CC2 Rade De Brest

Mrozinski Alexandre

```
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
```

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

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

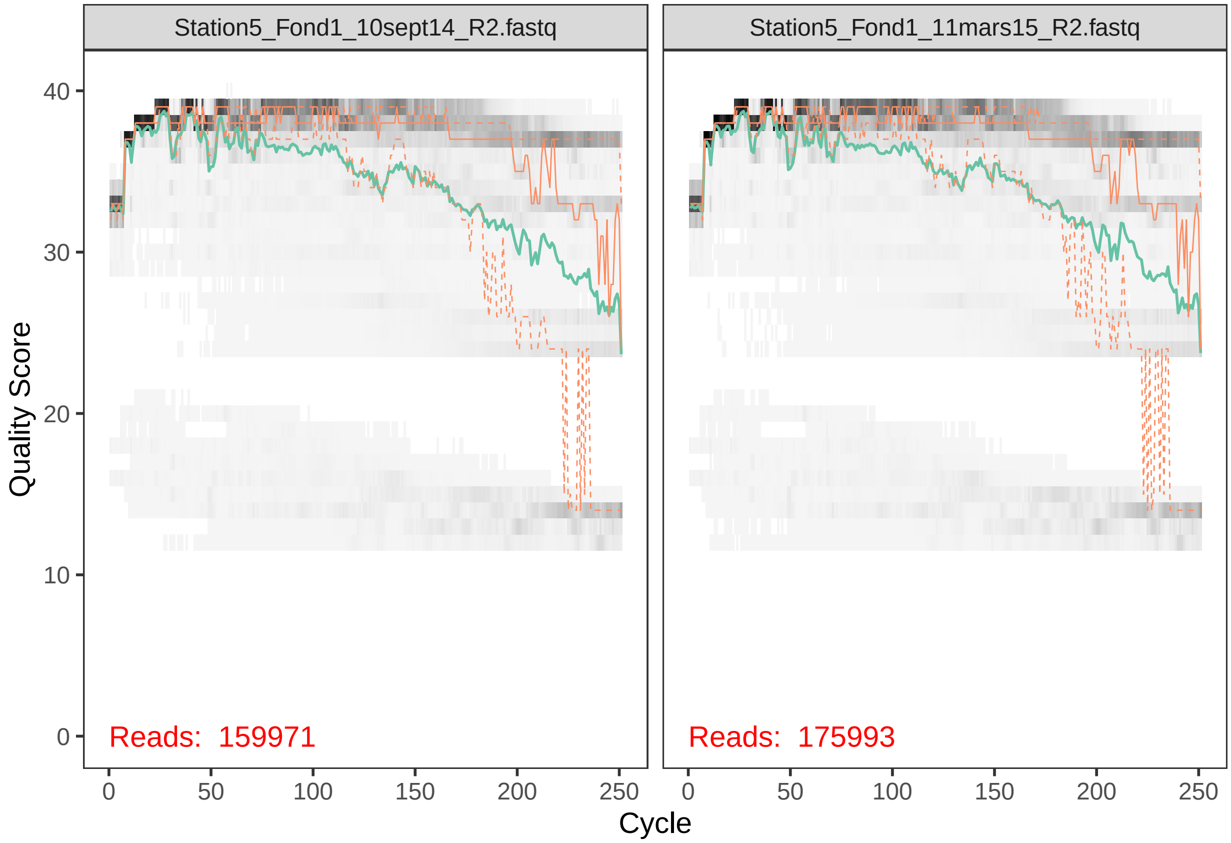
```
##  [1] "filtered"                           "Station5_Fond1_10sept14_R1.fastq"
##  [3] "Station5_Fond1_10sept14_R2.fastq"    "Station5_Fond1_11mars15_R1.fastq"
##  [5] "Station5_Fond1_11mars15_R2.fastq"    "Station5_Fond2_10sept14_R1.fastq"
##  [7] "Station5_Fond2_10sept14_R2.fastq"    "Station5_Fond2_11mars15_R1.fastq"
##  [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)
```

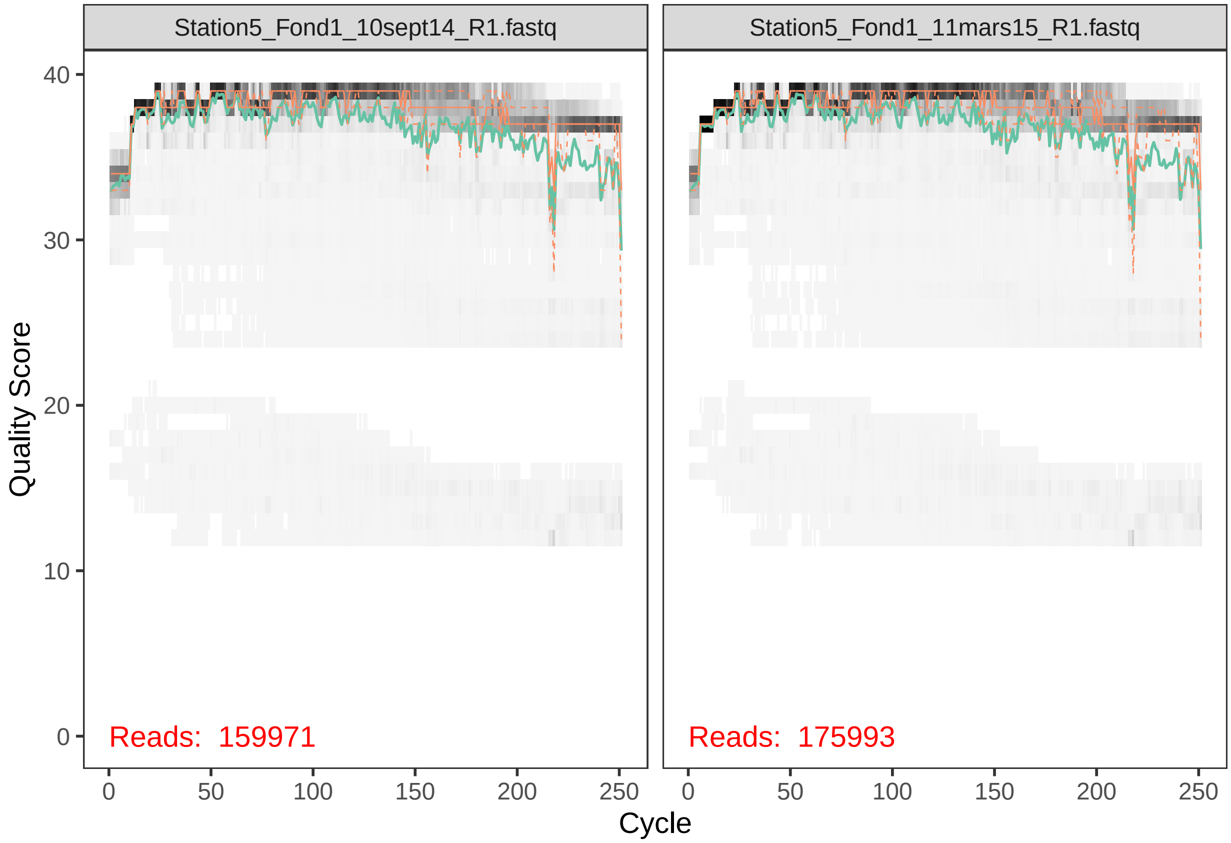
```
plotQualityProfile(fnRs[1:2])
```

```
## 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(240,190), 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    147535
## Station5_Fond1_11mars15_R1.fastq    175993    162532
## Station5_Fond2_10sept14_R1.fastq    197039    179732
## Station5_Fond2_11mars15_R1.fastq     87585     80998
## Station5_Fond3_10sept14_R1.fastq    117140    107720
## Station5_Median1_10sept14_R1.fastq  116519    108074
```

#Learn the Error Rates

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

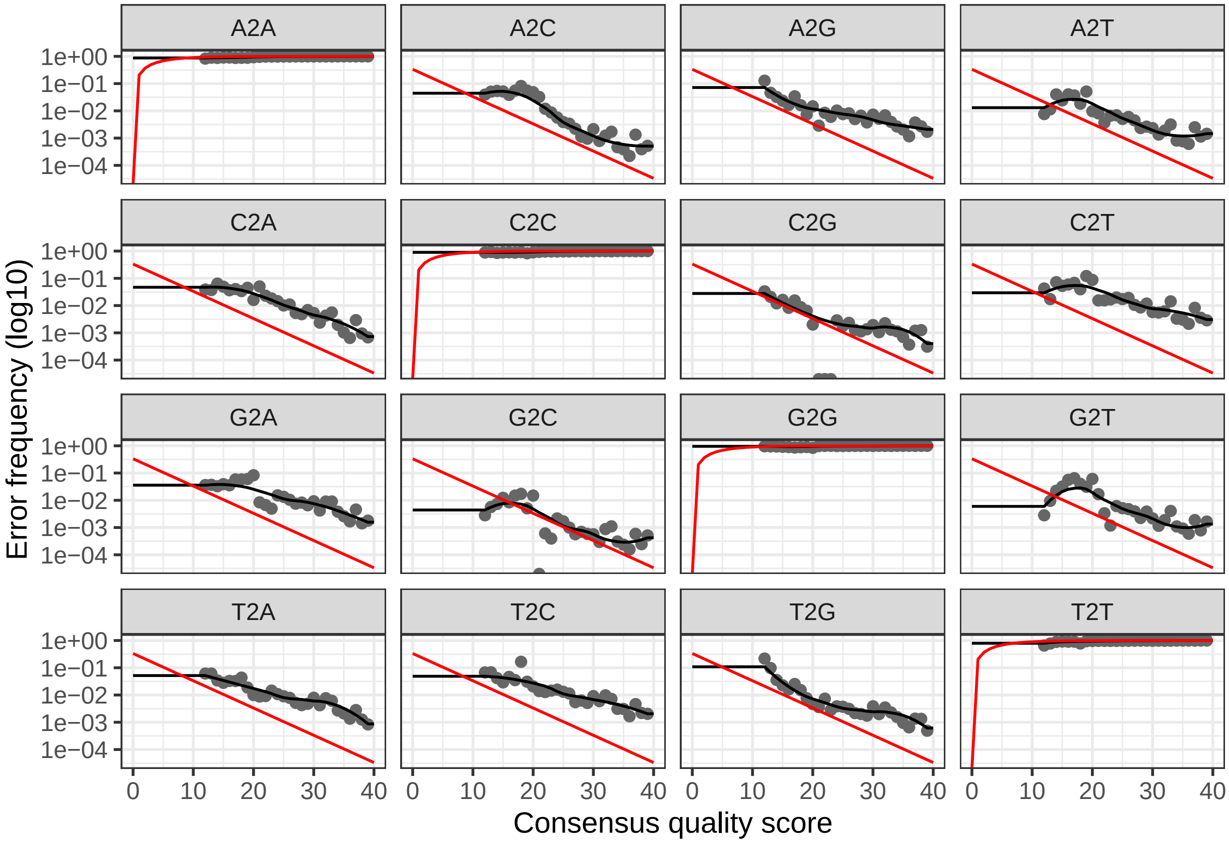
```
## 108735378 total bases in 489799 reads from 3 samples will be used for learning the error rates.
```

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

```
## 116704924 total bases in 678517 reads from 5 samples will be used for learning the error rates.
```

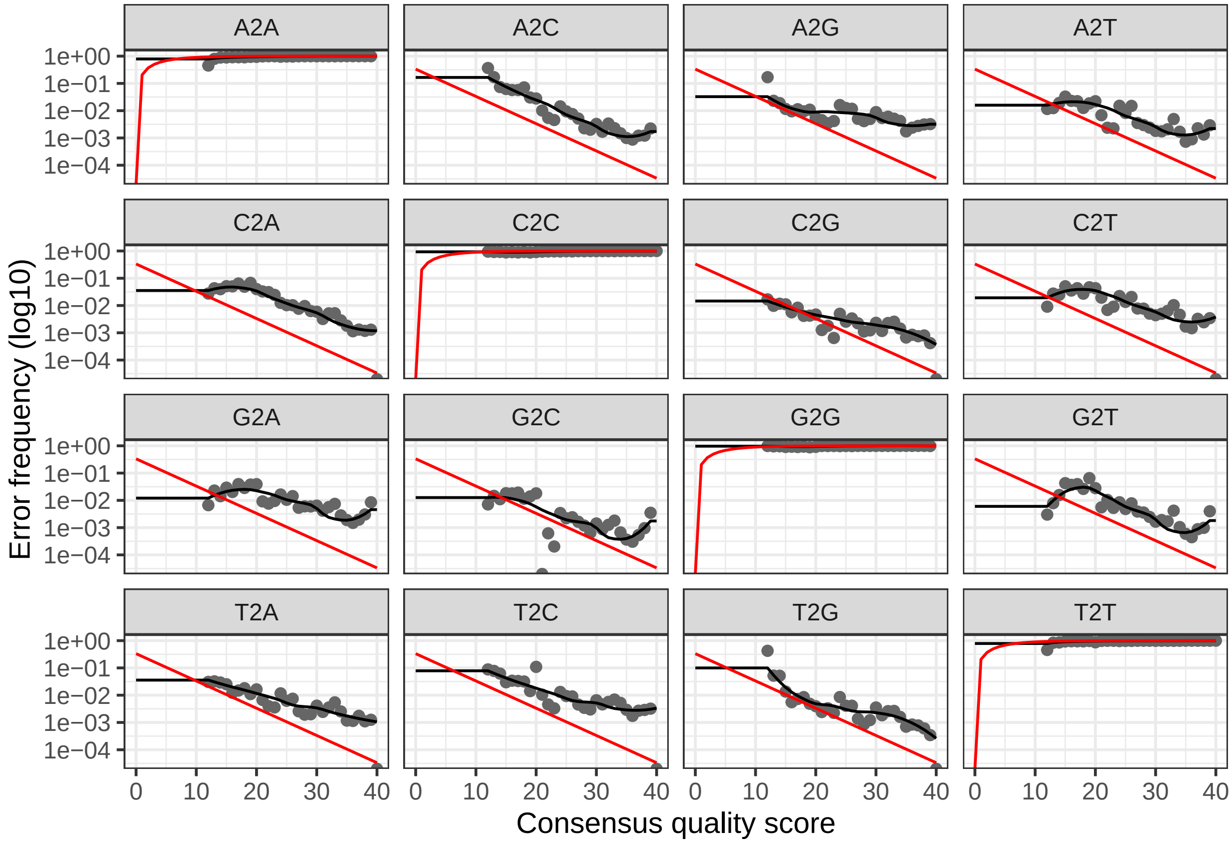
```
plotErrors(errFs, nominalQ=TRUE)
```

```
## 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 - 147535 reads in 38976 unique sequences.
## Sample 2 - 162532 reads in 36882 unique sequences.
## Sample 3 - 179732 reads in 48636 unique sequences.
## Sample 4 - 80998 reads in 20872 unique sequences.
## Sample 5 - 107720 reads in 31095 unique sequences.
## Sample 6 - 108074 reads in 29566 unique sequences.
## Sample 7 - 100124 reads in 26531 unique sequences.
## Sample 8 - 108790 reads in 27456 unique sequences.
## Sample 9 - 72045 reads in 18459 unique sequences.
## Sample 10 - 79849 reads in 21120 unique sequences.
## Sample 11 - 92833 reads in 25156 unique sequences.
```

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

```
## Sample 1 - 147535 reads in 44763 unique sequences.
## Sample 2 - 162532 reads in 40966 unique sequences.
## Sample 3 - 179732 reads in 54836 unique sequences.
## Sample 4 - 80998 reads in 22827 unique sequences.
## Sample 5 - 107720 reads in 34178 unique sequences.
## Sample 6 - 108074 reads in 31119 unique sequences.
## Sample 7 - 100124 reads in 28632 unique sequences.
## Sample 8 - 108790 reads in 28466 unique sequences.
## Sample 9 - 72045 reads in 21082 unique sequences.
## Sample 10 - 79849 reads in 21711 unique sequences.
## Sample 11 - 92833 reads in 27892 unique sequences.
```

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

```
## dada-class: object describing DADA2 denoising results
## 1022 sequence variants were inferred from 38976 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
## 865 sequence variants were inferred from 44763 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)
```

```
## 119776 paired-reads (in 5832 unique pairings) successfully merged out of 142625 (in 21937 pairings) i

## 141112 paired-reads (in 4783 unique pairings) successfully merged out of 158255 (in 16105 pairings) i

## 146248 paired-reads (in 8044 unique pairings) successfully merged out of 174368 (in 27989 pairings) i

## 68575 paired-reads (in 3032 unique pairings) successfully merged out of 78421 (in 9690 pairings) inpu

## 85589 paired-reads (in 3872 unique pairings) successfully merged out of 103901 (in 16752 pairings) in

## 88361 paired-reads (in 3898 unique pairings) successfully merged out of 104722 (in 14711 pairings) in

## 82504 paired-reads (in 3152 unique pairings) successfully merged out of 96988 (in 12777 pairings) inp

## 91011 paired-reads (in 3495 unique pairings) successfully merged out of 105437 (in 12689 pairings) in

## 60780 paired-reads (in 2186 unique pairings) successfully merged out of 69684 (in 8160 pairings) inpu

## 67559 paired-reads (in 2025 unique pairings) successfully merged out of 77715 (in 8532 pairings) inpu

## 76401 paired-reads (in 3547 unique pairings) successfully merged out of 89445 (in 12240 pairings) inp
```

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

```
##
## 1     TAATACGAAGGGACCTAGCGTAGTTCGGAATTACTGGGCTTAAAGAGTTCGTAGGTGGTTGAAAAAGTTAGTGGTGAAATCCCAGAGCTTAACT
## 2     TAATACGAAGGGACCTAGCGTAGTTCGGAATTACTGGGCTTAAAGAGTTCGTAGGTGGTTGAAAAAGTTGGTGGTGAAATCCCAGAGCTTAACT
## 3     TAATACGAAGGGACCTAGCGTAGTTCGGAATTACTGGGCTTAAAGAGTTCGTAGGTGGTTGAAAAAGTTGGTGGTGAAATCCCAGAGCTTAACT
## 4     TAATACGAAGGGACCTAGCGTAGTTCGGAATTACTGGGCTTAAAGAGTTCGTAGGTGGTTGAAAAAGTTAGTGGTGAAATCCCAGAGCTTAACT
## 5     TAATACGAAGGGACCTAGCGTAGTTCGGAATTACTGGGCTTAAAGAGTTCGTAGGTGGTTGAAAAAGTTGGTGGTGAAATCCCAGAGCTTAACT
## 6 TAATACGAGGGGTCCTAGCGTTGTCCGGATTTACTGGGCGTAAAGGGTACGTAGGCGTTTTAATAAGTTGTATGTTAAATATCTTAGCTTAACTAAGA
##   abundance forward reverse nmatch nmismatch nindel prefer accept
## 1      5218       1       2     19         0      0      2   TRUE
## 2      4153       2       1     19         0      0      2   TRUE
## 3      3777       3       1     19         0      0      2   TRUE
## 4      2508       1       1     19         0      0      2   TRUE
## 5      2201       2       2     19         0      0      2   TRUE
## 6      2176       6       9     15         0      0      1   TRUE
```

#Construct sequence table

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

```
## [1]    11 22274
```

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

```
##
##  232  358  359  368  369  370  371  372  373  374  375  376  377  378  379  380
##    1    1    1    1    1    4  208   28  177  228 5855 4447 2614 2944 3216  138
##  381  382
## 2313   97
```

#Remove chimeras

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

```
## Identified 20629 bimeras out of 22274 input sequences.
```

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

```
## [1]   11 1645
```

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

```
## [1] 0.770633
```

#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   147535    144485    145419 119776   89077
## Station5_Fond1_11mars15   175993   162532    159906    160607 141112  113032
## Station5_Fond2_10sept14   197039   179732    176245    177593 146248  105321
## Station5_Fond2_11mars15    87585    80998     79347     79864  68575   55221
## Station5_Fond3_10sept14   117140   107720    105293    106117  85589   65358
## Station5_Median1_10sept14 116519   108074    106071    106540  88361   66321
```

#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"
```

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

```
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)
station <- sapply(strsplit(samples.out, "_"), `[`, 2)

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


samdf <- data.frame(Profondeur=profondeur, Day=day)

samdf$Saison <- "Ete"
samdf$Saison[samdf$Day > "10sept14"] <- "Hiver"

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

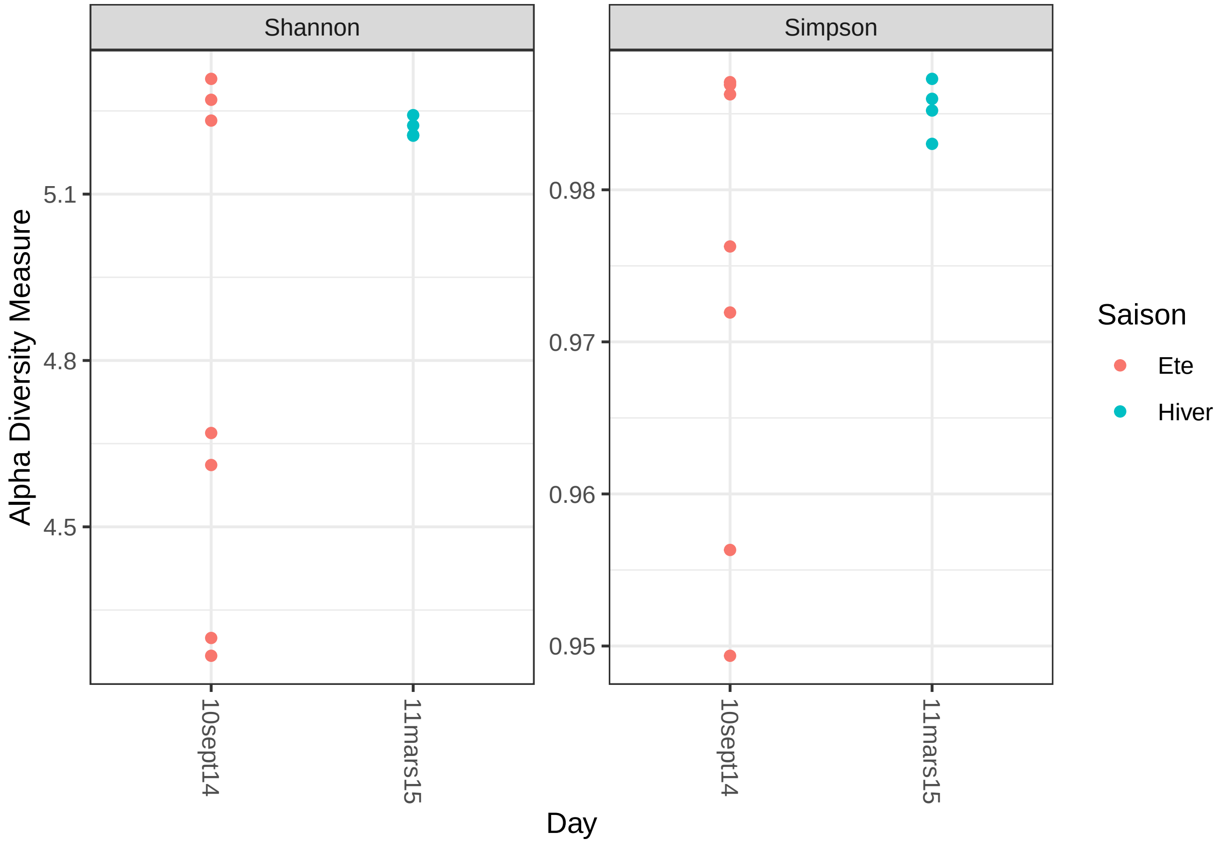
```
##                              Profondeur      Day Saison
## Station5_Fond1_10sept14              F 10sept14    Ete
## Station5_Fond1_11mars15              F 11mars15  Hiver
## Station5_Fond2_10sept14              F 10sept14    Ete
## Station5_Fond2_11mars15              F 11mars15  Hiver
## Station5_Fond3_10sept14              F 10sept14    Ete
## Station5_Median1_10sept14            M 10sept14    Ete
## Station5_Median2_10sept14            M 10sept14    Ete
## Station5_Surface1_10sept14           S 10sept14    Ete
## Station5_Surface1_11mars15           S 11mars15  Hiver
## Station5_Surface2_10sept14           S 10sept14    Ete
## Station5_Surface2_11mars15           S 11mars15  Hiver
```

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

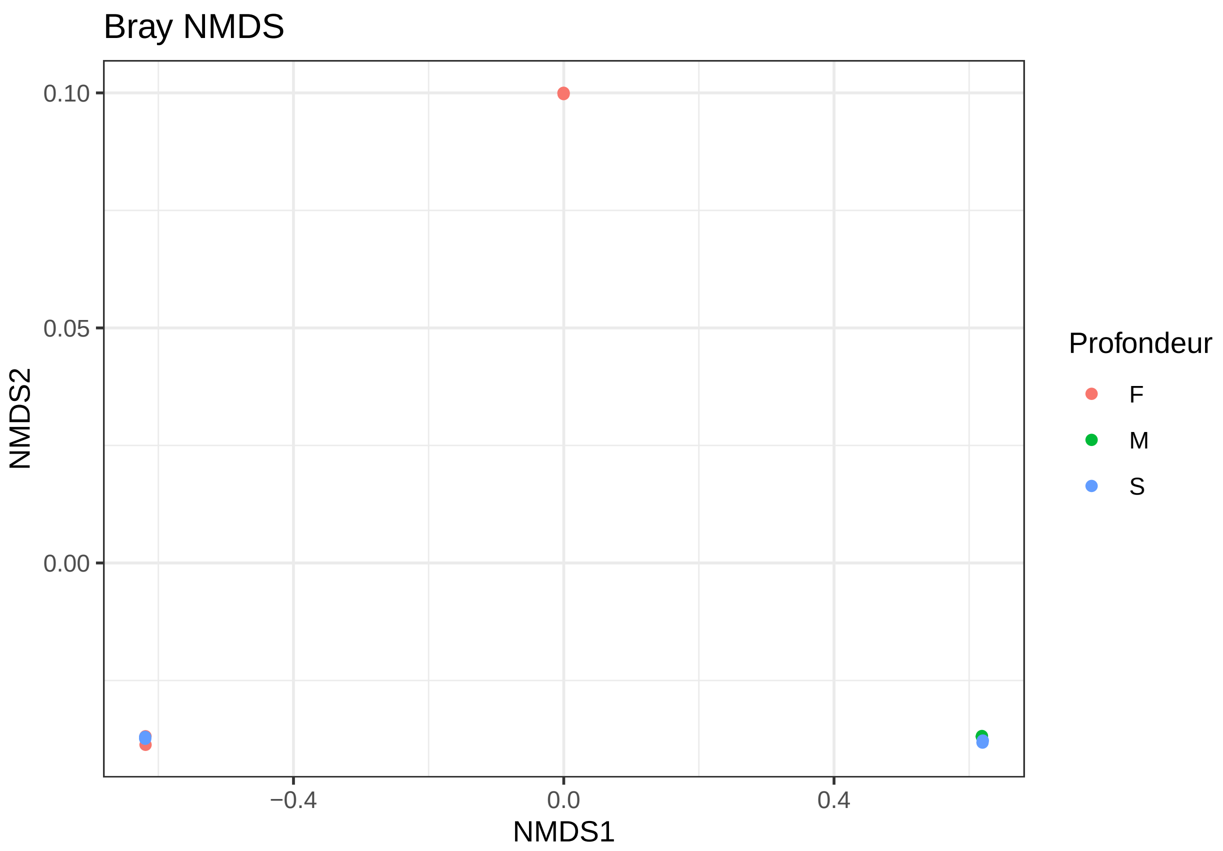
```r
dna <- Biostrings::DNAStringSet(taxa_names(ps))
names(dna) <- taxa_names(ps)
ps <- merge_phyloseq(ps, dna)
taxa_names(ps) <- paste0("ASV", seq(ntaxa(ps)))
ps
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1645 taxa and 11 samples ]
## sample_data() Sample Data:       [ 11 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 1645 taxa by 6 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 1645 reference sequences ]
```

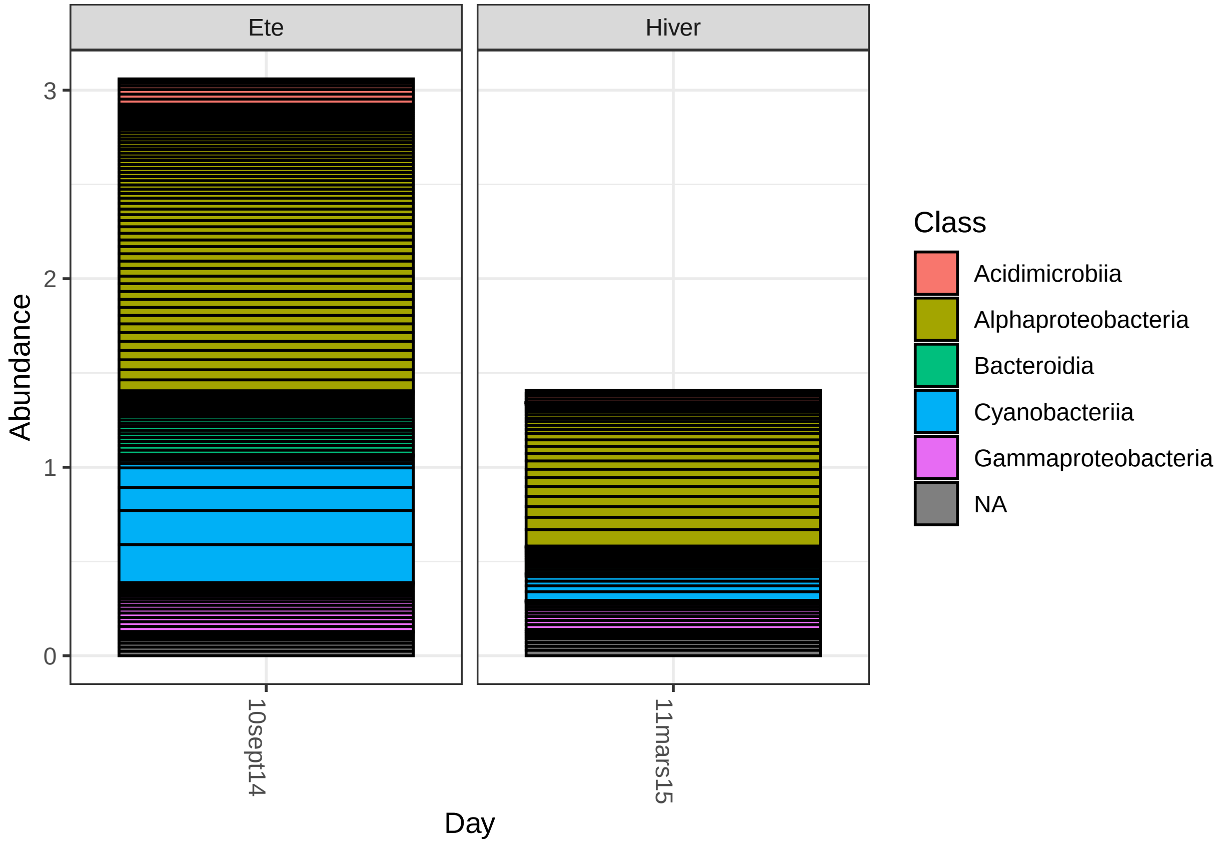
```r
plot_richness(ps, x="Day", measures=c("Shannon", "Simpson"), color="Saison")
```



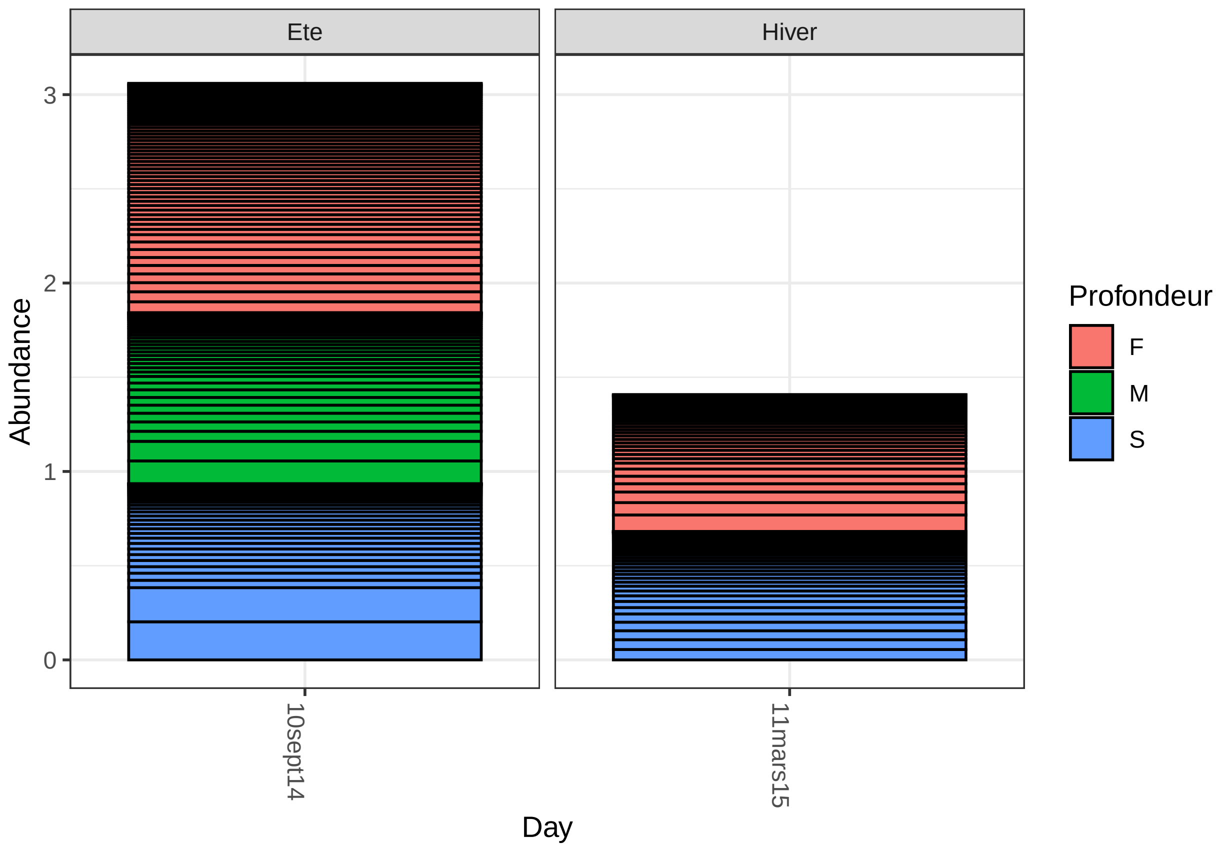
```r
plot_ordination(ps.prop, ord.nmds.bray, color="Profondeur", title="Bray NMDS")
```



```
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(~Saison, 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="Day", fill="Profondeur") + facet_wrap(~Saison, scales="free_x")
```


Ete
Hiver
Abundance
0
1
2
3
10sept14
11mars15
Day
Profondeur
F
M
S