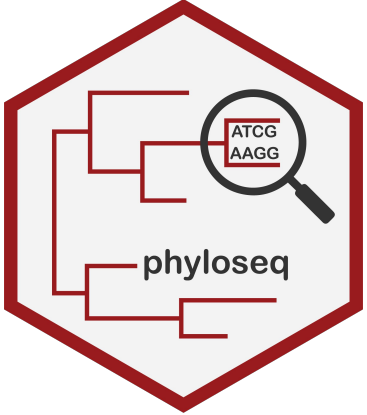


Introduction to R for Microbiome Data (part 2)
- Specifically Designed R Packages for Microbiome Data:
Phyloseq

Presenter: huyha

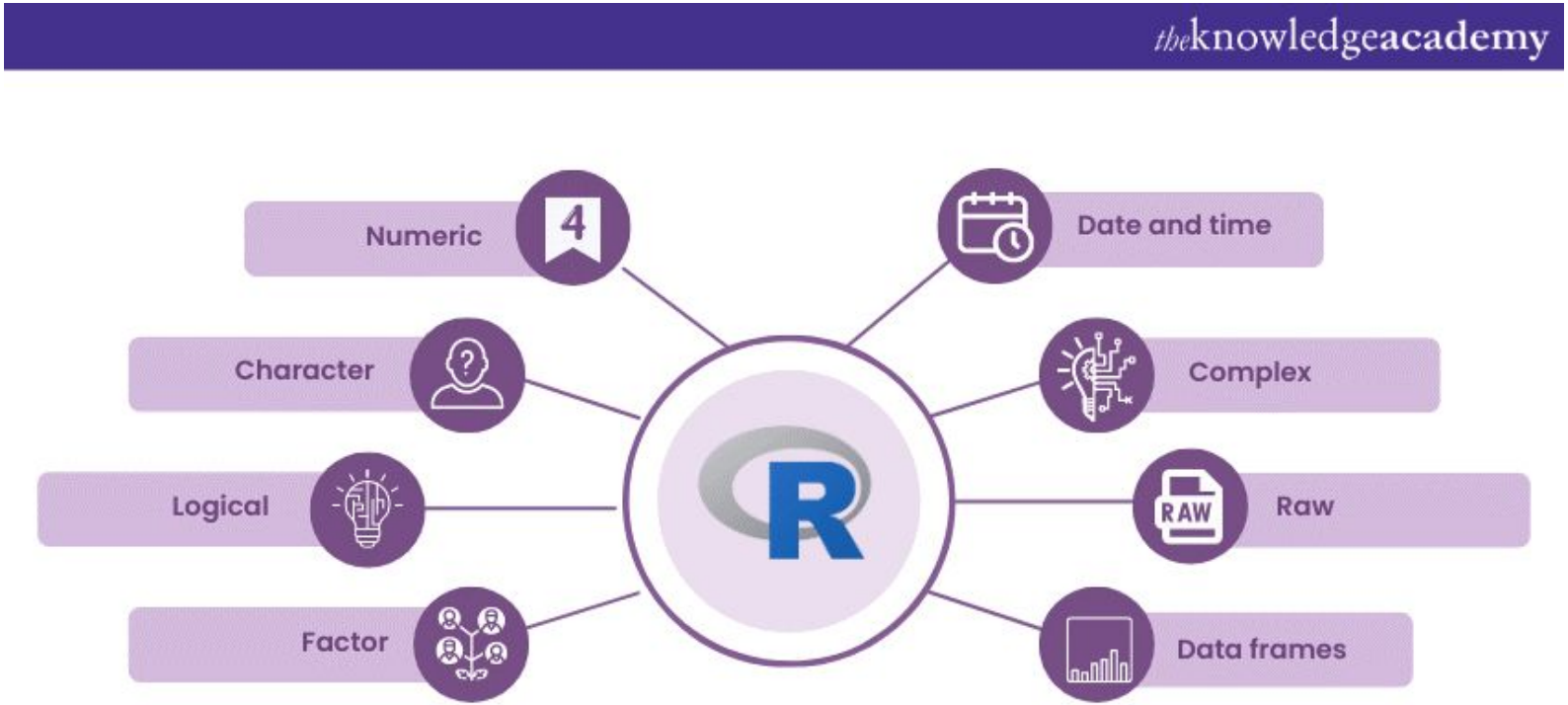
Contents

1. Review R data
2. How to install Phyloseq?
3. Introducing to Phyloseq package.
4. Some basic functions
5. Tutorial + practice
6. Several plots for microbiome data

Why use R ?

- Script vs. Menu driven software
  - Can be re-rerun with new data
- R works with vectors and tables
  - Z <- X + Y where X and Y are vectors
- "Tidy" worflow
  - Select(Tara_table, fraction="0.2-5")
  - Tidy "universe" : tidyverse
- Incredible graphics and plots
  - ggplot2 family
- Work environment
  - R studio
- Document your data processing
  - R markdown
- Share your data and workflow
  - GitHub

Data type


theknowledgeacademy
Numeric
4
Character
Logical
Factor
Date and time
Complex
Raw
RAW
Data frames
R

Types and struture of data

```r
# Numeric data
numeric_value <- 42
print(numeric_value)
```

```r
# Raw data
raw_value <- charToRaw("Hello")
print(raw_value)
```

```r
# Character data
char_value <- "Hello, World!"
print(char_value)
```

```r
# Date and time data
date_value <- as.Date("2022-01-01")
print(date_value)
```

```r
# Logical data
logical_value <- TRUE
print(logical_value)
```

```r
# Complex data
complex_value <- 3 + 4i
print(complex_value)
```

Structure of data

- Vector: store elements of the same data type
- Factor: represent categorical data in R
- Data_frame: two-dimensional labeled data structure with columns of potentially different
- Matrix: two-dimensional array-like structure consisting of rows and columns
- Lists: A list is a collection of objects (which can be vectors, matrices, dataframes, etc.) stored in a single variable
- Arrays: multi-dimensional generalizations of vectors and matrices
- Functions blocks of code that perform a specific task and can be called by their name
- Formulae: statistical models.
- Environments: store mappings from symbols to value
- Expressions: operations to be performed

Source: chatGPT

Structure of data

```r
# Single vector object
single_vector <- c(1, 2, 3, 4, 5)
```

```r
# Matrix
my_matrix <- matrix(1:9, nrow = 3, ncol = 3)
```

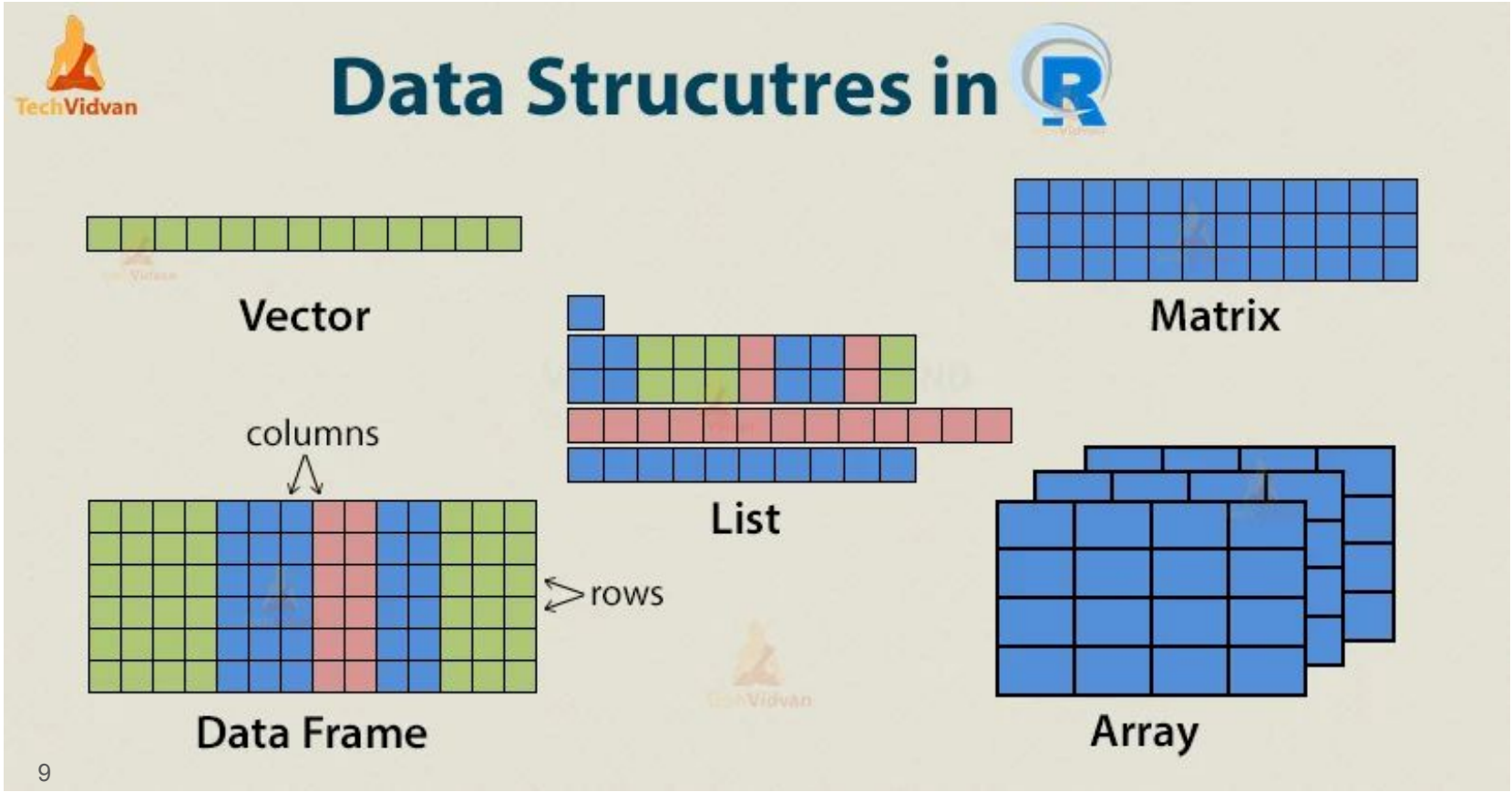
```r
# Single data frame object
single_df <- data.frame(
  Name = c("Alice", "Bob", "Charlie"),
  Age = c(30, 25, 40),
  Gender = c("Female", "Male", "Male")
)
```

Structure of data

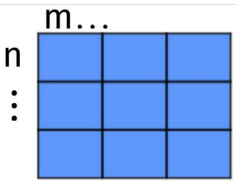
```r
# Create a list
my_list <- list(
  numeric_vector = c(1, 2, 3),
  character_vector = c("a", "b", "c"),
  my_matrix = matrix(1:9, nrow = 3, ncol = 3)
)
```

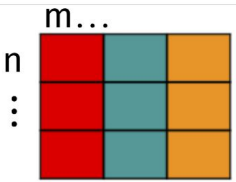
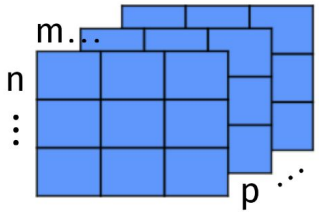
```r
# Create an array
my_array <- array(1:24, dim = c(2, 3, 4))
```

Data structures



	Dimensions	Mode (data "type")	Example
Vector		Identical	`c(10,0.2,34,48,53)`

Matrix		Identical	`matrix(c(1,2,3,`
			`11,12,13),`
			`nrow = 2,`
			`ncol = 3)`

Data frame		Can be different	`data.frame(x = 1:3,` `y = 5:7)`
Array		Identical	`array(data = 1:3,` `dim = c(2,4,2))`
List	Vector Matrix Data frame Array	Can be different	`list(x = cars[,1],` `y = cars[,2])`

Some question to test

```
#dataframe
1 <- data.frame(
    Name = c("Long", "The", "Cuong" "Dinh")
    Age = c(12, 15 , 16)
    Test =c(TRUE, FALSE, TRUE, TRUE)
    )

2. <- matrix(1:9, nrow=3, ncol=3)

3 <- c("Big", "Small","Big","Small","Small")
```

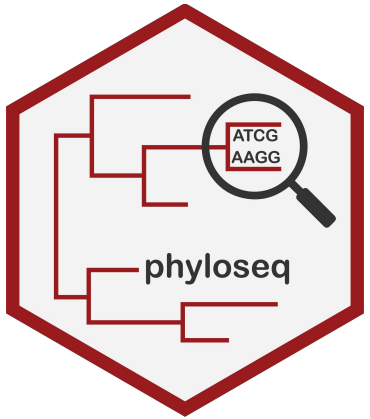
Some question to test

```r
#dataframe
1 <- data.frame(
    Name = c("Long", "The", "Cuong" "Dinh")
    Age = c(12, 15 , 16)
    Test =c(TRUE, FALSE, TRUE, TRUE)
    )

2. <- matrix(1:9, nrow=3, ncol=3)

3 <- c("Big", "Small","Big","Small","Small")
```

Introduction to **Phyloseq**



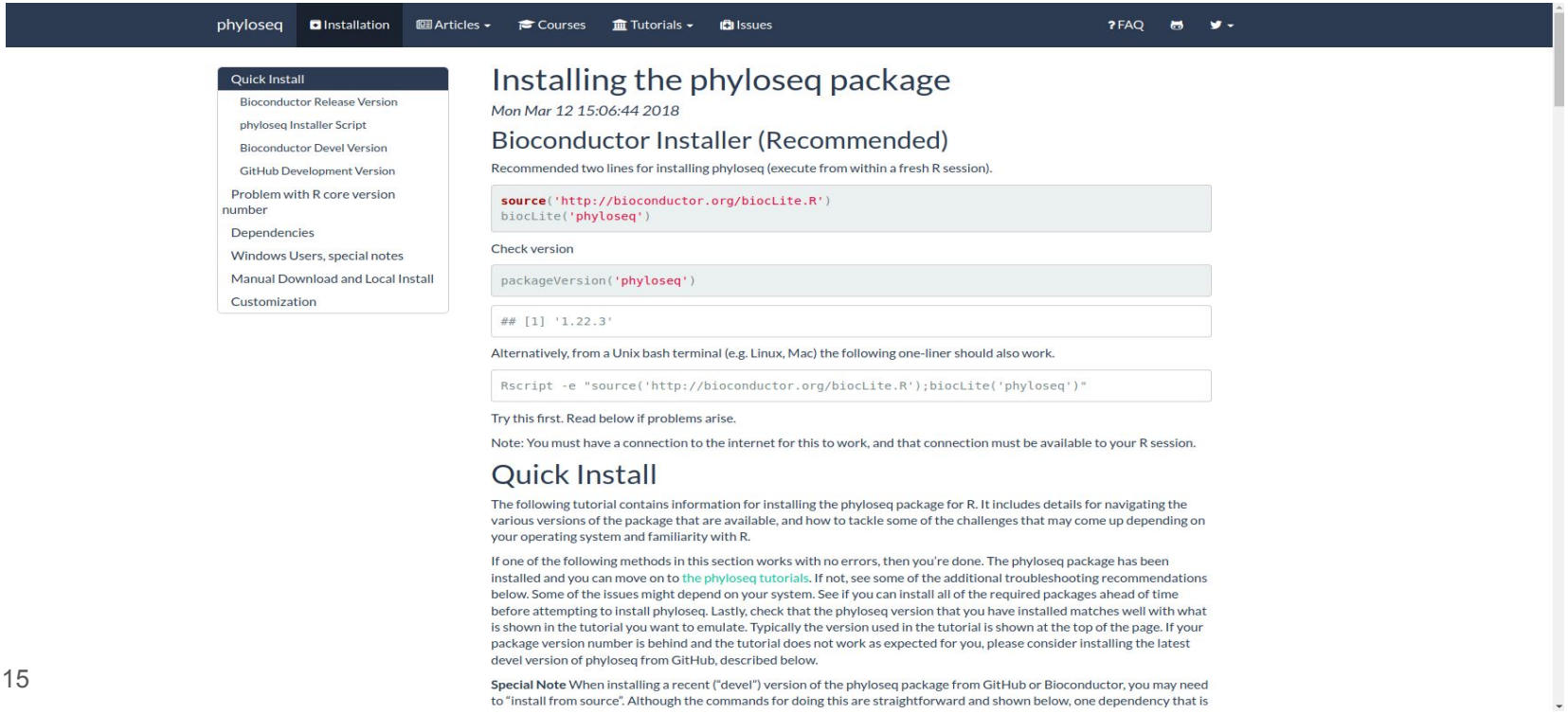
Why do we use phyloseq?

Phyloseq: A Tool for Microbiome Analysis in R

- **Challenge:** Analyzing microbiome data requires integrating difference type of data from ecology, genetics, phylogenetics, statistics & visualization.
- **Problem:** Existing R tools lack support for high-throughput microbiome data, making analyses difficult to **reproduce**.
- **Solution:** Phyloseq: an R package for object-oriented representation and analysis of microbiome census data.
- **Features:**
  - Imports data from various formats.
  - Supports calibration, filtering, diversity analysis, ordination methods, and more.
  - Creates publication-quality graphics.
  - Integrates with other R packages for open-source analysis techniques.
- **Benefits:**
  - Reproducible research: share and modify analyses easily.
  - Open-source and freely available on GitHub and Bioconductor.

How to install Phyloseq in Rstudio?

Link https://joey711.github.io/phyloseq/install.html



Phyloseq

OPEN ACCESS Freely available online

PLOS ONE

phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data

Paul J. McMurdie, Susan Holmes*

Department of Statistics, Stanford University, Stanford, California, United States of America

Abstract

Background: The analysis of microbial communities through DNA sequencing brings many challenges: the integration of different types of data with methods from ecology, genetics, phylogenetics, multivariate statistics, visualization and testing. With the increased breadth of experimental designs now being pursued, project-specific statistical analyses are often needed, and these analyses are often difficult (or impossible) for peer researchers to independently reproduce. The vast majority of the requisite tools for performing these analyses reproducibly are already implemented in R and its extensions (packages), but with limited support for high throughput microbiome census data.

Results: Here we describe a software project, phyloseq, dedicated to the object-oriented representation and analysis of microbiome census data in R. It supports importing data from a variety of common formats, as well as many analysis techniques. These include calibration, filtering, subsetting, agglomeration, multi-table comparisons, diversity analysis, parallelized Fast UniFrac, ordination methods, and production of publication-quality graphics; all in a manner that is easy to document, share, and modify. We show how to apply functions from other R packages to phyloseq-represented data, illustrating the availability of a large number of open source analysis techniques. We discuss the use of phyloseq with tools for reproducible research, a practice common in other fields but still rare in the analysis of highly parallel microbiome census data. We have made available all of the materials necessary to completely reproduce the analysis and figures included in this article, an example of best practices for reproducible research.

Conclusions: The phyloseq project for R is a new open-source software package, freely available on the web from both GitHub and Bioconductor.

Citation: McMurdie PJ, Holmes S (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8(4): e61217. doi:10.1371/journal.pone.0061217

Editor: Michael Watson, The Roslin Institute, University of Edinburgh, United Kingdom

Received October 17, 2012; **Accepted** March 6, 2013; **Published** April 22, 2013

Funding: This work was supported by grant NIH-R01GM086884. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: susan@stat.stanford.edu

Phyloseq sequencing workflow

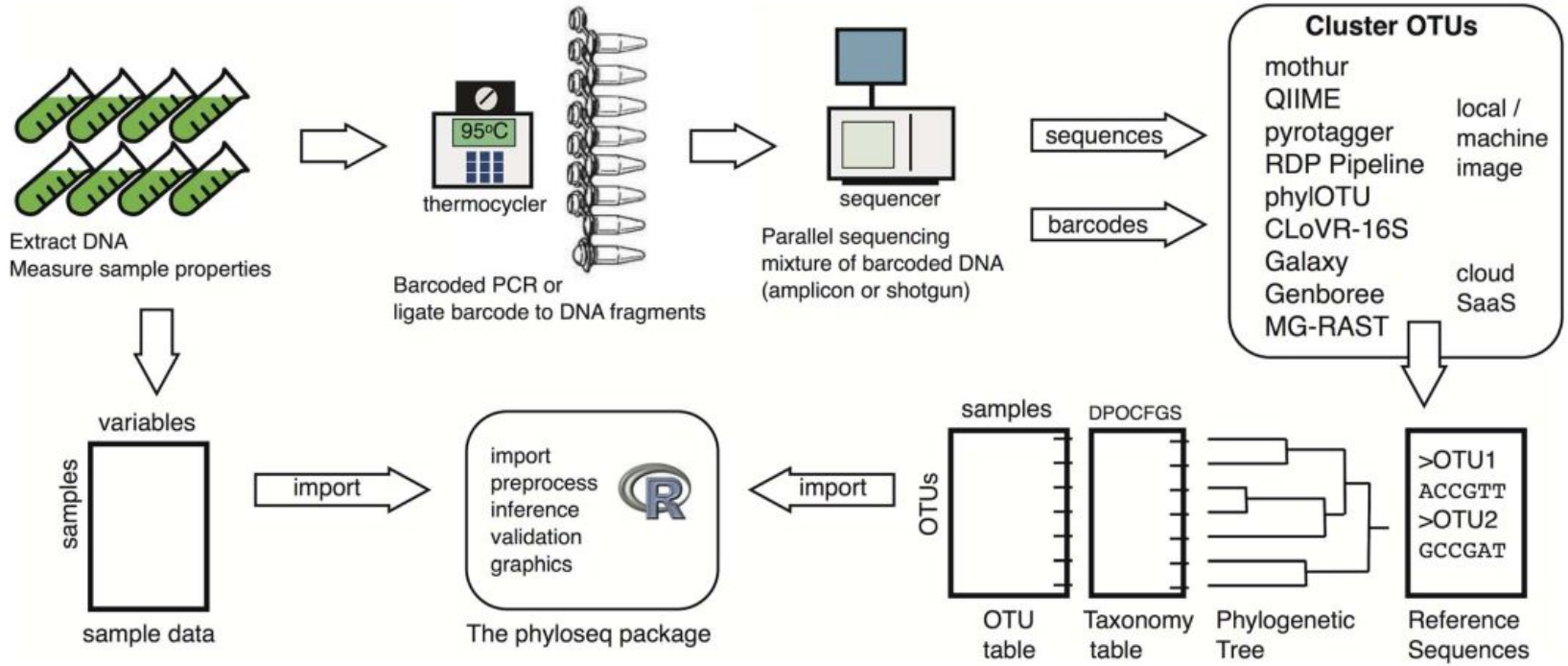


Figure 1. Example of a phylogenetic sequencing workflow. A diagram of an experimental and analysis workflow for amplicon or shotgun phylogenetic sequencing. The intended role for phyloseq is indicated.
doi:10.1371/journal.pone.0061217.g001

Phyloseq analysis workflow

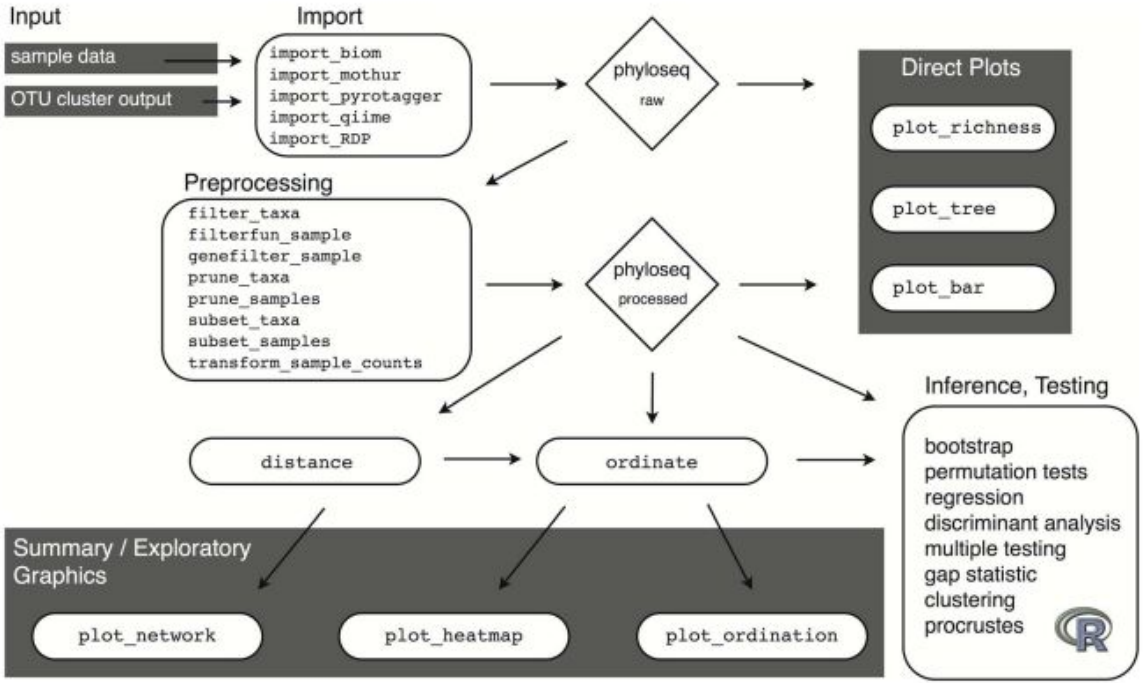


Figure 2. Analysis workflow using phyloseq. The workflow starts with the results of OTU clustering and independently-measured sample data (Input, top left), and ends at various analytic procedures available in R for inference and validation. In between are key functions for preprocessing and graphics. Rounded rectangles and diamond shapes represent functions and data objects, respectively, further described in Figure 3.
doi:10.1371/journal.pone.0061217.g002

Import data

- Necessary
  1. OTU table
  2. Taxonomy table
  3. Sample variables

- Optional
  - Tree
  - Sequences

Phyloseq class

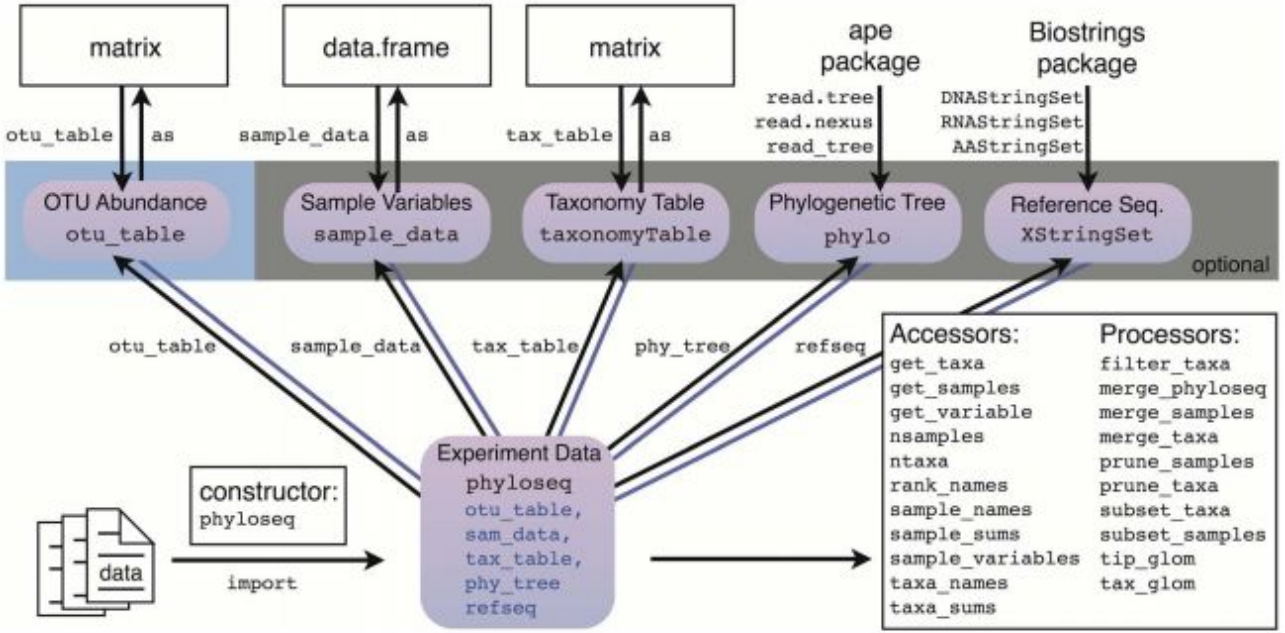
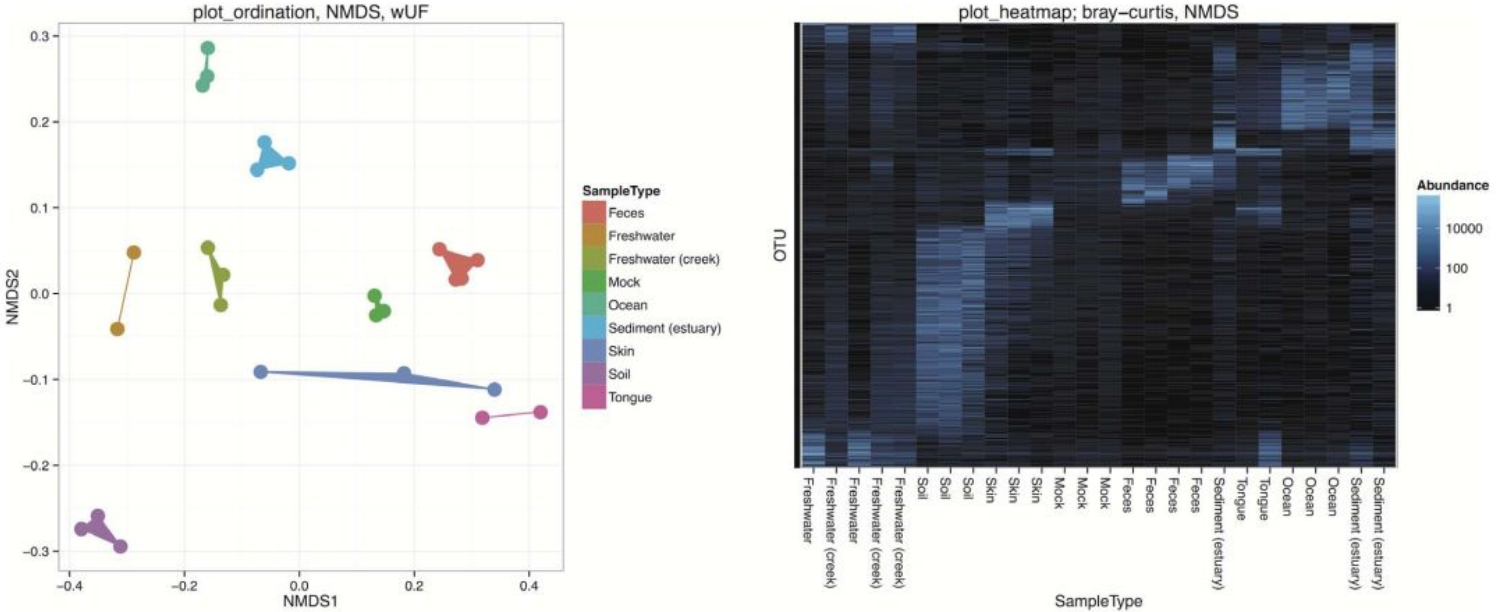


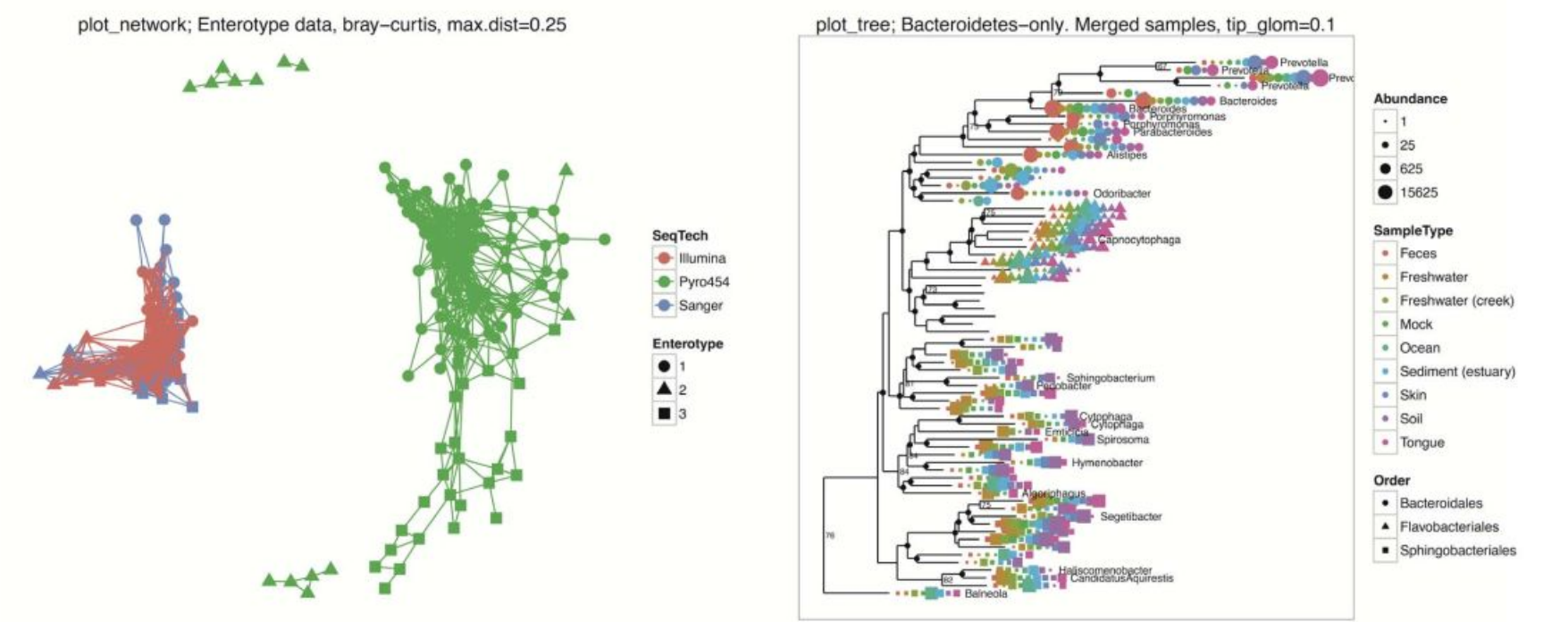
Figure 3. The "phyloseq" class. The phyloseq class is an experiment-level data storage class defined by the phyloseq package for representing phylogenetic sequencing data. Most functions in the phyloseq package expect an instance of this class as their primary argument. See the phyloseq manual [38] for a complete list of functions.
doi:10.1371/journal.pone.0061217.g003

Phyloseq
visualize function

An R Package for Microbiome Census Data



Phyloseq visualize function



Phyloseq visualize function

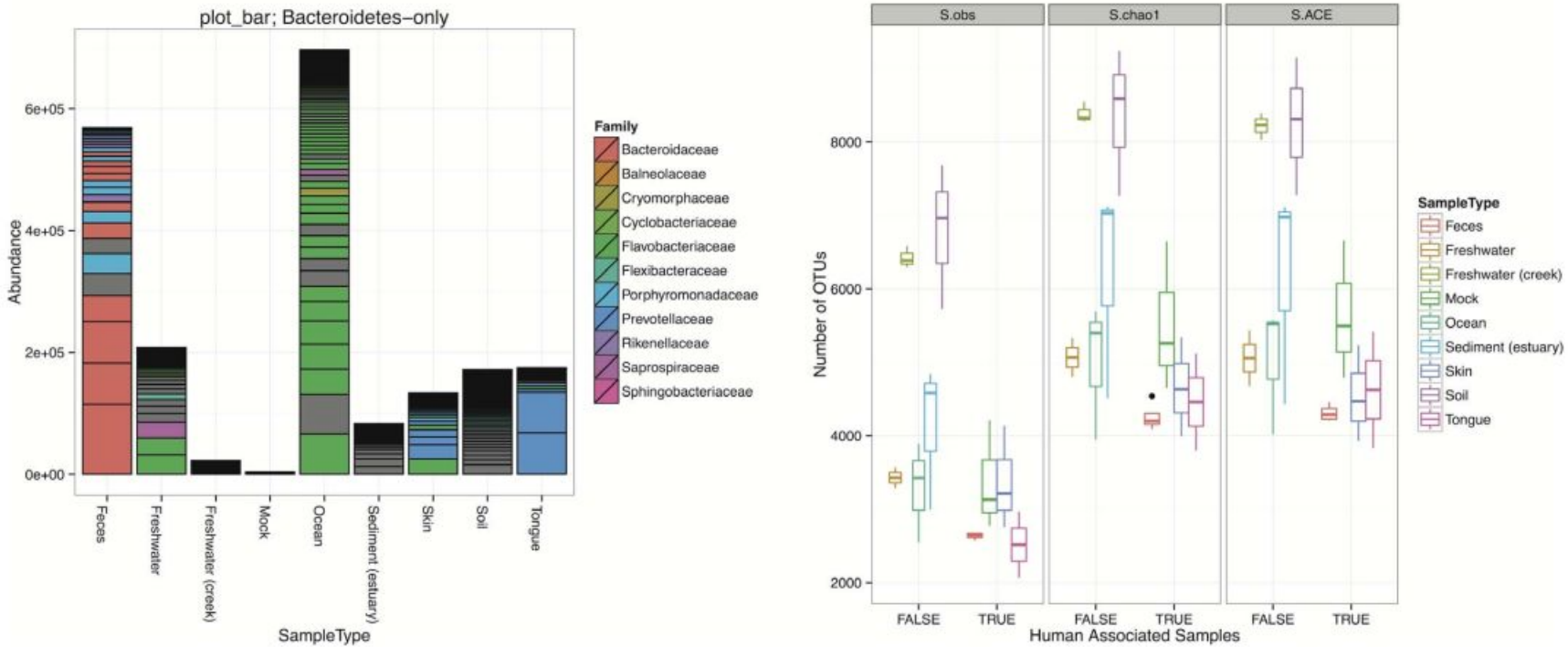


Figure 4. Graphic functions of the phyloseq package. The phyloseq class is an experiment-level data storage class defined by the phyloseq package for representing phylogenetic sequencing data. Most functions in the phyloseq package expect an instance of this class as their primary argument. See the phyloseq manual The Global Patterns [47] and Enterotypes [91] datasets are included with the phyloseq package. The Global

1. OTU_matrix import - example data

```
# Create a pretend OTU table that you read from a file, called otumat
otumat = matrix(sample(1:100, 100, replace = TRUE), nrow = 10, ncol = 10)
otumat
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]   96   50   36   35   59   80   83   63   38    35
##  [2,]   52   67   39   39   37   57   20   15   64    94
##  [3,]   94   18   15   11   14   75    1   12   42    58
##  [4,]   27   88   98  100   59   27   30   30   94    78
##  [5,]   26   66   93   85   41   30  100   41   92   100
##  [6,]   17   16   97   86   18   25   94   31   62    37
##  [7,]   63   19   16   43   89   25   17   63   15    82
##  [8,]   31   92   22   14   58    1   45    2   25    35
##  [9,]  100   33   19   77   43    1   14   69   42    18
## [10,]   13   35   80   43   34   45   24   47   71    72
```

Sample names and OTU names

```
rownames(otumat) <- paste0("OTU", 1:nrow(otumat))

colnames(otumat) <- paste0("Samplea 1:ncol(otumat))
  otumat
```

```
##       Sample1 Sample2 Sample3 Sample4 Sample5 Sample6 Sample7 Sample8
## OTU1       96      50      36      35      59      80      83      63
## OTU2       52      67      39      39      37      57      20      15
## OTU3       94      18      15      11      14      75       1      12
## OTU4       27      88      98     100      59      27      30      30
## OTU5       26      66      93      85      41      30     100      41
## OTU6       17      16      97      86      18      25      94      31
## OTU7       63      19      16      43      89      25      17      63
## OTU8       31      92      22      14      58       1      45       2
## OTU9      100      33      19      77      43       1      14      69
## OTU10      13      35      80      43      34      45      24      47
##       Sample9 Sample10
## OTU1       38       35
## OTU2       64       94
## OTU3       42       58
## OTU4       94       78
## OTU5       92      100
## OTU6       62       37
## OTU7       15       82
## OTU8       25       35
## OTU9       42       18
## OTU10      71       72
```

2. Taxa_matrix - example data

```
taxmat = matrix(sample(letters, 70, replace = TRUE), nrow = nrow(otumat), ncol = 7)
rownames(taxmat) <- rownames(otumat)
colnames(taxmat) <- c("Domain", "Phylum", "Class", "Order", "Family", "Genus", "Species")
  taxmat
```

```
##       Domain Phylum Class Order Family Genus Species
## OTU1  "x"    "d"    "q"   "v"   "l"    "k"   "i"
## OTU2  "a"    "d"    "x"   "a"   "k"    "o"   "r"
## OTU3  "h"    "a"    "h"   "c"   "d"    "j"   "k"
## OTU4  "t"    "f"    "j"   "e"   "n"    "y"   "o"
## OTU5  "o"    "q"    "s"   "w"   "d"    "y"   "j"
## OTU6  "e"    "r"    "p"   "k"   "b"    "v"   "t"
## OTU7  "m"    "l"    "y"   "u"   "b"    "y"   "q"
## OTU8  "d"    "o"    "w"   "g"   "p"    "w"   "v"
## OTU9  "f"    "o"    "a"   "n"   "l"    "u"   "e"
## OTU10 "h"    "r"    "d"   "j"   "u"    "f"   "a"
```

```
class(otumat)
```

```
## [1] "matrix"
```

```
class(taxmat)
```

```
## [1] "matrix"
```

2. Create phyloseq_object - example data

```
library("phyloseq")
OTU = otu_table(otumat, taxa_are_rows = TRUE)
TAX = tax_table(taxmat)
OTU # OTU view
```

```
## OTU Table:          [10 taxa and 10 samples]
##                      taxa are rows
##       Sample1 Sample2 Sample3 Sample4 Sample5 Sample6 Sample7 Sample8
## OTU1       96      50      36      35      59      80      83      63
## OTU2       52      67      39      39      37      57      20      15
## OTU3       94      18      15      11      14      75       1      12
## OTU4       27      88      98     100      59      27      30      30
## OTU5       26      66      93      85      41      30     100      41
## OTU6       17      16      97      86      18      25      94      31
## OTU7       63      19      16      43      89      25      17      63
## OTU8       31      92      22      14      58       1      45       2
## OTU9      100      33      19      77      43       1      14      69
## OTU10      13      35      80      43      34      45      24      47
##       Sample9 Sample10
## OTU1       38       35
## OTU2       64       94
## OTU3       42       58
## OTU4       94       78
## OTU5       92      100
## OTU6       62       37
## OTU7       15       82
## OTU8       25       35
## OTU9       42       18
## OTU10      71       72
```

2. Create phyloseq_object - example data

```
TAX
```

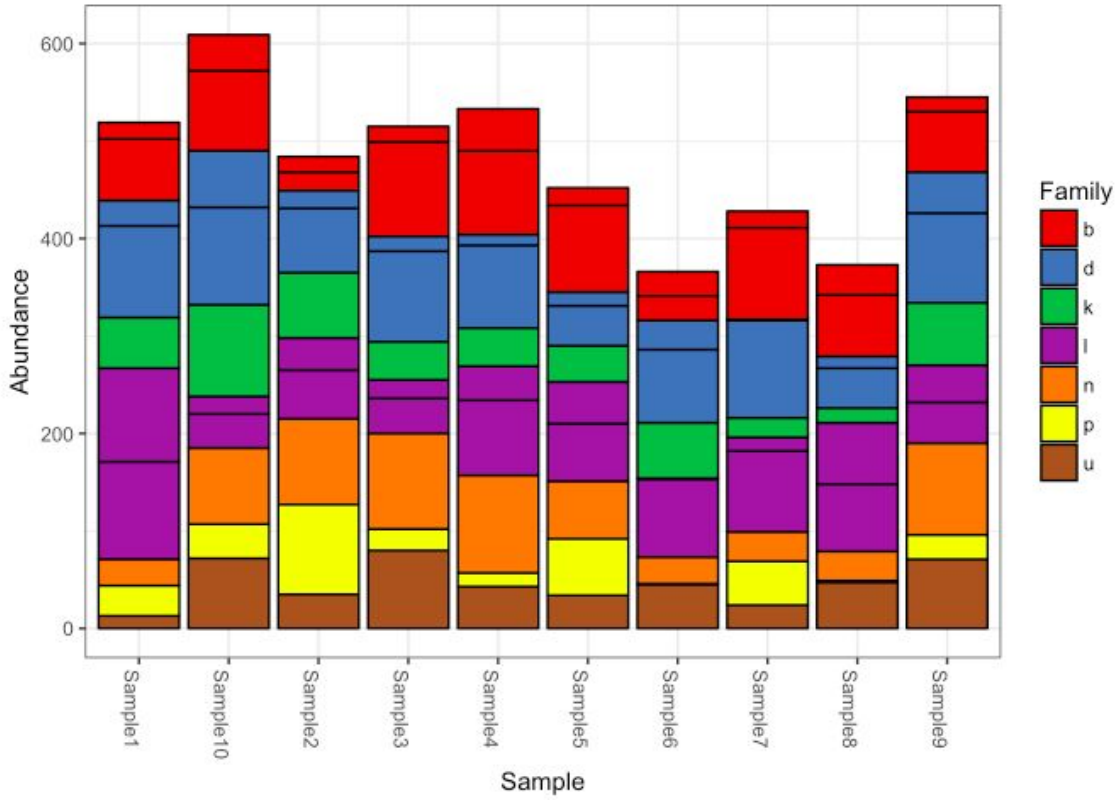
```
## Taxonomy Table:     [10 taxa by 7 taxonomic ranks]:
##       Domain Phylum Class Order Family Genus Species
## OTU1  "x"    "d"    "q"   "v"   "l"    "k"   "i"
## OTU2  "a"    "d"    "x"   "a"   "k"    "o"   "r"
## OTU3  "h"    "a"    "h"   "c"   "d"    "j"   "k"
## OTU4  "t"    "f"    "j"   "e"   "n"    "y"   "o"
## OTU5  "o"    "q"    "s"   "w"   "d"    "y"   "j"
## OTU6  "e"    "r"    "p"   "k"   "b"    "v"   "t"
## OTU7  "m"    "l"    "y"   "u"   "b"    "y"   "q"
## OTU8  "d"    "o"    "w"   "g"   "p"    "w"   "v"
## OTU9  "f"    "o"    "a"   "n"   "l"    "u"   "e"
## OTU10 "h"    "r"    "d"   "j"   "u"    "f"   "a"
```

```
physeq = phyloseq(OTU, TAX)
physeq
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 10 taxa and 10 samples ]
## tax_table()   Taxonomy Table:    [ 10 taxa by 7 taxonomic ranks ]
```

3.Barplot

```
plot_bar(physeq, fill = "Family")
```



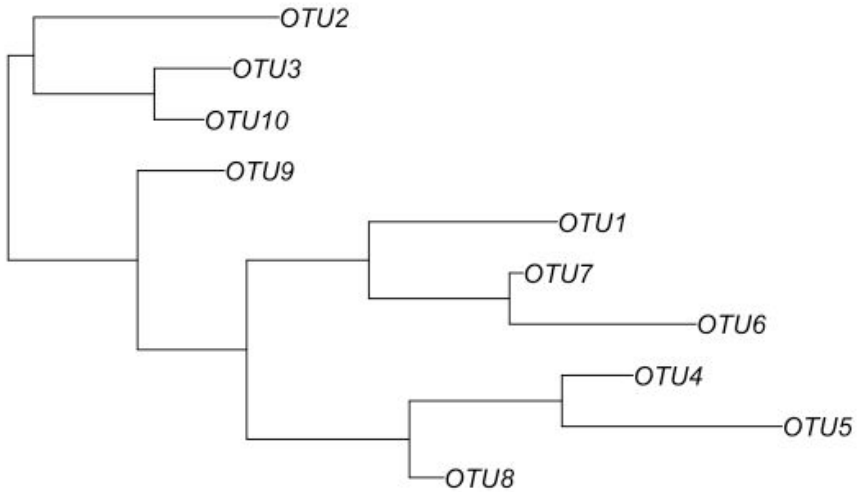
4. Sample data - meta data

```
sampledata = sample_data(data.frame(
  Location = sample(LETTERS[1:4], size=nsamples(physeq), replace=TRUE),
  Depth = sample(50:1000, size=nsamples(physeq), replace=TRUE),
  row.names=sample_names(physeq),
  stringsAsFactors=FALSE
))
  sampledata
```

```
##          Location Depth
## Sample1         D   337
## Sample2         B    74
## Sample3         D    68
## Sample4         C   397
## Sample5         B   142
## Sample6         D   970
## Sample7         D    69
## Sample8         C   253
## Sample9         A   497
## Sample10        D   237
```

5. Phylogenetic tree

```
library("ape")
random_tree = rtree(ntaxa(physeq), rooted=TRUE, tip.label=taxa_names(physeq))
plot(random_tree)
```



6. Merge data

```
physeq1 = merge_phyloseq(physeq, sampledata, random_tree)
physeq1
```

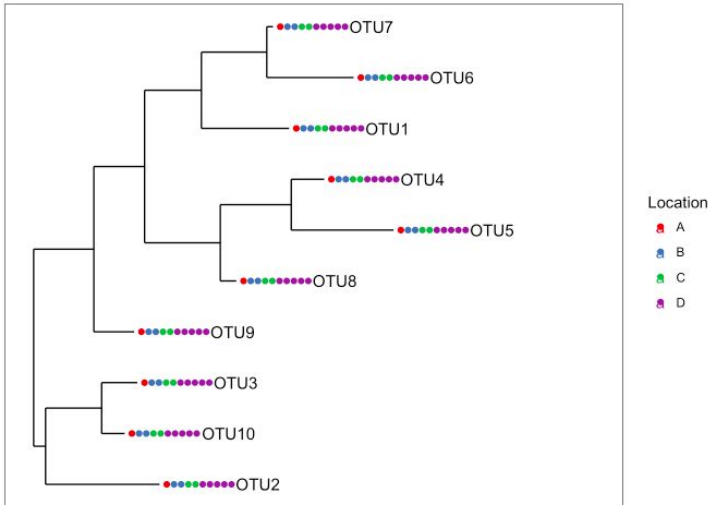
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 10 taxa and 10 samples ]
## sample_data() Sample Data:       [ 10 samples by 2 sample variables ]
## tax_table()   Taxonomy Table:    [ 10 taxa by 7 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 10 tips and 9 internal nodes ]
```

```
physeq2 = phyloseq(OTU, TAX, sampledata, random_tree)
physeq2
```

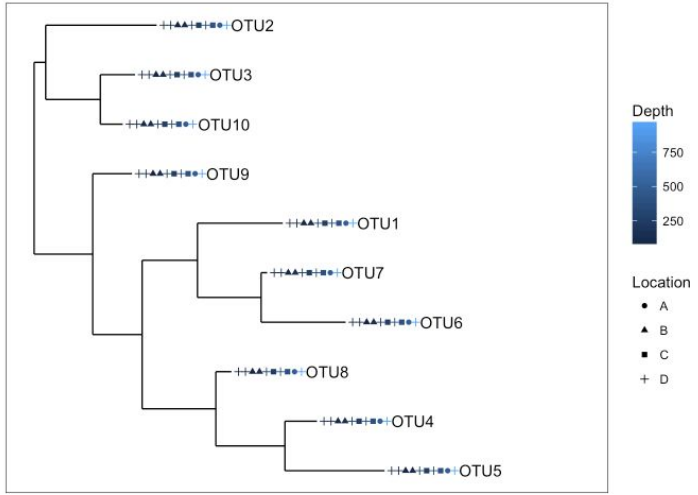
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 10 taxa and 10 samples ]
## sample_data() Sample Data:       [ 10 samples by 2 sample variables ]
## tax_table()   Taxonomy Table:    [ 10 taxa by 7 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 10 tips and 9 internal nodes ]
```

7. Treeplot with new combined data.

```
plot_tree(physeq1, color="Location", label.tips="taxa_names", ladderize="left", plot.margin=
0.3)
```



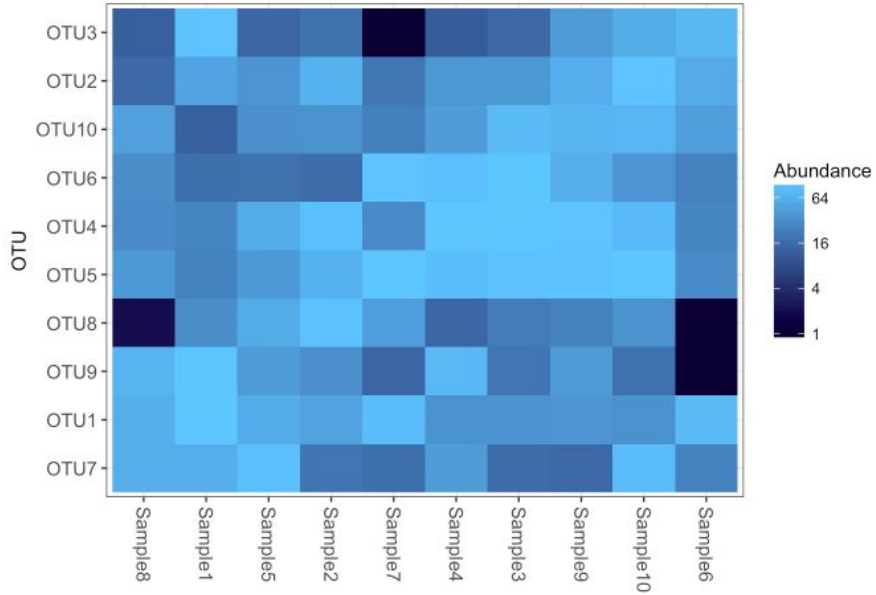

```
plot_tree(physeq1, color="Depth", shape="Location", label.tips="taxa_names", ladderize="righ
t", plot.margin=0.3)
```



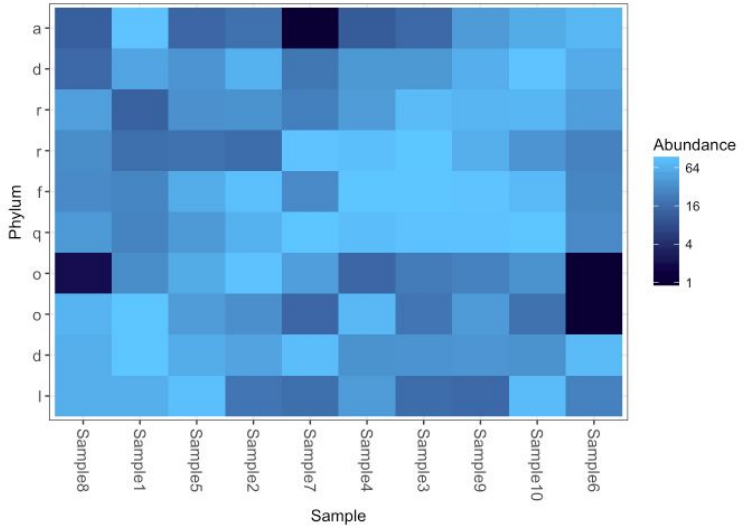
8. Heatmap

Now how about some heatmaps.

```
plot_heatmap(physeq1)
```




```
plot_heatmap(physeq1, taxa.label="Phylum")
```



9. How to save phyloseq_object.data

1. **Save the phyloseq object to an RData file**:
   This method saves the entire `phyloseq` object, including all its components, into a file that can be loaded back into R later.

R Copy code

```R
# Assuming your phyloseq object is called ps
save(ps, file = "phyloseq_object.RData")
```

To load the object back into R:

R Copy code

```R
load("phyloseq_object.RData")
```

9. How to save phyloseq_object.data

R

Copy code

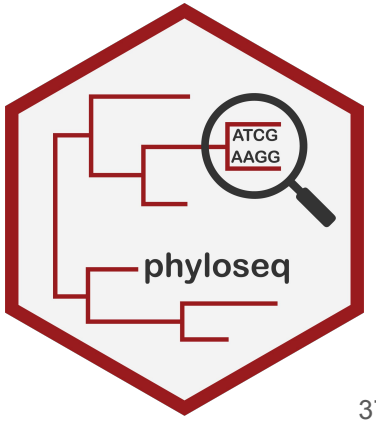
```r
# Save the OTU table
otu_table <- otu_table(ps)
write.csv(as.data.frame(otu_table), file = "otu_table.csv")

# Save the taxonomy table
tax_table <- tax_table(ps)
write.csv(as.data.frame(tax_table), file = "tax_table.csv")

# Save the sample data
sample_data <- sample_data(ps)
write.csv(as.data.frame(sample_data), file = "sample_data.csv")

# Save the phylogenetic tree (if applicable)
phy_tree <- phy_tree(ps)
if (!is.null(phy_tree)) {
  ape::write.tree(phy_tree, file = "phylogenetic_tree.nwk")
}
```

Practice and tutorial



1.Review data use

The ISME Journal (2018) 12:1360–1374
https://doi.org/10.1038/s41396-018-0050-z

ARTICLE

Small eukaryotic phytoplankton communities in tropical waters off Brazil are dominated by symbioses between Haptophyta and nitrogen-fixing cyanobacteria

Catherine Gérikas Ribeiro[1,2] · Adriana Lopes dos Santos[1,3] · Dominique Marie[1] · Frederico Pereira Brandini[2] · Daniel Vaulot[1]

Received: 22 May 2017 / Revised: 1 November 2017 / Accepted: 20 December 2017 / Published online: 9 February 2018

Abstract
Symbioses between eukaryotic algae and nitrogen-fixing cyanobacteria have been recognized in recent years as a key source of new nitrogen in the oceans. We investigated the composition of the small photosynthetic eukaryote communities associated with nitrogen-fixing cyanobacteria in the Brazilian South Atlantic Bight using a combination of flow cytometry sorting and high throughput sequencing of two genes: the V4 region of 18S rRNA and *nifH*. Two distinct eukaryotic communities were often encountered, one dominated by the Mamiellophyceae *Bathycoccus* and *Ostreococcus*, and one dominated by a prymnesiophyte known to live in symbiosis with the UCYN-A1 nitrogen-fixing cyanobacterium. Among *nifH* sequences, those from UCYN-A1 were most abundant but three other UCYN-A clades (A2, A3, A4) were also found. Network analysis confirmed the relation between A1 and A2 clades and their hypothesized hosts and pointed out to the potential association between novel clade A4 with *Braarudosphaera bigelowii*, previously hypothesized to host A2.

https://doi.org/10.1038/s41396-018-0050-z

Introduction

Small photosynthetic eukaryotes [1] are key component of the biomass and primary production in marine ecosystems [2–4]. In coastal waters, these small photosynthetic eukar-

prasinophytes clade VII) but also pelagophytes, chrysophytes, and prymnesiophytes [3, 6, 7]. While for some groups such as Mamiellophyceae, prasinophytes clade VII, or pelagophytes, numerous representatives have been isolated in culture [8], this is not the case for chrysophytes or

2. Objective

This study investigated the **tiny photosynthetic organism**s living alongside **nitrogen-fixing cyanobacteria** in the Brazilian South Atlantic Bight.

Symbioses between **eukaryotic algae** and **nitrogen-fixing cyanobacteria**

They used a combination of cell sorting and gene sequencing to identify these communities.

Use two genes: the **V4 region of 18S** rRNA and **nifH** to sequencing

3. Meta data

Transect	Station	Depth (m)	Picoplankton sorted samples				Nanoplankton sorted samples			
			Sample code	Sorted cells #	18S sequence #	nifH sequence #	Sample code	Sorted cells #	18S sequence #	nifH sequence #
0	6	45	1p	7651	19,466	137,117	1n	4845	95,054	*163*
			2p	7343	107,644	113,897	2n	3258	45,111	*143*
	19	0	3p	1005	134,873	92,500	3n	898	131,031	*116*
	21	0	5p	793	112,590	25,341	5n	660	24,696	77,360
	26	0	7p	907	22,348	26,969	7n	856	40,829	7667
1	81	140	9p	3181	44,610	*53*	9n	1235	19,193	*34*
			10p	3278	47,390	6241	10n	1232	53,230	*36*
	85	110	11p	16,312	31,899	10,201	11n	1615		
	86	105	13p	6366	59,626	11,954	13n	1007	46,001	21,316
	87	105	15p	6189	78,390	*1033*	15n	622	22,468	2678
2	96	5	120p	1150	76,182	23,147	120n	75	70,455	*93*
		30	121p	1737	71,785	23,706	121n	218	52,401	26,838
		50	122p	853	37,364	11,045	122n	234	78,740	15,543
	98	5	125p	3086	55,179	21,461	125n	1300	27,381	14,331
		50	126p	1217	30,406	10,140	126n	782	65,714	16,929
		85	127p	3420			127n	226	60,610	11,493
	101	5	140p	500	46,569	12,301	140n	366	48,126	25,286
		60	141p	1046	64,221	10,428	141n	485	30,081	21,302
		110	142p	641	89,797	17,156	142n	159	85,219	11,753
	106	5	155p	355	50,782	66,172	155n	18	54,162	20,674
		60	156p	1800	43,917	16,093	156n	300	55,065	14,447
		100	157p	6910	51,848	15,204	157n	1152	29,078	15,532
	114	5	165p	728	48,514	39,918	165n	226	50,732	14,706
		60	166p	660	62,897	28,107	166n	578	53,412	24,442
		80	167p	722	49,934	13,971	167n	390	31,424	20,616
	Bloom	0	tri01p	1002	36,576	7772	tri01n	194	14,162	11,792
			tri02p	744	46,889	17,259	tri02n	206	23,906	16,934
			tri03p	600	55,630	15,152	tri03n	218	34,892	21,523

Samples corresponding to sorted photosynthetic picoeukaryote populations are labeled with p and those corresponding to sorted photosynthetic nanoeukaryote populations are labeled with n. Samples with <2000 sequences (in italics) were not considered in the analysis

3. Sample location

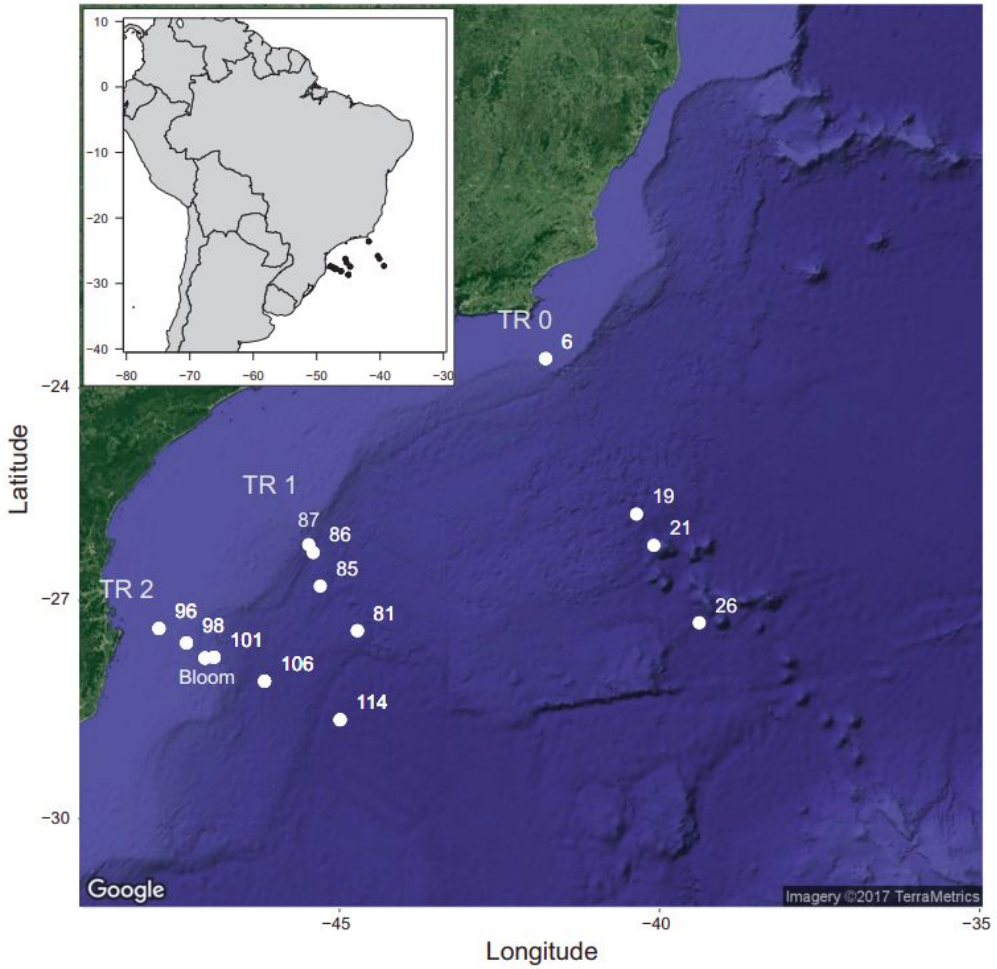
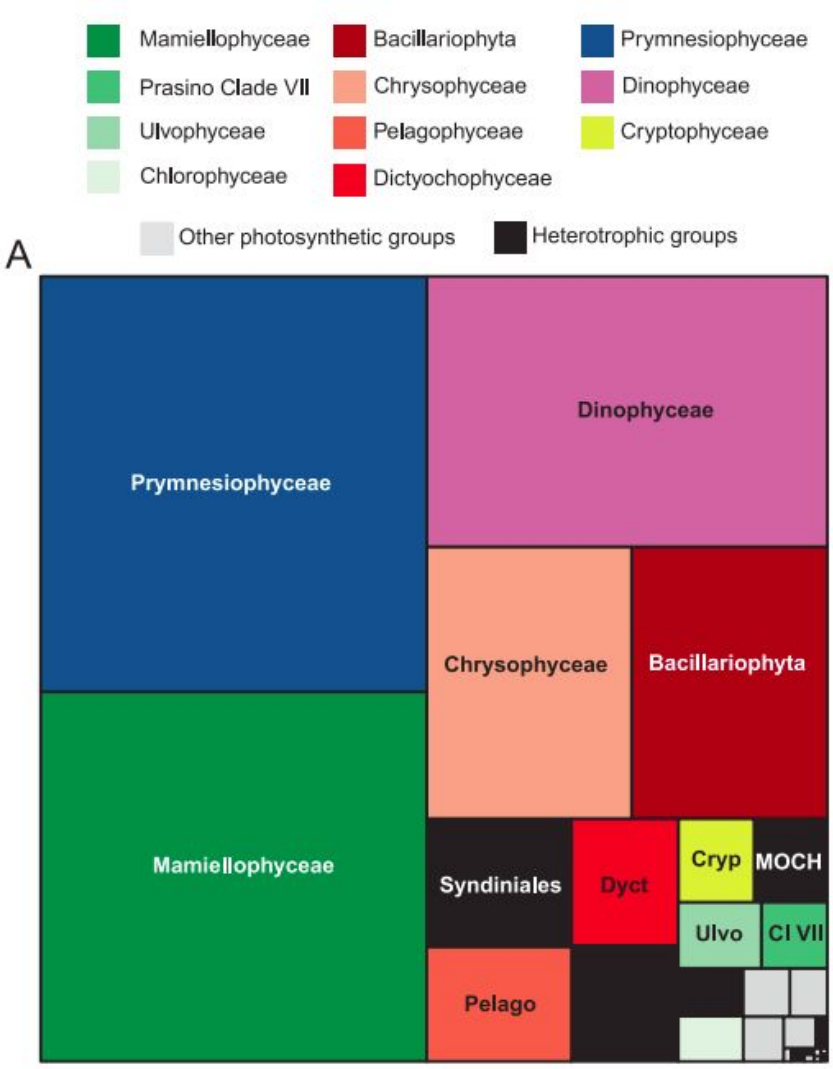
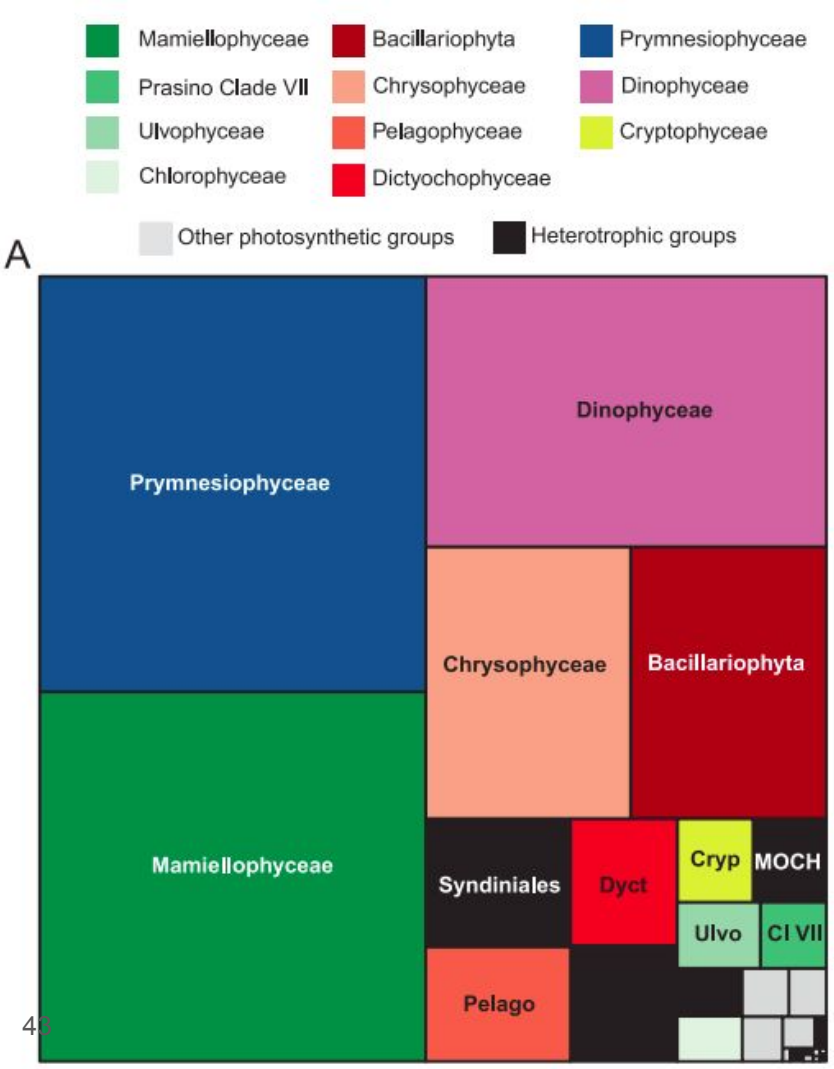


Fig. 1 Map of stations

4. 18S microbiome composition





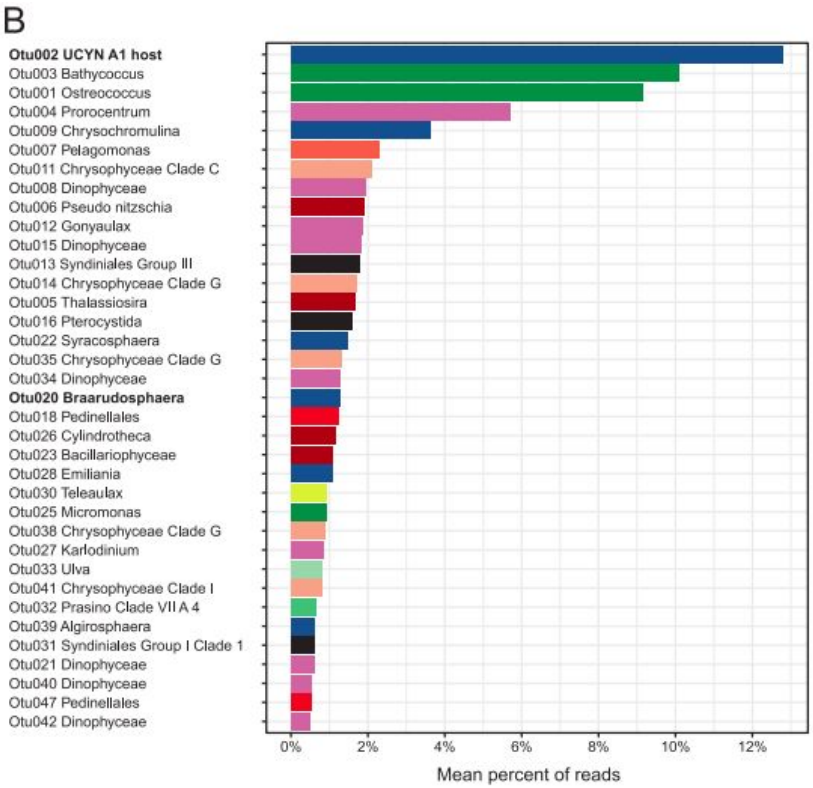
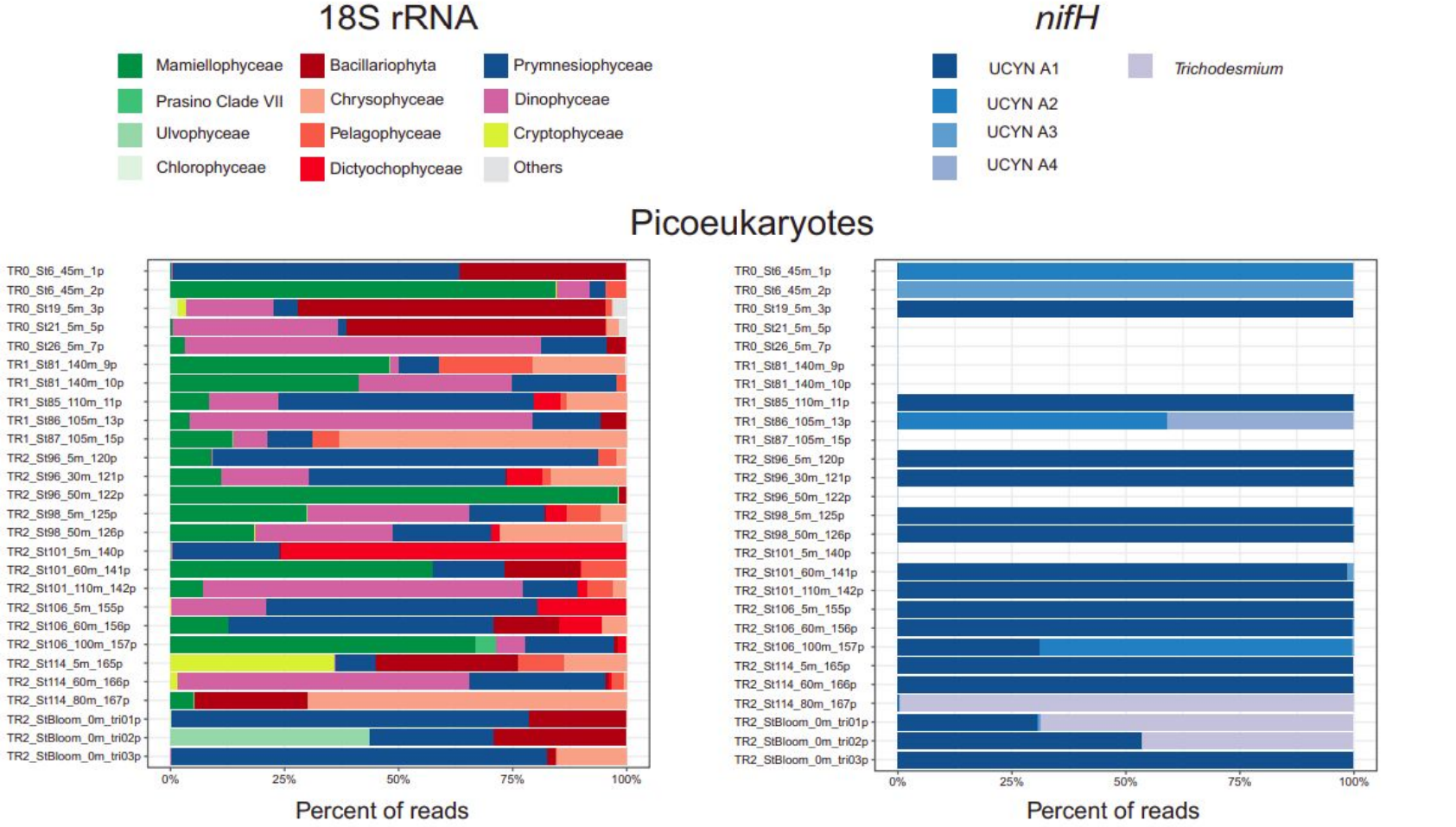


Fig. 2 a Mean relative contribution of each class to 18S rRNA sequences for all sorted samples (pico- and nanoeukaryotes). MOCH: Marine Ochrophyta. **b** Mean relative contribution for major 18S rRNA OTUs clustered at 98% similarity including both autotrophic (colored bars) and heterotrophic (black bars) groups. Major OTUs are defined as those that contribute to more than 20 % of reads in at least one sample

5. Composition between samples



6. Network analysis

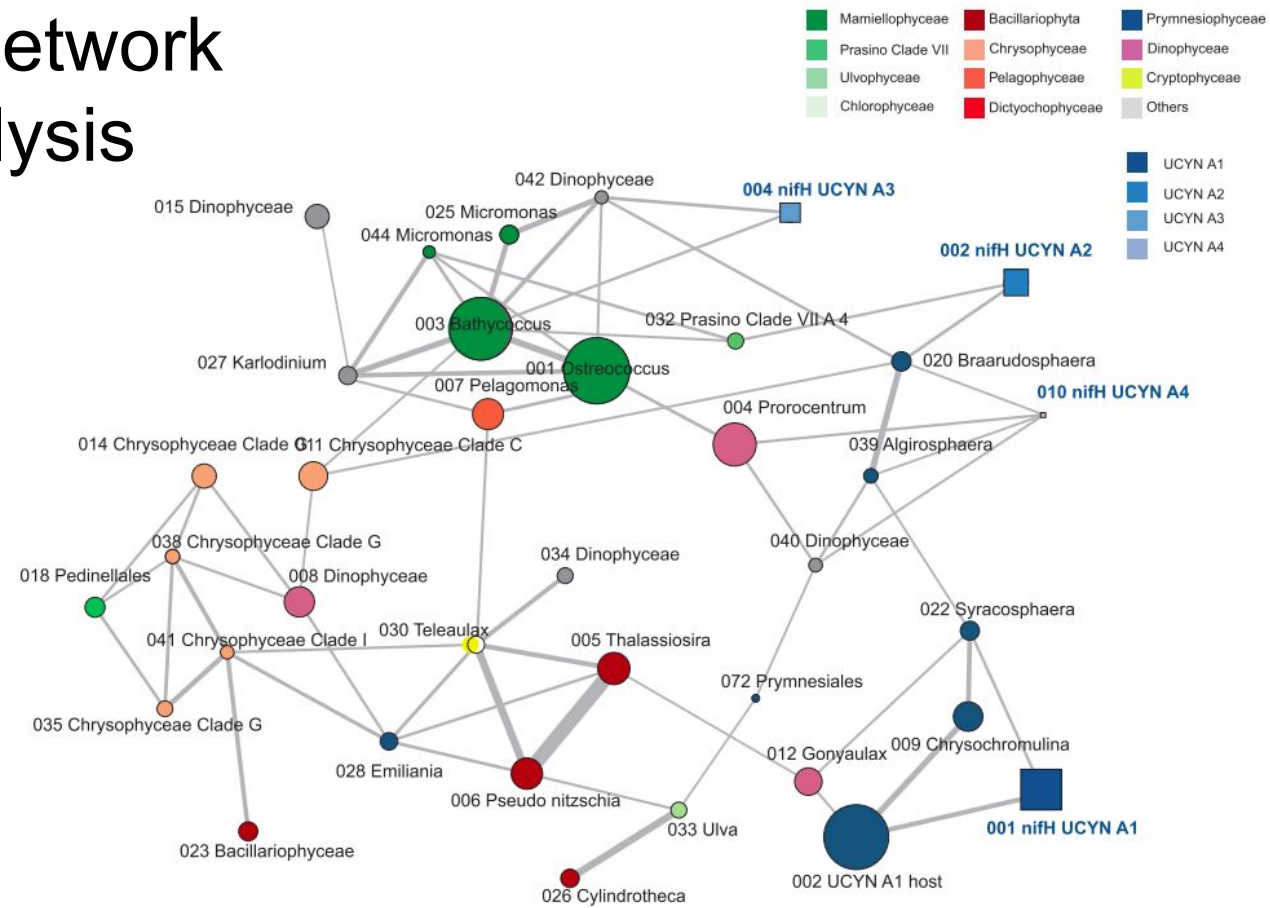
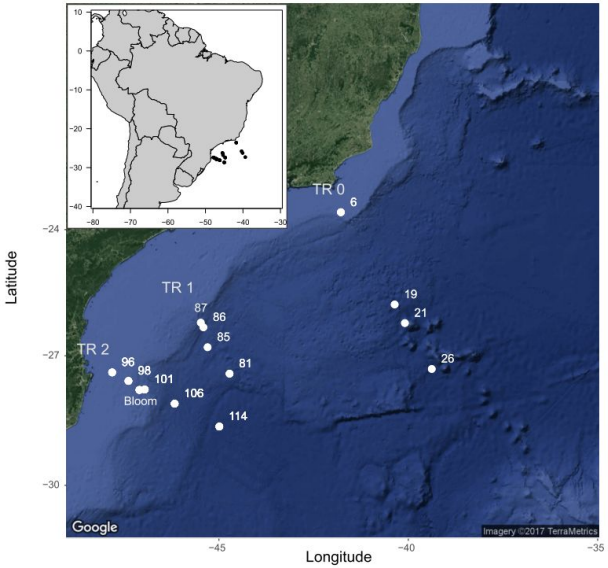
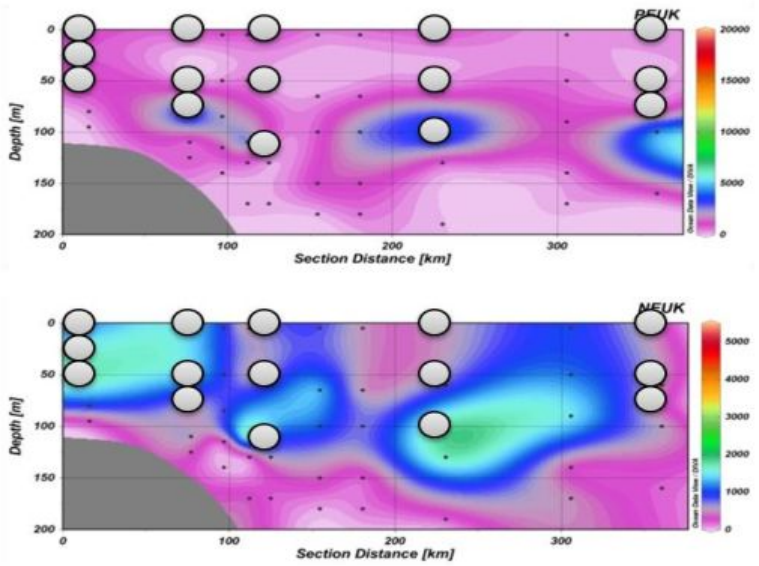


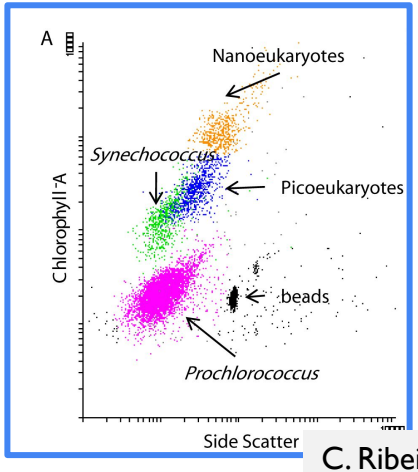
Fig. 7 Network analysis of the major autotrophic 18S rRNA (circles) and *nifH* (squares) OTUs (see Supplementary Table 4 and Supplementary Table 5) using SparCC correlation [43]. Colors of nodes correspond to taxonomic assignation. Size of nodes is proportional to number of reads obtained. Width of edges is proportional to correlation between OTUs. Only correlations >0.20 with pseudo p-values < 0.05 were considered

CARBOM cruise data



18S V4
Illumina 2*250 bp





Sorted samples

- Pico vs Nano
- Surf vs Deep

C. Ribeiro and A. Lopes dos Santos, 2018

Steps

- Import data
- Combine into "phyloseq" object
- Filter the data
  - taxonomy : e.g. only keep photosynthetic eukaryotes
  - samples : remove samples with low reads
  - abundance : remove OTUs with low abundance for some analyses
- Barplots
- Heatmaps
- Alpha diversity
- Multivariate analysis
- Network analysis

Step by step

1. Import data
   +Make sure class of dataset is correct
2. Create phyloseq object
   +Using phyloseq() or merege_phyloseq()
3. Filter data
   +subset_taxa()
   +subset_sample()
4. Normalize number of reads
5. Data visualize

Preparation

- Unzip files from phyloseq_tutorial.zip
- Copy to local or remote directory
- Open R studio either locally or on server
  - http://r.sb-roscoff.fr/

- Open « phyloseq_tutorial.Rmd» with R studio
- Set your working directory where the files are

1 - Otu table

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	otu	X10n	X10p	X11n	X11p	X120n	X120p	X121n	X121p	X122n	X122p	X125n	X125p	X126n	X126p	X127n	X13n	X13p	X140n	X140p	X141n
2	Otu001	13679	6292	42	2500	18850	5	43	7138	9432	10541	9	9772	1388	7	31538	38	2338	23	9	1358
3	Otu002	18	7134	38	9830	45	61420	182	23751	36	11	4535	3502	11018	5473	26	14411	38	19018	12	3080
4	Otu003	9939	8983	31	13	24620	19	19	16	12502	3831	4621	2240	9924	4052	9292	18	0	37	7	3680
5	Otu004	3675	4234	24	22	11	16	32967	35	6	18	6908	5	16	8702	24	11	37717	0	25	4196
6	Otu005	0	5	0	7	0	8	0	16	20166	0	0	2	5	8	2	16	0	13	0	0
7	Otu006	0	8	0	0	0	8	0	0	5	3	3	0	0	9	0	5	4	0	0	3
8	Otu007	4587	518	4	386	8775	5	6	1102	14336	0	0	3626	51	0	6	12	0	10	0	395
9	Otu008	1	8	2	4408	3	29	6	12355	0	0	0	0	0	9	3	1588	0	6	3	3
10	Otu009	115	914	3	325	0	629	1	834	5	0	1354	2108	1117	67	0	2010	1897	11227	1	3
11	Otu010	780	8	23810	12	3279	0	12	7	3027	0	2	4156	0	0	18	0	0	0	0	0
12	Otu011	0	3	2	2	0	13	5	5	4	7	3081	11	4	6804	0	3	11	0	5	0
13	Otu012	0	0	0	6	0	0	0	16	3	0	0	0	0	0	0	17	0	6	0	0
14	Otu013	6321	2471	2	0	12	3	0	0	4	20272	0	15	9	0	5	0	11	0	14	0
15	Otu014	0	82	4	3304	1	1667	4	9233	13	3	0	2707	0	0	3	4806	9	3	5	0
16	Otu015	0	12	0	3	7	25	1	6	10	0	4	2772	1	3	0	2	0	10	13	8052
17	Otu016	1	0	0	9	5	0	0	14	0	0	0	0	2654	0	0	6	1	1	0	0
18	Otu017	0	0	0	0	0	0	0	0	17	8	0	0	0	0	0	17	24	48	35210	4
19	Otu018	1	0	9	911	0	0	15	2702	6	4	342	2217	606	0	13	3846	4	6	8513	1
20	Otu019	0	0	13	0	0	0	29	0	0	0	0	0	0	0	11	0	0	5	4	0
21	Otu020	425	0	1	0	1706	0	8447	1	0	0	0	0	0	26	0	0	3490	0	2620	0
22	Otu021	0	4	0	0	0	10	0	0	0	0	2	0	0	4	0	0	0	0	0	4
23	Otu022	0	0	0	4987	0	0	0	6	90	1	1	524	0	467	0	4	8	6198	0	1
24	Otu023	4	0	1	0	0	3	0	0	0	0	0	0	3351	3	0	3910	1	2	3	0
25	Otu024	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	2	1	0
26	Otu025	69	0	0	0	290	0	0	0	21	0	118	2	9	513	2	0	0	2	0	0
27	Otu026	0	2	0	0	0	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0
28	Otu027	6	2304	0	0	5	0	0	0	57	4	0	14529	9597	2	6	0	0	0	0	0

2 - Taxonomy Table

	A	B	C	D	E	F	G	H	I
1	otu	Domain	Supergroup	Division	Class	Order	Family	Genus	
2	Otu001	Eukaryota	Archaeplastida	Chlorophyta	Mamiellophyceae	Mamiellales	Bathycoccaceae	Ostreococcus	
3	Otu002	Eukaryota	Hacrobia	Haptophyta	Prymnesiophyceae	Prymnesiophyceae_X	Braarudosphaeraceae	UCYN_A1_host	
4	Otu003	Eukaryota	Archaeplastida	Chlorophyta	Mamiellophyceae	Mamiellales	Bathycoccaceae	Bathycoccus	
5	Otu004	Eukaryota	Alveolata	Dinophyta	Dinophyceae	Dinophyceae_X	Dinophyceae_X	Prorocentrum	
6	Otu005	Eukaryota	Stramenopiles	Ochrophyta	Bacillariophyta	Mediophyceae	Mediophyceae_X	Thalassiosira	
7	Otu006	Eukaryota	Stramenopiles	Ochrophyta	Bacillariophyta	Bacillariophyceae	Bacillariophyceae_X	Pseudo_nitzschia	
8	Otu007	Eukaryota	Stramenopiles	Ochrophyta	Pelagophyceae	Pelagophyceae_X	Pelagophyceae_X	Pelagomonas	
9	Otu008	Eukaryota	Alveolata	Dinophyta	Dinophyceae	Dinophyceae_X	Dinophyceae_X	Dinophyceae_X	
10	Otu009	Eukaryota	Hacrobia	Haptophyta	Prymnesiophyceae	Prymnesiales	Chrysochromulinaceae	Chrysochromulina	
11	Otu010	Eukaryota	Opisthokonta	Metazoa	Craniata	Craniata_X	Craniata_XX	Craniata_XX_unclassified	
12	Otu011	Eukaryota	Stramenopiles	Ochrophyta	Chrysophyceae	Chrysophyceae_X	Chrysophyceae_Clade_C	Chrysophyceae_Clade_C_X	
13	Otu012	Eukaryota	Alveolata	Dinophyta	Dinophyceae	Dinophyceae_X	Dinophyceae_X	Gonyaulax	
14	Otu013	Eukaryota	Alveolata	Dinophyta	Syndiniales	Syndiniales_Group_III	Syndiniales_Group_III_X	Syndiniales_Group_III_X	
15	Otu014	Eukaryota	Stramenopiles	Ochrophyta	Chrysophyceae	Chrysophyceae_X	Chrysophyceae_Clade_G	Chrysophyceae_Clade_G_X	
16	Otu015	Eukaryota	Alveolata	Dinophyta	Dinophyceae	Dinophyceae_X	Dinophyceae_X	Dinophyceae_X	
17	Otu016	Eukaryota	Hacrobia	Centroheliozoa	Centroheliozoa_X	Pterocystida	Pterocystida_X	Pterocystida_X	
18	Otu017	Eukaryota	Opisthokonta	Fungi	Basidiomycota	Agaricomycotina	Agaricomycetes	Hyphodontia	
19	Otu018	Eukaryota	Stramenopiles	Ochrophyta	Dictyochophyceae	Dictyochophyceae_X	Pedinellales	Pedinellales_X	
20	Otu019	Eukaryota	Opisthokonta	Fungi	Basidiomycota	Agaricomycotina	Agaricomycetes	Itersonilia	
21	Otu020	Eukaryota	Hacrobia	Haptophyta	Prymnesiophyceae	Prymnesiophyceae_X	Braarudosphaeraceae	Braarudosphaera	
22	Otu021	Eukaryota	Alveolata	Dinophyta	Dinophyceae	Dinophyceae_X	Dinophyceae_X	Dinophyceae_X	
23	Otu022	Eukaryota	Hacrobia	Haptophyta	Prymnesiophyceae	Prymnesiophyceae_X	Prymnesiophyceae_X	Syracosphaera	
24	Otu023	Eukaryota	Stramenopiles	Ochrophyta	Bacillariophyta	Bacillariophyceae	Bacillariophyceae_X	Bacillariophyceae_X	
25	Otu024	Eukaryota	Archaeplastida	Streptophyta	Klebsormidiophyceae	Klebsormidiophyceae_X	Klebsormidiophyceae_XX	Klebsormidium	
26	Otu025	Eukaryota	Archaeplastida	Chlorophyta	Mamiellophyceae	Mamiellales	Mamiellaceae	Micromonas	
27	Otu026	Eukaryota	Stramenopiles	Ochrophyta	Bacillariophyta	Bacillariophyceae	Bacillariophyceae_X	Cylindrotheca	
28	Otu027	Eukaryota	Alveolata	Dinophyta	Dinophyceae	Suessiales	Suessiales_X	Karlodinium	
29	Otu028	Eukaryota	Hacrobia	Haptophyta	Prymnesiophyceae	Isochrysidales	Noelaerhabdaceae	Emiliania	
30	Otu029	Eukaryota	Opisthokonta	Fungi	Ascomycota	Saccharomycotina	Saccharomycetales	Debaryomyces	
31	Otu030	Eukaryota	Hacrobia	Cryptophyta	Cryptophyceae	Cryptophyceae_X	Cryptomonadales	Teleaulax	
32	Otu031	Eukaryota	Alveolata	Dinophyta	Syndiniales	Syndiniales_Group_I	Syndiniales_Group_I_Clade_1	Syndiniales_Group_I_Clade_1_X	
33	Otu032	Eukaryota	Archaeplastida	Chlorophyta	Prasino_Clade_VII	Prasino_Clade_VII_X	Prasino_Clade_VII_A	Prasino_Clade_VII_A_4_X	

3 - Sample table

	A	B	C	D	E	F	G	H	I	J	K	L
1	sample	fraction	Select_18S_nifH	total_18S	total_16S	total_nifH	sample_number	transect	station	depth	latitude	longitude
2	X10n	Nano	Yes	53230	8772	36	10	1	81	140	-27.42	-44.72
3	X10p	Pico	Yes	47390	4448	6241	10	1	81	140	-27.42	-44.72
4	X11n	Nano	No	24007	6193	3772	11	1	85	110	-26.8	-45.3
5	X11p	Pico	Yes	31899	14	10201	11	1	85	110	-26.8	-45.3
6	X120n	Nano	Yes	70455	5292	93	120	2	96	5	-27.39	-47.82
7	X120p	Pico	Yes	76182	53272	23147	120	2	96	5	-27.39	-47.82
8	X121n	Nano	Yes	52401	5958	26838	121	2	96	30	-27.39	-47.82
9	X121p	Pico	Yes	71785	10993	23706	121	2	96	30	-27.39	-47.82
10	X122n	Nano	Yes	78740	11730	15543	122	2	96	50	-27.39	-47.82
11	X122p	Pico	Yes	37364	11817	11045	122	2	96	50	-27.39	-47.82
12	X125n	Nano	Yes	27381	9	14331	125	2	98	5	-27.59	-47.39
13	X125p	Pico	Yes	55179	10419	21461	125	2	98	5	-27.59	-47.39
14	X126n	Nano	Yes	65714	15	16929	126	2	98	50	-27.59	-47.39
15	X126p	Pico	Yes	30406	3	10140	126	2	98	50	-27.59	-47.39
16	X127n	Nano	Yes	60610	9	11493	127	2	98	85	-27.59	-47.39
17	X13n	Nano	Yes	46001	33	21316	13	1	86	105	-26.33	-45.41
18	X13p	Pico	Yes	59626	7217	11954	13	1	86	105	-26.33	-45.41
19	X140n	Nano	Yes	48126	10428	25286	140	2	101	5	-27.79	-46.96
20	X140p	Pico	Yes	46569	10448	12301	140	2	101	5	-27.79	-46.96
21	X141n	Nano	Yes	30081	6394	21302	141	2	101	60	-27.79	-46.96
22	X141p	Pico	Yes	64221	11318	10428	141	2	101	60	-27.79	-46.96
23	X142n	Nano	Yes	85219	23243	11753	142	2	101	110	-27.79	-46.96
24	X142p	Pico	Yes	89797	9553	17156	142	2	101	110	-27.79	-46.96
25	X155n	Nano	Yes	54162	8237	20674	155	2	106	5	-28.12	-46.17
26	X155p	Pico	Yes	50782	7384	66172	155	2	106	5	-28.12	-46.17
27	X156n	Nano	Yes	55065	11371	14447	156	2	106	60	-28.12	-46.17
28	X156p	Pico	Yes	43917	9665	16093	156	2	106	60	-28.12	-46.17
29	X157n	Nano	Yes	29078	4978	15532	157	2	106	100	-28.12	-46.17
30	X157p	Pico	Yes	51848	9139	15204	157	2	106	100	-28.12	-46.17
31	X15n	Nano	Yes	22468	2887	2678	15	1	87	105	-26.22	-45.48
32	X15p	Pico	Yes	78390	13813	1033	15	1	87	105	-26.22	-45.48
33	X165n	Nano	Yes	50732	15337	14706	165	2	114	5	-28.65	-44.99
34	X165p	Pico	Yes	48514	10902	39918	165	2	114	5	-28.65	-44.99
35	X166n	Nano	Yes	53412	3411	24442	166	2	114	60	-28.65	-44.99

Import data

```
otu_mat<- read_excel("../data/CARBOM
data.xlsx", sheet = "OTU matrix")

  tax_mat<-
read_excel("../data/CARBOM
data.xlsx", sheet = "Taxonomy table")

  samples_df <-
read_excel("../data/CARBOM
data.xlsx", sheet = "Samples")
```

head(otu_mat)

A tibble: 6 × 56

otu <chr>	X10n <dbl>	X10p <dbl>	X11n <dbl>	X11p <dbl>	X120n <dbl>	X120p <dbl>	X121n <dbl>	X121p <dbl>	X122n <dbl>
Otu001	13679	6292	42	2500	18850	5	43	7138	9432
Otu002	18	7134	38	9830	45	61420	182	23751	36
Otu003	9939	8983	31	13	24620	19	19	16	12502
Otu004	3675	4234	24	22	11	16	32967	35	6
Otu005	0	5	0	7	0	8	0	16	20166
Otu006	0	8	0	0	0	8	0	0	5

head(tax_mat)

otu <chr>	Domain <chr>	Supergroup <chr>	Division <chr>	Class <chr>
Otu001	Eukaryota	Archaeplastida	Chlorophyta	Mamiellophyceae
Otu002	Eukaryota	Hacrobia	Haptophyta	Prymnesiophyceae
Otu003	Eukaryota	Archaeplastida	Chlorophyta	Mamiellophyceae
Otu004	Eukaryota	Alveolata	Dinophyta	Dinophyceae
Otu005	Eukaryota	Stramenopiles	Ochrophyta	Bacillariophyta
Otu006	Eukaryota	Stramenopiles	Ochrophyta	Bacillariophyta

head(samples_df)

sample <chr>	fraction <chr>	Select_18S_nifH <chr>	total_18S <dbl>	total_16S <dbl>	total_nifH <dbl>
X10n	Nano	Yes	53230	8772	36
X10p	Pico	Yes	47390	4448	6241
X11n	Nano	No	24007	6193	3772
X11p	Pico	Yes	31899	14	10201
X120n	Nano	Yes	70455	5292	93
X120p	Pico	Yes	76182	53272	23147

Import data

```
head(otu_mat)
```

A tibble: 6 × 56

otu <chr>	X10n <dbl>	X10p <dbl>	X11n <dbl>	X11p <dbl>	X120n <dbl>	X120p <dbl>	X121n <dbl>	X121p <dbl>	X122n <dbl>
Otu001	13679	6292	42	2500	18850	5	43	7138	9432
Otu002	18	7134	38	9830	45	61420	182	23751	36

```
otu_mat<- re
data.xlsx", s

 tax_mat<-
read_excel('
data.xlsx", s

 samples_d
read_excel('
data.xlsx", s
```

```
class(otu_mat)
class(samples_df)
class(tax_mat)
```

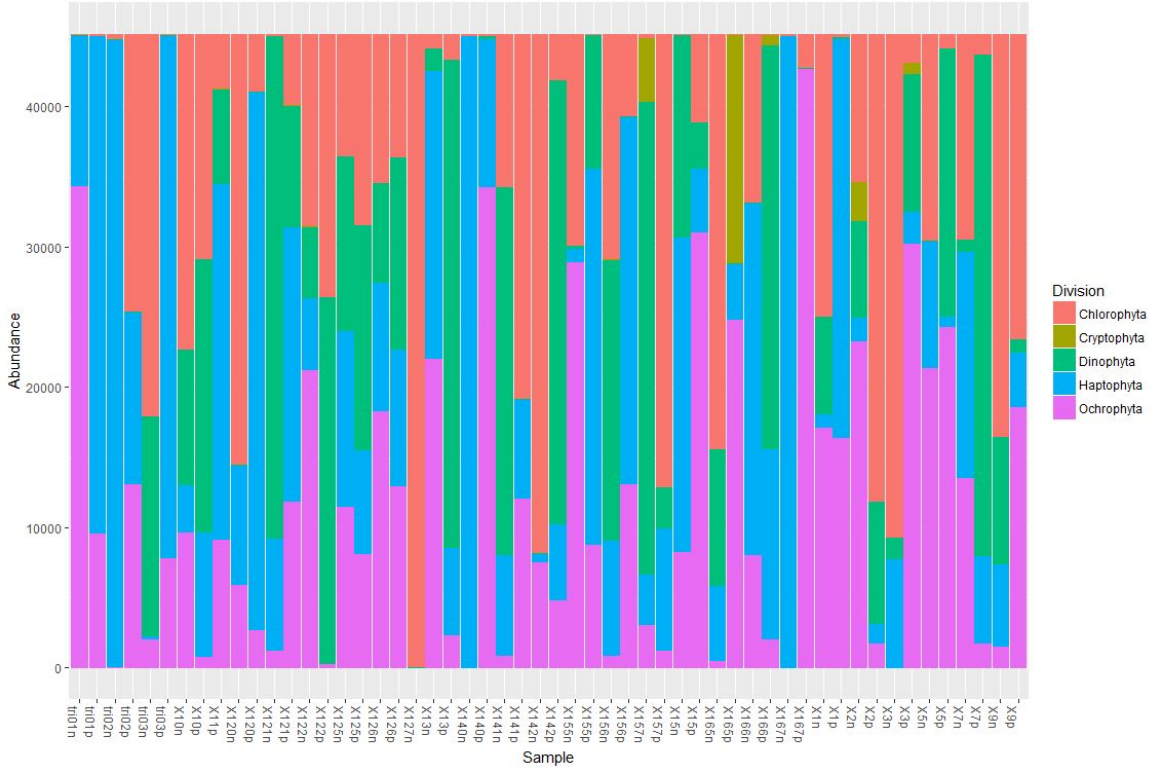
```
[1] "data.frame"
[1] "tbl_df"     "tbl"        "data.frame"
[1] "tbl_df"     "tbl"        "data.frame"
```

X10p	Pico	Yes	47390	4448	6241
X11n	Nano	No	24007	6193	3772
X11p	Pico	Yes	31899	14	10201
X120n	Nano	Yes	70455	5292	93
X120p	Pico	Yes	76182	53272	23147

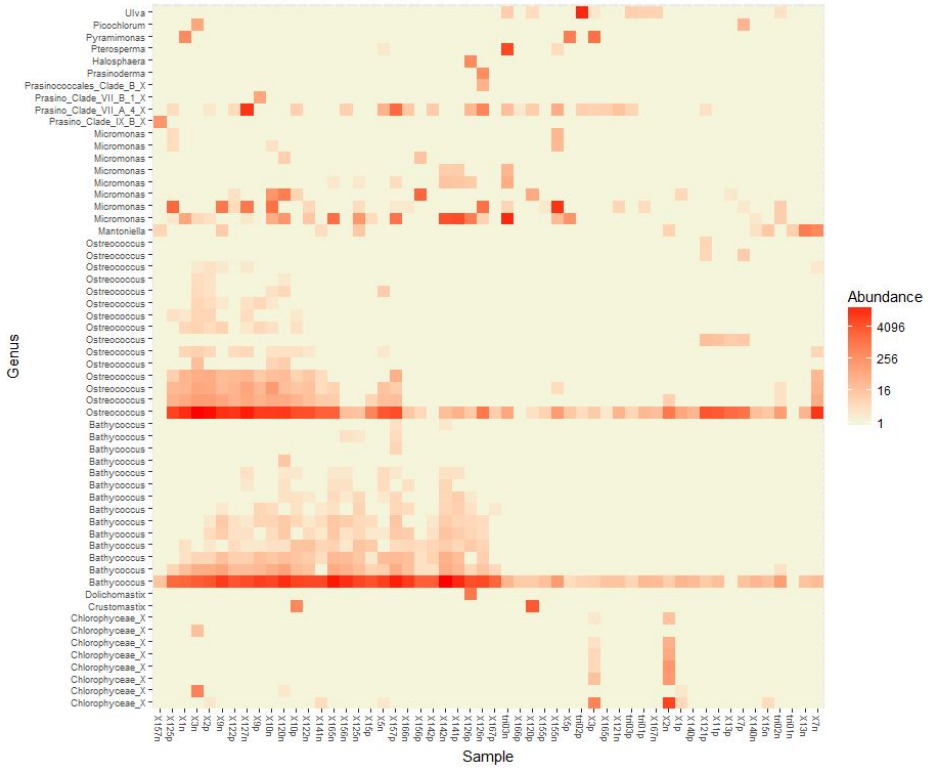
Phyloseq_object

carbom1	S4 (phyloseq::phyloseq)	S4 object of class phyloseq
otu_table	double [205 x 54] (phyloseq:	0 0 256 13339 7 62 0 6 8 7346 0 29 0 ...
tax_table	character [205 x 7] (phylose	'Eukaryota' 'Eukaryota' 'Eukaryota' 'Eukaryota' 'Eukaryota' 'Eukaryota' 'Stramen ...
sam_data	list [54 x 27] (phyloseq::sam	A data.frame with 54 rows and 27 columns
phy_tree	list [4] (S3: phylo)	List of length 4
refseq	NULL	Pairlist of length 0

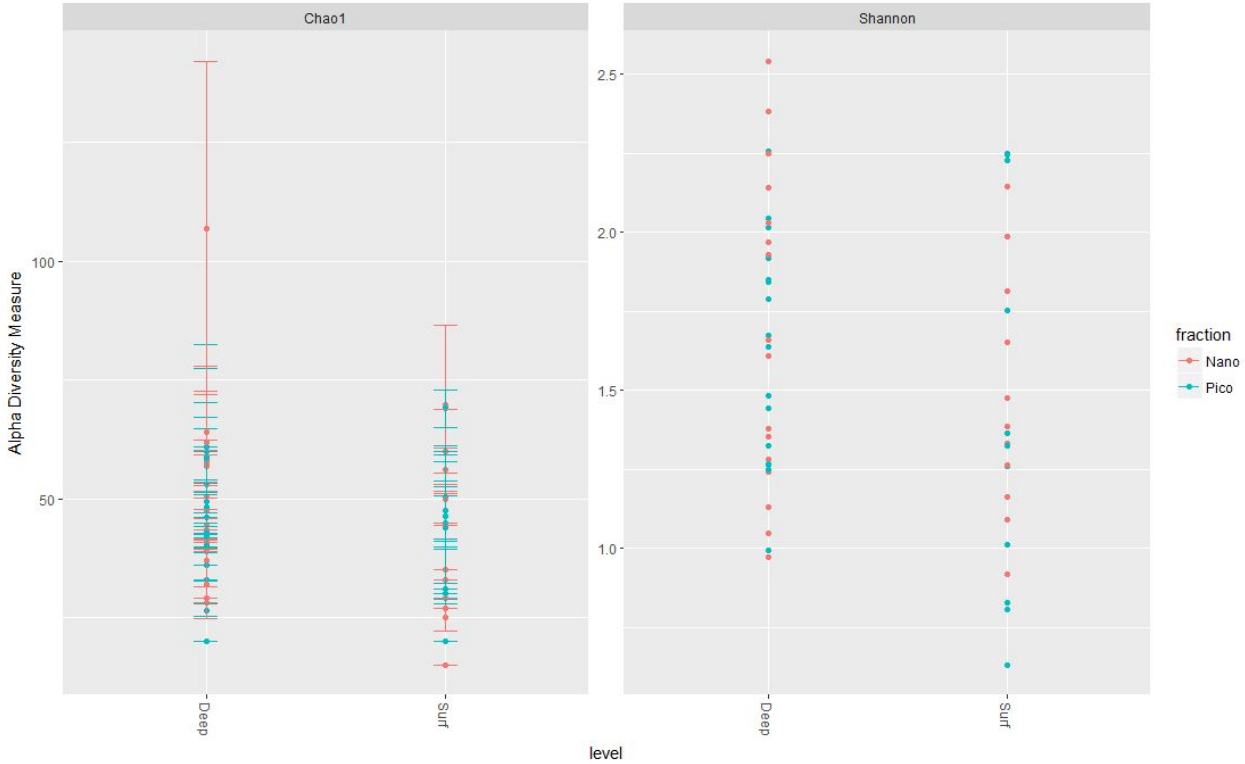
Barplots


Abundance
40000
30000
20000
10000
0
Sample
tri01n
tri01p
tri02n
tri02p
tri03n
tri03p
X10n
X10p
X11p
X120n
X120p
X121n
X121p
X122n
X122p
X125n
X125p
X126n
X126p
X127n
X13n
X13p
X140n
X140p
X141n
X141p
X142n
X142p
X155n
X155p
X156n
X156p
X157n
X157p
X15n
X15p
X165n
X165p
X166n
X166p
X167n
X167p
X1n
X1p
X2n
X2p
X3n
X3p
X5n
X5p
X7n
X7p
X9n
X9p
Division
Chlorophyta
Cryptophyta
Dinophyta
Haptophyta
Ochrophyta

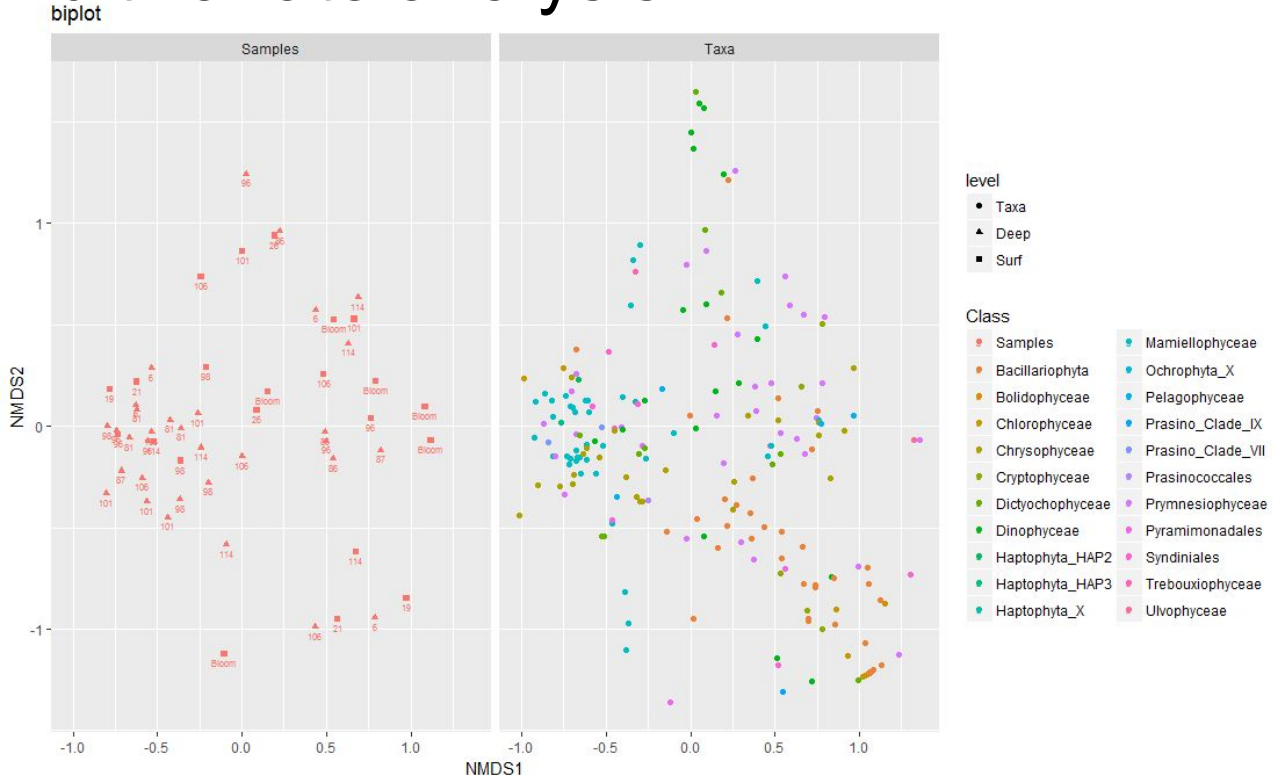
Heatmaps


Genus
Sample
Abundance
4096
256
16
1
Ulva
Picochlorum
Pyramimonas
Pterosperma
Halosphaera
Prasinoderma
Prasinococcales_Clade_B_X
Prasino_Clade_VII_B_1_X
Prasino_Clade_VII_A_4_X
Prasino_Clade_IX_B_X
Micromonas
Mantoniella
Ostreococcus
Bathycoccus
Dolichomastix
Crustomastix
Chlorophyceae_X

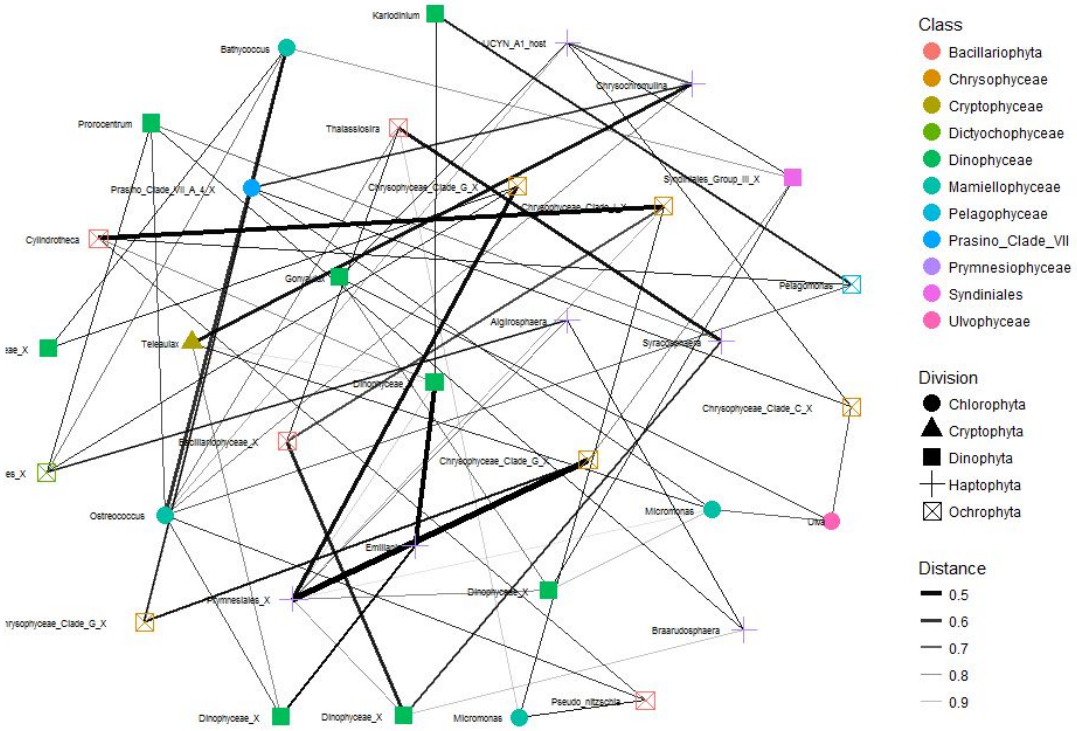
Alpha diversity



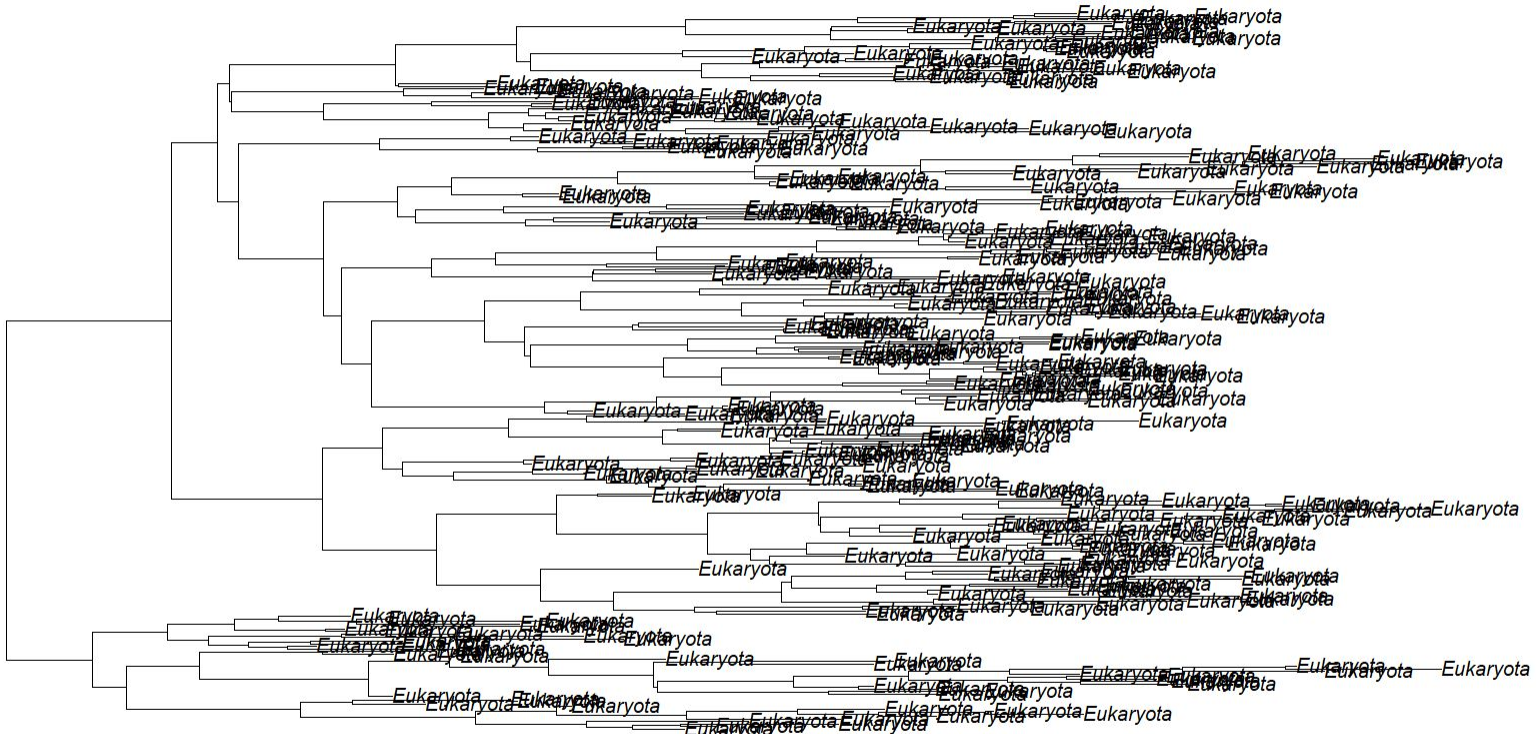
Multivariate analysis



Network analysis



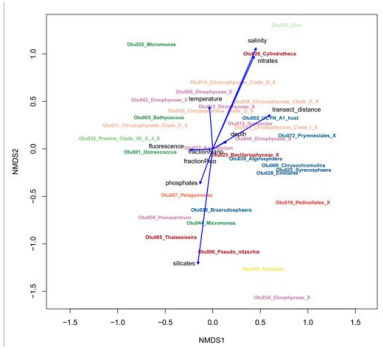
Phylogenetic tree

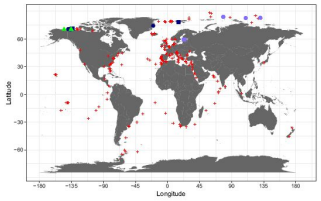


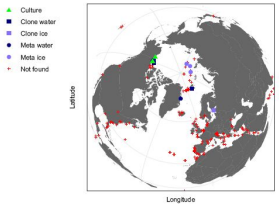
Other useful R packages

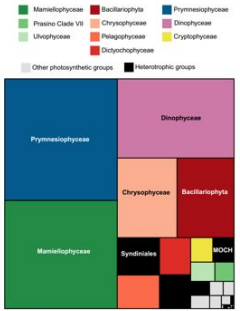
- ggplot2
  - Best package for plotting things
- vegan
  - Multivariate analyses
- dplyr
  - Sorting, filtering, reformating data
- maps
  - Plot the distribution of your data
- treemaps
  - Summary plots

- Tableau
  - Very intuitive software for data exploration









Summary

- Some usefull R function

Import, export data

Data struture and data type

Save, down, read data

- Some R package, specifically designed R package for microbiome data for analysis phylogenetic

What is OTU?

- How to use phyloseq package

Workflow