Decontam-SARS Samples

Brittany Seibert

# Before you begin:

These scripts were tailored for the analyses performed in:

Seibert et al, 2021, *Mild and severe SARS-CoV-2 infection induces respiratory and intestinal microbiome changes in the K18-hACE2 transgenic mouse model*

# 

# Load the needed packages

library(seqinr)  
library(decontam)  
library(phyloseq)  
library(ggplot2)

# 

# Information about Decontam

In addition to DADA2, @bejcal et al. also created a program for removing contaminants based on incorporated blanks called decontam (Nicole Davis et al. publication <https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0605-2>). Documentation is available at this website: <https://benjjneb.github.io/decontam/vignettes/decontam_intro.html>). Here, we will apply it without DNA concentrations – using prevalence of ASVs in the incorporated blanks – starting from our **count table** generated in dada2.

## Import Files from dada2

First, we will need to import the ASV count table, taxonomy file, and fasta file that were generated from dada2.

**The count table must be read in as a matrix to be used for decontam**

# ASV Count Table   
asv\_tab <- as.matrix(read.table("/Users/ASVs\_counts.tsv", sep = '\t', header = TRUE, row.names = 1))  
colnames(asv\_tab) # List the colnames   
  
# Taxonomy File  
asv\_tax\_dada2 <- as.matrix(read.table("/Users/ASVs\_taxonomy\_dada.tsv", sep = '\t', header = TRUE, row.names = 1))  
colnames(asv\_tax\_dada2) # List the colnames  
  
# Fasta File (package seqinr)  
asv\_fasta <- read.fasta("/Users/ASVs.fa")

## Set Controls

We will need to set the vector of the samples that are considered negative controls in our data set.

I will treat each negative control from the sequencing batch as their own negative control samples against all samples (n = 4)

# We will need the column number of the negative controls. These are the sampleIDs   
which(colnames(asv\_tab)=="BS.274") #column number for BS-274  
which(colnames(asv\_tab)=="BS.275") #column number for BS-275  
which(colnames(asv\_tab)=="BS.276") #column number for BS-276  
which(colnames(asv\_tab)=="BS.276") #column number for BS-277  
  
# Set the vector containing the negative controls (which are BS-274, BS-275, BS-276, BS-277) for decontam.   
# Negative Samples labeled = TRUE   
# True Samples labeled = FALSE  
vector\_for\_decontam <- c(rep(FALSE, 58), rep(TRUE, 2), rep(FALSE, 12), rep(TRUE, 2))  
vector\_for\_decontam

## Implement the Decontam algorithm using the Prevalence based method

*seqtab*: Integer matrix or phyloseq object. A feature table recording the observed abundances of each sequence variant (or OTU) in each sample. Rows should correspond to samples, and columns to sequences (or OTUs).  
*conc*: Required if performing frequency-based testing. A quantitative measure of the concentration of amplified DNA in each sample prior to sequencing.  
*neg*: Required if performing prevalence-based testing. TRUE if sample is a negative control, and FALSE if not (NA entries are not included in the testing).  
*method*: Default). frequency, prevalence or combined will be automatically selected based on whether just conc, just neg, or both were provided.  
*batch*: Default NULL. If provided, should be a vector of length equal to the number of input samples which specifies which batch each sample belongs to (eg. sequencing run). Contaminants identification will be performed independently within each batch

# Contaminants are identified by increased prevalence in negative controls. The default threshold for a contaminant is that it reaches a probability of 0.1 in the statistical test being performed.  
contam\_df <- isContaminant(t(asv\_tab), neg=vector\_for\_decontam)

# Report the number of ASVs that were not contaminants (FALSE) and those that were contaminants (TRUE)  
table(contam\_df$contaminant) #identified 14 as contaminants  
  
# Create vector containing the identified contaminant IDs  
contam\_asvs <- row.names(contam\_df[contam\_df$contaminant == TRUE, ])  
contam\_asvs  
  
# Look at the 3 contaminants (Which ASV they were and the taxonomic classification)  
contam <- asv\_tax\_dada2[row.names(asv\_tax\_dada2) %in% contam\_asvs, ]  
contam  
  
# Export the table of contaminants in an excel file   
write.csv(contam, "/Users/contaminants\_taxonomy.csv")

## Export new fasta file, count table, and taxonomy table without the contaminants

I will export both analysis thresholds just in case I want to compare the differences in future analysis.  
After looking into the ASVs I will use for future analysis in which only ASV 470 is removed since all of the other bacteria identified were biologically relevant.

# Fasta File  
contam\_indices <- which(asv\_fasta %in% paste0(">", "ASV\_470"))  
dont\_want <- sort(c(contam\_indices, contam\_indices + 1))  
asv\_fasta\_no\_contam\_470 <- asv\_fasta[- dont\_want]  
  
# ASV Count table  
asv\_tab\_no\_contam\_470 <- asv\_tab[!row.names(asv\_tab) %in% "ASV\_470", ]  
  
# Taxonomy File  
asv\_tax\_no\_contam\_470 <- asv\_tax\_dada2[!row.names(asv\_tax\_dada2) %in% "ASV\_470", ]  
  
# And now writing them out to files  
write(asv\_fasta\_no\_contam\_470, "/Users/ASVs\_no\_contam\_470.fa")  
  
write.table(asv\_tab\_no\_contam\_470, "/Users/asv\_tab\_no\_contam\_470.tsv", sep="\t", quote=F, col.names=NA)  
  
write.table(asv\_tax\_no\_contam\_470, "/Users/asv\_tax\_no\_contam\_470.tsv", sep="\t", quote=F, col.names=NA)