Sample-Processing and Setup

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*

Purpose:

The purpose of this script is to proccess the decontam samples to remove any variants with unknown phylas, are classified as Eukaryotes, Chloroplasts or Mitochondria, and to remove any samples with a coverage of less 10,000 reads/sample

Load the needed packages

```
library(phyloseq)
library(dplyr)
library(ggplot2)
```

Import files from decontam

First, we will need to import the ASV count table, taxonomy file from decontam and the metadata file.

In order to read the metadata into a phyloseq object you need to make the SampleID (first column) the row headers

```
# Set working directory
setwd = "/Users/baseibert/Perez_Lab/Projects/Microbiome/DataAnalysis/SARS/Analysis-Files/R_Files/"

# Import count table
count <-
read.table("/Users/baseibert/Perez_Lab/Projects/Microbiome/Projects/K18_SARS/DataAnalysis/Analysis_Files/R_Files/Decontam/asv_tab_no_contam_470.tsv", header=T, row.names=1, check.names=F, sep="\t")

# Import taxonomy file
taxa <-
as.matrix(read.table("/Users/baseibert/Perez_Lab/Projects/Microbiome/Projects/K18_SARS/DataAnalysis/Analysis_Files/R_Files/Decontam/asv_tax_no_contam_470.tsv", header=T, row.names=1, check.names=F, sep="\t"))

# Import metadata file
metadata <-
read.csv("/Users/baseibert/Perez_Lab/Projects/Microbiome/Projects/K18_SARS/DataAnalysis/Analysis_Files/R_Files/Import/Sars_metadata2.csv", header=T, row.names=1, check.names=F)
```

Filtering variants for ONLY BACTERIA

When the organisms from a set of samples are well-represented in the taxonomic reference database, it is advisable to filter variants in which a high-rank taxonomy could not be assigned. These are most likely sequence artifacts that don't exist in nature. Therefore, we will filter out any variants that are not classified as bacteria, have an unclassified phyla or are identified as Chloroplast or Mitochondria.

Import files into PhyloSeq

```
# Import count table into phyloseq
ps = otu_table(count, taxa_are_rows = TRUE)

# Import taxonomy table into phyloseq
tax = tax_table(taxa)

# Import metadata into phyloseq object
sample = sample_data(metadata)
```

```
# Merge the count table, taxonomy table and sample data
Phylo = merge_phyloseq(ps, tax, sample)
Phylo
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1594 taxa and 74 samples ]
## sample_data() Sample Data:       [ 74 samples by 12 sample variables ]
## tax_table()   Taxonomy Table:    [ 1594 taxa by 6 taxonomic ranks ]
```

First we will filter out any Phyla that are unassigned

```
# Create table, number of features for each phyla
table(tax_table(Phylo)[, "Phylum"], exclude = NULL)
```

```
##
##  Abditibacteriota   Acidobacteriota  Actinobacteriota      Bacteroidota
##                 1                 1                83                83
##  Campilobacterota       Chloroflexi     Cyanobacteria      Deinococcota
##                 1                 3                17                 3
##        Firmicutes   Gemmatimonadota    Proteobacteria Verrucomicrobiota
##              1197                 1               159                 4
##              <NA>
##                41
```

```
# Remove the ASVs and name as uncharacterized if the Phylum has an NA
Phylo.filter <- subset_taxa(Phylo, !is.na(Phylum) & !Phylum %in% c("", "uncharacterized"))
table(tax_table(Phylo.filter)[, "Phylum"], exclude = NULL)
```

```
##
##  Abditibacteriota   Acidobacteriota  Actinobacteriota      Bacteroidota
##                 1                 1                83                83
##  Campilobacterota       Chloroflexi     Cyanobacteria      Deinococcota
##                 1                 3                17                 3
##        Firmicutes   Gemmatimonadota    Proteobacteria Verrucomicrobiota
##              1197                 1               159                 4
```

Next we will filter out Eukaryotes, chloroplasts and mitochondria, because we only intended to amplify bacterial sequences

```
# Remove the ASVs and name as uncharacterized if the Phylum has an NA
table(tax_table(Phylo)[, "Kingdom"], exclude = NULL)
```

```
##
## Bacteria
##     1594
```

```
Phylo.filter
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1553 taxa and 74 samples ]
## sample_data() Sample Data:       [ 74 samples by 12 sample variables ]
## tax_table()   Taxonomy Table:    [ 1553 taxa by 6 taxonomic ranks ]
```

```
Phylo.filter1 <- Phylo.filter %>%
  subset_taxa(
    Kingdom == "Bacteria" &
    Class   != "Chloroplast" &
    Family  != "Mitochondria")
Phylo.filter1
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1282 taxa and 74 samples ]
## sample_data() Sample Data:       [ 74 samples by 12 sample variables ]
## tax_table()   Taxonomy Table:    [ 1282 taxa by 6 taxonomic ranks ]
```

Now I want to separate the samples from the negative and positive controls

```
# Remove the ASVs and name as uncharacterized if the Phylum has an NA
Phylo.samples <- Phylo.filter1 %>%
  subset_samples(sampleClass == "sample")
Phylo.samples
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1282 taxa and 69 samples ]
## sample_data() Sample Data:       [ 69 samples by 12 sample variables ]
## tax_table()   Taxonomy Table:    [ 1282 taxa by 6 taxonomic ranks ]

Phylo.control <- Phylo.filter1 %>%
  subset_samples(sampleClass == "Extraction" | sampleClass == "PCR" | sampleClass == "Mock")
Phylo.control

## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1282 taxa and 5 samples ]
## sample_data() Sample Data:       [ 5 samples by 12 sample variables ]
## tax_table()   Taxonomy Table:    [ 1282 taxa by 6 taxonomic ranks ]
```

After looking at the barplots, and doing the Grubbs test on outliers for sample 10, I will need to remove sample 10 (BS.208)
I will not remove the other 2 samples that have low coverage since they will be removed during rarification and not used in the taxonomic barplots

```
# Remove the ASVs and name as uncharacterized if the Phylum has an NA
Phylo.samples.final <- Phylo.samples %>%
  subset_samples(SampleID != "BS.208")
Phylo.samples.final

## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1282 taxa and 68 samples ]
## sample_data() Sample Data:       [ 68 samples by 12 sample variables ]
## tax_table()   Taxonomy Table:    [ 1282 taxa by 6 taxonomic ranks ]

# Save sample.rare to a file so that i can reload the rarified file
saveRDS(object = Phylo.samples.final, file = "Phylo.samples.final.rds")
```

Figure 1 : Coverage Boxplot

Then lets create a boxplot comparing negative controls to samples when we look at coverage. This could be a good way to show that the samples have much higher reads/sample compared to the negative controls, especially the lungs, since those can be easily contaminated. Negative controls included: 2 Extraction Negative controls, 2 PCR negative controls

```
# Make a data frame with a column for the read counts of each sample
read.counts <- data.frame(sum = sample_sums(Phylo.filter1))

# Add metadata to read.counts
read.counts <- tibble::rownames_to_column(read.counts, "SampleID")
read.counts.mt <-left_join(read.counts, metadata, by = "SampleID")

# Filter out the samples that will be removed from the analysis
read.counts.mt <- read.counts.mt %>%
  filter(SampleID != "BS.223") %>%        #this sample has very low coverage and was removed from
analysis
  filter(SampleID != "BS.219") %>%        #this sample has very low coverage and was removed from
analysis
  filter(SampleID != "BS.273") %>%        #this is the mock community and decided to not show it
  filter(SampleID != "BS.208")            #this sample was an outlier in the taxonomy analysis so it
was removed for analysis

# Calculate the means of each sample
means <- aggregate(sum ~  sampleClass + SampleType, read.counts.mt, mean)
means

##   sampleClass SampleType      sum
## 1      sample      Cecum 46826.59
## 2  Extraction Extraction  5824.00
## 3      sample       Lung 42216.84
## 4         PCR        PCR   199.50

# Convert sampleClass and SampleType to a factor with specific levels so that we can order them
read.counts.mt$sampleClass <- factor(read.counts.mt$sampleClass, levels = c("Extraction", "PCR",
"Mock", "sample"))
read.counts.mt$SampleType <- factor(read.counts.mt$SampleType, levels = c("Extraction", "PCR", "Mock",
"Cecum","Lung"))

# Assign the color groups for the graph
colorgroups = c("purple", "black","yellow","red")
```

```r
# Plot the boxplots of the coverage
ggplot(read.counts.mt, aes(x = sampleClass, y = sum, fill = SampleType)) +
  geom_boxplot() +
  theme_bw()+
  theme(panel.grid = element_blank())+
  scale_fill_manual(values = colorgroups)+
  scale_y_continuous(breaks = seq(0, 80000, by=10000), limits=c(0, 80000), labels = scales::comma)+
  ylab("Coverage (Reads / Sample)") +
  theme(axis.title.y = element_blank())
```