Dada2 pipeline in R

Marko Suokas

Load libraries

```
library(dada2);packageVersion("dada2")
```

```
[1] '1.32.0'
```

```
library(mia); packageVersion("mia")
```

```
[1] '1.12.0'
```

```
library(knitr);packageVersion("knitr")
```

```
[1] '1.48'
```

```
library(Biostrings);packageVersion("Biostrings")
```

```
[1] '2.72.1'
```

```
library(DECIPHER);packageVersion("DECIPHER")
```

```
[1] '3.0.0'
```

```
library(tidyverse);packageVersion("tidyverse")
```

```
[1] '2.0.0'
```

```
library(kableExtra);packageVersion("kableExtra")
```

```
[1] '1.4.0'
```

```
library(patchwork);packageVersion("patchwork")
```

```
[1] '1.2.0'
```

Parameters

```
#Path variables
path <- "data/reads"
training <- "~/feature_classifiers/SILVA_SSU_r138_2019.RData"
meta_file <- "data/metadata.tsv"
exportloc <- "result_tables"
#Truncation length and phix (Illumina)
truncation <- 245
phi <- FALSE
#Name of first column in metadata file
meta_1stcol <- "Sampleid"
#Create results directory
dir.create(exportloc)
```

Sample lists

```
#List files in path
list.files(path)
```

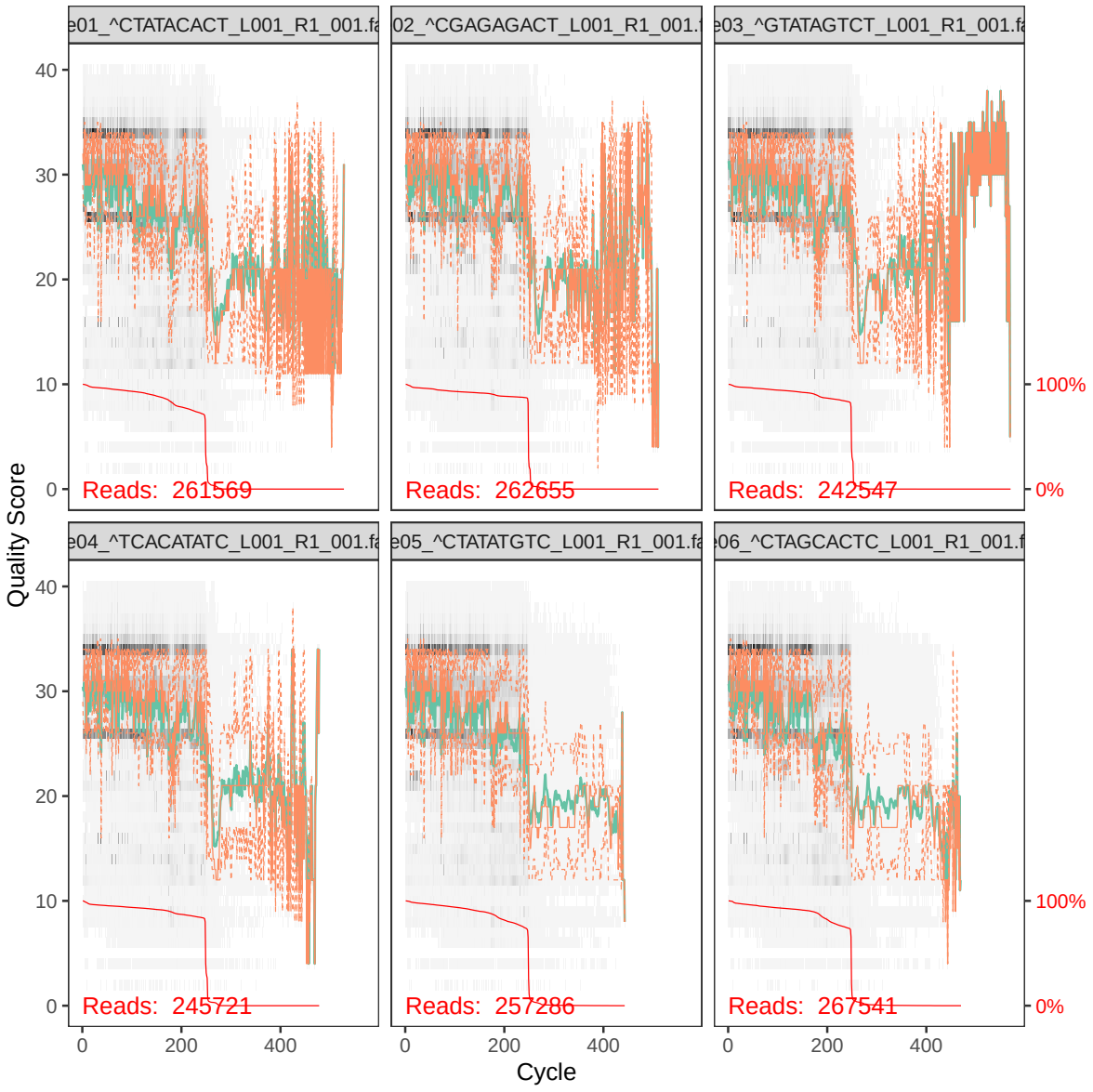
```
 [1] "Sample01_^CTATACACT_L001_R1_001.fastq.gz"
 [2] "Sample02_^CGAGAGACT_L001_R1_001.fastq.gz"
 [3] "Sample03_^GTATAGTCT_L001_R1_001.fastq.gz"
 [4] "Sample04_^TCACATATC_L001_R1_001.fastq.gz"
 [5] "Sample05_^CTATATGTC_L001_R1_001.fastq.gz"
 [6] "Sample06_^CTAGCACTC_L001_R1_001.fastq.gz"
 [7] "Sample07_^ATAGCTGTC_L001_R1_001.fastq.gz"
 [8] "Sample08_^ATACGACTC_L001_R1_001.fastq.gz"
 [9] "Sample09_^ACATGATCT_L001_R1_001.fastq.gz"
[10] "Sample10_^CTACGCATC_L001_R1_001.fastq.gz"
[11] "Sample11_^TCATGCGTC_L001_R1_001.fastq.gz"
[12] "Sample12_^TCATGTACT_L001_R1_001.fastq.gz"
[13] "Sample13_^CTACGTGCT_L001_R1_001.fastq.gz"
[14] "Sample14_^TCAGTATCT_L001_R1_001.fastq.gz"
[15] "Sample15_^GCAGTCGTC_L001_R1_001.fastq.gz"
[16] "Sample16_^TCAGTGCTC_L001_R1_001.fastq.gz"
[17] "Sample17_^AGCTACACT_L001_R1_001.fastq.gz"
[18] "Sample18_^CGCTAGTCT_L001_R1_001.fastq.gz"
[19] "filtered"
```

```
#Filenames have format: SAMPLENAME_R1_001.fastq
fnFs <- sort(list.files(path, pattern = "_R1_001.fastq", full.names = TRUE))
# Extract sample names, assuming pattern
sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
#Filtered files will be placed in filtered/ subdirectory
filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
names(filtFs) <- sample.names
```

Tip: If you have numbered samples, use 0X format. Otherwise you have problems in sort order.

Quality profile

```
# Base quality plot in first 6 samples
plotQualityProfile(fnFs[1:6])
```



Filtering and trimming reads

```
#Filtered files will be placed in filtered/ subdirectory
out <- filterAndTrim(fnFs, filtFs, truncLen=245,
                     maxN=0, maxEE=2, truncQ=2,
                     compress=TRUE, multithread=FALSE)
#Output is saved to rds file, so we don't have to recalculate, if we make changes
#If you are making changes to chunk, change eval = TRUE
saveRDS(out,"rds/out.rds")
```

```
#read rds file
out <- readRDS("rds/out.rds")
```

Considerations: The standard parameters are starting points. If you want to speed up downstream computation, consider tightening `maxEE`. If too few reads are passing the filter, consider relaxing `maxEE`.

For ITS sequencing, it is usually undesirable to truncate reads to a fixed length due to the large length variation at that locus. You can omit in this case truncLen parameter.

Learn error rates

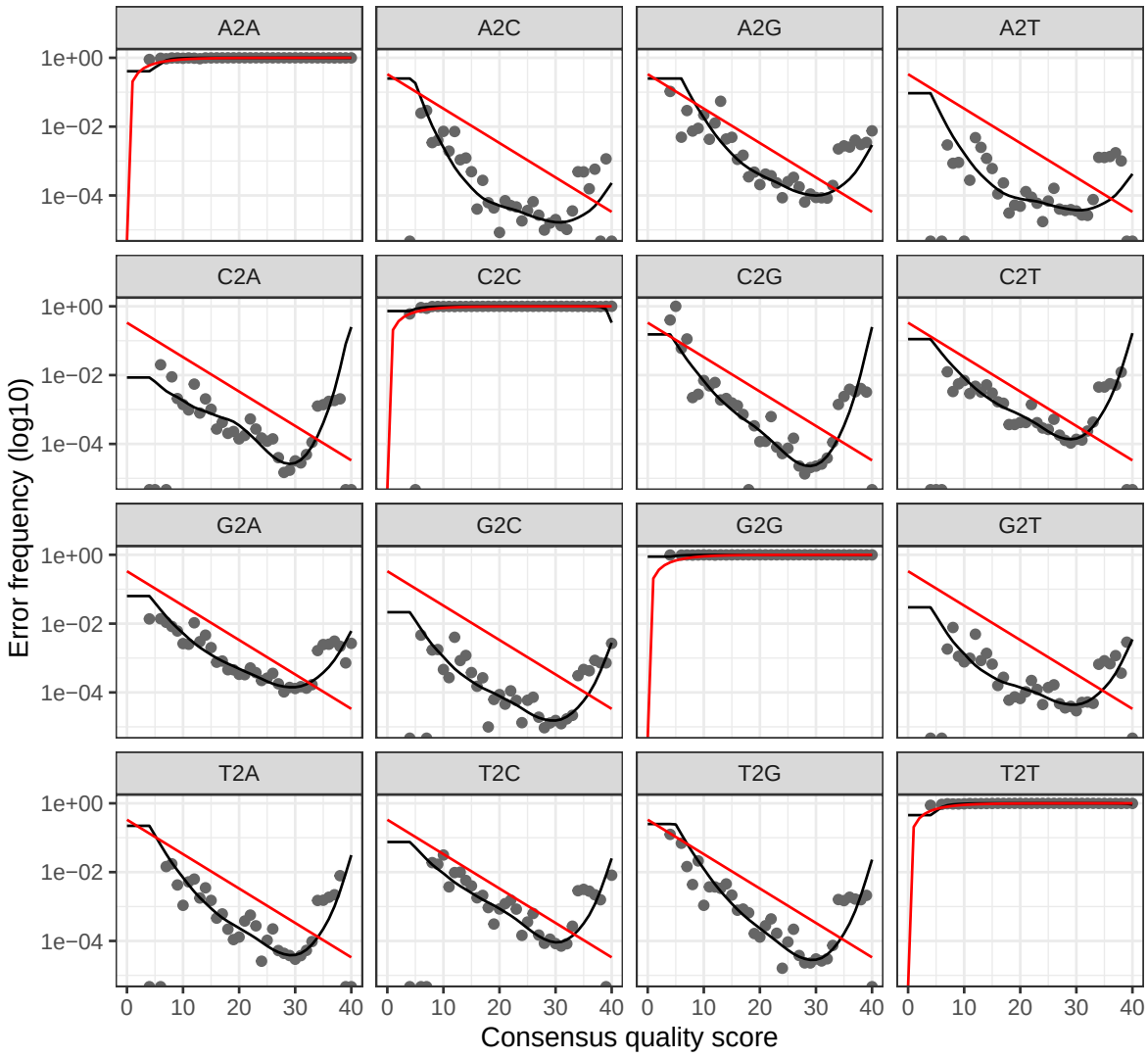
Step determinates error rate of dataset using *learnErrors* function.

```
# Forward read error rate
errF <- learnErrors(filtFs, multithread=TRUE)
# saverds
saveRDS(errF,"rds/errF.rds")
```

```
errF <- readRDS("rds/errF.rds")
```

Plot error profiles

```
# Plotting error rate profile for forward reads
plotErrors(errF, nominalQ=TRUE)
```



Denoise data

```
#denoise command
dadaFs <- dada(filtFs, err=errF, multithread=TRUE)
#save to rds file
saveRDS(dadaFs,"rds/dadaFs.rds")
```

```
#read rds
dadaFs <- readRDS("rds/dadaFs.rds")
```

Create ASV table

```
seqtab <- makeSequenceTable(dadaFs)
# Dimensions of ASV table
dim(seqtab)
```

```
[1]   18 1797
```

Remove chimeric variants

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

```
[1]   18 1558
```

Summary

Summary can help to pinpoint if at some stage, abnormal amount of the data is lost

```
getN <- function(x) sum(getUniques(x))
track <- cbind(out, sapply(dadaFs, getN), rowSums(seqtab.nochim),
               rowSums(seqtab.nochim != 0))
# If processing a single sample, replace sapply(dadaFs, getN) with getN(dadaFs)
colnames(track) <- c("Input", "Filtered", "Denoised", "Nonchimeric", "Variants")
rownames(track) <- sample.names
kable(track, caption="Denoising summary") %>%
  kable_styling(latex_options = c("HOLD_position","striped")) %>%
  row_spec(0, background="indigo", color="ivory")
```

Table 1: Denoising summary

	Input	Filtered	Denoised	Nonchimeric	Variants
Sample01	261569	137665	137385	136842	47
Sample02	262655	203573	203354	201985	59
Sample03	242547	174515	173923	172596	366
Sample04	245721	180881	180445	179564	380
Sample05	257286	147881	147568	146855	73
Sample06	267541	154404	154150	153195	57
Sample07	246034	147910	147210	144670	755
Sample08	265660	153425	152732	149897	727
Sample09	226624	186066	185870	185285	43
Sample10	216757	163175	162997	162556	36
Sample11	225419	164347	164114	163551	99
Sample12	232630	159525	159275	158958	110
Sample13	222337	155976	155772	154595	37
Sample14	232938	183999	183779	182886	42
Sample15	222467	162987	162557	162207	207
Sample16	227439	153865	153543	152411	181
Sample17	332584	138388	137843	136599	480
Sample18	403	247	228	228	8

Assign taxonomy

We use idTaxa from DECIPHER and Silva database to assign taxonomic information.

```
#Create a DNAStringSet from the ASVs
sequences <- DNAStringSet(getSequences(seqtab.nochim))
# CHANGE TO THE PATH OF YOUR TRAINING SET
load("~/feature_classifiers/SILVA_SSU_r138_2019.RData")
#IdTaxa
ids <- IdTaxa(sequences, trainingSet, strand="top", processors = 3, verbose = FALSE)
ranks <- c("domain", "phylum", "class", "order", "family", "genus", "species")
#Convert the output object of class "Taxa" to a matrix analogous to the output from assignTaxonomy
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)
#save end result to rds
saveRDS(taxid, "rds/taxid.rds")
```

```
taxid <- readRDS("rds/taxid.rds")
```

Create tse object

```
# project metadata
samples_meta <- read_tsv(meta_file, show_col_types = FALSE)
samples_meta <- samples_meta %>% tibble::column_to_rownames("Sampleid")
# representative sequences
repseq <- DNAStringSet(rownames(taxid))
# taxonomy
taxtable <- taxid
rownames(taxtable) <- paste0("ASV", seq(nrow(taxid)))
colnames(taxtable) <- c("Kingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species")
# counts
assay_data <- seqtab.nochim
colnames(assay_data) <- paste0("ASV",  seq(ncol(assay_data)))
assay_data <- t(assay_data)
# tse
tse <- TreeSummarizedExperiment(assays = list(counts = assay_data),
                                rowData = taxtable,
                                colData = samples_meta)
referenceSeq(tse) <- repseq
tse
```

```
class: TreeSummarizedExperiment
dim: 1558 18
metadata(0):
assays(1): counts
rownames(1558): ASV1 ASV2 ... ASV1557 ASV1558
rowData names(7): Kingdom Phylum ... Genus Species
colnames(18): Sample01 Sample02 ... Sample17 Sample18
colData names(5): Name Age Filtered Algae Group
reducedDimNames(0):
mainExpName: NULL
altExpNames(0):
rowLinks: NULL
rowTree: NULL
colLinks: NULL
colTree: NULL
referenceSeq: a DNAStringSet (1558 sequences)
```

Some pruning of data

```
#remove taxa with unknown kingdom
tse <- tse[!rowData(tse)$Kingdom %in% NA]
#remove chloroplastic
tse <- tse[!rowData(tse)$Order %in% "Chloroplast"]
#remove mitochondrial
tse <- tse[!rowData(tse)$Family %in% "Mitochondria"]
#remove negative control sample (Sample18)
tse <- tse[,!colData(tse)$Name %in% "control"]
#final object dimensions
dim(tse)
```

```
[1] 1417   17
```

In the end we have 17 samples and 1417 taxa

Writing data

Last step is to save data to suitable file formats.

```
saveRDS(tse, "rds/tse.rds")
```

All variant sequences are saved to fasta

```
tse %>% referenceSeq() %>% writeXStringSet(paste0(exportloc,"/repseq.fasta"),
                                           append=FALSE, compress=FALSE,
                                           format="fasta")
```

Taxonomy is brought from rowData and written as tsv

```
taxfile <- as.data.frame(rowData(tse))
taxfile %>% rownames_to_column(var = "Variant") %>%
  write_tsv(file=paste0(exportloc,"/taxonomy.tsv"))
```

Metadata is brought from colData and written to file

```
metadf <- data.frame(Sampleid = rownames(colData(tse)), colData(tse))
write_tsv(metadf, paste0(exportloc, "/metadata.tsv"))
```

Counts are brought from assays and written to file

```
ASV_counts <- as.data.frame(assays(tse)$counts)
ASV_counts %>% rownames_to_column(var= "Variant") %>%
write_tsv(file = paste0(exportloc, "/fishery_asvs.tsv"))
```