CC2 Rade De Brest

Mrozinski Alexandre

```
knitr::opts_chunk$set(echo=TRUE, eval=TRUE)
```

```
library(rmarkdown)
library(knitr)
library(phyloseq)
library(dada2)
```

```
## Loading required package: Rcpp
```

```
library(DECIPHER)
```

```
## Loading required package: Biostrings

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```
## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:phyloseq':
##
##     distance

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':
##
##     strsplit

## Loading required package: RSQLite

## Loading required package: parallel
```

```
library(phangorn)
```

```
## Loading required package: ape

##
## Attaching package: 'ape'

## The following object is masked from 'package:Biostrings':
##
##     complement
```

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:BiocGenerics':
##
##     combine
```

```
library(shiny)
library(miniUI)
library(caret)
```

```
## Loading required package: lattice
```

```
library(pls)
```

```
##
## Attaching package: 'pls'

## The following object is masked from 'package:caret':
##
##     R2

## The following object is masked from 'package:ape':
##
##     mvr

## The following object is masked from 'package:stats':
##
##     loadings
```

```
library(e1071)
library(ggplot2)
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:BiocGenerics':
##
##     combine
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:randomForest':
##
##     combine
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:Biostrings':
##
##     collapse, intersect, setdiff, setequal, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following object is masked from 'package:XVector':
##
##     slice

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggrepel)
#library(nlme)
library(devtools)
```

```
## Loading required package: usethis
```

```
library(reshape2)
library(PMA)
#library(structSSI)
library(ade4)
```

```
##
## Attaching package: 'ade4'
```

```
## The following object is masked from 'package:Biostrings':
##
##     score

## The following object is masked from 'package:BiocGenerics':
##
##     score
```

```
library(ggnetwork)
library(intergraph)
library(scales)
library(genefilter)
library(impute)
library(phyloseqGraphTest)
library(Biostrings)
```

```
wget pagesperso.univ-brest.fr/~maignien/teaching/M1-MFA/UE-Ecogenomique2/EcoG2_data_cc2.tar.gz
tar xzvf EcoG2_data_cc2.tar.gz
```

```
mkdir data
mv St_Stratif_11mars15/Station* data
mv St_Stratif_10sept14/Station* data

rm -d St_Stratif_11mars15
rm -d St_Stratif_10sept14
rm EcoG2_data_cc2.tar.gz
```

```
path <- "data"
```

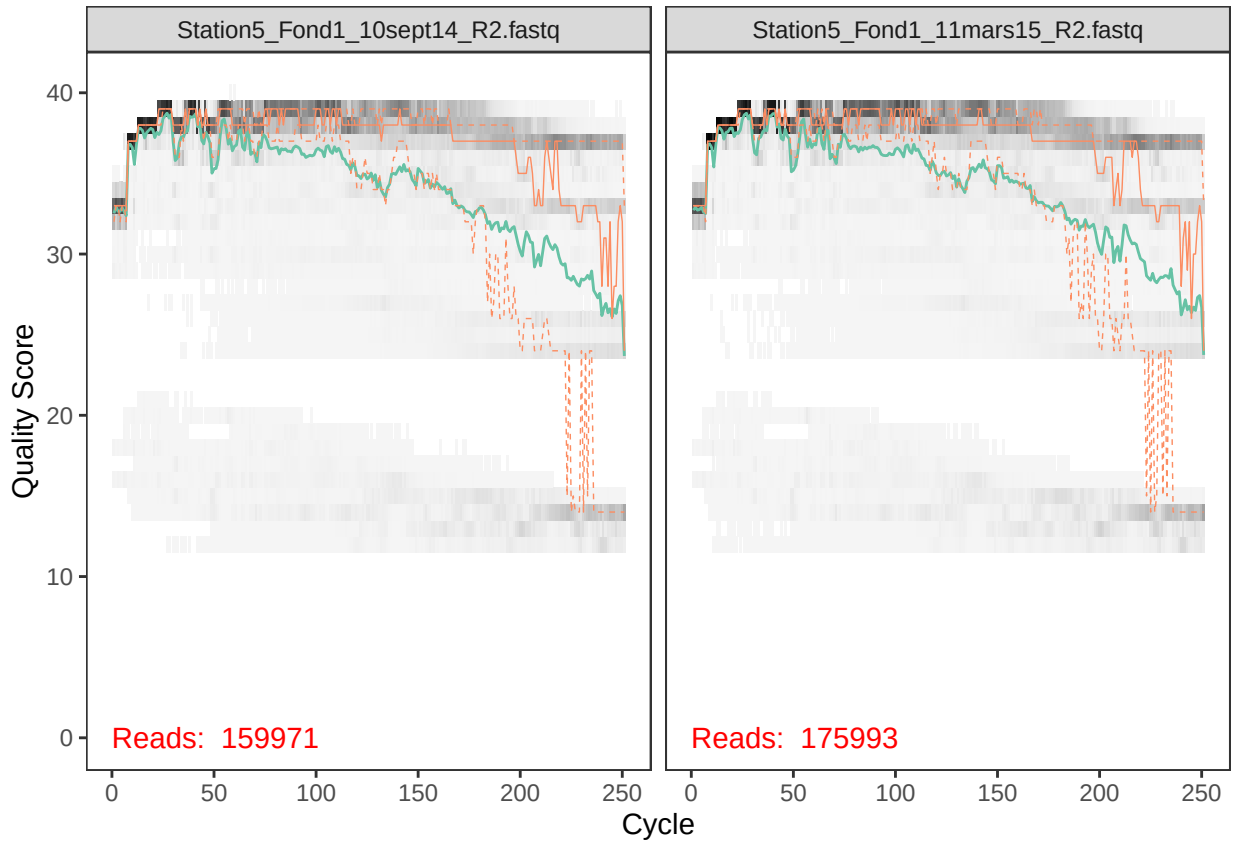
```
list.files(path)
```

```
##  [1] "filtered"                            "Station5_Fond1_10sept14_R1.fastq"
##  [3] "Station5_Fond1_10sept14_R2.fastq"    "Station5_Fond1_11mars15_R1.fastq"
##  [5] "Station5_Fond1_11mars15_R2.fastq"    "Station5_Fond2_10sept14_R1.fastq"
##  [7] "Station5_Fond2_10sept14_R2.fastq"    "Station5_Fond2_11mars15_R1.fastq"
##  [9] "Station5_Fond2_11mars15_R2.fastq"    "Station5_Fond3_10sept14_R1.fastq"
## [11] "Station5_Fond3_10sept14_R2.fastq"    "Station5_Median1_10sept14_R1.fastq"
## [13] "Station5_Median1_10sept14_R2.fastq"  "Station5_Median2_10sept14_R1.fastq"
## [15] "Station5_Median2_10sept14_R2.fastq"  "Station5_Surface1_10sept14_R1.fastq"
## [17] "Station5_Surface1_10sept14_R2.fastq" "Station5_Surface1_11mars15_R1.fastq"
## [19] "Station5_Surface1_11mars15_R2.fastq" "Station5_Surface2_10sept14_R1.fastq"
## [21] "Station5_Surface2_10sept14_R2.fastq" "Station5_Surface2_11mars15_R1.fastq"
## [23] "Station5_Surface2_11mars15_R2.fastq"
```

```
fnFs <- sort(list.files(path, pattern="_R1", full.names = TRUE))
fnRs <- sort(list.files(path, pattern="_R2", full.names = TRUE))

sample.names <- sapply(strsplit(basename(fnFs), "_R"), `[`, 1)
```

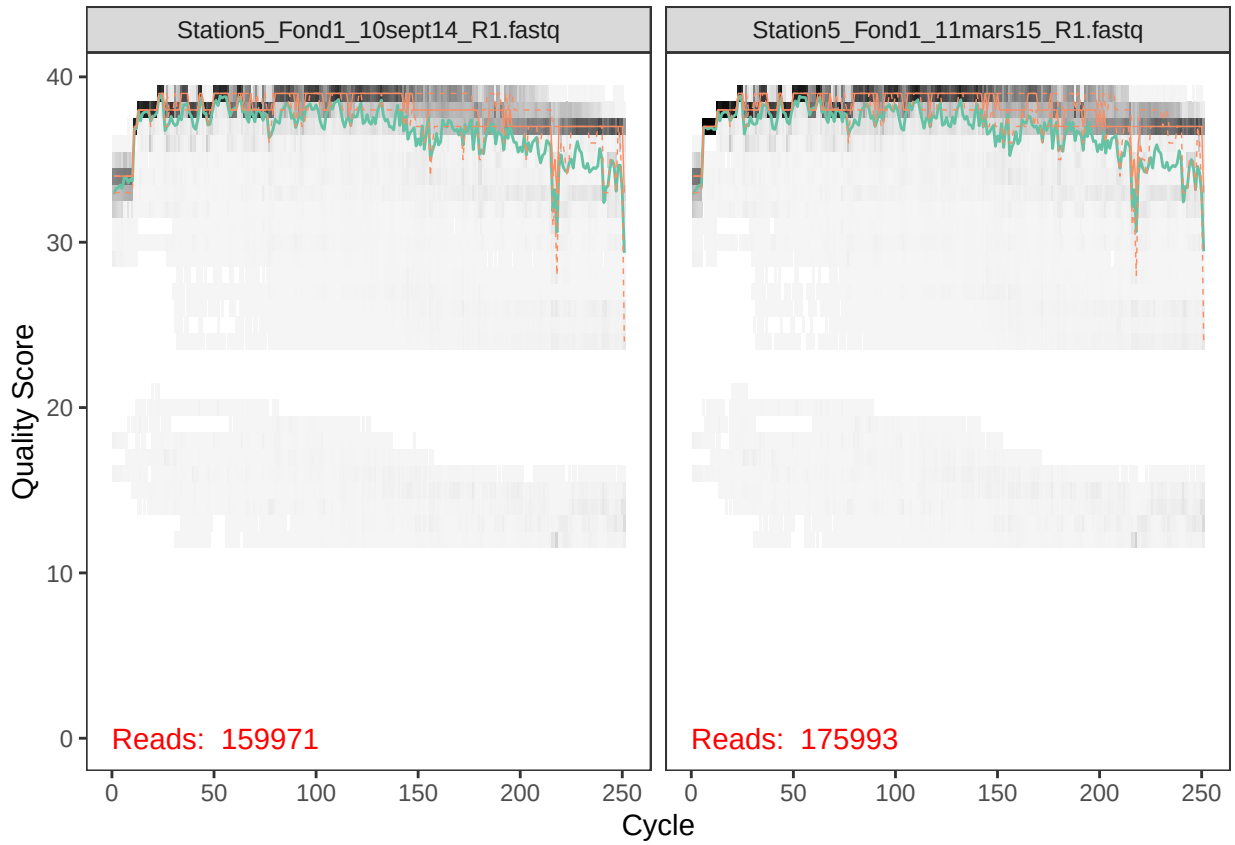
```
plotQualityProfile(fnRs[1:2])
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```




```
plotQualityProfile(fnFs[1:2])
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



Filter and trim

```
filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
filtRs <- file.path(path, "filtered", paste0(sample.names, "_R_filt.fastq.gz"))
names(filtFs) <- sample.names
names(filtRs) <- sample.names
```

```
out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(240,240), trimLeft=c(18,18),
              maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
              compress=TRUE, multithread=TRUE)
head(out)
```

```
##                                    reads.in reads.out
## Station5_Fond1_10sept14_R1.fastq     159971    134523
## Station5_Fond1_11mars15_R1.fastq     175993    149245
## Station5_Fond2_10sept14_R1.fastq     197039    163246
## Station5_Fond2_11mars15_R1.fastq      87585     74372
## Station5_Fond3_10sept14_R1.fastq     117140     98357
## Station5_Median1_10sept14_R1.fastq   116519     99668
```

Learn the Error Rates

```
errFs <- learnErrors(filtFs, multithread=TRUE)
```

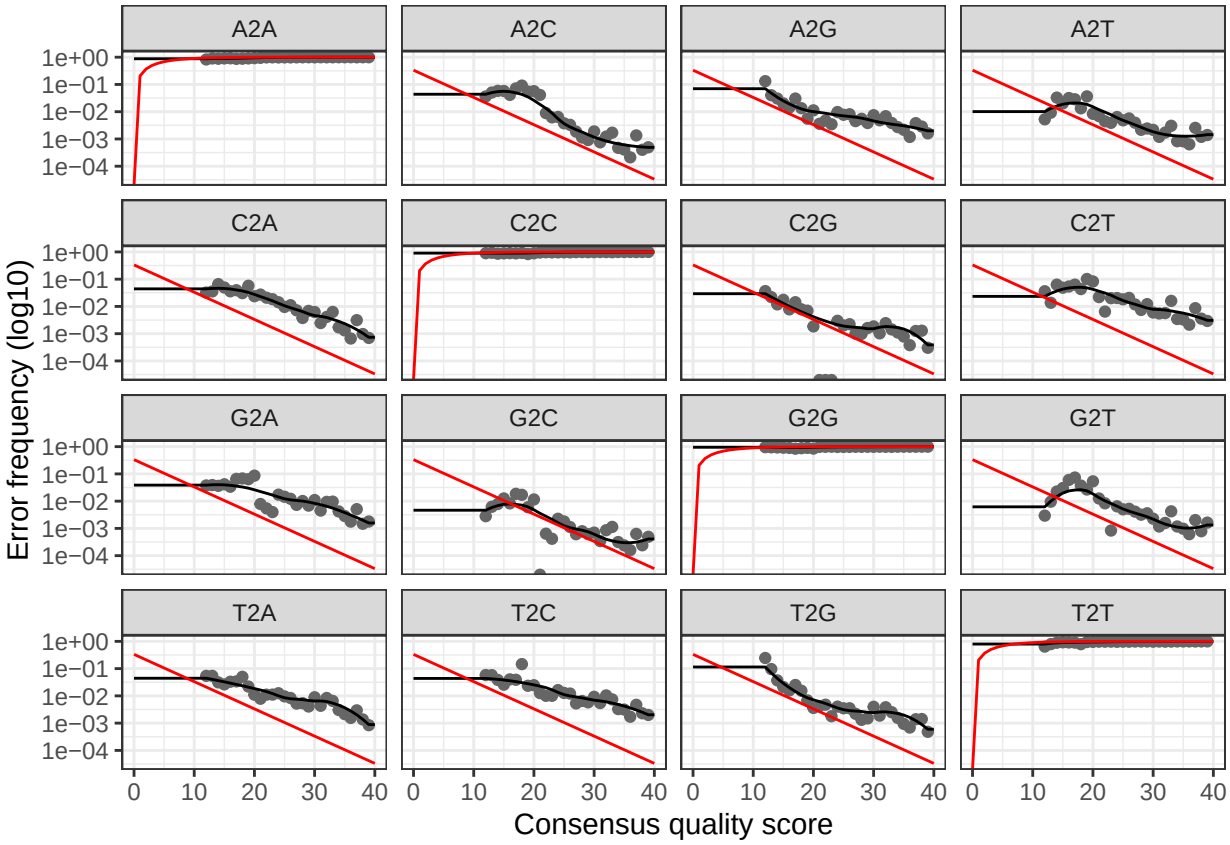
```
## 115747692 total bases in 521386 reads from 4 samples will be used for learning the error rates.
```

```
errRs <- learnErrors(filtRs, multithread=TRUE)
```

```
## 115747692 total bases in 521386 reads from 4 samples will be used for learning the error rates.
```

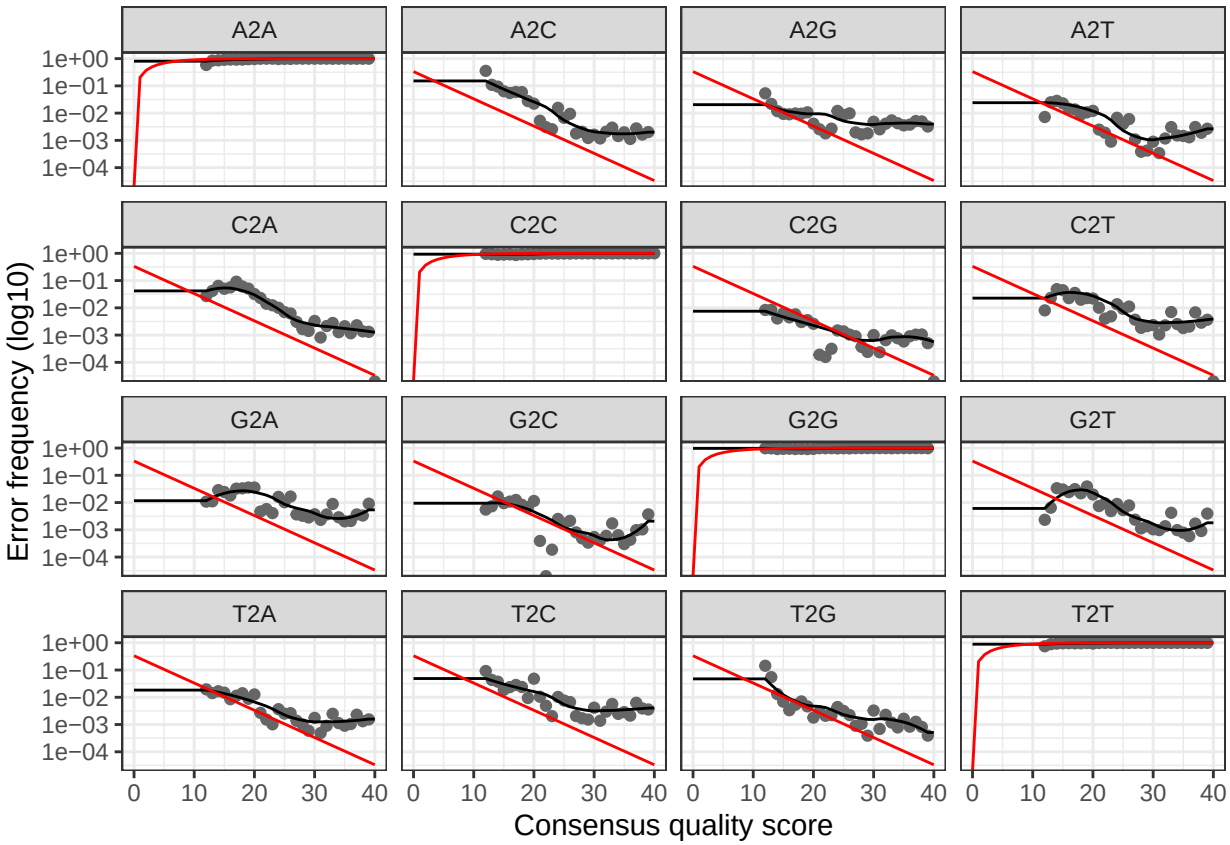
```
plotErrors(errFs, nominalQ=TRUE)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis
```




```
plotErrors(errRs, nominalQ=TRUE)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis
```



Sample Inference

```
dadaFs <- dada(filtFs, err=errFs, multithread=TRUE)
```

```
## Sample 1 - 134523 reads in 33780 unique sequences.
## Sample 2 - 149245 reads in 31809 unique sequences.
## Sample 3 - 163246 reads in 42025 unique sequences.
## Sample 4 - 74372 reads in 18151 unique sequences.
## Sample 5 - 98357 reads in 27114 unique sequences.
## Sample 6 - 99668 reads in 25888 unique sequences.
## Sample 7 - 92304 reads in 23153 unique sequences.
## Sample 8 - 100839 reads in 24107 unique sequences.
## Sample 9 - 65958 reads in 16059 unique sequences.
## Sample 10 - 72561 reads in 17955 unique sequences.
## Sample 11 - 84850 reads in 21970 unique sequences.
```

```
dadaRs <- dada(filtRs, err=errRs, multithread=TRUE)
```

```
## Sample 1 - 134523 reads in 56106 unique sequences.
## Sample 2 - 149245 reads in 52995 unique sequences.
## Sample 3 - 163246 reads in 67968 unique sequences.
## Sample 4 - 74372 reads in 29253 unique sequences.
```

```
## Sample 5 - 98357 reads in 42601 unique sequences.
## Sample 6 - 99668 reads in 39919 unique sequences.
## Sample 7 - 92304 reads in 36562 unique sequences.
## Sample 8 - 100839 reads in 37400 unique sequences.
## Sample 9 - 65958 reads in 26950 unique sequences.
## Sample 10 - 72561 reads in 27800 unique sequences.
## Sample 11 - 84850 reads in 35712 unique sequences.
```

```
dadaFs[[1]]
```

```
## dada-class: object describing DADA2 denoising results
## 986 sequence variants were inferred from 33780 input unique sequences.
## Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16
```

```
dadaRs[[1]]
```

```
## dada-class: object describing DADA2 denoising results
## 833 sequence variants were inferred from 56106 input unique sequences.
## Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16
```

Merge paired reads

```
mergers <- mergePairs(dadaFs, filtFs, dadaRs, filtRs, verbose=TRUE)
```

```
## 106593 paired-reads (in 3609 unique pairings) successfully merged out of 127613 (in 16602 pairings) i

## 127170 paired-reads (in 2941 unique pairings) successfully merged out of 143945 (in 12620 pairings) i

## 129083 paired-reads (in 4973 unique pairings) successfully merged out of 155344 (in 21473 pairings) i

## 60856 paired-reads (in 1740 unique pairings) successfully merged out of 70718 (in 7546 pairings) inpu

## 76064 paired-reads (in 2446 unique pairings) successfully merged out of 92630 (in 12579 pairings) inp

## 79142 paired-reads (in 2418 unique pairings) successfully merged out of 95096 (in 11430 pairings) inp

## 74335 paired-reads (in 1953 unique pairings) successfully merged out of 87930 (in 9744 pairings) inpu

## 82679 paired-reads (in 2272 unique pairings) successfully merged out of 96838 (in 10027 pairings) inp

## 54136 paired-reads (in 1296 unique pairings) successfully merged out of 62452 (in 6032 pairings) inpu

## 60370 paired-reads (in 1328 unique pairings) successfully merged out of 69794 (in 6607 pairings) inpu

## 67175 paired-reads (in 2019 unique pairings) successfully merged out of 80479 (in 9511 pairings) inpu
```

```
head(mergers[[1]])
```

```
##
## 1        TAATACGAAGGGACCTAGCGTAGTTCGGAATTACTGGGCTTAAAGAGTTCGTAGGTGGTTGAAAAAGTTAGTGGTGAAATCCCAGAGCTTA
## 2        TAATACGAAGGGACCTAGCGTAGTTCGGAATTACTGGGCTTAAAGAGTTCGTAGGTGGTTGAAAAAGTTGGTGGTGAAATCCCAGAGCTTA
## 3        TAATACGAAGGGACCTAGCGTAGTTCGGAATTACTGGGCTTAAAGAGTTCGTAGGTGGTTGAAAAAGTTGGTGGTGAAATCCCAGAGCTTA
## 4        TAATACGAAGGGACCTAGCGTAGTTCGGAATTACTGGGCTTAAAGAGTTCGTAGGTGGTTGAAAAAGTTAGTGGTGAAATCCCAGAGCTTA
## 5        TAATACGAAGGGACCTAGCGTAGTTCGGAATTACTGGGCTTAAAGAGTTCGTAGGTGGTTGAAAAAGTTGGTGGTGAAATCCCAGAGCTTA
## 6 TAATACATAGGGGTCAAGCGTTGTCCGGATTTATTGGGCGTAAAGAGCTCGTAGGCGGTTCAACAAGTCGGTCGTAAAAGTTTAGGGCTCAACCCTAA
##   abundance forward reverse nmatch nmismatch nindel prefer accept
## 1      4967       1       2     69          0      0      1   TRUE
## 2      3984       2       1     69          0      0      2   TRUE
## 3      3599       3       1     69          0      0      2   TRUE
## 4      2389       1       1     69          0      0      2   TRUE
## 5      2169       2       2     69          0      0      1   TRUE
## 6      2126       9       4     62          0      0      1   TRUE
```

Construct sequence table

```
seqtab <- makeSequenceTable(mergers)
dim(seqtab)
```

```
## [1]    11 13710
```

```
table(nchar(getSequences(seqtab)))
```

```
##
##  358  359  369  370  371  372  373  374  375  376  377  378  379  380  381  382
##    1    1    1    4  137   22  109  128 3923 1898 1620 2008 2251   77 1441   66
##  383  384  388  392  393  395  398  403  404  412  429  430  431  432
##    5    1    1    2    1    1    3    1    1    1    1    3    1    1
```

Remove chimeras

```
seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=TRUE, verbose=TRUE)
```

```
## Identified 12307 bimeras out of 13710 input sequences.
```

```
dim(seqtab.nochim)
```

```
## [1]   11 1403
```

```
sum(seqtab.nochim)/sum(seqtab)
```

```
## [1] 0.7891376
```

Track reads through the pipeline

```
getN <- function(x) sum(getUniques(x))
track <- cbind(out, sapply(dadaFs, getN), sapply(dadaRs, getN), sapply(mergers, getN), rowSums(seqtab.n
colnames(track) <- c("input", "filtered", "denoisedF", "denoisedR", "merged", "nonchim")
rownames(track) <- sample.names
head(track)
```

```
##                            input filtered denoisedF denoisedR merged nonchim
## Station5_Fond1_10sept14   159971   134523    131807    129952 106593   80898
## Station5_Fond1_11mars15   175993   149245    146865    145918 127170  103044
## Station5_Fond2_10sept14   197039   163246    159747    158439 129083   95568
## Station5_Fond2_11mars15    87585    74372     72825     71952  60856   50387
## Station5_Fond3_10sept14   117140    98357     96052     94566  76064   59308
## Station5_Median1_10sept14 116519    99668     97740     96815  79142   61499
```

Assign taxonomy

```
wget https://zenodo.org/record/4587955/files/silva_nr99_v138.1_train_set.fa.gz?download=1
```

```
taxa <- assignTaxonomy(seqtab.nochim, "silva_nr99_v138.1_train_set.fa.gz?download=1", multithread=TRUE)
```

```
taxa.print <- taxa
rownames(taxa.print) <- NULL
head(taxa.print)
```

```
##      Kingdom    Phylum             Class                 Order
## [1,] "Bacteria" "Proteobacteria"   "Alphaproteobacteria" "SAR11 clade"
## [2,] "Bacteria" "Cyanobacteria"    "Cyanobacteriia"      "Synechococcales"
## [3,] "Bacteria" "Proteobacteria"   "Alphaproteobacteria" "SAR11 clade"
## [4,] "Bacteria" "Proteobacteria"   "Alphaproteobacteria" "SAR11 clade"
## [5,] "Bacteria" "Proteobacteria"   "Alphaproteobacteria" "SAR11 clade"
## [6,] "Bacteria" "Actinobacteriota" "Acidimicrobiia"      "Actinomarinales"
##      Family             Genus
## [1,] "Clade I"          "Clade Ia"
## [2,] "Cyanobiaceae"     "Synechococcus CC9902"
## [3,] "Clade I"          "Clade Ia"
## [4,] "Clade I"          "Clade Ia"
## [5,] "Clade II"         NA
## [6,] "Actinomarinaceae" "Candidatus Actinomarina"
```

Test taxa 2

```
#test taxo 2
wget https://zenodo.org/record/4587955/files/silva_nr99_v138.1_wSpecies_train_set.fa.gz?download=1
```

```
taxa2 <- assignTaxonomy(seqtab.nochim, "silva_nr99_v138.1_wSpecies_train_set.fa.gz?download=1", multithr
```

```
taxa2.print <- taxa2
rownames(taxa2.print) <- NULL
head(taxa2.print)
```

```
##      Kingdom    Phylum             Class                  Order
## [1,] "Bacteria" "Proteobacteria"   "Alphaproteobacteria" "SAR11 clade"
## [2,] "Bacteria" "Cyanobacteria"    "Cyanobacteriia"      "Synechococcales"
## [3,] "Bacteria" "Proteobacteria"   "Alphaproteobacteria" "SAR11 clade"
## [4,] "Bacteria" "Proteobacteria"   "Alphaproteobacteria" "SAR11 clade"
## [5,] "Bacteria" "Proteobacteria"   "Alphaproteobacteria" "SAR11 clade"
## [6,] "Bacteria" "Actinobacteriota" "Acidimicrobiia"      "Actinomarinales"
##      Family             Genus                      Species
## [1,] "Clade I"          "Clade Ia"                 NA
## [2,] "Cyanobiaceae"     "Synechococcus CC9902"     NA
## [3,] "Clade I"          "Clade Ia"                 NA
## [4,] "Clade I"          "Clade Ia"                 NA
## [5,] "Clade II"         NA                         NA
## [6,] "Actinomarinaceae" "Candidatus Actinomarina" NA
```

Test supp taxo

```
wget http://www2.decipher.codes/Classification/TrainingSets/SILVA_SSU_r138_2019.RData
```

```
dna <- DNAStringSet(getSequences(seqtab.nochim))
load("SILVA_SSU_r138_2019.RData")
ids <- IdTaxa(dna, trainingSet, strand="top", processors=NULL, verbose=FALSE)
ranks <- c("domain", "phylum", "class", "order", "family", "genus", "species")

taxid <- t(sapply(ids, function(x) {
        m <- match(ranks, x$rank)
        taxa <- x$taxon[m]
        taxa[startsWith(taxa, "unclassified_")] <- NA
        taxa
}))
colnames(taxid) <- ranks; rownames(taxid) <- getSequences(seqtab.nochim)
```

Handoff to phyloseq

```
theme_set(theme_bw())
```

```
samples.out <- rownames(seqtab.nochim)
prof <- sapply(strsplit(samples.out, "_"), `[`, 2)

s_prof <- substr(prof,1,1)
day <- as.character(sapply(strsplit(samples.out, "_"), `[`, 3))
```

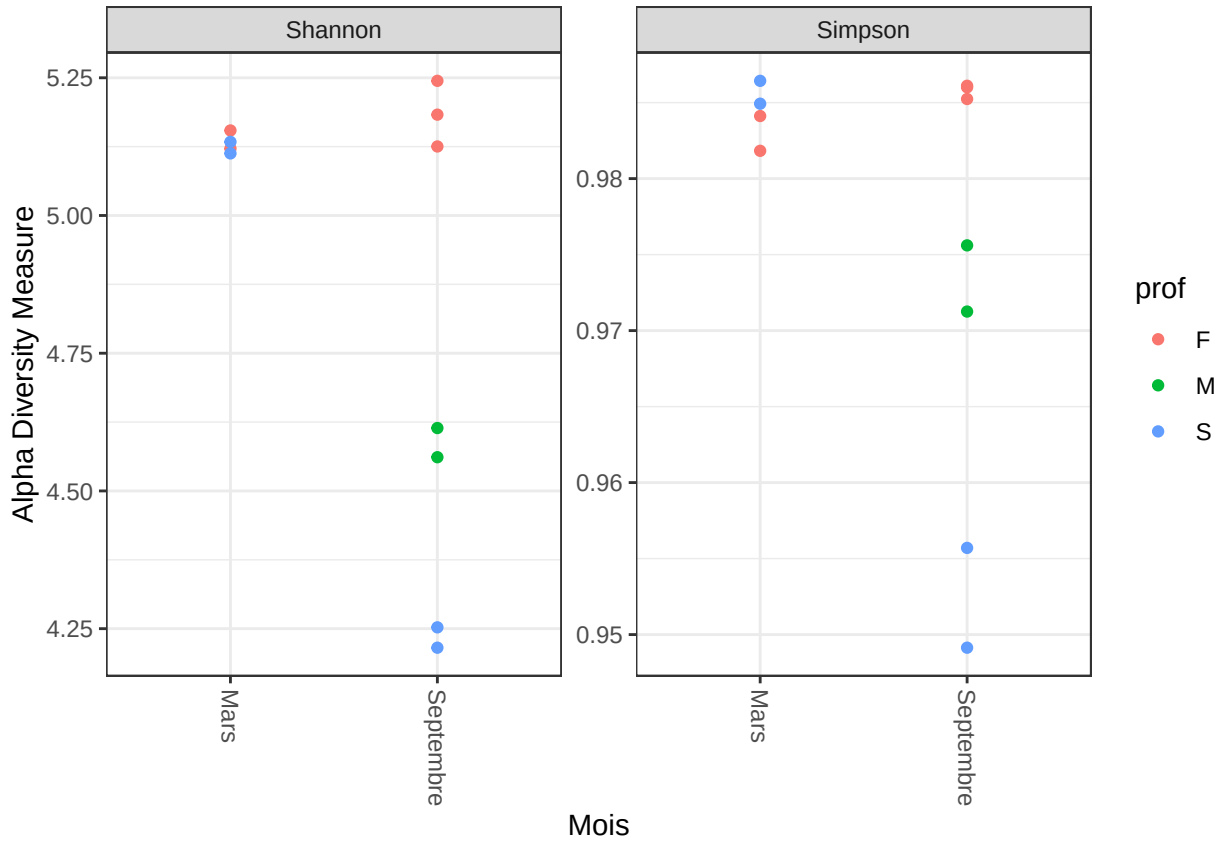
```
samdf <- data.frame(prof=s_prof, Day=day)


samdf$Mois <- "Septembre"
samdf$Mois[samdf$Day > "10sept14"] <- "Mars"

rownames(samdf) <- samples.out
```

```
ps <- phyloseq(otu_table(seqtab.nochim, taxa_are_rows=FALSE),
               sample_data(samdf),
               tax_table(taxa))
```

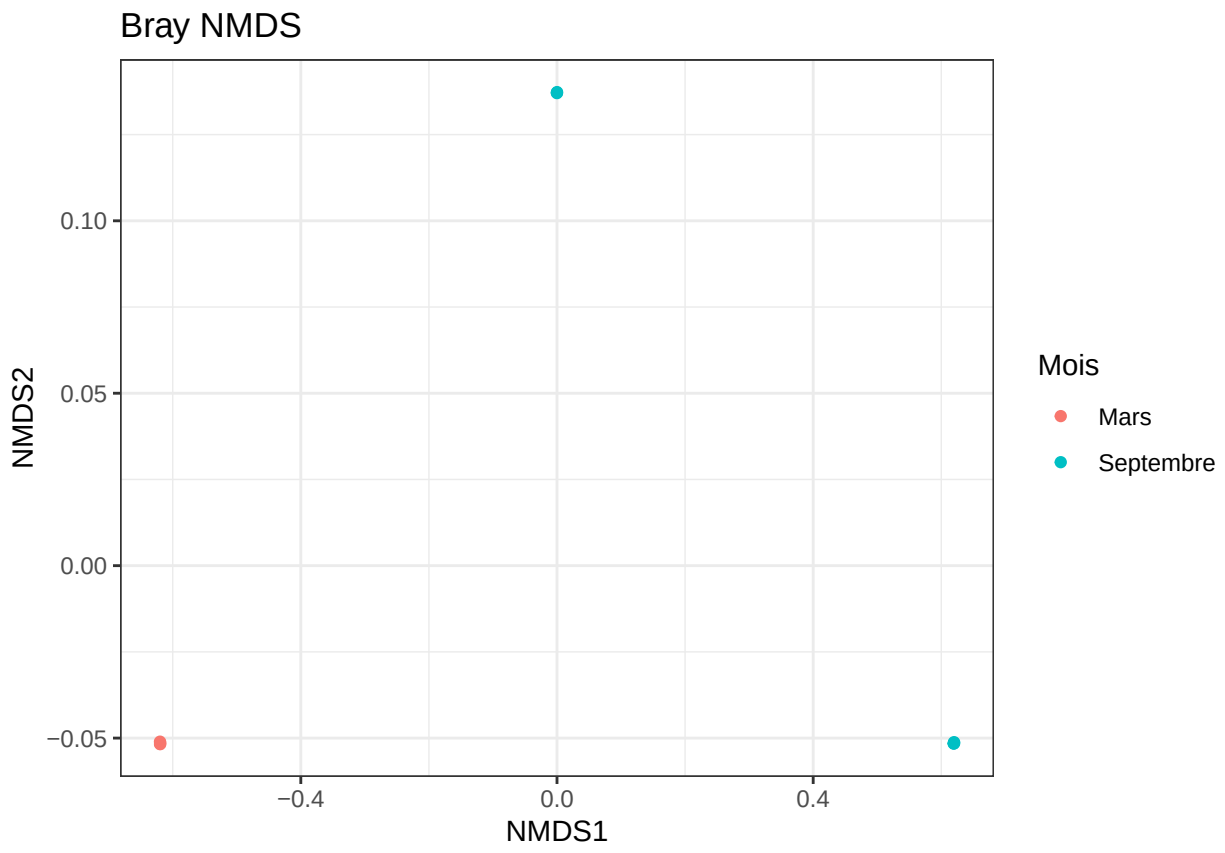
```
dna <- Biostrings::DNAStringSet(taxa_names(ps))
names(dna) <- taxa_names(ps)
ps <- merge_phyloseq(ps, dna)
taxa_names(ps) <- paste0("ASV", seq(ntaxa(ps)))
ps
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1403 taxa and 11 samples ]
## sample_data() Sample Data:       [ 11 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 1403 taxa by 6 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 1403 reference sequences ]
```

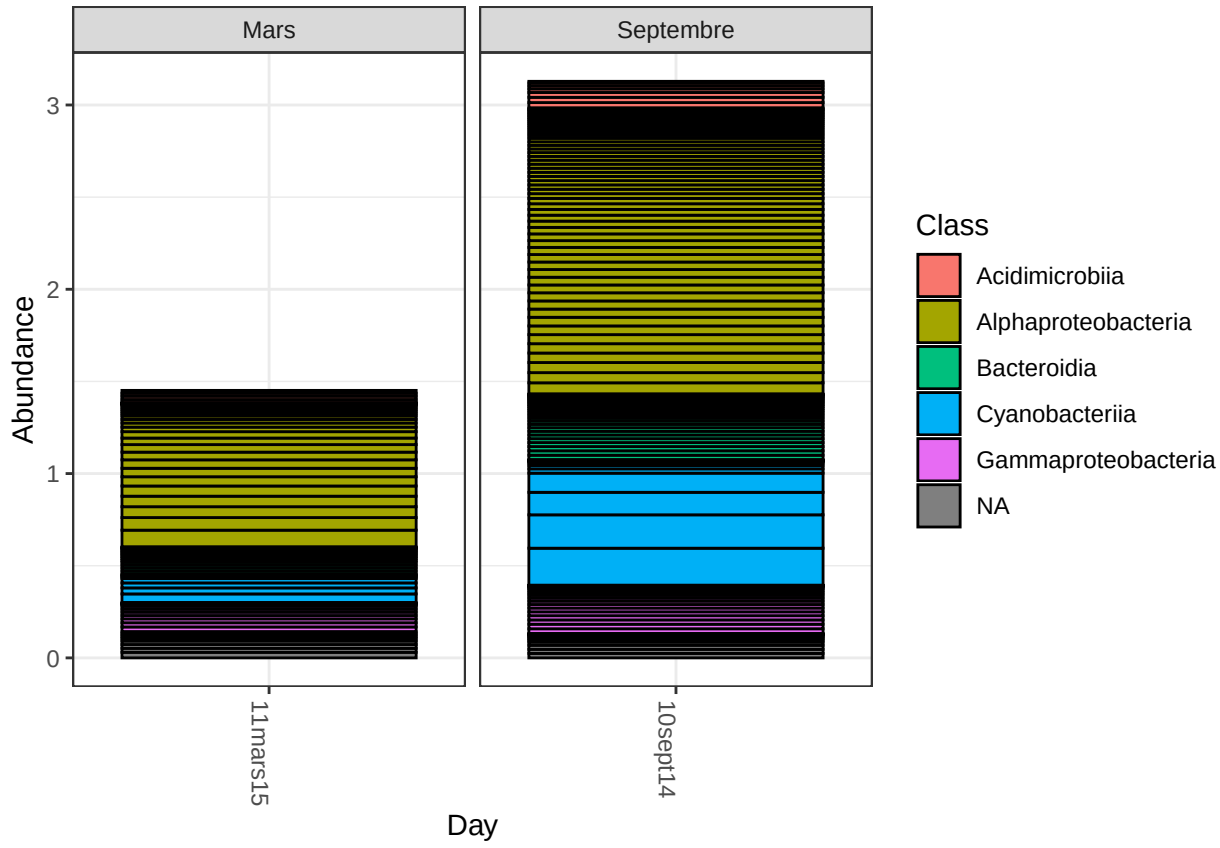
```
plot_richness(ps, x="Mois", measures=c("Shannon", "Simpson"), color="prof")
```



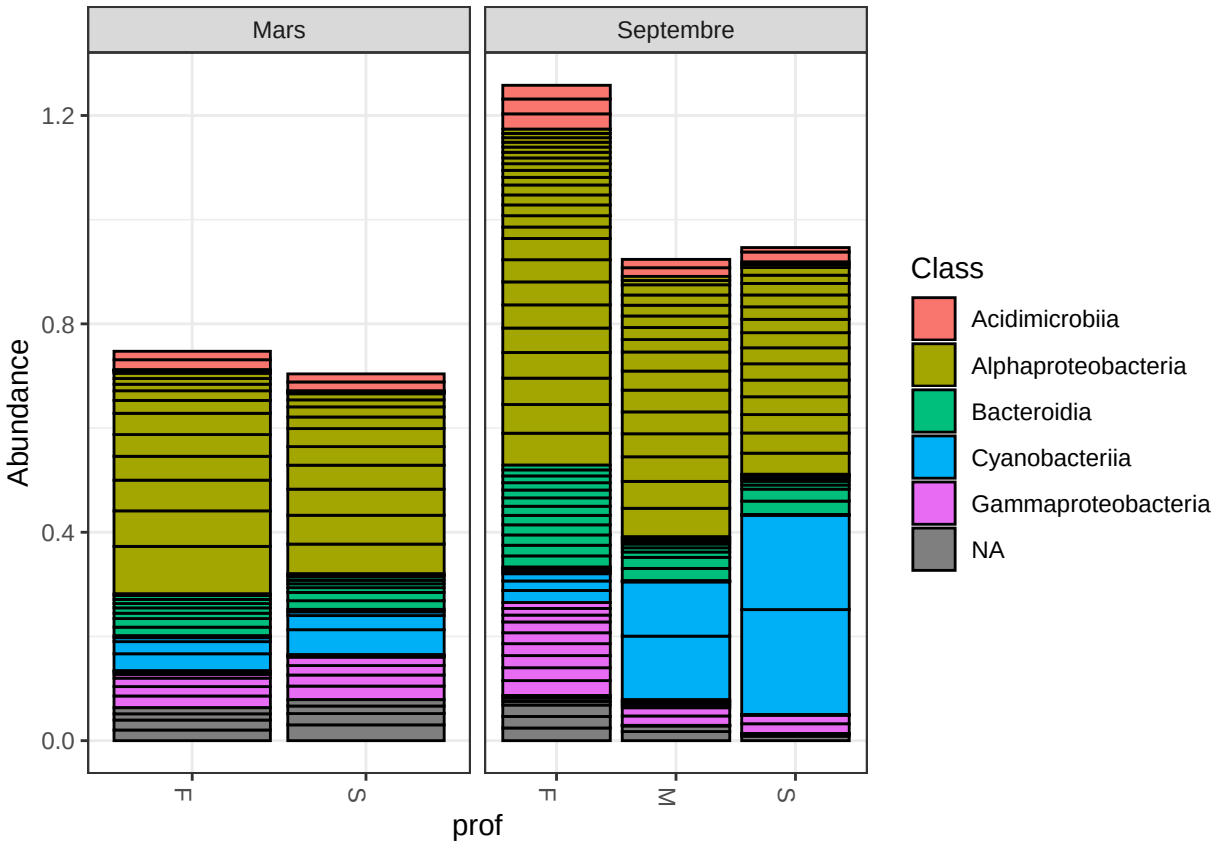
```
plot_ordination(ps.prop, ord.nmds.bray, color="Mois", title="Bray NMDS")
```




```
top20 <- names(sort(taxa_sums(ps), decreasing=TRUE))[1:20]
ps.top20 <- transform_sample_counts(ps, function(OTU) OTU/sum(OTU))
ps.top20 <- prune_taxa(top20, ps.top20)
plot_bar(ps.top20, x="Day", fill="Class") + facet_wrap(~Mois, scales="free_x")
```




```
top20 <- names(sort(taxa_sums(ps), decreasing=TRUE))[1:20]
ps.top20 <- transform_sample_counts(ps, function(OTU) OTU/sum(OTU))
ps.top20 <- prune_taxa(top20, ps.top20)
plot_bar(ps.top20, x="prof", fill="Class") + facet_wrap(~Mois, scales="free_x")
```



Question: Comment les communautées sont impactées par la profondeur et la période d'échantillobnage ?