Community ecology and multivariate analyses

Ramiro Logares
(ICM-CSIC, Barcelona)

Metabarcoding projects

1. Sampling and wet-lab

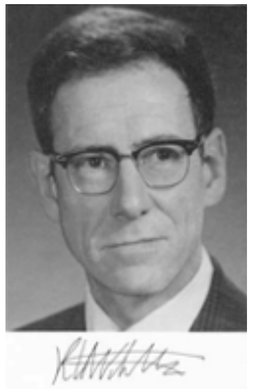
2. Sequencing and sequence processing

3. Ecological analyses: what questions do we want to answer?

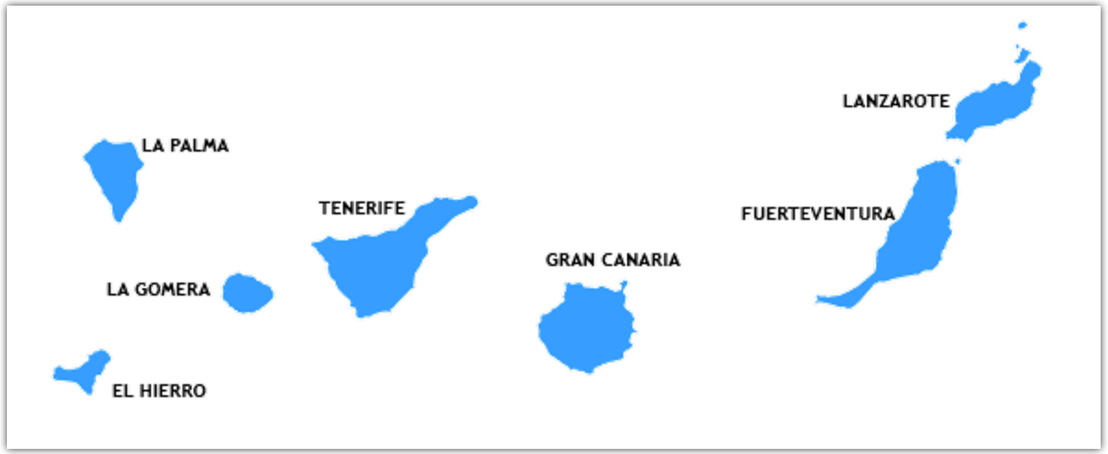
    - Alpha and Beta diversity, community structure and turnover, phylogenetic diversity, etc.

Diversity

- **Alpha**
  - **Richness**: number of species in a location / sample
  - **Evenness**: relative species abundance in a location / sample
- **Beta**
  - Species turnover across locations / time points / samples
- **Gamma**
  - Species in all analyzed locations / samples



Robert Whittaker

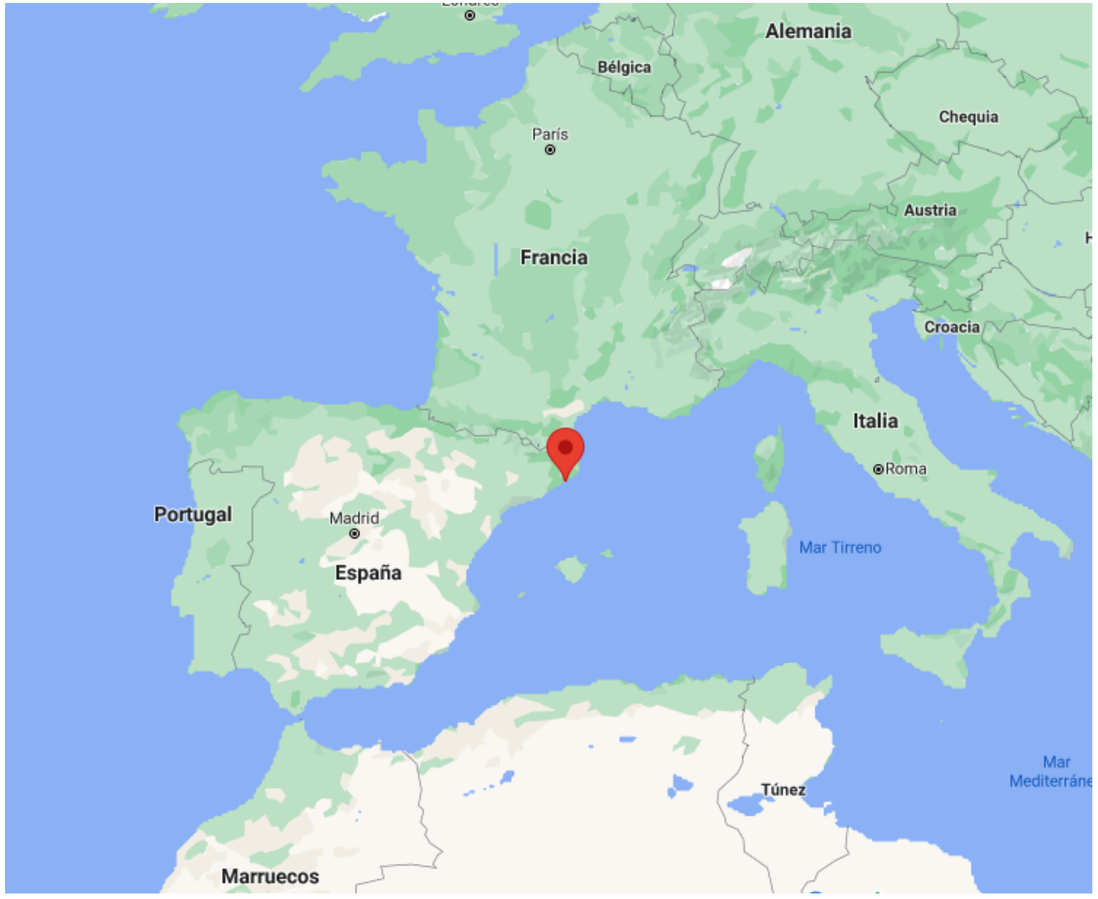


- **Alpha diversity:** number of species in each island
- **Beta diversity:** species change between islands
- **Gamma diversity:** species in all islands

Toy dataset

- Samples of the marine microbiome
  - Blanes Bay Microbial Observatory
- Community 18S rRNA gene (protists)
- 8 samples
  - January, April, July & October of 2004 and 2005

Blanes Bay Microbial Observatory



```
#########################
## Community ecology
#########################

# Full tutorial (a few things are not explained here due to time constraints):
https://github.com/krabberod/BIO9905MERG1_V25/blob/main/Lectures/community.ecology.intro/Comm.Ecology.R.BIO9905MERG1_V25.R

# Install packages (in case you didn't before)

install.packages("vegan")     # Community ecology functions
library(vegan)
# Several more packages are indicated in the tutorial

# Read dada2 otuput

otu.tab<-read_tsv("https://raw.githubusercontent.com/krabberod/BIO9905MERG1_V21/main/Dada2_Pipeline/dada2_results/OTU_table.tsv")

head(otu.tab)
names(otu.tab)

dim(otu.tab) # 2107   26
```

```
#Let's reorder the table
otu.tab<-otu.tab[,c(17,19:26,1:16,18)]

#We assign to rownames the OTU names

otu.tab <- column_to_rownames(otu.tab, var = "OTUNumber") # %>% as_tibble()

rownames(otu.tab)
dim(otu.tab) # 2107   25

otu.tab.simple<-otu.tab[,1:8] # We'll need this table for community ecology analyses

#We transpose the table, as this is how Vegan likes it


otu.tab.simple<-t(otu.tab.simple)
otu.tab.simple[1:5,1:5]


#          OTU_00001 OTU_00002 OTU_00004 OTU_00005 OTU_00006
# BL040126      4996     12348     11426         0      3958
# BL040419       739       684        97     16605      4702
# BL040719         0         0       166         0       806
# BL041019        78        74         0       184       286
# BL050120     30697     12885      5417         0      3739
```

Alpha diversity

Number of species in specific samples/location

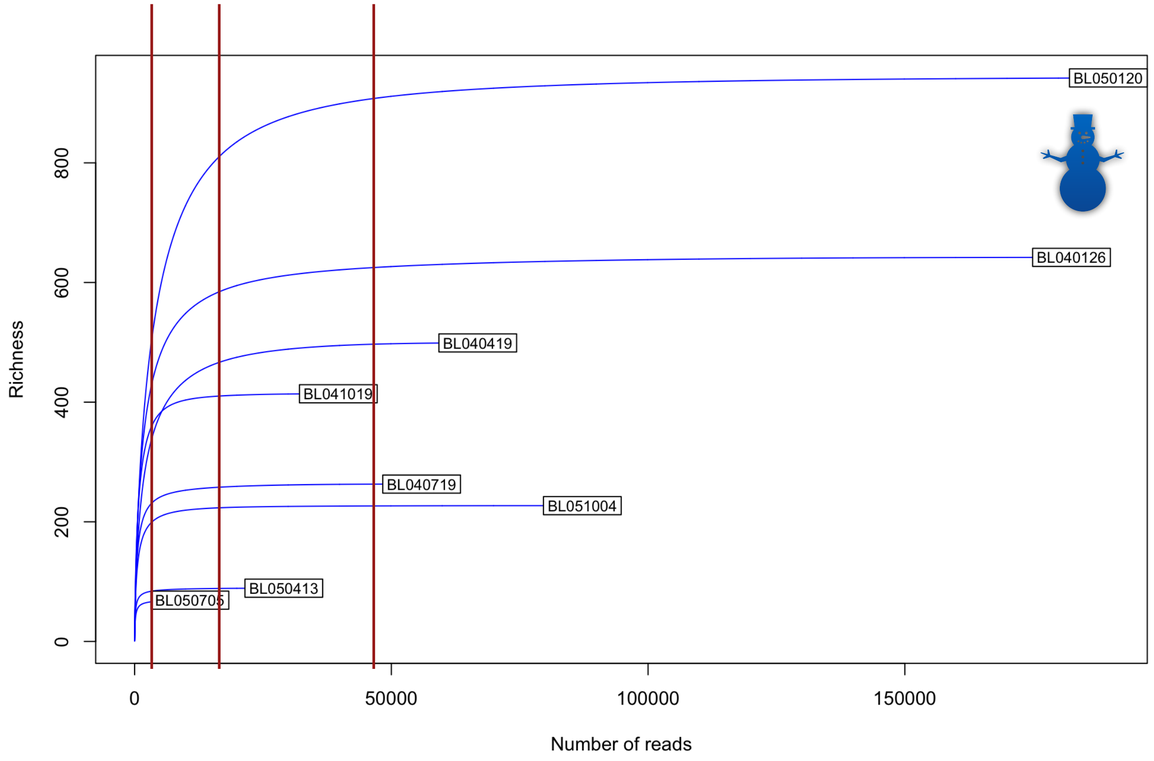
Richness estimates

```
richness<-estimateR(otu.tab.simple)

#             BL040126   BL040419   BL040719   BL041019   BL050120  BL050413  BL050705   BL051004
# S.obs     642.000000 499.000000 263.000000 414.000000 942.000000 89.000000 69.000000 227.000000
# S.chao1   642.000000 499.000000 263.000000 414.000000 943.250000 89.000000 69.000000 227.000000
# se.chao1    0.000000   0.000000   0.000000   0.000000   1.621617  0.000000  0.000000   0.000000
# S.ACE     642.000000 499.000000 263.000000 414.000000 943.399653 89.000000 69.000000 227.000000
# se.ACE      7.091415   9.573887   4.198497   6.011262  10.391615  2.539574  2.797514   2.604638

# Above are the estimators Chao and ACE and the species number.
```

Are we recovering all diversity?

Richness

```
# Rarefaction

#Let's calculate the number of reads per sample

rowSums(otu.tab.simple)

# BL040126 BL040419 BL040719 BL041019 BL050120 BL050413 BL050705 BL051004
#   182462    66827    55896    39672   189636    29053    10771    87192


rarecurve (otu.tab.simple, step=100, xlab= "Number of reads", ylab="Richness", col="blue")
```



What are these results telling us?

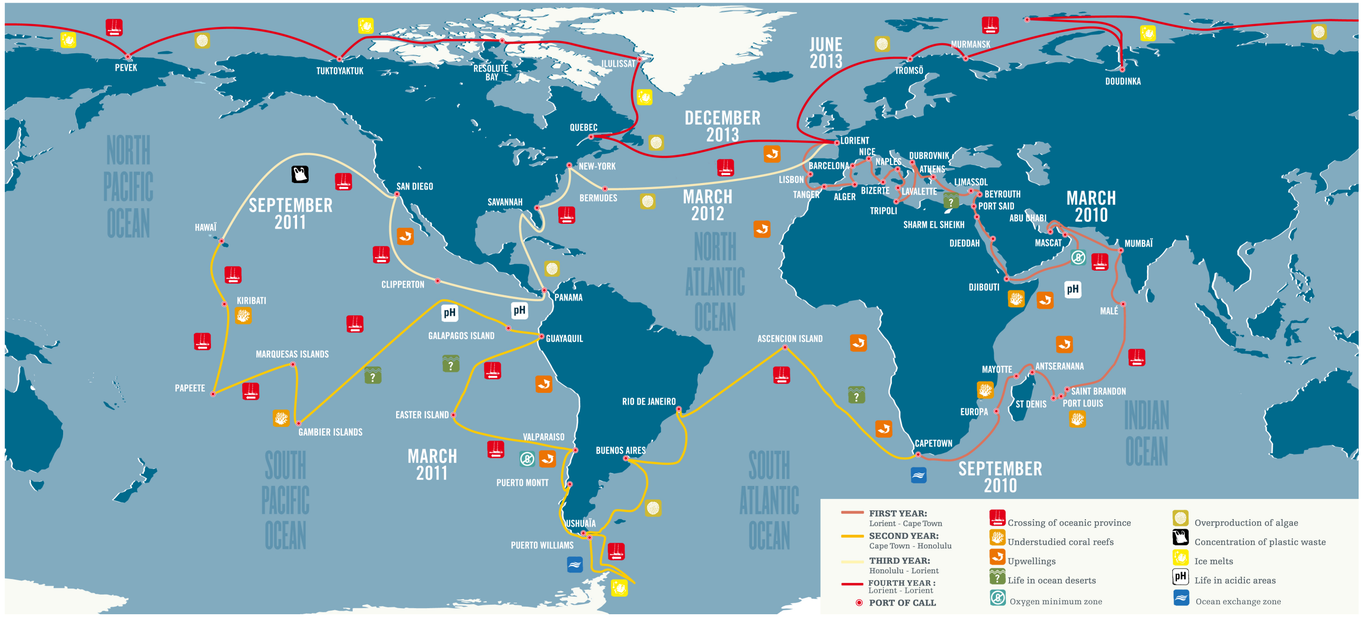
RESEARCH ARTICLE

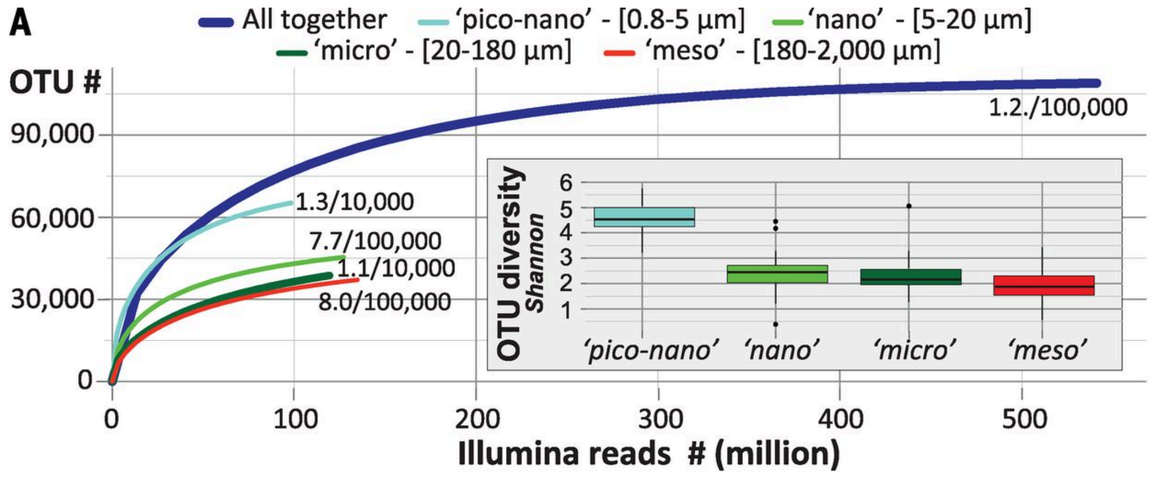
Eukaryotic plankton diversity in the sunlit ocean

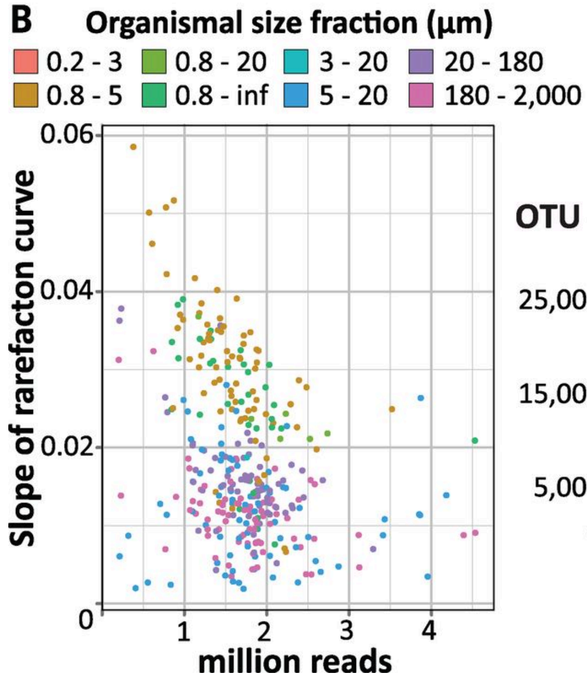
Colomban de Vargas[1,2,*,†], Stéphane Audic[1,2,†], Nicolas Henry[1,2,†], Johan Decelle[1,2,†], Frédéric Mahé[3,1,2,†], Ramiro Logares...

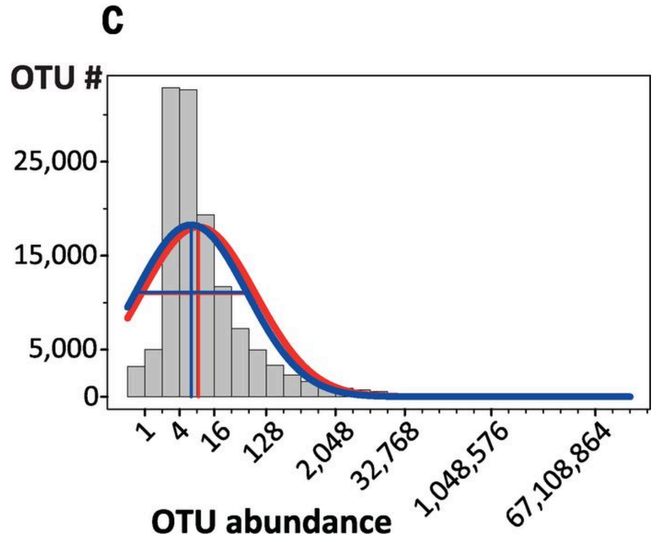
+ See all authors and affiliations

Science 22 May 2015:
Vol. 348, Issue 6237, 1261605
DOI: 10.1126/science.1261605





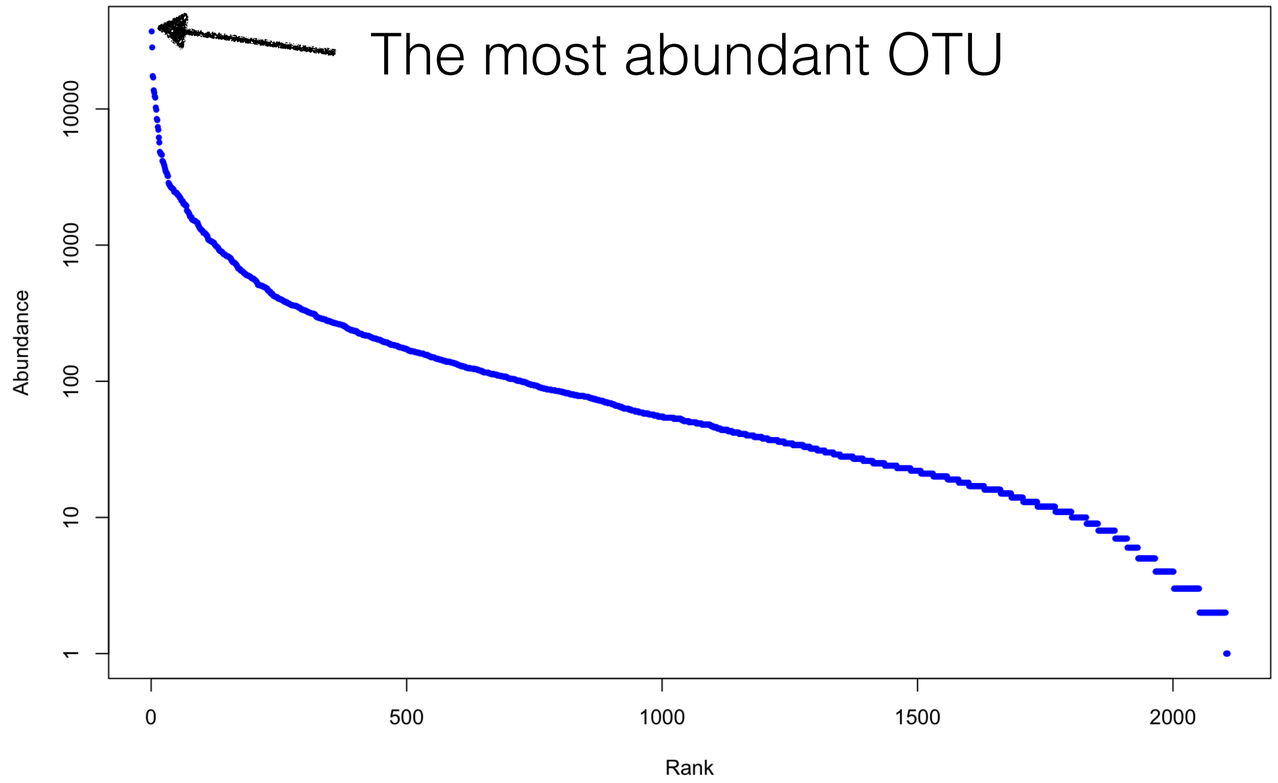




18S V9 (Swarm)

Evenness

```
#Evenness

plot(colSums(otu.tab.simple),log="y",xlab="Rank", ylab="Abundance", pch=19,
     cex=0.5, col="blue")
```



Few species highly abundant, while most species have a low abundance

Characteristic of microbiota
Why?

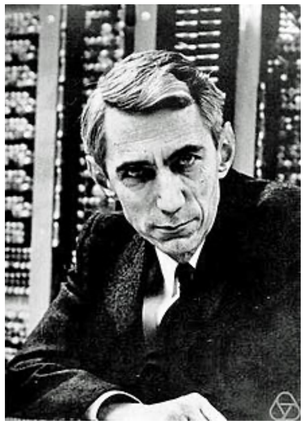
Shannon H index

-Considers richness and evenness

-Originally proposed by Claude Shannon in 1948 to quantify the entropy in strings of text.

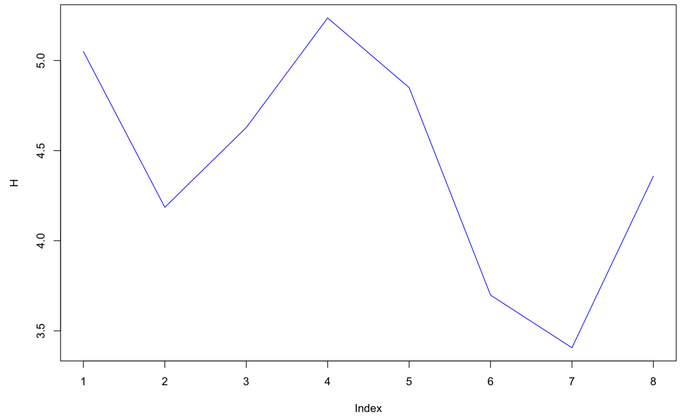
$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

p_i is the relative abundance of the ith species



Claude Shannon

```
#Shannon H index (considers richness and evenness)

H<-diversity(otu.tab.simple, index="shannon")

# BL040126 BL040419 BL040719 BL041019 BL050120 BL050413 BL050705 BL051004
# 5.049747 4.185494 4.627698 5.236017 4.849669 3.698185 3.406164 4.358232

plot(H, type="l", col="blue")
```



Pielou's index of evenness

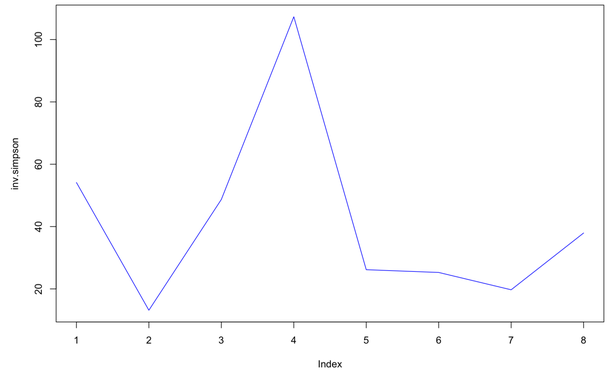
E.C. Pielou

$$J' = \frac{H'}{H'_{\max}}$$

Inverse Simpson's *D* index

Edward Simpson

$$\frac{1}{\lambda} = \frac{1}{\sum_{i=1}^{R} p_i^2} = {}^2D$$




```
#Pielou's index of evenness (range 0-1, 1 = maximum evenness)

# J=H/Hmax
# J=Shannon (H) / log(S=species richness)

J=H/log(rowSums(otu.tab.simple>0))

#  BL040126  BL040419  BL040719  BL041019  BL050120  BL050413  BL050705  BL051004
# 0.7811398 0.6737098 0.8305043 0.8689236 0.7081871 0.8238995 0.8044587 0.8033681


# Inverse Simpson's D index (richness+evenness. Larger values, larger diversity)

inv.simpson<-diversity(otu.tab.simple, "invsimpson")
plot(inv.simpson, type="l", col="blue")

# BL040126  BL040419  BL040719  BL041019  BL050120  BL050413  BL050705  BL051004
# 54.13768  13.15796  48.69382 107.30411  26.16040  25.27907  19.71550  37.93128
```

Beta diversity

Species change between samples / time points / locations

- Beta diversity analyses will investigate how communities change over different samples (time points, locations, etc.)

- Analyses can be biased if different samples have different sequencing efforts

- Different sequencing depths may bias the calculation of distances for multivariate analyses
    - One way to mitigate this is to subsample or “rarefy” samples to the same sequencing depth
        - However, it has been criticized due to the loss of information (but see below…)
    - Anyway, let’s try rarefying the samples to the same sequencing depth

```
#We rarefy all samples to the same sequencing depth to reduce biases

min(rowSums(otu.tab.simple)) # We calculate the sample with the minimum
amount of reads
# [1] 10771

otu.tab.simple.ss<-rrarefy(otu.tab.simple, 10771) #Samples are rarefied to
10771 reads per sample

rowSums(otu.tab.simple.ss) # We check the number of reads per sample

# BL040126 BL040419 BL040719 BL041019 BL050120 BL050413 BL050705 BL051004
#  10771    10771    10771    10771    10771    10771    10771    10771


#Check the dimensions of the tables
dim(otu.tab.simple)
# [1]    8 2107

dim(otu.tab.simple.ss)
# [1]    8 2107
```

```
#Tables have the same size, but after removing reads, several OTUs are
  left with cero abundance

length(which(colSums(otu.tab.simple)==0))
# [1] 0 #No OTU has an abundance sum that is 0, as expected


length(which(colSums(otu.tab.simple.ss)==0))
#  [1] 273 # A total of 273 OTUs were found in the rarefied table with cero
            abundance. Let's corroborate


which(colSums(otu.tab.simple.ss)==0) # Show the OTUs that have 0 abundance
 and their position in the table

# A small subsample of them

# OTU_00814 OTU_01076 OTU_01077 OTU_01232 OTU_01242
#    772       1020      1021       1166      1176
```

```
otu.tab.simple[,772] # This gives the abundance of the OTU_00814 across the different samples in the
                       table that is NOT subsampled


 # BL040126 BL040419 BL040719 BL041019 BL050120 BL050413 BL050705 BL051004
 #    0         0         0         0       88         0         0         0

otu.tab.simple.ss[,772] #  This gives the abundance of the OTU_00814 across the different samples in
                           the table that IS subsampled


 # BL040126 BL040419 BL040719 BL041019 BL050120 BL050413 BL050705 BL051004
 # 0         0         0         0         0         0         0         0
```

```
otu.tab.simple.ss.nocero<-otu.tab.simple.ss[,-(which(colSums(otu.tab.simple.ss)==0))]
 # Removes OTUs with cero abundance


 length(which(colSums(otu.tab.simple.ss.nocero)==0)) # Check that no cero abundance OTUs are left
 # [1] 0 # correct


 # Let's check dimensions

 dim(otu.tab.simple.ss)
 # [1]    8 2107


 dim(otu.tab.simple.ss.nocero)
 # [1]    8 1834

 # 2107-1834 = 273, This is the number of OTUs that we expected to be removed.
```

Distance metrics

- Statistical distance: distance between variables
- *Distance metrics in ecology: allow measuring the dissimilarity between communities composed by several species (OTUs)*
- Several distance metrics available in R
- Often used: Bray Curtis, Euclidean, Jaccard, Sorensen, Simpson

```
Distance metrics available in Vegan
"manhattan", "euclidean", "canberra", "clark", "bray", "kulczynski", "jaccard", "gower",
"altGower", "morisita", "horn", "mountford", "raup", "binomial", "chao", "cao", "mahalanobis",
"chisq" or "chord".
```

Bray Curtis distances for the rarefied datasets

```
# Distance metrics
# We calculate the Bray Curtis dissimilarities for the rarefied dataset


otu.tab.simple.ss.nozero.bray<-vegdist(otu.tab.simple.ss.nozero, method="bray")

as.matrix(otu.tab.simple.ss.nozero.bray)[1:5,1:5]


#            BL040126  BL040419  BL040719  BL041019  BL050120
# BL040126 0.0000000 0.8087457 0.9264692 0.8720639 0.5661498
# BL040419 0.8087457 0.0000000 0.9017733 0.8754062 0.8352985
# BL040719 0.9264692 0.9017733 0.0000000 0.7490484 0.9118002
# BL041019 0.8720639 0.8754062 0.7490484 0.0000000 0.8183084
# BL050120 0.5661498 0.8352985 0.9118002 0.8183084 0.0000000
```

Ordination

In ecological terms: ordination serves to summarise community data (such as species abundance data) by producing a low-dimensional ordination space in which *similar species and samples are plotted close together, and dissimilar species and samples are placed far apart*.

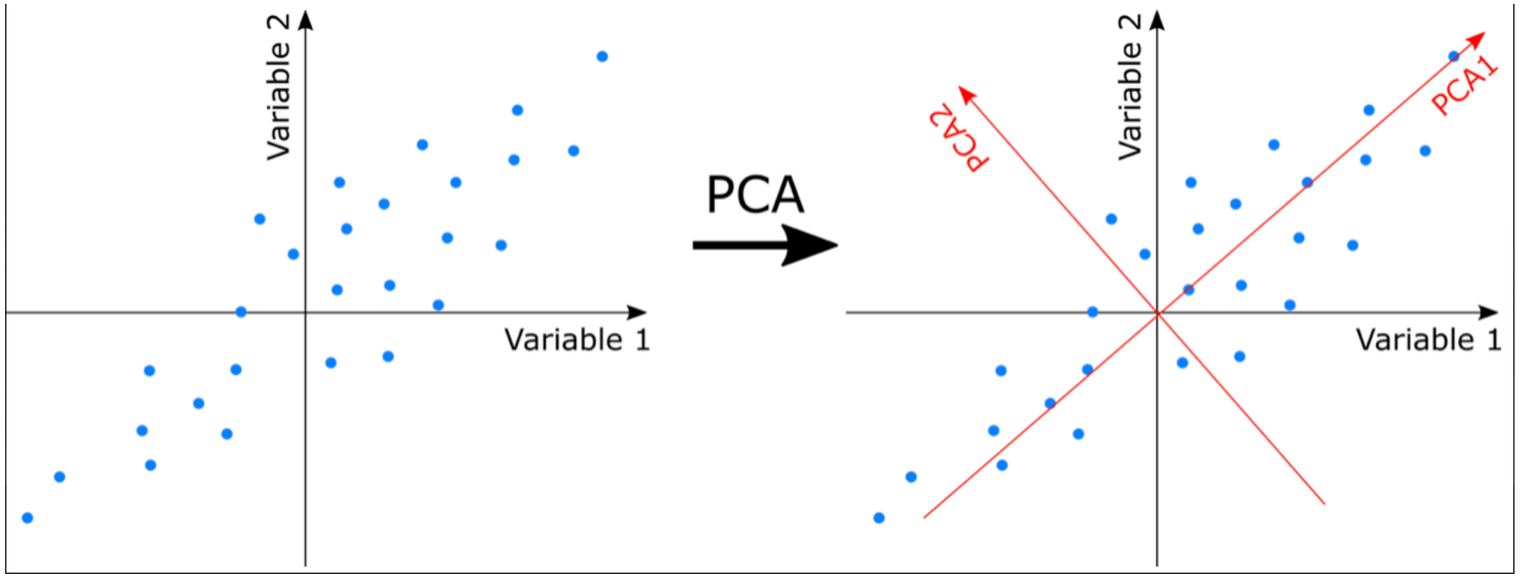
Ordination approaches

Two commonly used unconstrained techniques

- Principal Component Analysis (PCA)
- Non-metric Multidimensional Scaling (NMDS)

PCA

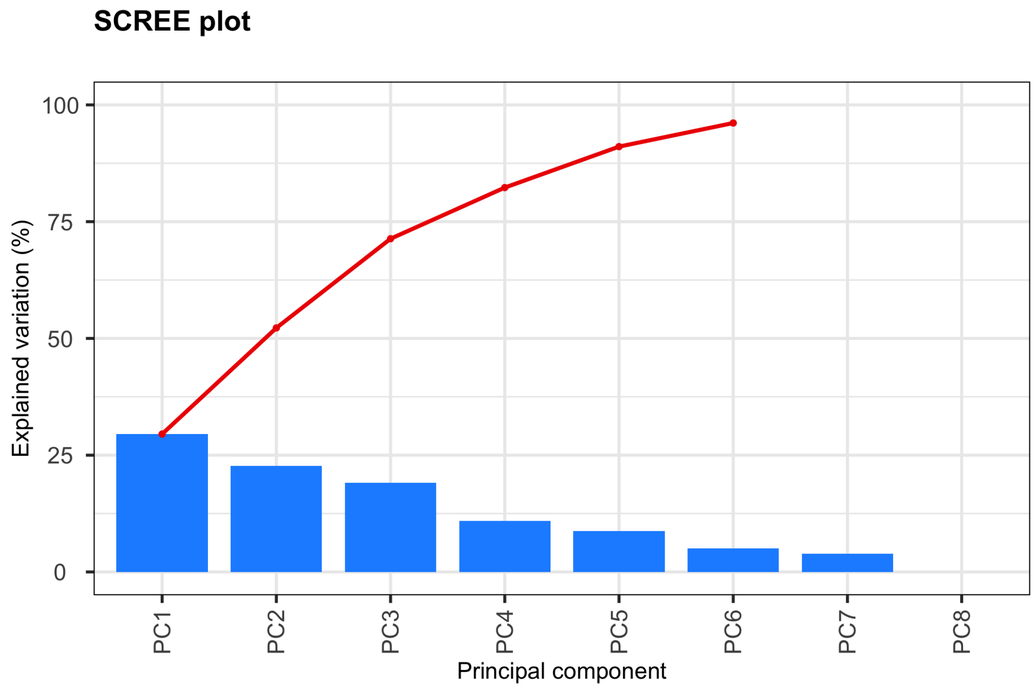
Rotates the original axes in order to maximise the 2D variability. The first principal component (PC) will be placed in the direction of the maximum variability and subsequent PCs will be generated in the same manner



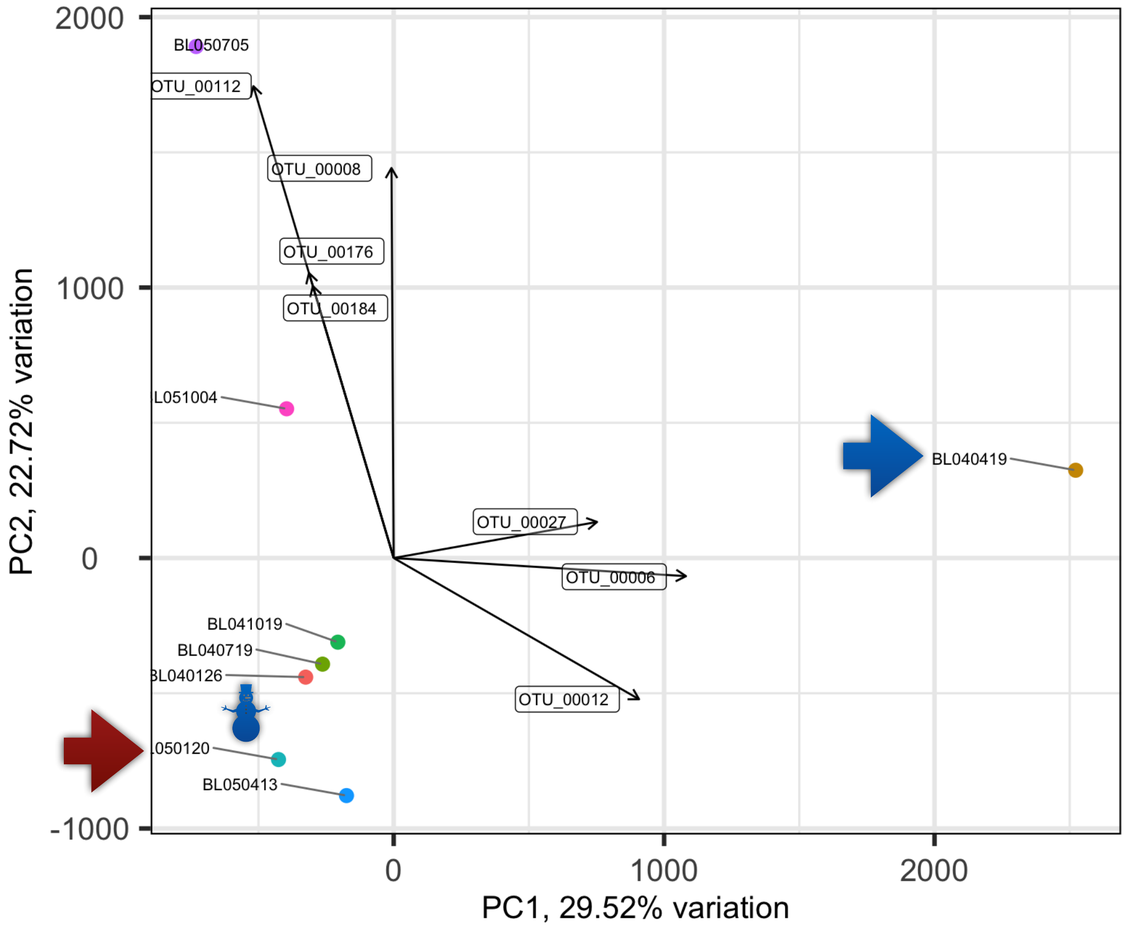
PCA

```
#Ordination and clustering

#PCA

# We install PCAtools
if (!requireNamespace('BiocManager', quietly = TRUE))
  install.packages('BiocManager')

BiocManager::install('PCAtools')

library(PCAtools)

#PCA rarefied table

otu.tab.simple.ss.nozero.pca<-pca(t(otu.tab.simple.ss.nozero), scale=FALSE) #
Runs de PCA

biplot(otu.tab.simple.ss.nozero.pca, showLoadings = T,
       lab=rownames(otu.tab.simple.ss.nozero)) # Plots de PCA

screeplot(otu.tab.simple.ss.nozero.pca, axisLabSize = 18, titleLabSize = 22)

# We plot the percentage of variance explained by each axis
```

PCA



Percentage of variance explain by each PC



Samples and OTUs are plotted. The arrows indicate the weight of each OTU in the different directions

Non-metric Multidimensional Scaling (NMDS)

- NMDS is more robust than PCA (e.g., is not affected by the arch effect)
- NMDS attempts to represent the pairwise dissimilarity between objects in a low-dimensional space
- Any distance metric can be used to build the distance matrix
- NMDS is a rank approach, meaning that ranks replace distances
- The stress value indicates how well the ordination summarises the observed distances among the samples

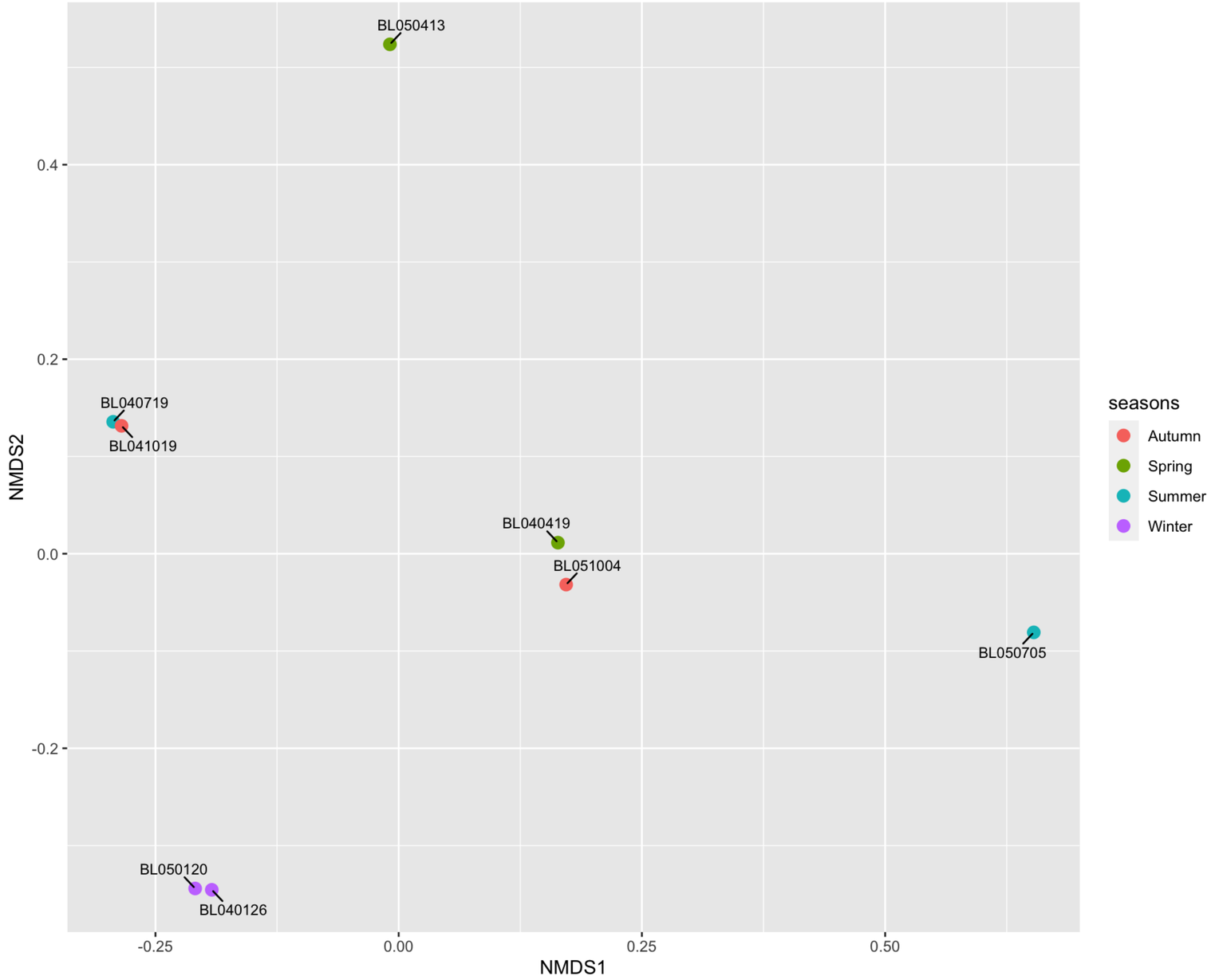
Non-metric Multidimensional Scaling (NMDS)

```
# We calculate NMDS for k(dimensions)=2

# Rarefied table (we use the dataframe to have access to sample and OTU names)

otu.tab.simple.ss.nozero.bray.nmds<-metaMDS(otu.tab.simple.ss.nozero, k=2,
trymax=100, trace=F, autotransform = F, distance="bray")

# Check Stress <0.2
```

```
# Simple plotting (Homework)


plot(otu.tab.simple.ss.nozero.bray.nmds, display="sites", type="n")

points(otu.tab.simple.ss.nozero.bray.nmds, display = "sites", col = "red", pch=19)

text(otu.tab.simple.ss.nozero.bray.nmds, display ="sites")



# Let's make nicer plots
# We define seasons for samples

seasons<-c("Winter","Spring","Summer","Autumn","Winter","Spring","Summer","Autumn")

months<-c("January","April","July","October","January","April","July","October")

library(ggplot2) # Generates nice plots

library(ggrepel) # Adds in to ggplot
```

```
# We generate a table of nmds scores and other features

otu.tab.simple.ss.nozero.bray.nmds.scores<-as.data.frame(scores(otu.tab.simple.ss.nozero.bray.nmds))
otu.tab.simple.ss.nozero.bray.nmds.scores$seasons<-seasons
otu.tab.simple.ss.nozero.bray.nmds.scores$months<-months
otu.tab.simple.ss.nozero.bray.nmds.scores$samples<-rownames(otu.tab.simple.ss.nozero.bray.nmds.scores)

#                 NMDS1       NMDS2 seasons  months  samples
# BL040126 -0.192087931 -0.34552707  Winter January BL040126
# BL040419  0.163687487  0.01138097  Spring   April BL040419
# BL040719 -0.293448084  0.13565597  Summer    July BL040719
# BL041019 -0.284857321  0.13150682  Autumn October BL041019
# BL050120 -0.209189049 -0.34417159  Winter January BL050120
# BL050413 -0.009003643  0.52375809  Spring   April BL050413
# BL050705  0.652757387 -0.08086158  Summer    July BL050705
# BL051004  0.172141153 -0.03174161  Autumn October BL051004


# Create the plot
p <- ggplot(otu.tab.simple.ss.nozero.bray.nmds.scores) +
  geom_point(mapping = aes(x = NMDS1, y = NMDS2, colour = seasons), size=3) +
  coord_fixed()+## need aspect ratio of 1!
  geom_text_repel(box.padding = 0.5, aes(x = NMDS1, y = NMDS2, label = samples),
            size = 3)
```

NMDS plots



What ordination axis corresponds to the largest gradient in our dataset (i.e. the gradient explaining most of the variance)?

Incorporating environmental data

- We aim to investigate whether environmental variability could explain community variance
- Environmental variables are standardized to have comparable ranges of variation
- For each data point (z-scores):

$$z = \frac{x - \mu}{\sigma}$$

Data point

Mean of all observations

Standard deviation of all observations

```
#Analyses using environmental variation

# We aim to investigate the environmental variation that may explain community variance.
# Read the environmental table
bbmo.metadata.course<-read_tsv("https://raw.githubusercontent.com/krabberod/BIO9905MERG1_V21/main/community.ecology/bbmo.metadata.course.tsv", col_names = T)

bbmo.metadata.course<-as.data.frame(bbmo.metadata.course)
rownames(bbmo.metadata.course)<-bbmo.metadata.course[,1]
bbmo.metadata.course<-bbmo.metadata.course[,-1]
```

#	BL040126	BL040419	BL040719	BL041019	BL050120	BL050413	BL050705	BL051004
# ENV_Temp	14	12.6	24	19.2	13	13	24	21.5
# ENV_SECCHI	14	6	24	12	19	18	22	17
# ENV_SAL_CTD	37.9	35.9	36.9	37.5	37	37.7	37.35	35.1
# ENV_CHL_total	1.1	1.4	0.4	0.3	0.5	2	0.1	0.6
# ENV_PO4	0.2	0.2	0.1	0.1	0.2	0.3	0.2	0.2
# ENV_NH4	0.3	1.5	1	0.5	1.1	2.1	1.4	1.5
# ENV_NO2	0.3	0.4	0.2	0.1	0.2	0.4	0.1	0.1
# ENV_NO3	1.5	2.5	0.1	0.4	1.1	3.3	0.2	2.4
# ENV_SI	1.8	6.1	1.4	1.4	2.6	3.4	1.8	1.6
# ENV_BACTERIA	854356	1046779	1654834	1083724	582655	788163	1127596	885144
# ENV_SYNECHOS	5927	1411	38741	30915.5	8253	4169	24823	33866
# ENV_Micromonas	9258	1424	203	730	4414	1543	505	573
# ENV_PNF_tot	11451	2266	1228	2811	5853	2506	1699	2052
# ENV_HNF_tot	329	1793	1357	822	420	669	1528	837
# ENV_Day_length_Hours_light	9.8	13.51	14.81	10.94	9.61	13.2	15.12	11.67
# Month	01_jan	04_apr	07_jul	10_oct	01_jan	04_apr	07_jul	10_oct
# Season	win	spr	sum	aut	win	spr	sum	aut
# Season_corr	win	spr	sum	aut	win	spr	sum	aut
# Year	2004	2004	2004	2004	2005	2005	2005	2005

```
#We transform variables 1:15 using z-scores to have comparable ranges of variation

bbmo.metadata.course.15vars<-bbmo.metadata.course[1:15,] #We select continuous variables

bbmo.metadata.course.15vars[]<- lapply(bbmo.metadata.course.15vars, as.character) #We transform
the datatype to characters

bbmo.metadata.course.15vars[]<- lapply(bbmo.metadata.course.15vars, as.numeric) #We transform to
numeric

#lapply: applies a function to all elements



bbmo.metadata.course.15vars.zscores<-scale(t(bbmo.metadata.course.15vars), center = T, scale = T)
bbmo.metadata.course.15vars.zscores[,1:5]



#            ENV_Temp  ENV_SECCHI ENV_SAL_CTD ENV_CHL_total    ENV_PO4
# BL040126 -0.7223777 -0.43425521  1.02225526     0.4644927  0.1950474
# BL040419 -0.9985084 -1.82387188 -1.06132234     0.9289853  0.1950474
# BL040719  1.2499845  1.30276563 -0.01953354    -0.6193235 -1.3653316
# BL041019  0.3032507 -0.78165938  0.60553974    -0.7741544 -1.3653316
# BL050120 -0.9196139  0.43425521  0.08464534    -0.4644927  0.1950474
# BL050413 -0.9196139  0.26055313  0.81389750     1.8579706  1.7554264
# BL050705  1.2499845  0.95536146  0.44927142    -1.0838162  0.1950474
# BL051004  0.7568940  0.08685104 -1.89475338    -0.3096618  0.1950474
```

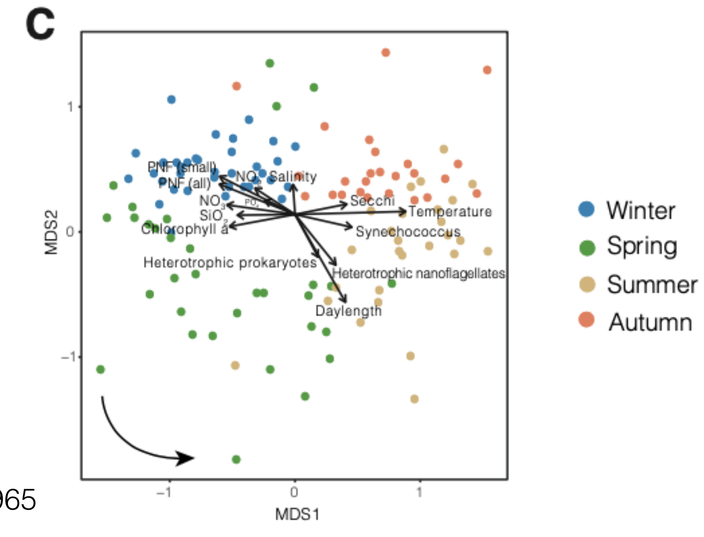
Unconstrained vs. Constrained ordination

- In **unconstrained ordination**, we first find the major compositional variation and then relate this variation to observed environmental variation (envfit e.g.)

- In **constrained ordination**, we do not want to display all or even most of the compositional variation, but only the variation that the used environmental variables, or constraints, can explain

Unconstrained ordination: Fitting vectors

- We correlate environmental variables with ordination axes
- The arrow points to the direction of most rapid change in the the environmental variable. Often this is called the direction of the gradient
- The length of the arrow is proportional to the correlation between ordination and environmental variable. Often this is called the strength of the gradient



doi: 10.1101/2021.03.18.435965

vegan (version 2.4-2)

envfit: Fits an Environmental Vector or Factor onto an Ordination

Description

The function fits environmental vectors or factors onto an ordination. The projections of points onto vectors have maximum correlation with corresponding environmental variables, and the factors show the averages of factor levels.

Usage

```
"envfit"(ord, env, permutations = 999, strata = NULL,  choices=c(1,2),  display = "sites", w  = weights(ord), na.rm = FALSE, ...)
"envfit"(formula, data, ...)
"plot"(x, choices = c(1,2), labels, arrow.mul, at = c(0,0),  axis = FALSE, p.max = NULL, col = "blue", bg, add = TRUE, ...)
"scores"(x, display, choices, ...)
vectorfit(X, P, permutations = 0, strata = NULL, w, ...)
factorfit(X, P, permutations = 0, strata = NULL, w, ...)
```

Constrained ordination

Redundancy Analysis (RDA): can be considered as a constrained version of PCA

Distance-based RDA: allows calculating RDA with a chosen distance

Selecting environmental variables that explain most community variance

- ***Forward selection:*** begins with an empty model and adds variables one by one. In each step forward, it adds one variable that gives the single best improvement to the model

- ***Backward elimination:*** starts with a model that includes all variables and eliminates variables with low explanatory power one by one

Selecting environmental variables that explain most community variance

- **Ordistep (Vegan):** Performs step-wise selection of environmental variables based on two criteria:
  - If their inclusion into the model leads to a significant increase of the explained variance
  - If the AIC (Akaike Information Criterion) of the new model is lower than the AIC of the more simple model
    - AIC: estimates the quality of models relative to other models (model selection). It is an estimator of prediction error

dbRDA

```
#Constrained Ordination

# Selection of the most important variables for distance-based redundancy analyses


mod0.rarefaction<-capscale(otu.tab.simple.ss.nozero.bray~1, as.data.frame(bbmo.metadata.course.15vars.zscores)) # model containing
                  only species matrix and intercept


mod1.rarefaction<-capscale(otu.tab.simple.ss.nozero.bray~ ., as.data.frame(bbmo.metadata.course.15vars.zscores)) # # model including
                       all variables from env matrix (the dot after tilde (~) means ALL!)

ordistep(mod0.rarefaction, scope = formula(mod1.rarefaction), perm.max = 1000, direction="forward")

# Start: otu.tab.simple.ss.nozero.bray ~ 1
#                               Df    AIC      F Pr(>F)
# + ENV_PNF_tot                  1 9.2535 1.3702  0.050 *
# + ENV_Day_length_Hours_light   1 9.1702 1.4474  0.055 .
# + ENV_Micromonas               1 9.2311 1.3909  0.055 .
# + ENV_BACTERIA                 1 9.4129 1.2248  0.110
# + ENV_Temp                     1 9.3168 1.3121  0.170
# + ENV_PO4                      1 9.4892 1.1562  0.195
# + ENV_HNF_tot                  1 9.4548 1.1870  0.240
# + ENV_SYNECHOS                 1 9.4613 1.1812  0.255
# + ENV_NH4                      1 9.5305 1.1193  0.285
# + ENV_NO3                      1 9.5767 1.0784  0.350
# + ENV_NO2                      1 9.6558 1.0087  0.380
# + ENV_CHL_total                1 9.5684 1.0857  0.385
# + ENV_SAL_CTD                  1 9.6782 0.9891  0.485
# + ENV_SI                       1 9.7718 0.9078  0.590
# + ENV_SECCHI                   1 9.8076 0.8770  0.675
# ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

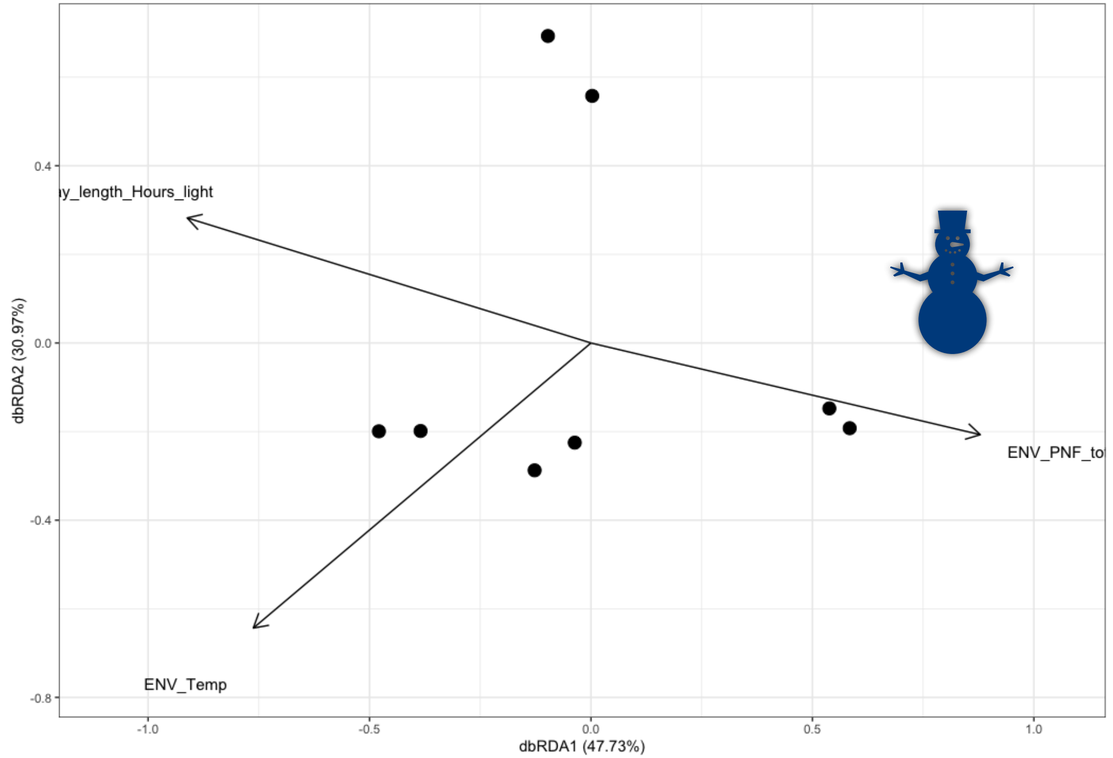
dbRDA

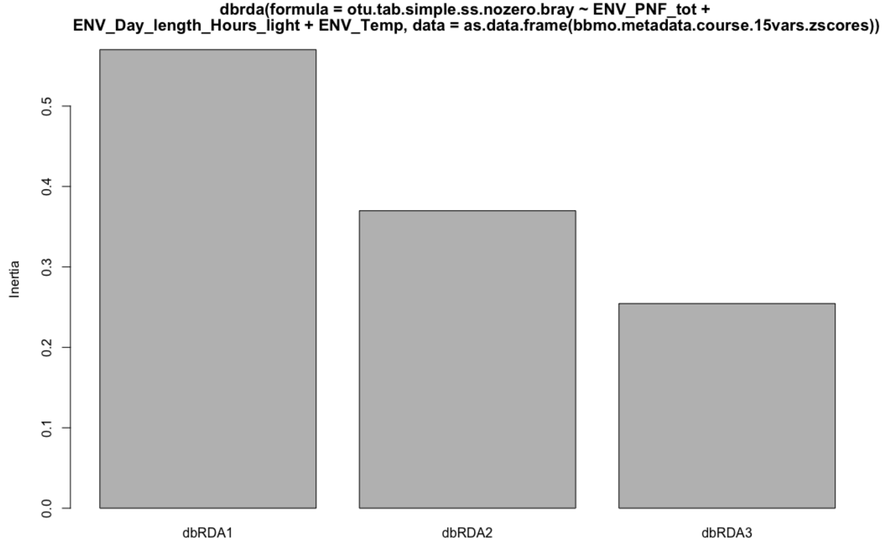
```
# Step: otu.tab.simple.ss.nozero.bray ~ ENV_PNF_tot
#                                Df     AIC      F Pr(>F)
# + ENV_SAL_CTD                   1  9.5007 1.2248  0.225
# + ENV_PO4                       1  9.4897 1.2333  0.235
# + ENV_CHL_total                 1  9.5584 1.1800  0.285
# + ENV_NO3                       1  9.6004 1.1477  0.285
# + ENV_NH4                       1  9.6031 1.1456  0.330
# + ENV_NO2                       1  9.7120 1.0625  0.370
# + ENV_Temp                      1  9.7212 1.0555  0.380
# + ENV_SYNECHOS                  1  9.7690 1.0195  0.465
# + ENV_SI                        1  9.7931 1.0013  0.600
# + ENV_BACTERIA                  1  9.8945 0.9258  0.620
# + ENV_HNF_tot                   1  9.9316 0.8983  0.645
# + ENV_SECCHI                    1  9.9584 0.8786  0.675
# + ENV_Day_length_Hours_light    1 10.0502 0.8115  0.720
# + ENV_Micromonas                1 10.0363 0.8216  0.745

# Call: capscale(formula = otu.tab.simple.ss.nozero.bray ~ ENV_PNF_tot, data = as.data.frame(bbmo.metadata.course.15vars.zscores))
# NB: the variables in this model are the ones that were selected.

# Inertia Proportion Rank
# Total          2.7072     1.0000       (Total variation in the distance matrix derived from the community data)
# Constrained    0.5033     0.1859    1  (The part of the total inertia that is explained by the explanatory variables in the model)
# Unconstrained  2.2039     0.8141    6  (The portion of the total inertia that is not captured by the model)
# Inertia is squared Bray distance

# Eigenvalues for constrained axes:
#   CAP1
# 0.5033

# Eigenvalues for unconstrained axes:
#  MDS1   MDS2   MDS3   MDS4   MDS5   MDS6
# 0.5958 0.4353 0.3912 0.2791 0.2778 0.2246
```

We use two more variables that are known to be important drivers of community variance

1. **Day-length**
2. **Temperature**

```
# We will use two more variables that we know they are important in this dataset

#We install ggord for nicer plots
library(devtools)
install_github('fawda123/ggord')
library(ggord)
library(ggplot2)

ggord(dbrda(formula = otu.tab.simple.ss.nozero.bray ~ ENV_PNF_tot+ENV_Day_length_Hours_light+ENV_Temp, data = as.data.frame(bbmo.metadata.course.
      15vars.zscores)))

screeplot(dbrda(formula = otu.tab.simple.ss.nozero.bray ~ ENV_PNF_tot+ENV_Day_length_Hours_light+ENV_Temp, data = as.data.frame(bbmo.metadata.
          course.15vars.zscores)))

dbrda(formula = otu.tab.simple.ss.nozero.bray ~ ENV_PNF_tot+ENV_Day_length_Hours_light+ENV_Temp, data = as.data.frame(bbmo.metadata.course.15vars.
      zscores))

# Call: dbrda(formula = otu.tab.simple.ss.nozero.bray ~ ENV_PNF_tot + ENV_Day_length_Hours_light + ENV_Temp, data =
#               as.data.frame(bbmo.metadata.course.15vars.zscores))
#                Inertia Proportion Rank
# Total          2.7072     1.0000
# Constrained    1.1945     0.4412    3   # The part of the total inertia that is explained by the explanatory variables in the model
# Unconstrained  1.5127     0.5588    4   # The portion of the total inertia that is not captured by the model

# Inertia is squared Bray distance  # Inertia = variance in species abundances


# Eigenvalues for constrained axes:
#   dbRDA1 dbRDA2 dbRDA3              # Amount of explained variation (inertia) along that specific canonical axis (sum = constrained inertia)
#   0.5701 0.3699 0.2545              # The first constrained axis (CAP1 or dbRDA1) captures the maximum variation in the response data that is
                                      linearly explainable by the predictors. The second constrained axis (CAP2/dbRDA2) captures the maximum
                                      remaining explainable variation, orthogonal to the first, and so on


# Eigenvalues for unconstrained axes:
#     MDS1   MDS2   MDS3   MDS4
#  0.5818 0.3960 0.2997 0.2353
```





dbRDA1 explains the largest portion of the model-constrained variation

Core microbiota BBMO: example

- What taxa constitute the interconnected core microbiota over 10 years at one marine location?
- How the diversity of core taxa changes over time? What are the seasonal patterns?
- What are their potential ecological interactions?

Krabberød *et al. Environmental Microbiome* (2022) 17:22
https://doi.org/10.1186/s40793-022-00417-1

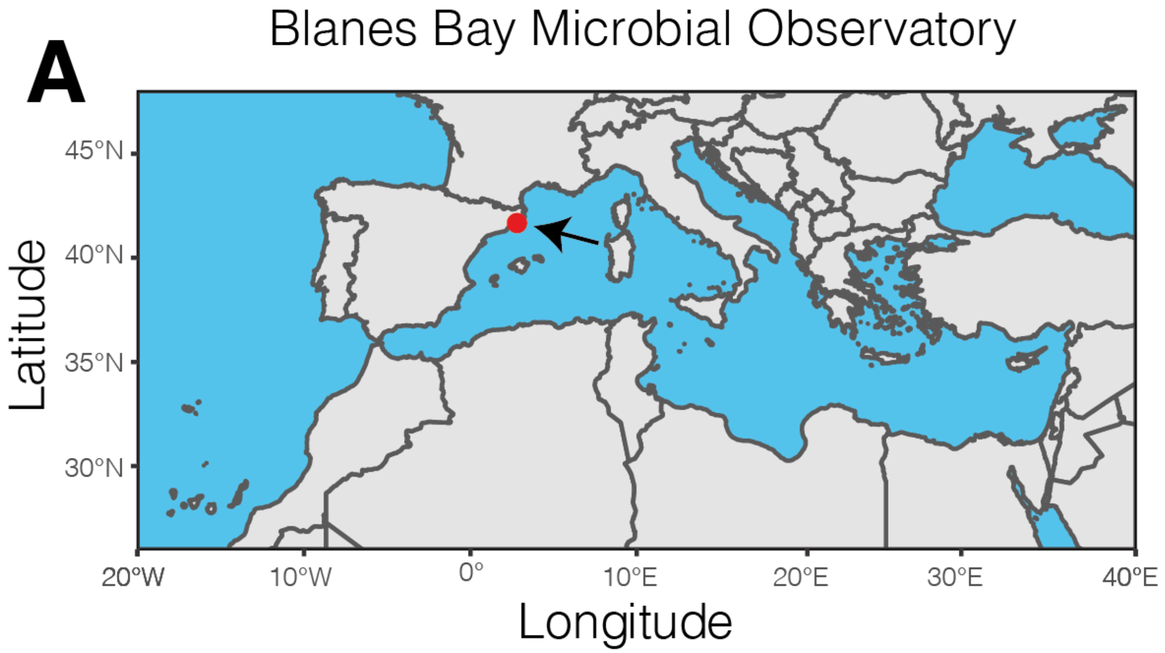
Environmental Microbiome

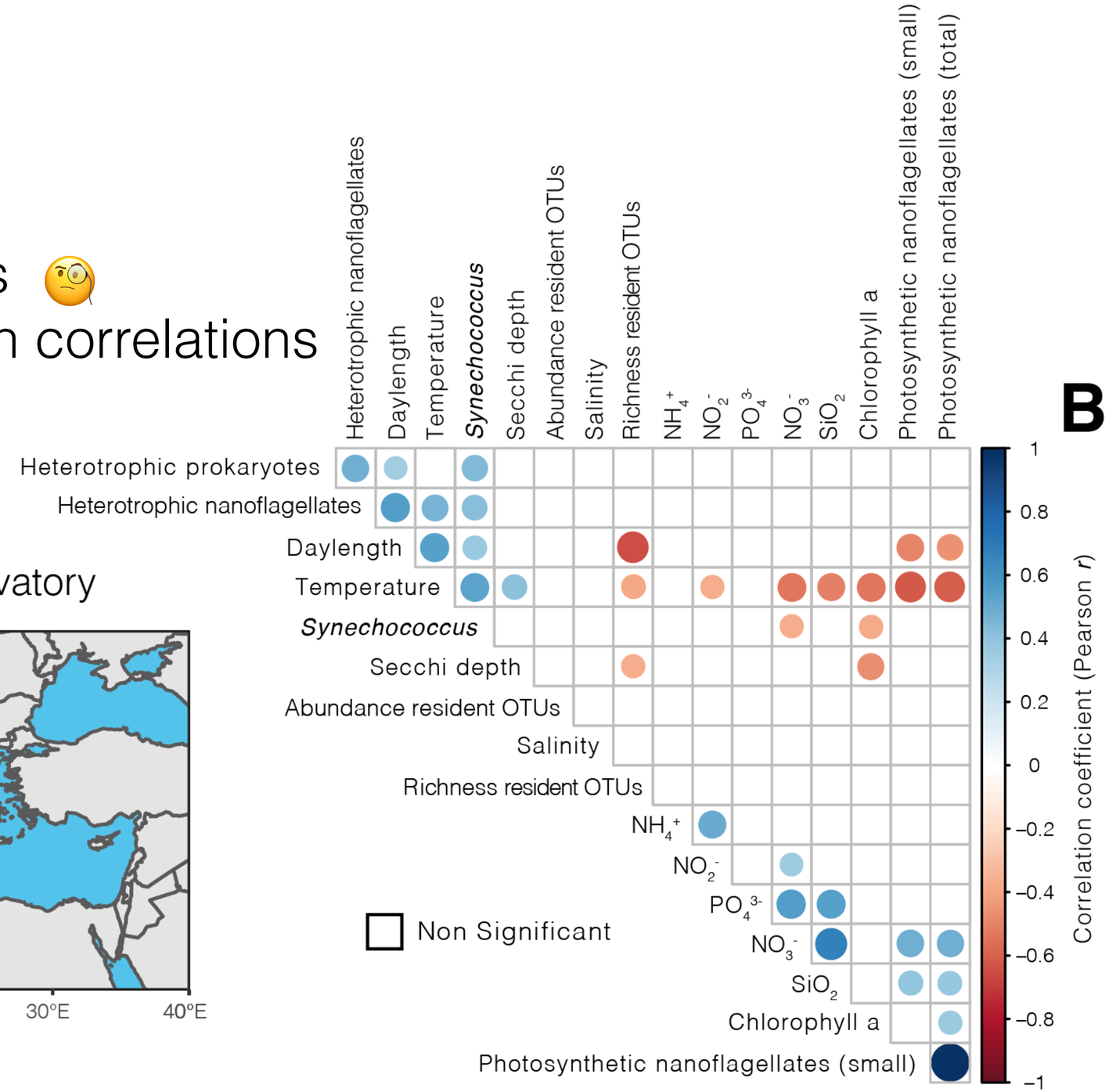
RESEARCH ARTICLE Open Access

Long-term patterns of an interconnected core marine microbiota

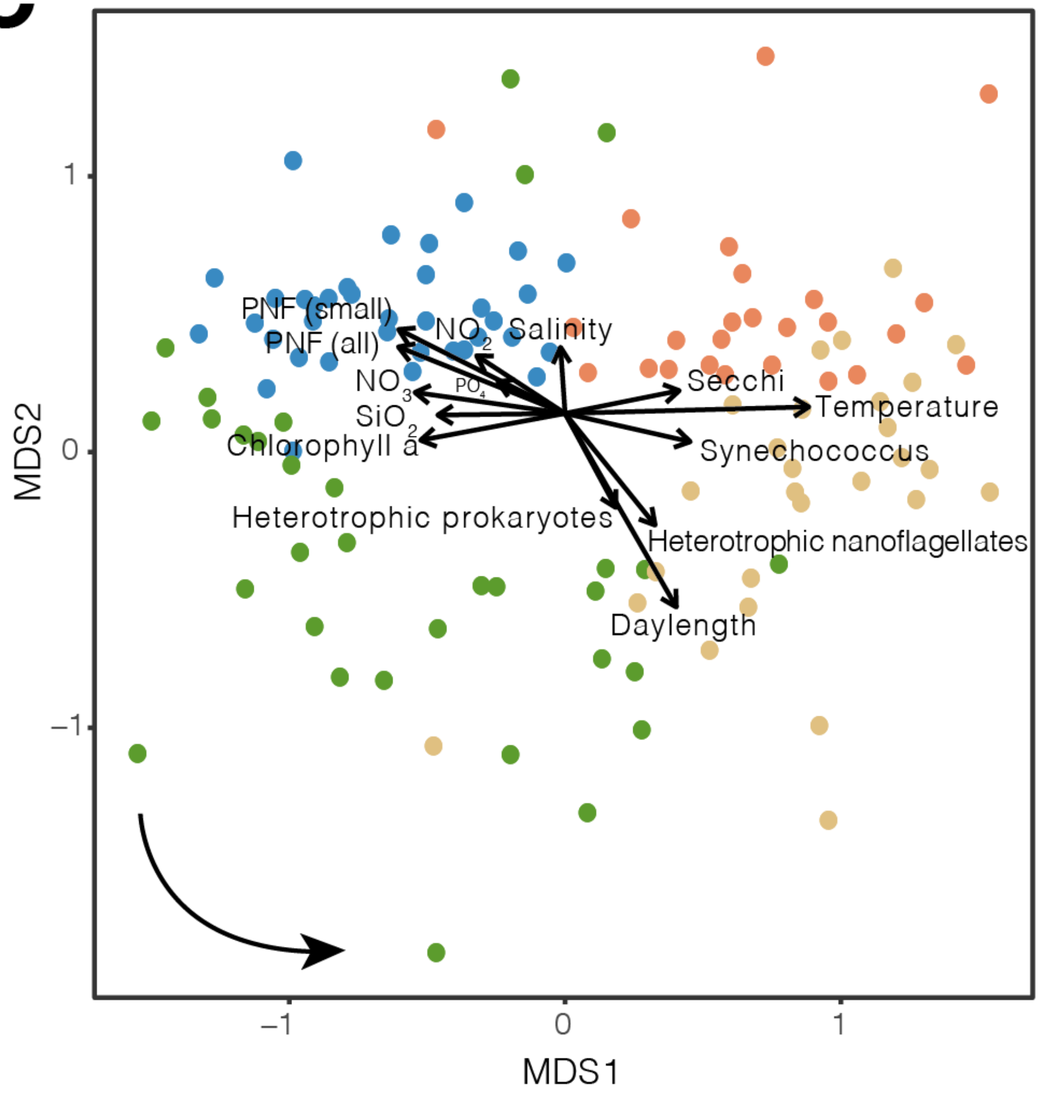
Anders K. Krabberød[1*], Ina M. Deutschmann[2], Marit F. M. Bjorbækmo[1], Vanessa Balagué[2], Caterina R. Giner[2], Isabel Ferrera[2,3], Esther Garcés[2], Ramon Massana[2], Josep M. Gasol[2,4] and Ramiro Logares[1,2*]

zscores 🧐
Pearson correlations

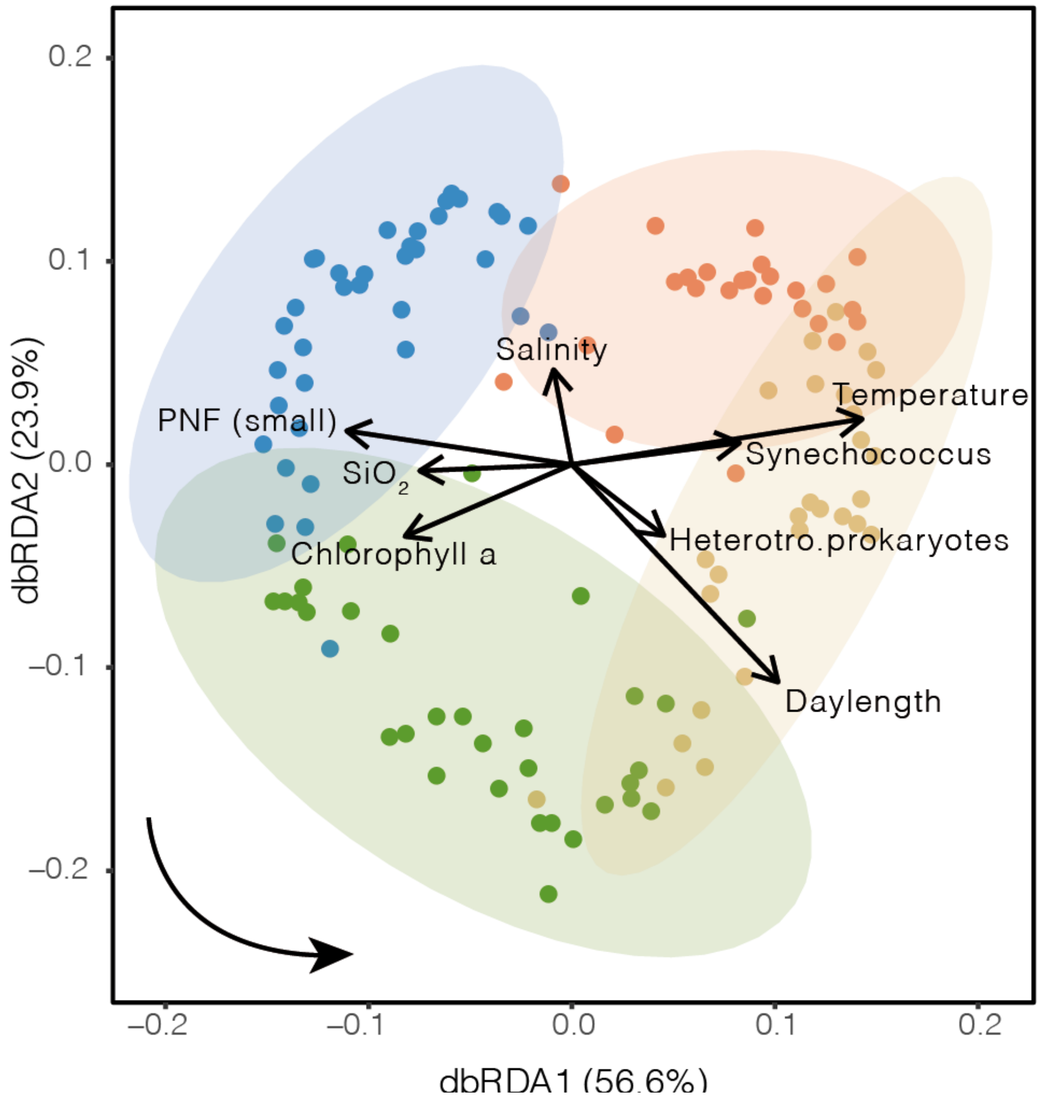




C



D



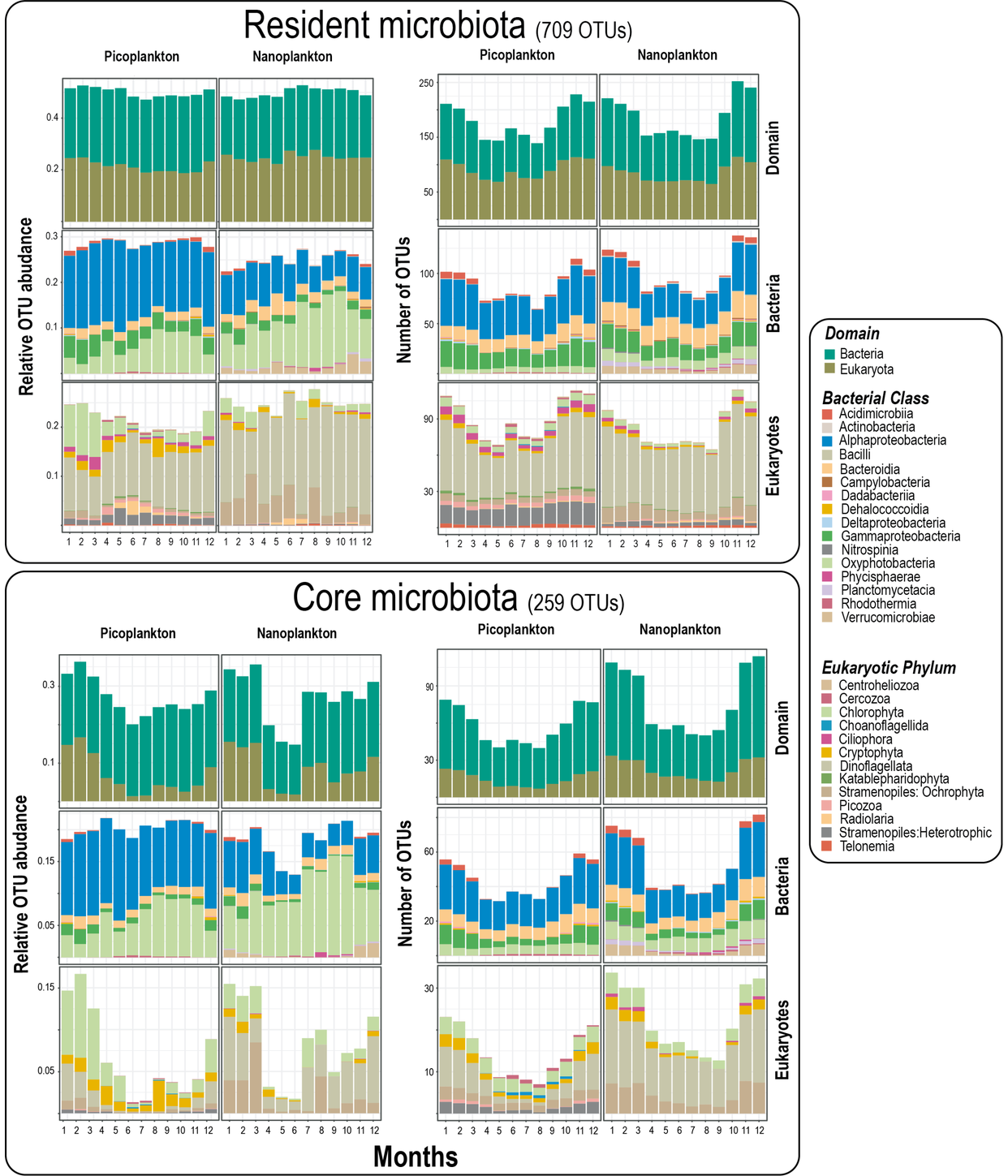
Winter
Spring
Summer
Autumn

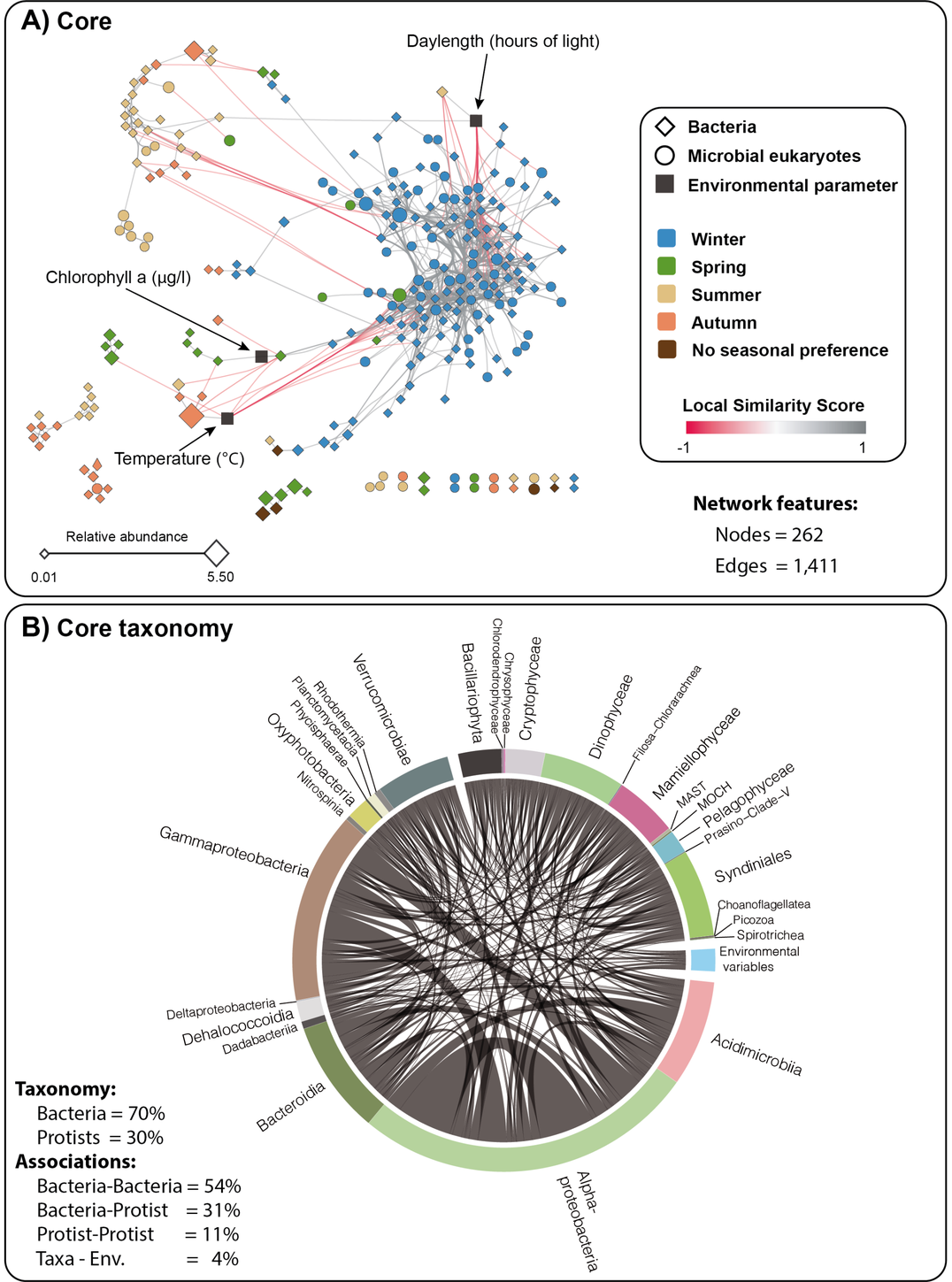
🧐

NMDS
envfit

dbRDA
Forward selection

Richness 🧐
Relative abundance




A) Core
Daylength (hours of light)
Chlorophyll a (µg/l)
Temperature (°C)
Relative abundance
0.01
5.50
Bacteria
Microbial eukaryotes
Environmental parameter
Winter
Spring
Summer
Autumn
No seasonal preference
Local Similarity Score
-1
1
Network features:
Nodes = 262
Edges = 1,411
B) Core taxonomy
Verrucomicrobiae
Bacillariophyta
Chrysophyceae
Chlorodendrophyceae
Cryptophyceae
Dinophyceae
Filosa-Chlorarachnea
Mamiellophyceae
MAST
MOCH
Pelagophyceae
Prasino-Clade-V
Syndiniales
Choanoflagellatea
Picozoa
Spirotrichea
Environmental variables
Acidimicrobiia
Alpha-proteobacteria
Bacteroidia
Dadabacteriia
Dehalococcoidia
Deltaproteobacteria
Gammaproteobacteria
Nitrospinia
Oxyphotobacteria
Physicisphaerae
Planctomycetacia
Rhodothermia
Taxonomy:
Bacteria = 70%
Protists = 30%
Associations:
Bacteria-Bacteria = 54%
Bacteria-Protist = 31%
Protist-Protist = 11%
Taxa - Env. = 4%

Conclusions of the study

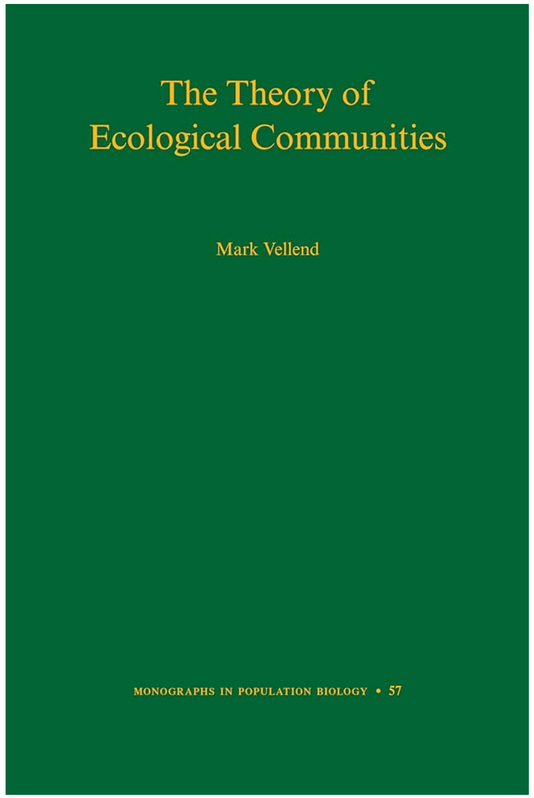
- The core microbiota included 259 Operational Taxonomic Units (OTUs), 182 bacteria, 77 protists, and 1411 strong and primarily positive (~ 95%) associations.

- The richness and abundance of core OTUs varied annually, decreasing in stratified warm waters and increasing in colder mixed waters.

- Most core OTUs preferred one season, mostly winter. This season displayed subnetworks with the highest connectivity.

Other things you could explore

- The relative importance of the main processes structuring microbiotas
  - Selection
  - Dispersal
  - Drift



The ISME Journal (2013) 7, 2069–2079

www.nature.com/ismej

ORIGINAL ARTICLE

Quantifying community assembly processes and identifying features that impose them

James C Stegen[1], Xueju Lin[1,2], Jim K Fredrickson[1], Xingyuan Chen[3], David W Kennedy[1], Christopher J Murray[4], Mark L Rockhold[3] and Allan Konopka[1]
[1]*Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA;* [2]*School of Biology, Georgia Institute of Technology, Atlanta, GA, USA;* [3]*Hydrology Group, Pacific Northwest National Laboratory, Richland, WA, USA and* [4]*Department of Geosciences, Pacific Northwest National Laboratory, Richland, WA, USA*

R code
https://github.com/stegen/Stegen_etal_ISME_2013

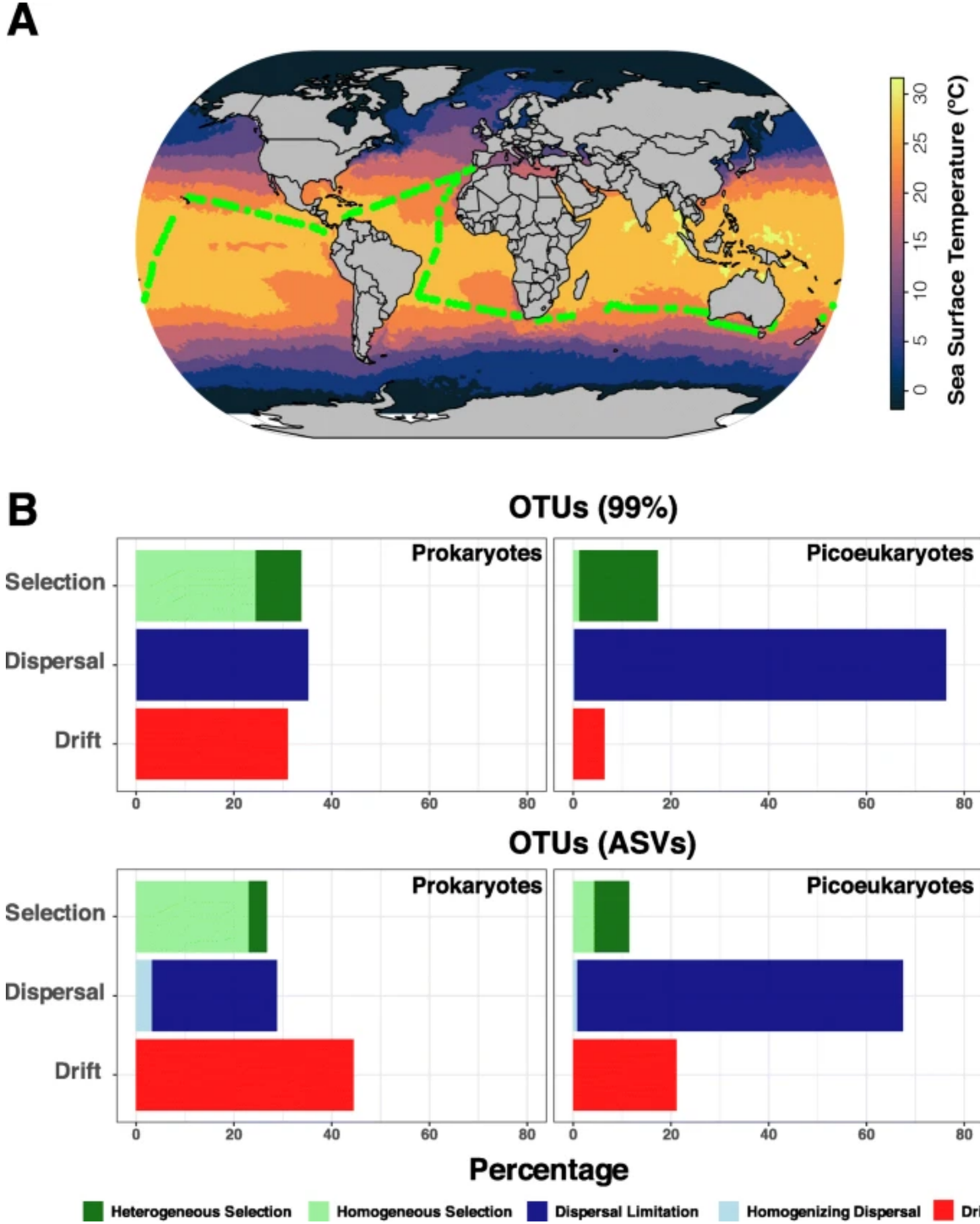
Logares *et al. Microbiome* (2020) 8:55
https://doi.org/10.1186/s40168-020-00827-8

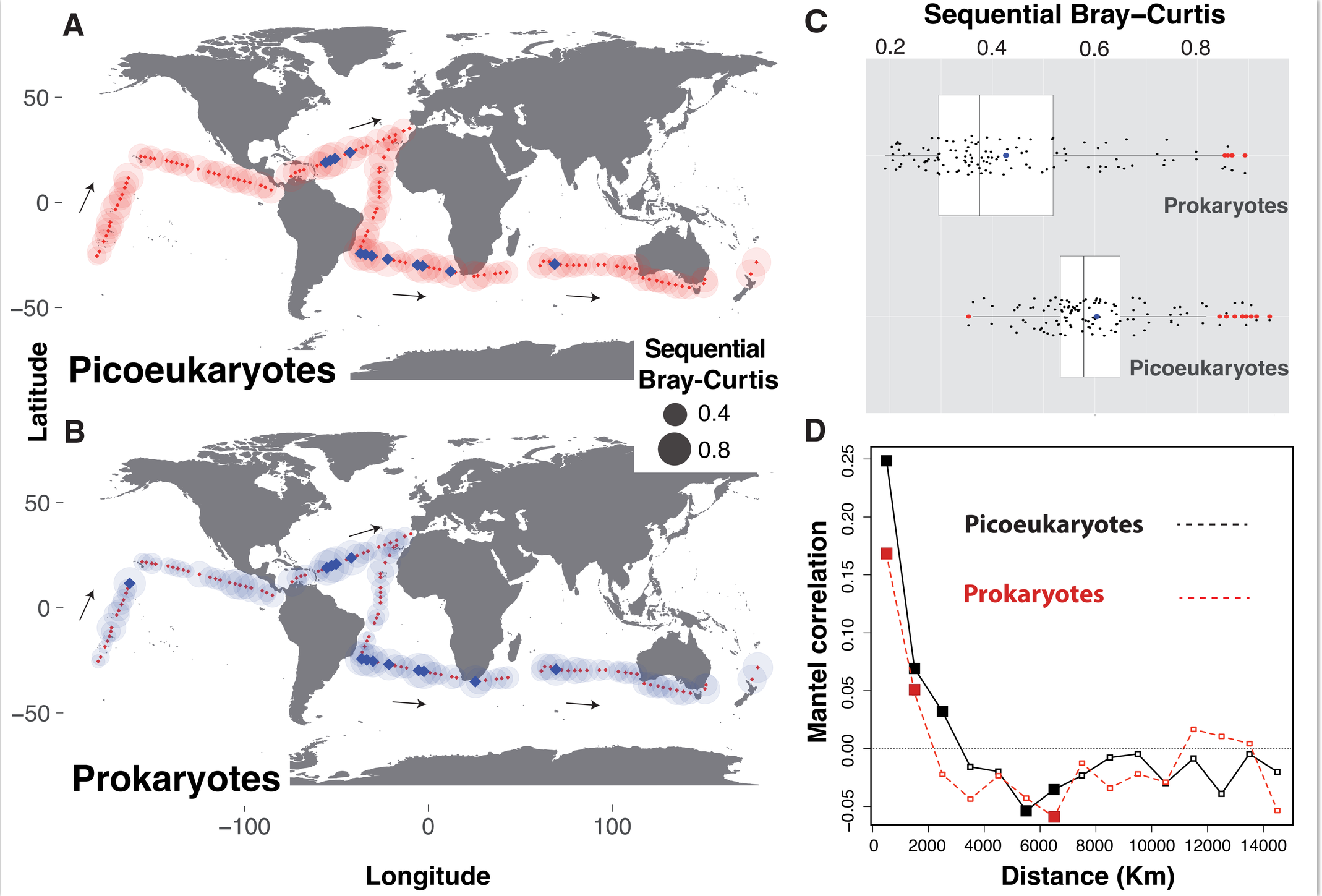
Microbiome

RESEARCH Open Access

Disentangling the mechanisms shaping the surface ocean microbiota

Ramiro Logares[1,2*], Ina M. Deutschmann[1], Pedro C. Junger[3], Caterina R. Giner[1,4], Anders K. Krabberød[2], Thomas S. B. Schmidt[5], Laura Rubinat-Ripoll[6], Mireia Mestre[1,7,8], Guillem Salazar[1,9], Clara Ruiz-González[1], Marta Sebastián[1,10], Colomban de Vargas[6], Silvia G. Acinas[1], Carlos M. Duarte[11], Josep M. Gasol[1,12] and Ramon Massana[1]




A
Picoeukaryotes
B
Prokaryotes
Sequential Bray-Curtis
0.4
0.8
Latitude
Longitude
50
0
−50
−100
0
100
C
Sequential Bray−Curtis
0.2
0.4
0.6
0.8
Prokaryotes
Picoeukaryotes
D
Mantel correlation
Distance (Km)
Picoeukaryotes
Prokaryotes
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
0
2000
4000
6000
8000
10000
12000
14000

Bray Curtis