DADA2 analysis of 16S rRNA gene amplicon sequencing reads–single end or forward only

Jianshu Zhao

2021-03-29

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(fig.width = 10)
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Introduction

Implementing DADA2 pipeline for resolving sequence variants from 16S rRNA gene amplicon ***paired-end*** sequencing reads, adopting the tutorial from https://benjjneb.github.io/dada2/tutorial.html and https://benjjneb.github.io/dada2/bigdata_paired.html with minor adjustments. This report captures all the workflow steps necessary to reproduce the analysis.

Load R packages:

```
## [1] '1.18.0'
```

```
## /Users/jianshuzhao/Github/dada2_wrapper/scripts
```

Get list of input fastq files.

```
# Variable 'input.path' containing path to input fastq files
# directory is inherited from wrapper script dada2_cli.r.

input.file.list <- grep("*fastq", list.files(input.path), value = T)
# input.path <- normalizePath('input/')

# List of input files

# Sort ensures forward/reverse reads are in same order
fnFs <- sort(grep("_R1.*\\.fastq", list.files(input.path), value = T))

# Extract sample names, allowing variable filenames; e.g.
# *_R1[_001].fastq[.gz]
sample.names <- gsub("_R1.*\\.fastq(\\.gz)?", "", fnFs, perl = T)

# Specify the full path to the fnFs and fnRs
fnFs <- file.path(input.path, fnFs)
```

Generate quality plots for FWD and REV reads and store in Read_QC folder.

```
# Create output folder NOTE: variable 'output.dir' containing
# name of output folder is inherited from wrapper script
# dada2_cli.r.
cwd <- getwd()

readQC.folder <- file.path(cwd, output.dir, "Read_QC")
ifelse(!dir.exists(readQC.folder), dir.create(readQC.folder,
    recursive = TRUE), FALSE)
```

```
## [1] FALSE
```

```
# Generate plots and save to folder in multi-page pdf

# Forward reads
fwd.qc.plots.list <- list()
for (i in 1:length(fnFs)) {
    fwd.qc.plots.list[[i]] <- plotQualityProfile(fnFs[i])
    rm(i)
}
# Save to file
pdf(paste0(readQC.folder, "/FWD_read_plot.pdf"), onefile = TRUE)
marrangeGrob(fwd.qc.plots.list, ncol = 2, nrow = 3, top = NULL)
```

```
dev.off()
```

```
## pdf
##   2
```

```
rm(fwd.qc.plots.list)
```

Trim and filter reads.

```
# Create filtered_input/ subdirectory for storing filtered
# fastq reads
filt_path <- file.path(cwd, output.dir, "filtered_input")
ifelse(!dir.exists(filt_path), dir.create(filt_path, recursive = TRUE),
    FALSE)
```

```
## [1] FALSE
```

```
# Define filenames for filtered input files
filtFs <- file.path(filt_path, paste0(sample.names, "_F_filt.fastq.gz"))
# Filter the forward and reverse reads: Note that: 1. Reads
# are both truncated and then filtered using the maxEE
# expected errors algorighm from UPARSE.  2. Reverse reads
# are truncated to shorter lengths than forward since they
# are much lower quality.  3. _Both_ reads must pass for the
# read pair to be output.  4. Output files are compressed by
# default.

rd.counts <- as.data.frame(filterAndTrim(fnFs, filtFs, truncLen = 225,
    maxN = 0, maxEE = 1, truncQ = 10, rm.phix = TRUE, compress = TRUE,
    multithread = TRUE))
# Table of before/after read counts
rd.counts$ratio <- round(rd.counts$reads.out/rd.counts$reads.in,
    digits = 2)
```

```
rd.counts
```

```
##                              reads.in reads.out ratio
## Orwoll_BI0023_BI_R1.fastq.gz    62724     52608  0.84
## Orwoll_BI0056_BI_R1.fastq.gz    55342     45875  0.83
## Orwoll_BI0131_BI_R1.fastq.gz    55144     46382  0.84
## Orwoll_BI0153_BI_R1.fastq.gz    44610     38107  0.85
## Orwoll_BI0215_BI_R1.fastq.gz    48227     40871  0.85
## Orwoll_BI0353_BI_R1.fastq.gz    54271     47997  0.88
```

```
# Write rd.counts table to file in readQC.folder
write.table(rd.counts, paste0(readQC.folder, "/Read_counts_after_filtering.tsv"),
    sep = "\t", quote = F, eol = "\n", col.names = NA)
```

Learn the error rates.

The DADA2 algorithm depends on a parametric error model (err) and every amplicon dataset has a different set of error rates. The learnErrors method learns the error model from the data, by alternating estimation of the error rates and inference of sample composition until they converge on a jointly consistent solution. As in many optimization problems, the algorithm must begin with an initial guess, for which the maximum possible error rates in this data are used (the error rates if only the most abundant sequence is correct and all the rest are errors).

```
set.seed(100)
# Filtered forward reads
errF <- learnErrors(filtFs, nread = 1e+06, multithread = TRUE)
```

```
## Warning in learnErrors(filtFs, nread = 1e+06, multithread = TRUE): The nreads
## parameter is DEPRECATED. Please update your code with the nbases parameter.
```

```
## 61164000 total bases in 271840 reads from 6 samples will be used for learning the error rates.
```

```
# Visualize the estimated error rates Forward
# plotErrors(errF, nominalQ=TRUE) Save to file
ggsave(paste0(readQC.folder, "/Error_rates_per_sample_single.pdf"),
    plotErrors(errF, nominalQ = TRUE), device = "pdf")
```

```
## Saving 10 x 4.5 in image
```

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

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

```
# Reverse plotErrors(errR, nominalQ=TRUE)
```

The error rates for each possible transition (eg. A->C, A->G, ...) are shown. Points are the observed error rates for each consensus quality score. The black line shows the estimated error rates after convergence. The red line shows the error rates expected under the nominal definition of the Q-value. If the black line (the estimated rates) fits the observed rates well, and the error rates drop with increased quality as expected, then everything looks reasonable and can proceed with confidence.

Infer Sequence Variants

This step consists of dereplication, sample inference, and merging of paired reads

Dereplication combines all identical sequencing reads into into "unique sequences" with a corresponding "abundance": the number of reads with that unique sequence. DADA2 retains a summary of the quality information associated with each unique sequence. The consensus quality profile of a unique sequence is the

average of the positional qualities from the dereplicated reads. These quality profiles inform the error model of the subsequent denoising step, significantly increasing DADA2's accuracy.

The sample inference step performs the core sequence-variant inference algorithm to the dereplicated data.

Spurious sequence variants are further reduced by merging overlapping reads. The core function here is mergePairs, which depends on the forward and reverse re.samples being in matching order at the time they were dereplicated.

```
# Sample inference of dereplicated reads, and merger of
# paired-end reads
mergers <- vector("list", length(sample.names))
names(mergers) <- sample.names
names(filtFs) <- sample.names
for (sam in sample.names) {
    cat("Processing:", sam, "\n")
    derepF <- derepFastq(filtFs[[sam]])
    ddF <- dada(derepF, err = errF, multithread = TRUE)
    mergers[[sam]] <- ddF
}
```

```
## Processing: Orwoll_BI0023_BI
## Sample 1 - 52608 reads in 11936 unique sequences.
## Processing: Orwoll_BI0056_BI
## Sample 1 - 45875 reads in 7970 unique sequences.
## Processing: Orwoll_BI0131_BI
## Sample 1 - 46382 reads in 8487 unique sequences.
## Processing: Orwoll_BI0153_BI
## Sample 1 - 38107 reads in 8321 unique sequences.
## Processing: Orwoll_BI0215_BI
## Sample 1 - 40871 reads in 9033 unique sequences.
## Processing: Orwoll_BI0353_BI
## Sample 1 - 47997 reads in 8590 unique sequences.
```

```
rm(derepF)
```

Construct sequence table

We can now construct a "sequence table" of our samples, a higher-resolution version of the "OTU table" produced by classical methods:

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

```
## [1]    6 1423
```

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

```
##
##  225
## 1423
```

```
# The sequence table is a matrix with rows corresponding to
# (and named by) the samples, and columns corresponding to
# (and named by) the sequence variants.
```

Remove chimeras

The core dada method removes substitution and indel errors, but chimeras remain. Fortunately, the accuracy of the sequences after denoising makes identifying chimeras simpler than it is when dealing with fuzzy OTUs: all sequences which can be exactly reconstructed as a bimera (two-parent chimera) from more abundant sequences.

```
# Remove chimeric sequences:
seqtab.nochim <- removeBimeraDenovo(seqtab, method = "consensus",
    multithread = TRUE, verbose = TRUE)
```

```
## Identified 674 bimeras out of 1423 input sequences.
```

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

```
## [1]   6 749
```

```
# ratio of chimeric sequence reads
1 - sum(seqtab.nochim)/sum(seqtab)
```

```
## [1] 0.09596727
```

```
# write sequence variants count table to file
write.table(t(seqtab.nochim), paste0(output.dir, "/all_samples_SV-counts.tsv"),
    sep = "\t", eol = "\n", quote = F, col.names = NA)
# write OTU table to file
saveRDS(seqtab.nochim, paste0(output.dir, "/seqtab_final.rds"))
```

IMPORTANT: Most of your **reads** should remain after chimera removal (it is not uncommon for a majority of **sequence variants** to be removed though). If most of your reads were removed as chimeric, upstream processing may need to be revisited. In almost all cases this is caused by primer sequences with ambiguous nucleotides that were not removed prior to beginning the DADA2 pipeline.

Track reads through the pipeline

As a final check of the progress, look at the number of reads that made it through each step in the pipeline. This is a great place to do a last sanity check. Outside of filtering (depending on how stringent you want to be) there should no step in which a majority of reads are lost. If a majority of reads failed to merge, you may need to revisit the truncLen parameter used in the filtering step and make sure that the truncated reads span your amplicon. If a majority of reads failed to pass the chimera check, you may need to revisit the removal of primers, as the ambiguous nucleotides in unremoved primers interfere with chimera identification.

```
getN <- function(x) sum(getUniques(x))
track <- cbind(rd.counts, sapply(mergers, getN), rowSums(seqtab),
    rowSums(seqtab.nochim))
colnames(track) <- c("input", "filtered", "ratio", "single",
    "tabled", "nonchim")
rownames(track) <- sample.names
# print table
track
```

```
##                  input filtered ratio single tabled nonchim
## Orwoll_BI0023_BI 62724    52608  0.84  51427  51427   46864
## Orwoll_BI0056_BI 55342    45875  0.83  45387  45387   39780
## Orwoll_BI0131_BI 55144    46382  0.84  45709  45709   43504
## Orwoll_BI0153_BI 44610    38107  0.85  37558  37558   34041
## Orwoll_BI0215_BI 48227    40871  0.85  40208  40208   36140
## Orwoll_BI0353_BI 54271    47997  0.88  47396  47396   41667
```

```
# save to file
write.table(track, paste0(readQC.folder, "/Read_counts_at_each_step.tsv"),
    sep = "\t", quote = F, eol = "\n", col.names = NA)
```

Align sequences and reconstruct phylogeny

Multiple sequence alignment of resolved sequence variants is used to generate a phylogenetic tree, which is required for calculating UniFrac beta-diversity distances between microbiome samples.

```
# Get sequences
seqs <- getSequences(seqtab.nochim)
names(seqs) <- seqs  # This propagates to the tip labels of the tree
# Multiple seqeuence alignment
mult <- msa(seqs, method = "ClustalOmega", type = "dna", order = "input")
```

```
## using Gonnet
```

```
# Save msa to file; convert first to phangorn object
phang.align <- as.phyDat(mult, type = "DNA", names = getSequences(seqtab.nochim))
write.phyDat(phang.align, format = "fasta", file = paste0(output.dir,
    "/msa.fasta"))

# Call FastTree (via 'system') to reconstruct phylogeny
if (unname(Sys.info()["sysname"]) == "Linux") {
    system(paste("../dependencies/FastTreeMP_Linux -gtr -nt ",
        output.dir, "/msa.fasta > ", output.dir, "/FastTree.tre",
        sep = ""))
} else {
    system(paste("../dependencies/FastTreeMP_Darwin -gtr -nt ",
        output.dir, "/msa.fasta > ", output.dir, "/FastTree.tre",
        sep = ""))
}

detach("package:phangorn", unload = TRUE)
detach("package:msa", unload = TRUE)
```

Assign taxonomy

The assignTaxonomy function takes a set of sequences and a training set of taxonomically classified sequences, and outputs the taxonomic assignments with at least minBoot bootstrap confidence. Formatted training datasets for taxonomic assignments can be downloaded from here https://benjjneb.github.io/dada2/training.html.

`assignTaxonomy( ... )` implements the RDP naive Bayesian classifier method described in Wang et al. 2007. In short, the kmer profile of the sequences to be classified are compared against the kmer profiles of all sequences in a training set of sequences with assigned taxonomies. The reference sequence with the most similar profile is used to assign taxonomy to the query sequence, and then a bootstrapping approach is used to assess the confidence assignment at each taxonomic level. The return value of assignTaxonomy(...) is a character matrix, with each row corresponding to an input sequence, and each column corresponding to a taxonomic level.

```
# Assign taxonomy:
taxa.gg13_8 <- assignTaxonomy(seqtab.nochim, "../reference_dbs_16S/gg_13_8_train_set_97.fa.gz",
    multithread = TRUE, tryRC = TRUE)
```

```
# Print first 6 rows of taxonomic assignment
unname(head(taxa.gg13_8))
```

```
##      [,1]          [,2]               [,3]
## [1,] "k__Bacteria" "p__Bacteroidetes"  "c__Bacteroidia"
## [2,] "k__Bacteria" "p__Proteobacteria" "c__Gammaproteobacteria"
## [3,] "k__Bacteria" "p__Bacteroidetes"  "c__Bacteroidia"
## [4,] "k__Bacteria" "p__Firmicutes"     "c__Clostridia"
## [5,] "k__Bacteria" "p__Bacteroidetes"  "c__Bacteroidia"
## [6,] "k__Bacteria" "p__Bacteroidetes"  "c__Bacteroidia"
##      [,4]                   [,5]                     [,6]
## [1,] "o__Bacteroidales"     "f__Bacteroidaceae"      "g__Bacteroides"
## [2,] "o__Enterobacteriales" "f__Enterobacteriaceae" "g__Klebsiella"
## [3,] "o__Bacteroidales"     "f__Bacteroidaceae"      "g__Bacteroides"
## [4,] "o__Clostridiales"     "f__Ruminococcaceae"     "g__Faecalibacterium"
## [5,] "o__Bacteroidales"     "f__Bacteroidaceae"      "g__Bacteroides"
## [6,] "o__Bacteroidales"     "f__Bacteroidaceae"      "g__Bacteroides"
##      [,7]
## [1,] "s__"
## [2,] "s__"
## [3,] NA
## [4,] "s__prausnitzii"
## [5,] "s__ovatus"
## [6,] "s__ovatus"
```

```
# Replace NAs in taxonomy assignment table with prefix
# corresponding to tax rank
taxa.gg13_8.2 <- replaceNA.in.assignedTaxonomy(taxa.gg13_8)

# Write taxa table to file
write.table(taxa.gg13_8.2, paste0(output.dir, "/all_samples_GG13-8-taxonomy.tsv"),
    sep = "\t", eol = "\n", quote = F, col.names = NA)
```

Merge OTU and GG13-8 taxonomy tables

```
otu.gg.tax.table <- merge(t(seqtab.nochim), taxa.gg13_8.2, by = "row.names")
rownames(otu.gg.tax.table) <- otu.gg.tax.table[, 1]
otu.gg.tax.table <- otu.gg.tax.table[, -1]

write.table(otu.gg.tax.table, paste0(output.dir, "/all_samples_SV-counts_and_GG13-8-taxonomy.tsv"),
    sep = "\t", eol = "\n", quote = F, col.names = NA)
```

For RDP and Silva, taxonomic assignment to species level is a two-step process. Fast and appropriate species-level assignment from 16S data is provided by the `assignSpecies( ... )` method. `assignSpecies( ... )` uses exact string matching against a reference database to assign Genus species binomials. In short, query sequence are compared against all reference sequences that had binomial genus-species nomenclature assigned, and the genus-species of all exact matches are recorded and returned if it is unambiguous.

The convenience function `addSpecies( ... )` takes as input a taxonomy table, and outputs a table with an added species column. Only those genus-species binomials which are consistent with the genus assigned in the provided taxonomy table are retained in the output. See here for more on taxonomic assignment https://benjjneb.github.io/dada2/assign.html.

Asign SILVA and RDP taxonomies and merge with OTU table

```
# Assign SILVA taxonomy
taxa.silva <- assignTaxonomy(seqtab.nochim, "../reference_dbs_16S/silva_nr_v128_train_set.fa.gz",
    multithread = TRUE)

# Replace NAs in taxonomy assignment table with prefix
# corresponding to tax rank
taxa.silva.2 <- replaceNA.in.assignedTaxonomy(taxa.silva)

# OMIT APPENDING SPECIES FOR SILVA DUE TO MEMORY CONSTRAINTS
# Append species. Note that appending the argument
# 'allowMultiple=3' will return up to 3 different matched
# species, but if 4 or more are matched it returns NA.
# taxa.silva.species <- addSpecies(taxa.silva,
# '/n/huttenhower_lab/data/dada2_reference_databases/silva_species_assignment_v128.fa.gz')

# Merge with OTU table and save to file
otu.silva.tax.table <- merge(t(seqtab.nochim), taxa.silva.2,
    by = "row.names")
rownames(otu.silva.tax.table) <- otu.silva.tax.table[, 1]
otu.silva.tax.table <- otu.silva.tax.table[, -1]

write.table(otu.silva.tax.table, paste0(output.dir, "/all_samples_SV-counts_and_SILVA-taxonomy.tsv"),
    sep = "\t", eol = "\n", quote = F, col.names = NA)

# Assign RDP taxonomy
taxa.rdp <- assignTaxonomy(seqtab.nochim, "../reference_dbs_16S/rdp_train_set_18.fa.gz",
    multithread = TRUE)
```

```
## Warning in .Call2("fasta_index", filexp_list, nrec, skip, seek.first.rec, :
## reading FASTA file ../reference_dbs_16S/rdp_train_set_18.fa.gz: ignored 9
## invalid one-letter sequence codes
```

```
# Replace NAs in taxonomy assignment table with prefix
# corresponding to tax rank
taxa.rdp.2 <- replaceNA.in.assignedTaxonomy(taxa.rdp)

# OMIT APPENDING SPECIES FOR RDP DUE TO MEMORY CONSTRAINTS
# Append species. Note that appending the argument
# 'allowMultiple=3' will return up to 3 different matched
# species, but if 4 or more are matched it returns NA.
# taxa.rdp.species <- addSpecies(taxa.rdp,
# '/n/huttenhower_lab/data/dada2_reference_databases/rdp_species_assignment_16.fa.gz')


# Merge with OTU table and save to file
otu.rdp.tax.table <- merge(t(seqtab.nochim), taxa.rdp.2, by = "row.names")
rownames(otu.rdp.tax.table) <- otu.rdp.tax.table[, 1]
otu.rdp.tax.table <- otu.rdp.tax.table[, -1]
# extract representative sequences and save to file

write.table(otu.rdp.tax.table, paste0(output.dir, "/all_samples_SV-counts_and_RDP-taxonomy.tsv"),
    sep = "\t", eol = "\n", quote = F, col.names = NA)
```

Session Info:

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: macOS  10.16
##
## Matrix products: default
## BLAS/LAPACK: /Users/jianshuzhao/Documents/miniconda3/lib/libopenblasp-r0.3.10.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] knitr_1.31                  rmarkdown_2.7
##  [3] gridExtra_2.3               dada2_1.18.0
##  [5] Rcpp_1.0.6                  Rsamtools_2.6.0
##  [7] Biostrings_2.58.0           XVector_0.30.0
##  [9] BiocParallel_1.24.1         SummarizedExperiment_1.20.0
## [11] Biobase_2.50.0              MatrixGenerics_1.2.1
## [13] GenomicRanges_1.42.0        GenomeInfoDb_1.26.4
## [15] IRanges_2.24.1              S4Vectors_0.28.1
## [17] BiocGenerics_0.36.0         ggplot2_3.3.3
## [19] GUniFrac_1.1                matrixStats_0.58.0
## [21] ape_5.4-1                   vegan_2.5-7
## [23] lattice_0.20-41             permute_0.9-5
## [25] RCurl_1.98-1.3              BiocManager_1.30.12
##
## loaded via a namespace (and not attached):
##  [1] png_0.1-7                digest_0.6.27            utf8_1.2.1
##  [4] R6_2.5.0                 plyr_1.8.6               ShortRead_1.48.0
##  [7] evaluate_0.14            pillar_1.5.1             zlibbioc_1.36.0
## [10] rlang_0.4.10             Matrix_1.3-2             labeling_0.4.2
## [13] splines_4.0.2            stringr_1.4.0            igraph_1.2.6
## [16] munsell_0.5.0            DelayedArray_0.16.3      xfun_0.22
## [19] compiler_4.0.2           pkgconfig_2.0.3          mgcv_1.8-34
## [22] htmltools_0.5.1          tibble_3.1.0             GenomeInfoDbData_1.2.4
## [25] codetools_0.2-18         quadprog_1.5-8           fansi_0.4.2
## [28] crayon_1.4.1             withr_2.4.1              GenomicAlignments_1.26.0
## [31] MASS_7.3-53.1            bitops_1.0-6             grid_4.0.2
## [34] nlme_3.1-152             gtable_0.3.0             lifecycle_1.0.0
## [37] formatR_1.8              magrittr_2.0.1           scales_1.1.1
## [40] RcppParallel_5.0.3       stringi_1.5.3            farver_2.1.0
## [43] hwriter_1.3.2            reshape2_1.4.4           latticeExtra_0.6-29
## [46] ellipsis_0.3.1           vctrs_0.3.6              fastmatch_1.1-0
## [49] RColorBrewer_1.1-2       tools_4.0.2              glue_1.4.2
## [52] jpeg_0.1-8.1             yaml_2.2.1               colorspace_2.0-0
## [55] cluster_2.1.1
```