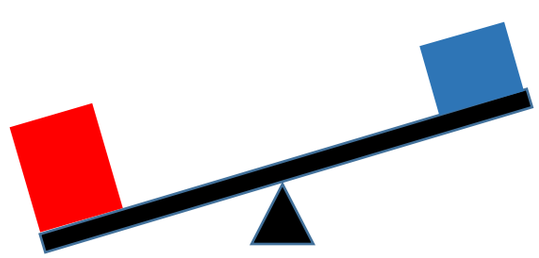
Theory

Compositional analyses ask instead: how does ***share of community*** *change?*

Red doubled
or
Blue halved

$$balance = \log\left(\frac{100}{100}\right) = 0$$

Time point 1



$$balance = \log\left(\frac{100}{50}\right) = \log\left(\frac{200}{100}\right) = \log\left(\frac{2/3}{1/3}\right) = \log 2$$

Time point 2

Jamie Morton *et al*. 2017, mSystems

Theory

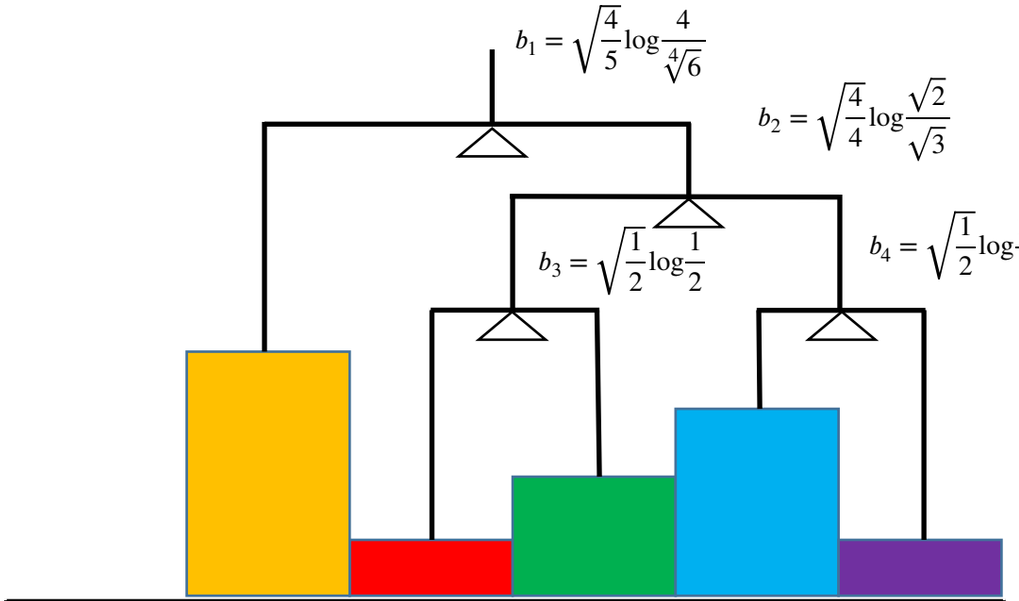
Balance trees

Use a **rooted tree** to construct a *basis* for projecting count data using the **Isometric Log Ratio** (ILR) transform

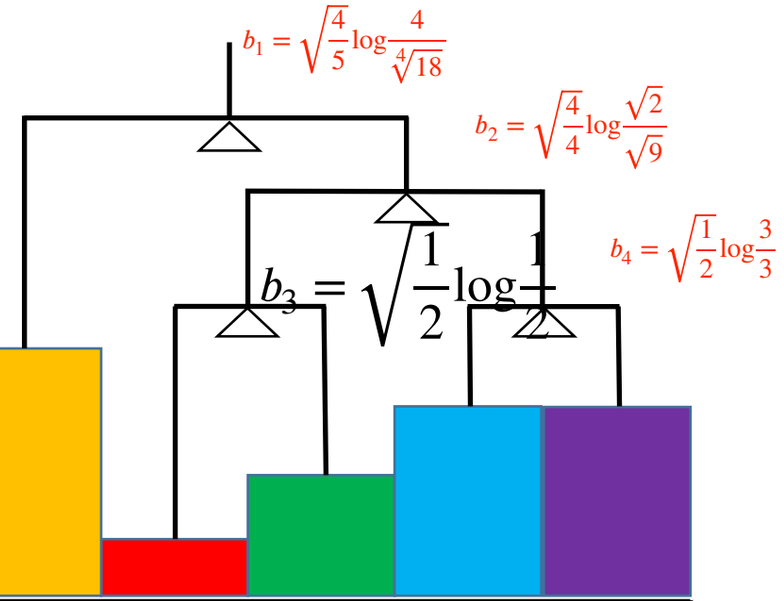
4 — PhyloFactor

Theory

Balance trees



Proportions	4/11	1/11	2/11	3/11	1/11
Percentages	36%	9%	18%	27%	9%
Counts	4	1	2	3	1



Proportions	4/13	1/13	2/13	3/13	3/13
Percentages	30%	7%	14%	21%	21%
Counts	4	1	2	3	3

$$b_i = \sqrt{\frac{|i_L||i_R|}{|i_L| + |i_R|}}\log\left(\frac{g(i_L)}{g(i_R)}\right)$$

Egozcue (2003)

Jamie Morton *et al*. 2017, mSystems

4 — PhyloFactor

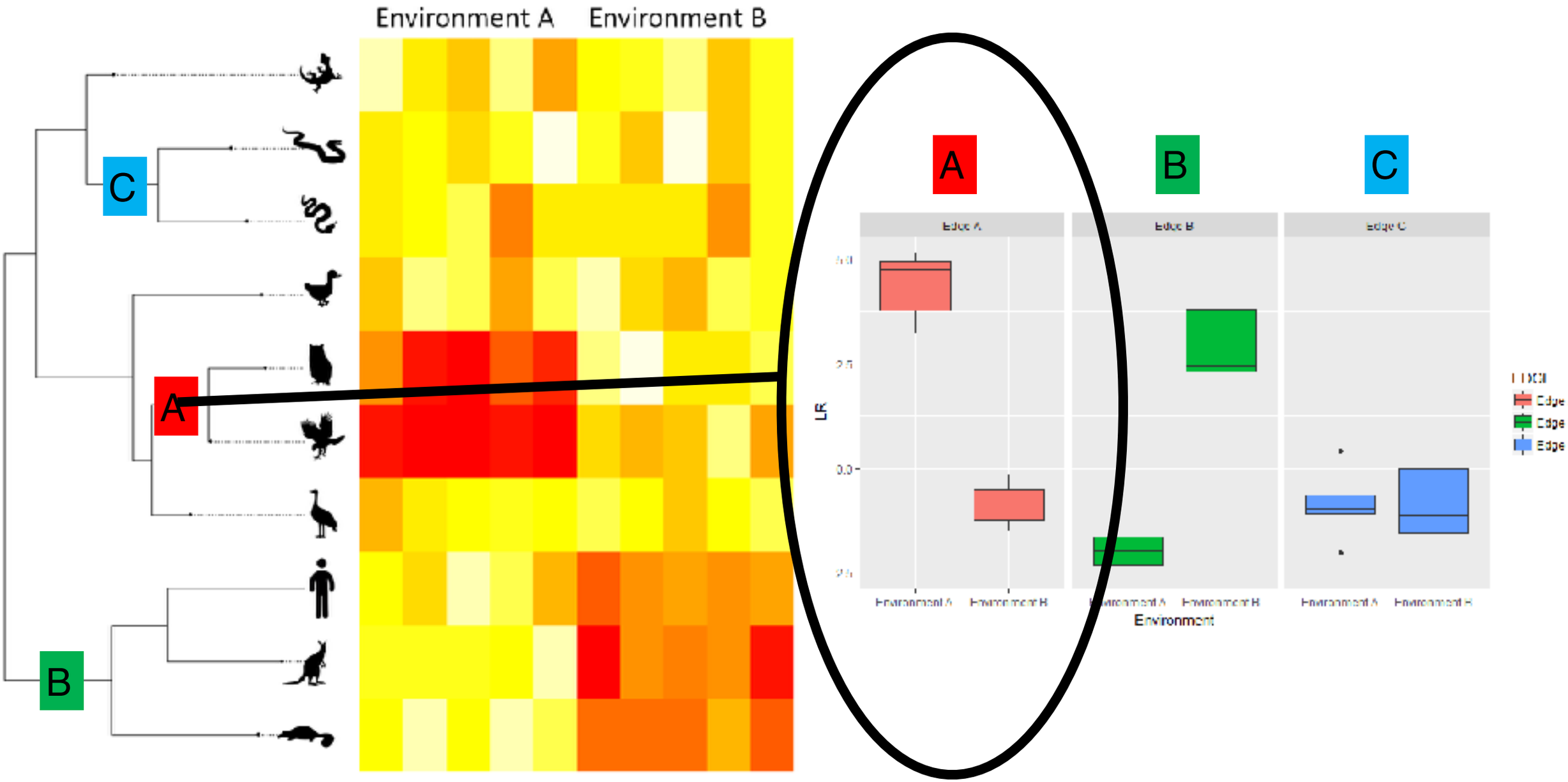
Theory

PhyloFactorization

Uses an **unrooted tree** to iteratively construct a ILR basis, identifying ‘most important’ edge on tree (balance) at each iteration

4 — PhyloFactor

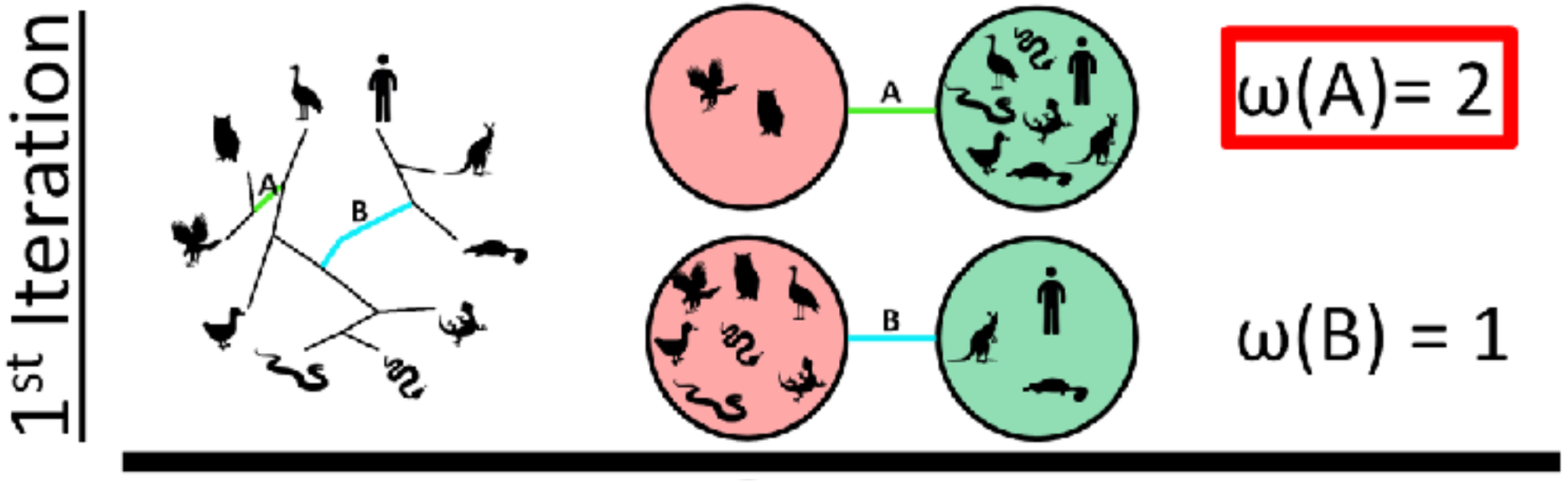
Theory

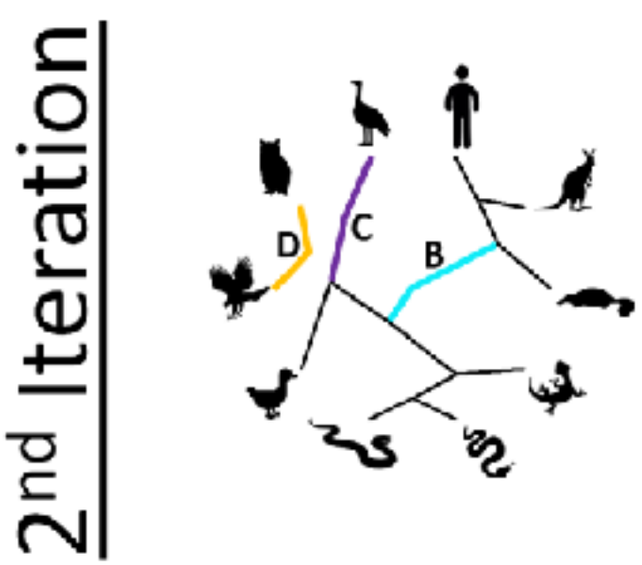


Washburne et al. (2017) *PeerJ*

4 — PhyloFactor

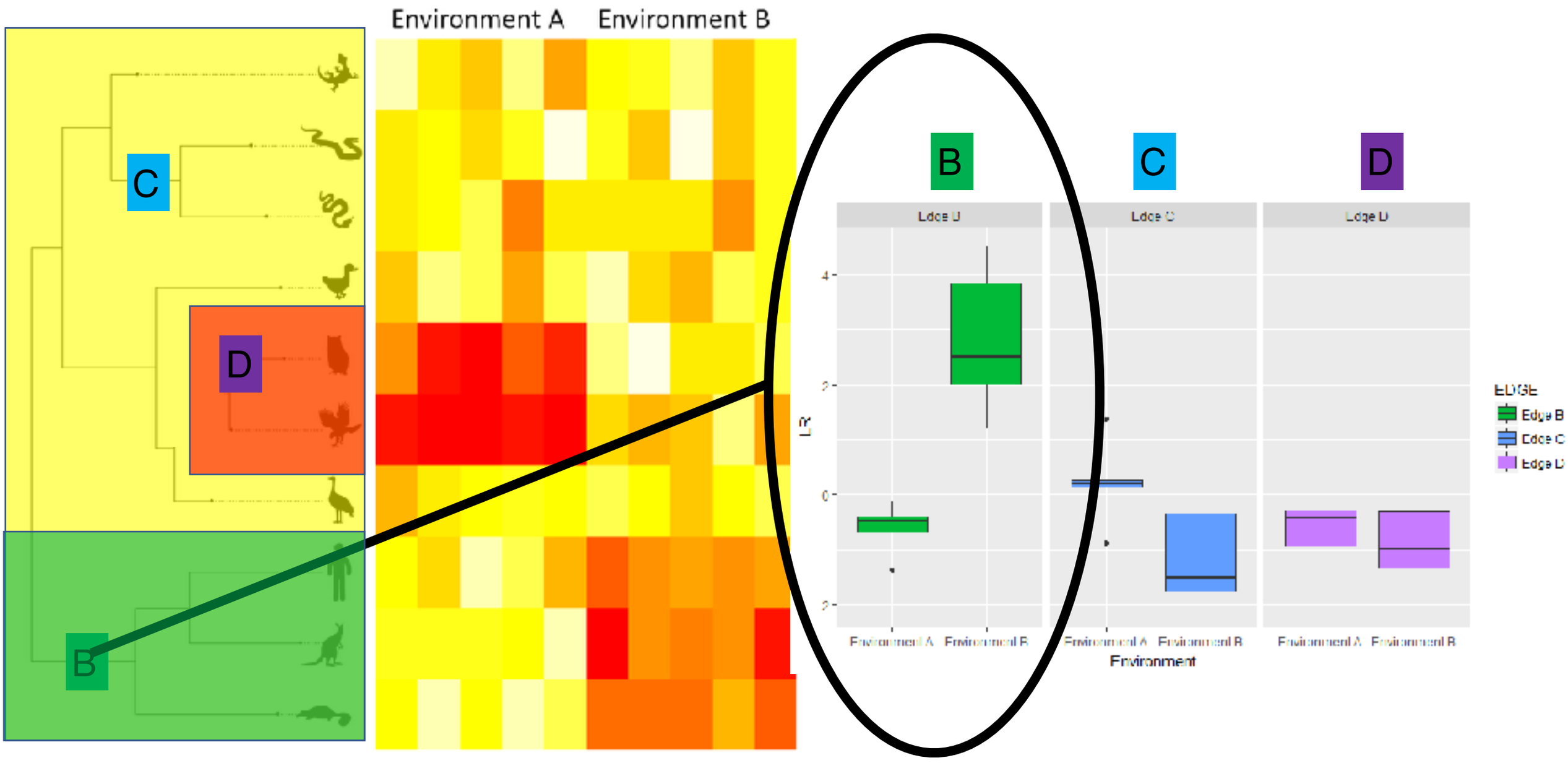
Theory





4 — PhyloFactor

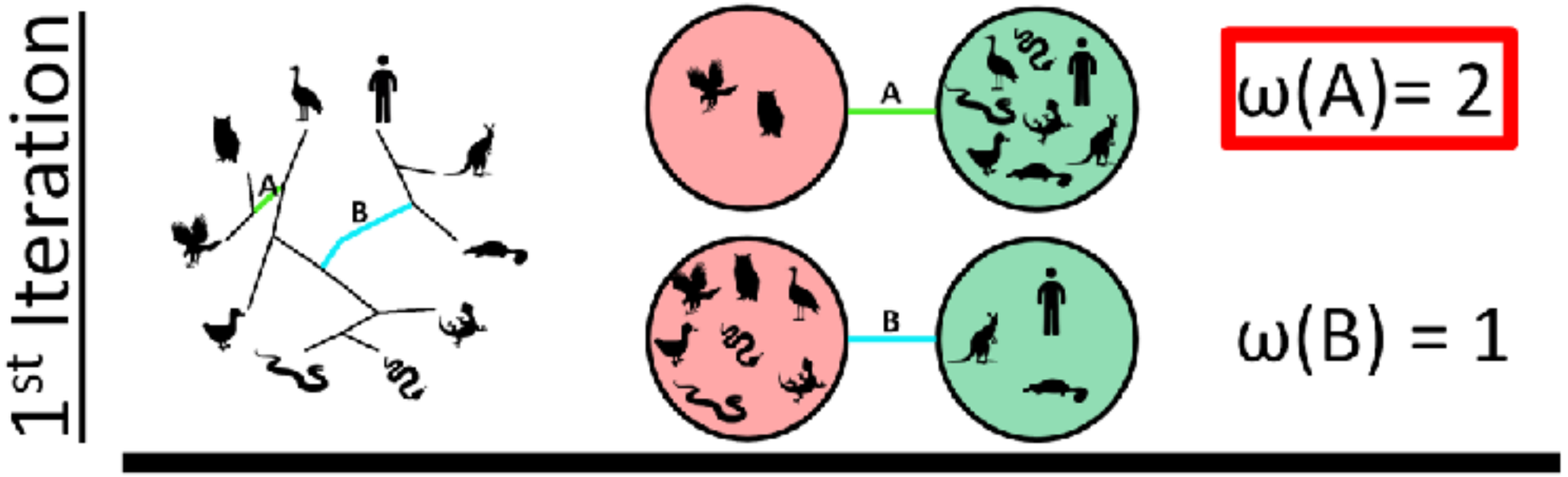
Theory

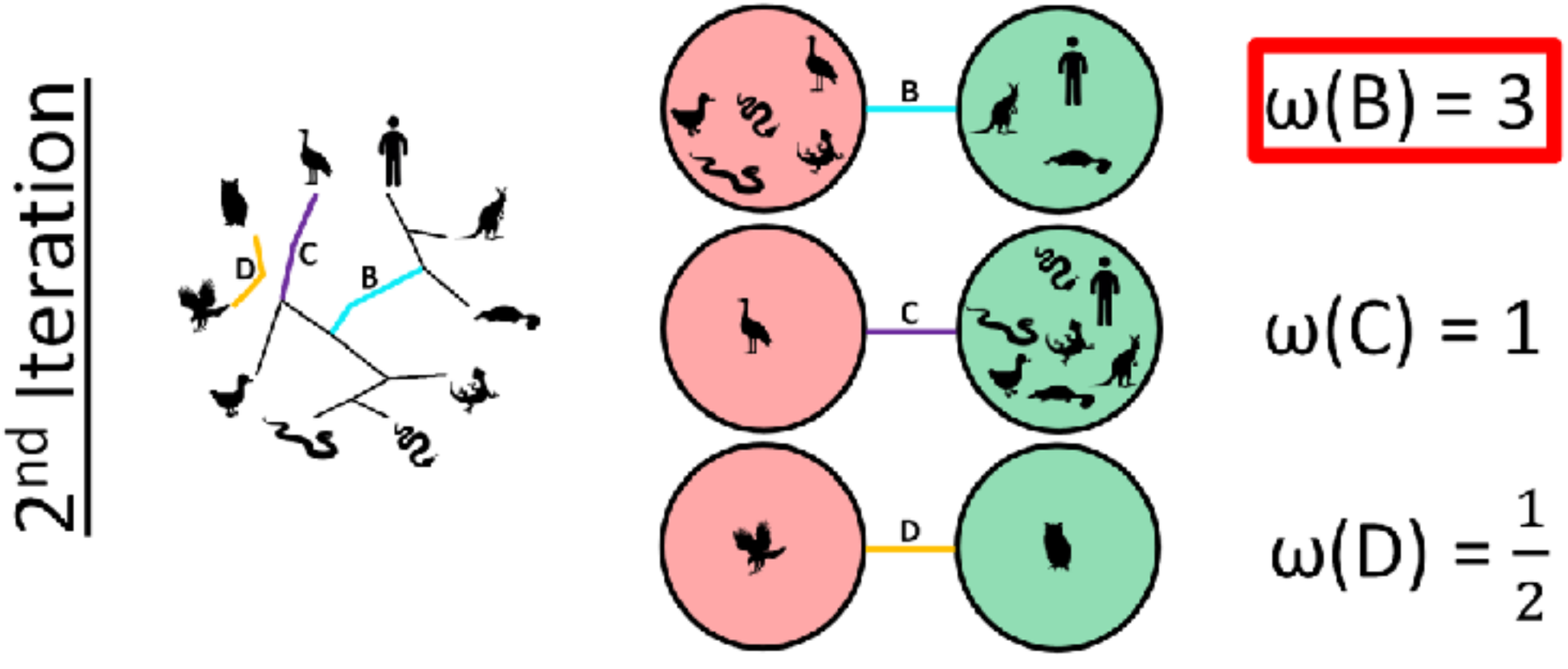


Washburne et al. (2017) *PeerJ*

4 — PhyloFactor

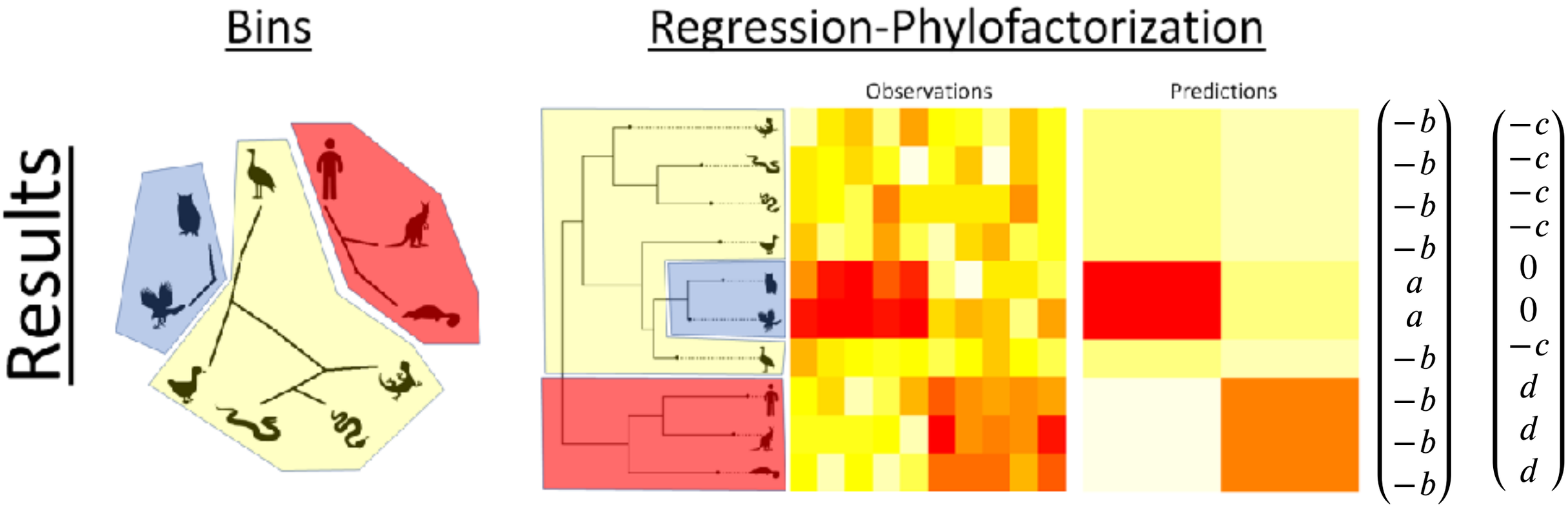
Theory





4 — PhyloFactor

Theory



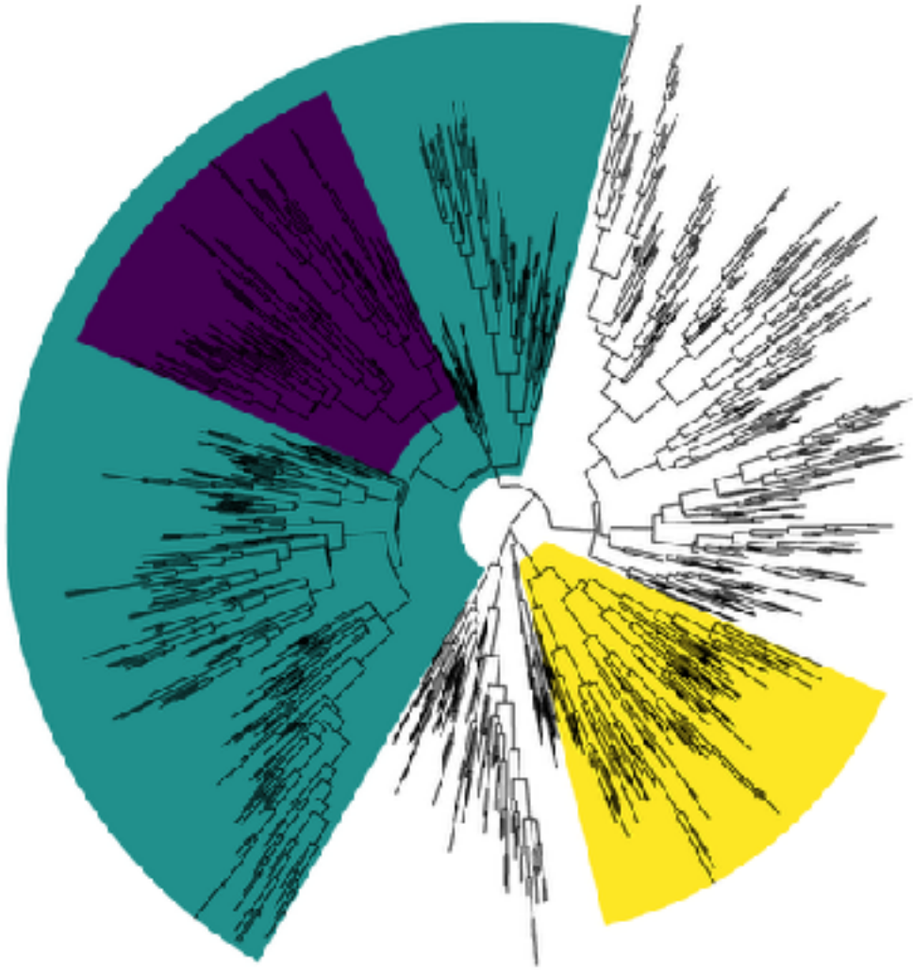
Washburne et al. (accepted) Phylofactorization: a graph partitioning algorithm & … *Ecological Monographs*

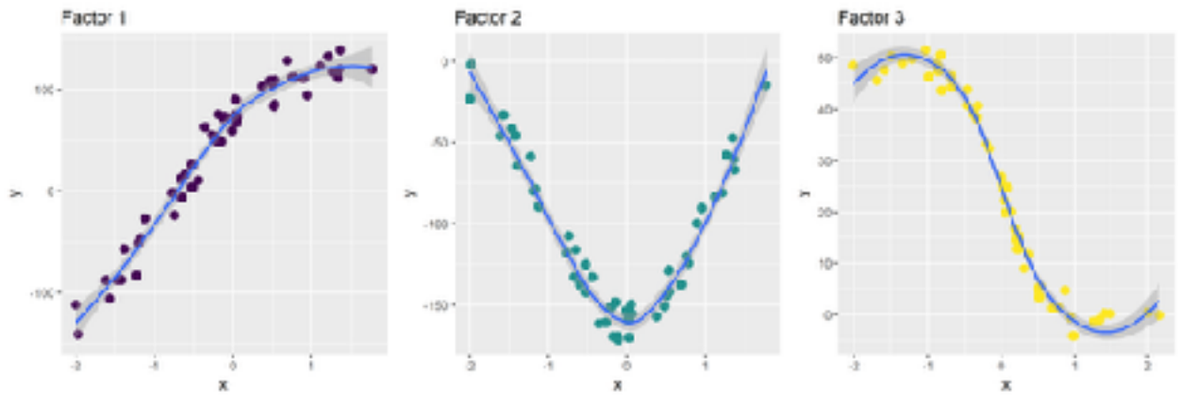
Washburne et al. (2017) *PeerJ*

Theory

PhyloFactorization

ILR balances are **independent** and **~normal:** allow parametric modeling of phylogenetic variables with more complex models

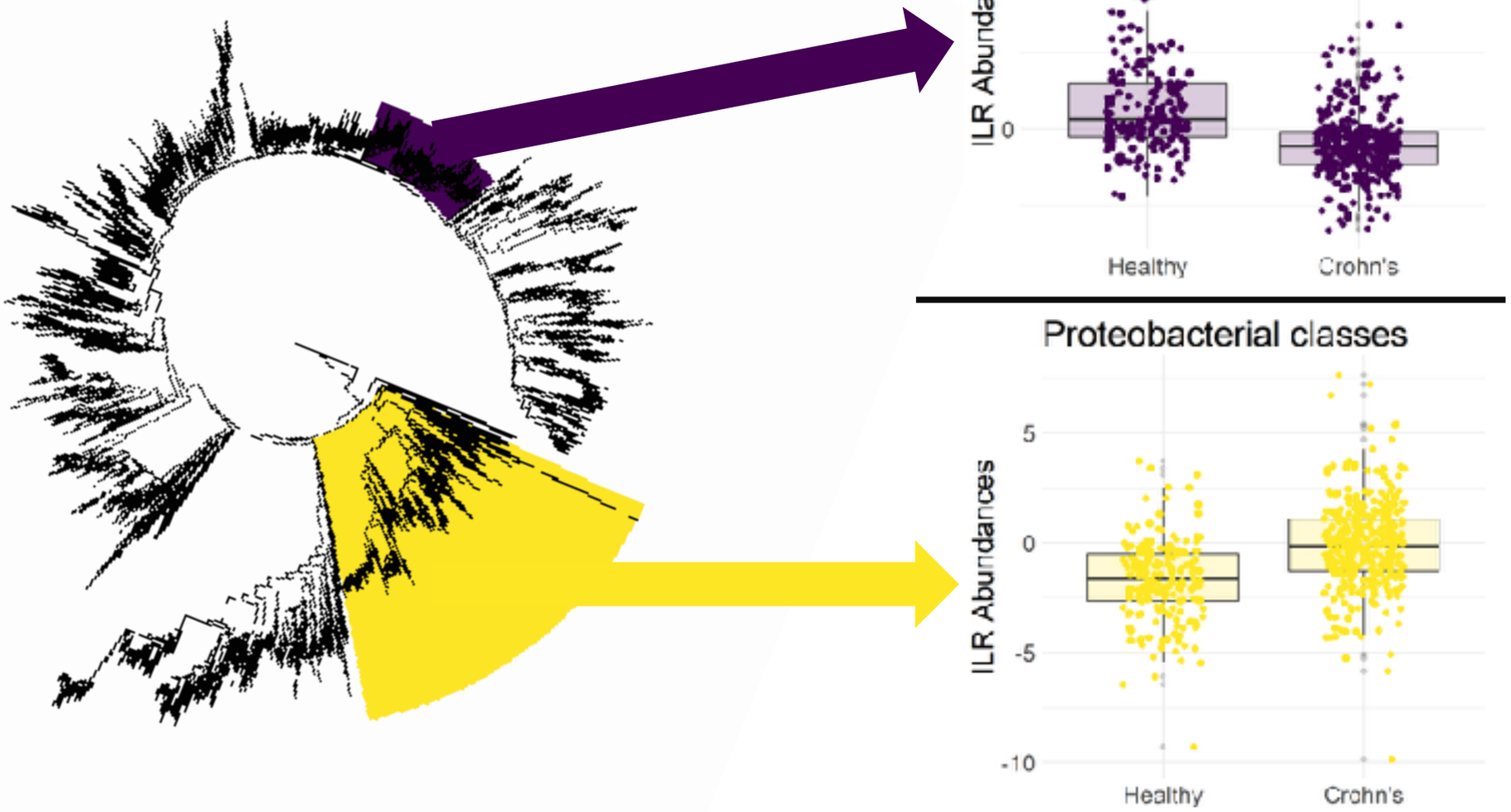




Theory

Example 1: Chron's Disease

`Formula: CD~Data`

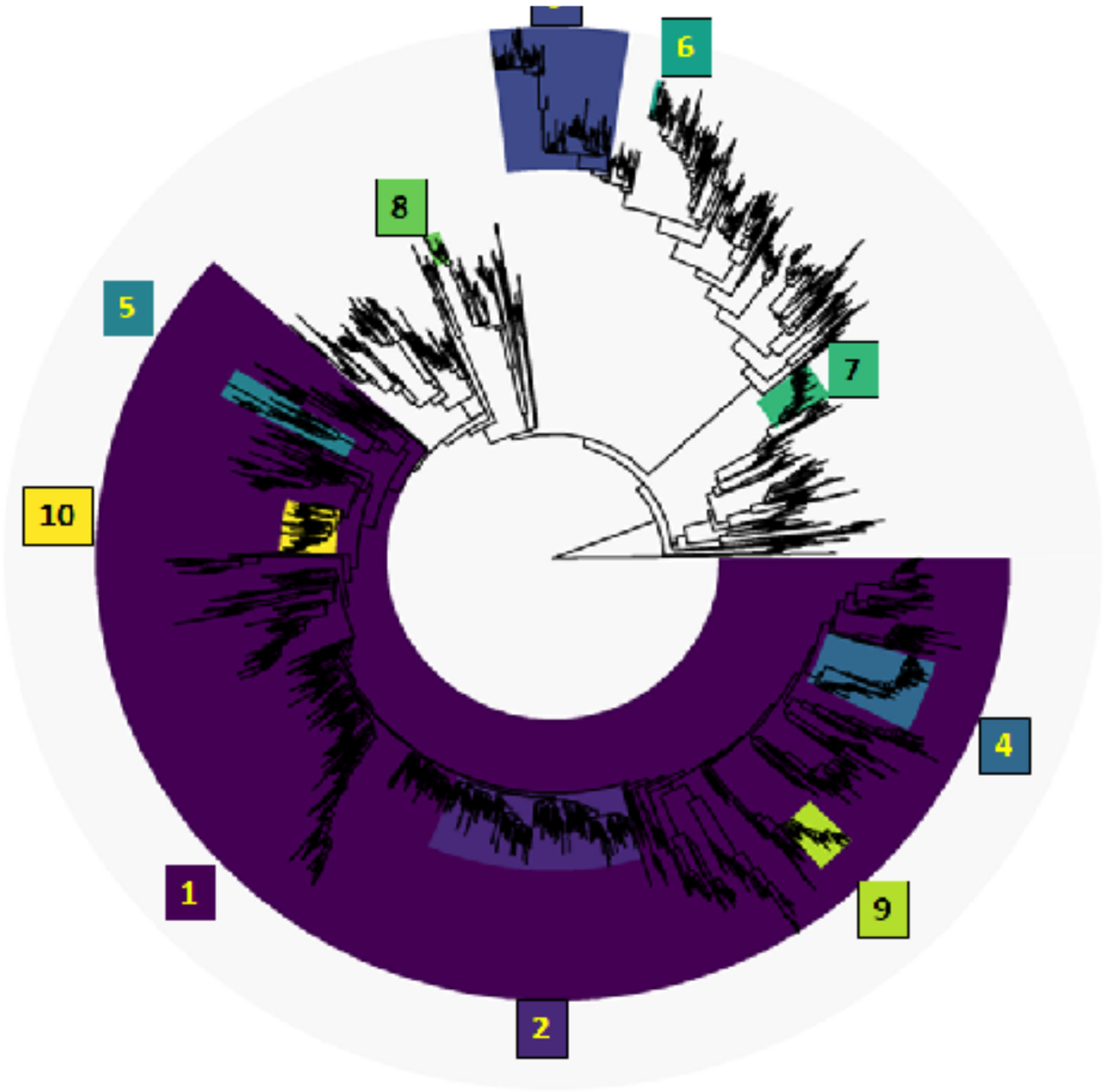


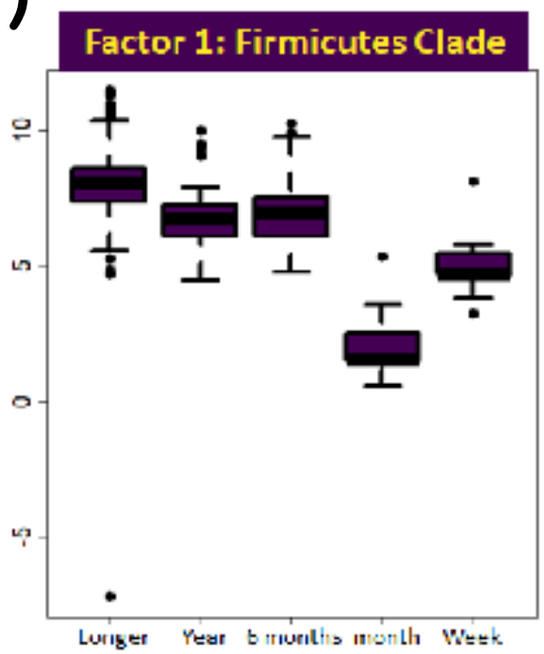
Vasquez Y et al. (2017) Guiding longitudinal sampling in IBD cohorts. *Gut*

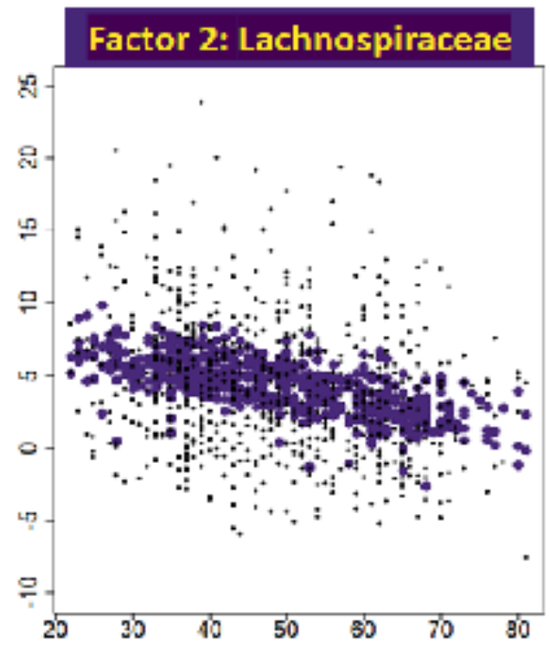
Theory

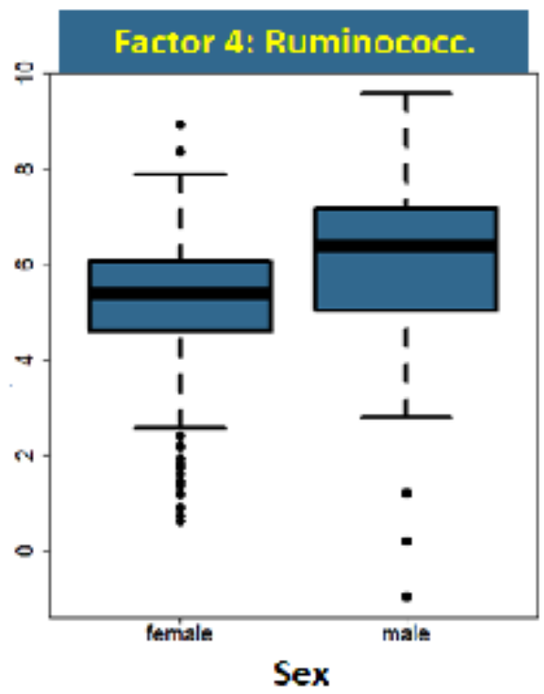
Example 2: American Gut

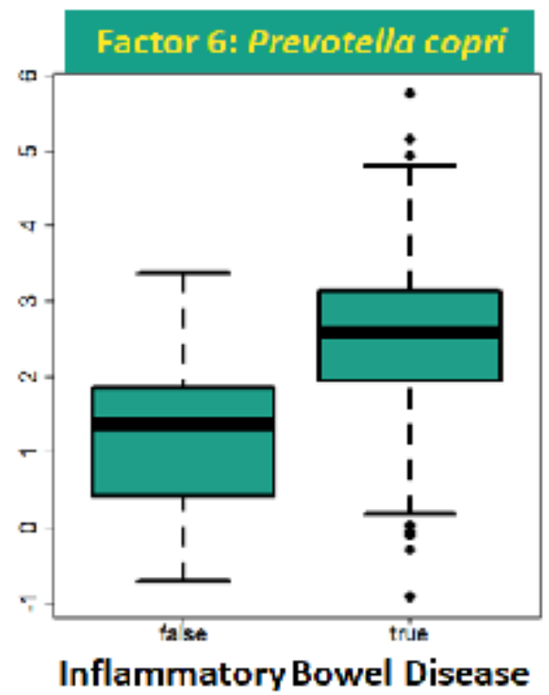
`Formula: var(Data)`







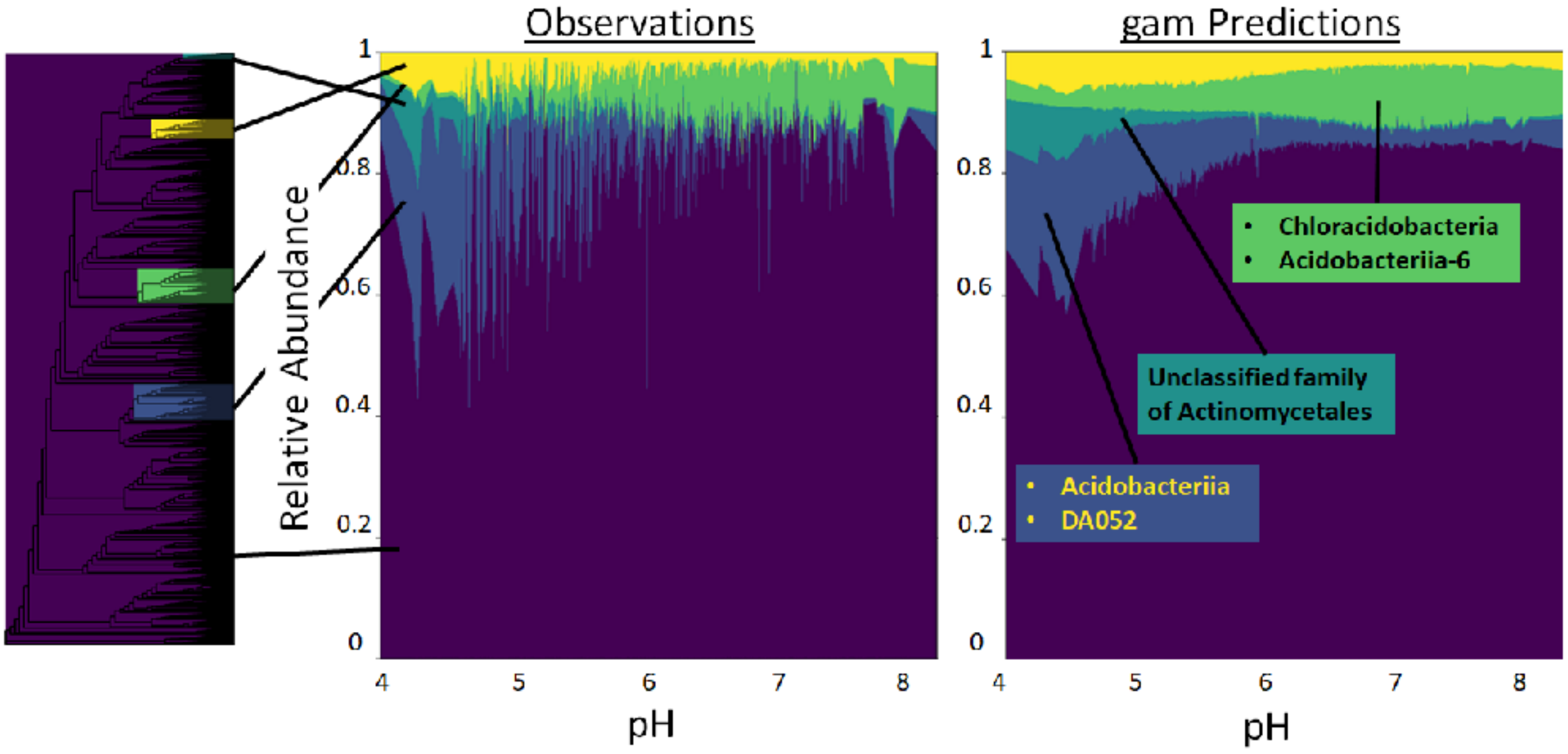




Theory

Example 3: Central Park Soils (gam)

`Formula: Data~s(C)+s(pH)+s(N)`



Code lines 427 — 724

Import data	`> read.table()` `> read.tree()`
Simulating phylo variable	`> rlnorm()`
Testing phylo variable with 2-sample test	`> twoSampleFactor()`
Testing distribution of OTUs	`> phylofactor()`
Summarizing factor taxa	`> pf.taxa()`
Visualizing factors on trees	`> pf.tree()` `> pf.heatmap()`
Predicting data on trees	`> predict()`

4 — PhyloFactor

```
##        phylofactor object from function PhyloFactor
##        --------------------------------------------
## Method                     : glm
## Choice                     : var
## Formula                    : Data ~ Depth_m
## Number of species          : 4315
## Number of factors          : 5
## Frac Explained Variance    : 0.0624
## Largest non-remainder bin : 473
## Number of singletons       : 0
## Paraphyletic Remainder     : 3250 species
##
## ------------------------------------------------------------
## Factor Table:
##                                    Group1                            Group2
## Factor 1 473 member Monophyletic clade 3842 member Monophyletic clade
## Factor 2 141 member Monophyletic clade 3701 member Paraphyletic clade
## Factor 3 337 member Monophyletic clade 3364 member Paraphyletic clade
## Factor 4 114 member Monophyletic clade 3250 member Paraphyletic clade
## Factor 5  38 member Monophyletic clade  435 member Paraphyletic clade
##             ExpVar      F    Pr(>F)
## Factor 1 0.0178330 12.333 0.0170660
## Factor 2 0.0129200 24.596 0.0042502
## Factor 3 0.0136960 45.169 0.0011049
## Factor 4 0.0105830 42.647 0.0012596
## Factor 5 0.0073844 22.799 0.0049935
```

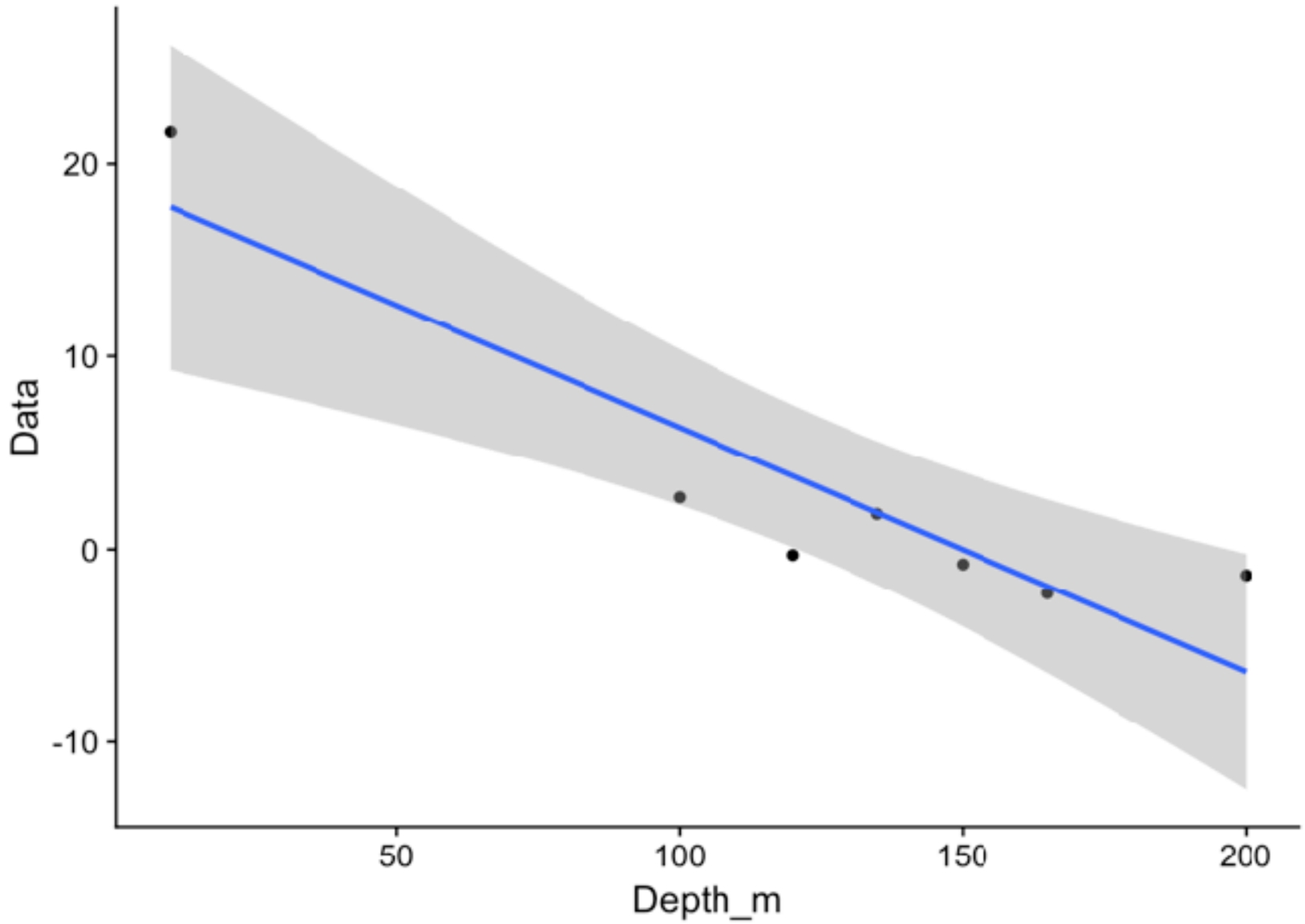
```
##  [1] "Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;PS1_clade"
##  [2]
"Bacteria;Proteobacteria;Alphaproteobacteria;Alphaproteobacteria_unclassified;Alphaproteobacteria_uncla
a_unclassified"
##  [3] "Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiales_unclassified;Rhizobiales_u
##  [4] "Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Rhodobacteraceae_
##  [5] "Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Frigidibacter"
##  [6] "Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Phyllobacteriaceae"
##  [7] "Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Hyphomonadaceae;Hellea"
##  [8] "Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Octadecabacter"
##  [9] "Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Litoreibacter"
## [10] "Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Amylibacter"
```

Practice

4 — PhyloFactor

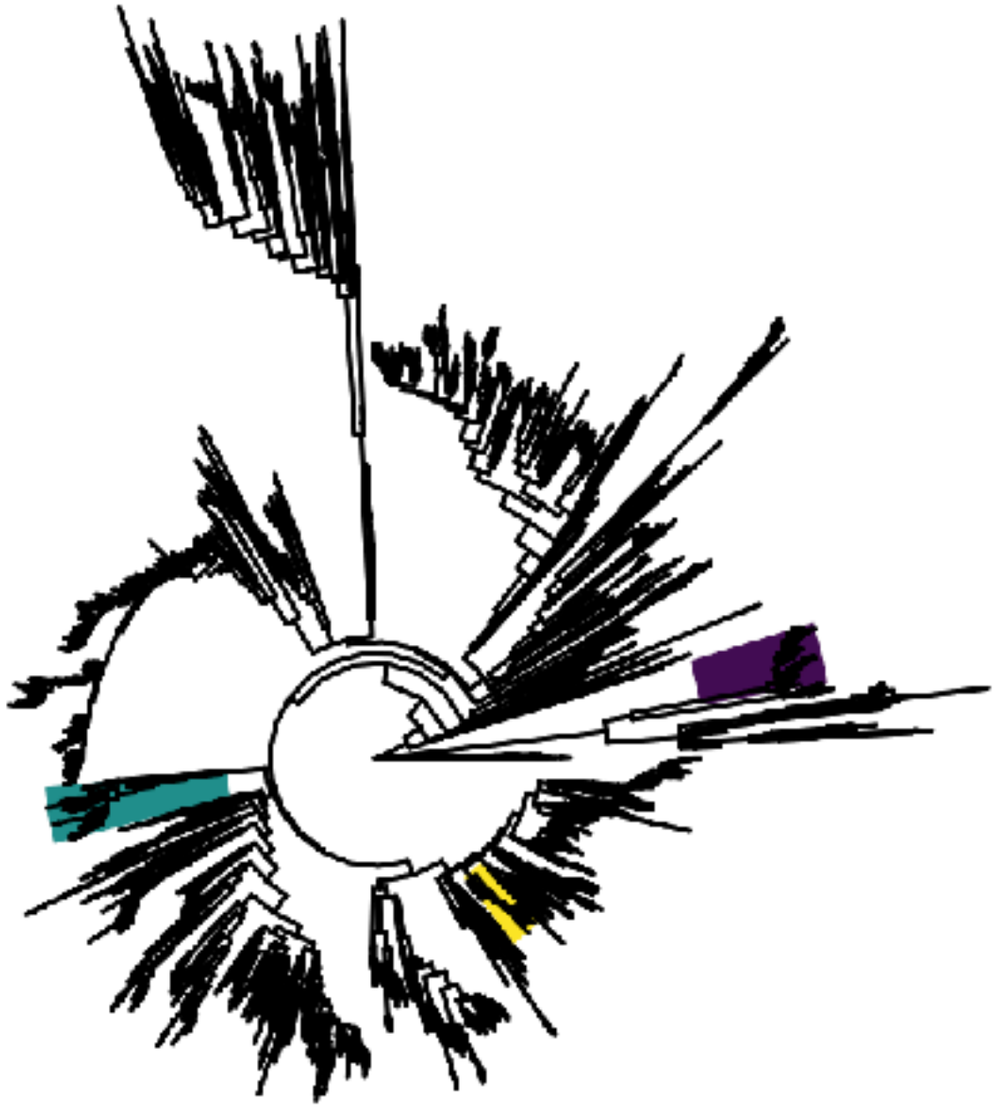
Practice

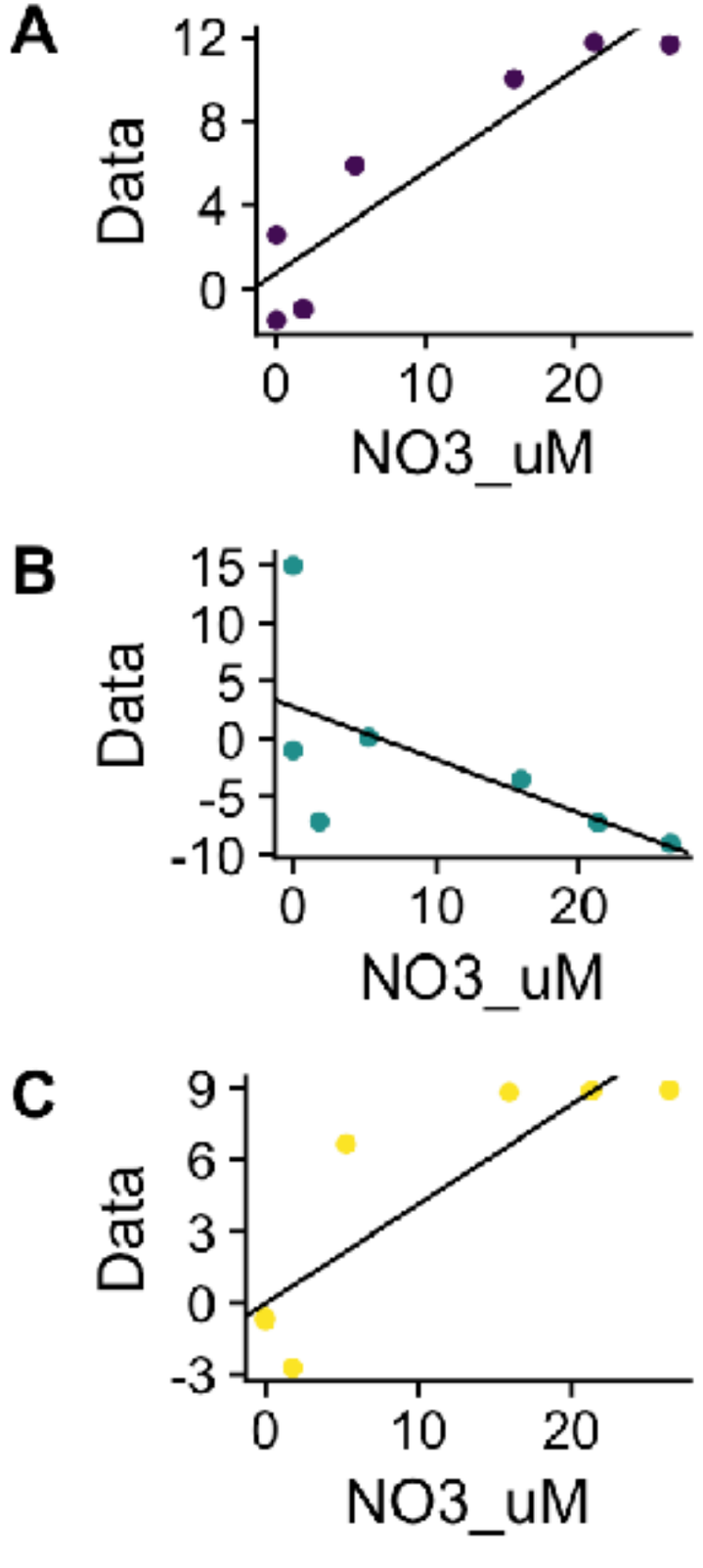
4 — PhyloFactor



4 — PhyloFactor

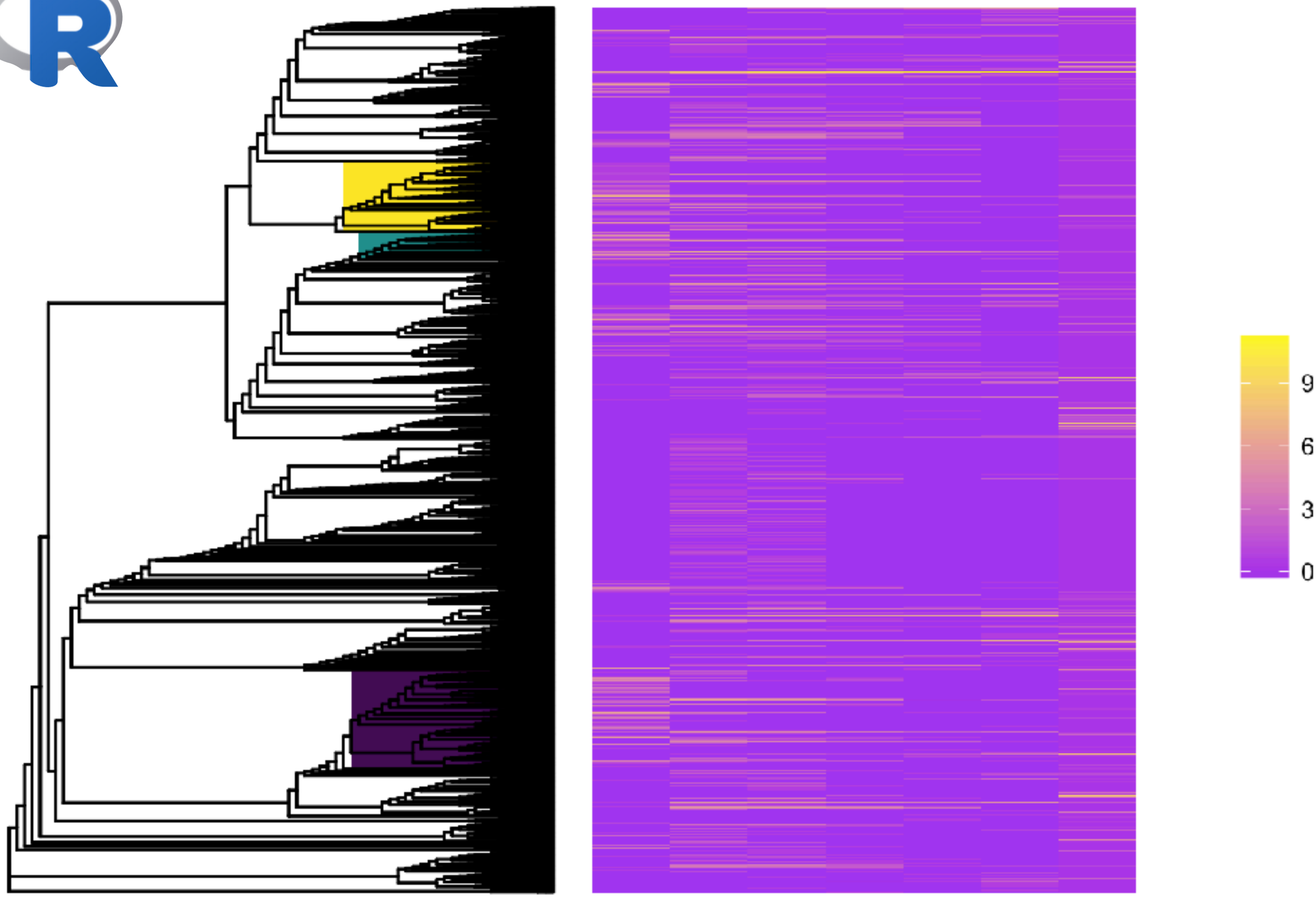
Phylogeny





Practice

4 — PhyloFactor



Practice

4 — PhyloFactor