Microbiome DADA2

Jair

2024-04-16

Load required packages

```
library(dada2)
```

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

Load sequences

```
sequences <- "sequences"
list.files(sequences)
```

```
##  [1] "filtered"                       "L1S105_9_L001_R1_001.fastq.gz"
##  [3] "L1S140_6_L001_R1_001.fastq.gz"  "L1S208_10_L001_R1_001.fastq.gz"
##  [5] "L1S257_11_L001_R1_001.fastq.gz" "L1S281_5_L001_R1_001.fastq.gz"
##  [7] "L1S57_13_L001_R1_001.fastq.gz"  "L1S76_12_L001_R1_001.fastq.gz"
##  [9] "L1S8_8_L001_R1_001.fastq.gz"    "L2S155_25_L001_R1_001.fastq.gz"
## [11] "L2S175_27_L001_R1_001.fastq.gz" "L2S204_1_L001_R1_001.fastq.gz"
## [13] "L2S222_23_L001_R1_001.fastq.gz" "L2S240_7_L001_R1_001.fastq.gz"
## [15] "L2S309_33_L001_R1_001.fastq.gz" "L2S357_15_L001_R1_001.fastq.gz"
## [17] "L2S382_34_L001_R1_001.fastq.gz" "L3S242_19_L001_R1_001.fastq.gz"
## [19] "L3S294_16_L001_R1_001.fastq.gz" "L3S313_32_L001_R1_001.fastq.gz"
## [21] "L3S341_18_L001_R1_001.fastq.gz" "L3S360_4_L001_R1_001.fastq.gz"
## [23] "L3S378_24_L001_R1_001.fastq.gz" "L4S112_26_L001_R1_001.fastq.gz"
## [25] "L4S137_21_L001_R1_001.fastq.gz" "L4S63_31_L001_R1_001.fastq.gz"
## [27] "L5S104_28_L001_R1_001.fastq.gz" "L5S155_2_L001_R1_001.fastq.gz"
## [29] "L5S174_29_L001_R1_001.fastq.gz" "L5S203_3_L001_R1_001.fastq.gz"
## [31] "L5S222_17_L001_R1_001.fastq.gz" "L5S240_14_L001_R1_001.fastq.gz"
## [33] "L6S20_20_L001_R1_001.fastq.gz"  "L6S68_30_L001_R1_001.fastq.gz"
## [35] "L6S93_22_L001_R1_001.fastq.gz"  "MANIFEST"
## [37] "metadata.yml"
```

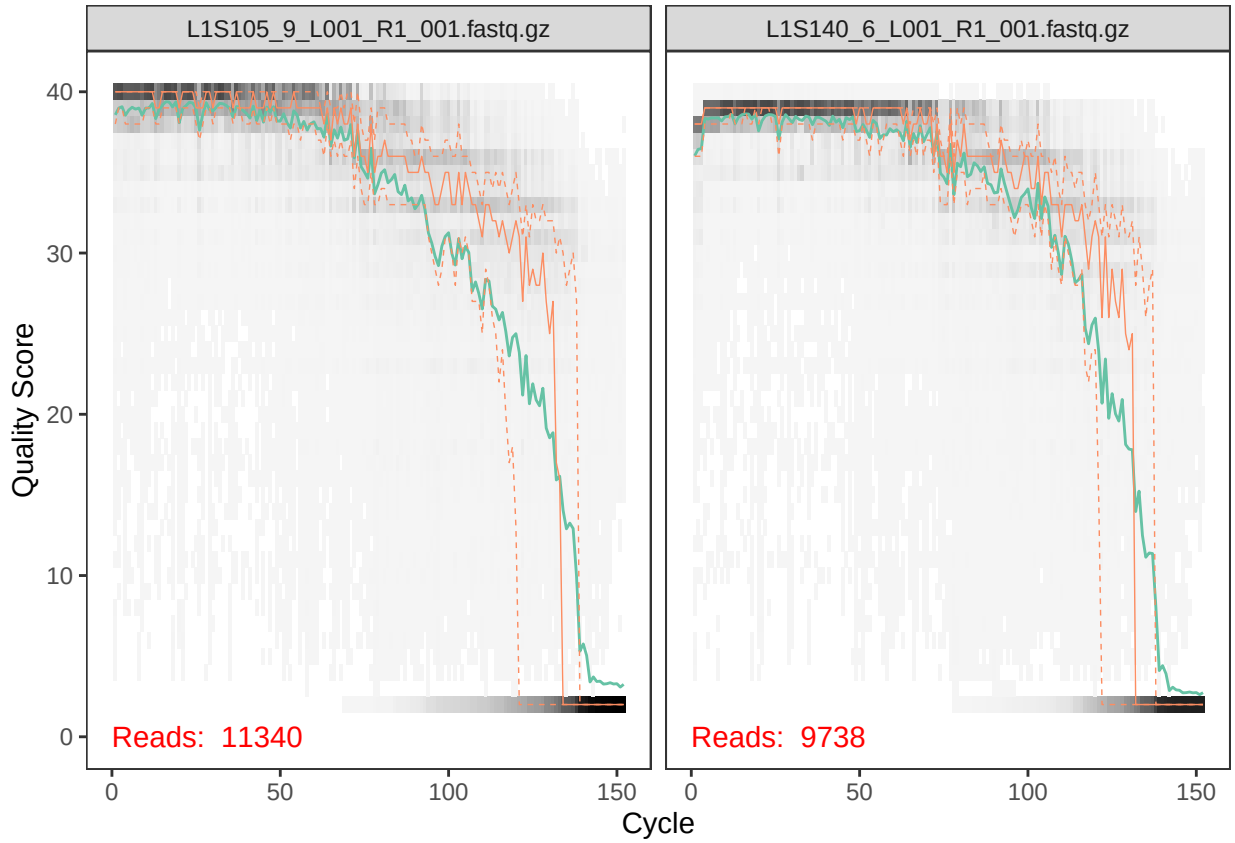
fastq filenames have format: SAMPLENAME_R1_001.fastq and SAMPLENAME_R2_001.fastq

```
fnFs <- sort(list.files(sequences, pattern="_R1_001.fastq", full.names = TRUE))

# Extract sample names, assuming filenames have format: SAMPLENAME_XXX.fastq
sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
```

Inspect read quality

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




```
# Place filtered files in filtered/ subdirectory
filtFs <- file.path(sequences, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
names(filtFs) <- sample.names
#According to our QualityProfile, the quality decreases at 120
out <- filterAndTrim(fnFs, filtFs, truncLen=c(120),
              maxN=0, maxEE=c(2), truncQ=2, rm.phix=TRUE,
              compress=TRUE, multithread=TRUE)
head(out)
```

```
##                                reads.in reads.out
## L1S105_9_L001_R1_001.fastq.gz     11340      8571
## L1S140_6_L001_R1_001.fastq.gz      9738      7677
## L1S208_10_L001_R1_001.fastq.gz    11337      9261
## L1S257_11_L001_R1_001.fastq.gz     8216      6705
## L1S281_5_L001_R1_001.fastq.gz      8907      7067
## L1S57_13_L001_R1_001.fastq.gz     11752      9299
```

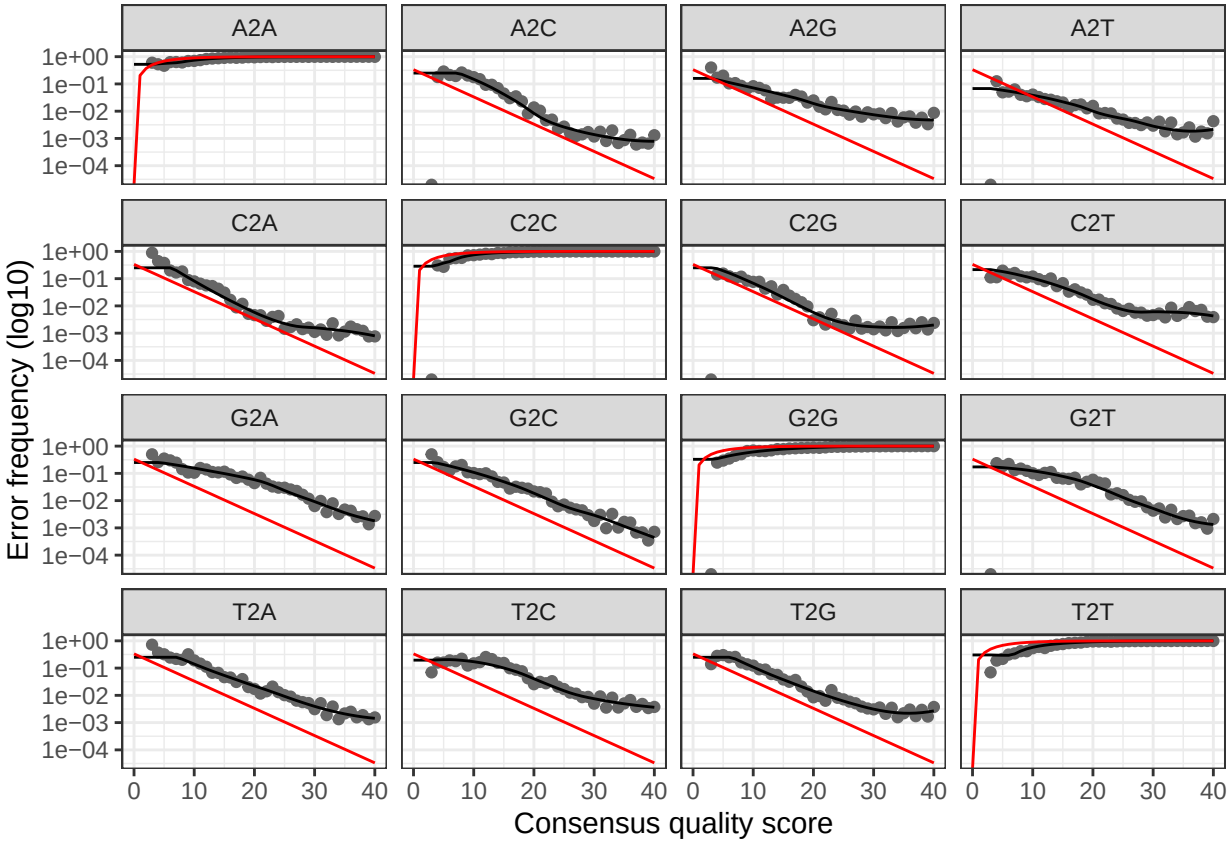
```
errF <- learnErrors(filtFs, multithread=TRUE)
```

```
## 19539480 total bases in 162829 reads from 34 samples will be used for learning the error rates.
```

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

```
## Warning in scale_y_log10(): log-10 transformation introduced infinite values.
```

```
## log-10 transformation introduced infinite values.
```



Sample inference

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

```
## Sample 1 - 8571 reads in 2110 unique sequences.
## Sample 2 - 7677 reads in 1728 unique sequences.
## Sample 3 - 9261 reads in 2490 unique sequences.
## Sample 4 - 6705 reads in 1940 unique sequences.
## Sample 5 - 7067 reads in 2144 unique sequences.
## Sample 6 - 9299 reads in 2317 unique sequences.
## Sample 7 - 8395 reads in 1967 unique sequences.
## Sample 8 - 7663 reads in 1573 unique sequences.
## Sample 9 - 4112 reads in 1272 unique sequences.
## Sample 10 - 4546 reads in 1325 unique sequences.
## Sample 11 - 3379 reads in 1131 unique sequences.
## Sample 12 - 3485 reads in 1574 unique sequences.
## Sample 13 - 5183 reads in 1104 unique sequences.
## Sample 14 - 1550 reads in 641 unique sequences.
## Sample 15 - 2526 reads in 874 unique sequences.
## Sample 16 - 4279 reads in 1281 unique sequences.
## Sample 17 - 970 reads in 246 unique sequences.
## Sample 18 - 1313 reads in 483 unique sequences.
## Sample 19 - 1191 reads in 460 unique sequences.
```

```
## Sample 20 - 1109 reads in 478 unique sequences.
## Sample 21 - 1132 reads in 603 unique sequences.
## Sample 22 - 1358 reads in 379 unique sequences.
## Sample 23 - 8603 reads in 2252 unique sequences.
## Sample 24 - 10064 reads in 2146 unique sequences.
## Sample 25 - 10096 reads in 2882 unique sequences.
## Sample 26 - 2253 reads in 448 unique sequences.
## Sample 27 - 1828 reads in 379 unique sequences.
## Sample 28 - 1969 reads in 407 unique sequences.
## Sample 29 - 2133 reads in 459 unique sequences.
## Sample 30 - 2556 reads in 468 unique sequences.
## Sample 31 - 1817 reads in 380 unique sequences.
## Sample 32 - 7087 reads in 983 unique sequences.
## Sample 33 - 6169 reads in 1033 unique sequences.
## Sample 34 - 7483 reads in 1272 unique sequences.
```

Amplicon sequence variant table (ASV) table (819 ASVs detected)

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

```
## [1]  34 819
```

```
# Inspect distribution of sequence lengths
table(nchar(getSequences(seqtab)))
```

```
##
## 120
## 819
```

#Removal of chimeras

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

```
## Identified 48 bimeras out of 819 input sequences.
```

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

```
## [1]  34 771
```

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

```
## [1] 0.9652497
```

Tracking reads through the pipeline

```
getN <- function(x) sum(getUniques(x))
# Only use dada results for forward reads, simplify if only one sample
if(length(dadaFs) == 1) {
    denoisedF <- getN(dadaFs)
} else {
    denoisedF <- sapply(dadaFs, getN)
}

# Assemble the tracking matrix
```

```
track <- cbind(out, denoisedF, rowSums(seqtab.nochim))
colnames(track) <- c("input", "filtered", "denoisedF", "nonchim")
rownames(track) <- sample.names
head(track)
```

```
##        input filtered denoisedF nonchim
## L1S105 11340     8571      8499    7780
## L1S140  9738     7677      7605    7163
## L1S208 11337     9261      9152    8152
## L1S257  8216     6705      6627    6388
## L1S281  8907     7067      6976    6615
## L1S57  11752     9299      9260    8702
```

Assigning taxa using the Silva reference database

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

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

```
##      Kingdom    Phylum           Class                  Order
## [1,] "Bacteria" "Bacteroidota"   "Bacteroidia"          "Bacteroidales"
## [2,] "Bacteria" "Proteobacteria" "Gammaproteobacteria" "Burkholderiales"
## [3,] "Bacteria" "Firmicutes"     "Bacilli"              "Lactobacillales"
## [4,] "Bacteria" "Bacteroidota"   "Bacteroidia"          "Bacteroidales"
## [5,] "Bacteria" "Bacteroidota"   "Bacteroidia"          "Bacteroidales"
## [6,] "Bacteria" "Proteobacteria" "Gammaproteobacteria" "Enterobacterales"
##      Family             Genus
## [1,] "Bacteroidaceae"   "Bacteroides"
## [2,] "Neisseriaceae"    "Neisseria"
## [3,] "Streptococcaceae" "Streptococcus"
## [4,] "Bacteroidaceae"   "Bacteroides"
## [5,] "Bacteroidaceae"   "Bacteroides"
## [6,] "Pasteurellaceae"  "Haemophilus"
```