# Semester project graphs

2025-04

This R Markdown file shows the graphs, with the corresponding code, that ended up in my final term paper.

The manuscript is available in full here

The entire analysis code is available at github.com/emilykibbler/amp_seq_microbiota

```r
# Fig 1: Number of reads at each filtering step

plotData <- as.data.frame(track) %>%
  gather(type, totals, reads.in, filtered, nonchimeras)

plotData$type <- factor(plotData$type, levels = c("reads.in", "filtered", "nonchimeras"),
                        labels = c("Unfiltered reads", "Filtered and trimmed reads", "Nonchimeric reads

plotData$Sample_type <- factor(plotData$Sample_type, levels = c("experimental", "negative"),
                               labels = c("Patient", "Lab negative control"))

ggplot(plotData, aes(x = Sample_type, y = as.numeric(totals))) +
  geom_boxplot(aes(color = Sample_type)) +
  geom_jitter(aes(color = Sample_type), width = 0.1) +
  facet_grid(cols = vars(type)) +
  scale_color_hue(name = "Sample type") +
  ylab("Reads") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 0, size = 10),
        axis.title.x = element_blank(),
        panel.border = element_rect(color = "gray", fill = NA, linewidth = 1),
        legend.position = "top",
        legend.title = element_text(size = 14),
        legend.text = element_text(size = 12),
        panel.background = element_rect(fill = "gray85"),
        plot.title = element_text(size = 16, face = "bold")) +
  ggtitle("Reads by filtering step")
```

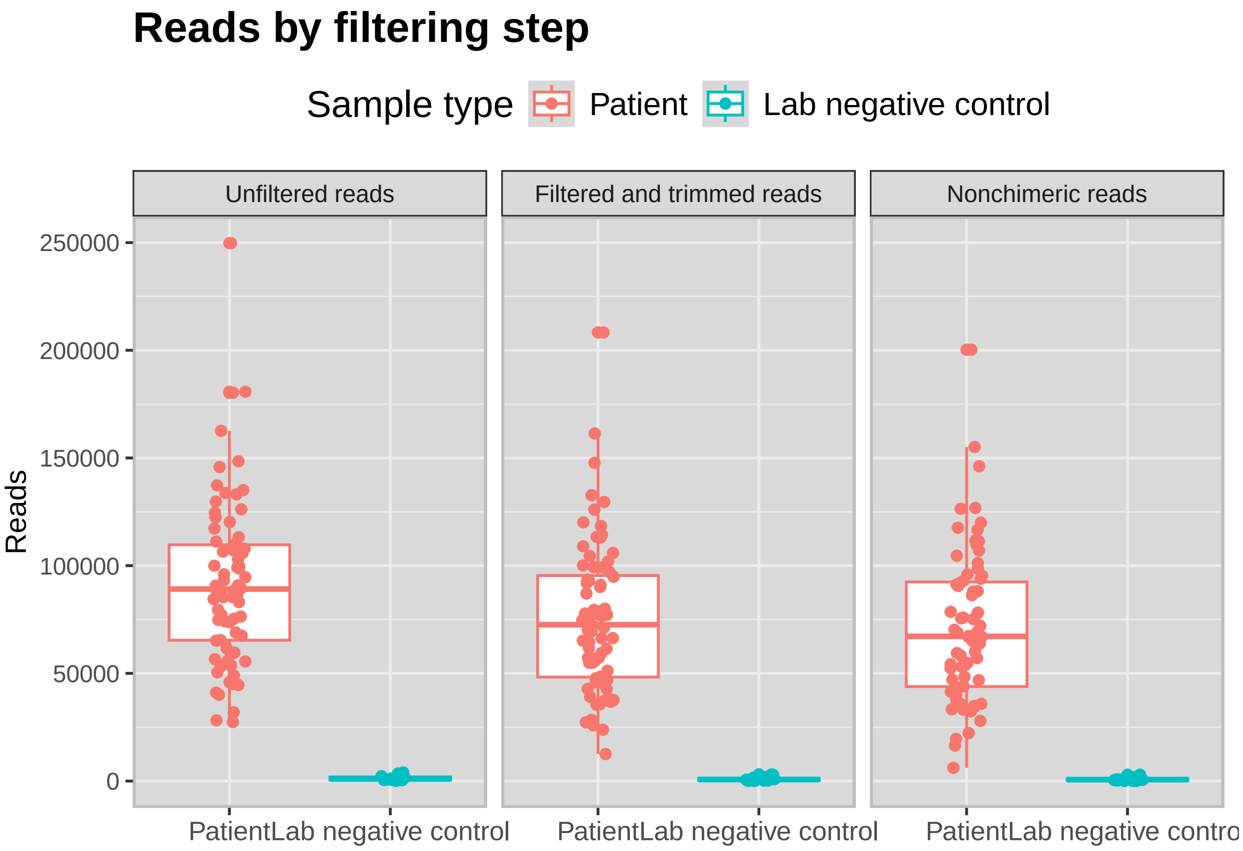

Figure 1. Reads per sample at each QA step.

```
# Fig 2: Venn diagram comparing cleaning methods

# Data using the decontam method: subset by treatment group
phylo_decontam_rar_atb <- subset_and_trim(phylo_decontam_rar, "Group", "Antibiotics")
phylo_decontam_rar_no_atb <- subset_and_trim(phylo_decontam_rar, "Group", "No antibiotics")
# Pull the unique SVs from those subsets
decontam_atb_SVs <- row.names(as.data.frame(phylo_decontam_rar_atb@tax_table))
decontam_no_atb_SVs <- row.names(as.data.frame(phylo_decontam_rar_no_atb@tax_table))

# Data using the Ishaq clean method: subset by treatment group
phylo_clean_rar_atb <- subset_and_trim(clean_phylo_rarified, "Group", "Antibiotics")
phylo_clean_rar_no_atb <- subset_and_trim(clean_phylo_rarified, "Group", "No antibiotics")
# Then pull out the SVs
clean_atb_SVs <- row.names(as.data.frame(phylo_clean_rar_atb@tax_table))
clean_no_atb_SVs <- row.names(as.data.frame(phylo_clean_rar_no_atb@tax_table))

clean_SVs <- c(clean_atb_SVs, clean_no_atb_SVs)
clean_SVs <- unique(clean_SVs)

decontam_SVs <- c(decontam_no_atb_SVs, decontam_atb_SVs)
decontam_SVs <- unique(decontam_SVs)

list(
  "Clean, +antibiotics" = clean_atb_SVs,
  "Decontam, +antibiotics" = decontam_atb_SVs,
  "Clean, -antibiotics" = clean_no_atb_SVs,
  "Decontam, -antibiotics" = decontam_no_atb_SVs
) %>%
  ggVennDiagram() +
    scale_y_continuous(expand = expansion(mult = .1)) +
    scale_x_continuous(expand = expansion(mult = .2)) +
    theme(plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
          legend.position = "bottom") +
    ggplot2::scale_fill_gradient(low = "blue",high = "yellow") +
    ggtitle("SVs -- After Decontam/Clean and Rarification")
```

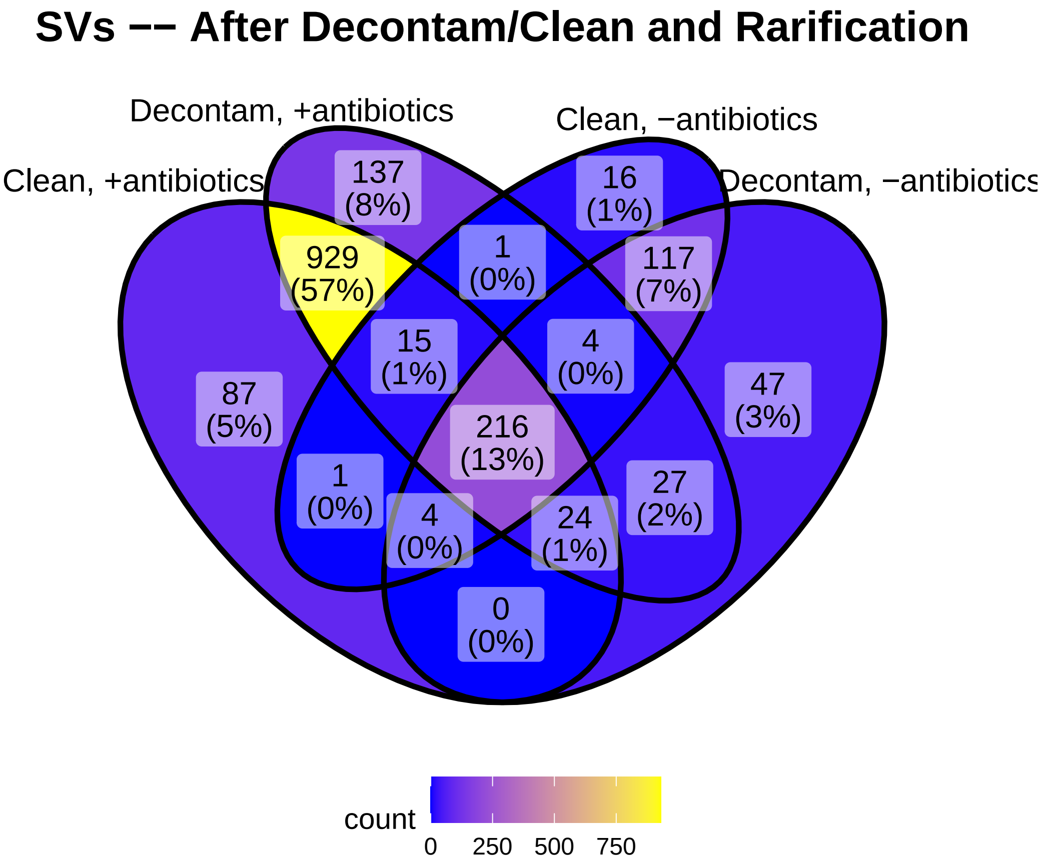

Figure 2. Cleaned (Ishaq method) versus decontaminated (decontam R package) data generally agree, with 929 SVs in common kept and 117 SVs in common discarded. Decontam method was less stringent overall and discarded fewer SVs (Decontam maintained 1521 SVs and Ishaq method maintained 1414). No skew was seen between removal of the SVs in either treatment group; SVs were removed proportionally. Percentages represent the proportion of the relevant SVs in the set of the diagram compared to total unique SVs maintained by either method (1625).

```
# Fig 3: SV venn diagram of SVs found, by method order

dat <- list("Decontamination then Species Assignment" = row.names(phylo_decontam_rar@tax_table),
  "Species Assignment then Decontamination" = row.names(phylo_species_then_decontam_rar@tax_table))

p <- venn.diagram(x = dat,
             filename = NULL,
             imagetype = "png",
             fill = c("yellow", "purple"),
             main = "SVs Found, by Order of Cleaning Steps",
             main.cex = 1.5,
             cat.pos = c(320, 140),
             cat.dist = c(0.1,0.1),
             cat.just = list(c(0.2,-0.1), c(0.8,0)))
grid::grid.draw(p)
```

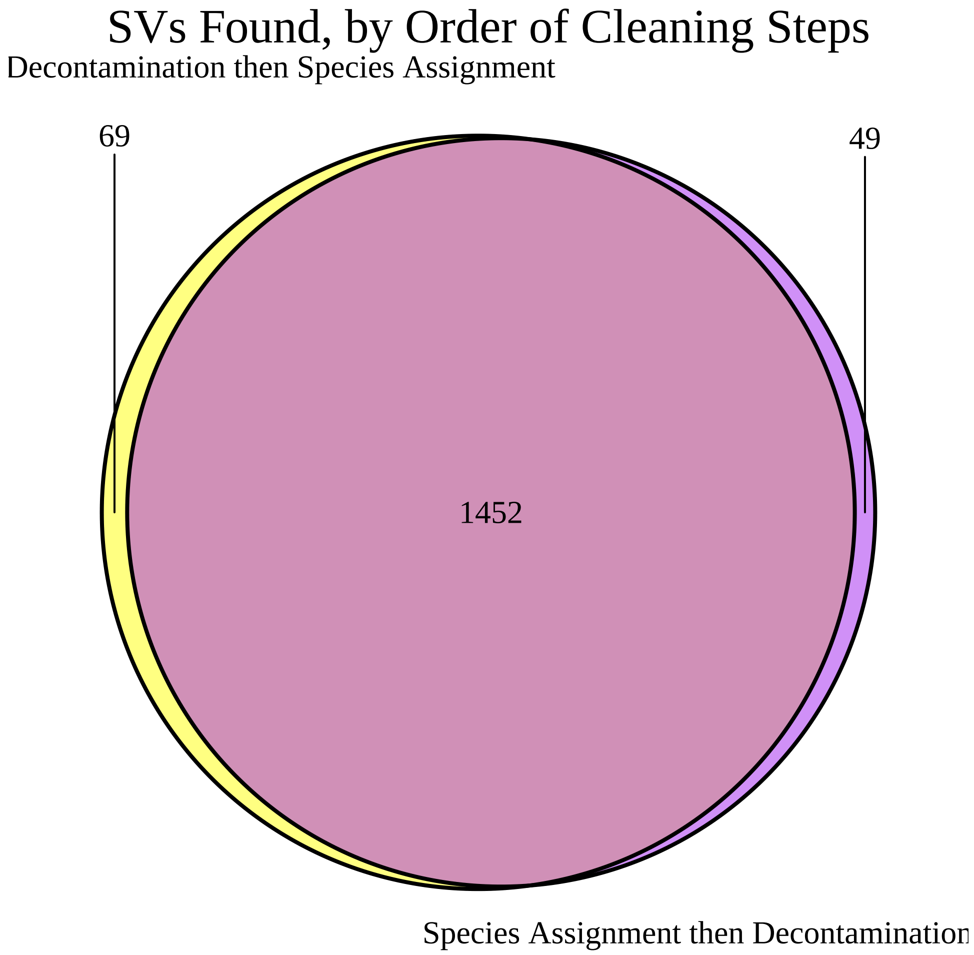

Figure 3. Small changes were observed when the order of operations was changed between species assignment and then decontamination or decontamination and then species assignment. Both orders were followed by rarefaction to 5000 reads. The majority of retained SVs were identical with either method.

```
# Fig 4: Patient metadata

# Make character strings prettier for the graph
patient_data_correlation_summary$Variable <- str_replace_all(patient_data_correlation_summary$Variable,
patient_data_correlation_summary$Variable <- str_to_title(patient_data_correlation_summary$Variable)
chisq_summary$variable <- str_replace_all(chisq_summary$variable, "_", " ")
chisq_summary$variable <- str_to_title(chisq_summary$variable)

p1 <-  subset(patient_data_correlation_summary, Variable != "Group" & Variable != "SampleID" ) %>%
        ggplot(aes(y = Variable, x = as.numeric(p.value), size = 2)) +
        geom_point(aes(color = as.numeric(estimate.cor))) +
        scale_size(guide = "none") +
        scale_color_gradient2(low = "blue",high = "red", midpoint = 0) +
        theme(panel.background = element_rect(fill = "gray"),
              axis.title.y = element_blank(),
              axis.text.y = element_text(size = 9)) +
        scale_x_continuous(breaks = seq(0, 1, 0.1)) +
        labs(color = "Correlation coefficient",
             x = "p-value",
             title = "Correlation of Numeric Patient Metrics",
             subtitle = "Treatment Group vs Control Group") +
        geom_vline(xintercept = 0.05, color = "red") +
        annotate("text", x = 0.03, y = 10, label = "p = 0.05", angle = 90)

p2 <-  subset(chisq_summary, variable != "Group" & variable != "SampleID" ) %>%
        ggplot(aes(y = variable, x = as.numeric(p.value), size = 2)) +
        geom_point(aes(color = as.numeric(`statistic.X-squared`))) +
        scale_size(guide = "none") +
        scale_color_gradient2(low = "blue",high = "red", midpoint = 0) +
        theme(panel.background = element_rect(fill = "gray"),
              axis.title.y = element_blank(),
              axis.text.y = element_text(size = 9)) +
        scale_x_continuous(breaks = seq(0, 1, 0.1)) +
        labs(color = "X-squared",
             x = "p-value",
             title = "Correlation of Categorical Patient Metrics",
             subtitle = "Treatment Group vs Control Group") +
        geom_vline(xintercept = 0.05, color = "red") +
        annotate("text", x = 0.03, y = 25, label = "p = 0.05", angle = 90)

ggarrange(plotlist = list(p1, p2),
          labels = c("A", "B"),
          nrow = 2,
          ncol = 1,
          heights = c(1, 1.25),
          align = "v")
```

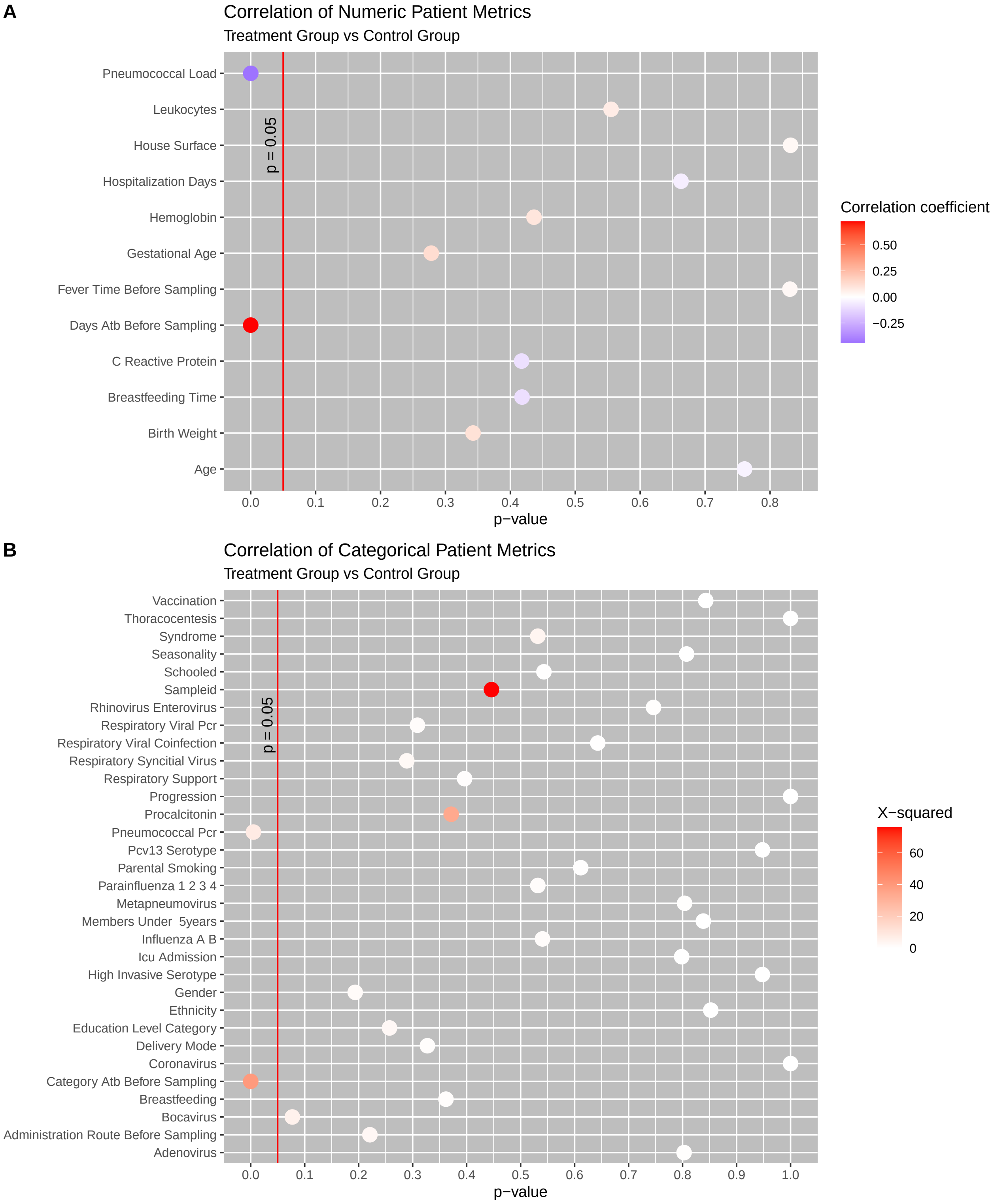

Figure 4. A: Numeric patient metadata correlated with group assignment. B: Categorical patient metadata measured for independence against the group assignment with chi-square tests.

```
# FIg 5: Bar charts of Chao1 and Shannon metrics

phylo_decontam_rar_df <- estimate_richness(phylo_decontam_rar, measures = c("Chao1", "Shannon"))
phylo_decontam_rar_df <- subset(phylo_decontam_rar_df, select = -se.chao1)
phylo_decontam_rar_df$Sample <- row.names(phylo_decontam_rar@sam_data)
phylo_decontam_rar_df$Group <- phylo_decontam_rar@sam_data$Group
phylo_decontam_rar_df <- pivot_longer(phylo_decontam_rar_df, c("Chao1", "Shannon"), names_to = "Metric"
phylo_decontam_rar_df$Group <- str_replace_all(phylo_decontam_rar_df$Group, "No antibiotics", "No antib
phylo_decontam_rar_df$Group <- str_replace_all(phylo_decontam_rar_df$Group, "Antibiotics", "Antibiotics
phylo_decontam_rar_df$Group <- factor(phylo_decontam_rar_df$Group,
                                      levels = c("No antibiotics (n=22)", "Antibiotics (n=54)"))

phylo_decontam_rar_df %>% ggplot(aes(x = Group, y = value)) +
  geom_boxplot(aes(fill = Group), outlier.shape = NA) +
  scale_fill_manual(values = c("#efe68a", "#679acc")) +
  facet_wrap(~Metric, scales = "free") +
  geom_point(position = position_jitter(width = 0.1)) +
  theme(legend.position = "none") +
  theme(plot.title = element_text(size = 14),
        axis.text.x = element_text(size = 10),
        panel.border = element_rect(color = "black", fill = NA, linewidth = 1)) +
  ylab("Alpha Diversity Metric") +
  ggtitle("Diversity Metrics: Cleaned and Rarefied Data")
```

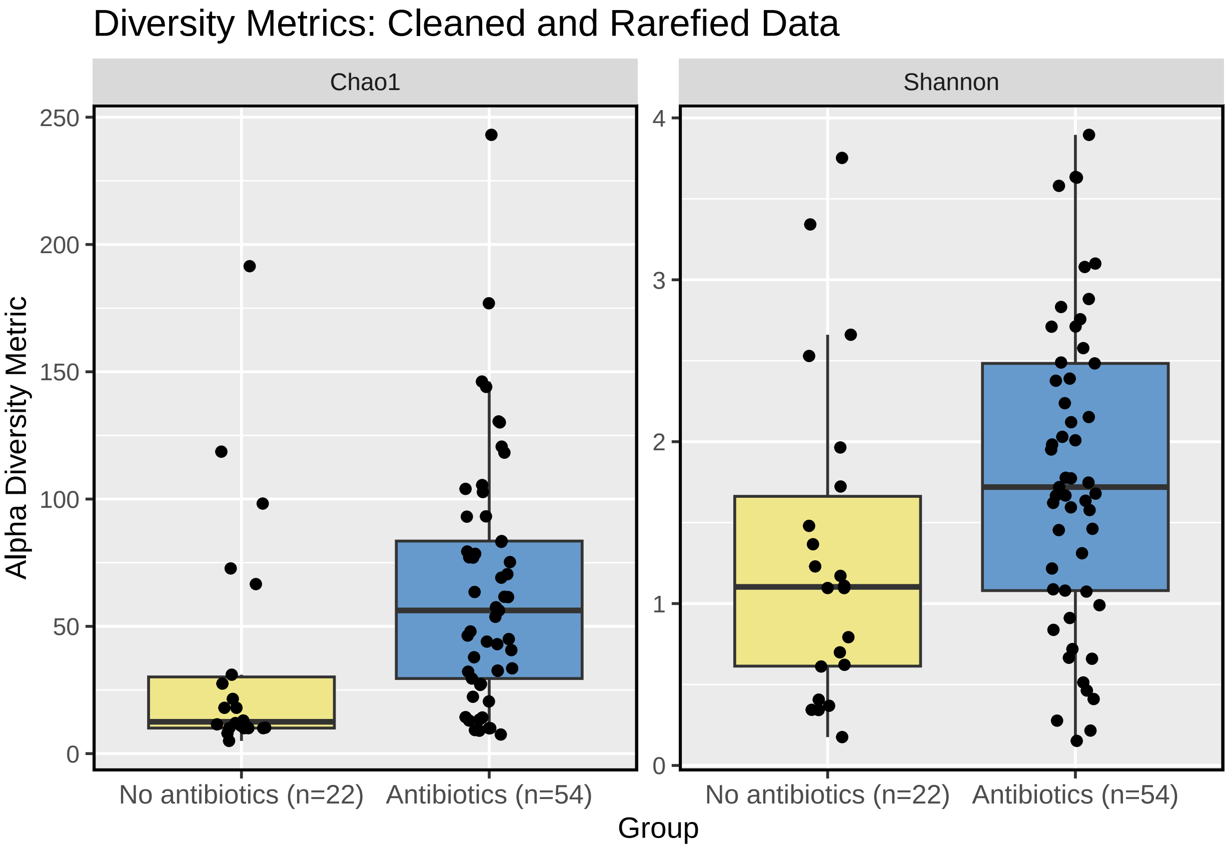

Figure 5. Alpha diversity metrics. This plot closely resembles figure 1 from Henares et al (supplemental figure 5). Minor differences that are apparent in the no-antibiotics group are likely due to the smaller n due to incomplete raw file availability. This is an encouraging finding that overall conclusions may not be majorly affected by the difference in rarefaction (12k in original paper, 5k in this analysis).

```
# Fig 6: Diversity metrics compared by pt dx

df <- estimate_richness(phylo_decontam_rar, measures = "Chao1")
df$SampleID <- phylo_decontam_rar@sam_data$SampleID
df <- full_join(df, as.data.frame(phylo_decontam_rar@sam_data), by = "SampleID")
df$Syndrome <- str_to_sentence(df$Syndrome)
df$Syndrome <- str_replace_all(df$Syndrome, "_", " ")

df %>% ggplot(aes(x = Syndrome, y = Chao1)) +
  geom_violin(trim = TRUE, aes(fill = Syndrome)) +
  geom_boxplot(outlier.shape = NA, width = 0.25) +
  geom_point(size = 1) +
  facet_grid(cols = vars(Group), switch = "x", scales = "free", space = "free") +
  theme(axis.text.x = element_blank(),
        axis.title.x = element_blank(),
        panel.border = element_rect(color = "black", fill = NA),
        legend.position = "top") +
  ylab("Chao1 Richness") +
  ggtitle("Alpha Diversity by Syndrome")
```

```
## Warning: Groups with fewer than two datapoints have been dropped.
## i Set `drop = FALSE` to consider such groups for position adjustment purposes.
```

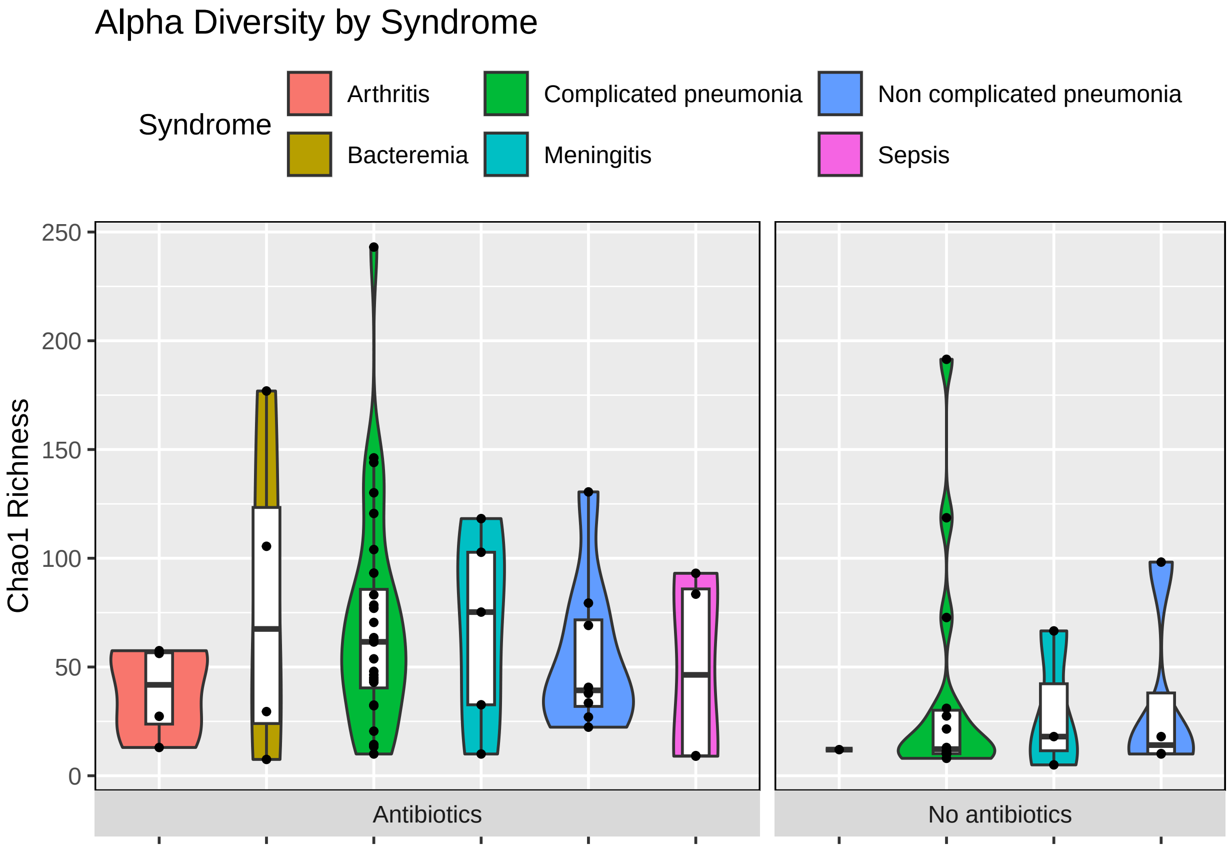

Figure 6. Chao1 richness, split by treatment and then by syndrome. No differences in alpha diversity are attributed to syndrome.

```
# Fig 7: Ordination plot, data before cleaning

phylo_ord <- ordinate(phylo,
                      method = "PCoA",
                      "bray", binary = TRUE)
plot_ordination(phylo, phylo_ord, type = "samples", color = "Group") +
  geom_point(size = 2.5) +
  theme(panel.background = element_rect(fill = "gray90", color = "black"),
    legend.position = "top") +
  ggtitle("Ordination Plot, Before Cleaning", subtitle = "Bray-curtis Distances")
```

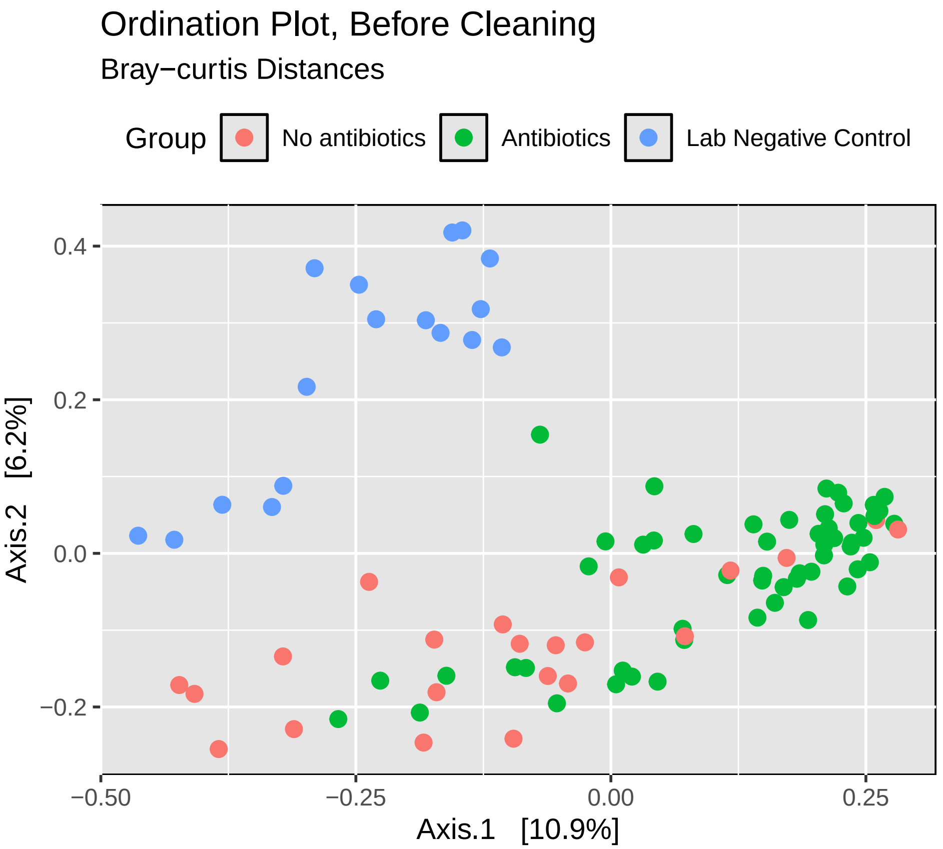

Figure 7. Ordination plot before data cleaning. A delineation is shown between treatment samples and lab negative controls; a smaller but still evident dispersion between the patient groups can be observed.

```
# Fig 8, core SV heat map

# _25 suffix means the data were created with a 25/100 frequency cutoff
atb_phylo.coreW_25 <- subset_samples(phylo.coreW_25, Group == "Antibiotics")
no_atb_phylo.coreW_25 <- subset_samples(phylo.coreW_25, Group == "No antibiotics")

# Aggregate the genera so we don't get a lot of lines separating all the SVs
plot.gen <- aggregate_taxa(phylo.coreW_25, "Genus")

prevalences <- seq(.05, 1, .05)
detections <- round(10^seq(log10(1e-4), log10(.2), length = 10), 3)

p1 <- plot_core(plot.gen,
```

```r
          plot.type = "heatmap",
          prevalences = prevalences,
          detections = detections,
          min.prevalence = 1/10000) +
  xlab("Detection Threshold (Relative Abundance (%))") +
  ylab("Bacterial SVs") +
  theme_minimal() + scale_fill_viridis() +
  ggtitle("Core SVs, All Patients")

atb_plot.gen <- aggregate_taxa(atb_phylo.coreW_25, "Genus")

p2 <- plot_core(atb_plot.gen,
          plot.type = "heatmap",
          prevalences = prevalences,
          detections = detections,
          min.prevalence = 1/10000) +
  xlab("Detection Threshold (Relative Abundance (%))") +
  ylab("Bacterial SVs") +
  theme_minimal() + scale_fill_viridis() +
  ggtitle("Core SVs, Antibiotics")

no_atb_plot.gen <- aggregate_taxa(no_atb_phylo.coreW_25, "Genus")

p3 <- plot_core(no_atb_plot.gen,
          plot.type = "heatmap",
          prevalences = prevalences,
          detections = detections, min.prevalence = 1/10000) +
  xlab("Detection Threshold (Relative Abundance (%))") +
  ylab("Bacterial SVs") +
  theme_minimal() + scale_fill_viridis() +
  ggtitle("Core SVs, No Antibiotics")

# Put the panels together in to one figure
ggarrange(plotlist = list(p1, p2, p3),
          labels = c("A", "B", "C"),
          common.legend = TRUE,
          nrow = 1,
          ncol = 3,
          legend = "bottom") %>%
  annotate_figure(top = text_grob("Core SVs; Frequency >1/10000 and Prevalence > 0.25", size = 16))
```

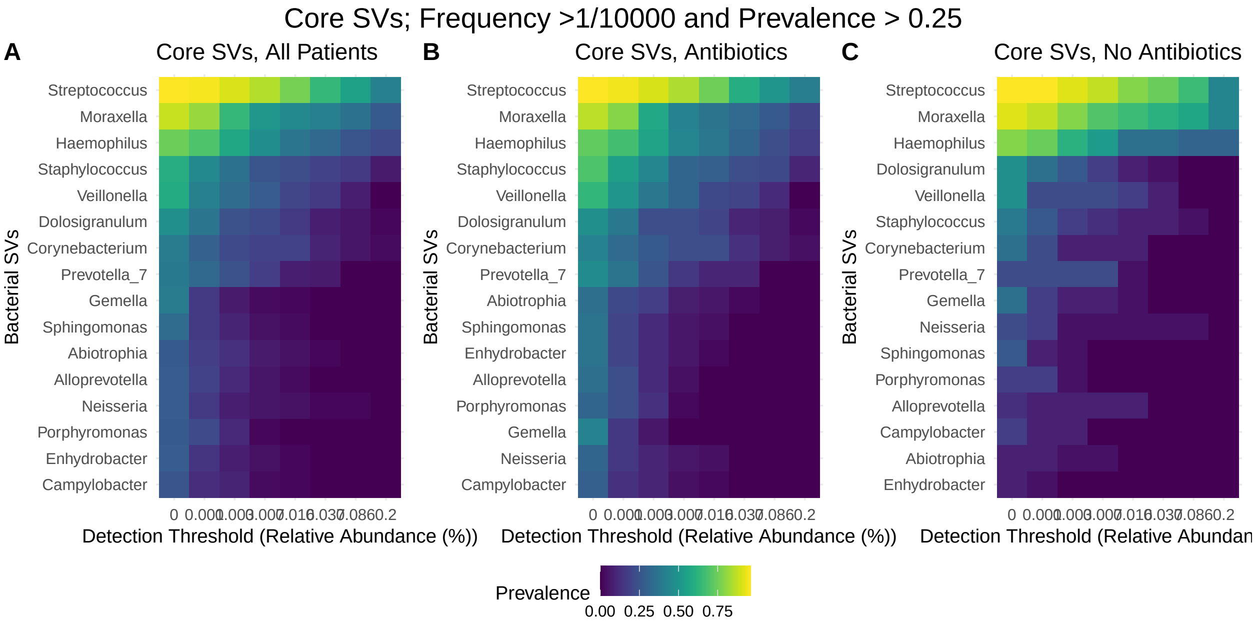

Figure 8. Core SVs, as defined by >1/10,000 frequency and >0.25 prevalence, then subset by treatment group for panels B and C. Streptococcus is more abundant in the group without antibiotic treatment. Lab negative controls are not represented as they were removed at this point in the workflow, that is, after decontamination and rarefaction.

```r
# Fig 9, core SVs

# _35 suffix means the data were generated with a 35/100 frequency cutoff
atb <- pivot_longer(as.data.frame(atb_phylo.coreW_35@otu_table),
                    cols = 1:ncol(as.data.frame(atb_phylo.coreW_35@otu_table)),
                    names_to = "SV",
                    values_to = "Abundance")
atb$Group <- "Antibiotics"


noatb <- pivot_longer(as.data.frame(no_atb_phylo.coreW_35@otu_table),
                      cols = 1:ncol(as.data.frame(no_atb_phylo.coreW_35@otu_table)),
                      names_to = "SV",
                      values_to = "Abundance")
noatb$Group <- "No antibiotics"

dat <- rbind(atb, noatb)

temp <- as.data.frame(atb_phylo.coreW_35@tax_table)
temp$SV <- row.names(temp)
dat <- left_join(dat, temp, by = "SV")

dat$wilcox <- NA
for (i in 1:nrow(dat)) {
  if (dat$SV[i] %in% core_SVs_sig_diff_on_wilcox_test) {
    dat$wilcox[i] <- 4
  }
}

dat$special <- NA
```

```r
# Streptococcus is the relevant pathologic bacteria with this disease
for (i in 1:nrow(dat)) {
  if (dat$Genus[i] == "Streptococcus") {
    dat$Genus[i] <- "Streptococcus +"
    dat$special[i] <- 3}
}

# Can edit where the stars go
for (i in 1:nrow(dat)) {
  if (!is.na(dat$special[i])) {
    dat$special[i] <- 2
  }
}
for (i in 1:nrow(dat)) {
  if (!is.na(dat$wilcox[i])) {
    dat$wilcox[i] <- 4.5
  }
}

dat %>% ggplot() +
  geom_boxplot(aes(x = SV,
                   y = Abundance,
                   color = Genus),
               outlier.shape = NA) +
  geom_point(aes(x = SV,
                 y = Abundance,
                 color = Genus),
             size = 1,
             position = position_jitter(width = 0.2),
             show.legend = FALSE) +
  geom_point(
    aes(x = SV,
        y = special),
    shape = "+",
    size = 5,
    color = "red",
    show.legend = FALSE) +
  geom_point(
    aes(x = SV, y = wilcox),
    shape = "*",
    size = 6,
    color = "red",
    show.legend = FALSE) +
  scale_y_continuous(trans = "log10", "Abundance, log10 scale", sec.axis = sec_axis(~ . , name = "Treat
  xlab("SV; Grouped by Phylum") +
  facet_grid(rows = vars(Group), cols = vars(Phylum), space = "free", scales = "free") +
  theme_bw() + # this puts the facet names in nice boxes
  theme(axis.text.x = element_blank(),
        legend.position = "bottom",
        legend.text = element_text(face = "italic"),
        panel.background = element_rect(fill = "gray84", color = "black")
  )  +
  ggtitle("Core SVs: >1/10,000 Frequency and >0.35 Prevalence")
```

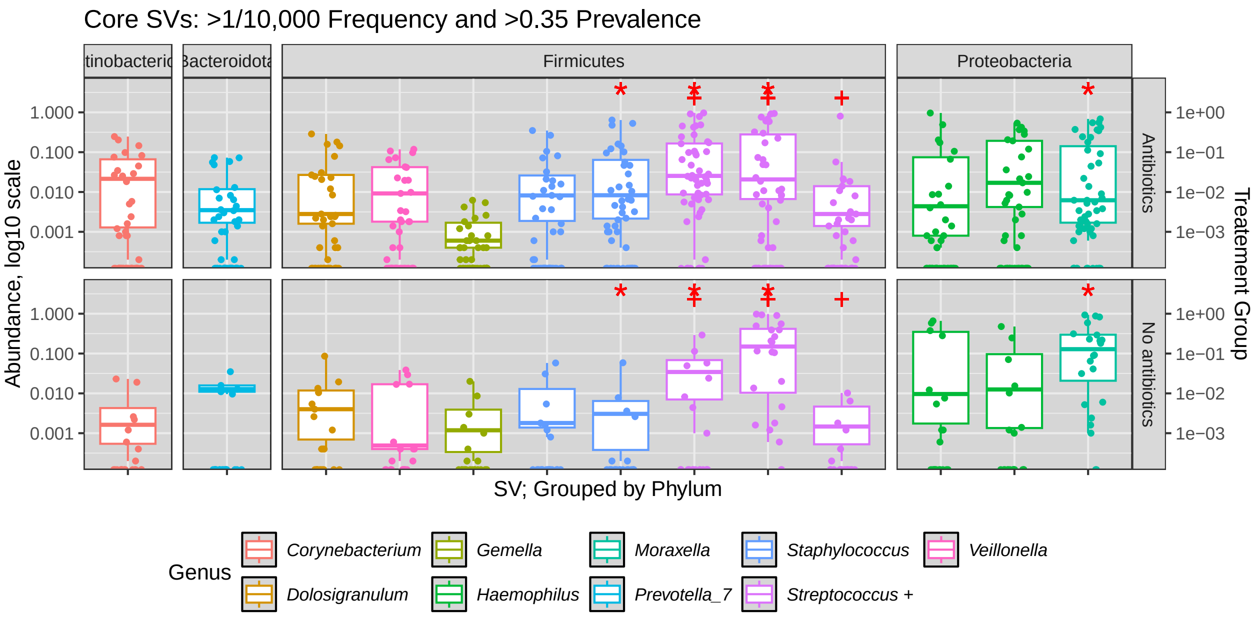

Figure 9. Core SVs, as defined as >10,000 frequency and >0.35 prevalence. Each box represents one SV. Since many SVs were not able to be assigned to the species level, each unique SV may or may not represent a different species. Asterisks indicate p < 0.05. Staphylococcus (presumed to be the cause of disease and the target of the antibiotics) SVs are highlighted with a red +.

```r
# Figure 10, significant SVs (by Wilcox test)

sig_SVs_adj <- wilcox_sig_SVs
for (i in 1:nrow(sig_SVs_adj)) {
  if (sig_SVs_adj$Abundance[i] == 0) {
    sig_SVs_adj$Abundance[i] <- 0.0001
  }
}


sig_SVs_adj$special <- NA
for (i in 1:nrow(sig_SVs_adj)) {
  if (sig_SVs_adj$Genus[i] == "Streptococcus") {
    sig_SVs_adj$Genus[i] <- "Streptococcus +"
    sig_SVs_adj$special[i] <- 1.1}
}

for (i in 1:nrow(sig_SVs_adj)) {
  if (!is.na(sig_SVs_adj$special[i])) {
    sig_SVs_adj$special[i] <- 2}
}

sig_SVs_adj %>% ggplot() +
  geom_boxplot(aes(x = SV,
                   y = Abundance,
                   color = Genus),
               outlier.shape = NA) +
  geom_point(aes(x = SV,
                 y = Abundance,
                 color = Genus),
             size = 1,
             position = position_jitter(width = 0.2),
             show.legend = FALSE) +
  geom_point(
    aes(x = SV,
        y = special),
    shape = "+",
    size = 6,
    color = "red",
    show.legend = FALSE) +
  scale_y_continuous(trans = "log10", "Abundance, log10 scale", sec.axis = sec_axis(~ . , name = "Treat
  xlab("SV; Grouped by Phylum") +
  facet_grid(rows = vars(Group), cols = vars(Phylum), space = "free", scales = "free") +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom",
        legend.text = element_text(face = "italic"),
        panel.background = element_rect(fill = "gray84", color = "black")
  )  +
  ggtitle("SVs with Significant Difference by Wilcoxon Test")
```

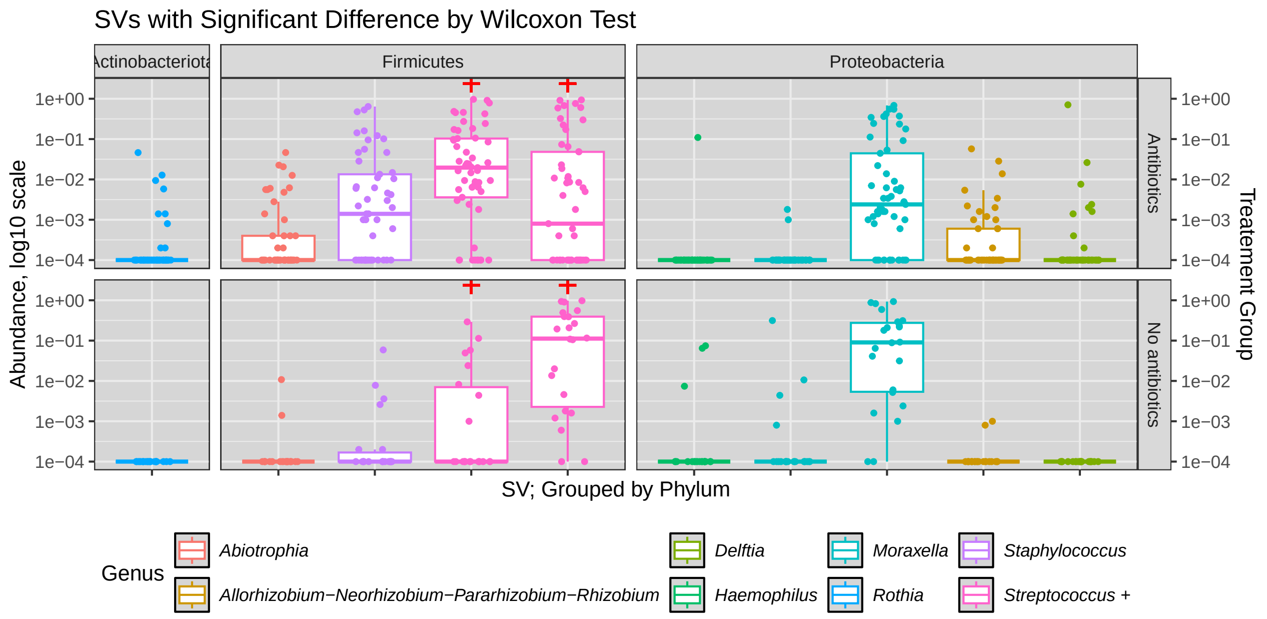

Figure 10. Wilcoxon tests between abundances in the antibiotic-treated group and the control group were performed for all SVs present. The 10 SVs shown here had p-values < 0.05. Since many SVs were not able to be assigned to the species level, each unique SV may or may not represent a different species. Staphylococcus (presumed to be the cause of disease and the target of the antibiotics) SVs are highlighted with a red +.

```
# Fig 11: NMDS

nmds_bray <- ordinate(phylo_decontam_rar,
                    method = "NMDS",
                    "bray",
                    binary = TRUE)
```

```
plot_ordination(phylo_decontam_rar,
                nmds_bray,
                type = "samples",
                color = "Group") +
  geom_point(aes(color = Group)) +
  theme(legend.position = "top",
        panel.border = element_rect(color = "gray", fill = NA, linewidth = 1)) +
  stat_ellipse(aes(group = Group, color = Group)) +
  labs(color = "Treatment Group",
       title = "NMDS Ordination",
       subtitle = "After Cleaning and Rarefaction; Bray-Curtis Distance")
```

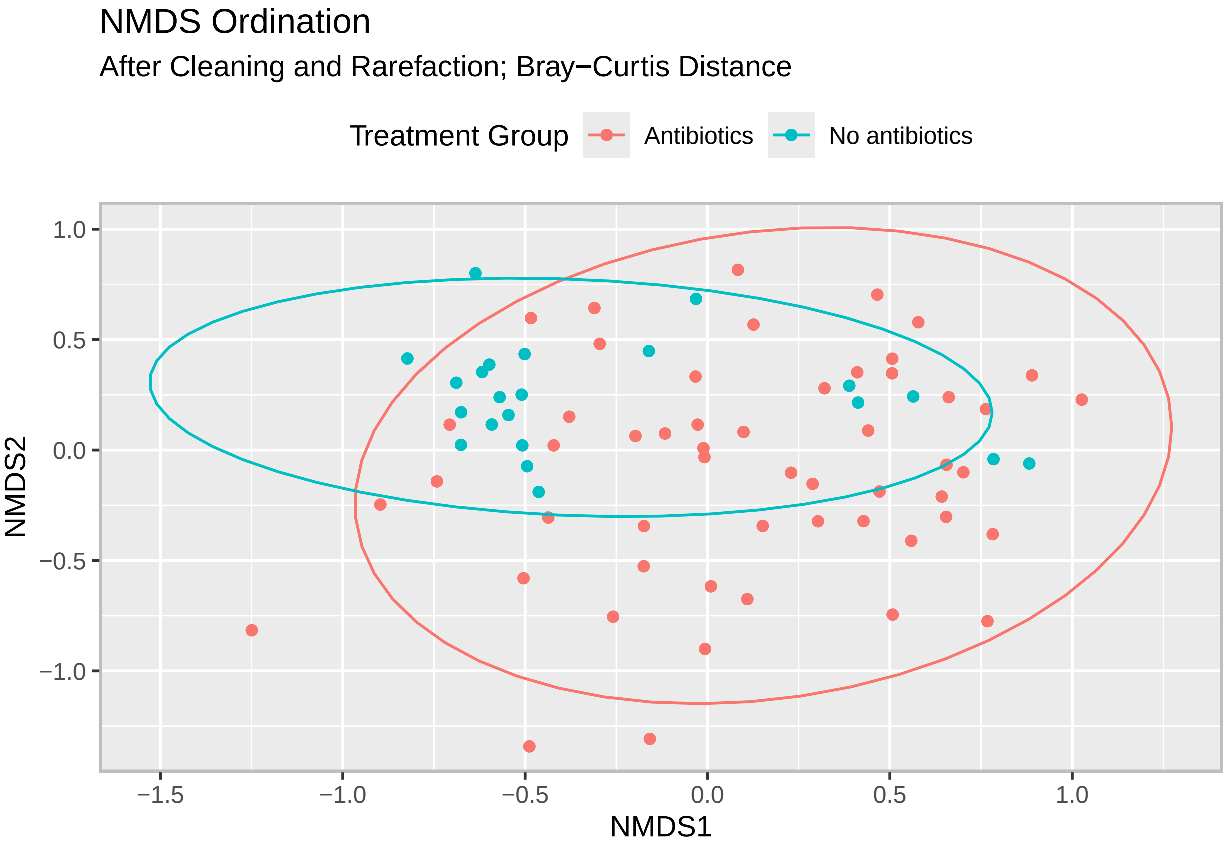

Figure 11. Nonmetric multidimensional plot (NMDS).

```
# Fig 12: DESeq results

ggplot(sigtab, aes(y = Genus, x = log2FoldChange, color = Phylum)) +
  geom_vline(xintercept = 0.0, color = "gray", linewidth = 0.5) +
  geom_point(aes(size = baseMean), position = position_jitter()) +
  scale_size_continuous(range = c(3, 8)) +
  theme(axis.text.x = element_text(hjust = 0, vjust = 0.5, size = 10),
        axis.text.y = element_text(size = 11),
        legend.title = element_text(size = 12),
        legend.text = element_text(size = 10),
        panel.border = element_rect(color = "gray", fill = NA, linewidth = 1)) +
  xlab("log2 Fold Change") +
  labs(size = "Mean Sequence Abundance",
       title = "Fold Change of Read Abundance",
       subtitle = "Antibiotic-treated Compared to Control Group")
```

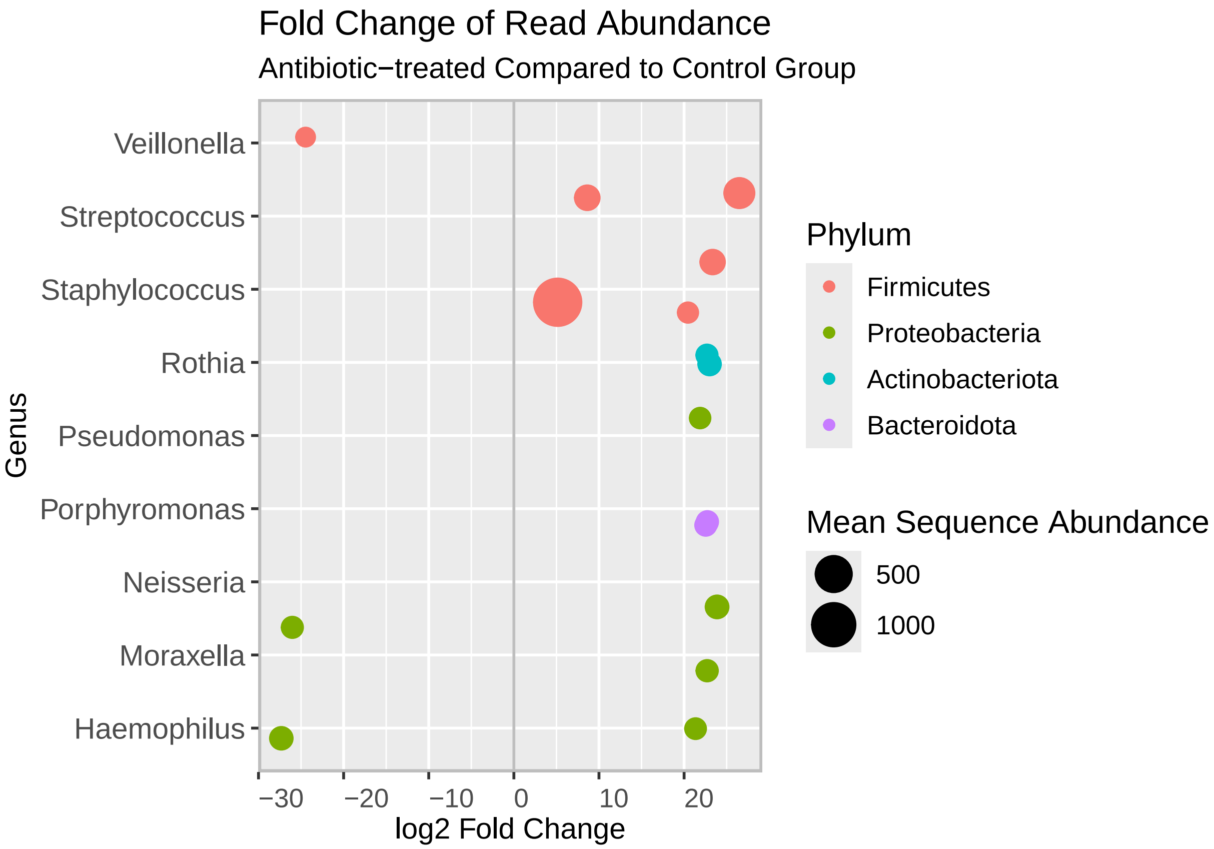

Figure 12. DESeq2 results. The majority (14 of 17) of taxa identified as significant are of increased abundance in the antibiotic-treated group.