



Ecological analysis of metabarcoding data

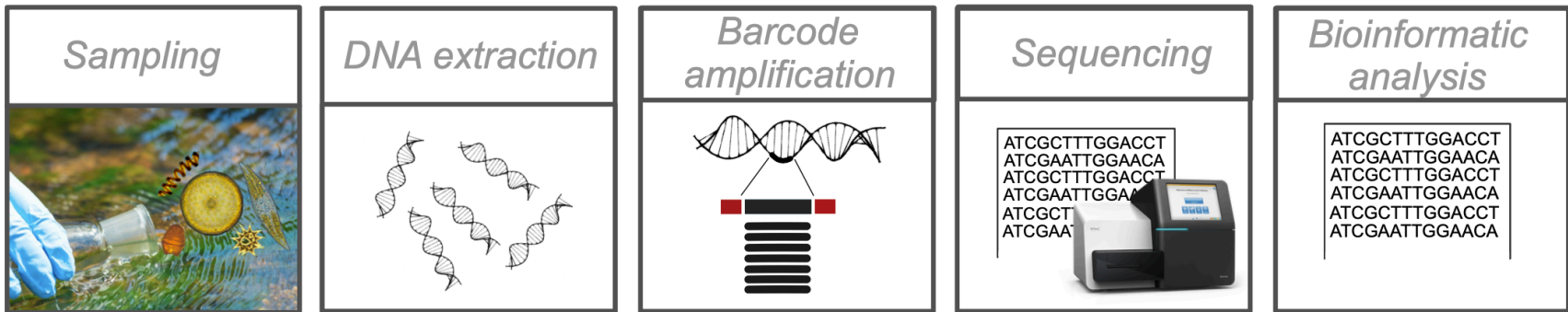
Introduction to R

Clarisse Lemonnier

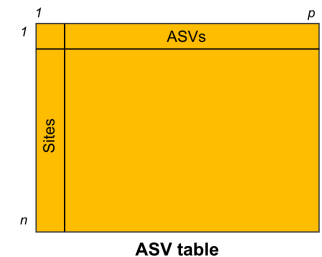


Previously in BIOLAWEB...

Theory on all the steps of the metabarcoding pipeline

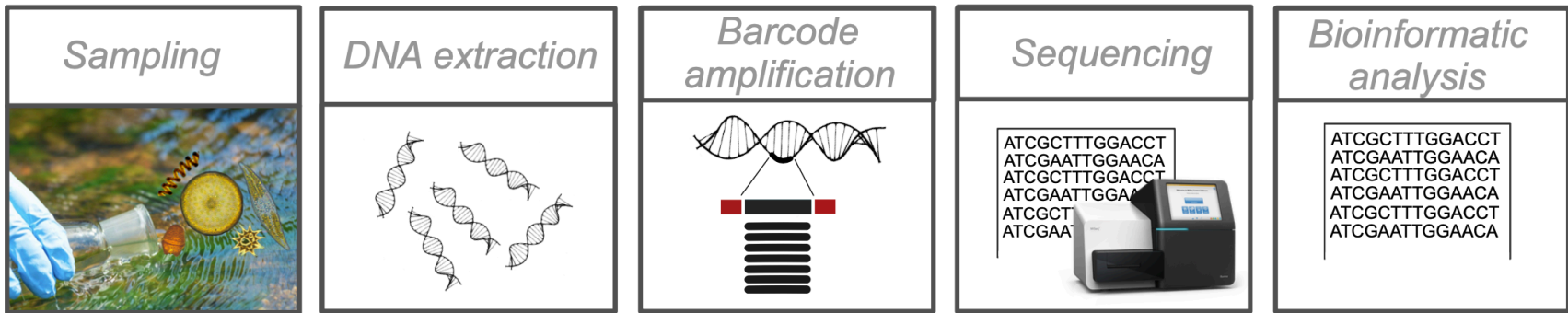


Raw data

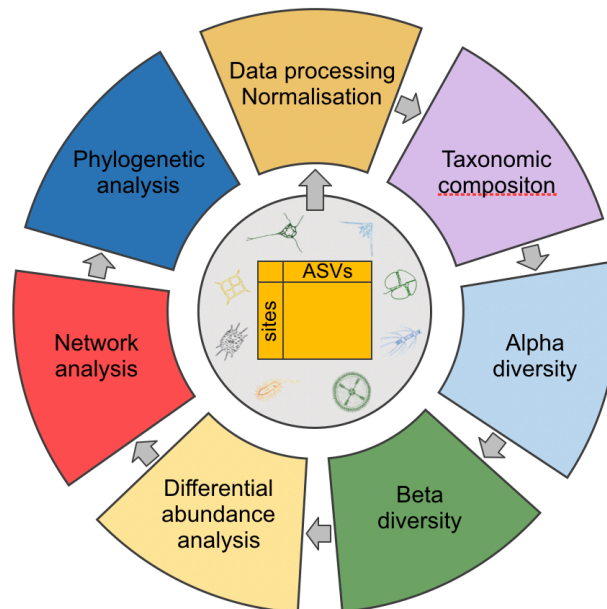


+ taxonomy

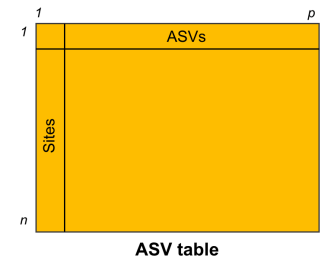
Now : Ecological analysis of metabarcoding data



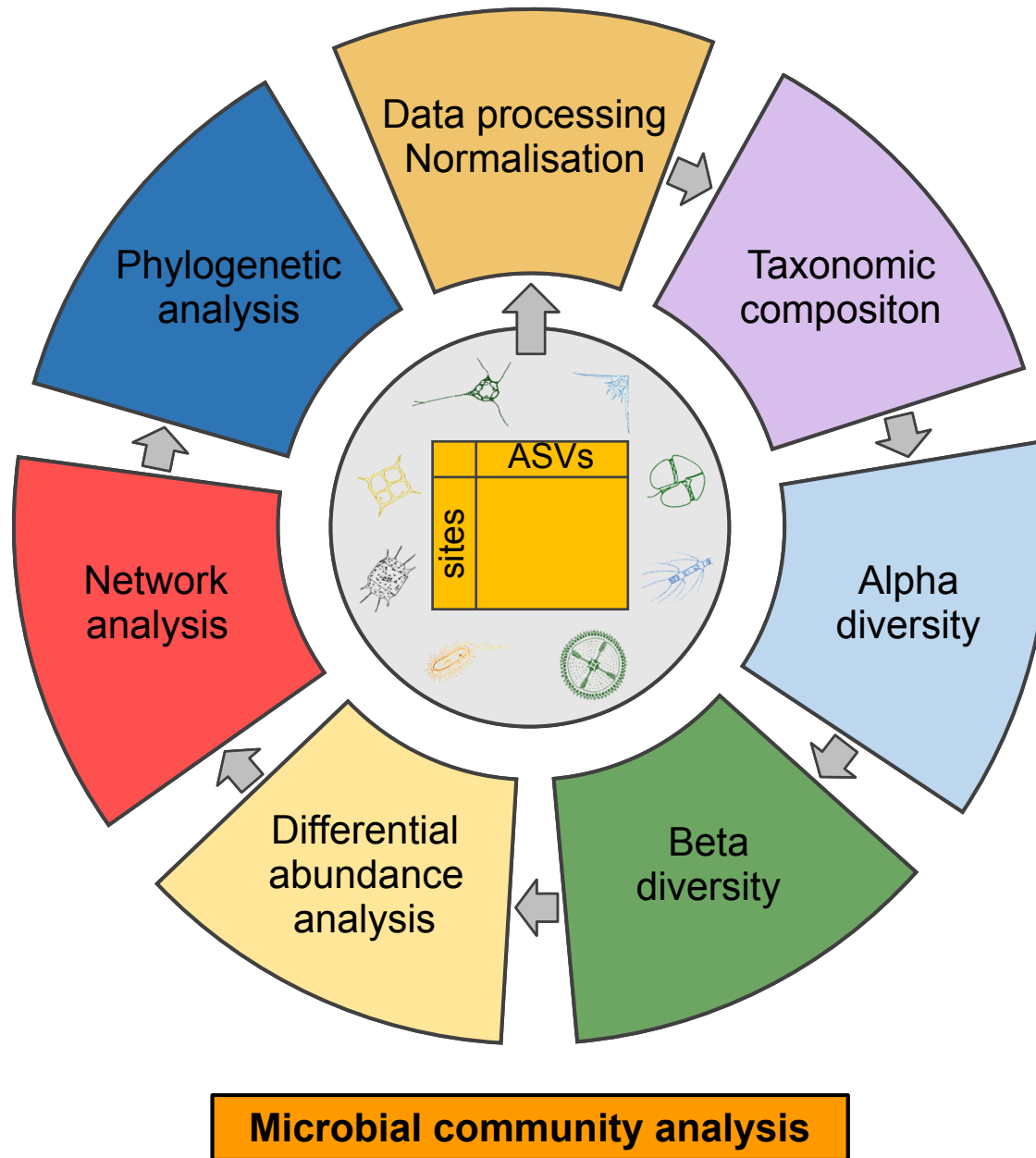
Theory
+
Practical sessions



Raw data



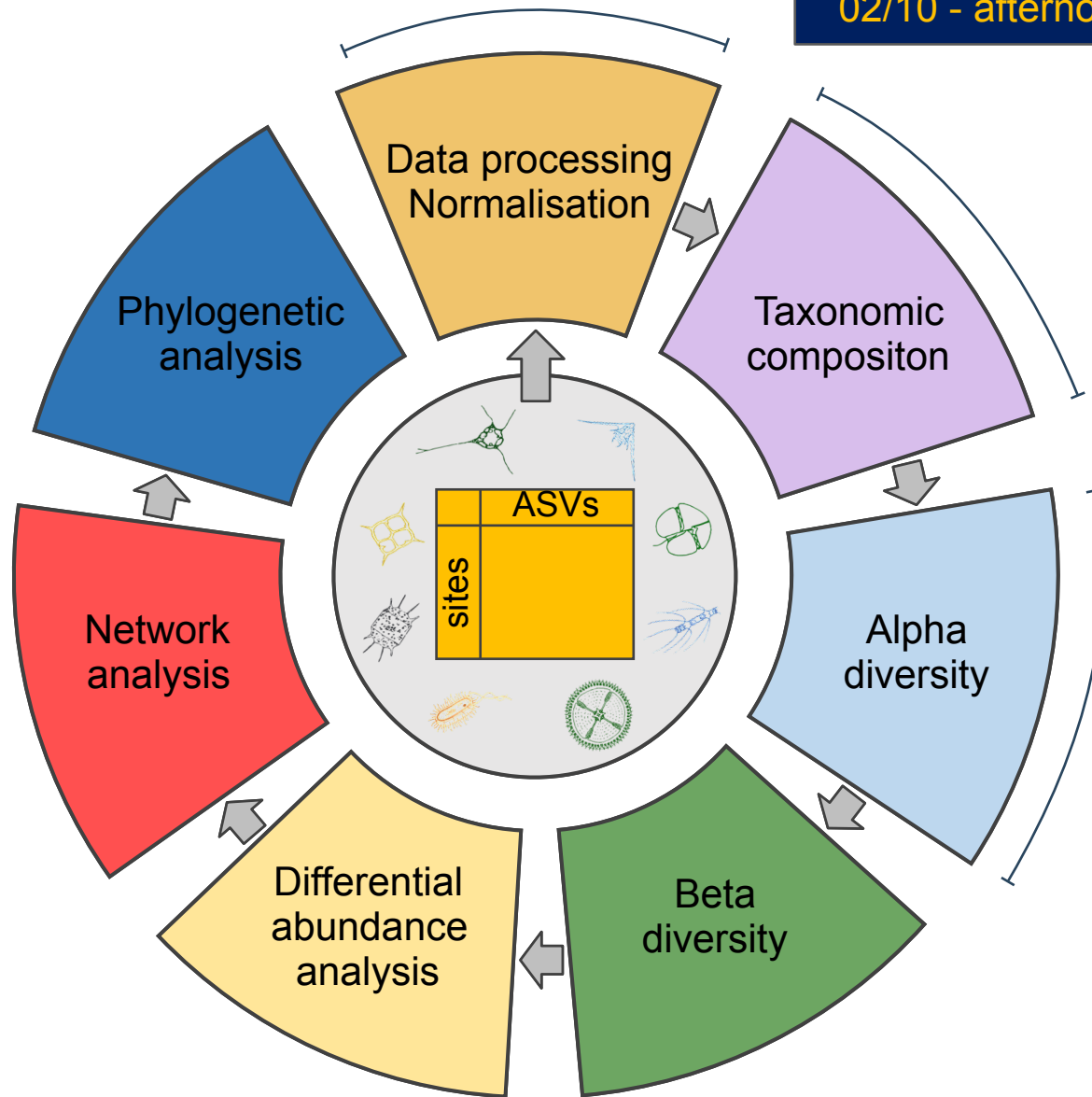
+ taxonomy



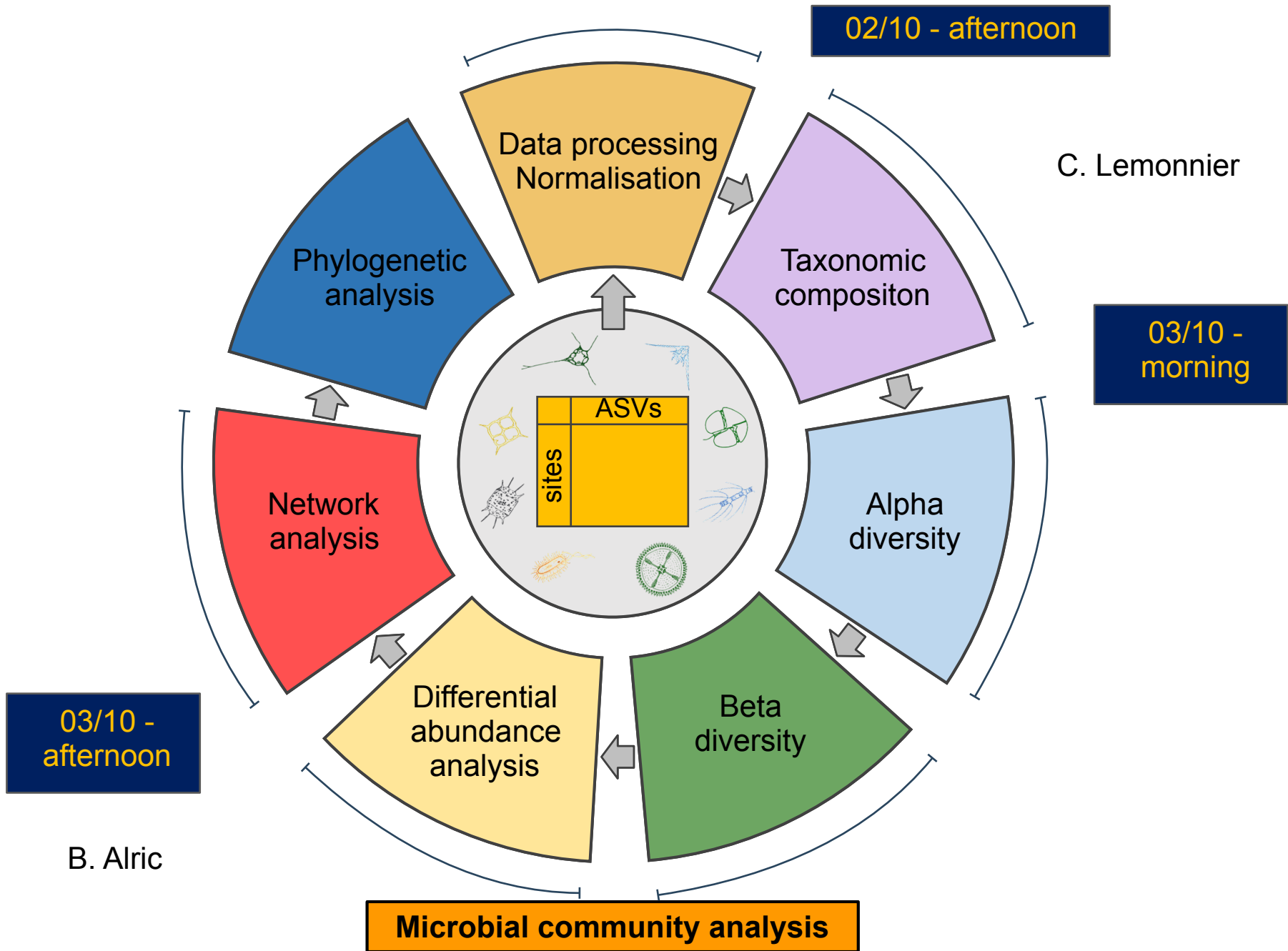
02/10 - afternoon

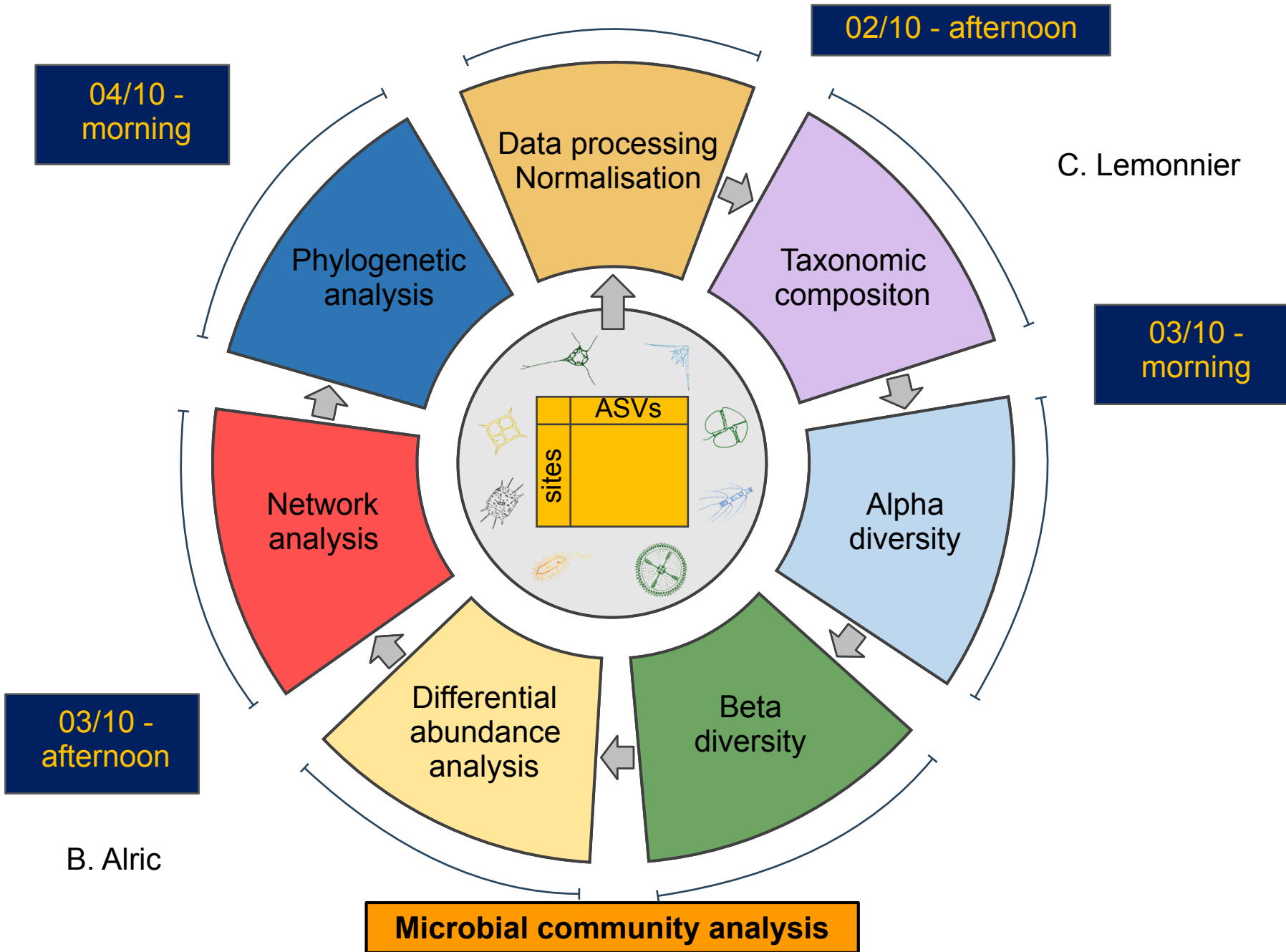
C. Lemonnier

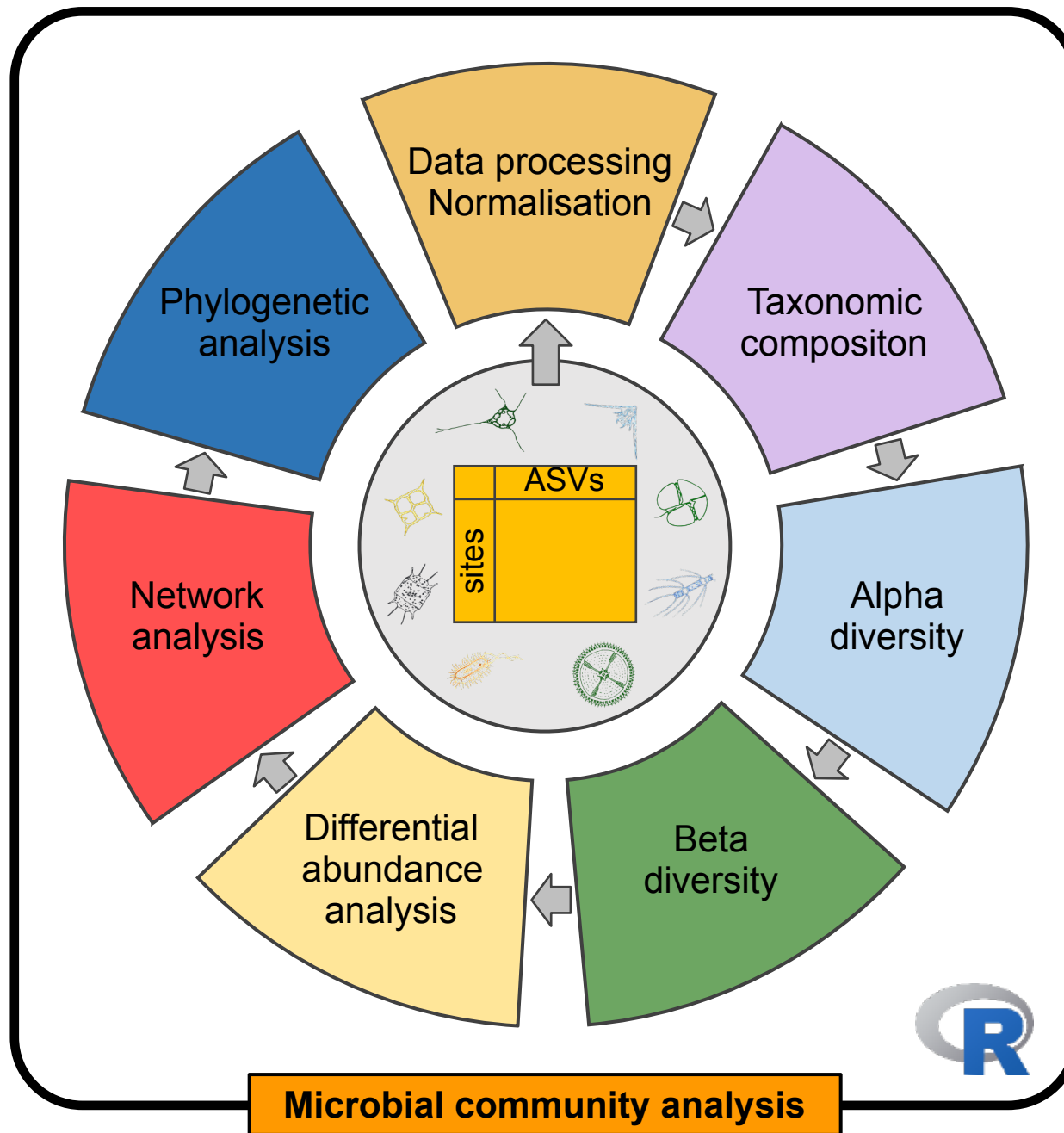
03/10 - morning

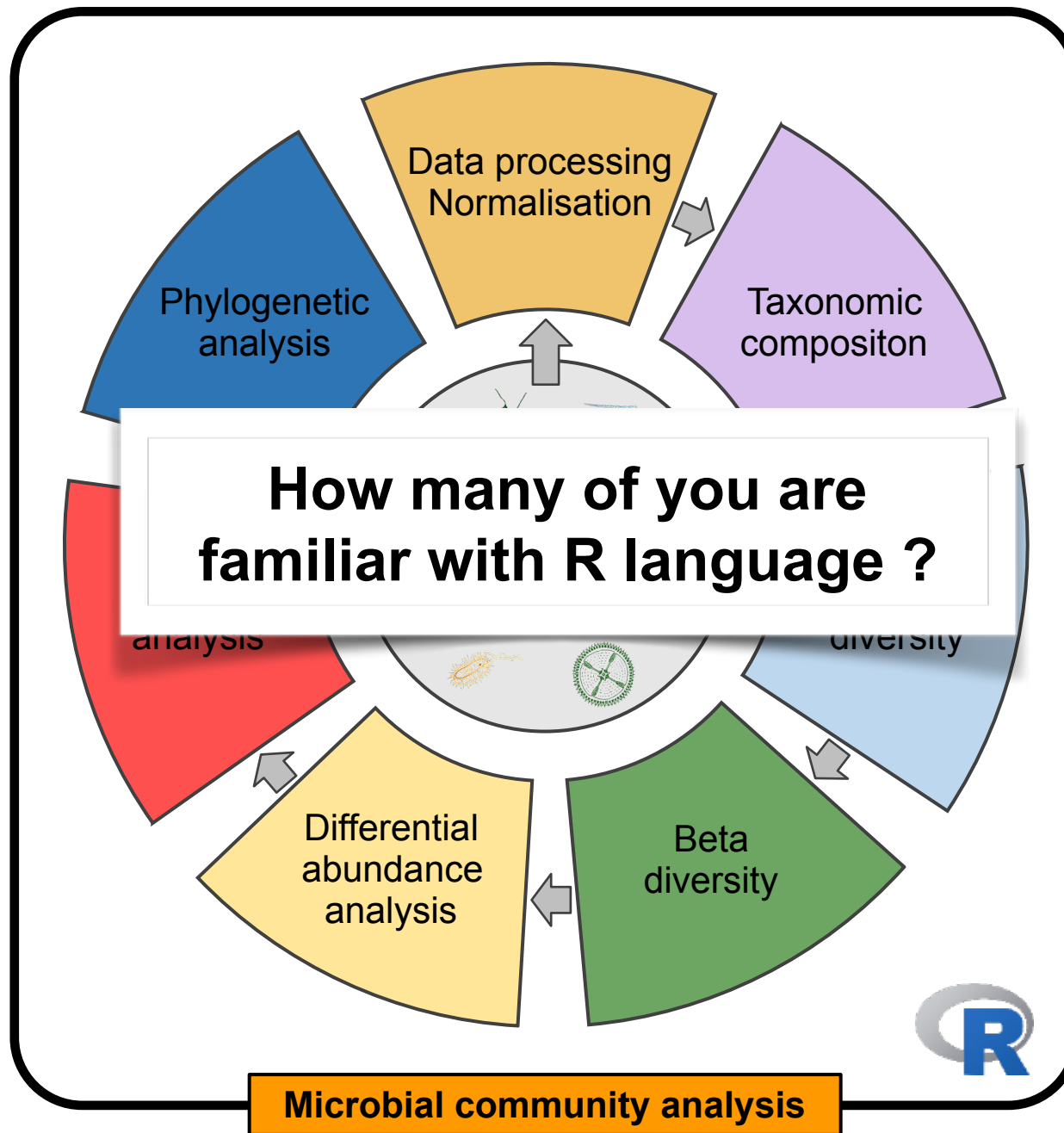


Microbial community analysis









```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
            1          2          3          4          5          6          7          8          9          10
(Intercept)  1.3831    5.2977    1.6105    0.4494    1.8797    0.2456    0.1248    0.0518    0.0116    0.0005
x1          5.2977    1.6105    0.4494    1.8797    0.2456    0.1248    0.0518    0.0116    0.0005
x2          1.6105    0.4494    1.8797    0.2456    0.1248    0.0518    0.0116    0.0005
x3          0.4494    1.8797    0.2456    0.1248    0.0518    0.0116    0.0005
groupB      1.8797    0.2456    0.1248    0.0518    0.0116    0.0005

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

What is R ?

- R is an **open source** programming language

- It was designed for **statistical computing, data analysis** and **graphic display**

- It works in all operating system



- It is widely used in academic research

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
(Intercept)  1.3831  0.24566  0.2027  0.0000  0.0000
          x1    5.29776  0.12483  0.0000  0.0000  0.0000
          x2    1.61055  0.0518  0.0000  0.0000  0.0000
          x3    0.44947  0.0000  0.0000  0.0000  0.0000
        groupB  1.87978  0.1000  0.0000  0.0000  0.0000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

What can we do with R?

Analytics

Basic mathematics

Statistical tests

Big data analysis

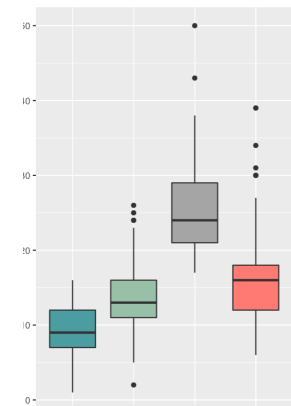
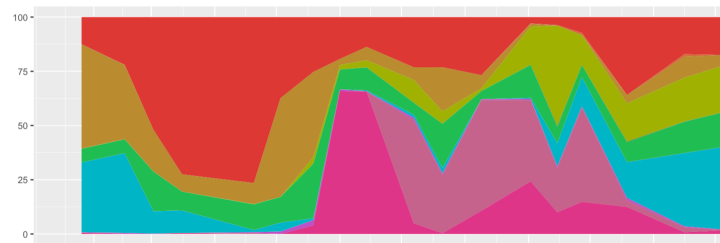
Statistical modeling

Machine learning

Graphics and visualisation

Static graphics

Cartography



```
Call:
lm(formula = y ~ ., data = data)

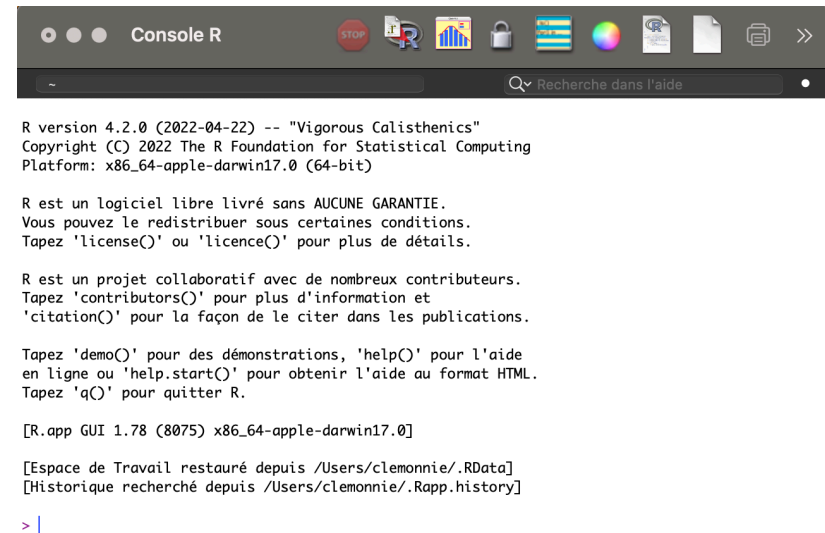
Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
(Intercept)  1.3831    2.4566    0.2483
x1           5.2977    1.2483    0.0000
x2           1.6105    0.5518    0.0000
x3           0.4494    0.0000    0.0000
groupB       1.8797    0.1600    0.0000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

R console

R is used through a command line console



```
Console R
~
Recherche dans l'aide

R version 4.2.0 (2022-04-22) -- "Vigorous Calisthenics"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

[R.app GUI 1.78 (8075) x86_64-apple-darwin17.0]

[Espace de Travail restauré depuis /Users/clemennie/.RData]
[Historique recherché depuis /Users/clemennie/.Rapp.history]

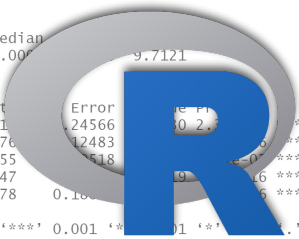
> |
```

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.38311    0.24566   5.634 0.000001**
x1           5.29776    0.12483  42.441 <2e-16***
x2           1.61055    0.05518  29.187 <2e-16***
x3           0.44947    0.01609  27.936 <2e-16***
groupB      1.87978    0.16005  11.745 <2e-16***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```



R console

R is used through a command line console

Set to the working directory

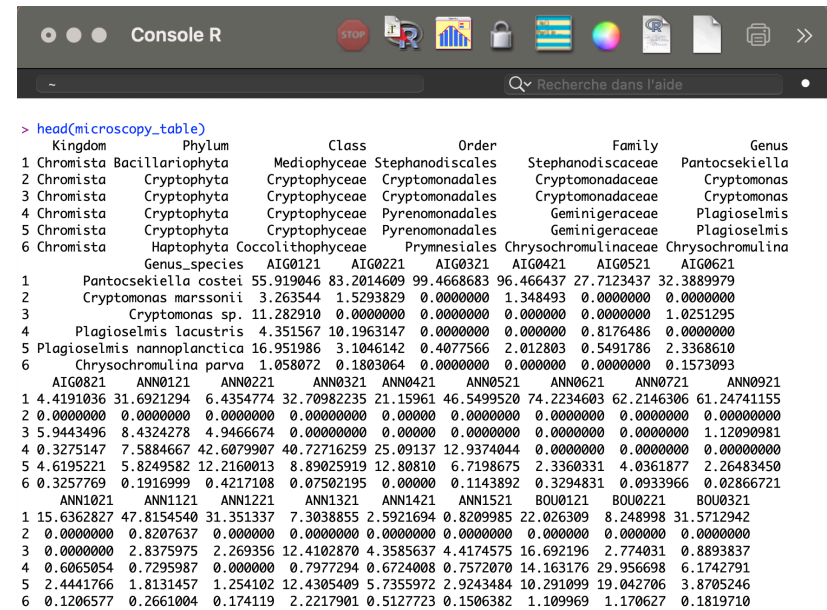
```
> setwd(dir=~/Desktop/R_Introduction/)
```

Load a table

```
> Microscopy_table <- read.table("Microscopy_table.csv", header=T, sep=";")
```

Visualize the table

```
> head(microscopy_table)
```

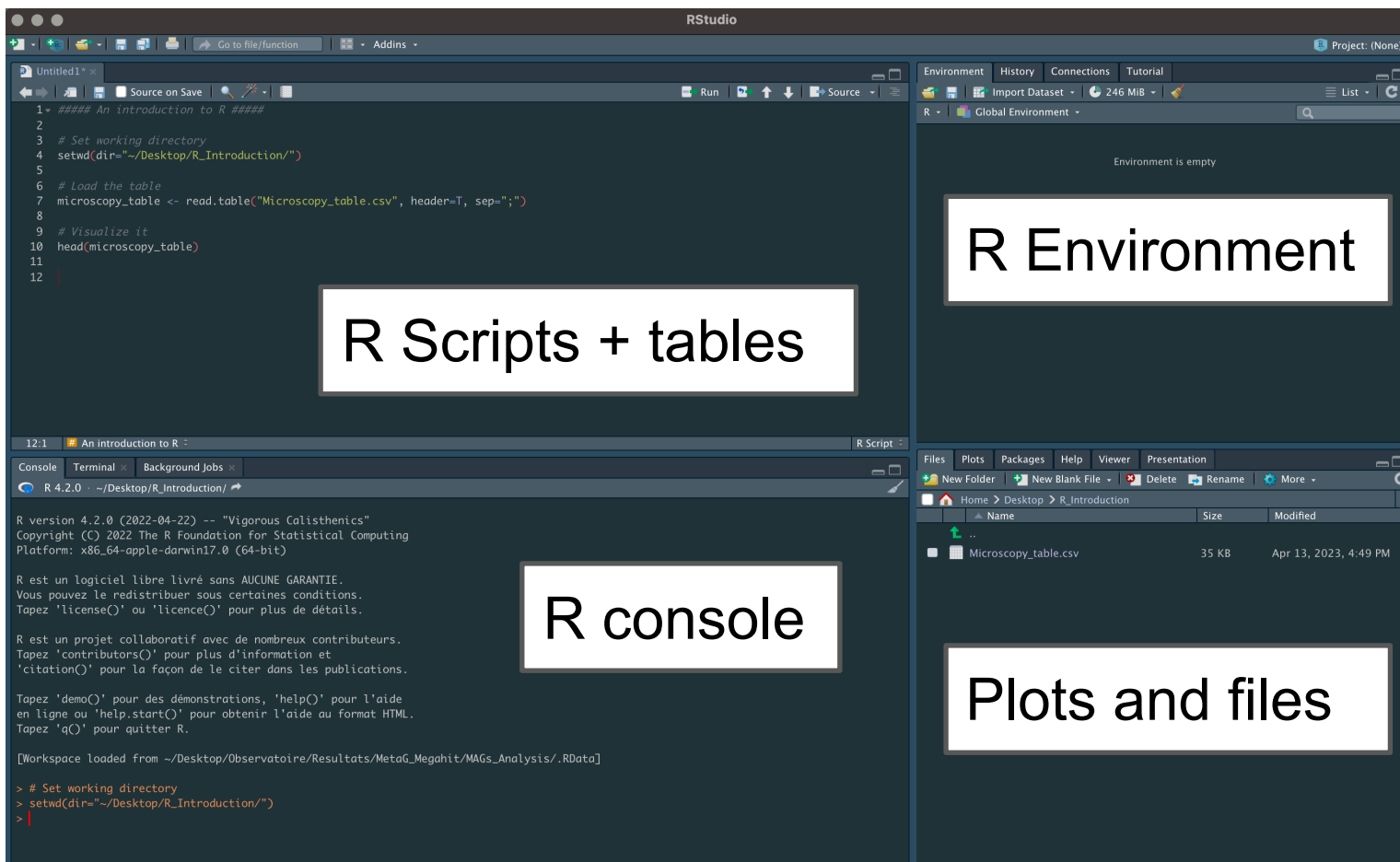


```
> head(microscopy_table)
```

	Kingdom	Phylum	Class	Order	Family	Genus
1	Chromista	Bacillariophyta	Mediophyceae	Stephanodiscales	Stephanodiscaeae	Pantocsekiella
2	Chromista	Cryptophyta	Cryptophyceae	Cryptomonadales	Cryptomonadaceae	Cryptomonas
3	Chromista	Cryptophyta	Cryptophyceae	Cryptomonadales	Cryptomonadaceae	Cryptomonas
4	Chromista	Cryptophyta	Cryptophyceae	Pyrenomonadales	Geminigeraceae	Plagioselmis
5	Chromista	Cryptophyta	Cryptophyceae	Pyrenomonadales	Geminigeraceae	Plagioselmis
6	Chromista	Haptophyta	Coccolithophyceae	Prymnesiales	Chrysochromulinaceae	Chrysochromulina



Is an integrated development environment (IDE)



The screenshot shows the RStudio IDE interface with four callout boxes highlighting key components:

- R Scripts + tables**: Points to the source editor showing R code for reading and visualizing a CSV file.
- R Environment**: Points to the Environment pane showing the Global Environment.
- R console**: Points to the Console pane showing the R version and command prompt.
- Plots and files**: Points to the Files pane showing the file explorer.

```
1 ##### An introduction to R #####
2
3 # Set working directory
4 setwd(dir=~ /Desktop/R_Introduction/")
5
6 # Load the table
7 microscopy_table <- read.table("Microscopy_table.csv", header=T, sep=";")
8
9 # Visualize it
10 head(microscopy_table)
11
12 |
```

R version 4.2.0 (2022-04-22) -- "Vigorous Calisthenics"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

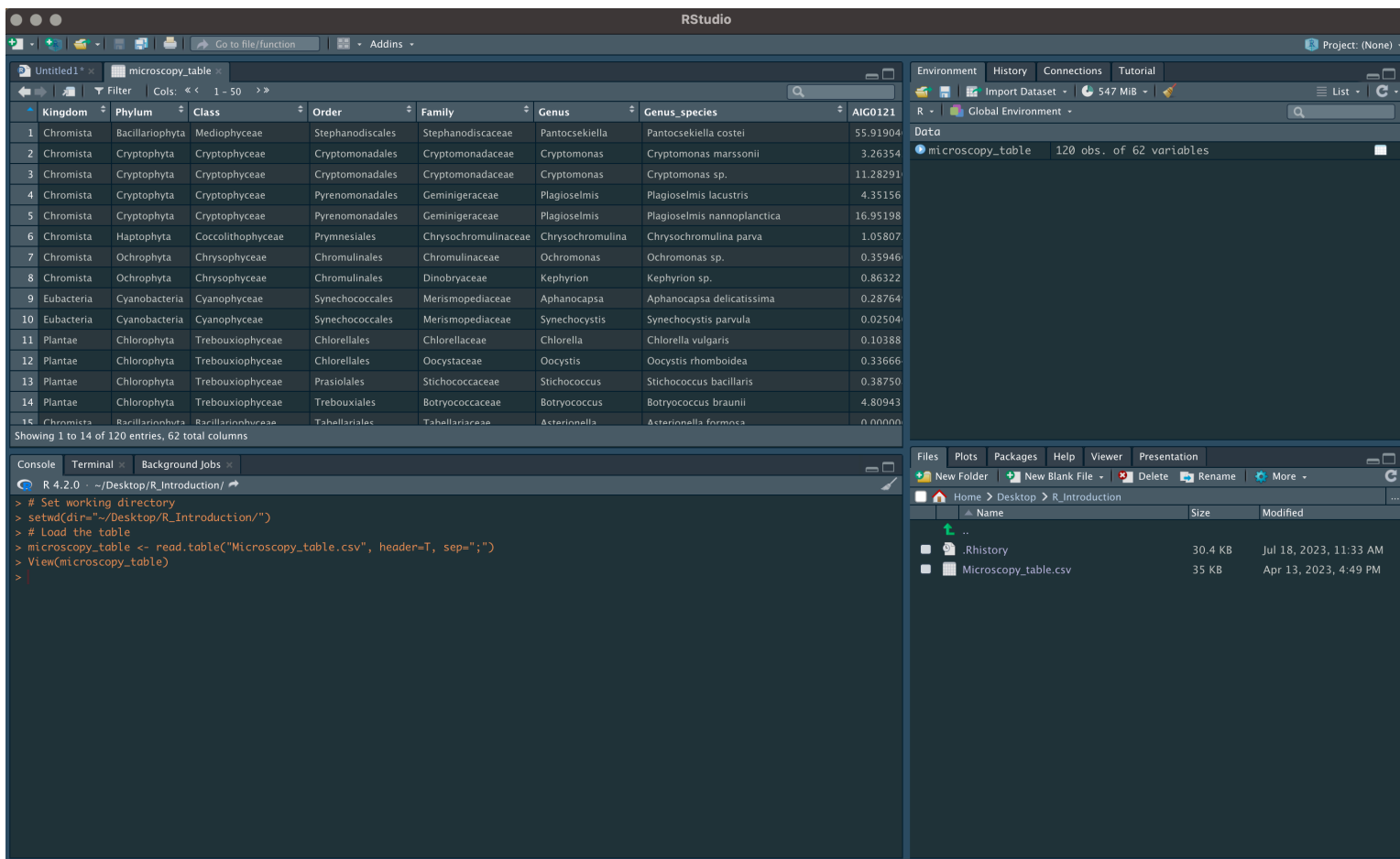
[Workspace loaded from ~/Desktop/Observatoire/Resultats/MetaG_Megahit/MAGs_Analysis/.RData]

```
> # Set working directory
> setwd(dir=~ /Desktop/R_Introduction/")
> |
```

Name	Size	Modified
..		
Microscopy_table.csv	35 KB	Apr 13, 2023, 4:49 PM



Is an integrated development environment (IDE)



The screenshot displays the RStudio IDE interface. The main window shows a data table with columns for taxonomic classification and abundance. The console window at the bottom left shows the R code used to load the data. The file explorer at the bottom right shows the current directory structure.

Kingdom	Phylum	Class	Order	Family	Genus	Genus_species	AIG0121	
1	Chromista	Bacillariophyta	Mediophyceae	Stephanodiscales	Stephanodisceae	Pantocsekiella	Pantocsekiella costei	55.91904
2	Chromista	Cryptophyta	Cryptophyceae	Cryptomonadales	Cryptomonadaceae	Cryptomonas	Cryptomonas marssonii	3.26354
3	Chromista	Cryptophyta	Cryptophyceae	Cryptomonadales	Cryptomonadaceae	Cryptomonas	Cryptomonas sp.	11.28291
4	Chromista	Cryptophyta	Cryptophyceae	Pyrenomonadales	Geminigeraceae	Plagioselmis	Plagioselmis lacustris	4.35156
5	Chromista	Cryptophyta	Cryptophyceae	Pyrenomonadales	Geminigeraceae	Plagioselmis	Plagioselmis nannoplanctica	16.95198
6	Chromista	Haptophyta	Coccolithophyceae	Prymniales	Chrysochromulinaceae	Chrysochromulina	Chrysochromulina parva	1.05807
7	Chromista	Ochrophyta	Chrysophyceae	Chromulinales	Chromulinaceae	Ochromonas	Ochromonas sp.	0.35946
8	Chromista	Ochrophyta	Chrysophyceae	Chromulinales	Dinobryaceae	Kephyrion	Kephyrion sp.	0.86322
9	Eubacteria	Cyanobacteria	Cyanophyceae	Synechococcales	Merismopediaceae	Aphanocapsa	Aphanocapsa delicatissima	0.28764
10	Eubacteria	Cyanobacteria	Cyanophyceae	Synechococcales	Merismopediaceae	Synechocystis	Synechocystis parvula	0.02504
11	Plantae	Chlorophyta	Trebouxiophyceae	Chlorellales	Chlorellaceae	Chlorella	Chlorella vulgaris	0.10388
12	Plantae	Chlorophyta	Trebouxiophyceae	Chlorellales	Oocystaceae	Oocystis	Oocystis rhomboidea	0.33666
13	Plantae	Chlorophyta	Trebouxiophyceae	Prasiolales	Stichococcaceae	Stichococcus	Stichococcus bacillaris	0.38750
14	Plantae	Chlorophyta	Trebouxiophyceae	Trebouxiales	Botryococcaceae	Botryococcus	Botryococcus braunii	4.80943
15	Chromista	Bacillariophyta	Bacillariophyceae	Tabellariales	Tabellariaceae	Asterionella	Asterionella formosa	0.00000

```

R 4.2.0 ~./Desktop/R_Introduction/
> # Set working directory
> setwd(dir=~./Desktop/R_Introduction/")
> # Load the table
> microscopy_table <- read.table("Microscopy_table.csv", header=T, sep=";")
> View(microscopy_table)
>
  
```

Files Explorer: Home > Desktop > R_Introduction

Name	Size	Modified
.Rhistory	30.4 KB	Jul 18, 2023, 11:33 AM
Microscopy_table.csv	35 KB	Apr 13, 2023, 4:49 PM

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
            1          2          3          4          5          6          7
(Intercept)  1.3831    5.2977    1.6105    0.4494    1.8797    0.1600    0.1600
x1           1.2456    1.2483    1.2483    1.2483    1.2483    1.2483    1.2483
x2           1.2456    1.2483    1.2483    1.2483    1.2483    1.2483    1.2483
x3           1.2456    1.2483    1.2483    1.2483    1.2483    1.2483    1.2483
groupB      1.2456    1.2483    1.2483    1.2483    1.2483    1.2483    1.2483
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

R Language

Command line example

```
> ASV_table <- read.table("ASV_table.txt", header=T, row.names=1, sep="\t")
```



```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
            1          2          3          4          5          6          7
(Intercept)  1.3831    2.4566    3.0248    4.1248    5.2977    6.1611    7.1611
x1           5.2977    6.1611    7.1611    8.1611    9.1611   10.1611   11.1611
x2           1.61055   2.4566    3.0248    4.1248    5.2977    6.1611    7.1611
x3           0.44947   1.61055   2.4566    3.0248    4.1248    5.2977    6.1611
groupB       1.87978   0.1611    0.1611    0.1611    0.1611    0.1611    0.1611
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

R Language

Command line example

```
> ASV_table <- read.table("ASV_table.txt", header=T, row.names=1, sep="\t")
```

function

arguments

A function() is made to perform a specific task.

It might works with arguments

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.38311    0.24566   5.635 0.000001**
x1           5.29776    0.12483  42.483 <2e-16***
x2           1.61055    0.05518  29.196 <2e-16***
x3           0.44947    0.01609  28.006 <2e-16***
groupB       1.87978    0.16005  11.746 5.5e-15***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

R Language

Command line example

```
> ASV_table <- read.table("ASV_table.txt", header=T, row.names=1, sep="\t")
```

↓
variable

↘
function

↘ ↘ ↘
arguments

A variable in R is the memory allocated to the stockage of a specific **object**

It is assigned to an object using either **=** or **<-**

A function() is made to perform a specific task.

It might works with arguments

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

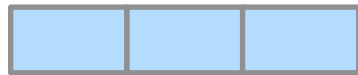
Coefficients:
            Estimate      Error Pr
(Intercept)  1.3831      0.24566  0.2
x1           5.29776      0.12483  0.0
x2           1.61055      0.0518  0.0
x3           0.44947      0.016  0.0
groupB       1.87978      0.16  0.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

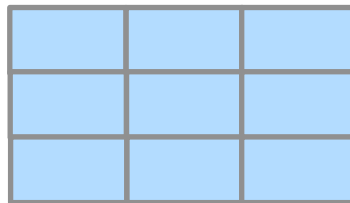


R Language

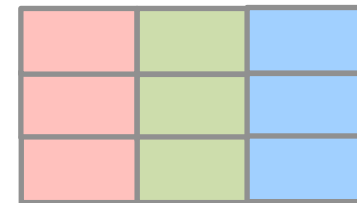
Different type of objects in R



Vector



Matrix



Data frame

```
[1] "Monday"  "Tuesday"
[3] "Wednesday" "Thursday"
[5] "Friday"  "Saturday"
[7] "Sunday"
```

	ANN0121	ANN0221
ASV1	3317	3583
ASV2	1040	359
ASV3	0	9
ASV4	673	509
ASV5	2698	3342

	Month	Date	day
ANN0121	January	19/01/21	19
ANN0221	February	09/02/21	40
ANN0321	March	09/03/21	68
ANN0421	April	01/04/21	91
ANN0521	April	15/04/21	105

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

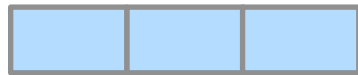
Coefficients:
            Estimate      Error    Pr(>|t|)
(Intercept)  1.3831      2.4566    0.276
x1           5.2977      1.2483    0.000
x2           1.6105      1.518    0.000
x3           0.4494      1.6    0.000
groupB       1.8797      0.16    0.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

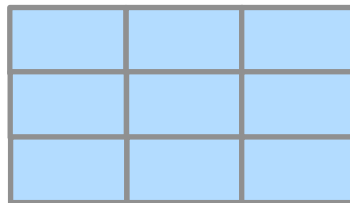


R Language

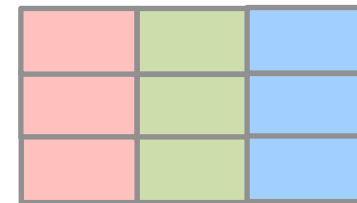
Different type of objects in R



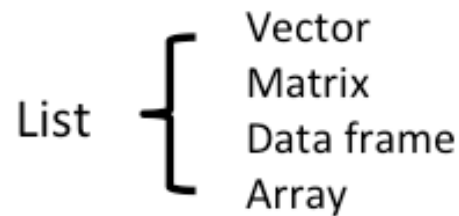
Vector



Matrix



Data frame



```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
(Intercept)  1.3831  0.24566  0.2483  0.16518  0.16518
x1           5.29776  0.12483  0.12483  0.12483  0.12483
x2           1.61055  0.16518  0.16518  0.16518  0.16518
x3           0.44947  0.16518  0.16518  0.16518  0.16518
groupB       1.87978  0.16518  0.16518  0.16518  0.16518
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

R Language

Different classes of objects in R

Character

« word »

Numeric

« 4.356 »

Integer

« 4 » or « 4L »

Date

« 01/10/2023 »

Logical

« TRUE » « FALSE »

Get the class of the object A

`class(A)`

Change A to a specific class

`as.numeric(A)`

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.0085  9.7121

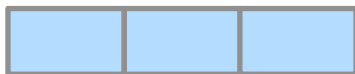
Coefficients:
            Estimate      Error Pr
(Intercept)  1.3831      24566      0.2
x1           5.2977      12483      0.0
x2           1.6105      1518      0.0
x3           0.4494      16      0.0
groupB      1.8797      0.16      0.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```



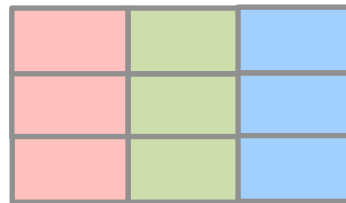
R Language

Create objects



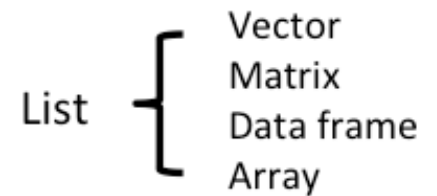
Vector

```
vec1 <- c(« » , « » , « » )
```



Data frame

```
df <- data.frame(col1, col2)
```



```
list1 <- list(vec1, df)
```

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

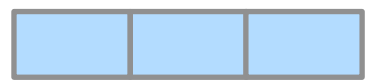
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.38311     0.24566   5.634 <0.0001**
x1           5.29776     0.12483  42.441 <0.0001**
x2           1.61055     0.05518  29.186 <0.0001**
x3           0.44947     0.01609  27.936 <0.0001**
groupB      1.87978     0.16055  11.710 <0.0001**
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

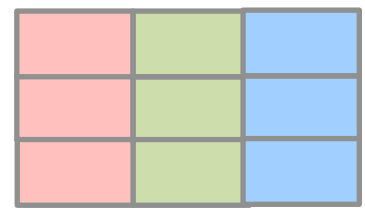


R Language

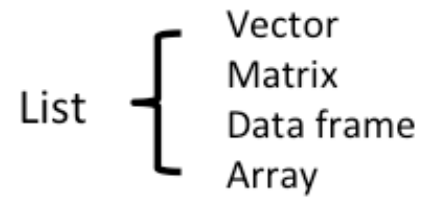
Manipulate objects using « [] »



Vector



Data frame



Vec1[2]

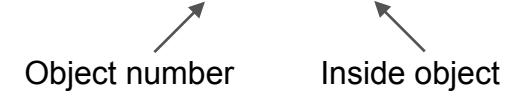
Vec1[c(1,3)]

df[, c(2,3)]



df\$col1

list1[[2]][,c(1:10)]



```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
            1          2          3          4          5
(Intercept)  1.3831    5.2977    1.6105    0.4494    1.8797
            Error term

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

R Language

Basic operations and functions

Different operators

R operators	
+	Addition
*	Multiplication
/	Division
-	Soustraction

Logical operators	
&	And
	Or
!	Not

Relational operators	
>	More than
<	Less than
>=	More or equal
!=	Not equal to

Miscellaneous oper.	
%in%	in

Different functions

subset()
sort()
order()
factor()
str()
print()
plot()
merge()
na.omit()
 ...

To do a specific mathematical or logical operation

To do a specific task


```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

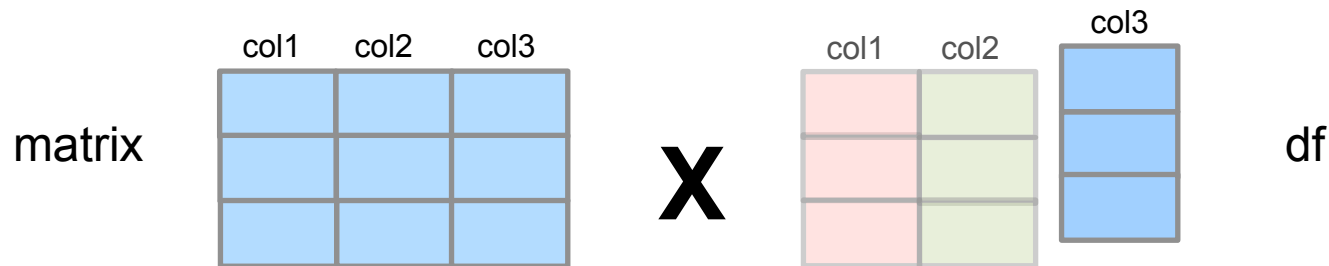
Coefficients:
(Intercept)  1.3831  0.24566  0.2483  0.1618  0.1618
          x1     5.29776  0.12483  0.1618  0.1618
          x2     1.61055  0.1618  0.1618  0.1618
          x3     0.44947  0.1618  0.1618  0.1618
        groupB     1.87978  0.1618  0.1618  0.1618
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

R Language

Basic operations and functions

Multiply one matrix by one column of a data frame



```
new_matrix <- matrix * df$col3
```

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

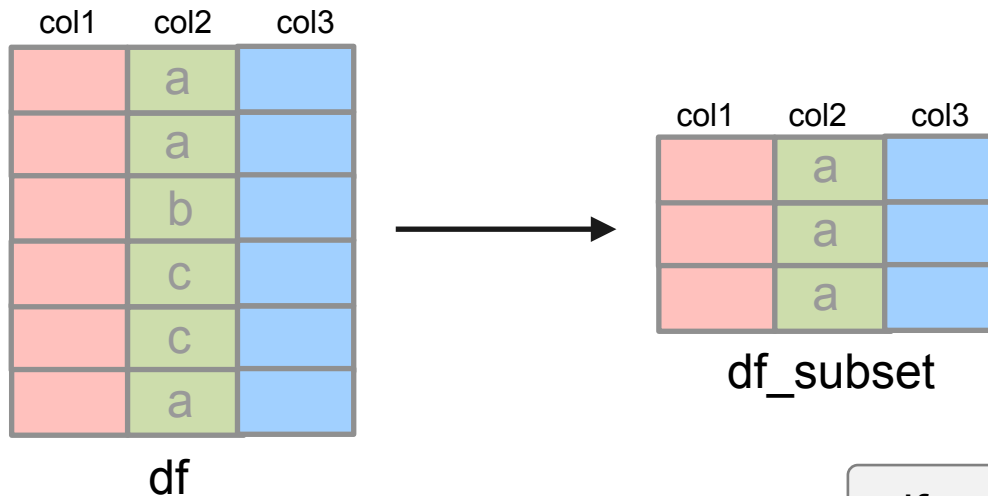
Coefficients:
(Intercept)  1.3831  1.24566  0.24566
x1           5.29776  1.2483  1.2483
x2           1.61055  1.518  1.518
x3           0.44947  0.16  0.16
groupB      1.87978  0.16  0.16
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

R Language

Basic operations and functions

Keep only rows that are in category « a »



```
df_subset <- subset(df, col2 == « a »)
```

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
            1          2          3          4          5          6
(Intercept)  1.3831    5.2977    1.6105    0.4494    1.8797    0.16...
x1           1.3831    5.2977    1.6105    0.4494    1.8797    0.16...
x2           1.3831    5.2977    1.6105    0.4494    1.8797    0.16...
x3           1.3831    5.2977    1.6105    0.4494    1.8797    0.16...
groupB      1.3831    5.2977    1.6105    0.4494    1.8797    0.16...
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

R Packages

Combination of different functions for a specific purpose

Most of the time they are available in CRAN



But sometimes (particularly if they are recent) you can directly install them from Bioconductor or Github



```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.3831     0.24566   5.634 0.0001*
x1           5.2977     0.12483  42.436 <0.0001*
x2           1.6105     0.05518  29.186 <0.0001*
x3           0.44947    0.016666  26.976 <0.0001*
groupB       1.87978    0.16667  11.225 <0.0001*
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

R Packages

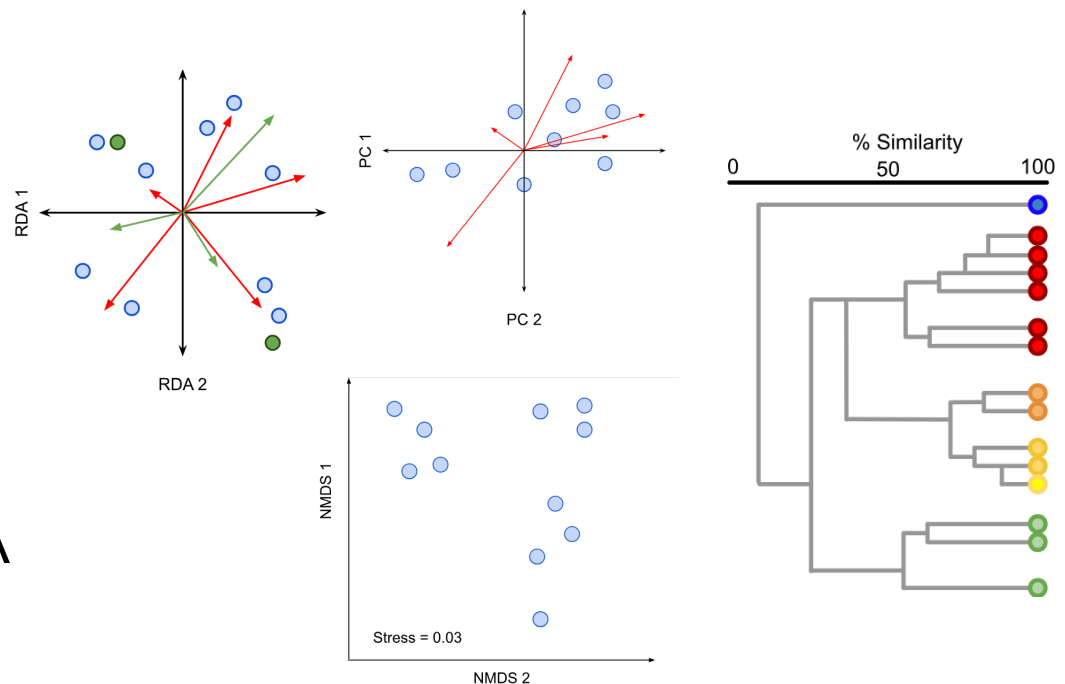
Community ecology : vegan package / ade4 package

Test hypothesis

Multivariate analysis

Constrained analysis

Composant analysis, PCA

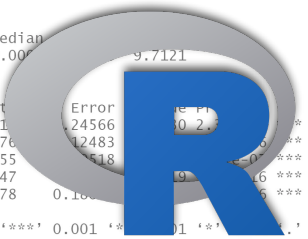


```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

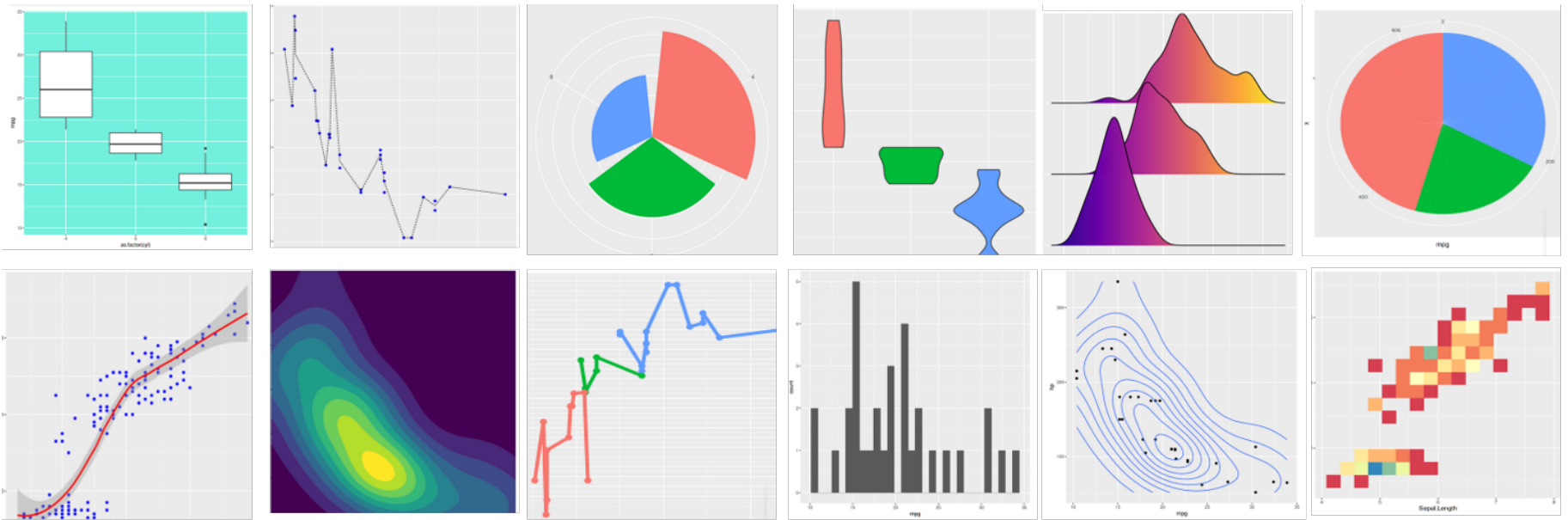
Coefficients:
            1          2          3          4          5
(Intercept)  1.3831  1.24566  0.2483  0.1618  0.1618
x1           5.29776  1.2483  0.1618  0.1618  0.1618
x2           1.61055  0.1618  0.1618  0.1618  0.1618
x3           0.44947  0.1618  0.1618  0.1618  0.1618
groupB      1.87978  0.1618  0.1618  0.1618  0.1618
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```



R Packages

Graphic display : ggplot2



```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
(Intercept)  1.3831    2.4566    0.2111
x1           5.2977    1.2483    0.0000
x2           1.6105    0.7518    0.0000
x3           0.4494    0.0000    0.0000
groupB      1.8797    0.1000    0.0000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```



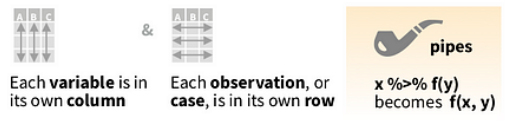
R Packages

Data manipulation : Dplyr

Data Transformation with dplyr :: CHEAT SHEET

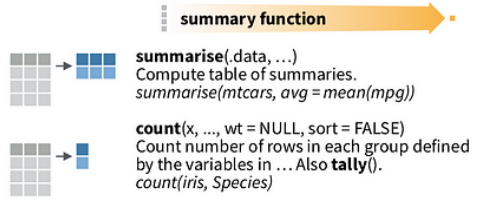


dplyr functions work with pipes and expect **tidy data**. In tidy data:



Summarise Cases

These apply **summary functions** to columns to create a new table of summary statistics. Summary functions take vectors as input and return one value (see back).

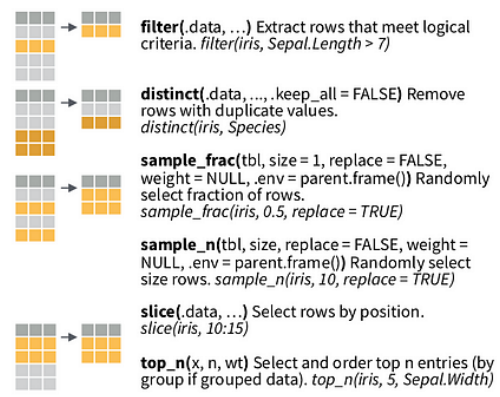


VARIATIONS

Manipulate Cases

EXTRACT CASES

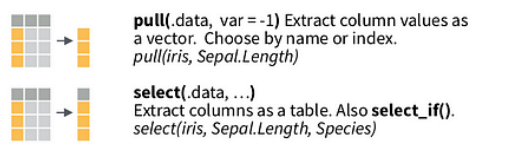
Row functions return a subset of rows as a new table.



Manipulate Variables

EXTRACT VARIABLES

Column functions return a set of columns as a new vector or table.

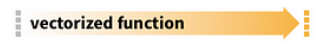


Use these helpers with **select** (), e.g. *select(iris, starts_with("Sepal"))*

contains(match) **num_range**(prefix, range) ; e.g. *mpg:cyl*
ends_with(match) **one_of**(...) ; e.g. *-Species*
matches(match) **starts_with**(match)

MAKE NEW VARIABLES

These apply **vectorized functions** to columns. Vectorized funs take vectors as input and return vectors of the same length as output (see back).

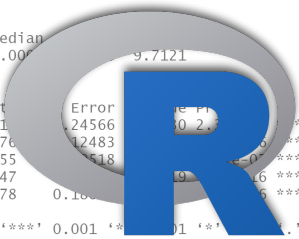


```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9711  9.7121

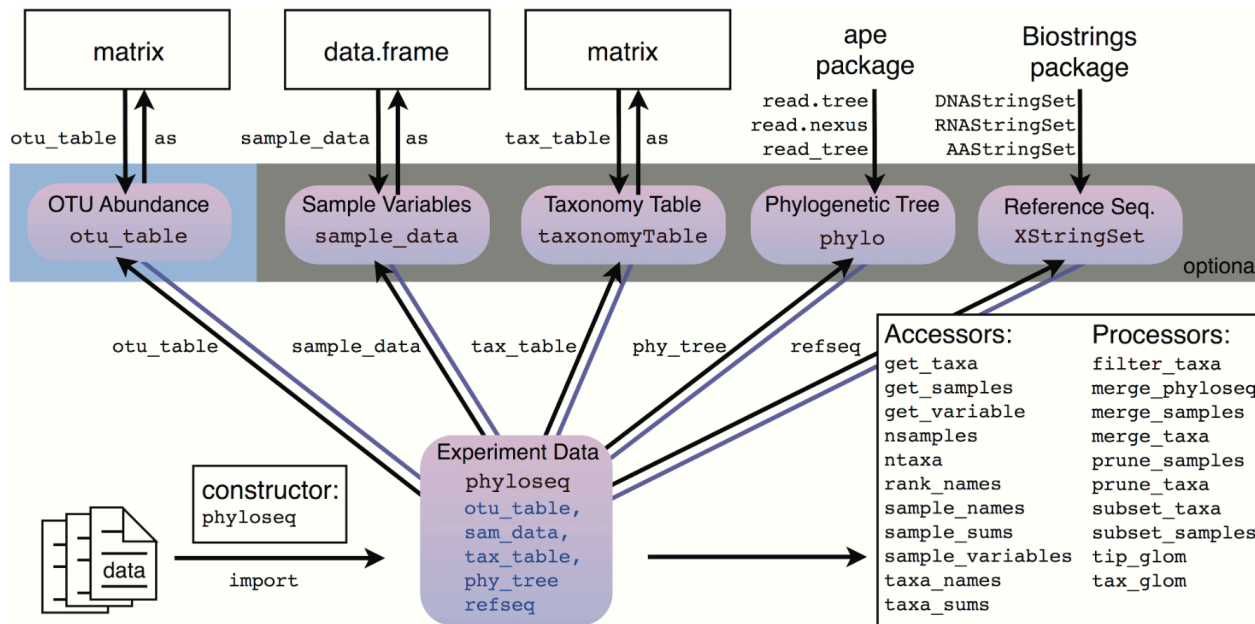
Coefficients:
(Intercept)  1.3831  0.24566  0.02777  0.00000  0.00000
          x1      5.29776  0.12483  0.00000  0.00000  0.00000
          x2      1.61055  0.05518  0.00000  0.00000  0.00000
          x3      0.44947  0.00000  0.00000  0.00000  0.00000
        groupB  1.87978  0.10000  0.00000  0.00000  0.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```



R Packages

A package for metabarcoding data analysis : phyloseq



```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.0085  9.7121

Coefficients:
(Intercept)  1.3831    2.4566    0.2483    0.1618
          x1     5.2977     1.6105     0.4494     1.8797
          x2     1.6105     0.4494     1.8797
          x3     0.4494     1.8797
    groupB     1.8797     0.1618

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```



R Help

Learning how to code in R language could be the subject of an entire workshop

It takes time to understand how the R function works, how to have the good code that will do what we want...

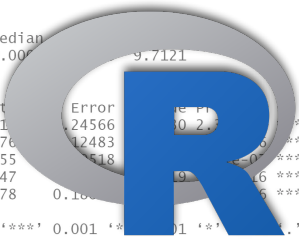
An important reflex : search for help


```
Call:
lm(formula = y ~ ., data = data)

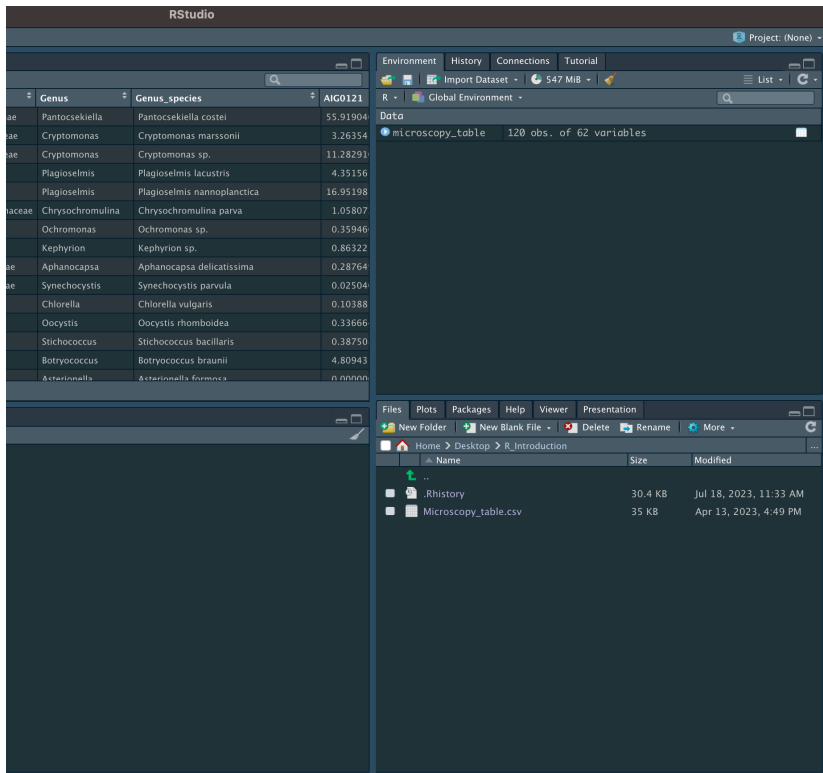
Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
(Intercept)  1.3831  0.24566  0.02777  0.00000  0.00000
x1           5.29776  0.12483  0.00000  0.00000  0.00000
x2           1.61055  0.0518  0.00000  0.00000  0.00000
x3           0.44947  0.00000  0.00000  0.00000  0.00000
groupB       1.87978  0.10000  0.00000  0.00000  0.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

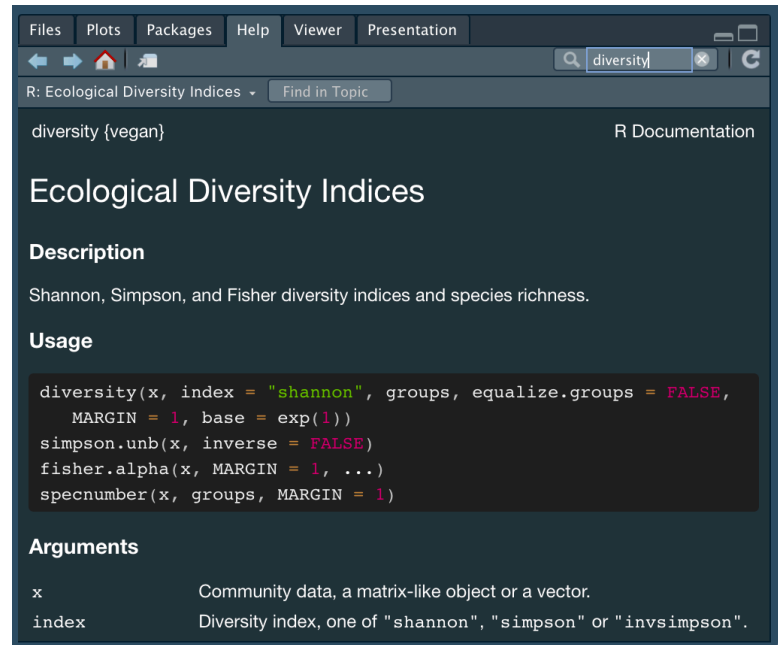


R Help



The screenshot shows the RStudio environment. The top pane displays a data table with columns 'Genus' and 'Genus_species'. The bottom pane shows a file explorer with files like '.Rhistory' and 'Microscopy_table.csv'.

Genus	Genus_species	Abundance
Pantocsekiella	Pantocsekiella costei	55.91904
Cryptomonas	Cryptomonas marssonii	3.26354
Cryptomonas	Cryptomonas sp.	11.28291
Plagioelmis	Plagioelmis lacustris	4.35156
Plagioelmis	Plagioelmis nannoplantica	16.95198
Chrysochromulina	Chrysochromulina parva	1.05807
Ochromonas	Ochromonas sp.	0.35946
Kephyrion	Kephyrion sp.	0.86322
Aphanocapsa	Aphanocapsa delicatissima	0.28764
Synechocystis	Synechocystis parvula	0.02504
Chlorella	Chlorella vulgaris	0.10388
Oocystis	Oocystis rhomboidea	0.33666
Stichococcus	Stichococcus bacillaris	0.38750
Botryococcus	Botryococcus braunii	4.80943
Actinonella	Actinonella formosa	0.00000



The screenshot shows the R Documentation page for the 'diversity' function in the 'vegan' package. It includes a description, usage examples, and arguments.

Ecological Diversity Indices

Description

Shannon, Simpson, and Fisher diversity indices and species richness.

Usage

```
diversity(x, index = "shannon", groups, equalize.groups = FALSE,
          MARGIN = 1, base = exp(1))
simpson.unb(x, inverse = FALSE)
fisher.alpha(x, MARGIN = 1, ...)
specnumber(x, groups, MARGIN = 1)
```

Arguments

- x**: Community data, a matrix-like object or a vector.
- index**: Diversity index, one of "shannon", "simpson" or "invsimpson".

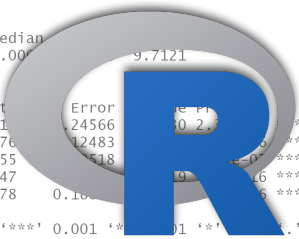


```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.7859 -2.0085  0.0000  2.9712  9.7121

Coefficients:
(Intercept)  1.3831  0.2456  0.0271  0.0000  0.0000
x1           5.2977  0.1248  0.0000  0.0000  0.0000
x2           1.6105  0.0518  0.0000  0.0000  0.0000
x3           0.4494  0.0000  0.0000  0.0000  0.0000
groupB      1.8797  0.1000  0.0000  0.0000  0.0000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

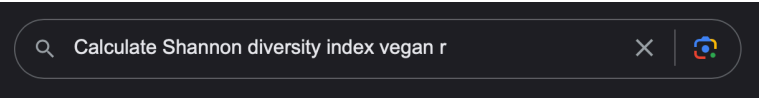
Residual standard error: 2.817 on 995 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7873
F-statistic: 925.6 on 4 and 995 DF,  p-value: < 2.2e-16
```



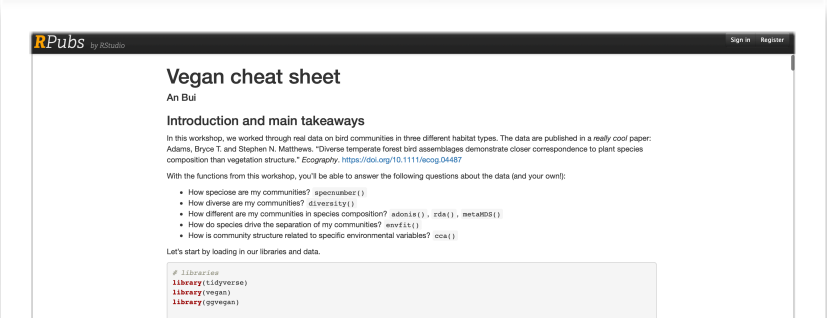
R Help

Any question has an answer on internet !!

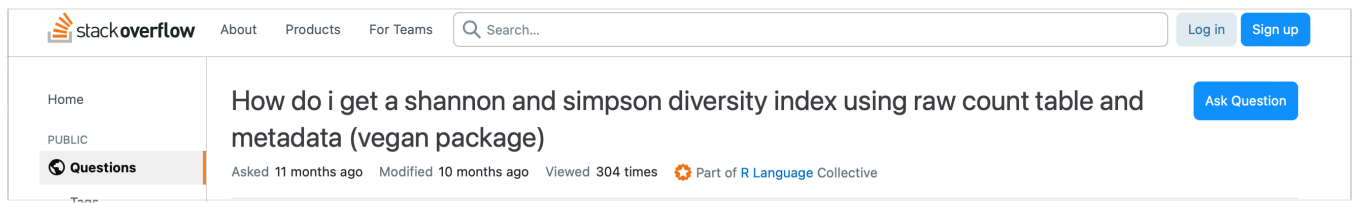
Official documentation of
Vegan package



Tutorial made by other scientists



Forums



Now let's practice !