Staphylome - tuf gene sequence analysis

Anna Ingham, Statens Serum Institut, Copenhagen

August 2020

```
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=55), tidy=TRUE) #to ensure line breaks in pdf output
```

Load packages

Read phyloseq objects with sample and mock data

```
ps <- readRDS("phyloseq_objects_for_publication/phy_obj_tuf.RData")
ps
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 523 taxa and 361 samples ]
## sample_data() Sample Data:       [ 361 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 523 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 523 reference sequences ]
```

```
ps_mocks <- readRDS("phyloseq_objects_for_publication/phy_obj_tuf_mocks.RData")
ps_mocks
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 127 taxa and 18 samples ]
## tax_table()   Taxonomy Table:    [ 127 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 127 reference sequences ]
```

Total average read count per sample (MAPPED ONLY):

```
psn <- prune_samples(sample_data(ps)$Sample_type == "Nose",
    ps)
print("Nose")
```

```
## [1] "Nose"
```

```
summary(sample_sums(psn))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      38   20163   37022   38416   51048  141715
```

```
psg <- prune_samples(sample_data(ps)$Sample_type == "Groin",
    ps)
print("Groin")
```

```
## [1] "Groin"
```

```
summary(sample_sums(psg))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      30   11822   28529   30820   44944   89572
```

```
psg <- prune_samples(sample_data(ps)$Sample_type == "Operation_site",
    ps)
print("OP site")
```

```
## [1] "OP site"
```

```
summary(sample_sums(psg))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      34    1424   13469   20328   31988   78206
```

Barplots of the mock samples

Read count per mock sample

```
sample_sums(ps_mocks)
```

```
##       pos_dx10_a_tuf_4_S94      pos_dx10_a_tuf_6_S114
##                      91436                      98216
##      pos_dx10_b_tuf_5_S104     pos_dx100_a_tuf_7_S124
##                      63118                     211154
##     pos_dx100_a_tuf_9_S144     pos_dx100_b_tuf_8_S134
##                      98695                      84205
##   pos_dx1000_a_tuf_10_S154    pos_dx1000_a_tuf_12_S95
##                      89679                      76330
##   pos_dx1000_b_tuf_11_S164  pos_dx10000_a_tuf_13_S105
##                      41970                       2874
##  pos_dx10000_a_tuf_15_S125  pos_dx10000_b_tuf_14_S115
##                      34737                      30868
## pos_dx100000_a_tuf_16_S135 pos_dx100000_a_tuf_18_S155
##                       9067                       4164
## pos_dx100000_b_tuf_17_S145        pos_x1_a_tuf_1_S143
##                      51069                      53415
##        pos_x1_a_tuf_3_S163        pos_x1_b_tuf_2_S153
##                      51668                      45689
```

```
summary(sample_sums(ps_mocks))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2874   36545   52542   63242   88310  211154
```

Minimum sample size is >2000, so no need exclude mock samples below cutoff

Agglomerate on Species level

```
rank_names(ps_mocks)
```

```
## [1] "Genus_Species" "Seq_number"
```

```
ps_mocks_gs <- tax_glom(ps_mocks, taxrank = "Genus_Species")
ps_mocks_gs <- prune_taxa(taxa_sums(ps_mocks_gs) != 0, ps_mocks_gs)
ps_mocks_gs
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 13 taxa and 18 samples ]
## tax_table()   Taxonomy Table:    [ 13 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 13 reference sequences ]
```

```
taxa_sums(ps_mocks_gs)
```

```
##   ASV1   ASV2   ASV3   ASV4   ASV7   ASV9  ASV10  ASV11  ASV15  ASV26  ASV60
## 197089 162784 259222 220306      7     80 151353 147376    120      5      2
##  ASV80 ASV674
##      9      1
```

```
sample_sums(ps_mocks_gs)
```

```
##       pos_dx10_a_tuf_4_S94      pos_dx10_a_tuf_6_S114
##                      91436                      98216
##      pos_dx10_b_tuf_5_S104     pos_dx100_a_tuf_7_S124
##                      63118                     211154
##     pos_dx100_a_tuf_9_S144     pos_dx100_b_tuf_8_S134
##                      98695                      84205
##   pos_dx1000_a_tuf_10_S154    pos_dx1000_a_tuf_12_S95
##                      89679                      76330
##   pos_dx1000_b_tuf_11_S164  pos_dx10000_a_tuf_13_S105
##                      41970                       2874
##  pos_dx10000_a_tuf_15_S125  pos_dx10000_b_tuf_14_S115
##                      34737                      30868
## pos_dx100000_a_tuf_16_S135 pos_dx100000_a_tuf_18_S155
##                       9067                       4164
## pos_dx100000_b_tuf_17_S145        pos_x1_a_tuf_1_S143
##                      51069                      53415
##        pos_x1_a_tuf_3_S163        pos_x1_b_tuf_2_S153
##                      51668                      45689
```

Convert to relative abundance

```
ps_mocks_gs_rel = transform_sample_counts(ps_mocks_gs, function(x) x/sum(x))
summary(sample_sums(ps_mocks_gs_rel))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       1       1       1       1
```

Subset top genera in mocks

```
SpecMock = names(sort(taxa_sums(ps_mocks_gs_rel), TRUE)[1:6])
```

to data frame

```
p_df_mo <- psmelt(ps_mocks_gs_rel)
p_df_mo$Genus_Species <- as.character(p_df_mo$Genus_Species)
p_df_mo$Genus_Species[!(p_df_mo$OTU %in% SpecMock)] <- "x_Other"
```

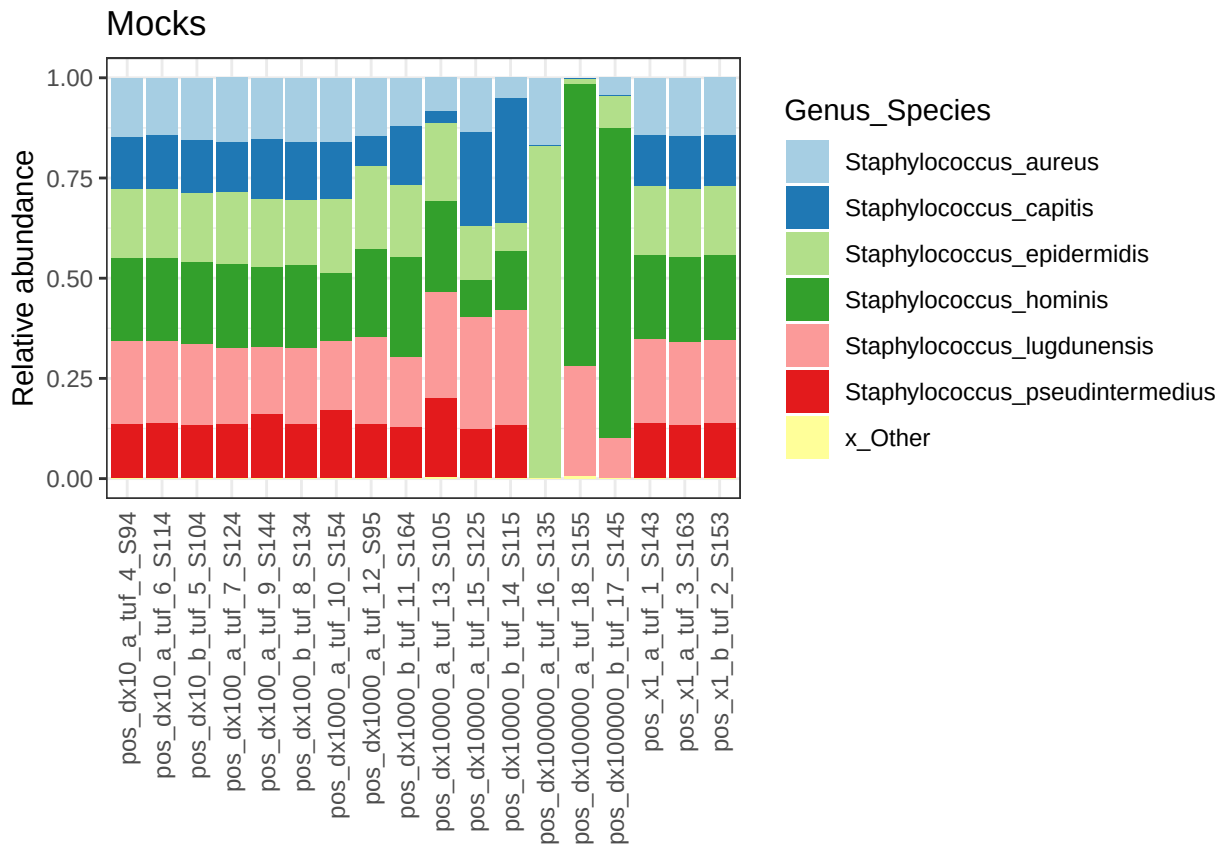
```
p_df_mo <- p_df_mo %>% mutate(Dilution = ifelse(grepl("_x1_",
    Sample), "x1", ifelse(grepl("_dx10_", Sample), "x10",
    ifelse(grepl("_dx100_", Sample), "x100", ifelse(grepl("_dx1000_",
        Sample), "x1000", ifelse(grepl("_dx10000_", Sample),
        "x10000", "x100000"))))))
```

Define color code

```
staph_col <- c(Staphylococcus_aureus = "#A6CEE3", Staphylococcus_capitis = "#1F78B4",
    Staphylococcus_epidermidis = "#B2DF8A", Staphylococcus_hominis = "#33A02C",
    Staphylococcus_lugdunensis = "#FB9A99", Staphylococcus_warneri = "#FF7F00",
    Staphylococcus_simulans = "#FDBF6F", Staphylococcus_pseudintermedius = "#E31A1C",
    Staphylococcus_pasteuri = "#CAB2D6", Staphylococcus_haemolyticus = "#6A3D9A",
    Staphylococcus_saprophyticus = "#B15928", x_Other = "#FFFF99",
    Staphylococcus_caprae = "#E7298A", Staphylococcus_sciuri = "#A6761D")
```

Barplots of relative abundance

```
pmo <- ggplot(p_df_mo, aes(x = Sample, y = Abundance, fill = Genus_Species)) +
    geom_bar(stat = "identity", width = 0.9) + scale_fill_manual(values = staph_col) +
    theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 90, hjust = 1,
            vjust = 0.5), strip.background = element_rect(fill = "white"),
        strip.text.y = element_text(size = 10, face = "bold")) +
    ylab("Relative abundance") + ggtitle("Mocks")

pmo
```



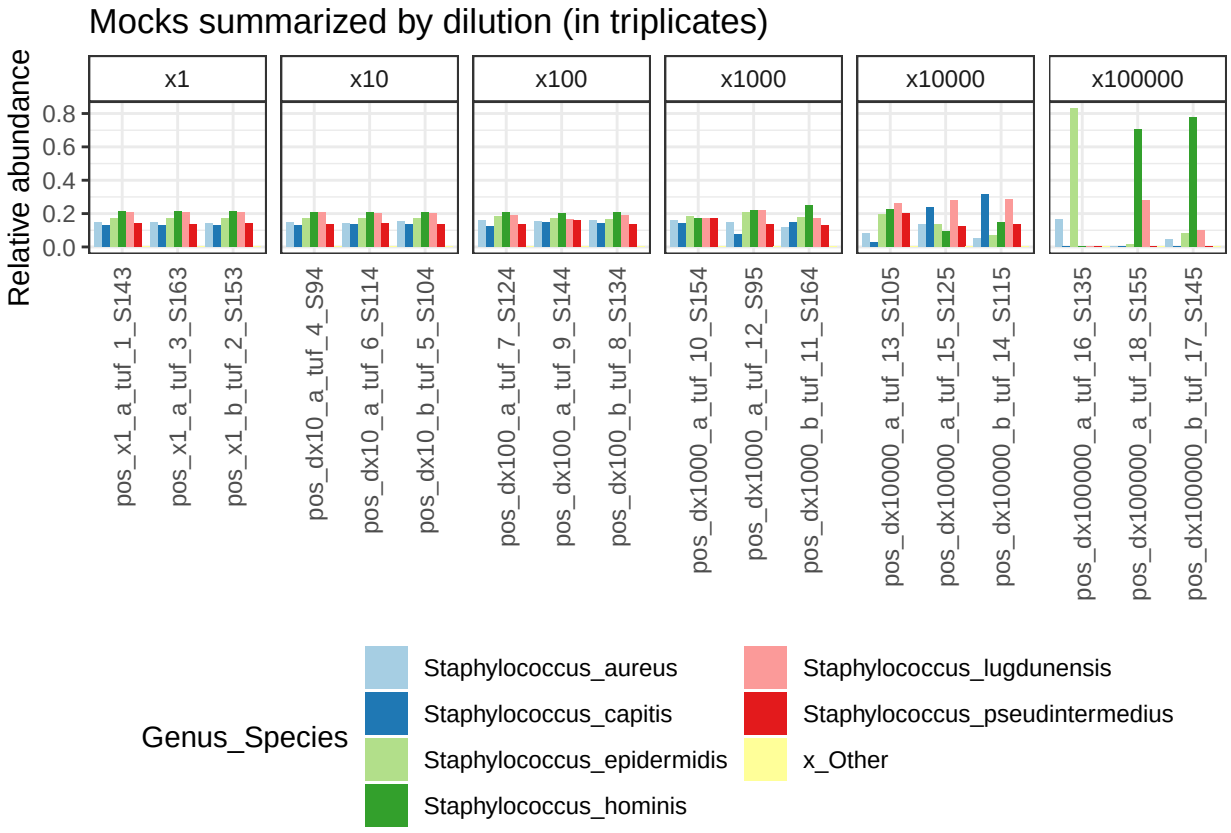
Other needs to be summed before dodging

```
p_df_mo_S <- p_df_mo %>% group_by(Dilution, Sample, Genus_Species) %>%
    summarise(summed_abund = sum(Abundance))
```

```
## `summarise()` regrouping output by 'Dilution', 'Sample' (override with `.groups` argument)
```

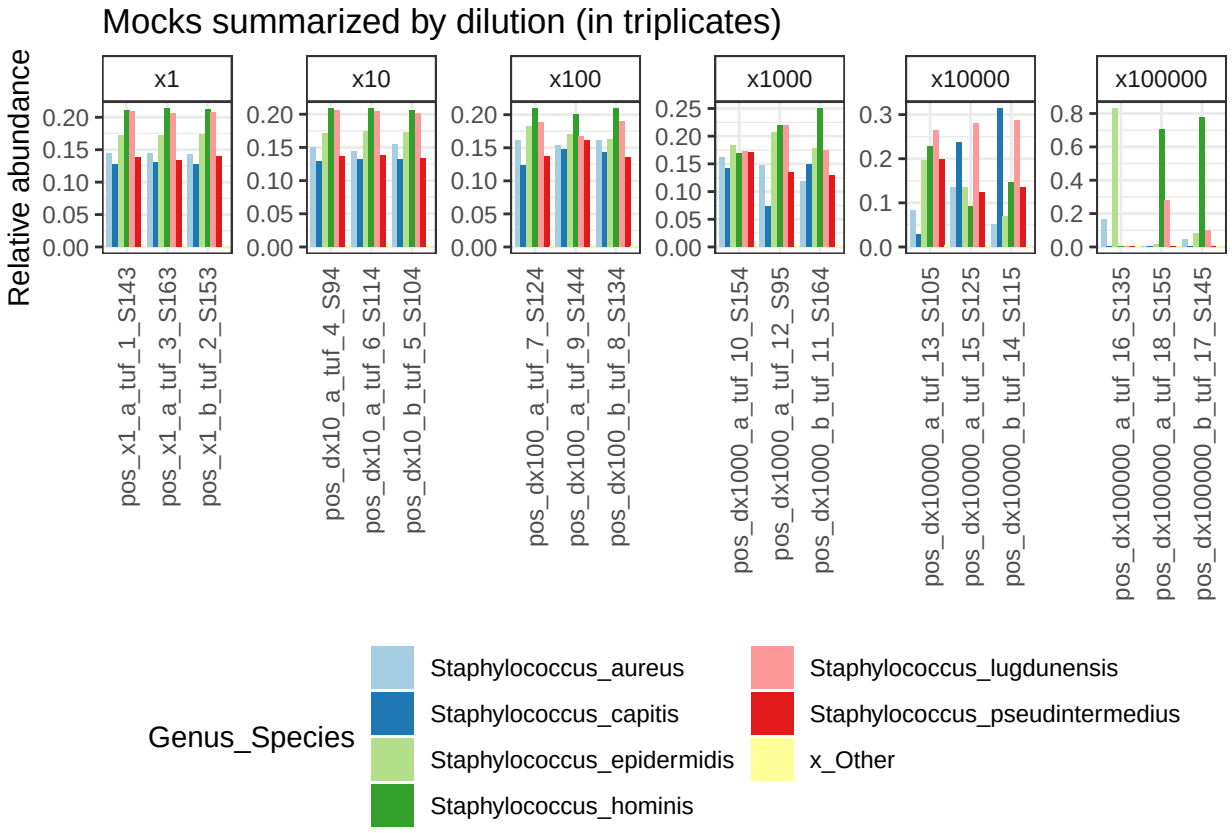
With the same y-axis

```
ggplot(p_df_mo_S, aes(x = Sample, y = summed_abund, fill = Genus_Species)) +
    geom_bar(stat = "identity", width = 1, position = "dodge") +
    scale_fill_manual(values = staph_col) + theme(axis.title.x = element_blank(),
    axis.ticks.x = element_blank(), axis.text.x = element_text(angle = 90,
        hjust = 1, vjust = 0.5), strip.background = element_rect(fill = "white"),
    strip.text.y = element_text(size = 10, face = "bold"),
    legend.position = "bottom") + ylab("Relative abundance") +
    ggtitle("Mocks summarized by dilution (in triplicates)") +
    facet_wrap(Dilution ~ ., scales = "free_x", nrow = 1) +
    guides(fill = guide_legend(ncol = 2))
```



With individaul y-axes

```
ggplot(p_df_mo_S, aes(x = Sample, y = summed_abund, fill = Genus_Species,
    group = Genus_Species)) + geom_bar(stat = "identity",
    width = 1, position = "dodge") + scale_fill_manual(values = staph_col) +
    theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 90, hjust = 1,
            vjust = 0.5), strip.background = element_rect(fill = "white"),
        strip.text.y = element_text(size = 10, face = "bold"),
        legend.position = "bottom") + ylab("Relative abundance") +
    ggtitle("Mocks summarized by dilution (in triplicates)") +
    facet_wrap(Dilution ~ ., scales = "free", nrow = 1) +
    guides(fill = guide_legend(ncol = 2))
```



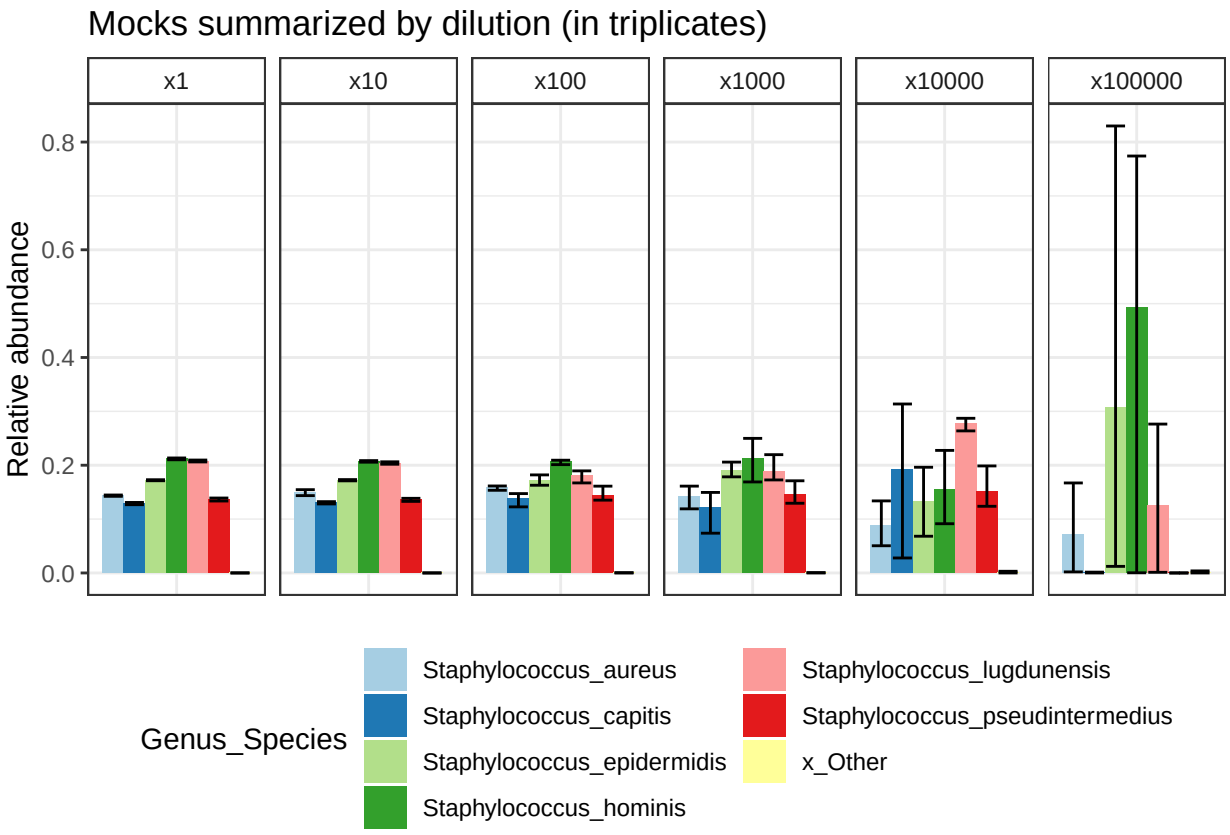
Summarize relative abundance by dilution (average of triplicates) and plot with error bars (where max = abundance in the triplacte with the highest abundance for a given species, and min = abundance in the triplacte with the lowest abundance for a given species):

```
p_df_mo1 <- p_df_mo %>% group_by(Genus_Species, Dilution) %>%
    summarise(mean = mean(Abundance, na.rm = TRUE), min = min(Abundance,
        na.rm = TRUE), max = max(Abundance, na.rm = TRUE))
```

```
## `summarise()` regrouping output by 'Genus_Species' (override with `.groups` argument)
```

With the same y-axis

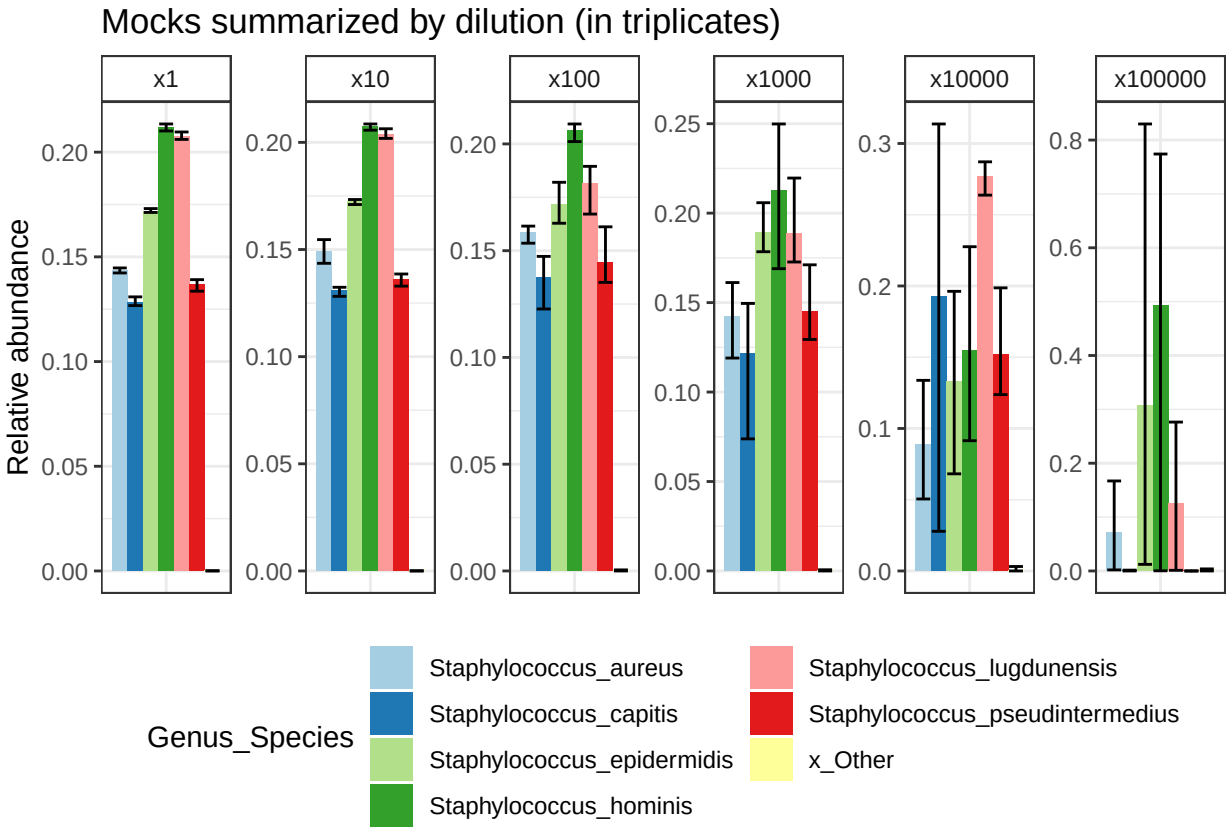
```
mp <- ggplot(p_df_mo1, aes(x = Dilution, y = mean, fill = Genus_Species)) +
    geom_bar(stat = "identity", width = 1, position = "dodge") +
    scale_fill_manual(values = staph_col) + theme(axis.title.x = element_blank(),
    axis.ticks.x = element_blank(), axis.text.x = element_blank(),
    strip.background = element_rect(fill = "white"), strip.text.y = element_text(size = 10,
        face = "bold"), legend.position = "bottom") + ylab("Relative abundance") +
    ggtitle("Mocks summarized by dilution (in triplicates)") +
    geom_errorbar(data = p_df_mo1, aes(x = Dilution, ymin = min,
        ymax = max), position = position_dodge(1)) + facet_wrap(Dilution ~
    ., scales = "free_x", nrow = 1)
mp + guides(fill = guide_legend(ncol = 2))
```




```
ggsave(filename = "plots/mock_bars.pdf", plot = mp, device = cairo_pdf,
    width = 297, height = 105, units = "mm")
```

With individaul y-axes

```
ggplot(p_df_mo1, aes(x = Dilution, y = mean, fill = Genus_Species)) +
    geom_bar(stat = "identity", width = 1, position = "dodge") +
    scale_fill_manual(values = staph_col) + theme(axis.title.x = element_blank(),
    axis.ticks.x = element_blank(), axis.text.x = element_blank(),
    strip.background = element_rect(fill = "white"), strip.text.y = element_text(size = 10,
        face = "bold"), legend.position = "bottom") + ylab("Relative abundance") +
    ggtitle("Mocks summarized by dilution (in triplicates)") +
    geom_errorbar(data = p_df_mo1, aes(x = Dilution, ymin = min,
        ymax = max), position = position_dodge(1)) + facet_wrap(Dilution ~
    ., scales = "free", nrow = 1) + guides(fill = guide_legend(ncol = 2))
```



Split by body site

Nose

```
ps_samp_m_nose <- prune_samples(sample_data(ps)$Sample_type ==
    "Nose", ps)
ps_samp_m_nose <- prune_taxa(taxa_sums(ps_samp_m_nose) !=
    0, ps_samp_m_nose)
ps_samp_m_nose
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 324 taxa and 161 samples ]
## sample_data() Sample Data:       [ 161 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 324 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 324 reference sequences ]
```

```
sample_data(ps_samp_m_nose)$Sample_type
```

```
##   [1] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
##  [11] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
##  [21] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
##  [31] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
```

```
##  [41] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
##  [51] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
##  [61] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
##  [71] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
##  [81] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
##  [91] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
## [101] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
## [111] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
## [121] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
## [131] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
## [141] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
## [151] "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose" "Nose"
## [161] "Nose"
```

```
length(unique(sample_data(ps_samp_m_nose)$Patient_ID))
```

```
## [1] 82
```

Which patients have both, a before and an after sample from the nose

```
table(sample_data(ps_samp_m_nose)$Patient_ID)
```

```
##
## P01 P02 P03 P04 P05 P06 P07 P08 P09 P10 P11 P12 P13 P14 P15 P16 P17 P18 P19 P20
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P21 P22 P23 P24 P25 P26 P27 P28 P29 P30 P31 P33 P34 P35 P36 P37 P38 P39 P40 P41
##   2   2   1   2   2   2   2   2   2   2   2   1   2   2   2   2   2   2   2   2
## P42 P43 P44 P45 P46 P47 P48 P49 P50 P51 P52 P53 P54 P55 P56 P57 P58 P59 P60 P61
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P62 P63 P64 P65 P66 P67 P68 P69 P70 P71 P72 P73 P74 P75 P76 P77 P78 P79 P80 P81
##   2   2   2   2   2   2   2   2   2   2   2   2   1   2   2   2   2   2   2   2
## P82 P83
##   2   2
```

exclude patients with only one time point

```
ps_samp_m_nose <- prune_samples(!sample_data(ps_samp_m_nose)$Patient_ID %in%
    c("P23", "P33", "P74"), ps_samp_m_nose)
ps_samp_m_nose
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 324 taxa and 158 samples ]
## sample_data() Sample Data:       [ 158 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 324 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 324 reference sequences ]
```

```
table(sample_data(ps_samp_m_nose)$Patient_ID)
```

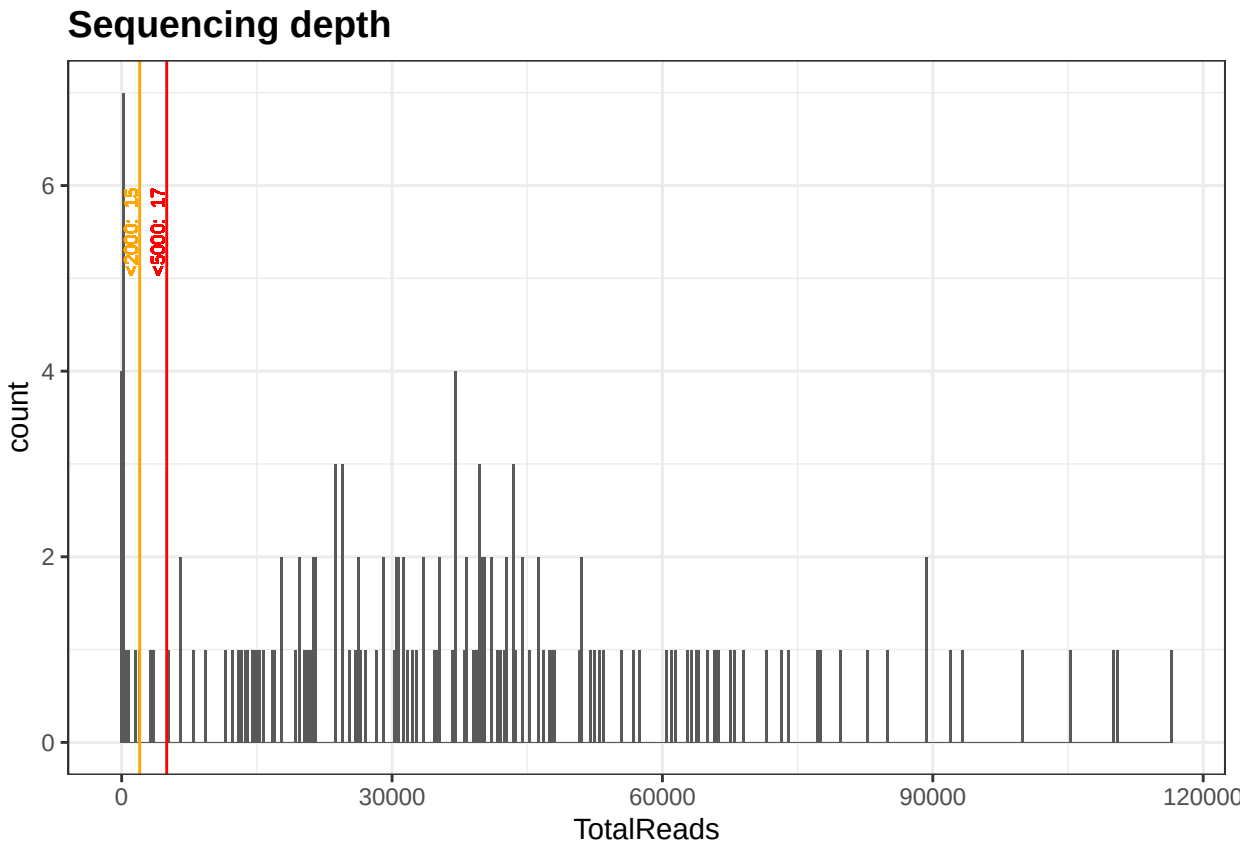
```
##
## P01 P02 P03 P04 P05 P06 P07 P08 P09 P10 P11 P12 P13 P14 P15 P16 P17 P18 P19 P20
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P21 P22 P24 P25 P26 P27 P28 P29 P30 P31 P34 P35 P36 P37 P38 P39 P40 P41 P42 P43
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P44 P45 P46 P47 P48 P49 P50 P51 P52 P53 P54 P55 P56 P57 P58 P59 P60 P61 P62 P63
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P64 P65 P66 P67 P68 P69 P70 P71 P72 P73 P75 P76 P77 P78 P79 P80 P81 P82 P83
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
```

```
length(unique(sample_data(ps_samp_m_nose)$Patient_ID))
```

```
## [1] 79
```

Seq depth

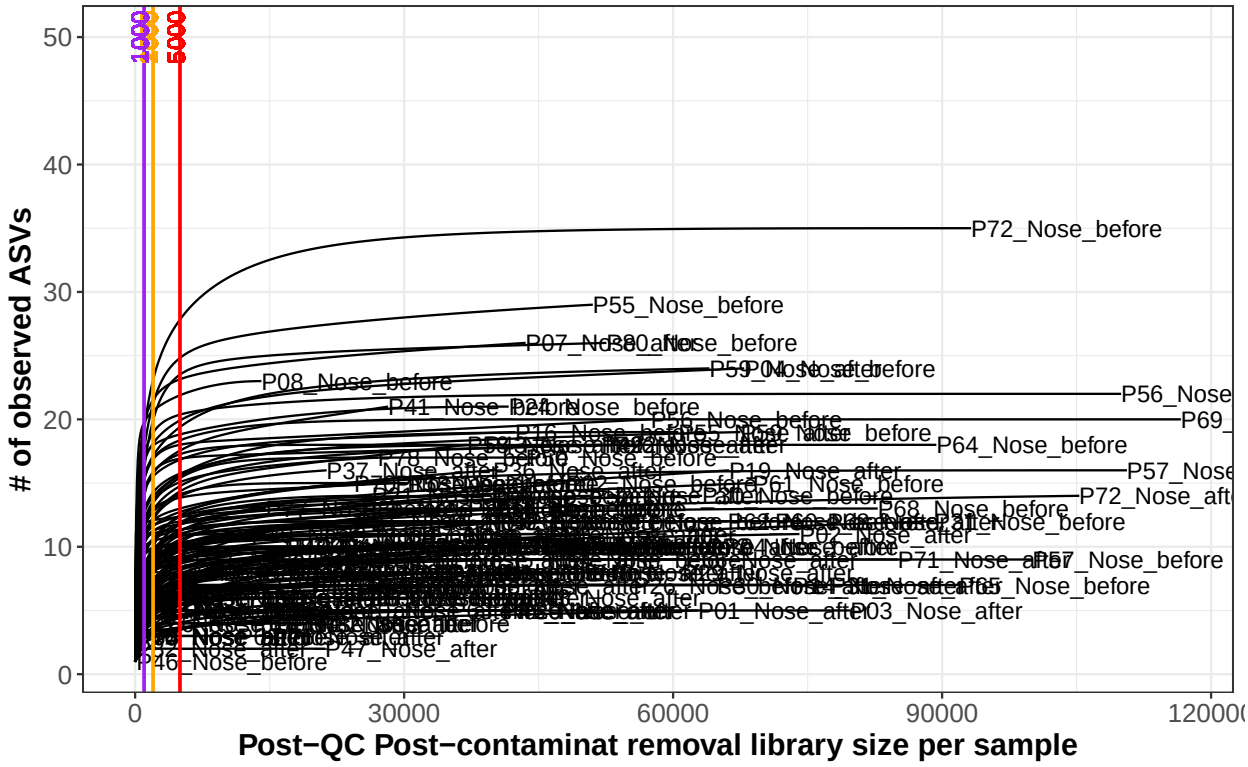
```
sdt = data.table::data.table(as(sample_data(ps_samp_m_nose),
    "data.frame"), TotalReads = sample_sums(ps_samp_m_nose),
    keep.rownames = TRUE)
data.table::setnames(sdt, "rn", "SampleID")
pSeqDepth = ggplot(sdt, aes(TotalReads)) + geom_histogram(binwidth = 250) +
    geom_vline(xintercept = 5000, color = "red") + geom_vline(xintercept = 2000,
    color = "orange") + geom_text(aes(x = 1000, label = paste("<2000: ",
    nrow(sdt[sdt$TotalReads < 2000])), y = 5.5), colour = "orange",
    angle = 90, size = 2.5) + geom_text(aes(x = 4000, label = paste("<5000: ",
    nrow(sdt[sdt$TotalReads < 5000])), y = 5.5), colour = "red",
    angle = 90, size = 2.5) + ggtitle("Sequencing depth") +
    theme(plot.title = element_text(size = 14, face = "bold"))

pSeqDepth
```



Do the rarefaction curves justify that we remove samples with reads <2000?

Rarefaction curves

```
p2 <- p2 + theme(panel.background = element_blank(), axis.title.x = element_text(size = 14,
    face = "bold"), axis.title.y = element_text(size = 14,
    face = "bold"), axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12), legend.title = element_text(size = 16,
        face = "bold"), legend.text = element_text(size = 16),
    strip.text.x = element_text(angle = 0, face = "bold",
        size = 12), strip.background = element_rect(fill = "white")) +
    xlab("Post-QC Post-contaminat removal library size per sample") +
    ylab("# of observed ASVs") + geom_vline(xintercept = 5000,
    color = "red", size = 0.8) + geom_vline(xintercept = 2000,
    color = "orange", size = 0.8) + geom_vline(xintercept = 1000,
    color = "purple", size = 0.8) + geom_text(aes(x = 4550,
    label = "5000", y = 50), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 50), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 50), colour = "purple",
        angle = 90, size = 4)
p2
```

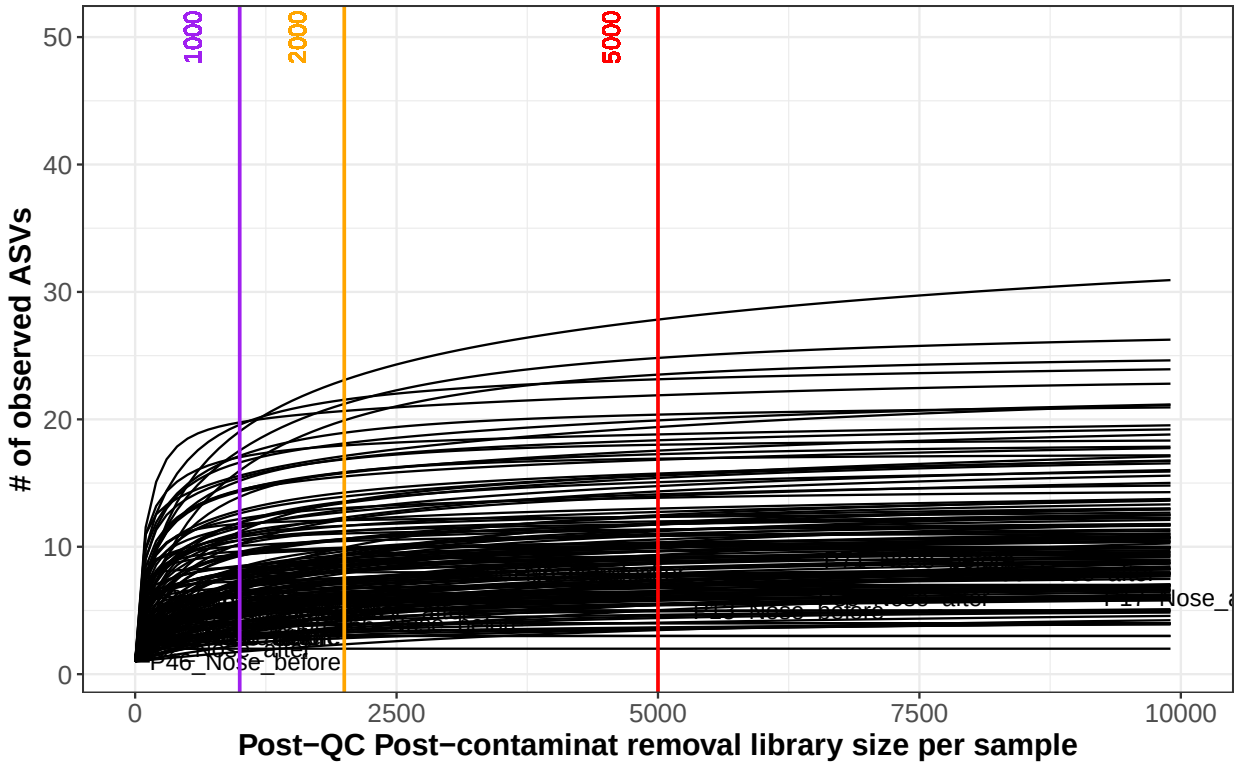


Zoom

```
p2 + xlim(0, 10000)
```

```
## Warning: Removed 136 rows containing missing values (geom_text).
```

```
## Warning: Removed 45637 row(s) containing missing values (geom_path).
```



How do the rarefaction curves look on species level?

Rarefaction curves

```
ps_samp_m_nose_gen <- tax_glom(ps_samp_m_nose, taxrank = "Genus_Species")
ps_samp_m_nose_gen
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 34 taxa and 158 samples ]
## sample_data() Sample Data:       [ 158 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 34 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 34 reference sequences ]
```

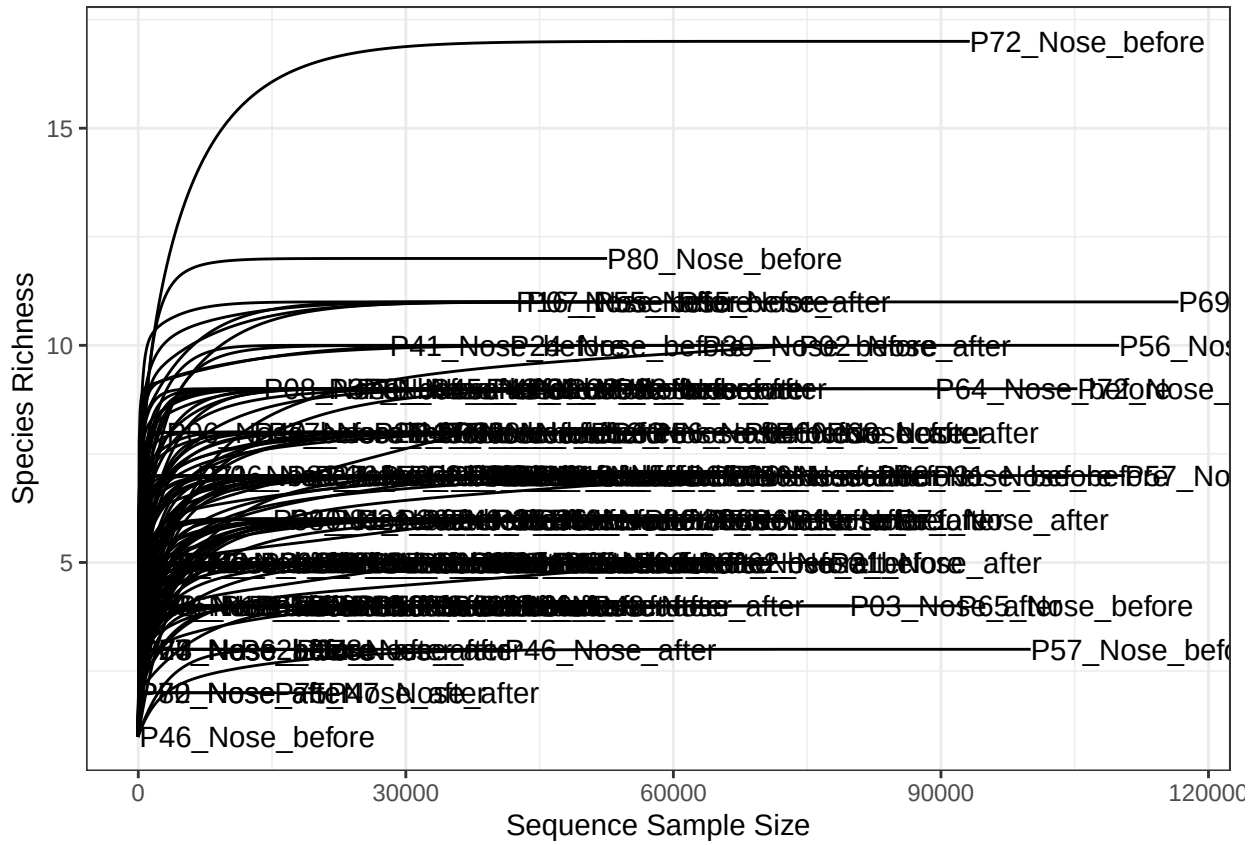
```
ps_samp_m_nose_gen <- prune_taxa(taxa_sums(ps_samp_m_nose_gen) !=
    0, ps_samp_m_nose_gen)
set.seed(123)
p3 <- ggrare(ps_samp_m_nose_gen, step = 100, se = FALSE,
    label = "Sample")
```

```
## rarefying sample P61_Nose_before
## rarefying sample P61_Nose_after
## rarefying sample P70_Nose_before
## rarefying sample P70_Nose_after
## rarefying sample P71_Nose_before
## rarefying sample P71_Nose_after
## rarefying sample P72_Nose_before
```

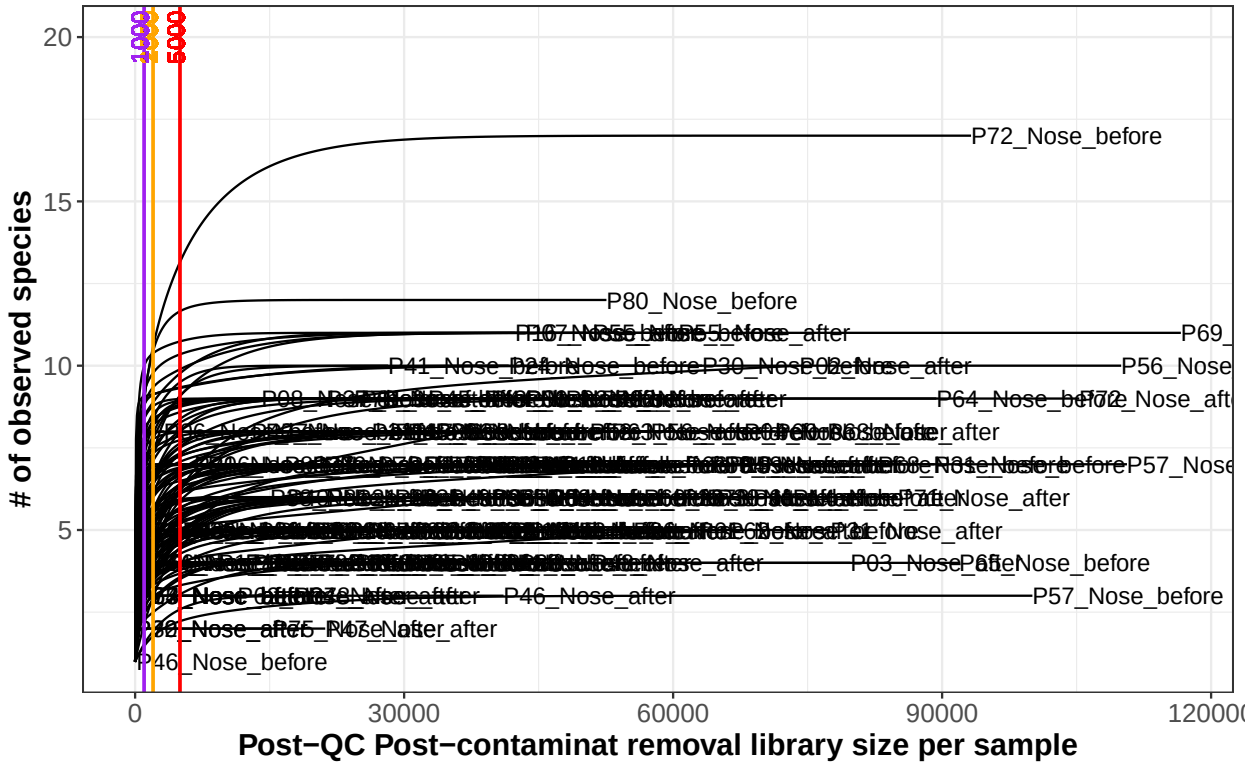
```
## rarefying sample P72_Nose_after
## rarefying sample P73_Nose_before
## rarefying sample P73_Nose_after
## rarefying sample P28_Nose_before
## rarefying sample P11_Nose_before
## rarefying sample P05_Nose_before
## rarefying sample P02_Nose_before
## rarefying sample P16_Nose_before
## rarefying sample P18_Nose_before
## rarefying sample P07_Nose_before
## rarefying sample P10_Nose_before
## rarefying sample P03_Nose_before
## rarefying sample P18_Nose_after
## rarefying sample P17_Nose_before
## rarefying sample P12_Nose_before
## rarefying sample P04_Nose_before
## rarefying sample P15_Nose_before
## rarefying sample P14_Nose_before
## rarefying sample P01_Nose_before
## rarefying sample P06_Nose_before
## rarefying sample P29_Nose_before
## rarefying sample P09_Nose_before
## rarefying sample P19_Nose_before
## rarefying sample P05_Nose_after
## rarefying sample P11_Nose_after
## rarefying sample P20_Nose_before
## rarefying sample P10_Nose_after
## rarefying sample P08_Nose_before
## rarefying sample P16_Nose_after
## rarefying sample P04_Nose_after
## rarefying sample P15_Nose_after
## rarefying sample P13_Nose_before
## rarefying sample P19_Nose_after
## rarefying sample P14_Nose_after
## rarefying sample P03_Nose_after
## rarefying sample P17_Nose_after
## rarefying sample P09_Nose_after
## rarefying sample P07_Nose_after
## rarefying sample P30_Nose_before
## rarefying sample P12_Nose_after
## rarefying sample P20_Nose_after
## rarefying sample P02_Nose_after
## rarefying sample P08_Nose_after
## rarefying sample P40_Nose_before
## rarefying sample P13_Nose_after
## rarefying sample P31_Nose_before
## rarefying sample P42_Nose_before
## rarefying sample P06_Nose_after
## rarefying sample P28_Nose_after
## rarefying sample P40_Nose_after
## rarefying sample P38_Nose_before
## rarefying sample P39_Nose_before
## rarefying sample P41_Nose_before
## rarefying sample P29_Nose_after
```

```
## rarefying sample P31_Nose_after
## rarefying sample P43_Nose_before
## rarefying sample P44_Nose_before
## rarefying sample P45_Nose_before
## rarefying sample P46_Nose_before
## rarefying sample P30_Nose_after
## rarefying sample P42_Nose_after
## rarefying sample P38_Nose_after
## rarefying sample P39_Nose_after
## rarefying sample P41_Nose_after
## rarefying sample P46_Nose_after
## rarefying sample P45_Nose_after
## rarefying sample P01_Nose_after
## rarefying sample P47_Nose_before
## rarefying sample P43_Nose_after
## rarefying sample P44_Nose_after
## rarefying sample P47_Nose_after
## rarefying sample P48_Nose_before
## rarefying sample P48_Nose_after
## rarefying sample P49_Nose_before
## rarefying sample P49_Nose_after
## rarefying sample P50_Nose_before
## rarefying sample P21_Nose_before
## rarefying sample P22_Nose_before
## rarefying sample P24_Nose_before
## rarefying sample P25_Nose_before
## rarefying sample P26_Nose_before
## rarefying sample P27_Nose_before
## rarefying sample P22_Nose_after
## rarefying sample P25_Nose_after
## rarefying sample P21_Nose_after
## rarefying sample P24_Nose_after
## rarefying sample P50_Nose_after
## rarefying sample P27_Nose_after
## rarefying sample P26_Nose_after
## rarefying sample P34_Nose_before
## rarefying sample P35_Nose_before
## rarefying sample P35_Nose_after
## rarefying sample P34_Nose_after
## rarefying sample P36_Nose_before
## rarefying sample P37_Nose_before
## rarefying sample P37_Nose_after
## rarefying sample P36_Nose_after
## rarefying sample P51_Nose_before
## rarefying sample P51_Nose_after
## rarefying sample P52_Nose_before
## rarefying sample P52_Nose_after
## rarefying sample P53_Nose_before
## rarefying sample P53_Nose_after
## rarefying sample P54_Nose_before
## rarefying sample P54_Nose_after
## rarefying sample P55_Nose_before
## rarefying sample P55_Nose_after
## rarefying sample P56_Nose_before
```

```
## rarefying sample P57_Nose_before
## rarefying sample P57_Nose_after
## rarefying sample P56_Nose_after
## rarefying sample P58_Nose_before
## rarefying sample P58_Nose_after
## rarefying sample P59_Nose_before
## rarefying sample P59_Nose_after
## rarefying sample P60_Nose_before
## rarefying sample P60_Nose_after
## rarefying sample P75_Nose_before
## rarefying sample P75_Nose_after
## rarefying sample P76_Nose_before
## rarefying sample P76_Nose_after
## rarefying sample P77_Nose_before
## rarefying sample P77_Nose_after
## rarefying sample P78_Nose_before
## rarefying sample P78_Nose_after
## rarefying sample P62_Nose_before
## rarefying sample P62_Nose_after
## rarefying sample P79_Nose_before
## rarefying sample P79_Nose_after
## rarefying sample P80_Nose_before
## rarefying sample P80_Nose_after
## rarefying sample P81_Nose_before
## rarefying sample P81_Nose_after
## rarefying sample P82_Nose_before
## rarefying sample P82_Nose_after
## rarefying sample P83_Nose_before
## rarefying sample P83_Nose_after
## rarefying sample P63_Nose_before
## rarefying sample P63_Nose_after
## rarefying sample P64_Nose_before
## rarefying sample P64_Nose_after
## rarefying sample P65_Nose_before
## rarefying sample P65_Nose_after
## rarefying sample P66_Nose_before
## rarefying sample P66_Nose_after
## rarefying sample P67_Nose_before
## rarefying sample P67_Nose_after
## rarefying sample P68_Nose_before
## rarefying sample P68_Nose_after
## rarefying sample P69_Nose_before
## rarefying sample P69_Nose_after
```




```
p3 <- p3 + theme(panel.background = element_blank(), axis.title.x = element_text(size = 14,
    face = "bold"), axis.title.y = element_text(size = 14,
    face = "bold"), axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12), legend.title = element_text(size = 16,
        face = "bold"), legend.text = element_text(size = 16),
    strip.text.x = element_text(angle = 0, face = "bold",
        size = 12), strip.background = element_rect(fill = "white")) +
    xlab("Post-QC Post-contaminat removal library size per sample") +
    ylab("# of observed species") + geom_vline(xintercept = 5000,
    color = "red", size = 0.8) + geom_vline(xintercept = 2000,
    color = "orange", size = 0.8) + geom_vline(xintercept = 1000,
    color = "purple", size = 0.8) + geom_text(aes(x = 4550,
    label = "5000", y = 20), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 20), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 20), colour = "purple",
        angle = 90, size = 4)
p3
```

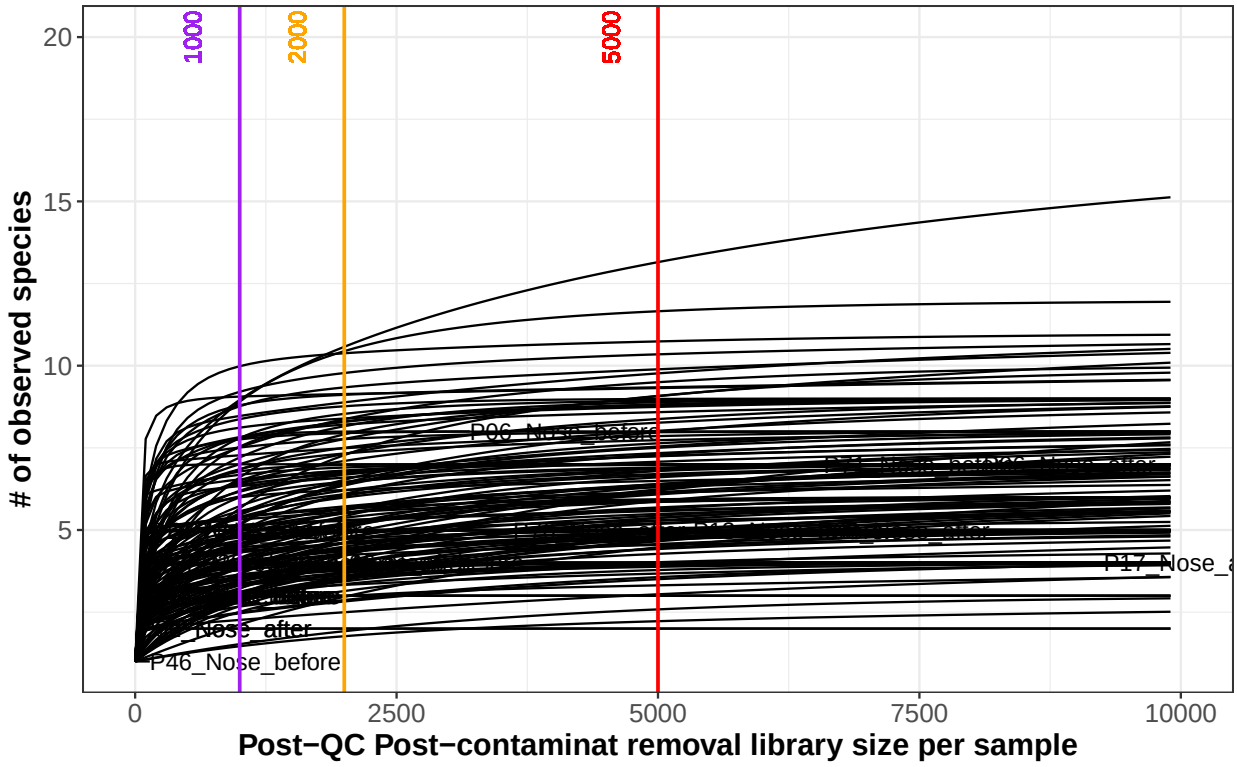


Zoom

```
p3 + xlim(0, 10000)
```

```
## Warning: Removed 136 rows containing missing values (geom_text).
```

```
## Warning: Removed 45637 row(s) containing missing values (geom_path).
```



Exclude samples with <2000 counts

```
summary(sample_sums(ps_samp_m_nose))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      38   19868   35968   37673   50860  116624
```

```
ps_samp_m_nose_tu <- prune_samples(!sample_sums(ps_samp_m_nose) <
    2000, ps_samp_m_nose)
ps_samp_m_nose_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 324 taxa and 143 samples ]
## sample_data() Sample Data:       [ 143 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 324 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 324 reference sequences ]
```

```
summary(sample_sums(ps_samp_m_nose_tu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3200   24141   38230   41582   52837  116624
```

Now which patients have only one time point left?

```
table(sample_data(ps_samp_m_nose_tu)$Patient_ID)
```

```
##
## P01 P02 P03 P04 P05 P06 P07 P08 P09 P10 P11 P12 P13 P14 P15 P16 P17 P18 P19 P20
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P21 P22 P24 P25 P26 P27 P28 P29 P30 P31 P34 P35 P36 P37 P38 P39 P40 P41 P42 P43
##   2   2   1   2   2   1   2   2   2   2   2   2   2   2   1   2   2   2   2   2
## P44 P45 P46 P47 P48 P49 P50 P51 P52 P53 P54 P55 P56 P57 P58 P59 P60 P61 P62 P63
##   1   2   1   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P64 P65 P66 P68 P69 P70 P71 P72 P73 P75 P76 P77 P78 P79 P80 P81 P82 P83
##   1   2   1   2   2   1   2   2   1   1   1   2   2   1   2   2   1   2
```

```
length(unique(sample_data(ps_samp_m_nose_tu)$Patient_ID))
```

```
## [1] 78
```

```
ps_samp_m_nose_tu <- prune_samples(!sample_data(ps_samp_m_nose_tu)$Patient_ID %in%
    c("P24", "P27", "P38", "P44", "P46", "P64", "P66", "P70",
        "P73", "P75", "P76", "P79", "P82"), ps_samp_m_nose_tu)
ps_samp_m_nose_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 324 taxa and 130 samples ]
## sample_data() Sample Data:       [ 130 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 324 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 324 reference sequences ]
```

```
length(unique(sample_data(ps_samp_m_nose_tu)$Patient_ID))
```

```
## [1] 65
```

65 patients left

Read count after removing samples <2000 and patients with only 1 time point:

```
print("Nose 65 paients")
```

```
## [1] "Nose 65 paients"
```

```
summary(sample_sums(ps_samp_m_nose_tu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3200   24580   38606   42107   52974  116624
```

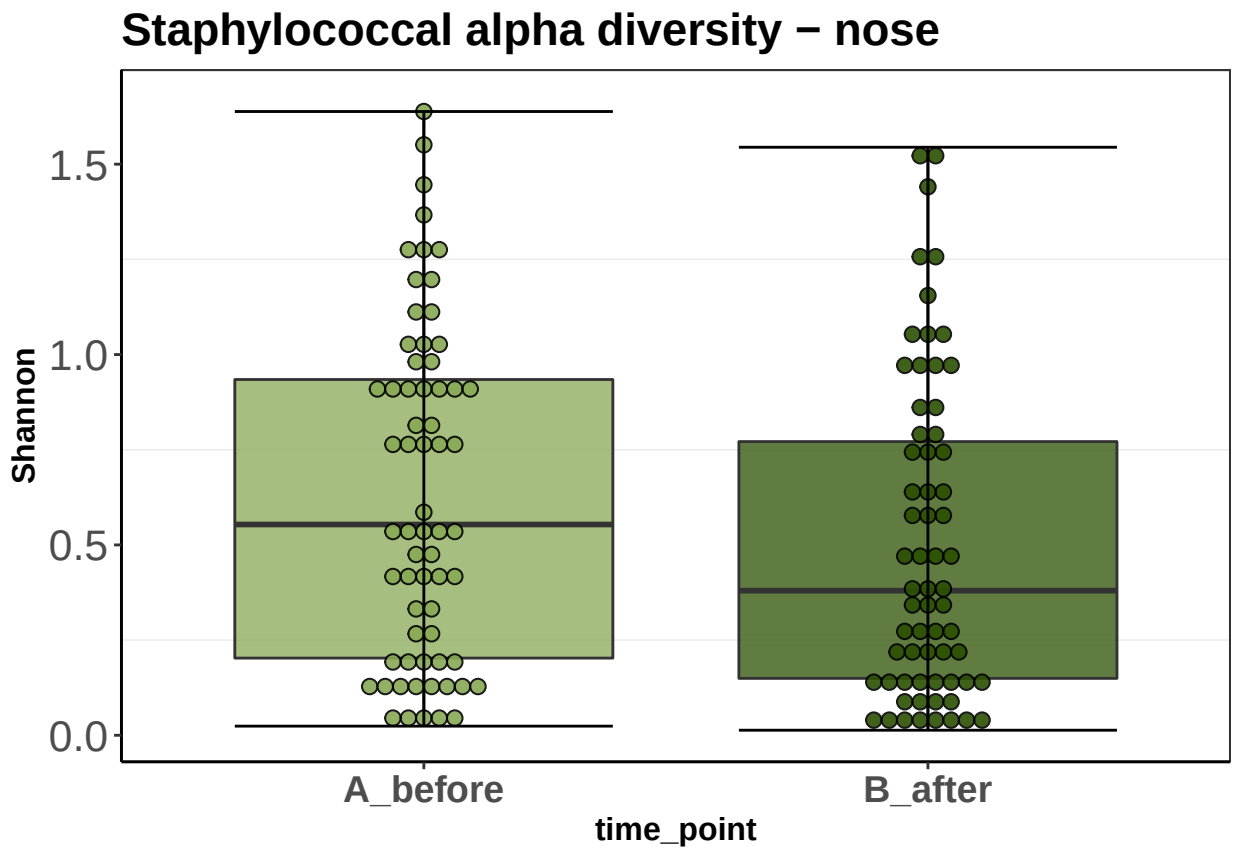
Alpha diversity

```
#### Add diversity measures to the phyloseq object as
#### variables
alpha_div_raw <- estimate_richness(ps_samp_m_nose_tu, measures = c("Observed",
    "Chao1", "Shannon", "InvSimpson"))
rownames(alpha_div_raw) <- gsub("X", "", rownames(alpha_div_raw))
ps_samp_m_nose_tu <- merge_phyloseq(ps_samp_m_nose_tu, sample_data(alpha_div_raw))
df_ps_samp_m_nose_tu <- as(sample_data(ps_samp_m_nose_tu),
    "data.frame")
```

Shannon diversity over time:

```
plot_g_Shannon <- ggplot(df_ps_samp_m_nose_tu, aes(x = time_point,
    y = Shannon, fill = time_point)) + geom_boxplot(outlier.color = "NA",
    alpha = 0.75) + geom_dotplot(binaxis = "y", stackdir = "center",
    alpha = 0.9, position = position_dodge(0.75), dotsize = 0.75) +
    theme(axis.title.y = element_text(size = 12, face = "bold"),
        axis.text.y = element_text(size = 16), axis.text.x = element_text(size = 14,
            face = "bold", angle = 0), axis.title.x = element_text(size = 12,
            face = "bold"), legend.position = "none", panel.grid.major = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"),
        strip.text.x = element_text(angle = 0, face = "bold",
            size = 12), strip.text.y = element_text(angle = 0,
            face = "bold", size = 12), strip.background = element_rect(fill = "white"),
        title = element_text(size = 14, face = "bold")) +
    stat_boxplot(geom = "errorbar") + scale_fill_manual(values = c("#88a954",
    "#2b5000")) + ggtitle("Staphylococcal alpha diversity - nose")
plot_g_Shannon
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggsave(filename = "plots/alpha_div_nose_tuf.pdf", plot = plot_g_Shannon,
    device = cairo_pdf, width = 297, height = 210, units = "mm")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```

Paired Wilcoxon signed rank test

```
df_ps_samp_m_nose_tu_c <- dcast(df_ps_samp_m_nose_tu, Patient_ID ~
    time_point, value.var = "Shannon", drop = FALSE)
wilcox.test(df_ps_samp_m_nose_tu_c$A_before, df_ps_samp_m_nose_tu_c$B_after,
    paired = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  df_ps_samp_m_nose_tu_c$A_before and df_ps_samp_m_nose_tu_c$B_after
## V = 1328, p-value = 0.09563
## alternative hypothesis: true location shift is not equal to 0
```

Staphyloccocal alpha diversity in the nose does not decrease significantly, but there is a trend.

Agglomerate on species level

```
ps_samp_m_nose_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 324 taxa and 130 samples ]
## sample_data() Sample Data:       [ 130 samples by 8 sample variables ]
## tax_table()   Taxonomy Table:    [ 324 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 324 reference sequences ]
```

```
ps_samp_m_nose_tu_gs <- tax_glom(ps_samp_m_nose_tu, taxrank = "Genus_Species")
ps_samp_m_nose_tu_gs <- prune_taxa(taxa_sums(ps_samp_m_nose_tu_gs) !=
    0, ps_samp_m_nose_tu_gs)
ps_samp_m_nose_tu_gs
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 32 taxa and 130 samples ]
## sample_data() Sample Data:       [ 130 samples by 8 sample variables ]
## tax_table()   Taxonomy Table:    [ 32 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 32 reference sequences ]
```

Convert to relative abundance

```
ps_samp_m_nose_tu_gs_rel = transform_sample_counts(ps_samp_m_nose_tu_gs,
    function(x) x/sum(x))
summary(sample_sums(ps_samp_m_nose_tu_gs_rel))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       1       1       1       1
```

Subset top 10 species

```
Species10 = names(sort(taxa_sums(ps_samp_m_nose_tu_gs_rel),
    TRUE)[1:10])
```

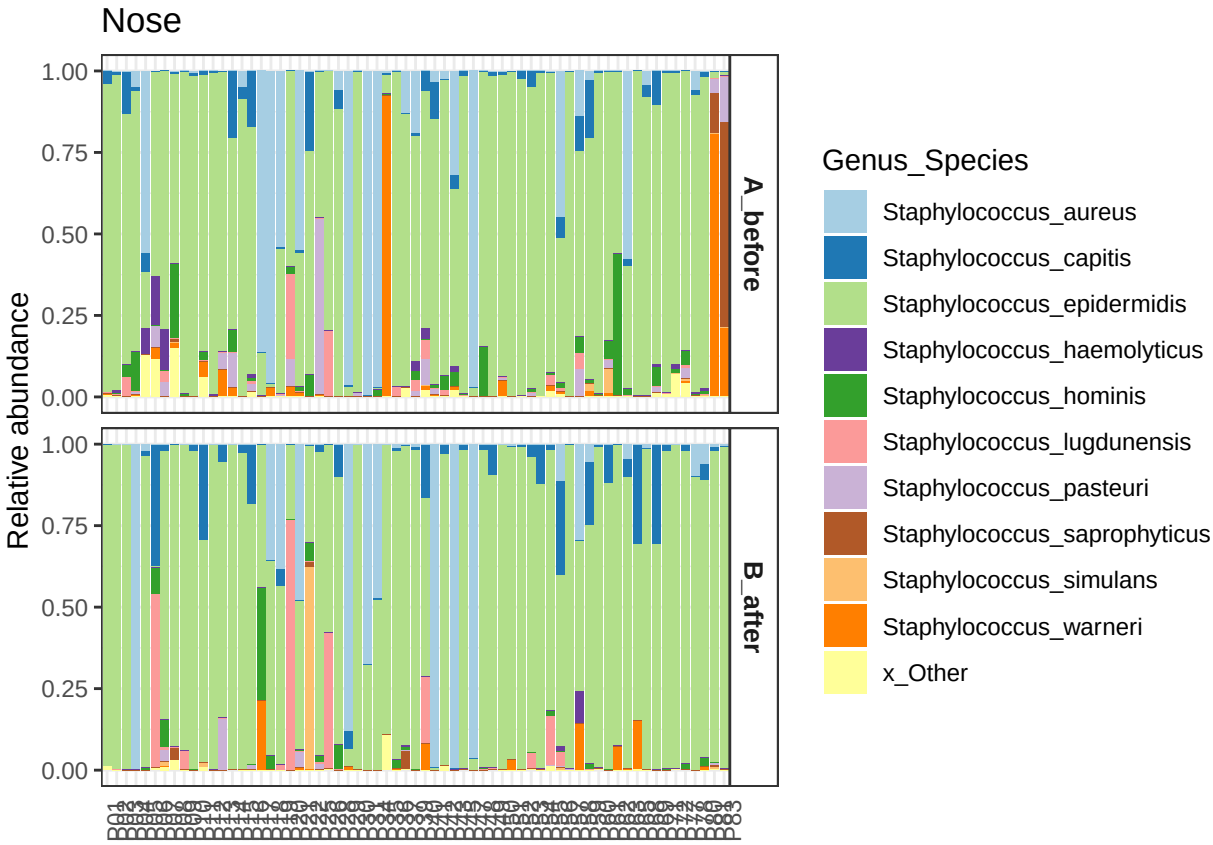
to data frame

```
p_df_o <- psmelt(ps_samp_m_nose_tu_gs_rel)
p_df_o$Genus_Species <- as.character(p_df_o$Genus_Species)
p_df_o$Genus_Species[!(p_df_o$OTU %in% Species10)] <- "x_Other"
```

Barplots of relative abundance

The patients are not in the same order here as in the heatmap, because in the heatmap they are ordered by clustering and here just by number (see ordered version below)

```
np <- ggplot(p_df_o, aes(x = Patient_ID, y = Abundance,
    fill = Genus_Species)) + geom_bar(stat = "identity",
    width = 0.9) + facet_grid(time_point ~ ., scales = "free") +
    scale_fill_manual(values = staph_col) + theme(axis.title.x = element_blank(),
    axis.ticks.x = element_blank(), axis.text.x = element_text(angle = 90),
    strip.background = element_rect(fill = "white"), strip.text.y = element_text(size = 10,
        face = "bold")) + ylab("Relative abundance") + ggtitle("Nose")

np
```



Patient-wise plots

```
ggplot(p_df_o, aes(x = time_point, y = Abundance, fill = Genus_Species)) +
    geom_bar(stat = "identity", width = 0.9) + facet_wrap(. ~
    Patient_ID, nrow = 5) + scale_fill_manual(values = staph_col) +
    theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 90), strip.background = element_rect(fill = "white"),
        strip.text.y = element_text(size = 10, face = "bold"),
        legend.position = "bottom") + ylab("Relative abundance") +
    ggtitle("Nose")
```

Subset the top 10 genera (without other)

```
ps_samp_m_nose_tu_gs_rel <- prune_taxa(taxa_names(ps_samp_m_nose_tu_gs_rel) %in%
    Species10, ps_samp_m_nose_tu_gs_rel)
ps_samp_m_nose_tu_gs_rel
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 10 taxa and 130 samples ]
## sample_data() Sample Data:       [ 130 samples by 8 sample variables ]
## tax_table()   Taxonomy Table:    [ 10 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 10 reference sequences ]
```

to data frame

```
p_df <- psmelt(ps_samp_m_nose_tu_gs_rel)
p_df_d <- dcast(p_df, Patient_ID + Genus_Species ~ time_point,
    value.var = "Abundance", drop = FALSE)
```

Calculate relative change in each patient for each species

```
p_df_d <- p_df_d %>% mutate(Percent_point_change = B_after -
    A_before)
p_df_d$Percent_point_change <- p_df_d$Percent_point_change *
    100
```

to matrix

```
p_df_d_m <- acast(p_df_d[, c(1, 2, 5)], Genus_Species ~
    Patient_ID, value.var = "Percent_point_change")
```

Visualize in a heatmap

```
pdf(file = "plots/Nose_heatmap_tuf.pdf", width = 11.69,
    height = 8.27)

heatmap.2(p_df_d_m, scale = "none", col = bluered(100),
    trace = "none", density.info = "histogram", margin = c(6,
        15), cexRow = 1, cexCol = 0.75, adjCol = 1, key.xlab = "Relative abundance change \nin percent
    keysize = 0.7, key.title = NA, main = "NOSE")

dev.off()
```

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

Which of the top 10 Staph species do significantly change from before to after?

(Paired Wilcoxon test)

```
wilc_df <- p_df_d %>% group_by(Genus_Species) %>% summarise(wilcox_p_value = wilcox.test(A_before,
    B_after, paired = TRUE)$p.value)
```

```
## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## `summarise()` ungrouping output (override with `.groups` argument)
```

```
wilc_df$BH_adjusted_wilcox_p_value <- p.adjust(wilc_df$wilcox_p_value,
    method = "BH")

wilc_df
```

```
## # A tibble: 10 x 3
##    Genus_Species               wilcox_p_value BH_adjusted_wilcox_p_value
```

```
##    <chr>                                <dbl>                    <dbl>
##  1 Staphylococcus_aureus             0.136                    0.227
##  2 Staphylococcus_capitis            0.0887                   0.200
##  3 Staphylococcus_epidermidis        0.205                    0.293
##  4 Staphylococcus_haemolyticus       0.00339                  0.0170
##  5 Staphylococcus_hominis            0.00586                  0.0195
##  6 Staphylococcus_lugdunensis        0.829                    0.896
##  7 Staphylococcus_pasteuri           0.000299                 0.00299
##  8 Staphylococcus_saprophyticus      0.896                    0.896
##  9 Staphylococcus_simulans           0.399                    0.498
## 10 Staphylococcus_warneri            0.100                    0.200
```

S. haemolyticus, S. hominis and S. pasteuri have a significant *overall* change in the nose, also after multiple testing correction.

Do they *overall* decrease or increase?

```
p_df_d %>% group_by(Genus_Species) %>% summarise(Mean_percent_point_change = mean(B_after) -
    mean(A_before))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 10 x 2
##    Genus_Species                Mean_percent_point_change
##    <chr>                                           <dbl>
##  1 Staphylococcus_aureus                         -0.0245
##  2 Staphylococcus_capitis                         0.0189
##  3 Staphylococcus_epidermidis                     0.0548
##  4 Staphylococcus_haemolyticus                   -0.00594
##  5 Staphylococcus_hominis                        -0.0127
##  6 Staphylococcus_lugdunensis                     0.0219
##  7 Staphylococcus_pasteuri                       -0.0176
##  8 Staphylococcus_saprophyticus                  -0.00996
##  9 Staphylococcus_simulans                        0.00858
## 10 Staphylococcus_warneri                        -0.0257
```

The 3 significant species decrease after treatment.

Write results to table for use in 16S script (for Staph genus and species correlation analyis)

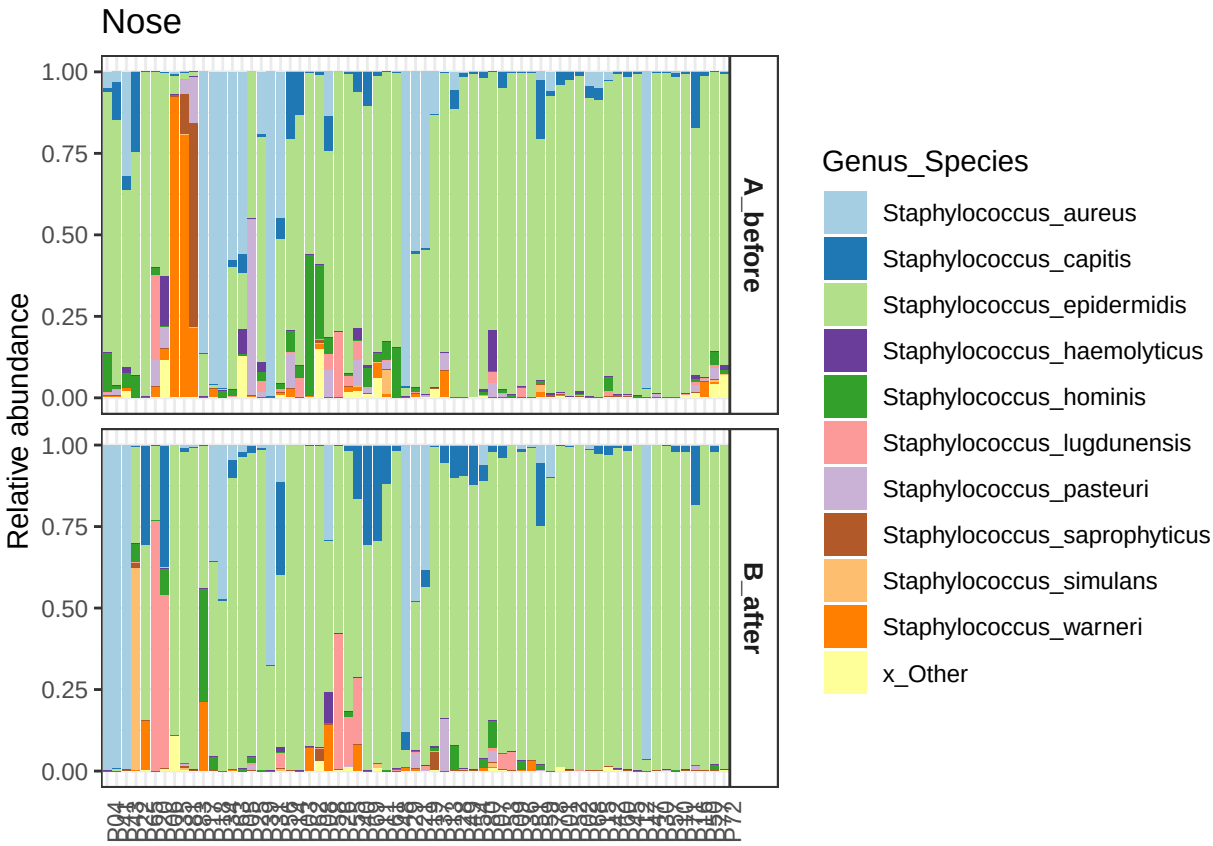
```
p_df_d_tuf_nose <- p_df_d %>% select(Patient_ID, Genus_Species,
    Percent_point_change) %>% pivot_wider(names_from = Genus_Species,
    values_from = Percent_point_change)
write.table(p_df_d_tuf_nose, file = "tables/p_df_d_tuf_nose.csv",
    sep = ";", row.names = FALSE)
```

Make a version of the barplots with the same order of patients as in the heatmap

```
positions <- rownames(hmm$carpet)

np1 <- ggplot(p_df_o, aes(x = Patient_ID, y = Abundance,
```

```
    fill = Genus_Species)) + geom_bar(stat = "identity",
    width = 0.9) + facet_grid(time_point ~ ., scales = "free") +
    scale_fill_manual(values = staph_col) + theme(axis.title.x = element_blank(),
    axis.ticks.x = element_blank(), axis.text.x = element_text(angle = 90),
    strip.background = element_rect(fill = "white"), strip.text.y = element_text(size = 10,
        face = "bold")) + ylab("Relative abundance") + ggtitle("Nose") +
    scale_x_discrete(limits = positions)
```

```
np1
```




```
ggsave(filename = "plots/nose_bars_tuf_ordered_IDs.pdf",
    plot = np1, device = cairo_pdf, width = 297, height = 210,
    units = "mm")
```

Groin

```
ps_samp_m_groin <- prune_samples(sample_data(ps)$Sample_type ==
    "Groin", ps)
ps_samp_m_groin <- prune_taxa(taxa_sums(ps_samp_m_groin) !=
    0, ps_samp_m_groin)
ps_samp_m_groin
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 264 taxa and 126 samples ]
## sample_data() Sample Data:       [ 126 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 264 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 264 reference sequences ]
```

```
sample_data(ps_samp_m_groin)$Sample_type
```

```
##   [1] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
##  [10] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
##  [19] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
##  [28] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
##  [37] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
##  [46] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
##  [55] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
##  [64] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
##  [73] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
##  [82] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
##  [91] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
## [100] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
## [109] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
## [118] "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin" "Groin"
```

```
length(unique(sample_data(ps_samp_m_groin)$Patient_ID))
```

```
## [1] 65
```

Which patients have both, a before and an after sample from the groin

```
table(sample_data(ps_samp_m_groin)$Patient_ID)
```

```
##
## P01 P02 P03 P04 P05 P06 P07 P09 P10 P11 P12 P13 P15 P16 P17 P18 P19 P20 P21 P22
##   2   2   2   2   2   2   2   2   2   2   1   2   2   1   2   2   1   2   2   2
## P23 P24 P25 P26 P27 P28 P32 P33 P35 P36 P37 P39 P43 P45 P47 P48 P49 P50 P51 P53
##   2   2   2   2   2   2   1   2   2   2   2   2   2   2   2   2   2   2   2   2
## P54 P55 P61 P62 P63 P64 P65 P66 P67 P68 P69 P70 P71 P72 P73 P74 P75 P76 P77 P78
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P79 P80 P81 P82 P83
##   2   2   2   2   2
```

```
ps_samp_m_groin <- prune_samples(!sample_data(ps_samp_m_groin)$Patient_ID %in%
    c("P12", "P16", "P19", "P32"), ps_samp_m_groin)
ps_samp_m_groin
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 264 taxa and 122 samples ]
## sample_data() Sample Data:       [ 122 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 264 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 264 reference sequences ]
```

```
table(sample_data(ps_samp_m_groin)$Patient_ID)
```

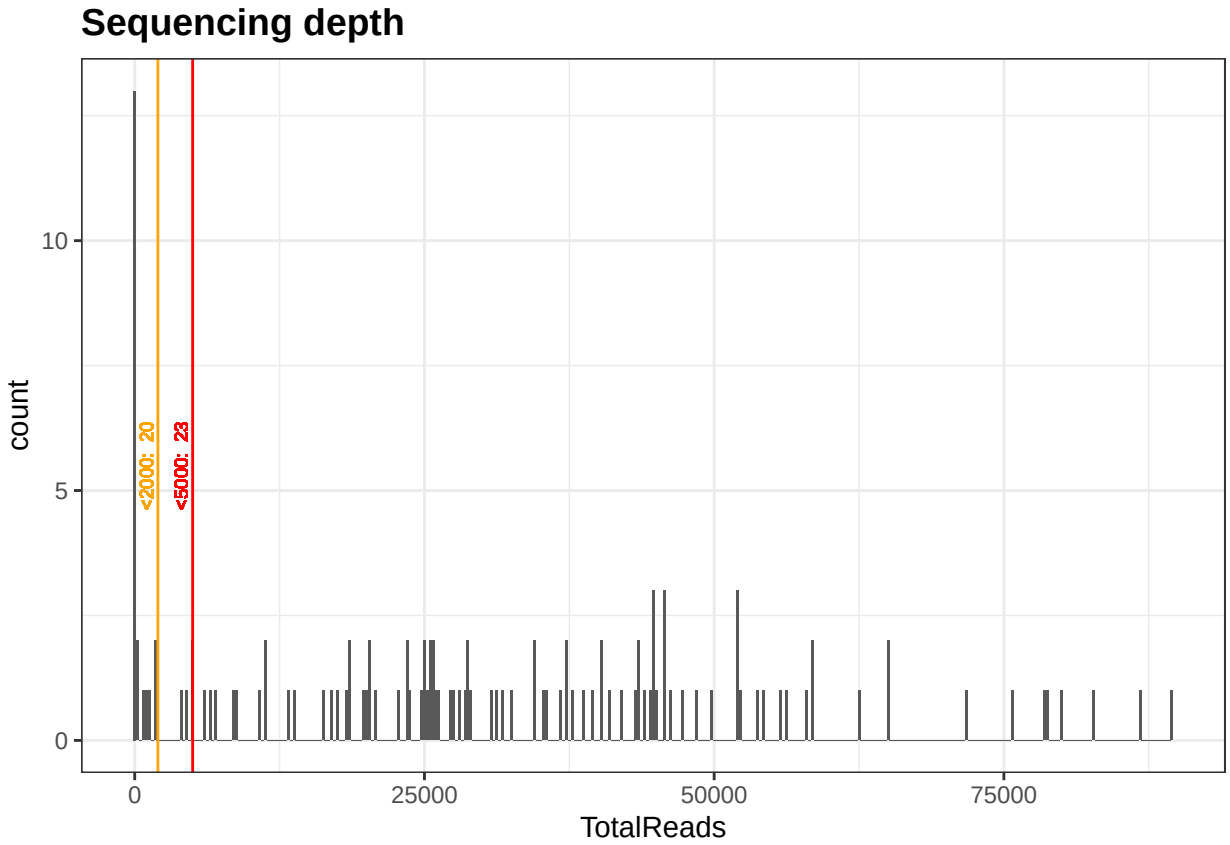
```
##
## P01 P02 P03 P04 P05 P06 P07 P09 P10 P11 P13 P15 P17 P18 P20 P21 P22 P23 P24 P25
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P26 P27 P28 P33 P35 P36 P37 P39 P43 P45 P47 P48 P49 P50 P51 P53 P54 P55 P61 P62
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P63 P64 P65 P66 P67 P68 P69 P70 P71 P72 P73 P74 P75 P76 P77 P78 P79 P80 P81 P82
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P83
##   2
```

```
length(unique(sample_data(ps_samp_m_groin)$Patient_ID))
```

```
## [1] 61
```

Seq depth

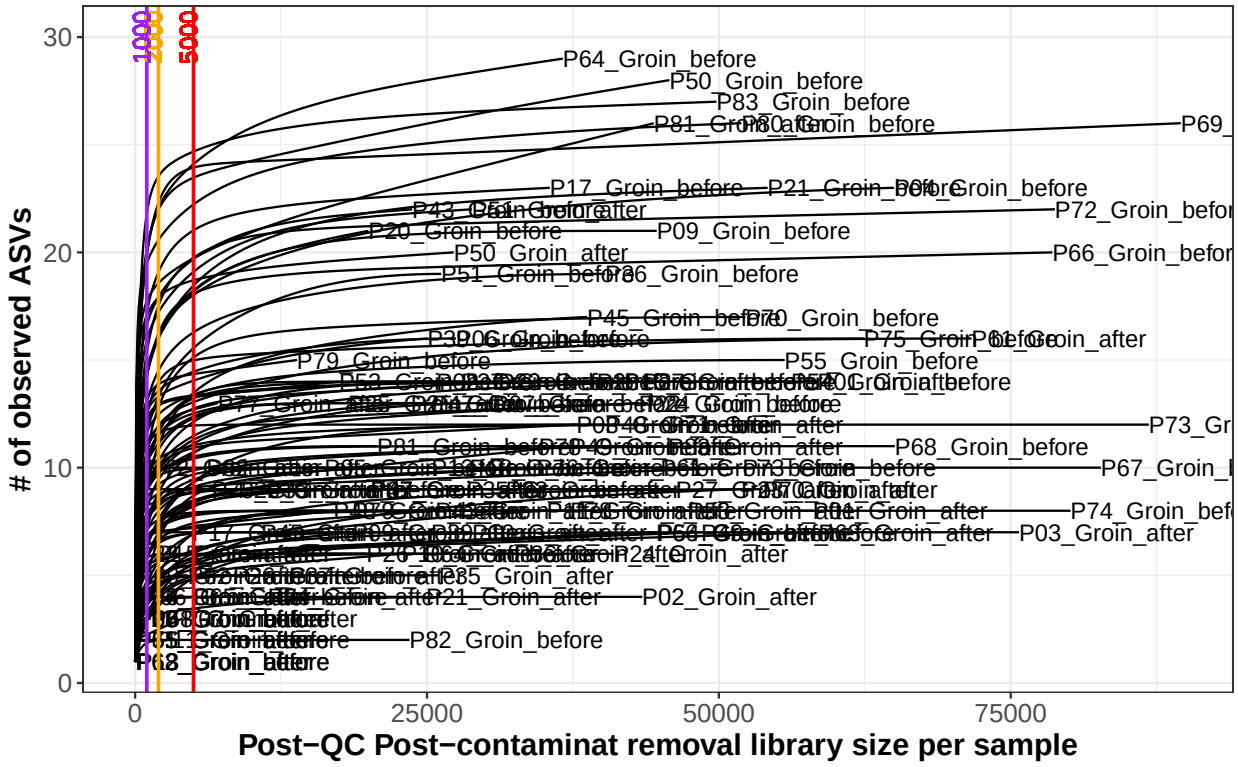
```
sdt = data.table::data.table(as(sample_data(ps_samp_m_groin),
    "data.frame"), TotalReads = sample_sums(ps_samp_m_groin),
    keep.rownames = TRUE)
data.table::setnames(sdt, "rn", "SampleID")
pSeqDepth = ggplot(sdt, aes(TotalReads)) + geom_histogram(binwidth = 250) +
    geom_vline(xintercept = 5000, color = "red") + geom_vline(xintercept = 2000,
    color = "orange") + geom_text(aes(x = 1000, label = paste("<2000: ",
    nrow(sdt[sdt$TotalReads < 2000])), y = 5.5), colour = "orange",
    angle = 90, size = 2.5) + geom_text(aes(x = 4000, label = paste("<5000: ",
    nrow(sdt[sdt$TotalReads < 5000])), y = 5.5), colour = "red",
    angle = 90, size = 2.5) + ggtitle("Sequencing depth") +
    theme(plot.title = element_text(size = 14, face = "bold"))

pSeqDepth
```



Do the rarefaction curves justify that we remove samples with reads <2000?

Rarefaction curves

```
p20 <- p20 + theme(panel.background = element_blank(), axis.title.x = element_text(size = 14,
    face = "bold"), axis.title.y = element_text(size = 14,
    face = "bold"), axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12), legend.title = element_text(size = 16,
        face = "bold"), legend.text = element_text(size = 16),
    strip.text.x = element_text(angle = 0, face = "bold",
        size = 12), strip.background = element_rect(fill = "white")) +
    xlab("Post-QC Post-contaminat removal library size per sample") +
    ylab("# of observed ASVs") + geom_vline(xintercept = 5000,
    color = "red", size = 0.8) + geom_vline(xintercept = 2000,
    color = "orange", size = 0.8) + geom_vline(xintercept = 1000,
    color = "purple", size = 0.8) + geom_text(aes(x = 4550,
    label = "5000", y = 30), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 30), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 30), colour = "purple",
        angle = 90, size = 4)
p20
```

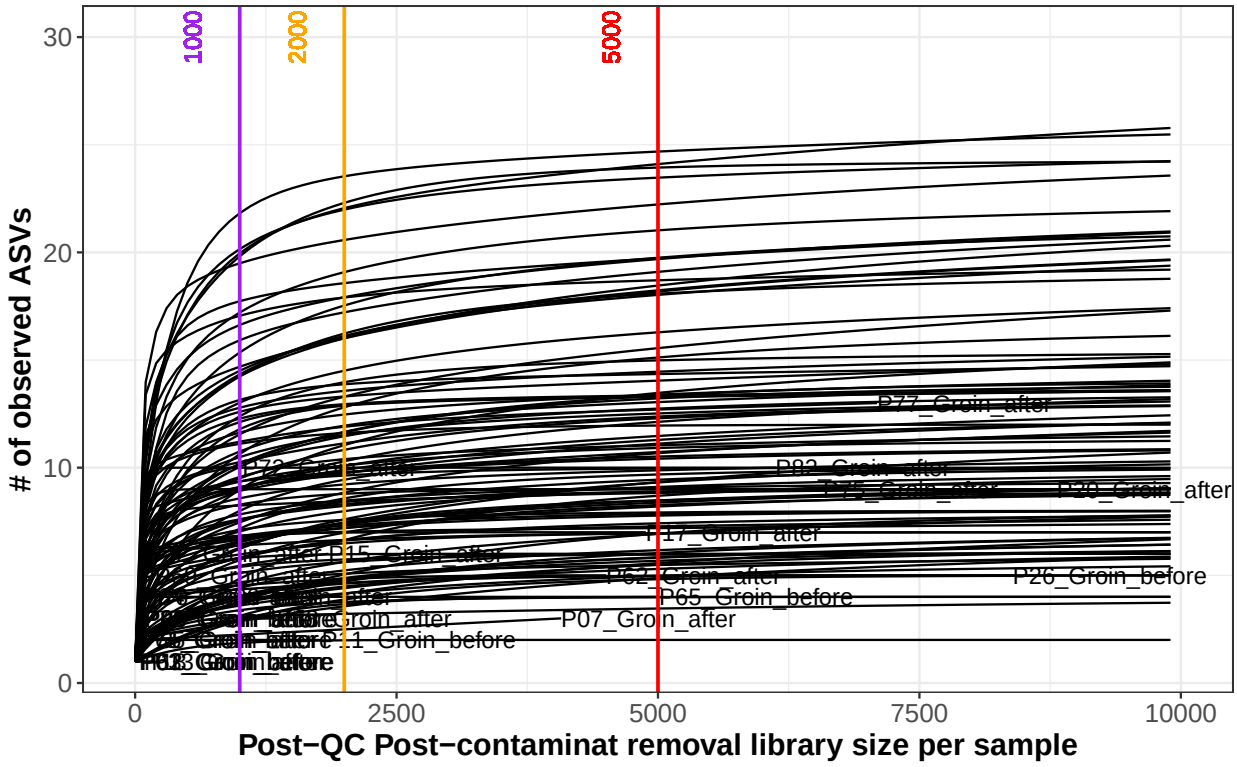


Zoom

```
p20 + xlim(0, 10000)
```

```
## Warning: Removed 93 rows containing missing values (geom_text).
```

```
## Warning: Removed 27195 row(s) containing missing values (geom_path).
```



How do the rarefaction curves look on species level?

Rarefaction curves

```
ps_samp_m_groin_gen <- tax_glom(ps_samp_m_groin, taxrank = "Genus_Species")
ps_samp_m_groin_gen
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 26 taxa and 122 samples ]
## sample_data() Sample Data:       [ 122 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 26 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 26 reference sequences ]
```

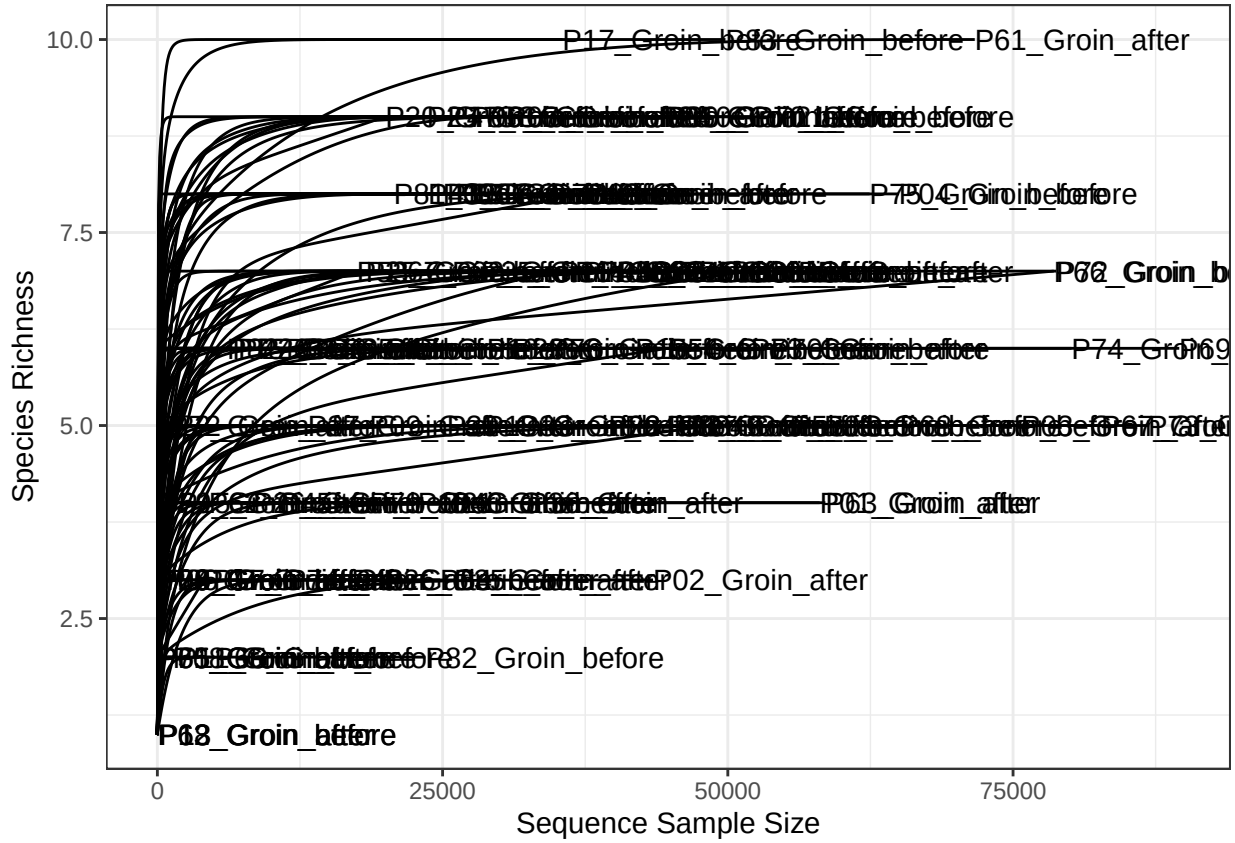
```
ps_samp_m_groin_gen <- prune_taxa(taxa_sums(ps_samp_m_groin_gen) !=
    0, ps_samp_m_groin_gen)
set.seed(123)
p30 <- ggrare(ps_samp_m_groin_gen, step = 100, se = FALSE,
    label = "Sample")
```

```
## rarefying sample P61_Groin_before
## rarefying sample P61_Groin_after
## rarefying sample P70_Groin_before
## rarefying sample P70_Groin_after
## rarefying sample P71_Groin_before
## rarefying sample P71_Groin_after
## rarefying sample P72_Groin_before
```

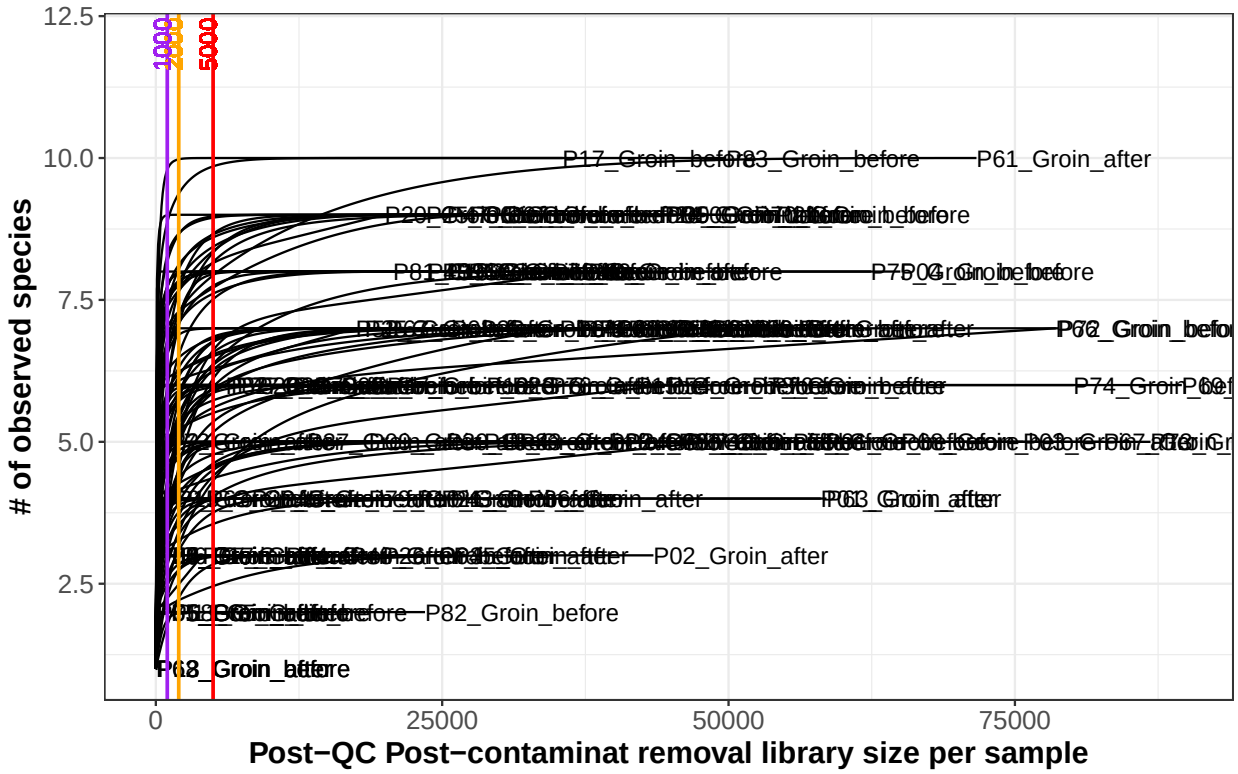
```
## rarefying sample P72_Groin_after
## rarefying sample P73_Groin_before
## rarefying sample P73_Groin_after
## rarefying sample P28_Groin_before
## rarefying sample P11_Groin_before
## rarefying sample P05_Groin_before
## rarefying sample P02_Groin_before
## rarefying sample P18_Groin_before
## rarefying sample P07_Groin_before
## rarefying sample P10_Groin_before
## rarefying sample P03_Groin_before
## rarefying sample P18_Groin_after
## rarefying sample P17_Groin_before
## rarefying sample P04_Groin_before
## rarefying sample P15_Groin_before
## rarefying sample P01_Groin_before
## rarefying sample P06_Groin_before
## rarefying sample P09_Groin_before
## rarefying sample P05_Groin_after
## rarefying sample P11_Groin_after
## rarefying sample P20_Groin_before
## rarefying sample P10_Groin_after
## rarefying sample P04_Groin_after
## rarefying sample P15_Groin_after
## rarefying sample P13_Groin_before
## rarefying sample P03_Groin_after
## rarefying sample P17_Groin_after
## rarefying sample P09_Groin_after
## rarefying sample P07_Groin_after
## rarefying sample P20_Groin_after
## rarefying sample P02_Groin_after
## rarefying sample P13_Groin_after
## rarefying sample P06_Groin_after
## rarefying sample P28_Groin_after
## rarefying sample P39_Groin_before
## rarefying sample P43_Groin_before
## rarefying sample P45_Groin_before
## rarefying sample P39_Groin_after
## rarefying sample P45_Groin_after
## rarefying sample P01_Groin_after
## rarefying sample P47_Groin_before
## rarefying sample P43_Groin_after
## rarefying sample P47_Groin_after
## rarefying sample P48_Groin_before
## rarefying sample P48_Groin_after
## rarefying sample P49_Groin_before
## rarefying sample P49_Groin_after
## rarefying sample P50_Groin_before
## rarefying sample P21_Groin_before
## rarefying sample P22_Groin_before
## rarefying sample P23_Groin_before
## rarefying sample P24_Groin_before
## rarefying sample P25_Groin_before
## rarefying sample P26_Groin_before
```

```
## rarefying sample P27_Groin_before
## rarefying sample P23_Groin_after
## rarefying sample P22_Groin_after
## rarefying sample P25_Groin_after
## rarefying sample P21_Groin_after
## rarefying sample P24_Groin_after
## rarefying sample P50_Groin_after
## rarefying sample P27_Groin_after
## rarefying sample P26_Groin_after
## rarefying sample P33_Groin_before
## rarefying sample P35_Groin_before
## rarefying sample P35_Groin_after
## rarefying sample P36_Groin_before
## rarefying sample P37_Groin_before
## rarefying sample P37_Groin_after
## rarefying sample P36_Groin_after
## rarefying sample P33_Groin_after
## rarefying sample P51_Groin_before
## rarefying sample P51_Groin_after
## rarefying sample P53_Groin_before
## rarefying sample P53_Groin_after
## rarefying sample P54_Groin_before
## rarefying sample P54_Groin_after
## rarefying sample P55_Groin_before
## rarefying sample P55_Groin_after
## rarefying sample P74_Groin_before
## rarefying sample P74_Groin_after
## rarefying sample P75_Groin_before
## rarefying sample P75_Groin_after
## rarefying sample P76_Groin_before
## rarefying sample P76_Groin_after
## rarefying sample P77_Groin_before
## rarefying sample P77_Groin_after
## rarefying sample P78_Groin_before
## rarefying sample P78_Groin_after
## rarefying sample P62_Groin_before
## rarefying sample P62_Groin_after
## rarefying sample P79_Groin_before
## rarefying sample P79_Groin_after
## rarefying sample P80_Groin_before
## rarefying sample P80_Groin_after
## rarefying sample P81_Groin_before
## rarefying sample P81_Groin_after
## rarefying sample P82_Groin_before
## rarefying sample P82_Groin_after
## rarefying sample P83_Groin_before
## rarefying sample P83_Groin_after
## rarefying sample P63_Groin_before
## rarefying sample P63_Groin_after
## rarefying sample P64_Groin_before
## rarefying sample P64_Groin_after
## rarefying sample P65_Groin_before
## rarefying sample P65_Groin_after
## rarefying sample P66_Groin_before
```

```
## rarefying sample P66_Groin_after
## rarefying sample P67_Groin_before
## rarefying sample P67_Groin_after
## rarefying sample P68_Groin_before
## rarefying sample P68_Groin_after
## rarefying sample P69_Groin_before
## rarefying sample P69_Groin_after
```




```
p30 <- p30 + theme(panel.background = element_blank(), axis.title.x = element_text(size = 14,
    face = "bold"), axis.title.y = element_text(size = 14,
    face = "bold"), axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12), legend.title = element_text(size = 16,
        face = "bold"), legend.text = element_text(size = 16),
    strip.text.x = element_text(angle = 0, face = "bold",
        size = 12), strip.background = element_rect(fill = "white")) +
    xlab("Post-QC Post-contaminat removal library size per sample") +
    ylab("# of observed species") + geom_vline(xintercept = 5000,
    color = "red", size = 0.8) + geom_vline(xintercept = 2000,
    color = "orange", size = 0.8) + geom_vline(xintercept = 1000,
    color = "purple", size = 0.8) + geom_text(aes(x = 4550,
    label = "5000", y = 12), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 12), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 12), colour = "purple",
        angle = 90, size = 4)
p30
```

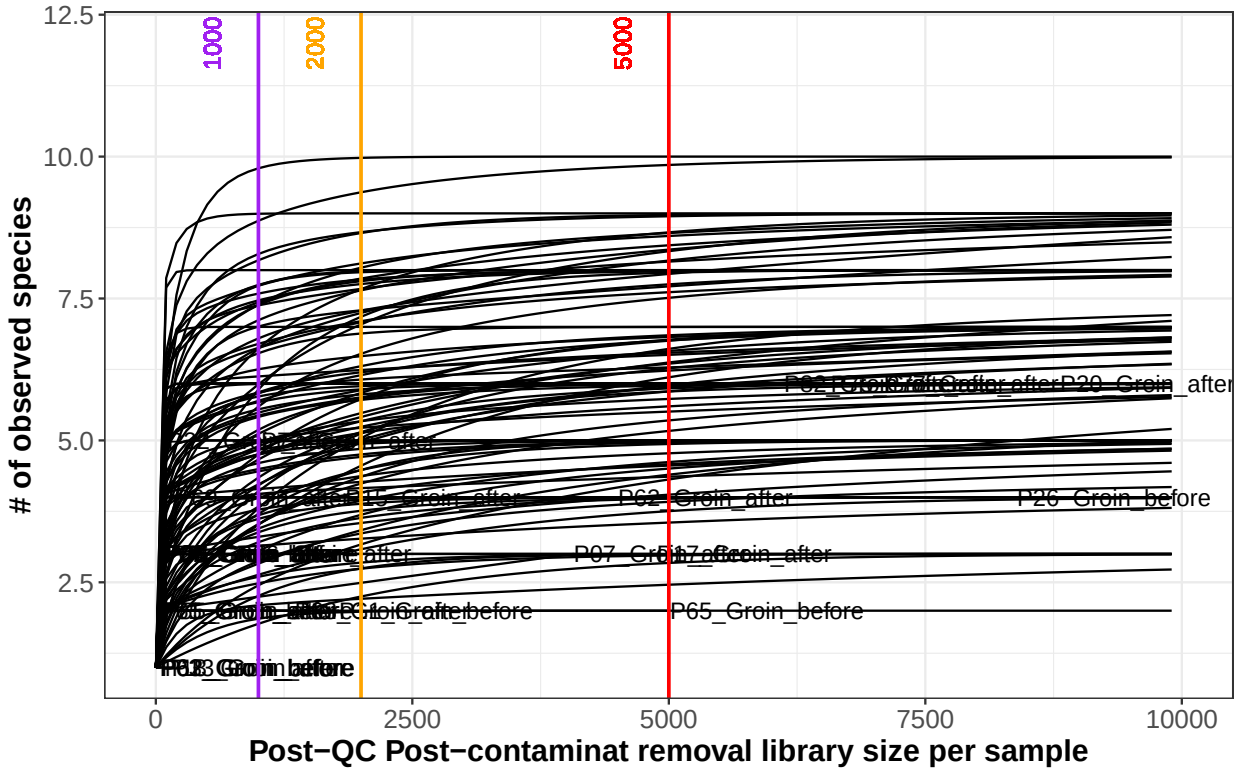


Zoom

```
p30 + xlim(0, 10000)
```

```
## Warning: Removed 93 rows containing missing values (geom_text).
```

```
## Warning: Removed 27195 row(s) containing missing values (geom_path).
```



Exclude samples with <2000 counts

```
summary(sample_sums(ps_samp_m_groin))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      30   11311   27708   30322   44723   89572
```

```
ps_samp_m_groin_tu <- prune_samples(!sample_sums(ps_samp_m_groin) <
    2000, ps_samp_m_groin)
ps_samp_m_groin_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 264 taxa and 102 samples ]
## sample_data() Sample Data:       [ 102 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 264 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 264 reference sequences ]
```

```
summary(sample_sums(ps_samp_m_groin_tu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4074   22993   34488   36189   46187   89572
```

Now which patients have only one time point left?

```
table(sample_data(ps_samp_m_groin_tu)$Patient_ID)
```

```
##
## P01 P02 P03 P04 P05 P06 P07 P09 P10 P11 P13 P15 P17 P18 P20 P21 P22 P23 P24 P25
##   2   2   2   1   1   1   2   2   2   1   1   1   2   1   2   2   1   2   2   2
## P26 P27 P28 P33 P35 P36 P37 P39 P43 P45 P47 P48 P49 P50 P51 P53 P54 P55 P61 P62
##   2   2   1   2   2   2   2   2   2   2   2   2   2   2   2   2   2   1   2   1
## P63 P64 P65 P66 P67 P68 P69 P70 P71 P72 P73 P74 P75 P76 P77 P78 P79 P80 P81 P82
##   1   2   1   2   2   1   1   2   1   1   2   2   2   1   1   2   2   1   2   2
## P83
##   2
```

```
length(unique(sample_data(ps_samp_m_groin_tu)$Patient_ID))
```

```
## [1] 61
```

```
ps_samp_m_groin_tu <- prune_samples(!sample_data(ps_samp_m_groin_tu)$Patient_ID %in%
    c("P04", "P05", "P06", "P11", "P13", "P15", "P18", "P22",
        "P28", "P55", "P62", "P63", "P65", "P68", "P69",
        "P71", "P72", "P76", "P77", "P80"), ps_samp_m_groin_tu)
ps_samp_m_groin_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 264 taxa and 82 samples ]
## sample_data() Sample Data:       [ 82 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 264 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 264 reference sequences ]
```

```
length(unique(sample_data(ps_samp_m_groin_tu)$Patient_ID))
```

```
## [1] 41
```

41 patients left

Read count after removing samples <2000 and patients with only 1 time point:

```
print("Groin 41 paients")
```

```
## [1] "Groin 41 paients"
```

```
summary(sample_sums(ps_samp_m_groin_tu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4074   21251   31442   35029   45493   86836
```

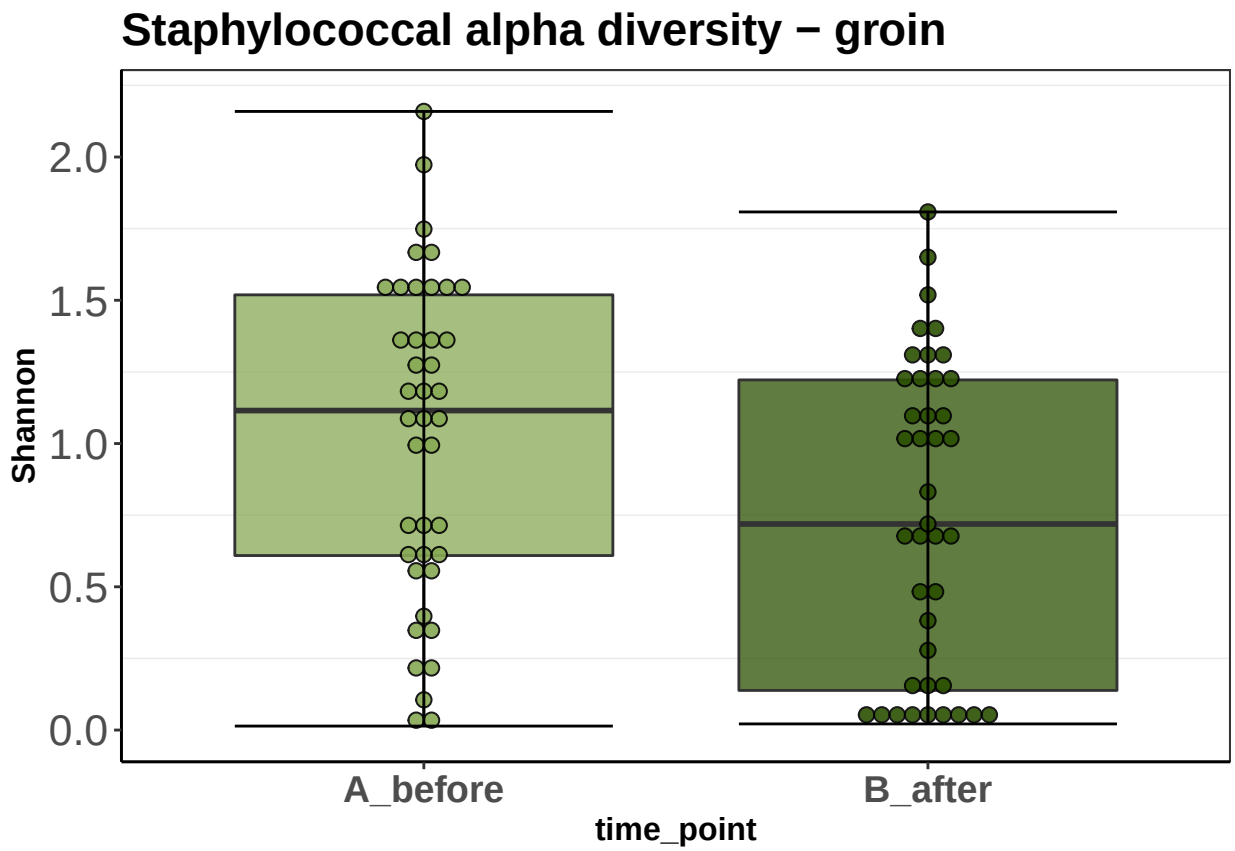
Alpha diversity

```
#### Add diversity measures to the phyloseq object as
#### variables
alpha_div_raw <- estimate_richness(ps_samp_m_groin_tu, measures = c("Observed",
    "Chao1", "Shannon", "InvSimpson"))
rownames(alpha_div_raw) <- gsub("X", "", rownames(alpha_div_raw))
ps_samp_m_groin_tu <- merge_phyloseq(ps_samp_m_groin_tu,
    sample_data(alpha_div_raw))
df_ps_samp_m_groin_tu <- as(sample_data(ps_samp_m_groin_tu),
    "data.frame")
```

Shannon diversity over time:

```
plot_g_Shannon <- ggplot(df_ps_samp_m_groin_tu, aes(x = time_point,
    y = Shannon, fill = time_point)) + geom_boxplot(outlier.color = "NA",
    alpha = 0.75) + geom_dotplot(binaxis = "y", stackdir = "center",
    alpha = 0.9, position = position_dodge(0.75), dotsize = 0.75) +
    theme(axis.title.y = element_text(size = 12, face = "bold"),
        axis.text.y = element_text(size = 16), axis.text.x = element_text(size = 14,
            face = "bold", angle = 0), axis.title.x = element_text(size = 12,
            face = "bold"), legend.position = "none", panel.grid.major = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"),
        strip.text.x = element_text(angle = 0, face = "bold",
            size = 12), strip.text.y = element_text(angle = 0,
            face = "bold", size = 12), strip.background = element_rect(fill = "white"),
        title = element_text(size = 14, face = "bold")) +
    stat_boxplot(geom = "errorbar") + scale_fill_manual(values = c("#88a954",
    "#2b5000")) + ggtitle("Staphylococcal alpha diversity - groin")
plot_g_Shannon
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggsave(filename = "plots/alpha_div_groin_tuf.pdf", plot = plot_g_Shannon,
    device = cairo_pdf, width = 297, height = 210, units = "mm")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```

Paired Wilcoxon signed rank test

```
df_ps_samp_m_groin_tu_c <- dcast(df_ps_samp_m_groin_tu,
    Patient_ID ~ time_point, value.var = "Shannon", drop = FALSE)
wilcox.test(df_ps_samp_m_groin_tu_c$A_before, df_ps_samp_m_groin_tu_c$B_after,
    paired = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  df_ps_samp_m_groin_tu_c$A_before and df_ps_samp_m_groin_tu_c$B_after
## V = 600, p-value = 0.02744
## alternative hypothesis: true location shift is not equal to 0
```

Staphyloccocal alpha diversity in the groin decreases significantly.

Agglomerate on species level

```
ps_samp_m_groin_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 264 taxa and 82 samples ]
## sample_data() Sample Data:       [ 82 samples by 8 sample variables ]
## tax_table()   Taxonomy Table:    [ 264 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 264 reference sequences ]
```

```
ps_samp_m_groin_tu_gs <- tax_glom(ps_samp_m_groin_tu, taxrank = "Genus_Species")
ps_samp_m_groin_tu_gs <- prune_taxa(taxa_sums(ps_samp_m_groin_tu_gs) !=
    0, ps_samp_m_groin_tu_gs)
ps_samp_m_groin_tu_gs
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 24 taxa and 82 samples ]
## sample_data() Sample Data:       [ 82 samples by 8 sample variables ]
## tax_table()   Taxonomy Table:    [ 24 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 24 reference sequences ]
```

Convert to relative abundance

```
ps_samp_m_groin_tu_gs_rel = transform_sample_counts(ps_samp_m_groin_tu_gs,
    function(x) x/sum(x))
summary(sample_sums(ps_samp_m_groin_tu_gs_rel))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       1       1       1       1
```

Subset top 10 species

```
Species10 = names(sort(taxa_sums(ps_samp_m_groin_tu_gs_rel),
    TRUE)[1:10])
```

to data frame

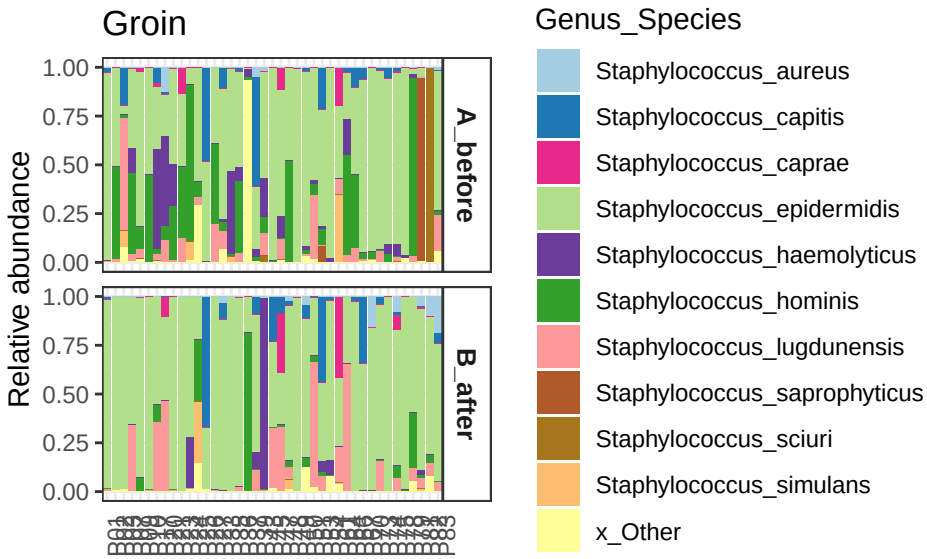
```
p_df_o_g <- psmelt(ps_samp_m_groin_tu_gs_rel)
p_df_o_g$Genus_Species <- as.character(p_df_o_g$Genus_Species)
p_df_o_g$Genus_Species[!(p_df_o_g$OTU %in% Species10)] <- "x_Other"
```

Barplots of relative abundance

The patients are not in the same order here as in the heatmap, because in the heatmap they are ordered by clustering and here just by number (see ordered version below).

```
gp <- ggplot(p_df_o_g, aes(x = Patient_ID, y = Abundance,
    fill = Genus_Species)) + geom_bar(stat = "identity",
    width = 0.9) + facet_grid(time_point ~ ., scales = "free") +
    scale_fill_manual(values = staph_col) + theme(axis.title.x = element_blank(),
    axis.ticks.x = element_blank(), axis.text.x = element_text(angle = 90),
    strip.background = element_rect(fill = "white"), strip.text.y = element_text(size = 10,
        face = "bold")) + ylab("Relative abundance") + ggtitle("Groin")
```

```
gp
```



Patient-wise plots

```
ggplot(p_df_o_g, aes(x = time_point, y = Abundance, fill = Genus_Species)) +
    geom_bar(stat = "identity", width = 0.9) + facet_wrap(. ~
    Patient_ID, nrow = 5) + scale_fill_manual(values = staph_col) +
    theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 90), strip.background = element_rect(fill = "white"),
        strip.text.y = element_text(size = 10, face = "bold"),
        legend.position = "bottom") + ylab("Relative abundance") +
    ggtitle("Groin")
```

Subset the top 10 genera (without other)

```
ps_samp_m_groin_tu_gs_rel <- prune_taxa(taxa_names(ps_samp_m_groin_tu_gs_rel) %in%
    Species10, ps_samp_m_groin_tu_gs_rel)
ps_samp_m_groin_tu_gs_rel
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 10 taxa and 82 samples ]
## sample_data() Sample Data:       [ 82 samples by 8 sample variables ]
## tax_table()   Taxonomy Table:    [ 10 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 10 reference sequences ]
```

to data frame

```
p_df <- psmelt(ps_samp_m_groin_tu_gs_rel)
p_df_d <- dcast(p_df, Patient_ID + Genus_Species ~ time_point,
    value.var = "Abundance", drop = FALSE)
```

Calculate relative change in each patient for each species

```
p_df_d <- p_df_d %>% mutate(Percent_point_change = B_after -
    A_before)
p_df_d$Percent_point_change <- p_df_d$Percent_point_change *
    100
```

to matrix

```
p_df_d_m <- acast(p_df_d[, c(1, 2, 5)], Genus_Species ~
    Patient_ID, value.var = "Percent_point_change")
```

Visualize in a heatmap

```
pdf(file = "plots/Groin_heatmap_tuf.pdf", width = 11.69,
    height = 8.27)

heatmap.2(p_df_d_m, scale = "none", col = bluered(100),
    trace = "none", density.info = "histogram", margin = c(6,
        15), cexRow = 1, cexCol = 0.75, adjCol = 1, key.xlab = "Relative abundance change \nin percent p
    keysize = 0.7, key.title = NA, main = "GROIN")

dev.off()
```

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

Which of the top 10 Staph species do significantly change from before to after?

(Paired Wilcoxon test)

```
wilc_df <- p_df_d %>% group_by(Genus_Species) %>% summarise(wilcox_p_value = wilcox.test(A_before,
    B_after, paired = TRUE)$p.value)
```

```
## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

wilc_df$BH_adjusted_wilcox_p_value <- p.adjust(wilc_df$wilcox_p_value,
    method = "BH")

wilc_df
```

```
## # A tibble: 10 x 3
##    Genus_Species                wilcox_p_value BH_adjusted_wilcox_p_value
##    <chr>                                 <dbl>                      <dbl>
##  1 Staphylococcus_aureus                 0.724                      0.787
##  2 Staphylococcus_capitis                0.526                      0.658
##  3 Staphylococcus_caprae                 0.351                      0.501
##  4 Staphylococcus_epidermidis            0.0179                     0.0893
##  5 Staphylococcus_haemolyticus           0.0713                     0.178
##  6 Staphylococcus_hominis                0.00102                    0.0102
##  7 Staphylococcus_lugdunensis            0.244                      0.407
##  8 Staphylococcus_saprophyticus          0.0310                     0.103
##  9 Staphylococcus_sciuri                 0.787                      0.787
## 10 Staphylococcus_simulans               0.127                      0.255
```

S. epidermidis, S. hominis and S. saprolyticus have a significant *overall* change in the groin. After multiple testing correction, only S. hominis is still significant.

Do they *overall* decrease or increase?

```
p_df_d %>% group_by(Genus_Species) %>% summarise(Mean_percent_point_change = mean(B_after) -
    mean(A_before))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 10 x 2
##    Genus_Species                Mean_percent_point_change
##    <chr>                                            <dbl>
##  1 Staphylococcus_aureus                          0.0111
##  2 Staphylococcus_capitis                         0.00409
##  3 Staphylococcus_caprae                          0.0103
##  4 Staphylococcus_epidermidis                     0.157
##  5 Staphylococcus_haemolyticus                   -0.0279
##  6 Staphylococcus_hominis                        -0.118
##  7 Staphylococcus_lugdunensis                     0.0415
##  8 Staphylococcus_saprophyticus                  -0.0253
##  9 Staphylococcus_sciuri                         -0.0252
## 10 Staphylococcus_simulans                       -0.00458
```

S. epi increases, the other 3 significant species decrease after treatment in the groin.

Write results to table for use in 16S script (for Staph genus and species correlation analyis)

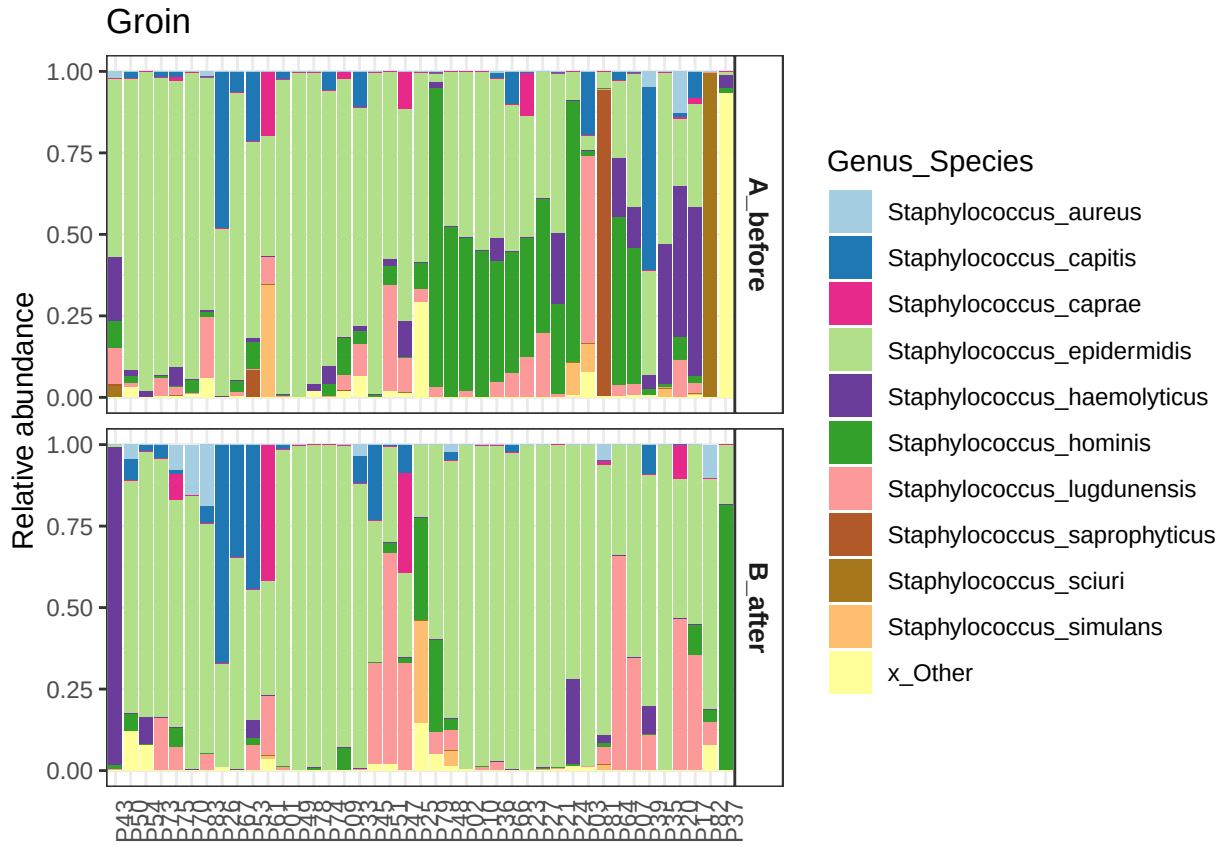
```
p_df_d_tuf_groin <- p_df_d %>% select(Patient_ID, Genus_Species,
    Percent_point_change) %>% pivot_wider(names_from = Genus_Species,
    values_from = Percent_point_change)
write.table(p_df_d_tuf_groin, file = "tables/p_df_d_tuf_groin.csv",
    sep = ";", row.names = FALSE)
```

Make a version of the barplots with the same order of patients as in the heatmap

```
positions <- rownames(hmm$carpet)
```

```
gr <- ggplot(p_df_o_g, aes(x = Patient_ID, y = Abundance,
    fill = Genus_Species)) + geom_bar(stat = "identity",
    width = 0.9) + facet_grid(time_point ~ ., scales = "free") +
    scale_fill_manual(values = staph_col) + theme(axis.title.x = element_blank(),
    axis.ticks.x = element_blank(), axis.text.x = element_text(angle = 90),
    strip.background = element_rect(fill = "white"), strip.text.y = element_text(size = 10,
        face = "bold")) + ylab("Relative abundance") + ggtitle("Groin") +
    scale_x_discrete(limits = positions)
```

```
gr
```




```
ggsave(filename = "plots/groin_bars_tuf_ordered_IDs.pdf",
    plot = gr, device = cairo_pdf, width = 297, height = 210,
    units = "mm")
```

Operation_site

```
ps_samp_m_Operation_site <- prune_samples(sample_data(ps)$Sample_type ==
    "Operation_site", ps)
```

```
ps_samp_m_Operation_site <- prune_taxa(taxa_sums(ps_samp_m_Operation_site) !=
    0, ps_samp_m_Operation_site)
ps_samp_m_Operation_site
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 164 taxa and 74 samples ]
## sample_data() Sample Data:       [ 74 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 164 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 164 reference sequences ]
```

```
sample_data(ps_samp_m_Operation_site)$Sample_type
```

```
##  [1] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
##  [5] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
##  [9] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [13] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [17] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [21] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [25] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [29] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [33] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [37] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [41] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [45] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [49] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [53] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [57] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [61] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [65] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [69] "Operation_site" "Operation_site" "Operation_site" "Operation_site"
## [73] "Operation_site" "Operation_site"
```

```
length(unique(sample_data(ps_samp_m_Operation_site)$Patient_ID))
```

```
## [1] 37
```

Which patients have both, a before and an after sample from Operation_site

```
table(sample_data(ps_samp_m_Operation_site)$Patient_ID)
```

```
##
## P01 P04 P09 P12 P15 P17 P21 P23 P30 P31 P32 P35 P37 P53 P61 P62 P63 P64 P65 P66
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P67 P68 P69 P70 P71 P72 P73 P74 P75 P76 P77 P78 P79 P80 P81 P82 P83
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
```

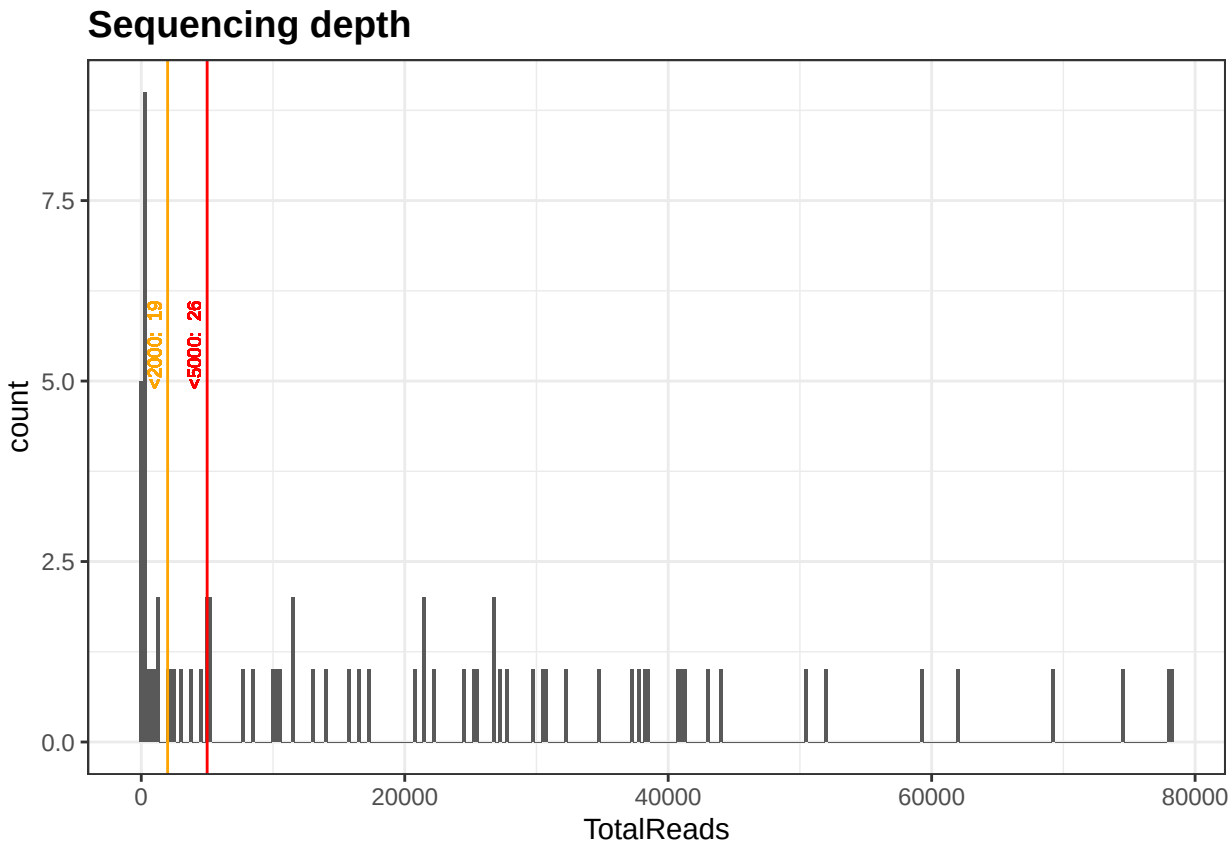
```
length(unique(sample_data(ps_samp_m_Operation_site)$Patient_ID))
```

```
## [1] 37
```

Seq depth

```
sdt = data.table::data.table(as(sample_data(ps_samp_m_Operation_site),
    "data.frame"), TotalReads = sample_sums(ps_samp_m_Operation_site),
    keep.rownames = TRUE)
data.table::setnames(sdt, "rn", "SampleID")
pSeqDepth = ggplot(sdt, aes(TotalReads)) + geom_histogram(binwidth = 250) +
    geom_vline(xintercept = 5000, color = "red") + geom_vline(xintercept = 2000,
    color = "orange") + geom_text(aes(x = 1000, label = paste("<2000: ",
    nrow(sdt[sdt$TotalReads < 2000])), y = 5.5), colour = "orange",
    angle = 90, size = 2.5) + geom_text(aes(x = 4000, label = paste("<5000: ",
    nrow(sdt[sdt$TotalReads < 5000])), y = 5.5), colour = "red",
    angle = 90, size = 2.5) + ggtitle("Sequencing depth") +
    theme(plot.title = element_text(size = 14, face = "bold"))
```

```
pSeqDepth
```

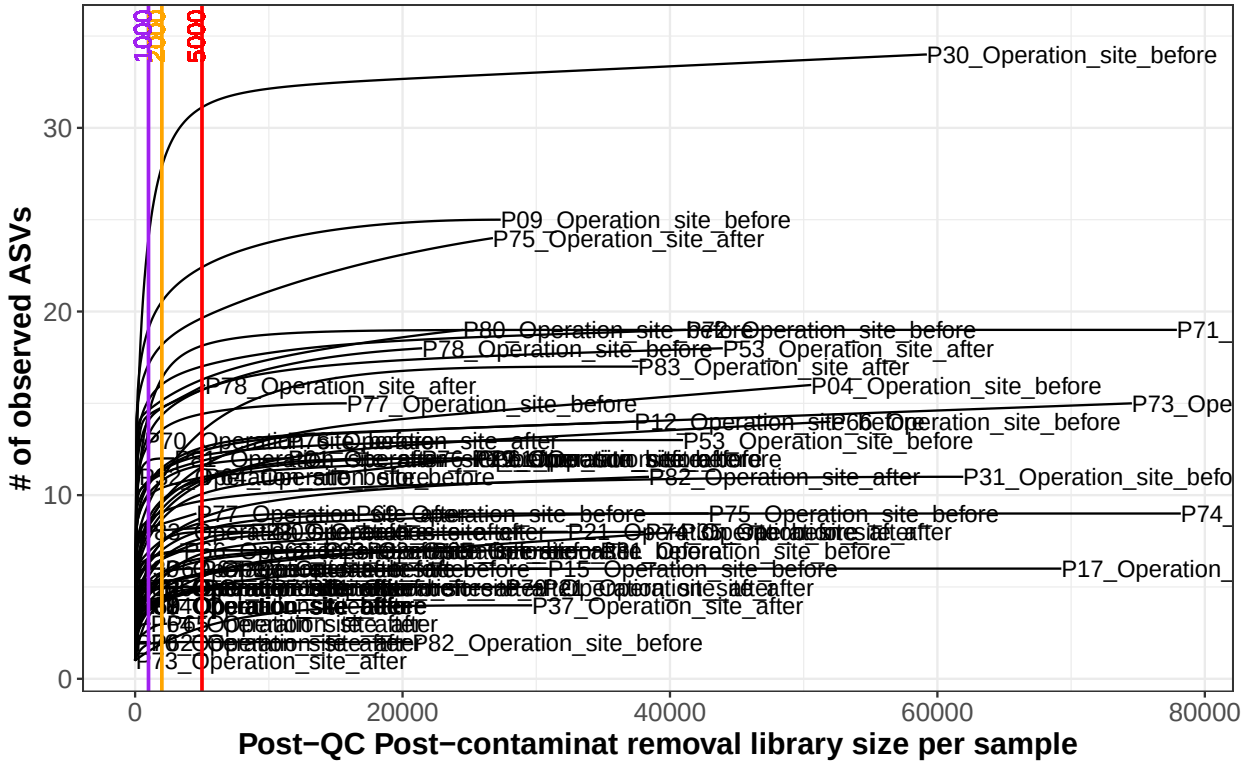


Do the rarefaction curves justify that we remove samples with reads <2000?

Rarefaction curves

```
p200 <- p200 + theme(panel.background = element_blank(),
    axis.title.x = element_text(size = 14, face = "bold"),
```

```
    axis.title.y = element_text(size = 14, face = "bold"),
    axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12),
    legend.title = element_text(size = 16, face = "bold"),
    legend.text = element_text(size = 16), strip.text.x = element_text(angle = 0,
        face = "bold", size = 12), strip.background = element_rect(fill = "white")) +
    xlab("Post-QC Post-contaminat removal library size per sample") +
    ylab("# of observed ASVs") + geom_vline(xintercept = 5000,
    color = "red", size = 0.8) + geom_vline(xintercept = 2000,
    color = "orange", size = 0.8) + geom_vline(xintercept = 1000,
    color = "purple", size = 0.8) + geom_text(aes(x = 4550,
    label = "5000", y = 35), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 35), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 35), colour = "purple",
        angle = 90, size = 4)
p200
```

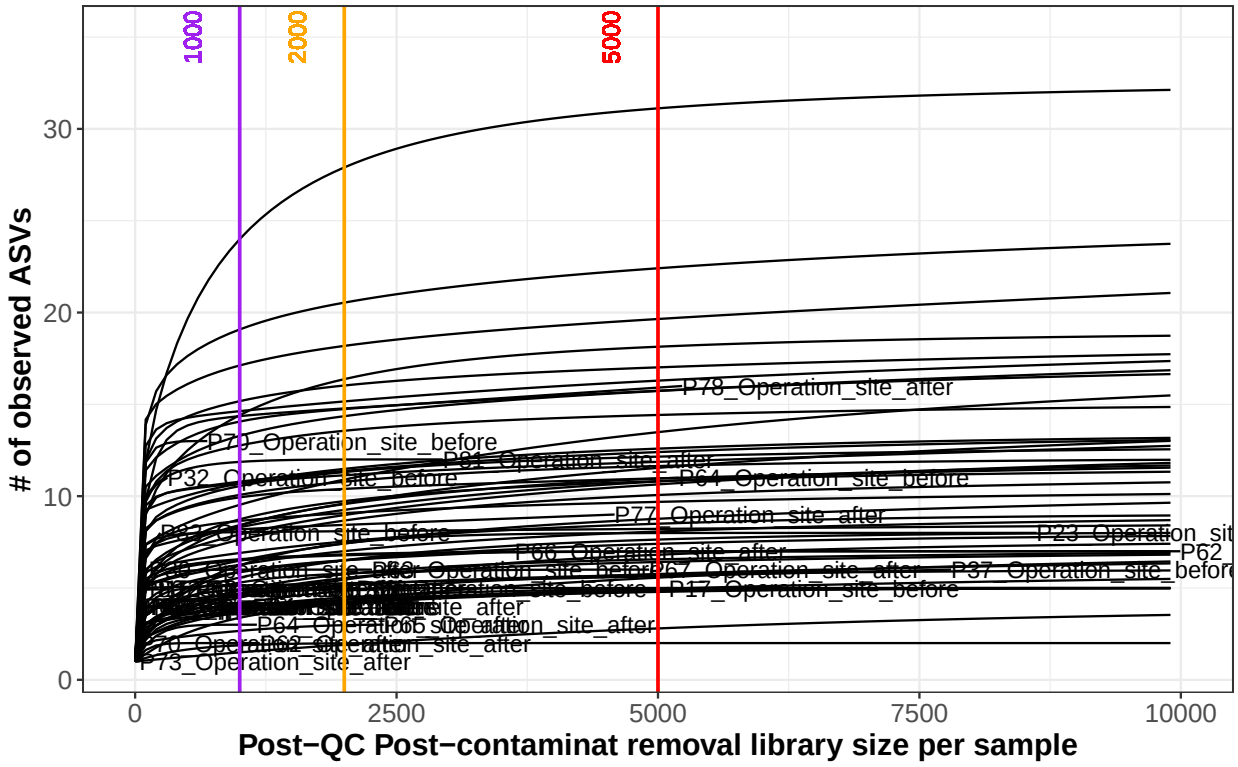


Zoom

```
p200 + xlim(0, 10000)
```

```
## Warning: Removed 42 rows containing missing values (geom_text).
```

```
## Warning: Removed 10189 row(s) containing missing values (geom_path).
```



How do the rarefaction curves look on species level?

Rarefaction curves

```
ps_samp_m_Operation_site_gen <- tax_glom(ps_samp_m_Operation_site,
    taxrank = "Genus_Species")
ps_samp_m_Operation_site_gen
```

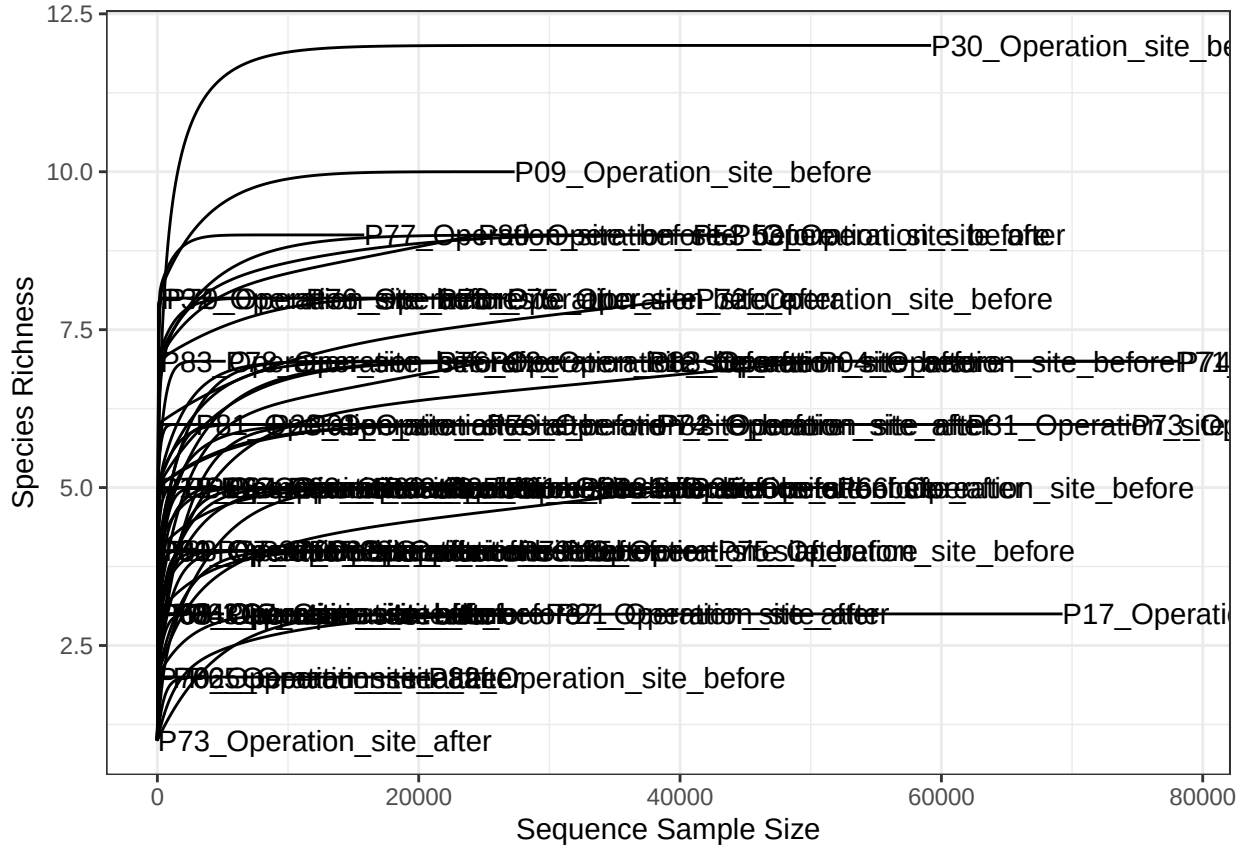
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 24 taxa and 74 samples ]
## sample_data() Sample Data:       [ 74 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 24 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 24 reference sequences ]
```

```
ps_samp_m_Operation_site_gen <- prune_taxa(taxa_sums(ps_samp_m_Operation_site_gen) !=
    0, ps_samp_m_Operation_site_gen)
set.seed(123)
p300 <- ggrare(ps_samp_m_Operation_site_gen, step = 100,
    se = FALSE, label = "Sample")
```

```
## rarefying sample P61_Operation_site_before
## rarefying sample P61_Operation_site_after
## rarefying sample P70_Operation_site_before
## rarefying sample P70_Operation_site_after
## rarefying sample P71_Operation_site_before
## rarefying sample P71_Operation_site_after
```

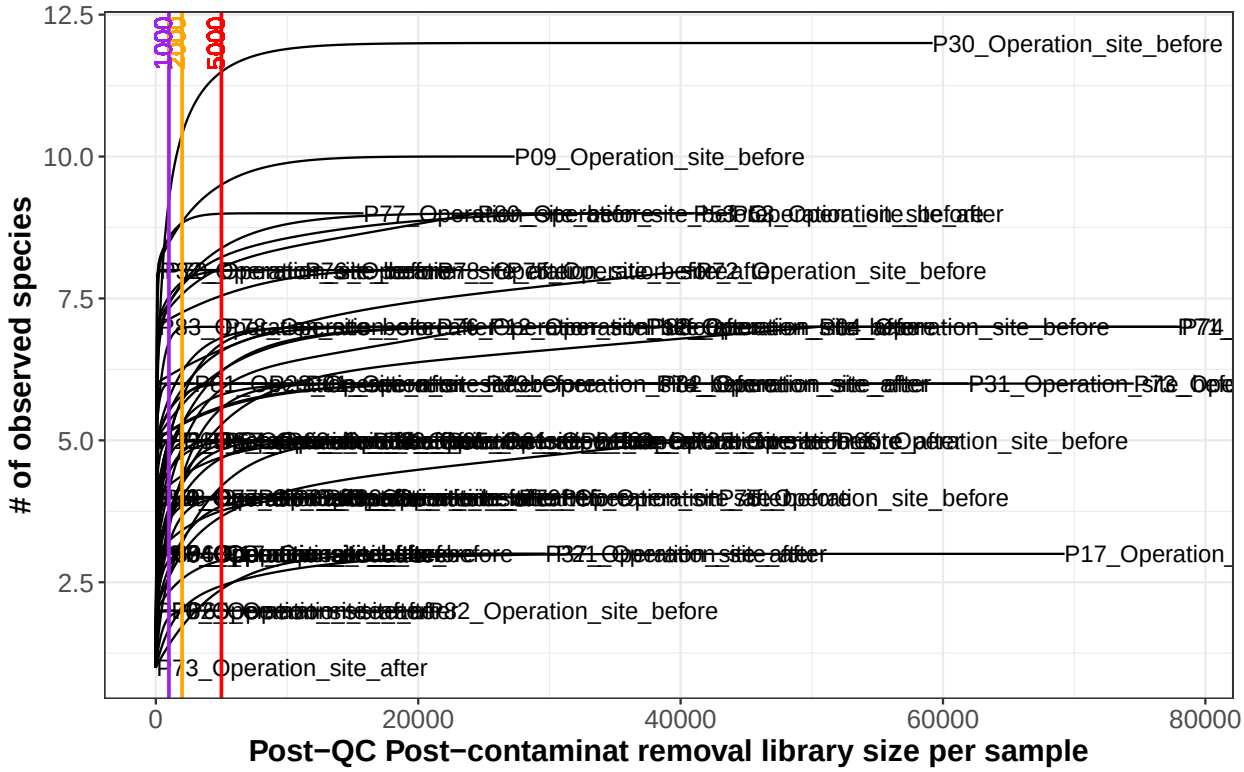
```
## rarefying sample P72_Operation_site_before
## rarefying sample P72_Operation_site_after
## rarefying sample P73_Operation_site_before
## rarefying sample P73_Operation_site_after
## rarefying sample P17_Operation_site_before
## rarefying sample P12_Operation_site_before
## rarefying sample P04_Operation_site_before
## rarefying sample P15_Operation_site_before
## rarefying sample P01_Operation_site_before
## rarefying sample P09_Operation_site_before
## rarefying sample P04_Operation_site_after
## rarefying sample P15_Operation_site_after
## rarefying sample P17_Operation_site_after
## rarefying sample P09_Operation_site_after
## rarefying sample P30_Operation_site_before
## rarefying sample P12_Operation_site_after
## rarefying sample P31_Operation_site_before
## rarefying sample P31_Operation_site_after
## rarefying sample P30_Operation_site_after
## rarefying sample P01_Operation_site_after
## rarefying sample P21_Operation_site_before
## rarefying sample P23_Operation_site_before
## rarefying sample P23_Operation_site_after
## rarefying sample P21_Operation_site_after
## rarefying sample P32_Operation_site_before
## rarefying sample P35_Operation_site_before
## rarefying sample P35_Operation_site_after
## rarefying sample P37_Operation_site_before
## rarefying sample P37_Operation_site_after
## rarefying sample P32_Operation_site_after
## rarefying sample P53_Operation_site_before
## rarefying sample P53_Operation_site_after
## rarefying sample P74_Operation_site_before
## rarefying sample P74_Operation_site_after
## rarefying sample P75_Operation_site_before
## rarefying sample P75_Operation_site_after
## rarefying sample P76_Operation_site_before
## rarefying sample P76_Operation_site_after
## rarefying sample P77_Operation_site_before
## rarefying sample P77_Operation_site_after
## rarefying sample P78_Operation_site_before
## rarefying sample P78_Operation_site_after
## rarefying sample P62_Operation_site_before
## rarefying sample P62_Operation_site_after
## rarefying sample P79_Operation_site_before
## rarefying sample P79_Operation_site_after
## rarefying sample P80_Operation_site_before
## rarefying sample P80_Operation_site_after
## rarefying sample P81_Operation_site_before
## rarefying sample P81_Operation_site_after
## rarefying sample P82_Operation_site_before
## rarefying sample P82_Operation_site_after
## rarefying sample P83_Operation_site_before
## rarefying sample P83_Operation_site_after
```

```
## rarefying sample P63_Operation_site_before
## rarefying sample P63_Operation_site_after
## rarefying sample P64_Operation_site_before
## rarefying sample P64_Operation_site_after
## rarefying sample P65_Operation_site_before
## rarefying sample P65_Operation_site_after
## rarefying sample P66_Operation_site_before
## rarefying sample P66_Operation_site_after
## rarefying sample P67_Operation_site_before
## rarefying sample P67_Operation_site_after
## rarefying sample P68_Operation_site_before
## rarefying sample P68_Operation_site_after
## rarefying sample P69_Operation_site_before
## rarefying sample P69_Operation_site_after
```




```
p300 <- p300 + theme(panel.background = element_blank(),
    axis.title.x = element_text(size = 14, face = "bold"),
    axis.title.y = element_text(size = 14, face = "bold"),
    axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12),
    legend.title = element_text(size = 16, face = "bold"),
    legend.text = element_text(size = 16), strip.text.x = element_text(angle = 0,
        face = "bold", size = 12), strip.background = element_rect(fill = "white")) +
    xlab("Post-QC Post-contaminat removal library size per sample") +
    ylab("# of observed species") + geom_vline(xintercept = 5000,
    color = "red", size = 0.8) + geom_vline(xintercept = 2000,
```

```
    color = "orange", size = 0.8) + geom_vline(xintercept = 1000,
    color = "purple", size = 0.8) + geom_text(aes(x = 4550,
    label = "5000", y = 12), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 12), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 12), colour = "purple",
        angle = 90, size = 4)
p300
```

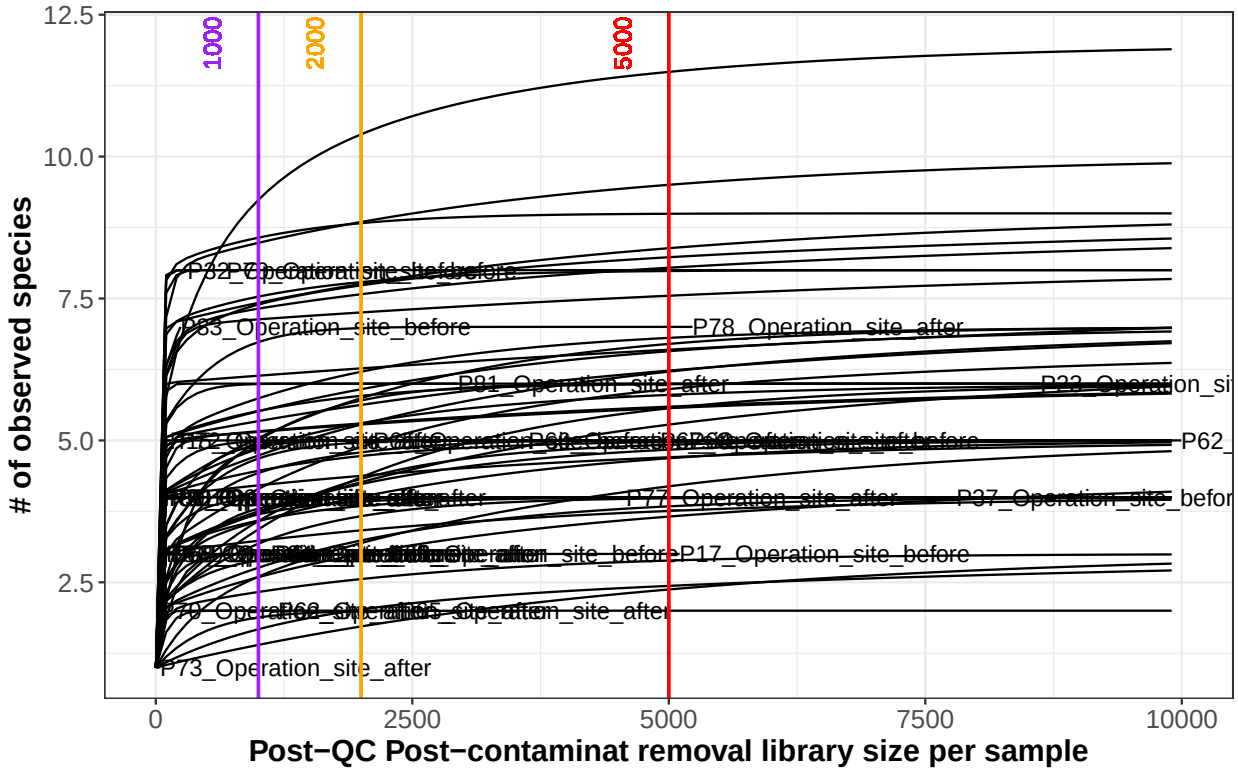


Zoom

```
p300 + xlim(0, 10000)
```

```
## Warning: Removed 42 rows containing missing values (geom_text).
```

```
## Warning: Removed 10189 row(s) containing missing values (geom_path).
```



Exclude samples with <2000 counts

```
summary(sample_sums(ps_samp_m_Operation_site))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      34    1424   13469   20328   31988   78206
```

```
ps_samp_m_Operation_site_tu <- prune_samples(!sample_sums(ps_samp_m_Operation_site) <
    2000, ps_samp_m_Operation_site)
ps_samp_m_Operation_site_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 164 taxa and 55 samples ]
## sample_data() Sample Data:       [ 55 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 164 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 164 reference sequences ]
```

```
summary(sample_sums(ps_samp_m_Operation_site_tu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2118   10312   25186   27226   38302   78206
```

Now which patients have only one time point left?

```
table(sample_data(ps_samp_m_Operation_site_tu)$Patient_ID)
```

```
##
## P01 P04 P09 P12 P15 P17 P21 P23 P30 P31 P32 P35 P37 P53 P61 P62 P63 P64 P65 P66
##   1   1   2   2   1   2   2   2   1   2   1   2   2   2   1   1   1   1   2   2
## P67 P69 P71 P72 P73 P74 P75 P76 P77 P78 P79 P80 P81 P82 P83
##   1   1   2   1   1   2   2   2   2   2   2   1   2   2   1
```

```
length(unique(sample_data(ps_samp_m_Operation_site_tu)$Patient_ID))
```

```
## [1] 35
```

```
ps_samp_m_Operation_site_tu <- prune_samples(!sample_data(ps_samp_m_Operation_site_tu)$Patient_ID %in%
    c("P01", "P04", "P15", "P30", "P32", "P61", "P62", "P63",
        "P64", "P67", "P69", "P72", "P73", "P80", "P83"),
    ps_samp_m_Operation_site_tu)
ps_samp_m_Operation_site_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 164 taxa and 40 samples ]
## sample_data() Sample Data:       [ 40 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 164 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 164 reference sequences ]
```

```
length(unique(sample_data(ps_samp_m_Operation_site_tu)$Patient_ID))
```

```
## [1] 20
```

20 patients left with 2 time points

Read count after removing samples <2000 and patients with only 1 time point:

```
print("OP site 20 paients")
```

```
## [1] "OP site 20 paients"
```

```
summary(sample_sums(ps_samp_m_Operation_site_tu))
```

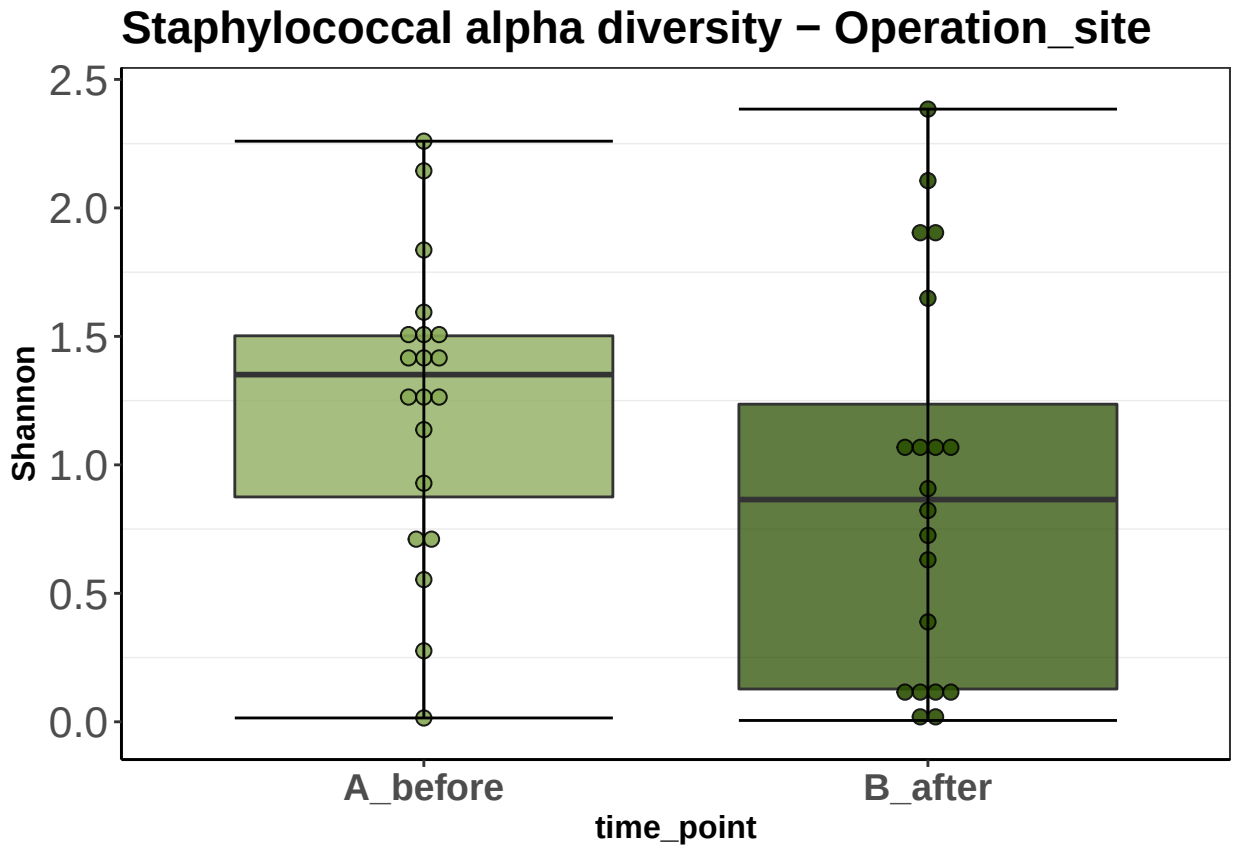
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2118   10387   25296   27114   38238   78206
```

Alpha diversity

```
#### Add diversity measures to the phyloseq object as
#### variables
alpha_div_raw <- estimate_richness(ps_samp_m_Operation_site_tu,
    measures = c("Observed", "Chao1", "Shannon", "InvSimpson"))
rownames(alpha_div_raw) <- gsub("X", "", rownames(alpha_div_raw))
ps_samp_m_Operation_site_tu <- merge_phyloseq(ps_samp_m_Operation_site_tu,
    sample_data(alpha_div_raw))
df_ps_samp_m_Operation_site_tu <- as(sample_data(ps_samp_m_Operation_site_tu),
    "data.frame")
```

Shannon diversity over time:

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```

Paired Wilcoxon signed rank test

```
df_ps_samp_m_Operation_site_tu_c <- dcast(df_ps_samp_m_Operation_site_tu,
    Patient_ID ~ time_point, value.var = "Shannon", drop = FALSE)
wilcox.test(df_ps_samp_m_Operation_site_tu_c$A_before, df_ps_samp_m_Operation_site_tu_c$B_after,
    paired = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  df_ps_samp_m_Operation_site_tu_c$A_before and df_ps_samp_m_Operation_site_tu_c$B_after
## V = 148, p-value = 0.114
## alternative hypothesis: true location shift is not equal to 0
```

Staphyloccocal alpha diversity at the Operation_site does not decrease significantly.

Agglomerate on species level

```
ps_samp_m_Operation_site_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 164 taxa and 40 samples ]
## sample_data() Sample Data:       [ 40 samples by 8 sample variables ]
## tax_table()   Taxonomy Table:    [ 164 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 164 reference sequences ]
```

```
ps_samp_m_Operation_site_tu_gs <- tax_glom(ps_samp_m_Operation_site_tu,
    taxrank = "Genus_Species")
ps_samp_m_Operation_site_tu_gs <- prune_taxa(taxa_sums(ps_samp_m_Operation_site_tu_gs) !=
    0, ps_samp_m_Operation_site_tu_gs)
ps_samp_m_Operation_site_tu_gs
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 17 taxa and 40 samples ]
## sample_data() Sample Data:       [ 40 samples by 8 sample variables ]
## tax_table()   Taxonomy Table:    [ 17 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 17 reference sequences ]
```

Convert to relative abundance

```
ps_samp_m_Operation_site_tu_gs_rel = transform_sample_counts(ps_samp_m_Operation_site_tu_gs,
    function(x) x/sum(x))
summary(sample_sums(ps_samp_m_Operation_site_tu_gs_rel))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       1       1       1       1
```

Subset top 10 species

```
Species10 = names(sort(taxa_sums(ps_samp_m_Operation_site_tu_gs_rel),
    TRUE)[1:10])
```

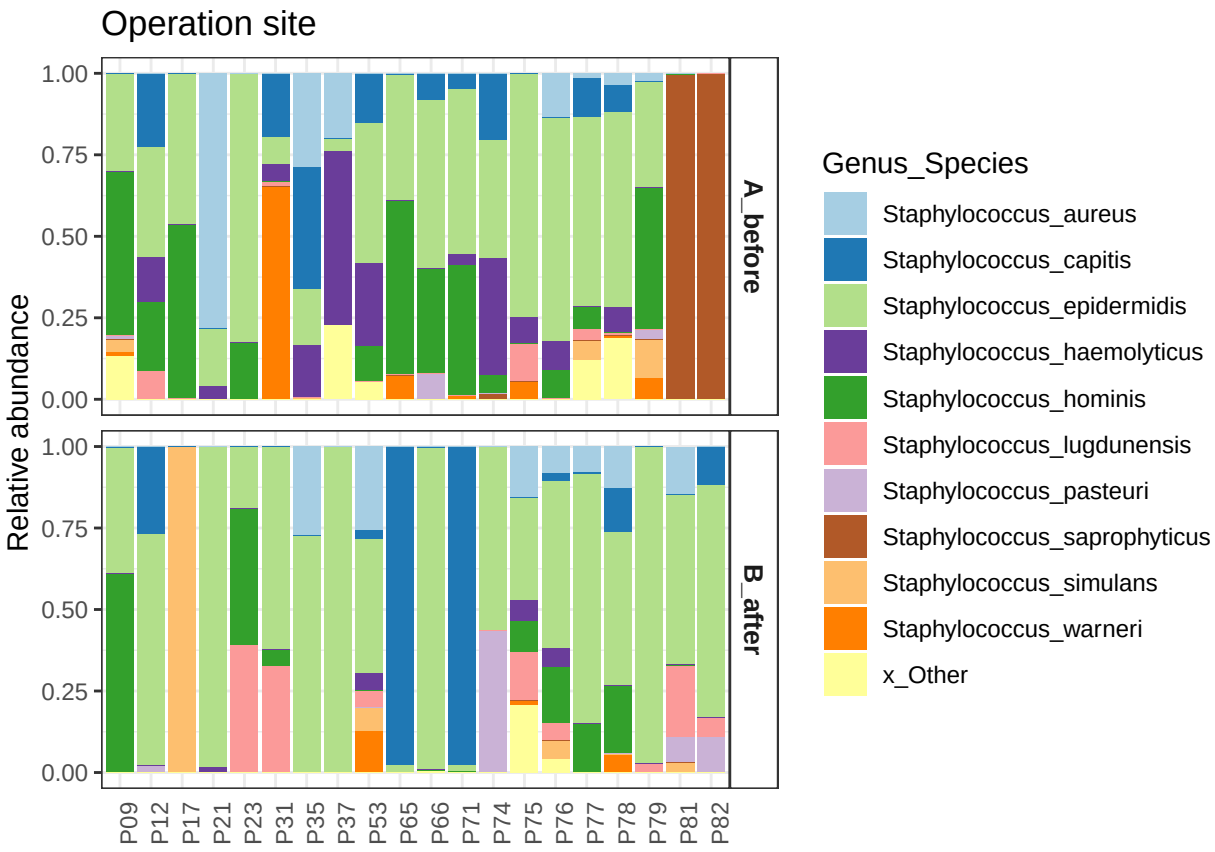
to data frame

```
p_df_o_op <- psmelt(ps_samp_m_Operation_site_tu_gs_rel)
p_df_o_op$Genus_Species <- as.character(p_df_o_op$Genus_Species)
p_df_o_op$Genus_Species[!(p_df_o_op$OTU %in% Species10)] <- "x_Other"
```

Barplots of relative abundance

The patients are not in the same order here as in the heatmap, because in the heatmap they are ordered by clustering and here just by number (see ordered version below).

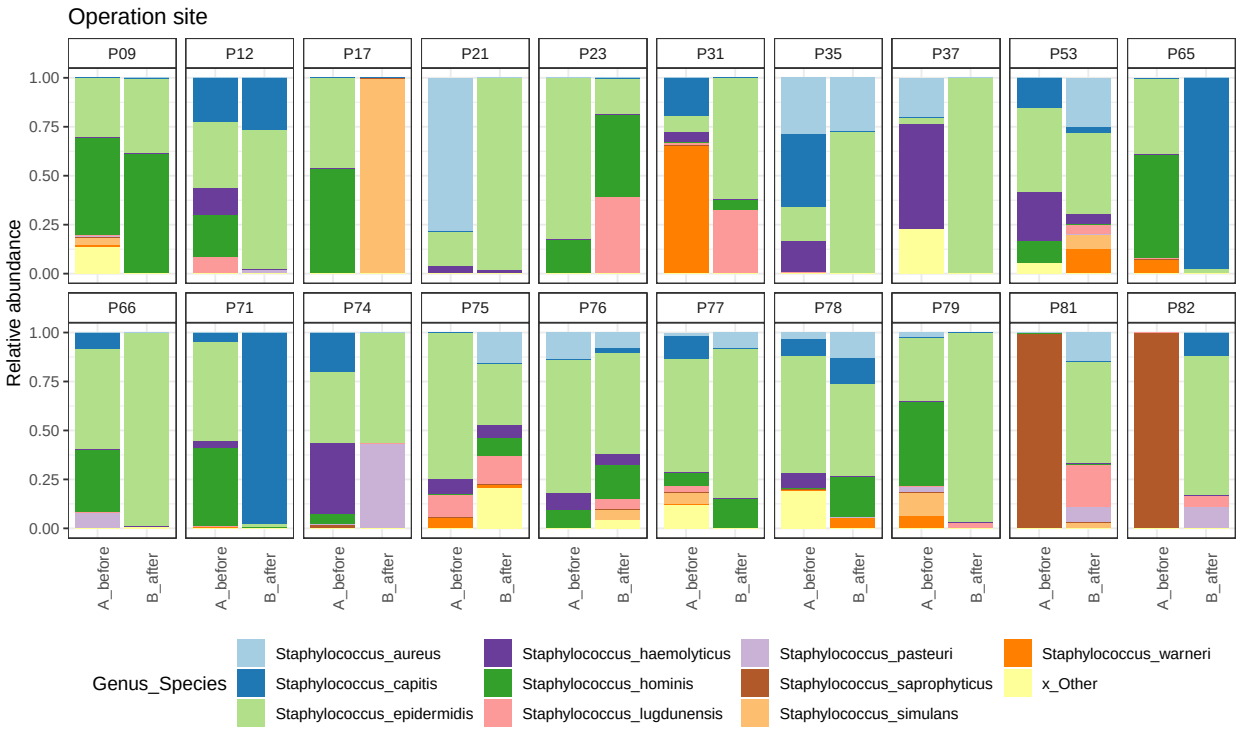
```
op <- ggplot(p_df_o_op, aes(x = Patient_ID, y = Abundance,
    fill = Genus_Species)) + geom_bar(stat = "identity",
    width = 0.9) + facet_grid(time_point ~ ., scales = "free") +
    scale_fill_manual(values = staph_col) + theme(axis.title.x = element_blank(),
    axis.ticks.x = element_blank(), axis.text.x = element_text(angle = 90),
    strip.background = element_rect(fill = "white"), strip.text.y = element_text(size = 10,
        face = "bold")) + ylab("Relative abundance") + ggtitle("Operation site")
op
```




```
ggsave(filename = "plots/OP_bars_tuf.pdf", plot = op, device = cairo_pdf,
    width = 297, height = 210, units = "mm")
```

Patient-wise plots

```
ggplot(p_df_o_op, aes(x = time_point, y = Abundance, fill = Genus_Species)) +
    geom_bar(stat = "identity", width = 0.9) + facet_wrap(. ~
    Patient_ID, nrow = 2) + scale_fill_manual(values = staph_col) +
    theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 90), strip.background = element_rect(fill = "white"),
        strip.text.y = element_text(size = 10, face = "bold"),
        legend.position = "bottom") + ylab("Relative abundance") +
    ggtitle("Operation site")
```



Subset the top 10 genera (without other)

```
ps_samp_m_Operation_site_tu_gs_rel <- prune_taxa(taxa_names(ps_samp_m_Operation_site_tu_gs_rel) %in%
    Species10, ps_samp_m_Operation_site_tu_gs_rel)
ps_samp_m_Operation_site_tu_gs_rel
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 10 taxa and 40 samples ]
## sample_data() Sample Data:       [ 40 samples by 8 sample variables ]
## tax_table()   Taxonomy Table:    [ 10 taxa by 2 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 10 reference sequences ]
```

To data frame

```
p_df <- psmelt(ps_samp_m_Operation_site_tu_gs_rel)
p_df_d <- dcast(p_df, Patient_ID + Genus_Species ~ time_point,
    value.var = "Abundance", drop = FALSE)
```

Calculate relative change in each patient for each species

```
p_df_d <- p_df_d %>% mutate(Percent_point_change = B_after -
    A_before)
p_df_d$Percent_point_change <- p_df_d$Percent_point_change *
    100
```

To matrix

```
p_df_d_m <- acast(p_df_d[, c(1, 2, 5)], Genus_Species ~
    Patient_ID, value.var = "Percent_point_change")
```

Visualize in a heatmap

```
pdf(file = "plots/OP_heatmap_tuf.pdf", width = 11.69, height = 8.27)

heatmap.2(p_df_d_m, scale = "none", col = bluered(100),
    trace = "none", density.info = "histogram", margin = c(6,
        15), cexRow = 1, cexCol = 0.75, adjCol = 1, key.xlab = "Relative abundance change \nin percent p
    keysize = 0.7, key.title = NA, main = "OPERATION SITE")

dev.off()
```

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

Which of the top 10 Staph species do significantly change from before to after?

(Paired Wilcoxon test)

```
wilc_df <- p_df_d %>% group_by(Genus_Species) %>% summarise(wilcox_p_value = wilcox.test(A_before,
    B_after, paired = TRUE)$p.value)
```

```
## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes
```

```
## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## Warning in wilcox.test.default(A_before, B_after, paired = TRUE): cannot compute
## exact p-value with zeroes

## `summarise()` ungrouping output (override with `.groups` argument)
```

```
wilc_df$BH_adjusted_wilcox_p_value <- p.adjust(wilc_df$wilcox_p_value,
    method = "BH")

wilc_df
```

```
## # A tibble: 10 x 3
##    Genus_Species                wilcox_p_value BH_adjusted_wilcox_p_value
##    <chr>                                 <dbl>                      <dbl>
##  1 Staphylococcus_aureus                0.601                      0.752
##  2 Staphylococcus_capitis               1                          1
##  3 Staphylococcus_epidermidis           0.114                      0.380
##  4 Staphylococcus_haemolyticus          0.00386                    0.0386
##  5 Staphylococcus_hominis               0.384                      0.638
##  6 Staphylococcus_lugdunensis           0.0826                     0.380
##  7 Staphylococcus_pasteuri              0.447                      0.638
##  8 Staphylococcus_saprophyticus         0.402                      0.638
##  9 Staphylococcus_simulans              0.969                      1
## 10 Staphylococcus_warneri               0.168                      0.420
```

S. haemolyticus has a significant *overall* change at the OP site, also after multiple testing correction.

Do they *overall* decrease or increase?

```
p_df_d %>% group_by(Genus_Species) %>% summarise(Mean_percent_point_change = mean(B_after) -
    mean(A_before))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

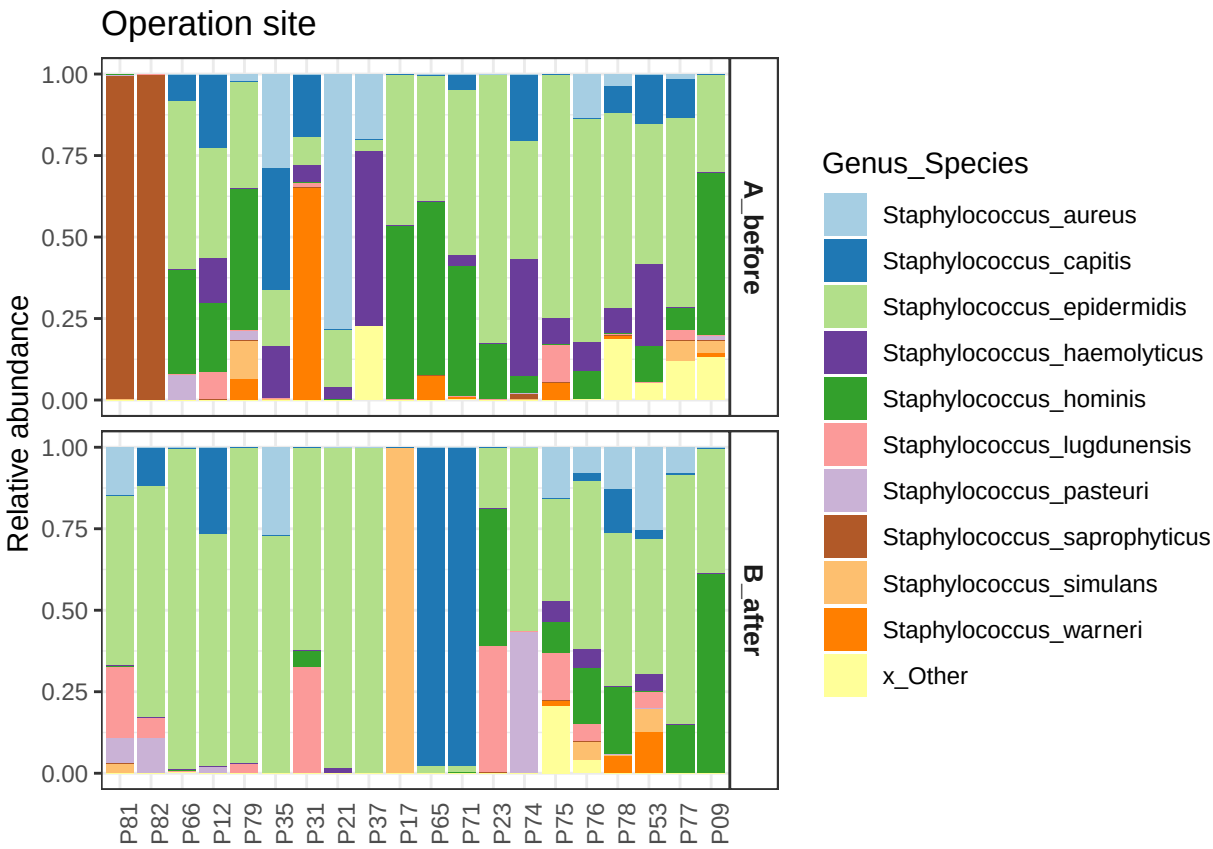
```
## # A tibble: 10 x 2
##    Genus_Species                Mean_percent_point_change
##    <chr>                                            <dbl>
##  1 Staphylococcus_aureus                          -0.0181
##  2 Staphylococcus_capitis                          0.0525
```

```
##  3 Staphylococcus_epidermidis                0.167
##  4 Staphylococcus_haemolyticus              -0.0817
##  5 Staphylococcus_hominis                   -0.0857
##  6 Staphylococcus_lugdunensis                0.0510
##  7 Staphylococcus_pasteuri                   0.0260
##  8 Staphylococcus_saprophyticus             -0.101
##  9 Staphylococcus_simulans                   0.0467
## 10 Staphylococcus_warneri                   -0.0341
```

S. haemolyticus decreases after treatment.

Make a version of the barplots with the same order of patients as in the heatmap

```
positions <- rownames(hmm$carpet)

np1 <- ggplot(p_df_o_op, aes(x = Patient_ID, y = Abundance,
    fill = Genus_Species)) + geom_bar(stat = "identity",
    width = 0.9) + facet_grid(time_point ~ ., scales = "free") +
    scale_fill_manual(values = staph_col) + theme(axis.title.x = element_blank(),
    axis.ticks.x = element_blank(), axis.text.x = element_text(angle = 90),
    strip.background = element_rect(fill = "white"), strip.text.y = element_text(size = 10,
        face = "bold")) + ylab("Relative abundance") + ggtitle("Operation site") +
    scale_x_discrete(limits = positions)

np1
```



```
ggsave(filename = "plots/OP_site_bars_tuf_ordered_IDs.pdf",
    plot = np1, device = cairo_pdf, width = 297, height = 210,
    units = "mm")
```

PCoA of Nose, Groin and OP site, before vs after surgery

```
ps_n_g <- merge_phyloseq(ps_samp_m_nose_tu, ps_samp_m_groin_tu,
    ps_samp_m_Operation_site_tu)
sample_data(ps_n_g)$group_tp <- paste(sample_data(ps_n_g)$Sample_type,
    sample_data(ps_n_g)$time_point)
```

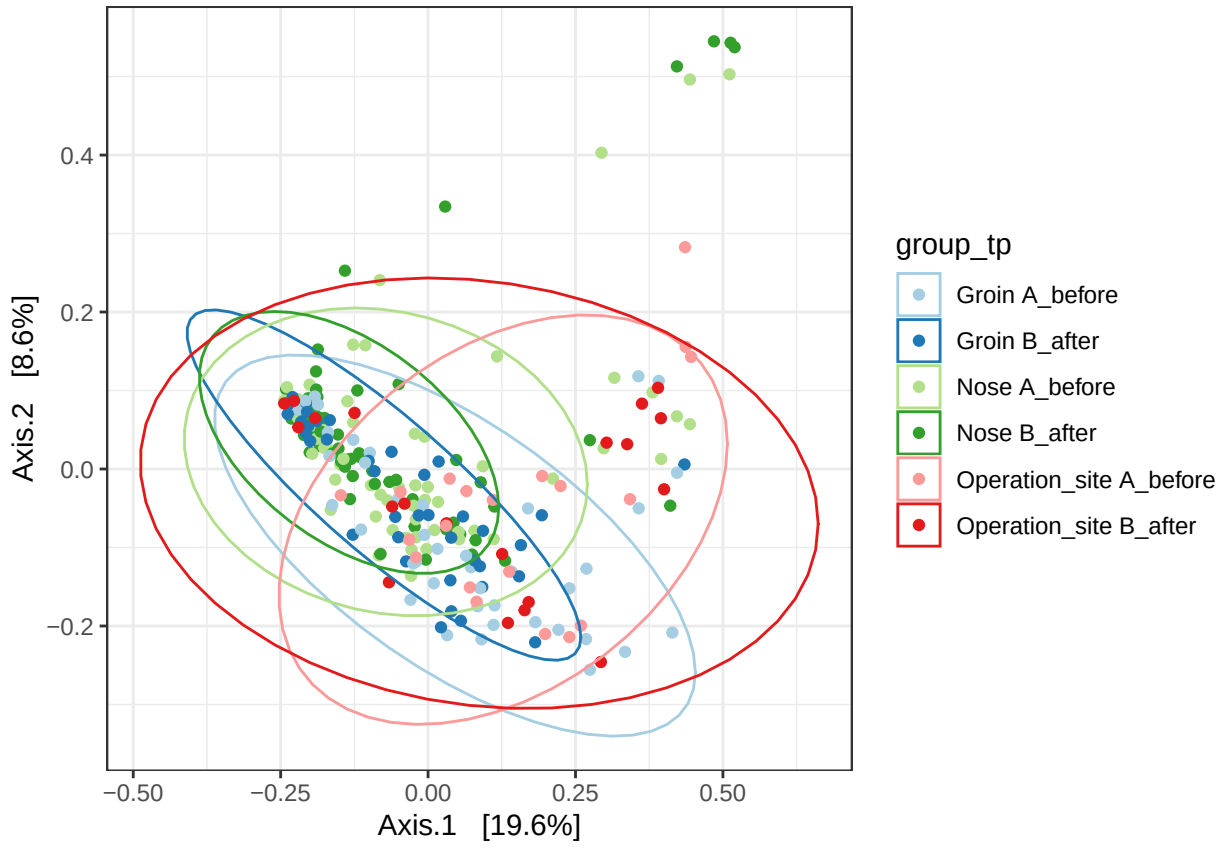
Hellinger transform before ordination

```
ps_n_g_hell <- transform_sample_counts(ps_n_g, function(x) sqrt(x/sum(x)))  #hellinger transform

ps_n_g_ord <- ordinate(ps_n_g_hell, method = "PCoA", distance = "bray")
```

PCoA

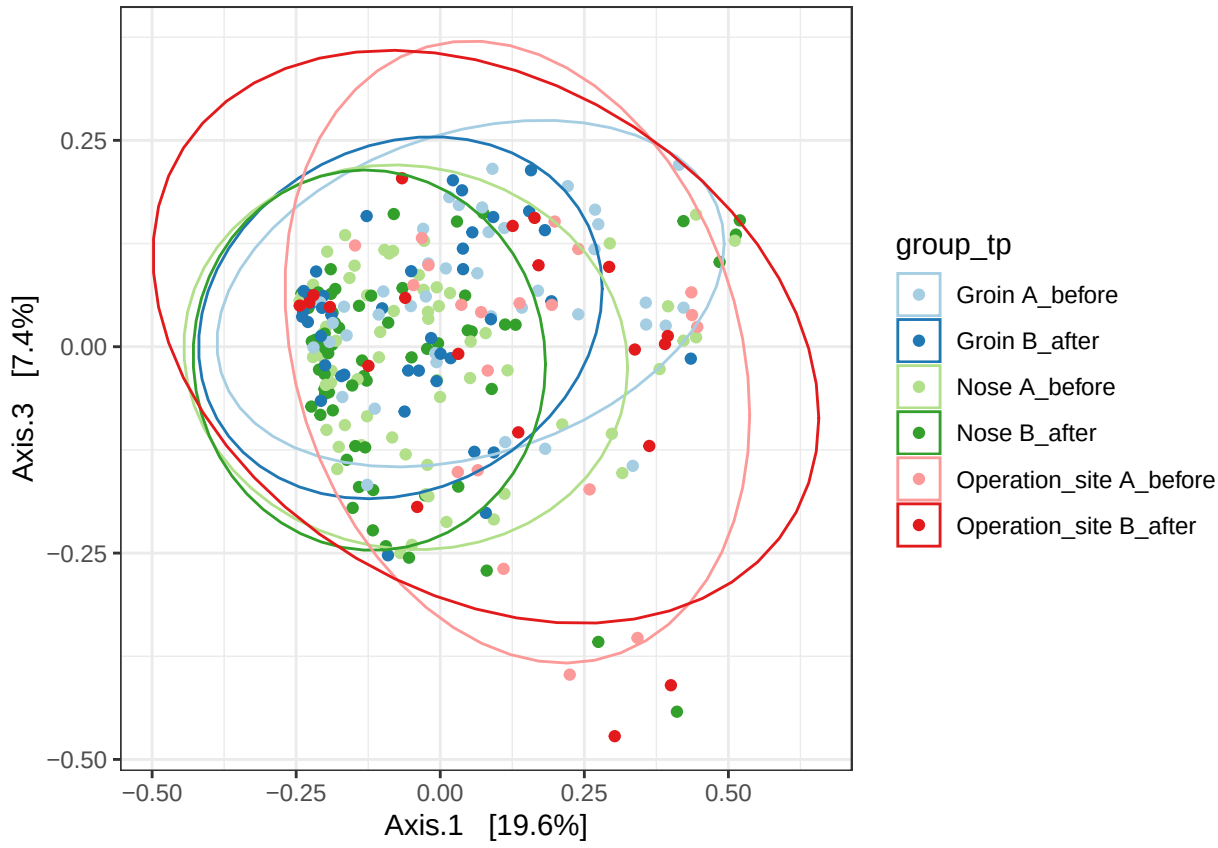
PCoA - Axis 1,2

```
ord_plot <- plot_ordination(ps_n_g_hell, ps_n_g_ord, type = "samples",
    color = "group_tp", axes = 1:2)
ord_plot + stat_ellipse(geom = "polygon", type = "t", alpha = 0,
    aes(fill = group_tp)) + scale_color_brewer(palette = "Paired",
    type = "div") + scale_fill_brewer(palette = "Paired",
    type = "div")
```



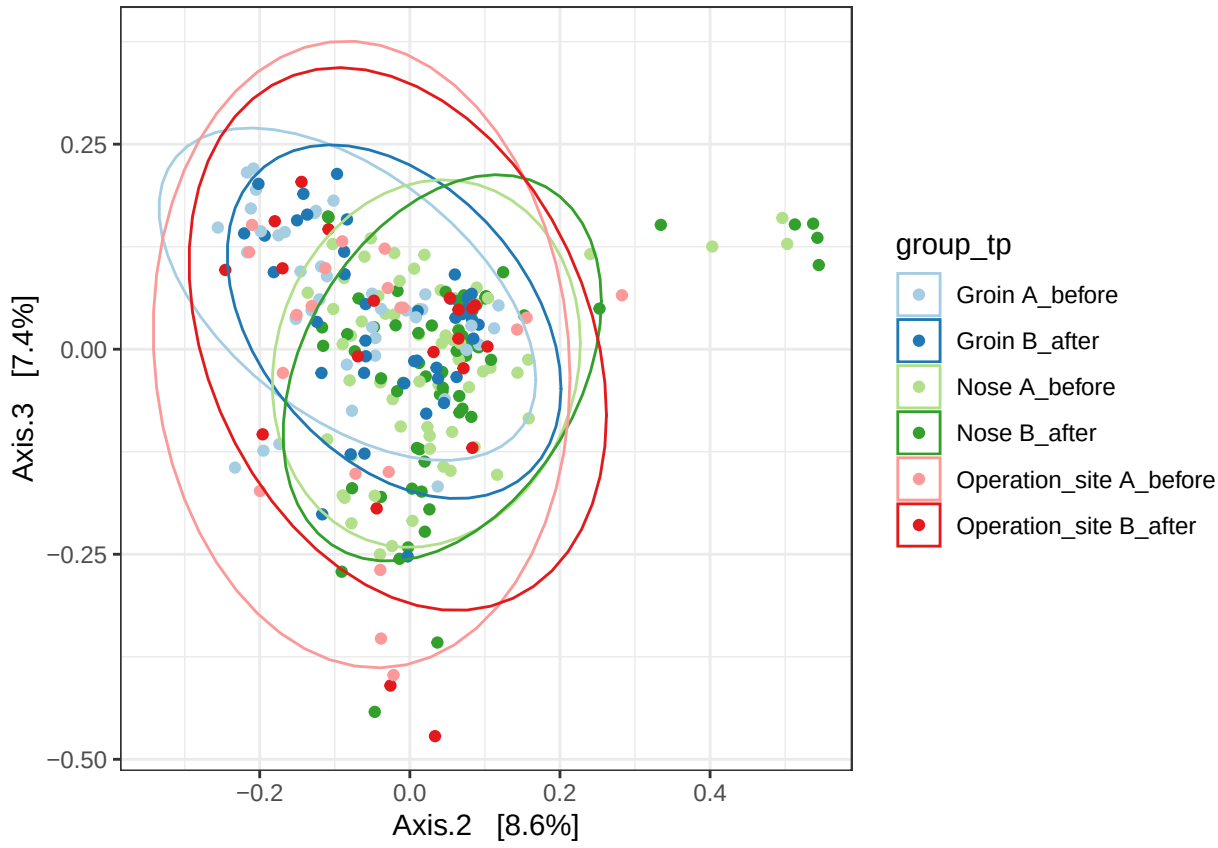
PCoA - Axis 1,3

```
ord_plot <- plot_ordination(ps_n_g_hell, ps_n_g_ord, type = "samples",
    color = "group_tp", axes = c(1, 3))
ord_plot + stat_ellipse(geom = "polygon", type = "t", alpha = 0,
    aes(fill = group_tp)) + scale_color_brewer(palette = "Paired",
    type = "div") + scale_fill_brewer(palette = "Paired",
    type = "div")
```



PCoA - Axis 2,3

```
ord_plot <- plot_ordination(ps_n_g_hell, ps_n_g_ord, type = "samples",
    color = "group_tp", axes = c(2, 3))
ord_plot + stat_ellipse(geom = "polygon", type = "t", alpha = 0,
    aes(fill = group_tp)) + scale_color_brewer(palette = "Paired",
    type = "div") + scale_fill_brewer(palette = "Paired",
    type = "div")
```



Ordinate

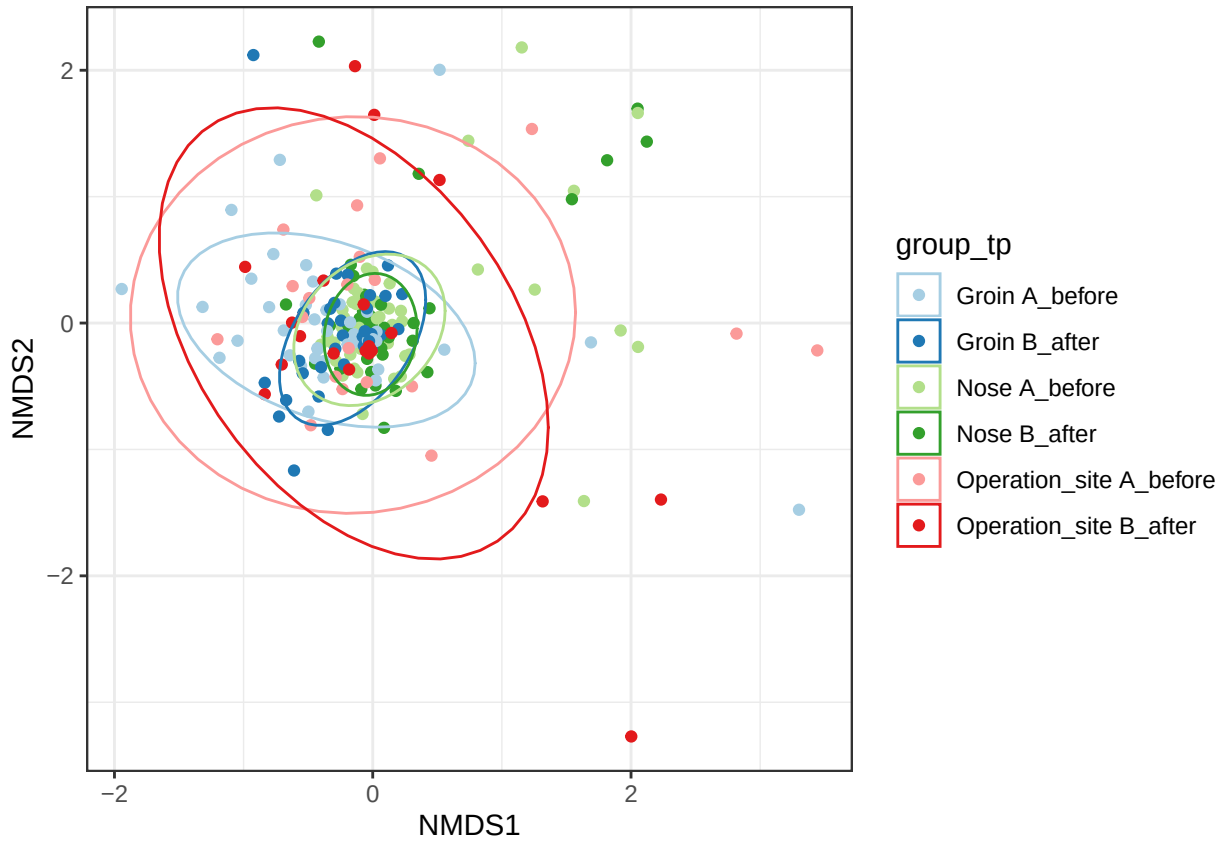
```
ps_n_g_ord <- ordinate(ps_n_g_hell, method = "NMDS", distance = "bray")
```

```
## Run 0 stress 0.2117936
## Run 1 stress 0.2205301
## Run 2 stress 0.210547
## ... New best solution
## ... Procrustes: rmse 0.04813254  max resid 0.3207883
## Run 3 stress 0.2104255
## ... New best solution
## ... Procrustes: rmse 0.05411986  max resid 0.2375853
## Run 4 stress 0.2188744
## Run 5 stress 0.2105311
## ... Procrustes: rmse 0.05079381  max resid 0.2644759
## Run 6 stress 0.2151032
## Run 7 stress 0.2105434
## ... Procrustes: rmse 0.03953748  max resid 0.2961556
## Run 8 stress 0.2122864
## Run 9 stress 0.220615
## Run 10 stress 0.216989
## Run 11 stress 0.2180765
## Run 12 stress 0.213956
## Run 13 stress 0.208735
```

```
## ... New best solution
## ... Procrustes: rmse 0.04559277  max resid 0.2165524
## Run 14 stress 0.2151668
## Run 15 stress 0.2109476
## Run 16 stress 0.2119815
## Run 17 stress 0.2161556
## Run 18 stress 0.2098081
## Run 19 stress 0.2115362
## Run 20 stress 0.2091301
## ... Procrustes: rmse 0.04063366  max resid 0.4248194
## *** No convergence -- monoMDS stopping criteria:
##     12: no. of iterations >= maxit
##      8: stress ratio > sratmax
```

NMDS

```
ord_plot <- plot_ordination(ps_n_g, ps_n_g_ord, type = "samples",
    color = "group_tp")
ord_plot + stat_ellipse(geom = "polygon", type = "t", alpha = 0,
    aes(fill = group_tp)) + scale_color_brewer(palette = "Paired",
    type = "div") + scale_fill_brewer(palette = "Paired",
    type = "div")
```



Are the within group variations homogenous?

```
df <- as(sample_data(ps_n_g_hell), "data.frame")

bray_dist <- phyloseq::distance(ps_n_g_hell, method = "bray")

bo <- betadisper(bray_dist, group = df$group_tp)
anova(bo)
```

```
## Analysis of Variance Table
##
## Response: Distances
##            Df Sum Sq  Mean Sq F value    Pr(>F)
## Groups      5 0.9639 0.192786  6.1424 2.201e-05 ***
## Residuals 246 7.7210 0.031386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No, they are not (p<0.05), therefore the adonis() test needs to be interpreted with caution.

Permutational Multivariate Analysis of Variance Using Distance Matrices (adonis())

Is the bacterial community different depending on group and time point?

```
set.seed(123)
vegan::adonis(bray_dist ~ group_tp, data = df)
```

```
##
## Call:
## vegan::adonis(formula = bray_dist ~ group_tp, data = df)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## group_tp    5     3.606 0.72122  3.7269 0.07042  0.001 ***
## Residuals 246    47.605 0.19352         0.92958
## Total     251    51.211                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P<0.05, therefore the bacterial community is different based on body site and time point.

Is the difference attributable to body site, time point or an interaction of both?

```
set.seed(123)
vegan::adonis2(bray_dist ~ Sample_type + time_point + Sample_type:time_point,
    data = df, strata = Patient_ID:time_point)
```

```
## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## vegan::adonis2(formula = bray_dist ~ Sample_type + time_point + Sample_type:time_point, data = df, s
##                          Df SumOfSqs      R2      F Pr(>F)
## Sample_type               2    2.327 0.04544 6.0124  0.001 ***
## time_point                1    0.778 0.01519 4.0200  0.001 ***
## Sample_type:time_point    2    0.501 0.00979 1.2949  0.141
## Residual                246   47.605 0.92958
## Total                   251   51.211 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The difference is driven by BOTH sample type and time point. BUT as mentioned, the groups have different within group variances and are therefore not really comparable by adonis.

split the data by sample type and then check if there is a difference between time points within groups

Nose

```
ps_n_g_hell_NOSE <- prune_samples(sample_data(ps_n_g_hell)$Sample_type ==
    "Nose", ps_n_g_hell)
ps_n_g_hell_NOSE <- prune_taxa(taxa_sums(ps_n_g_hell_NOSE) !=
    0, ps_n_g_hell_NOSE)

df <- as(sample_data(ps_n_g_hell_NOSE), "data.frame")

bray_dist <- phyloseq::distance(ps_n_g_hell_NOSE, method = "bray")

bo <- betadisper(bray_dist, group = df$time_point)
anova(bo)
```

```
## Analysis of Variance Table
##
## Response: Distances
##            Df Sum Sq  Mean Sq F value Pr(>F)
## Groups      1 0.0928 0.092758  2.2909 0.1326
## Residuals 128 5.1826 0.040489
```

Within group variatons are not different, adonis can be used.

Is the nasal community different depending on time point?

```
set.seed(123)
vegan::adonis(bray_dist ~ time_point, data = df)
```

```
##
## Call:
## vegan::adonis(formula = bray_dist ~ time_point, data = df)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##             Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## time_point   1     0.253 0.25296  1.5373 0.01187  0.125
## Residuals  128    21.062 0.16455         0.98813
## Total      129    21.315                1.00000
```

Nasal community is not different before and after

Groin

```
ps_n_g_hell_GROIN <- prune_samples(sample_data(ps_n_g_hell)$Sample_type ==
    "Groin", ps_n_g_hell)
ps_n_g_hell_GROIN <- prune_taxa(taxa_sums(ps_n_g_hell_GROIN) !=
    0, ps_n_g_hell_GROIN)

df <- as(sample_data(ps_n_g_hell_GROIN), "data.frame")

bray_dist <- phyloseq::distance(ps_n_g_hell_GROIN, method = "bray")

bo <- betadisper(bray_dist, group = df$time_point)
anova(bo)
```

```
## Analysis of Variance Table
##
## Response: Distances
##           Df  Sum Sq  Mean Sq F value  Pr(>F)
## Groups     1 0.14826 0.148256  6.5721 0.01223 *
## Residuals 80 1.80467 0.022558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Within group variatons are different, so interpret adonis with caution.

Is the groin community different depending on time point?

```
set.seed(123)
vegan::adonis(bray_dist ~ time_point, data = df)
```

```
##
## Call:
## vegan::adonis(formula = bray_dist ~ time_point, data = df)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## time_point  1    0.6681 0.66805  3.4335 0.04115  0.001 ***
## Residuals  80   15.5655 0.19457         0.95885
## Total      81   16.2336                1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Groin community is different before and after (but CAVE: different within group variations)

Operation site

```
ps_n_g_hell_Operation_site <- prune_samples(sample_data(ps_n_g_hell)$Sample_type ==
    "Operation_site", ps_n_g_hell)
ps_n_g_hell_Operation_site <- prune_taxa(taxa_sums(ps_n_g_hell_Operation_site) !=
    0, ps_n_g_hell_Operation_site)

df <- as(sample_data(ps_n_g_hell_Operation_site), "data.frame")

bray_dist <- phyloseq::distance(ps_n_g_hell_Operation_site,
    method = "bray")

bo <- betadisper(bray_dist, group = df$time_point)
anova(bo)
```

```
## Analysis of Variance Table
##
## Response: Distances
##           Df  Sum Sq   Mean Sq F value Pr(>F)
## Groups     1 0.00163 0.0016328   0.077 0.7829
## Residuals 38 0.80556 0.0211989
```

Within group variatons are not different.

Is the Operation_site community different depending on time point?

```
set.seed(123)
vegan::adonis(bray_dist ~ time_point, data = df)
```

```
##
## Call:
## vegan::adonis(formula = bray_dist ~ time_point, data = df)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## time_point  1    0.3581 0.35810  1.2397 0.03159  0.223
## Residuals  38   10.9770 0.28887         0.96841
## Total      39   11.3351                1.00000
```

OP site community is not different before and after