Dada2 Processing

Anastasia Poluzerova

2023-06-07

```
knitr::opts_chunk$set(fig.width=14, fig.height=8)

library('dada2')
library('phyloseq')
library('dplyr')



set.seed(1609)
setwd('/home/nastasista/Metagenomics')
```

Read files and data

```
path <- '/home/nastasista/Metagenomics/data_met/sequences'
list.files(path)
```

```
##  [1] "Abacumov-B-1_S1_L001_R1_001.fastq.gz"
##  [2] "Abacumov-B-1_S1_L001_R2_001.fastq.gz"
##  [3] "Abacumov-B-13_S13_L001_R1_001.fastq.gz"
##  [4] "Abacumov-B-13_S13_L001_R2_001.fastq.gz"
##  [5] "Abacumov-B-14_S14_L001_R1_001.fastq.gz"
##  [6] "Abacumov-B-14_S14_L001_R2_001.fastq.gz"
##  [7] "Abacumov-B-15_S15_L001_R1_001.fastq.gz"
##  [8] "Abacumov-B-15_S15_L001_R2_001.fastq.gz"
##  [9] "Abacumov-B-16_S16_L001_R1_001.fastq.gz"
## [10] "Abacumov-B-16_S16_L001_R2_001.fastq.gz"
## [11] "Abacumov-B-2_S2_L001_R1_001.fastq.gz"
## [12] "Abacumov-B-2_S2_L001_R2_001.fastq.gz"
## [13] "Abacumov-B-25_S25_L001_R1_001.fastq.gz"
## [14] "Abacumov-B-25_S25_L001_R2_001.fastq.gz"
## [15] "Abacumov-B-26_S26_L001_R1_001.fastq.gz"
## [16] "Abacumov-B-26_S26_L001_R2_001.fastq.gz"
## [17] "Abacumov-B-27_S27_L001_R1_001.fastq.gz"
## [18] "Abacumov-B-27_S27_L001_R2_001.fastq.gz"
## [19] "Abacumov-B-28_S28_L001_R1_001.fastq.gz"
## [20] "Abacumov-B-28_S28_L001_R2_001.fastq.gz"
## [21] "Abacumov-B-3_S3_L001_R1_001.fastq.gz"
## [22] "Abacumov-B-3_S3_L001_R2_001.fastq.gz"
## [23] "Abacumov-B-37_S37_L001_R1_001.fastq.gz"
## [24] "Abacumov-B-37_S37_L001_R2_001.fastq.gz"
## [25] "Abacumov-B-38_S38_L001_R1_001.fastq.gz"
## [26] "Abacumov-B-38_S38_L001_R2_001.fastq.gz"
## [27] "Abacumov-B-39_S39_L001_R1_001.fastq.gz"
## [28] "Abacumov-B-39_S39_L001_R2_001.fastq.gz"
## [29] "Abacumov-B-4_S4_L001_R1_001.fastq.gz"
## [30] "Abacumov-B-4_S4_L001_R2_001.fastq.gz"
## [31] "Abacumov-B-40_S40_L001_R1_001.fastq.gz"
## [32] "Abacumov-B-40_S40_L001_R2_001.fastq.gz"
## [33] "Abacumov-B-49_S49_L001_R1_001.fastq.gz"
## [34] "Abacumov-B-49_S49_L001_R2_001.fastq.gz"
## [35] "Abacumov-B-50_S50_L001_R1_001.fastq.gz"
## [36] "Abacumov-B-50_S50_L001_R2_001.fastq.gz"
## [37] "Abacumov-B-51_S51_L001_R1_001.fastq.gz"
## [38] "Abacumov-B-51_S51_L001_R2_001.fastq.gz"
## [39] "Abacumov-B-52_S52_L001_R1_001.fastq.gz"
## [40] "Abacumov-B-52_S52_L001_R2_001.fastq.gz"
## [41] "Abacumov-B-61_S61_L001_R1_001.fastq.gz"
## [42] "Abacumov-B-61_S61_L001_R2_001.fastq.gz"
## [43] "Abacumov-B-62_S62_L001_R1_001.fastq.gz"
## [44] "Abacumov-B-62_S62_L001_R2_001.fastq.gz"
## [45] "Abacumov-B-63_S63_L001_R1_001.fastq.gz"
## [46] "Abacumov-B-63_S63_L001_R2_001.fastq.gz"
## [47] "Abacumov-B-64_S64_L001_R1_001.fastq.gz"
## [48] "Abacumov-B-64_S64_L001_R2_001.fastq.gz"
## [49] "filtered"
```

```
metadata <- read.csv('data_met/map.csv')
metadata$SampleID <- paste(metadata$Source, metadata$Site, metadata$Horizont, metadata$Repeat, sep=".")
metadata
```

```
##         Filename             Source Site Horizont Repeat
## 1   Abacumov-B-1 Self-growing Dumps   B1       AY      1
## 2   Abacumov-B-2 Self-growing Dumps   B1       AY      2
## 3   Abacumov-B-3 Self-growing Dumps   B1       AY      3
## 4   Abacumov-B-4 Self-growing Dumps   B1       AY      4
## 5  Abacumov-B-13          Litostrat   B2        C      1
## 6  Abacumov-B-14          Litostrat   B2        C      2
## 7  Abacumov-B-15          Litostrat   B2        C      3
## 8  Abacumov-B-16          Litostrat   B2        C      4
## 9  Abacumov-B-25 Coal Mine Terricon   B3        C      1
## 10 Abacumov-B-26 Coal Mine Terricon   B3        C      2
## 11 Abacumov-B-27 Coal Mine Terricon   B3        C      3
## 12 Abacumov-B-28 Coal Mine Terricon   B3        C      4
## 13 Abacumov-B-37    Local Reference   B4       AY      1
## 14 Abacumov-B-38    Local Reference   B4       AY      2
## 15 Abacumov-B-39    Local Reference   B4       AY      3
## 16 Abacumov-B-40    Local Reference   B4       AY      4
## 17 Abacumov-B-49        Embryo Sand   B5       AY      1
## 18 Abacumov-B-50        Embryo Sand   B5       AY      2
## 19 Abacumov-B-51        Embryo Sand   B5       AY      3
## 20 Abacumov-B-52        Embryo Sand   B5       AY      4
## 21 Abacumov-B-61 Regional Reference   B6       AY      1
## 22 Abacumov-B-62 Regional Reference   B6       AY      2
## 23 Abacumov-B-63 Regional Reference   B6       AY      3
## 24 Abacumov-B-64 Regional Reference   B6       AY      4
##                      SampleID
## 1  Self-growing Dumps.B1.AY.1
## 2  Self-growing Dumps.B1.AY.2
## 3  Self-growing Dumps.B1.AY.3
## 4  Self-growing Dumps.B1.AY.4
## 5            Litostrat.B2.C.1
## 6            Litostrat.B2.C.2
## 7            Litostrat.B2.C.3
## 8            Litostrat.B2.C.4
## 9   Coal Mine Terricon.B3.C.1
## 10  Coal Mine Terricon.B3.C.2
## 11  Coal Mine Terricon.B3.C.3
## 12  Coal Mine Terricon.B3.C.4
## 13    Local Reference.B4.AY.1
## 14    Local Reference.B4.AY.2
## 15    Local Reference.B4.AY.3
## 16    Local Reference.B4.AY.4
## 17        Embryo Sand.B5.AY.1
## 18        Embryo Sand.B5.AY.2
## 19        Embryo Sand.B5.AY.3
## 20        Embryo Sand.B5.AY.4
## 21 Regional Reference.B6.AY.1
## 22 Regional Reference.B6.AY.2
## 23 Regional Reference.B6.AY.3
## 24 Regional Reference.B6.AY.4
```

Run DADA2 pipeline

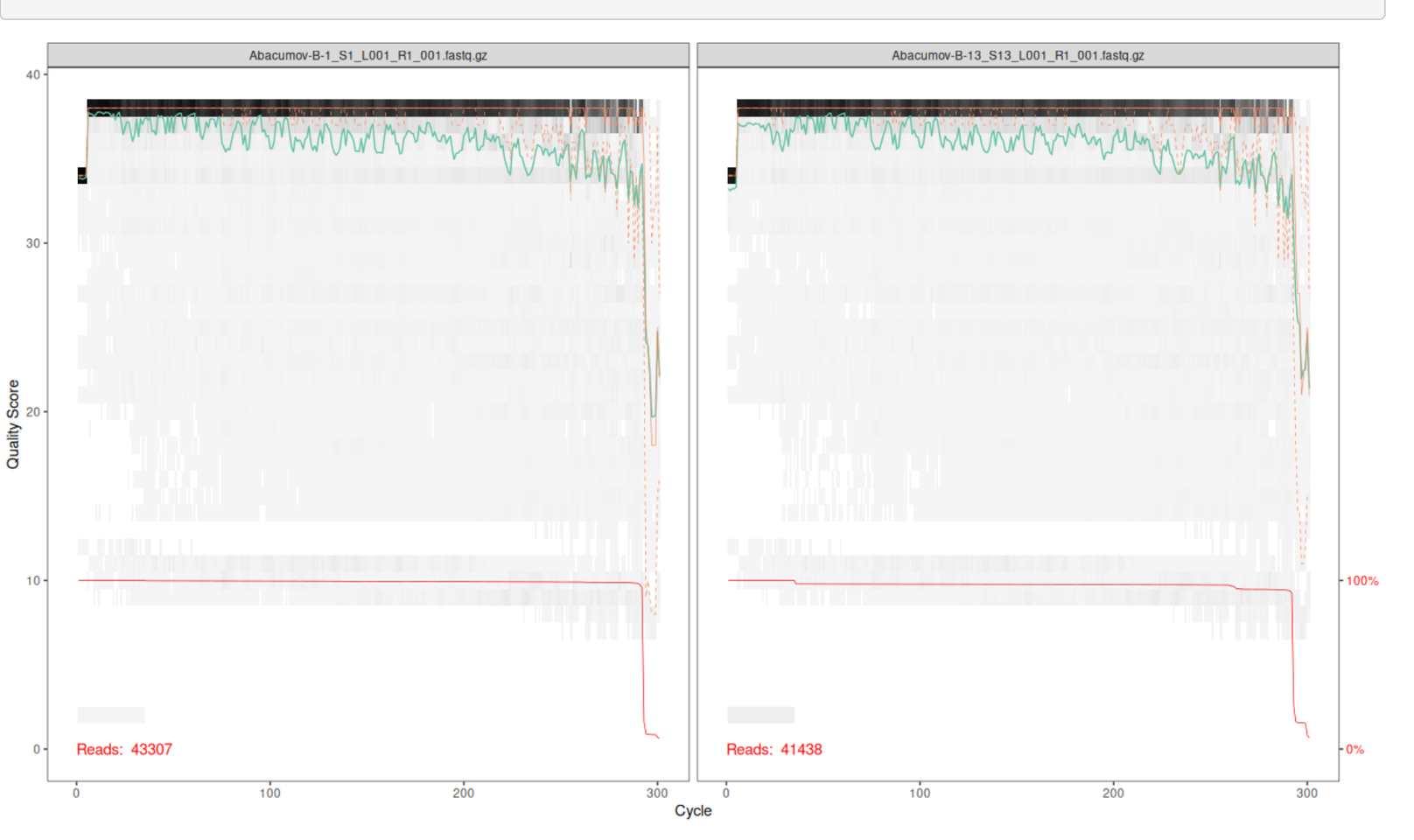
A realisation of a basic tutorial from https://benjjneb.github.io/dada2/tutorial.html

```
# Forward and reverse fastq filenames have format: SAMPLENAME_R1_001.fastq and SAMPLENAME_R2_001.fastq
fnFs <- sort(list.files(path, pattern="_R1_001.fastq", full.names = TRUE))
fnRs <- sort(list.files(path, pattern="_R2_001.fastq", full.names = TRUE))
# Extract sample names, assuming filenames have format: SAMPLENAME_XXX.fastq
sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
sample.names
```

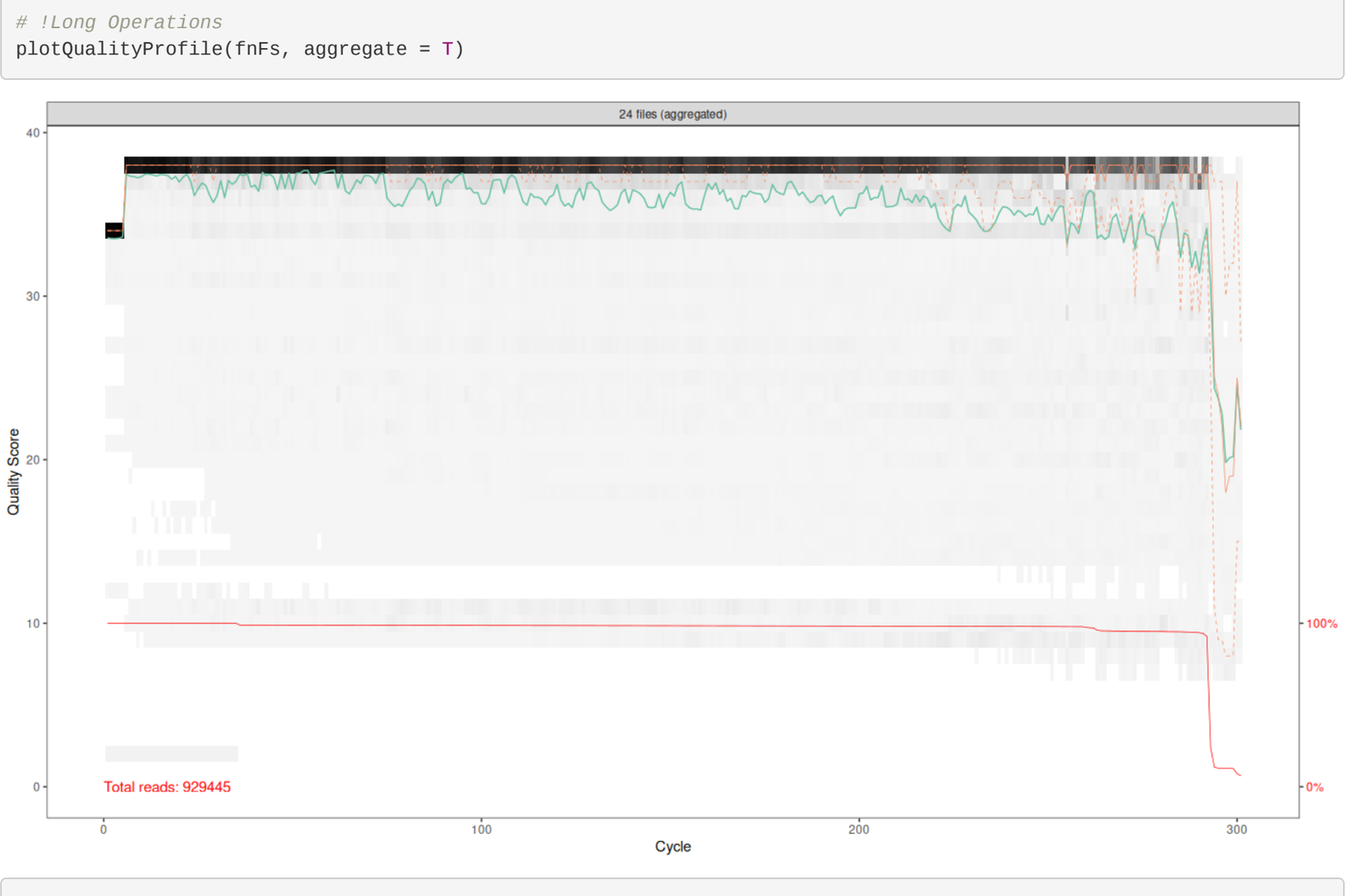
```
##  [1] "Abacumov-B-1"  "Abacumov-B-13" "Abacumov-B-14" "Abacumov-B-15"
##  [5] "Abacumov-B-16" "Abacumov-B-2"  "Abacumov-B-25" "Abacumov-B-26"
##  [9] "Abacumov-B-27" "Abacumov-B-28" "Abacumov-B-3"  "Abacumov-B-37"
## [13] "Abacumov-B-38" "Abacumov-B-39" "Abacumov-B-4"  "Abacumov-B-40"
## [17] "Abacumov-B-49" "Abacumov-B-50" "Abacumov-B-51" "Abacumov-B-52"
## [21] "Abacumov-B-61" "Abacumov-B-62" "Abacumov-B-63" "Abacumov-B-64"
```

Quality plot

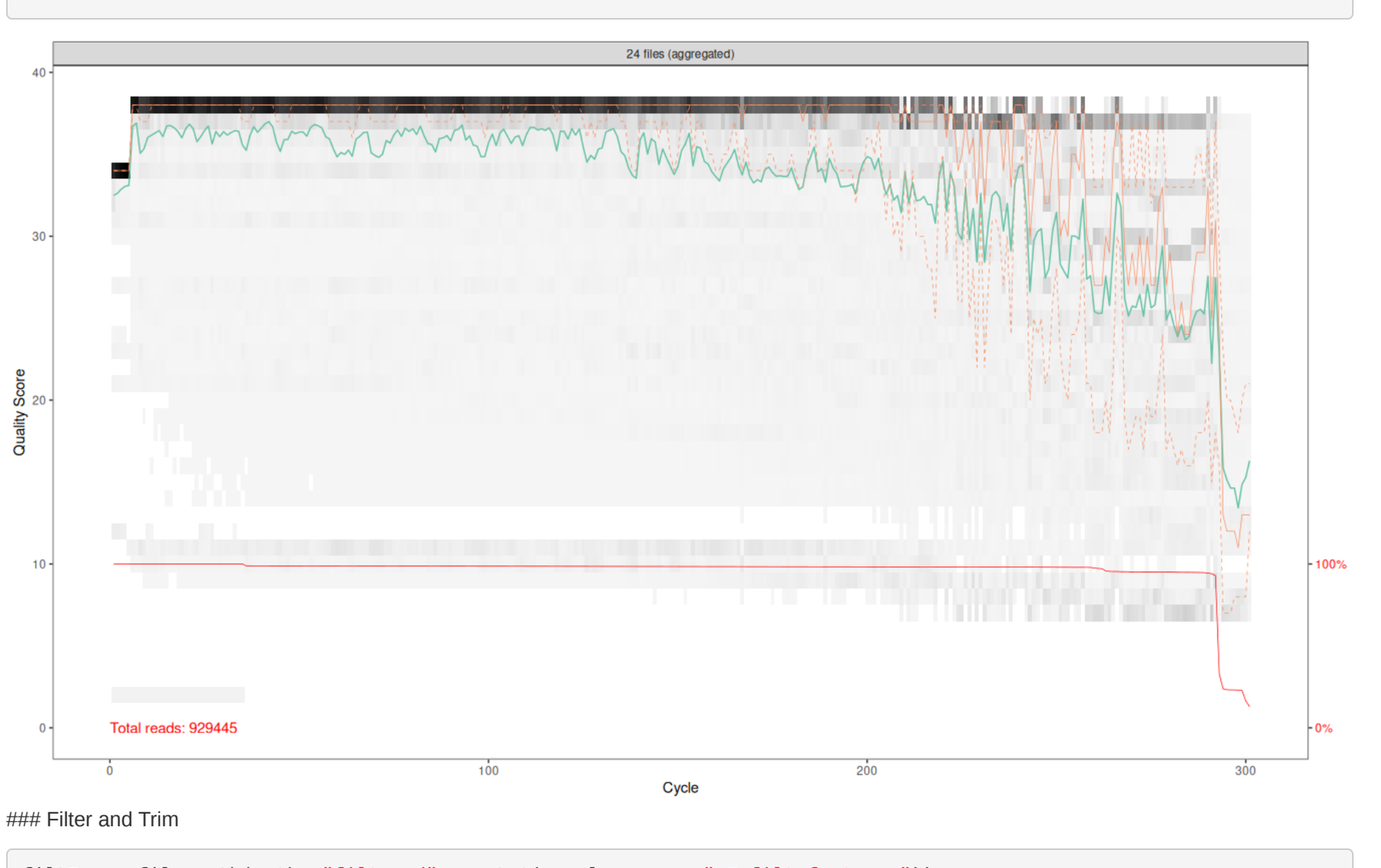
```
plotQualityProfile(fnFs[1:2])
```




```
# !Long Operations
plotQualityProfile(fnFs, aggregate = T)
```




```
plotQualityProfile(fnRs, aggregate = T)
```



Filter and Trim

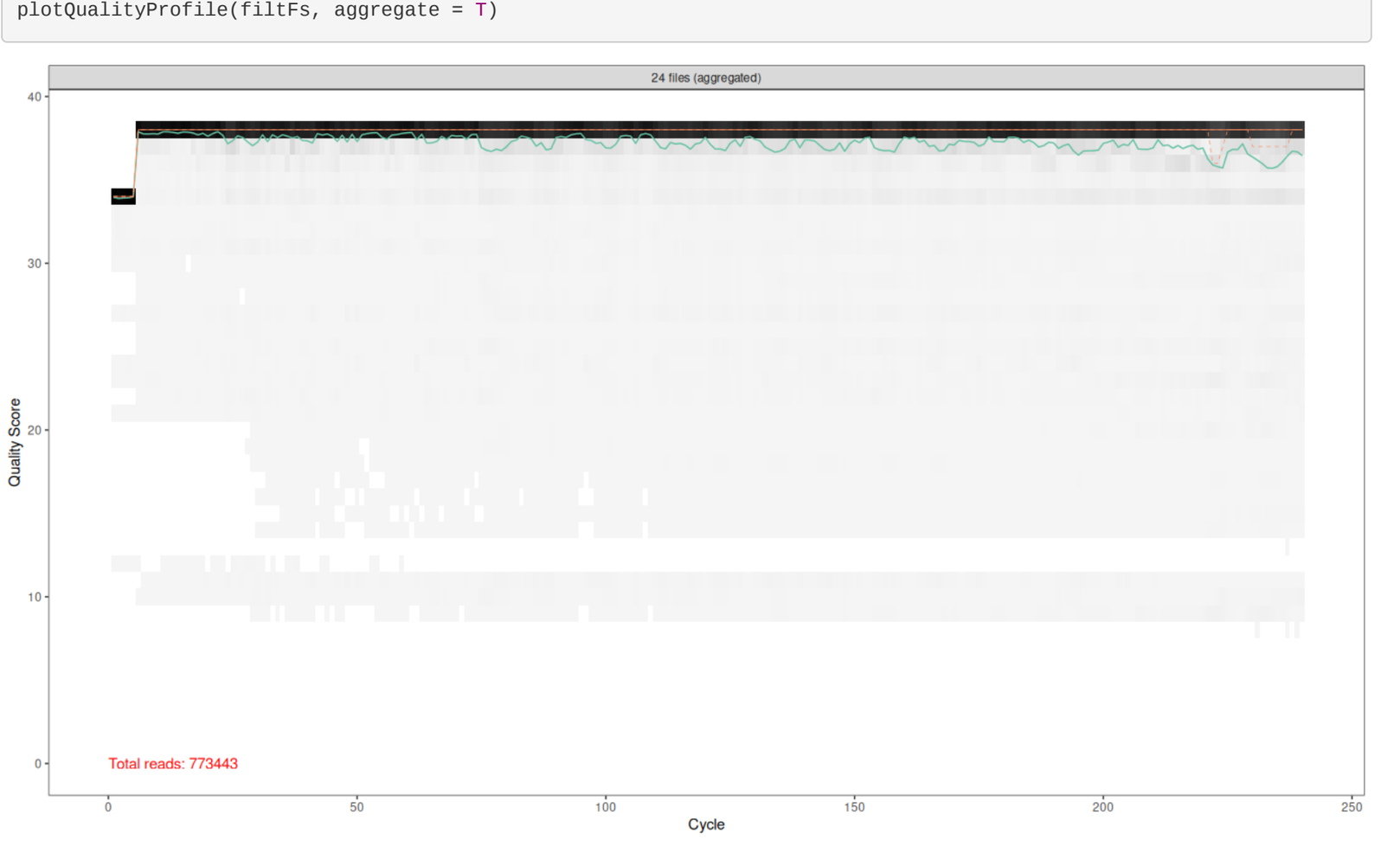
```
filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
filtRs <- file.path(path, "filtered", paste0(sample.names, "_R_filt.fastq.gz"))
names(filtFs) <- sample.names
names(filtRs) <- sample.names

out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(240,180),
              maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
              compress=TRUE, multithread=TRUE)
out
```

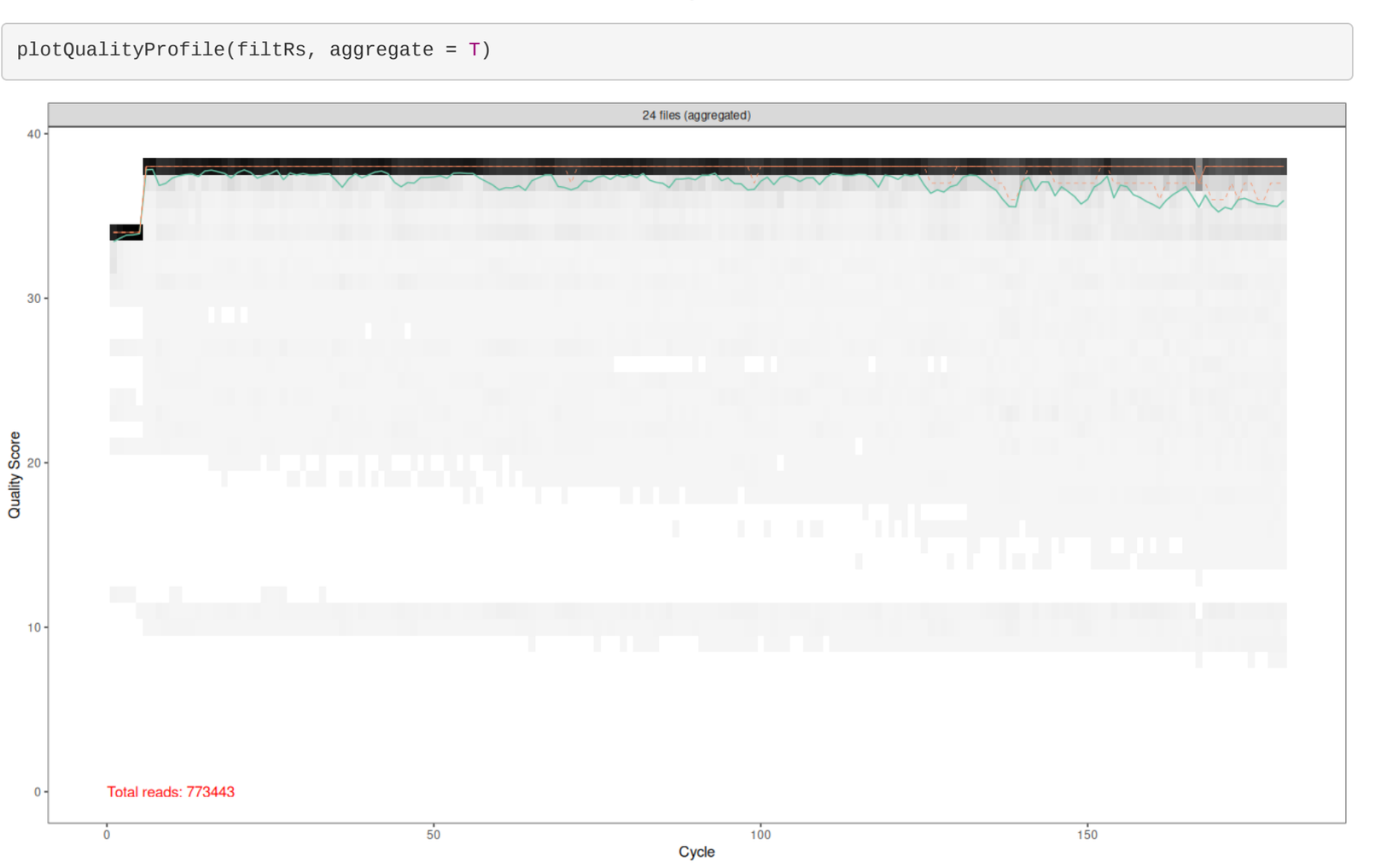
```
##                                       reads.in reads.out
## Abacumov-B-1_S1_L001_R1_001.fastq.gz      43307     36622
## Abacumov-B-13_S13_L001_R1_001.fastq.gz    41438     34160
## Abacumov-B-14_S14_L001_R1_001.fastq.gz    38072     31002
## Abacumov-B-15_S15_L001_R1_001.fastq.gz    46395     37918
## Abacumov-B-16_S16_L001_R1_001.fastq.gz    47274     38962
## Abacumov-B-2_S2_L001_R1_001.fastq.gz      45972     39386
## Abacumov-B-25_S25_L001_R1_001.fastq.gz    29444     22676
## Abacumov-B-26_S26_L001_R1_001.fastq.gz    21786     17052
## Abacumov-B-27_S27_L001_R1_001.fastq.gz    25002     19659
## Abacumov-B-28_S28_L001_R1_001.fastq.gz    25214     19501
## Abacumov-B-3_S3_L001_R1_001.fastq.gz      54999     47256
## Abacumov-B-37_S37_L001_R1_001.fastq.gz    31738     26915
## Abacumov-B-38_S38_L001_R1_001.fastq.gz    33322     27806
## Abacumov-B-39_S39_L001_R1_001.fastq.gz    27048     22885
## Abacumov-B-4_S4_L001_R1_001.fastq.gz      40115     34065
## Abacumov-B-40_S40_L001_R1_001.fastq.gz    28630     24332
## Abacumov-B-49_S49_L001_R1_001.fastq.gz    54922     45257
## Abacumov-B-50_S50_L001_R1_001.fastq.gz    41565     34229
## Abacumov-B-51_S51_L001_R1_001.fastq.gz    50422     42869
## Abacumov-B-52_S52_L001_R1_001.fastq.gz    38603     31780
## Abacumov-B-61_S61_L001_R1_001.fastq.gz    48250     41246
## Abacumov-B-62_S62_L001_R1_001.fastq.gz    39777     33175
## Abacumov-B-63_S63_L001_R1_001.fastq.gz    38230     32469
## Abacumov-B-64_S64_L001_R1_001.fastq.gz    37920     32221
```

Trimmed quality plot

```
# !Long Operations
plotQualityProfile(filtFs, aggregate = T)
```




```
plotQualityProfile(filtRs, aggregate = T)
```



Reads are trimmed fairly, everything is OK, go to the next step

Build a model and apply it

#графики вероятности перехода

```
# !Long Operation
errF <- learnErrors(filtFs, multithread=TRUE)
```

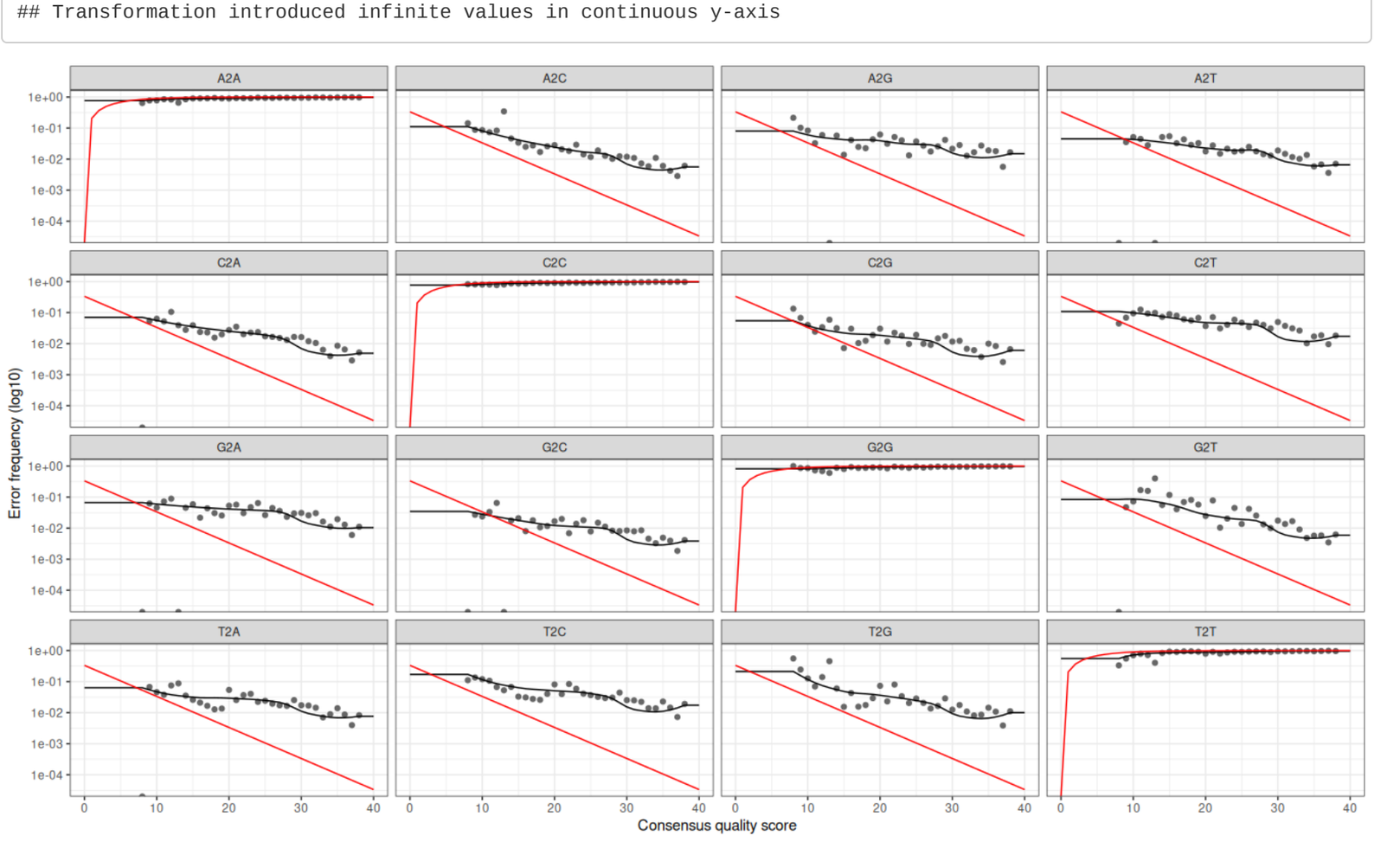
```
## 101232000 total bases in 421800 reads from 14 samples will be used for learning the error rates.
```

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

```
## 100742940 total bases in 559683 reads from 18 samples will be used for learning the error rates.
```

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

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




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

```
## Sample 1 - 36622 reads in 28087 unique sequences.
## Sample 2 - 34160 reads in 18735 unique sequences.
## Sample 3 - 31002 reads in 18643 unique sequences.
## Sample 4 - 37918 reads in 19870 unique sequences.
## Sample 5 - 38962 reads in 20810 unique sequences.
## Sample 6 - 39386 reads in 30781 unique sequences.
## Sample 7 - 22676 reads in 6818 unique sequences.
## Sample 8 - 17052 reads in 5663 unique sequences.
## Sample 9 - 19659 reads in 6237 unique sequences.
## Sample 10 - 19501 reads in 5706 unique sequences.
## Sample 11 - 47256 reads in 33296 unique sequences.
## Sample 12 - 26915 reads in 19311 unique sequences.
## Sample 13 - 27806 reads in 20185 unique sequences.
## Sample 14 - 22885 reads in 17428 unique sequences.
## Sample 15 - 34065 reads in 26006 unique sequences.
## Sample 16 - 24332 reads in 18733 unique sequences.
## Sample 17 - 45257 reads in 27334 unique sequences.
## Sample 18 - 34229 reads in 21495 unique sequences.
## Sample 19 - 42869 reads in 27611 unique sequences.
## Sample 20 - 31780 reads in 20178 unique sequences.
## Sample 21 - 41246 reads in 30101 unique sequences.
## Sample 22 - 33175 reads in 23318 unique sequences.
## Sample 23 - 32469 reads in 22700 unique sequences.
## Sample 24 - 32221 reads in 23605 unique sequences.
```

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

```
## Sample 1 - 36622 reads in 28652 unique sequences.
## Sample 2 - 34160 reads in 19562 unique sequences.
## Sample 3 - 31002 reads in 18627 unique sequences.
## Sample 4 - 37918 reads in 19876 unique sequences.
## Sample 5 - 38962 reads in 20775 unique sequences.
## Sample 6 - 39386 reads in 31028 unique sequences.
## Sample 7 - 22676 reads in 7053 unique sequences.
## Sample 8 - 17052 reads in 5659 unique sequences.
## Sample 9 - 19659 reads in 7019 unique sequences.
## Sample 10 - 19501 reads in 5888 unique sequences.
## Sample 11 - 47256 reads in 34628 unique sequences.
## Sample 12 - 26915 reads in 20489 unique sequences.
## Sample 13 - 27806 reads in 20955 unique sequences.
## Sample 14 - 22885 reads in 17871 unique sequences.
## Sample 15 - 34065 reads in 26746 unique sequences.
## Sample 16 - 24332 reads in 19350 unique sequences.
## Sample 17 - 45257 reads in 28159 unique sequences.
## Sample 18 - 34229 reads in 22116 unique sequences.
## Sample 19 - 42869 reads in 28277 unique sequences.
## Sample 20 - 31780 reads in 21109 unique sequences.
## Sample 21 - 41246 reads in 31022 unique sequences.
## Sample 22 - 33175 reads in 24219 unique sequences.
## Sample 23 - 32469 reads in 23696 unique sequences.
## Sample 24 - 32221 reads in 24015 unique sequences.
```

Merge reads and create table

```
# !Long Operation
mergers <- mergePairs(dadaFs, filtFs, dadaRs, filtRs, verbose=TRUE)
```

```
## 15729 paired-reads (in 416 unique pairings) successfully merged out of 27255 (in 4019 pairings) input.
```

```
## 22924 paired-reads (in 625 unique pairings) successfully merged out of 30302 (in 3396 pairings) input.
```

```
## 19335 paired-reads (in 524 unique pairings) successfully merged out of 27027 (in 3343 pairings) input.
```

```
## 26430 paired-reads (in 702 unique pairings) successfully merged out of 34157 (in 3582 pairings) input.
```

```
## 27267 paired-reads (in 730 unique pairings) successfully merged out of 34815 (in 3483 pairings) input.
```

```
## 16156 paired-reads (in 409 unique pairings) successfully merged out of 28750 (in 4321 pairings) input.
```

```
## 19847 paired-reads (in 475 unique pairings) successfully merged out of 21816 (in 1222 pairings) input.
```

```
## 14224 paired-reads (in 389 unique pairings) successfully merged out of 16302 (in 1077 pairings) input.
```

```
## 16700 paired-reads (in 416 unique pairings) successfully merged out of 18925 (in 1182 pairings) input.
```

```
## 17225 paired-reads (in 415 unique pairings) successfully merged out of 18993 (in 1045 pairings) input.
```

```
## 22615 paired-reads (in 620 unique pairings) successfully merged out of 36917 (in 5423 pairings) input.
```

```
## 11775 paired-reads (in 329 unique pairings) successfully merged out of 21544 (in 3228 pairings) input.
```

```
## 12074 paired-reads (in 349 unique pairings) successfully merged out of 22312 (in 3288 pairings) input.
```

```
## 9115 paired-reads (in 236 unique pairings) successfully merged out of 17755 (in 2532 pairings) input.
```

```
## 14492 paired-reads (in 393 unique pairings) successfully merged out of 25241 (in 3708 pairings) input.
```

```
## 9929 paired-reads (in 281 unique pairings) successfully merged out of 18440 (in 2615 pairings) input.
```

```
## 26537 paired-reads (in 764 unique pairings) successfully merged out of 38769 (in 4673 pairings) input.
```

```
## 19792 paired-reads (in 566 unique pairings) successfully merged out of 28803 (in 3524 pairings) input.
```

```
## 23682 paired-reads (in 611 unique pairings) successfully merged out of 36076 (in 4670 pairings) input.
```

```
## 17117 paired-reads (in 533 unique pairings) successfully merged out of 26351 (in 3344 pairings) input.
```

```
## 18224 paired-reads (in 443 unique pairings) successfully merged out of 33192 (in 5163 pairings) input.
```

```
## 14348 paired-reads (in 385 unique pairings) successfully merged out of 26975 (in 3871 pairings) input.
```

```
## 14376 paired-reads (in 386 unique pairings) successfully merged out of 26025 (in 3621 pairings) input.
```

```
## 13777 paired-reads (in 341 unique pairings) successfully merged out of 25769 (in 3546 pairings) input.
```

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

```
## [1]   24 7004
```

Taxonomy annotation

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

```
## Identified 1948 bimeras out of 7004 input sequences.
```

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

```
## [1]   24 5056
```

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

```
## [1] 0.8053931
```

```
getN <- function(x) sum(getUniques(x))
track <- cbind(out, sapply(dadaFs, getN), sapply(dadaRs, getN), sapply(mergers, getN), rowSums(seqtab.nochim))
# If processing a single sample, remove the sapply calls: e.g. replace sapply(dadaFs, getN) with getN(dadaFs)
colnames(track) <- c("input", "filtered", "denoisedF", "denoisedR", "merged", "nonchim")
rownames(track) <- sample.names
track
```

```
##               input filtered denoisedF denoisedR merged nonchim
## Abacumov-B-1  43307    36622     29113     32777  15729   14462
## Abacumov-B-13 41438    34160     31668     32193  22924   17630
## Abacumov-B-14 38072    31002     28282     29053  19335   15500
## Abacumov-B-15 46395    37918     35599     35901  26430   19483
## Abacumov-B-16 47274    38962     36181     36939  27267   19850
## Abacumov-B-2  45972    39386     30707     35254  16156   14720
## Abacumov-B-25 29444    22676     22213     22195  19847   11714
## Abacumov-B-26 21786    17052     16638     16612  14224    8413
## Abacumov-B-27 25002    19659     19210     19247  16700    9839
## Abacumov-B-28 25214    19501     19192     19214  17225    9896
## Abacumov-B-3  54999    47256     39718     42453  22615   20250
## Abacumov-B-37 31738    26915     22832     24440  11775   10392
## Abacumov-B-38 33322    27806     23564     25429  12074   10827
## Abacumov-B-39 27048    22885     18852     20676   9115    8157
## Abacumov-B-4  40115    34065     26982     30465  14492   13500
## Abacumov-B-40 28630    24332     19694     21767   9929    8925
## Abacumov-B-49 54922    45257     40839     41993  26537   21718
## Abacumov-B-50 41565    34229     30510     31488  19792   17003
## Abacumov-B-51 50422    42869     38023     39726  23682   19938
## Abacumov-B-52 38603    31780     27870     29113  17117   14984
## Abacumov-B-61 48250    41246     34823     38016  18224   16231
## Abacumov-B-62 39777    33175     28334     30630  14348   12889
## Abacumov-B-63 38230    32469     27559     29584  14376   12724
## Abacumov-B-64 37920    32221     27196     29463  13777   12192
```

Merging leads to losses in reads. Re-run more relaxed filtering

```
taxa <- assignTaxonomy(seqtab.nochim, "/home/nastasista/Metagenomics/silva_nr_v132_train_set.fa.gz", multithread=
TRUE)

taxa.print <- taxa # Removing sequence rownames for display only
rownames(taxa.print) <- NULL
head(taxa.print)
```

```
##      Kingdom    Phylum          Class              Order        Family Genus
## [1,] "Bacteria" "Chloroflexi"   "AD3"              NA           NA     NA
## [2,] "Bacteria" "Chloroflexi"   "AD3"              NA           NA     NA
## [3,] "Bacteria" "Chloroflexi"   "AD3"              NA           NA     NA
## [4,] "Bacteria" "Chloroflexi"   "AD3"              NA           NA     NA
## [5,] "Bacteria" "Cyanobacteria" "Oxyphotobacteria" "Nostocales" NA     NA
## [6,] "Bacteria" "Cyanobacteria" "Oxyphotobacteria" "Nostocales" NA     NA
```

```
rownames(metadata) <- metadata$Filename

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

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 5056 taxa and 24 samples ]
## sample_data() Sample Data:       [ 24 samples by 6 sample variables ]
## tax_table()   Taxonomy Table:    [ 5056 taxa by 6 taxonomic ranks ]
```

```
sample_names(ps)
```

```
##  [1] "Abacumov-B-1"  "Abacumov-B-13" "Abacumov-B-14" "Abacumov-B-15"
##  [5] "Abacumov-B-16" "Abacumov-B-2"  "Abacumov-B-25" "Abacumov-B-26"
##  [9] "Abacumov-B-27" "Abacumov-B-28" "Abacumov-B-3"  "Abacumov-B-37"
## [13] "Abacumov-B-38" "Abacumov-B-39" "Abacumov-B-4"  "Abacumov-B-40"
## [17] "Abacumov-B-49" "Abacumov-B-50" "Abacumov-B-51" "Abacumov-B-52"
## [21] "Abacumov-B-61" "Abacumov-B-62" "Abacumov-B-63" "Abacumov-B-64"
```

Rename phyloseq-object according to our needs

```
metadata
```

```
##                    Filename             Source Site Horizont Repeat
## Abacumov-B-1   Abacumov-B-1 Self-growing Dumps   B1       AY      1
## Abacumov-B-2   Abacumov-B-2 Self-growing Dumps   B1       AY      2
## Abacumov-B-3   Abacumov-B-3 Self-growing Dumps   B1       AY      3
## Abacumov-B-4   Abacumov-B-4 Self-growing Dumps   B1       AY      4
## Abacumov-B-13 Abacumov-B-13          Litostrat   B2        C      1
## Abacumov-B-14 Abacumov-B-14          Litostrat   B2        C      2
## Abacumov-B-15 Abacumov-B-15          Litostrat   B2        C      3
## Abacumov-B-16 Abacumov-B-16          Litostrat   B2        C      4
## Abacumov-B-25 Abacumov-B-25 Coal Mine Terricon   B3        C      1
## Abacumov-B-26 Abacumov-B-26 Coal Mine Terricon   B3        C      2
## Abacumov-B-27 Abacumov-B-27 Coal Mine Terricon   B3        C      3
## Abacumov-B-28 Abacumov-B-28 Coal Mine Terricon   B3        C      4
## Abacumov-B-37 Abacumov-B-37    Local Reference   B4       AY      1
## Abacumov-B-38 Abacumov-B-38    Local Reference   B4       AY      2
## Abacumov-B-39 Abacumov-B-39    Local Reference   B4       AY      3
## Abacumov-B-40 Abacumov-B-40    Local Reference   B4       AY      4
## Abacumov-B-49 Abacumov-B-49        Embryo Sand   B5       AY      1
## Abacumov-B-50 Abacumov-B-50        Embryo Sand   B5       AY      2
## Abacumov-B-51 Abacumov-B-51        Embryo Sand   B5       AY      3
## Abacumov-B-52 Abacumov-B-52        Embryo Sand   B5       AY      4
## Abacumov-B-61 Abacumov-B-61 Regional Reference   B6       AY      1
## Abacumov-B-62 Abacumov-B-62 Regional Reference   B6       AY      2
## Abacumov-B-63 Abacumov-B-63 Regional Reference   B6       AY      3
## Abacumov-B-64 Abacumov-B-64 Regional Reference   B6       AY      4
##                                 SampleID
## Abacumov-B-1  Self-growing Dumps.B1.AY.1
## Abacumov-B-2  Self-growing Dumps.B1.AY.2
## Abacumov-B-3  Self-growing Dumps.B1.AY.3
## Abacumov-B-4  Self-growing Dumps.B1.AY.4
## Abacumov-B-13           Litostrat.B2.C.1
## Abacumov-B-14           Litostrat.B2.C.2
## Abacumov-B-15           Litostrat.B2.C.3
## Abacumov-B-16           Litostrat.B2.C.4
## Abacumov-B-25  Coal Mine Terricon.B3.C.1
## Abacumov-B-26  Coal Mine Terricon.B3.C.2
## Abacumov-B-27  Coal Mine Terricon.B3.C.3
## Abacumov-B-28  Coal Mine Terricon.B3.C.4
## Abacumov-B-37    Local Reference.B4.AY.1
## Abacumov-B-38    Local Reference.B4.AY.2
## Abacumov-B-39    Local Reference.B4.AY.3
## Abacumov-B-40    Local Reference.B4.AY.4
## Abacumov-B-49        Embryo Sand.B5.AY.1
## Abacumov-B-50        Embryo Sand.B5.AY.2
## Abacumov-B-51        Embryo Sand.B5.AY.3
## Abacumov-B-52        Embryo Sand.B5.AY.4
## Abacumov-B-61 Regional Reference.B6.AY.1
## Abacumov-B-62 Regional Reference.B6.AY.2
## Abacumov-B-63 Regional Reference.B6.AY.3
## Abacumov-B-64 Regional Reference.B6.AY.4
```

```
## Rename Samples
new.names <- ps@sam_data %>%
  data.frame() %>%
  dplyr::select(Filename, SampleID) %>%
  arrange(Filename, levels = sample_names(ps))

if (all(sample_names(ps) == new.names$Filename)) {
  sample_names(ps) <- ps@sam_data$SampleID
  print("Renamed")
}
```

```
## [1] "Renamed"
```

```
sample_names(ps)
```

```
##  [1] "Self-growing Dumps.B1.AY.1" "Litostrat.B2.C.1"
##  [3] "Litostrat.B2.C.2"           "Litostrat.B2.C.3"
##  [5] "Litostrat.B2.C.4"           "Self-growing Dumps.B1.AY.2"
##  [7] "Coal Mine Terricon.B3.C.1"  "Coal Mine Terricon.B3.C.2"
##  [9] "Coal Mine Terricon.B3.C.3"  "Coal Mine Terricon.B3.C.4"
## [11] "Self-growing Dumps.B1.AY.3" "Local Reference.B4.AY.1"
## [13] "Local Reference.B4.AY.2"    "Local Reference.B4.AY.3"
## [15] "Self-growing Dumps.B1.AY.4" "Local Reference.B4.AY.4"
## [17] "Embryo Sand.B5.AY.1"        "Embryo Sand.B5.AY.2"
## [19] "Embryo Sand.B5.AY.3"        "Embryo Sand.B5.AY.4"
## [21] "Regional Reference.B6.AY.1" "Regional Reference.B6.AY.2"
## [23] "Regional Reference.B6.AY.3" "Regional Reference.B6.AY.4"
```

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

Save phyloseq-object and aquire checksum

```
saveRDS(ps, "ps.RData")
ps <- readRDS("ps.RData")

tools::md5sum("ps.RData")
```

```
##                           ps.RData
## "9767b2490c215ecc89058c1d259db9bd"
```