Staphylome - 16S rRNA 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 in phyloseq object

```
ps1 <- readRDS("phyloseq_objects_for_publication/phy_obj_16S.RData")
ps1
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1504 taxa and 372 samples ]
## sample_data() Sample Data:       [ 372 samples by 3 sample variables ]
## tax_table()   Taxonomy Table:    [ 1504 taxa by 7 taxonomic ranks ]
```

Summary of read counts per sample

```
print("16S before contaminant removal: ")
```

```
## [1] "16S before contaminant removal: "
```

```
print("All: ")
```

```
## [1] "All: "
```

```
summary(sample_sums(ps1))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      38    4159   11168   15733   22059   86693
```

```
print("Nose: ")
```

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

```
summary(sample_sums(ps1)[sample_data(ps1)$Sample_type ==
    "Nose"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      38    5200   12989   18015   27555   86693
```

```
print("Groin: ")
```

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

```
summary(sample_sums(ps1)[sample_data(ps1)$Sample_type ==
    "Groin"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     377    5663   11847   17376   25868   64556
```

```
print("OP site: ")
```

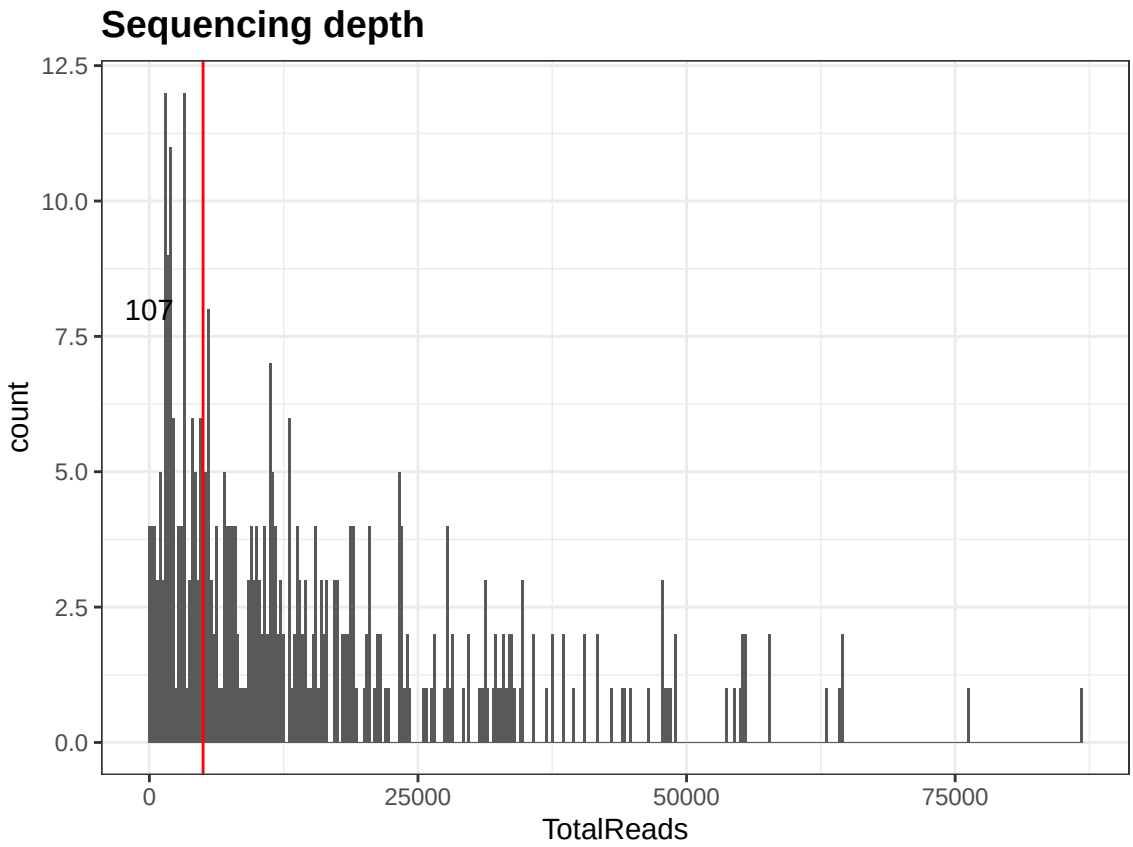
```
## [1] "OP site: "
```

```
summary(sample_sums(ps1)[sample_data(ps1)$Sample_type ==
    "Operation_site"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      44    2106    9666   10076   14884   57837
```

How does the sequencing depth look before decontam?

```
sdt = data.table::data.table(as(sample_data(ps1), "data.frame"),
    TotalReads = sample_sums(ps1), keep.rownames = TRUE)
data.table::setnames(sdt, "rn", "SampleID")
pSeqDepth = ggplot(sdt, aes(TotalReads)) + geom_histogram(binwidth = 250) +
    geom_vline(xintercept = 5000, color = "red") + annotate("text",
    label = nrow(sdt[sdt$TotalReads < 5000]), x = 0, y = 8,
    size = 4, colour = "black") + ggtitle("Sequencing depth") +
    theme(plot.title = element_text(size = 14, face = "bold"))

pSeqDepth
```



107 samples are already now below 5000 reads

DECONTAM

We do this for all body sites together, because the extractions were done together and nose and skin do not differ so much in biomass as e.g. feces would do.

Contaminants identified by prevalence method

Threshold 0.5 removes all contaminant ASVs that are more prevalent in controls than in samples

```
sample_data(ps1)$is.neg <- sample_data(ps1)$Sample_type %in%
    c("Extraction_control")
contamdf.prev <- isContaminant(ps1, method = "prevalence",
    neg = "is.neg", threshold = 0.5)
table(contamdf.prev$contaminant)
```

```
##
## FALSE  TRUE
##  1456    48
```

48 contaminants identified

Who are they?

```
ps.contam <- prune_taxa(contamdf.prev$contaminant, ps1)
ps.contam
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 48 taxa and 372 samples ]
## sample_data() Sample Data:       [ 372 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 48 taxa by 7 taxonomic ranks ]
```

```
tax_table(ps.contam)
```

```
## Taxonomy Table:     [48 taxa by 7 taxonomic ranks]:
##          Kingdom       Phylum               Class
## OTU_13   "d__Bacteria" "p__Actinobacteria"  "c__Actinobacteria"
## OTU_46   "d__Bacteria" "p__Actinobacteria"  "c__Actinobacteria"
## OTU_69   "d__Bacteria" "p__Actinobacteria"  "c__Actinobacteria"
## OTU_142  "d__Bacteria" "p__Actinobacteria"  "c__Actinobacteria"
## OTU_143  "d__Bacteria" "p__Actinobacteria"  "c__Actinobacteria"
## OTU_190  "d__Bacteria" "p__Actinobacteria"  "c__Actinobacteria"
## OTU_401  "d__Bacteria" "p__Bacteroidetes"   "c__Flavobacteriia"
## OTU_420  "d__Bacteria" "p__Bacteroidetes"   "c__Flavobacteriia"
## OTU_427  "d__Bacteria" "p__Bacteroidetes"   "c__Sphingobacteriia"
## OTU_432  "d__Bacteria" "p__Bacteroidetes"   "c__Sphingobacteriia"
## OTU_442  "d__Bacteria" "p__Bacteroidetes"   "c__Sphingobacteriia"
## OTU_456  "d__Bacteria" "p__Bacteroidetes"   "c__Sphingobacteriia"
## OTU_531  "d__Bacteria" "p__Firmicutes"      "c__Bacilli"
## OTU_577  "d__Bacteria" "p__Firmicutes"      "c__Bacilli"
## OTU_634  "d__Bacteria" "p__Firmicutes"      "c__Bacilli"
## OTU_708  "d__Bacteria" "p__Firmicutes"      "c__Clostridia"
## OTU_727  "d__Bacteria" "p__Firmicutes"      "c__Clostridia"
## OTU_763  "d__Bacteria" "p__Firmicutes"      "c__Clostridia"
## OTU_805  "d__Bacteria" "p__Firmicutes"      "c__Clostridia"
## OTU_832  "d__Bacteria" "p__Firmicutes"      "c__Clostridia"
## OTU_833  "d__Bacteria" "p__Firmicutes"      "c__Clostridia"
## OTU_910  "d__Bacteria" "p__Proteobacteria"  "c__Alphaproteobacteria"
## OTU_933  "d__Bacteria" "p__Proteobacteria"  "c__Alphaproteobacteria"
## OTU_961  "d__Bacteria" "p__Proteobacteria"  "c__Alphaproteobacteria"
## OTU_964  "d__Bacteria" "p__Proteobacteria"  "c__Alphaproteobacteria"
## OTU_1034 "d__Bacteria" "p__Proteobacteria"  "c__Alphaproteobacteria"
## OTU_1037 "d__Bacteria" "p__Proteobacteria"  "c__Alphaproteobacteria"
## OTU_1042 "d__Bacteria" "p__Proteobacteria"  "c__Alphaproteobacteria"
## OTU_1055 "d__Bacteria" "p__Proteobacteria"  "c__Alphaproteobacteria"
## OTU_1075 "d__Bacteria" "p__Proteobacteria"  "c__Alphaproteobacteria"
## OTU_1094 "d__Bacteria" "p__Proteobacteria"  "c__Alphaproteobacteria"
## OTU_1112 "d__Bacteria" "p__Proteobacteria"  "c__Betaproteobacteria"
## OTU_1128 "d__Bacteria" "p__Proteobacteria"  "c__Betaproteobacteria"
## OTU_1143 "d__Bacteria" "p__Proteobacteria"  "c__Betaproteobacteria"
## OTU_1165 "d__Bacteria" "p__Proteobacteria"  "c__Betaproteobacteria"
```

```
## OTU_1166 "d__Bacteria" "p__Proteobacteria"  "c__Betaproteobacteria"
## OTU_1173 "d__Bacteria" "p__Proteobacteria"  "c__Betaproteobacteria"
## OTU_1198 "d__Bacteria" "p__Proteobacteria"  "c__Betaproteobacteria"
## OTU_1309 "d__Bacteria" "p__Proteobacteria"  "c__Deltaproteobacteria"
## OTU_1360 "d__Bacteria" "p__Proteobacteria"  "c__Gammaproteobacteria"
## OTU_1386 "d__Bacteria" "p__Proteobacteria"  "c__Gammaproteobacteria"
## OTU_1460 "d__Bacteria" "p__Proteobacteria"  "c__Gammaproteobacteria"
## OTU_1470 "d__Bacteria" "p__Proteobacteria"  "c__Gammaproteobacteria"
## OTU_1474 "d__Bacteria" "p__Proteobacteria"  "c__Gammaproteobacteria"
## OTU_1495 "d__Bacteria" "p__Proteobacteria"  "c__Gammaproteobacteria"
## OTU_1499 "d__Bacteria" "p__Proteobacteria"  "c__Gammaproteobacteria"
## OTU_1507 "d__Bacteria" "p__Proteobacteria"  "c__Gammaproteobacteria"
## OTU_1565 "d__Bacteria" "p__Verrucomicrobia" "c__Opitutae"
##          Order
## OTU_13   "o__Actinomycetales"
## OTU_46   "o__Actinomycetales"
## OTU_69   "o__Actinomycetales"
## OTU_142  "o__Actinomycetales"
## OTU_143  "o__Actinomycetales"
## OTU_190  "o__Actinomycetales"
## OTU_401  "o__Flavobacteriales"
## OTU_420  "o__Flavobacteriales"
## OTU_427  "o__Sphingobacteriales"
## OTU_432  "o__Sphingobacteriales"
## OTU_442  "o__Sphingobacteriales"
## OTU_456  "o__Sphingobacteriales"
## OTU_531  "o__Bacillales"
## OTU_577  "o__Bacillales"
## OTU_634  "o__Lactobacillales"
## OTU_708  "o__Clostridiales"
## OTU_727  "o__Clostridiales"
## OTU_763  "o__Clostridiales"
## OTU_805  "o__Clostridiales"
## OTU_832  "o__Clostridiales"
## OTU_833  "o__Clostridiales"
## OTU_910  "o__Caulobacterales"
## OTU_933  "o__Rhizobiales"
## OTU_961  "o__Rhizobiales"
## OTU_964  "o__Rhizobiales"
## OTU_1034 "o__Rhodobacterales"
## OTU_1037 "o__Rhodobacterales"
## OTU_1042 "o__Rhodospirillales"
## OTU_1055 "o__Rhodospirillales"
## OTU_1075 "o__Sphingomonadales"
## OTU_1094 "o__Sphingomonadales"
## OTU_1112 "o__Burkholderiales"
## OTU_1128 "o__Burkholderiales"
## OTU_1143 "o__Burkholderiales"
## OTU_1165 "o__Burkholderiales"
## OTU_1166 "o__Burkholderiales"
## OTU_1173 "o__Burkholderiales"
## OTU_1198 "o__Burkholderiales"
## OTU_1309 "o__Myxococcales"
## OTU_1360 "o__Enterobacteriales"
```

```
## OTU_1386 "o__Enterobacteriales"
## OTU_1460 "o__Pseudomonadales"
## OTU_1470 "o__Pseudomonadales"
## OTU_1474 "o__Pseudomonadales"
## OTU_1495 "o__Pseudomonadales"
## OTU_1499 "o__Pseudomonadales"
## OTU_1507 "o__Pseudomonadales"
## OTU_1565 "o__Opitutales"
##          Family
## OTU_13   "f__Actinomycetaceae"
## OTU_46   "f__Corynebacteriaceae"
## OTU_69   "f__Corynebacteriaceae"
## OTU_142  "f__Microbacteriaceae"
## OTU_143  "f__Microbacteriaceae"
## OTU_190  "f__Nocardiaceae"
## OTU_401  "f__Flavobacteriaceae"
## OTU_420  "f__Flavobacteriaceae"
## OTU_427  "f__Chitinophagaceae"
## OTU_432  "f__Chitinophagaceae"
## OTU_442  "f__Chitinophagaceae"
## OTU_456  "f__Sphingobacteriaceae"
## OTU_531  "f__Bacillales_"
## OTU_577  "f__Staphylococcaceae"
## OTU_634  "f__Enterococcaceae"
## OTU_708  "f__Clostridiaceae-1"
## OTU_727  "f__Clostridiales_"
## OTU_763  "f__Lachnospiraceae"
## OTU_805  "f__Peptoniphilaceae"
## OTU_832  "f__Ruminococcaceae"
## OTU_833  "f__Ruminococcaceae"
## OTU_910  "f__Caulobacteraceae"
## OTU_933  "f__(Beijerinckiaceae_33%_Bradyrhizobiaceae_30%_Phyllobacteriaceae_24%_Hyphomicrobiaceae_6%
## OTU_961  "f__Bradyrhizobiaceae"
## OTU_964  "f__Bradyrhizobiaceae"
## OTU_1034 "f__Rhodobacteraceae"
## OTU_1037 "f__Rhodobacteraceae"
## OTU_1042 "f__(Rhodospirillaceae)"
## OTU_1055 "f__Rhodospirillaceae"
## OTU_1075 "f__Sphingomonadaceae"
## OTU_1094 "f__Sphingomonadaceae"
## OTU_1112 "f__Alcaligenaceae"
## OTU_1128 "f__Burkholderiaceae"
## OTU_1143 "f__Burkholderiaceae"
## OTU_1165 "f__Burkholderiaceae"
## OTU_1166 "f__Comamonadaceae"
## OTU_1173 "f__Comamonadaceae"
## OTU_1198 "f__Comamonadaceae"
## OTU_1309 "f__(Labilitrichaceae_35%_Polyangiaceae_33%_Phaselicystidaceae_12%_Kofleriaceae_11%_Cystoba
## OTU_1360 "f__Enterobacteriaceae"
## OTU_1386 "f__Enterobacteriaceae"
## OTU_1460 "f__Moraxellaceae"
## OTU_1470 "f__Moraxellaceae"
## OTU_1474 "f__Moraxellaceae"
## OTU_1495 "f__Pseudomonadaceae"
```

```
## OTU_1499 "f__Pseudomonadaceae"
## OTU_1507 "f__Pseudomonadaceae"
## OTU_1565 "f__Opitutaceae"
##          Genus
## OTU_13   "g__Actinomyces"
## OTU_46   "g__Corynebacterium"
## OTU_69   "g__Corynebacterium"
## OTU_142  "g__(Zimmermannella_31%_Amnibacterium_17%_Leifsonia_17%_Frigoribacterium_14%_Leucobacter_6%
## OTU_143  "g__Microbacterium"
## OTU_190  "g__Rhodococcus"
## OTU_401  "g__Cloacibacterium"
## OTU_420  "g__Flavobacterium"
## OTU_427  "g__(Vibrionimonas_36%_Sediminibacterium_33%_Asinibacterium_15%_Hydrobacter_14%_Terrimonas_
## OTU_432  "g__Asinibacterium"
## OTU_442  "g__Sediminibacterium"
## OTU_456  "g__Pedobacter"
## OTU_531  "g__Exiguobacterium"
## OTU_577  "g__Staphylococcus"
## OTU_634  "g__Enterococcus"
## OTU_708  "g__Clostridium"
## OTU_727  "g__Anaerococcus"
## OTU_763  "g__Blautia"
## OTU_805  "g__Peptoniphilus"
## OTU_832  "g__Faecalibacterium"
## OTU_833  "g__Fastidiosipila"
## OTU_910  "g__(Phenylobacterium_85%_Asticcacaulis_15%)"
## OTU_933  NA
## OTU_961  "g__Oligotropha"
## OTU_964  "g__Rhodoblastus"
## OTU_1034 "g__Paracoccus"
## OTU_1037 "g__Rhodobacter"
## OTU_1042 NA
## OTU_1055 "g__(Azospirillum_67%_Lacibacterium_16%_Skermanella_12%_Elstera_2%)"
## OTU_1075 "g__(Novosphingobium_63%_Sphingomonas_19%_Blastomonas_14%_Sphingobium_3%_Sphingopyxis_2%)"
## OTU_1094 "g__Sphingomonas"
## OTU_1112 "g__(Achromobacter_50%_Pelistega_28%_Basilea_11%_Bordetella_11%)"
## OTU_1128 "g__Burkholderia"
## OTU_1143 "g__Burkholderia"
## OTU_1165 "g__Ralstonia"
## OTU_1166 "g__(Acidovorax_35%_Curvibacter_29%_Comamonas_19%_Mitsuaria_13%_Diaphorobacter_3%)"
## OTU_1173 "g__Caldimonas"
## OTU_1198 "g__Pelomonas"
## OTU_1309 NA
## OTU_1360 "g__(Enterobacter_35%_Raoultella_21%_Klebsiella_10%_Citrobacter_7%_Pantoea_5%_Escherichia_4
## OTU_1386 "g__Klebsiella"
## OTU_1460 "g__Acinetobacter"
## OTU_1470 "g__Moraxella"
## OTU_1474 "g__Moraxella"
## OTU_1495 "g__Pseudomonas"
## OTU_1499 "g__Pseudomonas"
## OTU_1507 "g__Pseudomonas"
## OTU_1565 "g__Opitutus"
##          Species
## OTU_13   "s__(oris_37%_naeslundii_27%_odontolyticus_23%_viscosus_10%_israelii_2%_georgiae_1%)"
```

```
## OTU_46   "s__(striatum_29%_pseudogenitalium_9%_tuberculostearicum_8%_accolens_8%_tuscaniense_5%_propi
## OTU_69   "s__lipophiloflavum"
## OTU_142  NA
## OTU_143  "s__(oxydans_89%_trichothecenolyticum_6%_schleiferi_6%)"
## OTU_190  "s__erythropolis"
## OTU_401  "s__normanense"
## OTU_420  "s__suncheonense"
## OTU_427  NA
## OTU_432  "s__lactis"
## OTU_442  "s__(salmoneum_94%_goheungense_6%)"
## OTU_456  "s__cryoconitis"
## OTU_531  "s__(lactigenes_59%_mexicanum_32%_aurantiacum_5%_sibiricum_2%_acetylicum_2%)"
## OTU_577  "s__(epidermidis_58%_hominis_20%_caprae_9%_lugdunensis_4%_aureus_3%_nepalensis_3%_warneri_2
## OTU_634  "s__(faecalis_78%_casseliflavus_12%_hirae_3%_termitis_2%_faecium_2%_devriesei_2%)"
## OTU_708  "s__butyricum"
## OTU_727  "s__(octavius_33%_provenciensis_23%_pacaensis_20%_murdochii_16%_hydrogenalis_2%_obesiensis_1
## OTU_763  "s__(luti_83%_producta_17%)"
## OTU_805  "s__duerdenii"
## OTU_832  "s__prausnitzii"
## OTU_833  "s__sanguinis"
## OTU_910  NA
## OTU_933  NA
## OTU_961  "s__(carboxidovorans)"
## OTU_964  "s__acidophilus"
## OTU_1034 "s__(limosus_17%_saliphilus_15%_siganidrum_14%_sphaerophysae_12%_marcusii_10%_marinus_10%_y
## OTU_1037 "s__(blasticus)"
## OTU_1042 NA
## OTU_1055 NA
## OTU_1075 NA
## OTU_1094 "s__echinoides"
## OTU_1112 NA
## OTU_1128 "s__(caledonica_52%_xenovorans_10%_graminis_5%_cepacia_5%_hospita_4%_terricola_4%_insulsa_4
## OTU_1143 "s__sprentiae"
## OTU_1165 "s__solanacearum"
## OTU_1166 NA
## OTU_1173 "s__hydrothermale"
## OTU_1198 "s__(saccharophila_78%_puraquae_9%_soli_7%_aquatica_6%)"
## OTU_1309 NA
## OTU_1360 NA
## OTU_1386 "s__variicola"
## OTU_1460 "s__junii"
## OTU_1470 "s__(catarrhalis_59%_nonliquefaciens_20%_cuniculi_19%_lincolnii_2%)"
## OTU_1474 "s__lincolnii"
## OTU_1495 "s__fluorescens"
## OTU_1499 "s__marginalis"
## OTU_1507 "s__veronii"
## OTU_1565 "s__terrae"
```

How abundant are the identified contaminants in the controls vs samples?

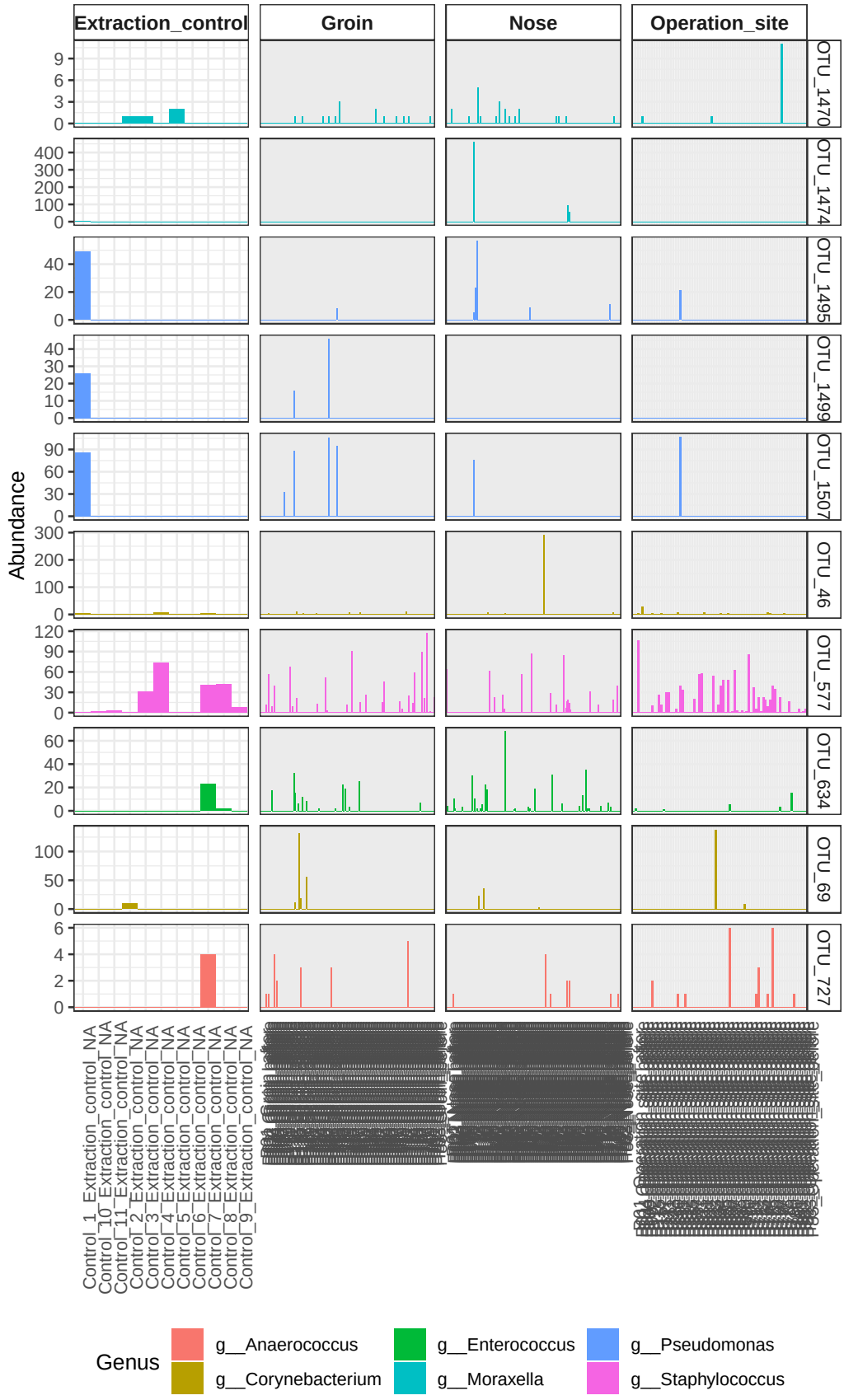
```
ps1
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1504 taxa and 372 samples ]
## sample_data() Sample Data:       [ 372 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 1504 taxa by 7 taxonomic ranks ]
```

```
ps1_contam <- prune_taxa(contamdf.prev$contaminant, ps1)
ps1_contam
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 48 taxa and 372 samples ]
## sample_data() Sample Data:       [ 372 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 48 taxa by 7 taxonomic ranks ]
```

```
ps1_contam_df <- psmelt(ps1_contam)

gens <- c("g__Moraxella", "g__Corynebacterium", "g__Anaerococcus",
    "g__Enterococcus", "g__Staphylococcus", "g__Pseudomonas")

ggplot(data = ps1_contam_df[ps1_contam_df$Genus %in% gens,
    ], aes(x = Sample, y = Abundance, fill = Genus)) + geom_bar(stat = "identity",
    width = 1) + # scale_fill_manual(values = getPalette(colourCount)) +
theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
    axis.text.x = element_text(angle = 90), legend.position = "bottom",
    strip.background = element_rect(fill = "white"), strip.text.x = element_text(size = 10,
        face = "bold")) + guides(fill = guide_legend(nrow = 2)) +
    facet_grid(OTU ~ Sample_type, scales = "free")
```


Extraction_control
Groin
Nose
Operation_site
OTU_1470
OTU_1474
OTU_1495
OTU_1499
OTU_1507
OTU_46
OTU_577
OTU_634
OTU_69
OTU_727
Abundance
Control_1_Extraction_control_NA
Control_10_Extraction_control_NA
Control_11_Extraction_control_NA
Control_2_Extraction_control_NA
Control_3_Extraction_control_NA
Control_4_Extraction_control_NA
Control_5_Extraction_control_NA
Control_6_Extraction_control_NA
Control_7_Extraction_control_NA
Control_8_Extraction_control_NA
Control_9_Extraction_control_NA
Genus
g__Anaerococcus
g__Corynebacterium
g__Enterococcus
g__Moraxella
g__Pseudomonas
g__Staphylococcus

Remove the contaminants identified with prevalence method except for those belonging to the genera Moraxella, Staphylococcus, Anaerococcus, Enterococcus, and Corynebacterium

```
gens1 <- c("g__Moraxella", "g__Corynebacterium", "g__Anaerococcus",
    "g__Enterococcus", "g__Staphylococcus")

ps1_contam
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 48 taxa and 372 samples ]
## sample_data() Sample Data:       [ 372 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 48 taxa by 7 taxonomic ranks ]
```

```
ps1_contam1 <- subset_taxa(ps1_contam, !Genus %in% gens1)
ps1_contam1
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 41 taxa and 372 samples ]
## sample_data() Sample Data:       [ 372 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 41 taxa by 7 taxonomic ranks ]
```

```
ps1
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1504 taxa and 372 samples ]
## sample_data() Sample Data:       [ 372 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 1504 taxa by 7 taxonomic ranks ]
```

```
ps1_clean <- prune_taxa(!taxa_names(ps1) %in% taxa_names(ps1_contam1),
    ps1)
ps1_clean
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1463 taxa and 372 samples ]
## sample_data() Sample Data:       [ 372 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 1463 taxa by 7 taxonomic ranks ]
```

Exclude controls

```
ps1_clean <- prune_samples(!sample_data(ps1_clean)$Sample_type ==
    "Extraction_control", ps1_clean)
ps1_clean <- prune_taxa(taxa_sums(ps1_clean) != 0, ps1_clean)
ps1_clean
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1454 taxa and 361 samples ]
## sample_data() Sample Data:       [ 361 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 1454 taxa by 7 taxonomic ranks ]
```

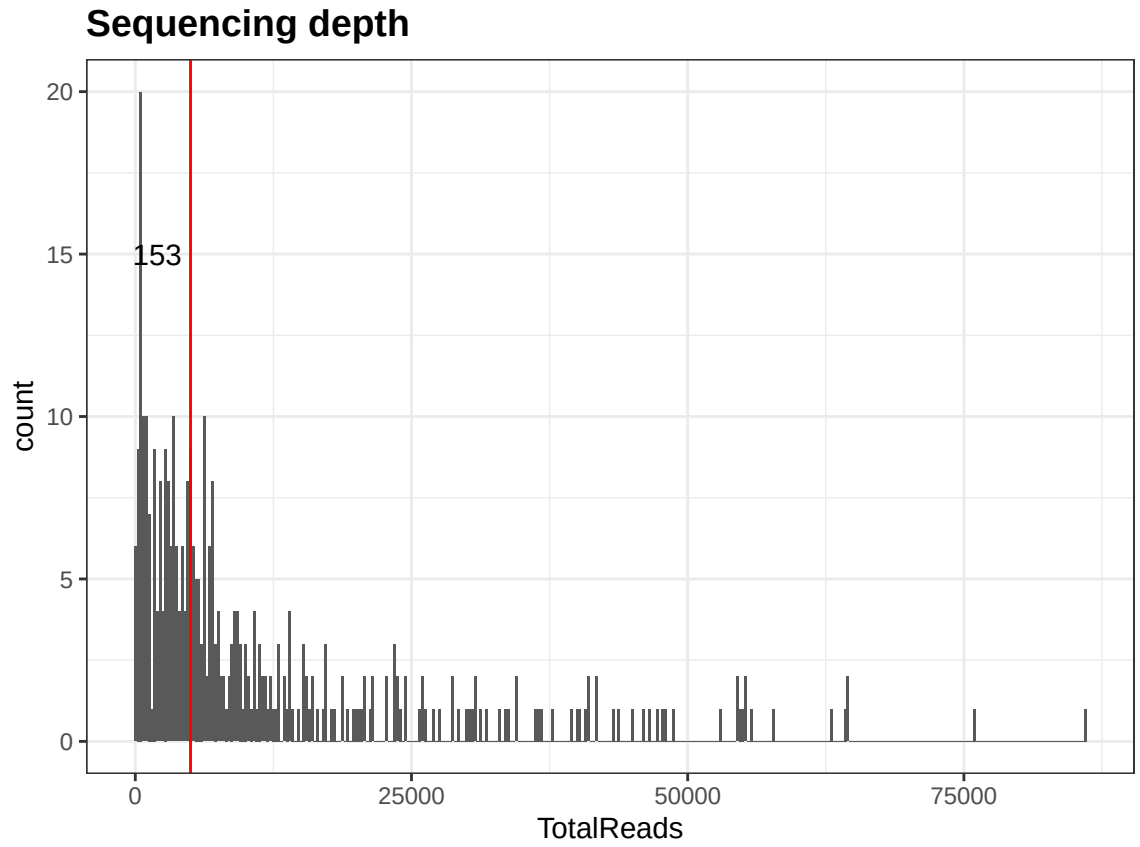
```
summary(sample_sums(ps1_clean))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      19    2702    6172   12356   15409   86082
```

How does the sequencing depth look after decontam?

```
sdt = data.table::data.table(as(sample_data(ps1_clean),
    "data.frame"), TotalReads = sample_sums(ps1_clean),
    keep.rownames = TRUE)
data.table::setnames(sdt, "rn", "SampleID")
pSeqDepth = ggplot(sdt, aes(TotalReads)) + geom_histogram(binwidth = 250) +
    geom_vline(xintercept = 5000, color = "red") + annotate("text",
    label = nrow(sdt[sdt$TotalReads < 5000]), x = 2000,
    y = 15, size = 4, colour = "black") + ggtitle("Sequencing depth") +
    theme(plot.title = element_text(size = 14, face = "bold"))
```

```
pSeqDepth
```



153 samples are now below 5000 reads

Exclude remaining environmental contaminants manually (and those not classified to at least Order-level)

```
# unique(tax_table(ps1_clean)[,'Order'])
ps1_clean <- subset_taxa(ps1_clean, (Order != "NA"))
# unique(tax_table(ps1_clean)[,'Order'])

# unique(tax_table(ps1_clean)[,'Kingdom'])
ps1_clean <- subset_taxa(ps1_clean, (Kingdom != "d__Archaea"))
# unique(tax_table(ps1_clean)[,'Kingdom'])

# unique(tax_table(ps1_clean)[,'Phylum'])
ps1_clean <- subset_taxa(ps1_clean, (Phylum != "p__Cyanobacteria_Chloroplast" &
    Phylum != "p__Chloroflexi" & Phylum != "p__Deinococcus-Thermus" &
    Phylum != "p__Planctomycetes"))
# unique(tax_table(ps1_clean)[,'Phylum'])

# unique(tax_table(ps1_clean)[,'Class'])

# unique(tax_table(ps1_clean)[,'Order'])
ps1_clean <- subset_taxa(ps1_clean, (Order != "o__(Rhodospirillales_92%_Caulobacterales_4%_Sphingomonada
    Order != "o__Rhizobiales" & Order != "o__Rhodobacterales" &
    Order != "o__Rhodospirillales" & Order != "o__Rickettsiales" &
    Order != "o__Rhodocyclales" & Order != "o__Oceanospirillales"))
# unique(tax_table(ps1_clean)[,'Order'])

# unique(tax_table(ps1_clean)[,'Genus'])
ps1_clean <- subset_taxa(ps1_clean, (Genus != "g__Ralstonia" &
    Genus != "g__Burkholderia"))
# unique(tax_table(ps1_clean)[,'Genus'])

ps1_clean
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1241 taxa and 361 samples ]
## sample_data() Sample Data:       [ 361 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 1241 taxa by 7 taxonomic ranks ]
```

I excluded the following taxa manually:

Those which were not classified to more than "Class" level

"d___Archaea"

"p___Cyanobacteria_Chloroplast"

"p___Chloroflexi"

"p___Deinococcus-Thermus"

"p___Planctomycetes"

"o___(Rhodospirillales_92%_Caulobacterales_4%_Sphingomonadales_4%)"

"o___Rhizobiales"

"o___Rhodobacterales"

"o___Rhodospirillales"

"o___Rickettsiales"

"o___Rhodocyclales"

"o___Oceanospirillales"

"g___Ralstonia" (the remaining ones that were not found by decontam)

"g___Burkholderia" (the remaining ones that were not found by decontam)

213 additional OTUs removed

Summary of read counts per sample after decontam

```
print("16S after contaminant removal: ")
```

```
## [1] "16S after contaminant removal: "
```

```
print("All: ")
```

```
## [1] "All: "
```

```
summary(sample_sums(ps1_clean))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      17    1532    5088   11605   14316   86008
```

```
print("Nose: ")
```

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

```
summary(sample_sums(ps1_clean)[sample_data(ps1_clean)$Sample_type ==
    "Nose"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      17    3130    7319   14171   19730   86008
```

```
print("Groin: ")
```

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

```
summary(sample_sums(ps1_clean)[sample_data(ps1_clean)$Sample_type ==
    "Groin"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      21    2910    6842   13127   18534   64387
```

```
print("OP site: ")
```

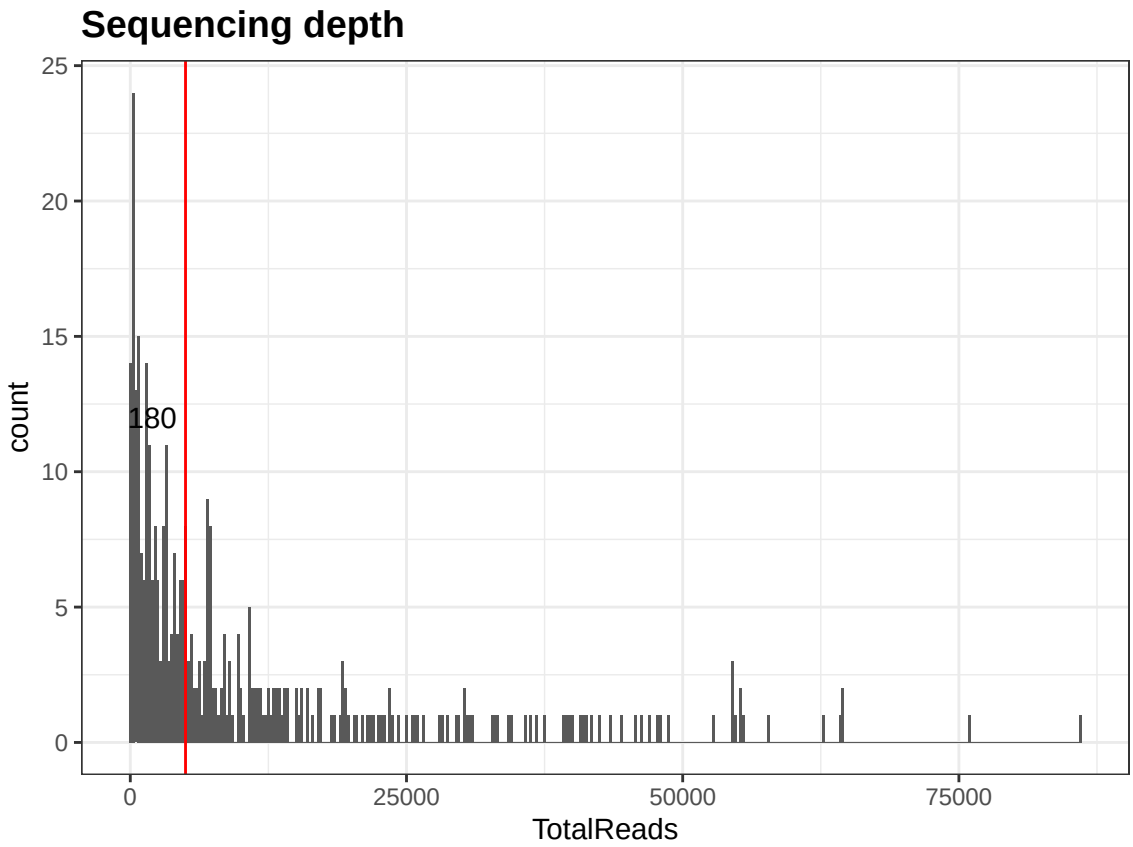
```
## [1] "OP site: "
```

```
summary(sample_sums(ps1_clean)[sample_data(ps1_clean)$Sample_type ==
    "Operation_site"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      31     311    1080    3429    2276   55445
```

How does the sequencing depth look after additional manual decontam?

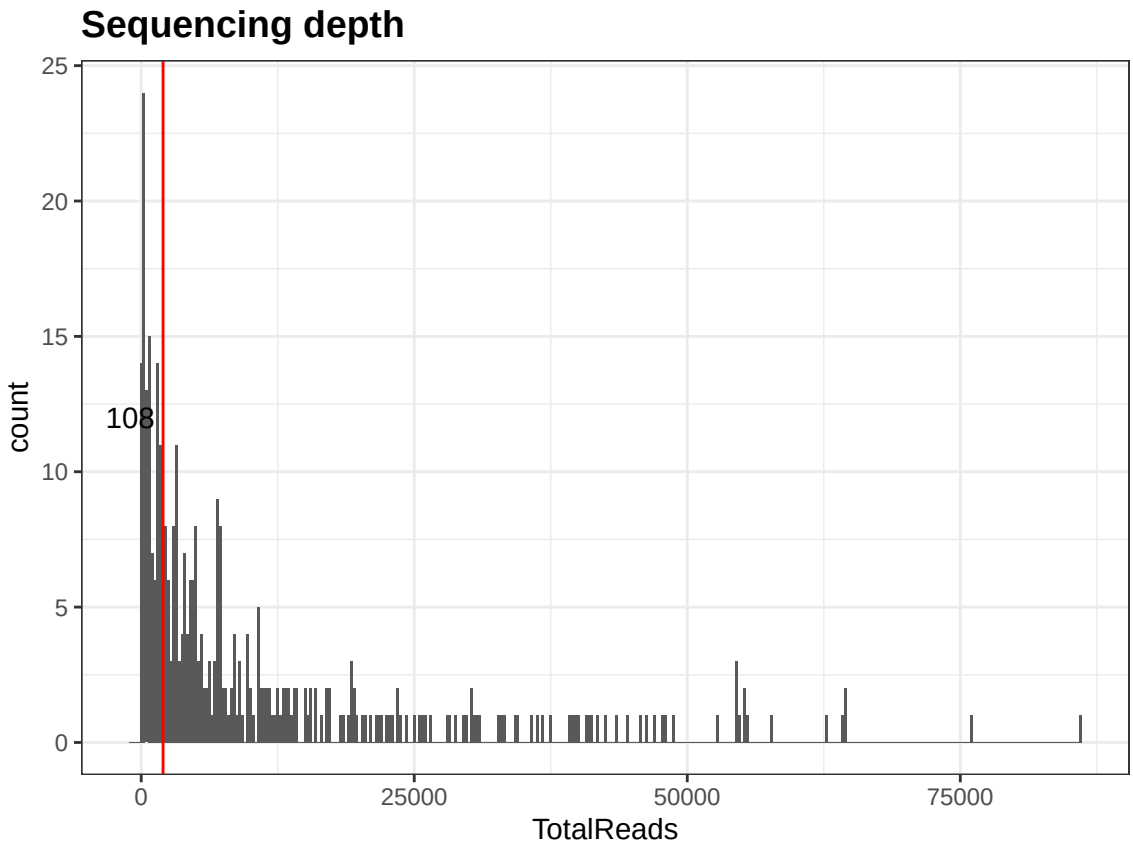
```
sdt = data.table::data.table(as(sample_data(ps1_clean),
    "data.frame"), TotalReads = sample_sums(ps1_clean),
    keep.rownames = TRUE)
data.table::setnames(sdt, "rn", "SampleID")
pSeqDepth = ggplot(sdt, aes(TotalReads)) + geom_histogram(binwidth = 250) +
    geom_vline(xintercept = 5000, color = "red") + annotate("text",
    label = nrow(sdt[sdt$TotalReads < 5000]), x = 2000,
    y = 12, size = 4, colour = "black") + ggtitle("Sequencing depth") +
    theme(plot.title = element_text(size = 14, face = "bold"))

pSeqDepth
```



180 samples are now below 5000 reads

How many are below 2000?

```
sdt = data.table::data.table(as(sample_data(ps1_clean),
    "data.frame"), TotalReads = sample_sums(ps1_clean),
    keep.rownames = TRUE)
data.table::setnames(sdt, "rn", "SampleID")
pSeqDepth = ggplot(sdt, aes(TotalReads)) + geom_histogram(binwidth = 250) +
    geom_vline(xintercept = 2000, color = "red") + annotate("text",
    label = nrow(sdt[sdt$TotalReads < 2000]), x = -1000,
    y = 12, size = 4, colour = "black") + ggtitle("Sequencing depth") +
    theme(plot.title = element_text(size = 14, face = "bold"))

pSeqDepth
```



108 samples are < 2000 reads

Split by body site

Nose

```
ps1_clean_nose <- prune_samples(sample_data(ps1_clean)$Sample_type ==
    "Nose", ps1_clean)
ps1_clean_nose <- prune_taxa(taxa_sums(ps1_clean_nose) !=
    0, ps1_clean_nose)
ps1_clean_nose
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 845 taxa and 161 samples ]
## sample_data() Sample Data:       [ 161 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 845 taxa by 7 taxonomic ranks ]
```

```
sample_data(ps1_clean_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"
```

Number of patients with nose samples overall

```
length(unique(sample_data(ps1_clean_nose)$Patient_ID))
```

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

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

```
table(sample_data(ps1_clean_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 3 patients with only one time point

```
ps1_clean_nose <- prune_samples(!sample_data(ps1_clean_nose)$Patient_ID %in%
    c("P23", "P33", "P74"), ps1_clean_nose)
ps1_clean_nose
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 845 taxa and 158 samples ]
## sample_data() Sample Data:       [ 158 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 845 taxa by 7 taxonomic ranks ]
```

```
table(sample_data(ps1_clean_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
```

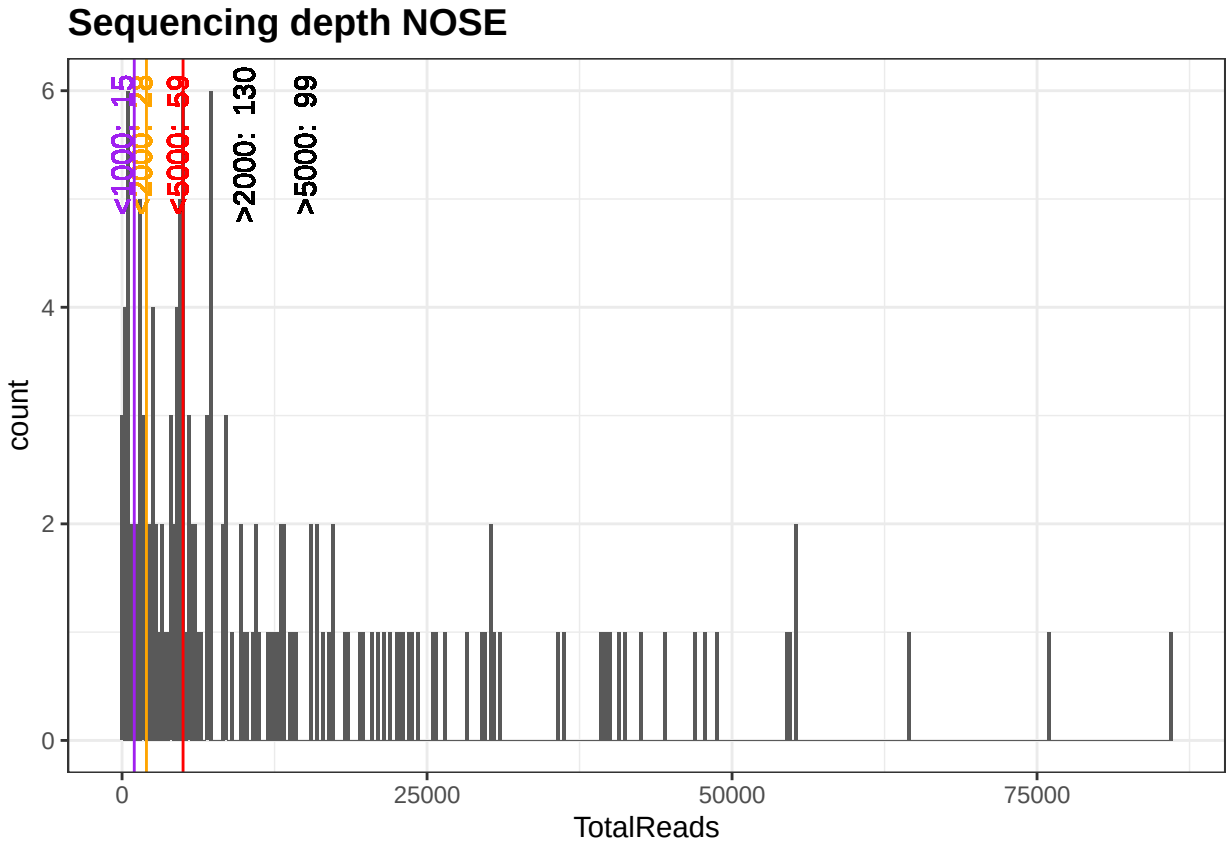
How many patients left with both a before and and after sample?

```
length(unique(sample_data(ps1_clean_nose)$Patient_ID))
```

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

How many samples have below 1000 / 2000 / 5000 reads?

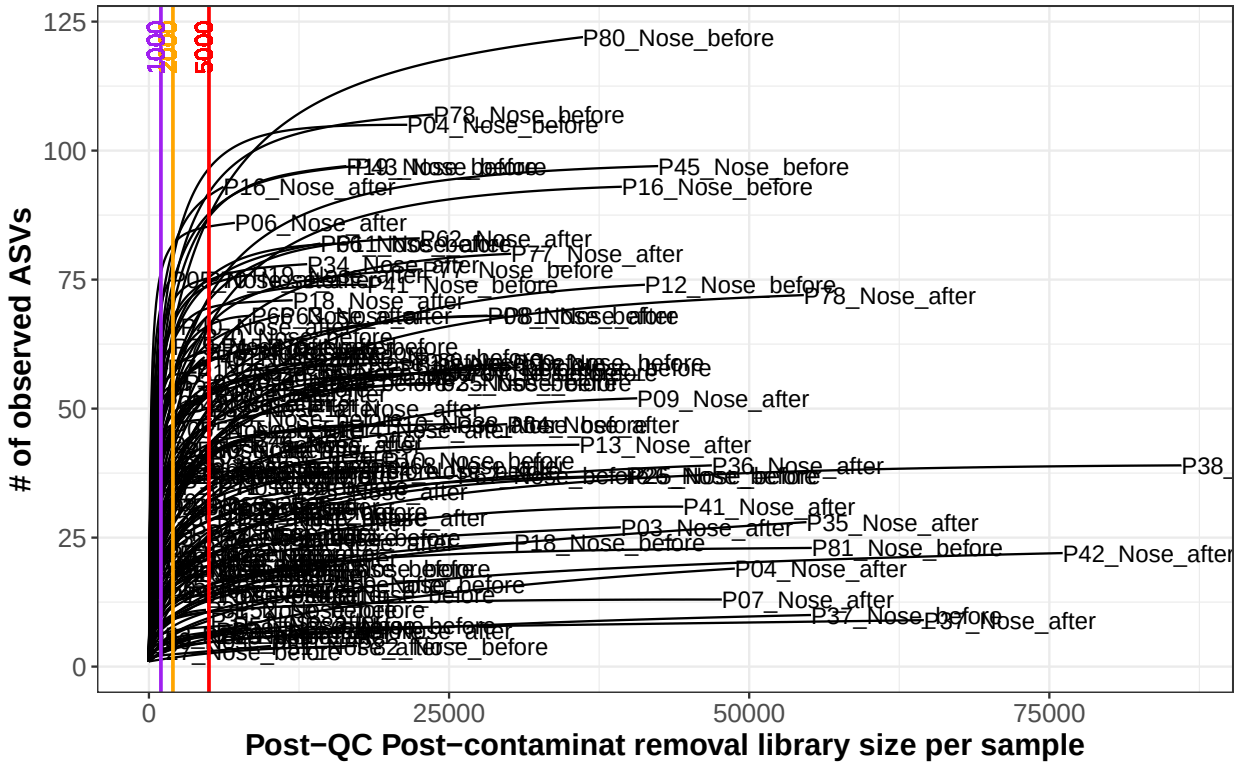
```
sdt = data.table::data.table(as(sample_data(ps1_clean_nose),
    "data.frame"), TotalReads = sample_sums(ps1_clean_nose),
    keep.rownames = TRUE)
data.table::setnames(sdt, "rn", "SampleID")
pSeqDepth = ggplot(sdt, aes(TotalReads)) + geom_histogram(binwidth = 250) +
    geom_vline(xintercept = 5000, color = "red") + geom_text(aes(x = 4550,
    label = paste("<5000: ", nrow(sdt[sdt$TotalReads < 5000])),
    y = 5.5), colour = "red", angle = 90) + geom_text(aes(x = 1550,
    label = paste("<2000: ", nrow(sdt[sdt$TotalReads < 2000])),
    y = 5.5), colour = "orange", angle = 90) + geom_text(aes(x = 0,
    label = paste("<1000: ", nrow(sdt[sdt$TotalReads < 1000])),
    y = 5.5), colour = "purple", angle = 90) + geom_text(aes(x = 10000,
    label = paste(">2000: ", nrow(sdt[sdt$TotalReads > 2000])),
    y = 5.5), colour = "black", angle = 90) + geom_text(aes(x = 15000,
    label = paste(">5000: ", nrow(sdt[sdt$TotalReads > 5000])),
    y = 5.5), colour = "black", angle = 90) + geom_vline(xintercept = 2000,
    color = "orange") + geom_vline(xintercept = 1000, color = "purple") +
    ggtitle("Sequencing depth NOSE") + theme(plot.title = element_text(size = 14,
    face = "bold"))

pSeqDepth
```



Do the rarefaction curves justify that we remove samples with reads <1000 / <2000?

Rarefaction curves

```
p1 <- p1 + 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 = 120), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 120), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 120), colour = "purple",
        angle = 90, size = 4)
p1
```

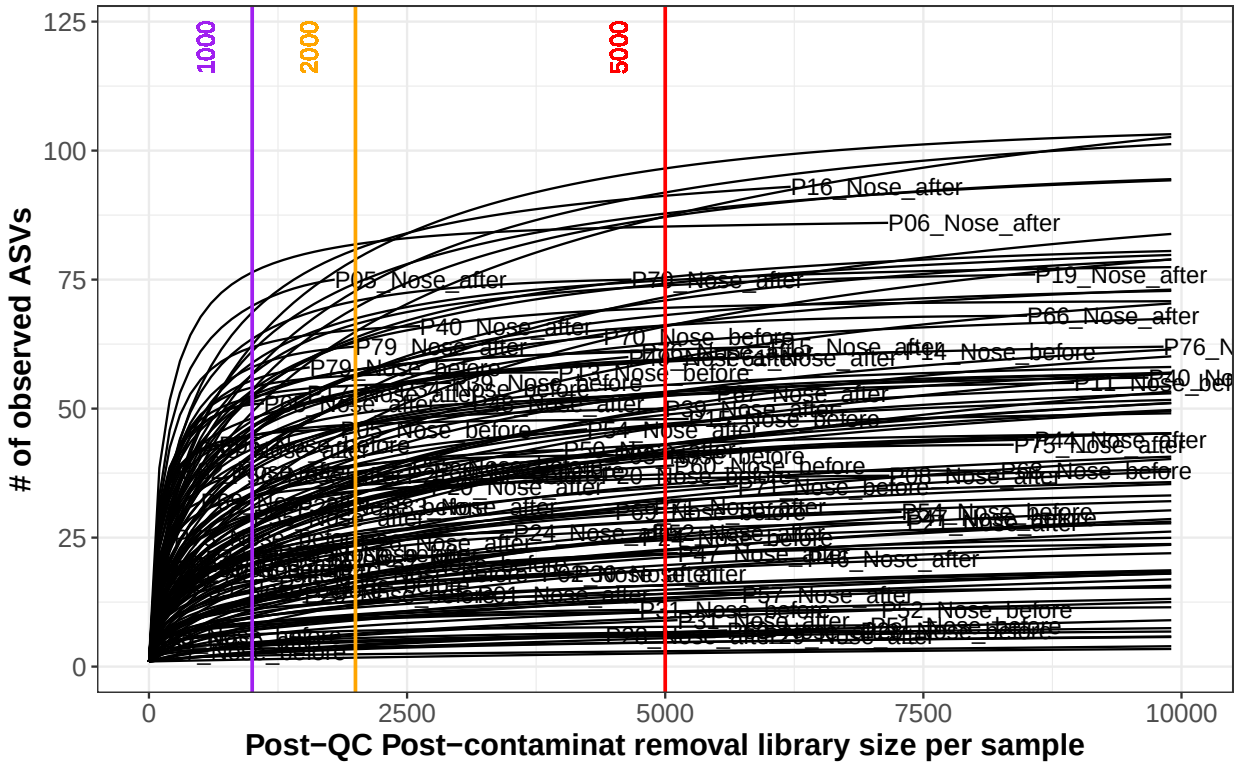


Zoom

```
p1 + xlim(0, 10000)
```

```
## Warning: Removed 69 rows containing missing values (geom_text).
```

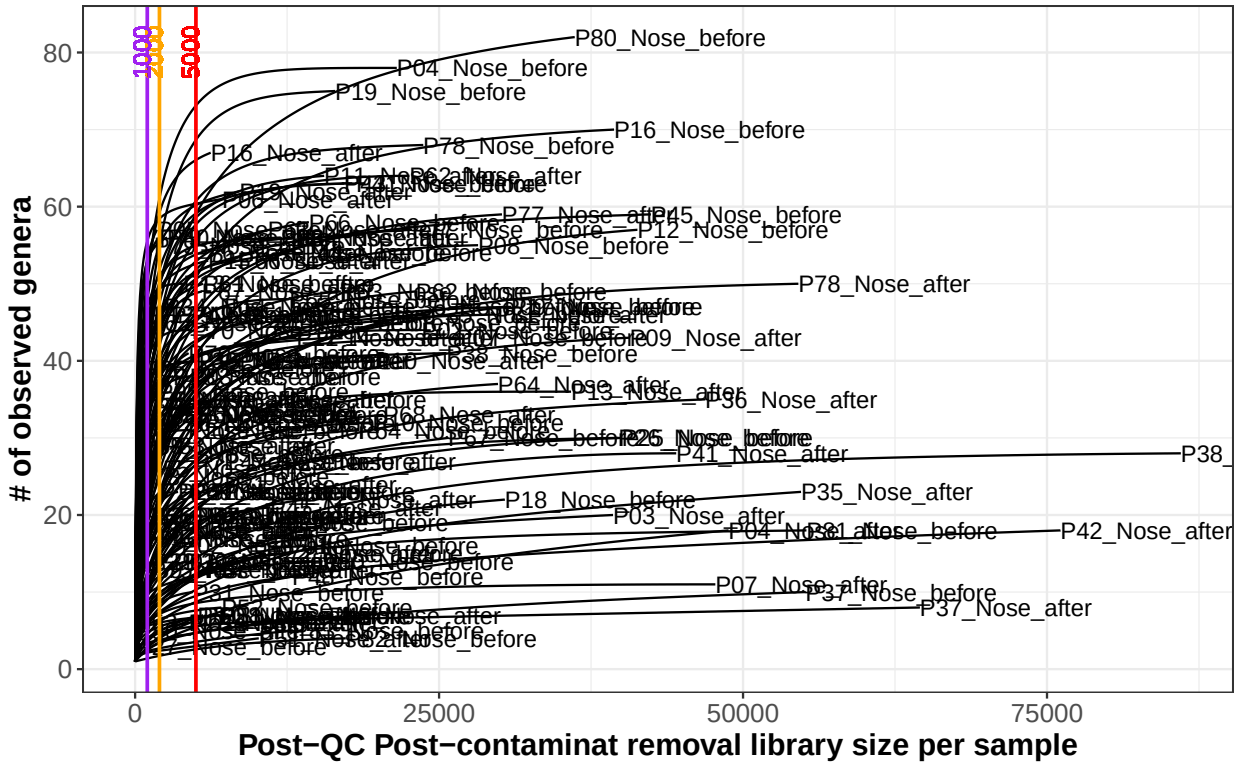
```
## Warning: Removed 12365 row(s) containing missing values (geom_path).
```



How do the rarefaction curves look on genus level?

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 genera") + 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 = 80), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 80), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 80), colour = "purple",
        angle = 90, size = 4)
p2
```

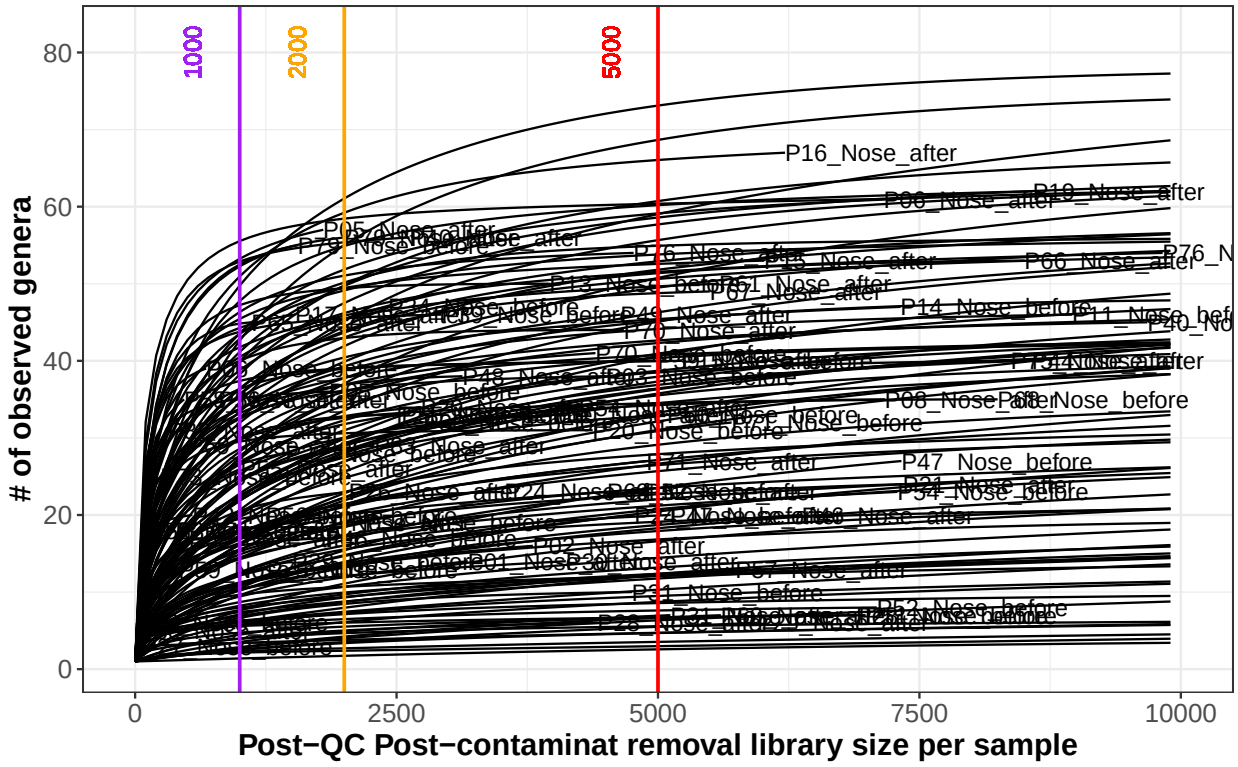


Zoom

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

```
## Warning: Removed 69 rows containing missing values (geom_text).
```

```
## Warning: Removed 12365 row(s) containing missing values (geom_path).
```



Exclude samples with <2000 reads

```
summary(sample_sums(ps1_clean_nose))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      17    3269    7324   14335   20241   86008
```

```
ps1_clean_nose_tu <- prune_samples(!sample_sums(ps1_clean_nose) <
    2000, ps1_clean_nose)
ps1_clean_nose_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 845 taxa and 130 samples ]
## sample_data() Sample Data:       [ 130 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 845 taxa by 7 taxonomic ranks ]
```

```
summary(sample_sums(ps1_clean_nose_tu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2042    5120   11011   17230   23435   86008
```

Now how many patients still have both time points left after removing samples with <2000 reads?

```
table(sample_data(ps1_clean_nose_tu)$Patient_ID)
```

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

```
length(unique(sample_data(ps1_clean_nose_tu)$Patient_ID))
```

```
## [1] 71
```

```
ps1_clean_nose_tu <- prune_samples(!sample_data(ps1_clean_nose_tu)$Patient_ID %in%
    c("P06", "P17", "P22", "P29", "P42", "P46", "P55", "P65",
        "P69", "P75", "P80", "P82"), ps1_clean_nose_tu)
ps1_clean_nose_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 845 taxa and 118 samples ]
## sample_data() Sample Data:       [ 118 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 845 taxa by 7 taxonomic ranks ]
```

```
length(unique(sample_data(ps1_clean_nose_tu)$Patient_ID))
```

```
## [1] 59
```

59 patients left with 2 time points

Read counts in the remaining patients with 2 samples >2000 reads

```
summary(sample_sums(ps1_clean_nose_tu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2042    5096   11174   17127   23435   86008
```

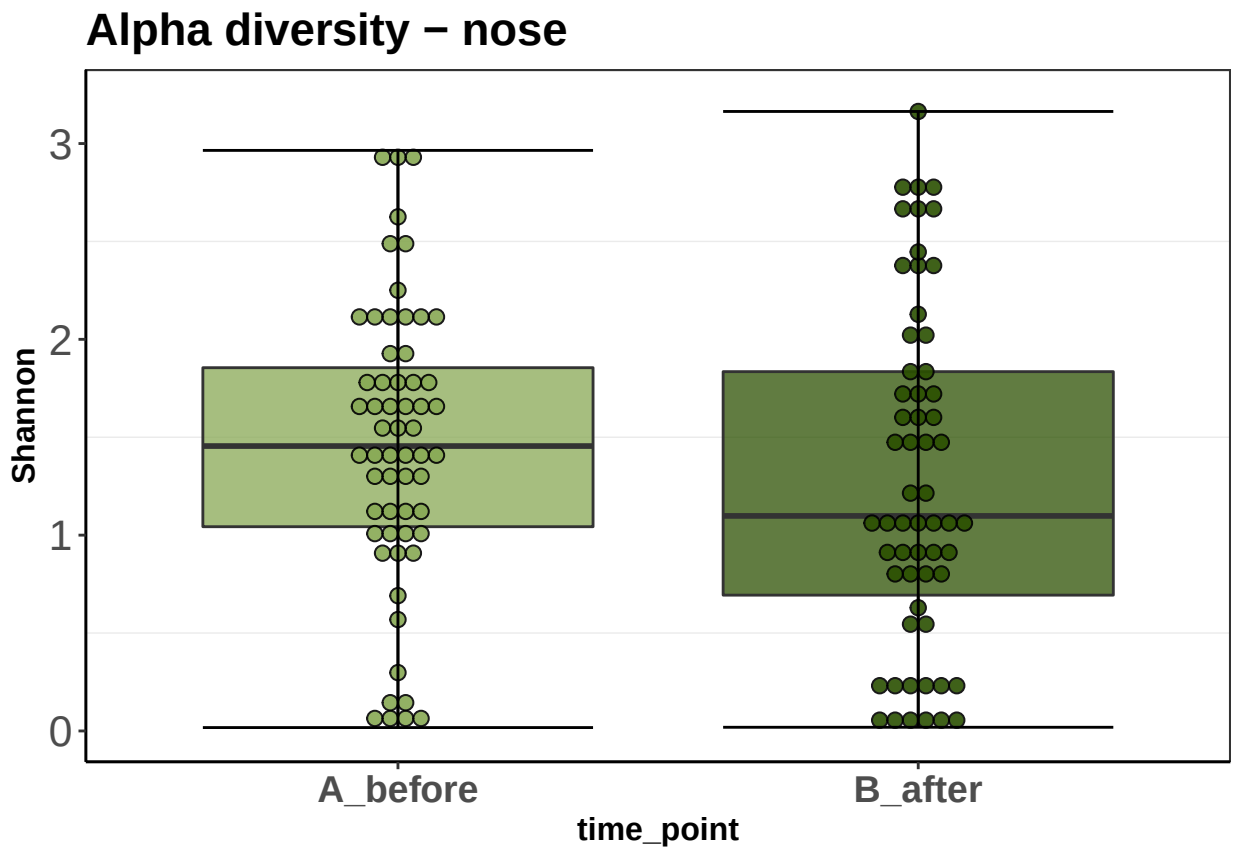
Alpha diversity

```
#### Add diversity measures to the phyloseq object as
#### variables
alpha_div_raw <- estimate_richness(ps1_clean_nose_tu, measures = c("Observed",
    "Chao1", "Shannon", "InvSimpson"))
rownames(alpha_div_raw) <- gsub("X", "", rownames(alpha_div_raw))
ps1_clean_nose_tu <- merge_phyloseq(ps1_clean_nose_tu, sample_data(alpha_div_raw))
df_ps1_clean_nose_tu <- as(sample_data(ps1_clean_nose_tu),
    "data.frame")
```

Shannon diversity over time:

```
plot_g_Shannon <- ggplot(df_ps1_clean_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("Alpha diversity - nose")
plot_g_Shannon
```

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




```
ggsave(filename = "plots/Nose_alpha_div_16S.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_ps1_clean_nose_tu_c <- dcast(df_ps1_clean_nose_tu, Patient_ID ~
    time_point, value.var = "Shannon", drop = FALSE)
wilcox.test(df_ps1_clean_nose_tu_c$A_before, df_ps1_clean_nose_tu_c$B_after,
    paired = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  df_ps1_clean_nose_tu_c$A_before and df_ps1_clean_nose_tu_c$B_after
## V = 1043, p-value = 0.2345
## alternative hypothesis: true location shift is not equal to 0
```

No significant change in alpha diversity in the nose.

Agglomerate on Genus level

```
ps1_clean_nose_tu_gs <- tax_glom(ps1_clean_nose_tu, taxrank = "Genus")
ps1_clean_nose_tu_gs
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 342 taxa and 118 samples ]
## sample_data() Sample Data:       [ 118 samples by 9 sample variables ]
## tax_table()   Taxonomy Table:    [ 342 taxa by 7 taxonomic ranks ]
```

```
ps1_clean_nose_tu_gs <- prune_taxa(taxa_sums(ps1_clean_nose_tu_gs) !=
    0, ps1_clean_nose_tu_gs)
ps1_clean_nose_tu_gs
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 328 taxa and 118 samples ]
## sample_data() Sample Data:       [ 118 samples by 9 sample variables ]
## tax_table()   Taxonomy Table:    [ 328 taxa by 7 taxonomic ranks ]
```

Convert to relative abundance

```
ps1_clean_nose_tu_gs_rel = transform_sample_counts(ps1_clean_nose_tu_gs,
    function(x) x/sum(x))
summary(sample_sums(ps1_clean_nose_tu_gs_rel))
```

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

Subset top 10 genera

```
Genus10 = names(sort(taxa_sums(ps1_clean_nose_tu_gs_rel),
    TRUE)[1:10])
```

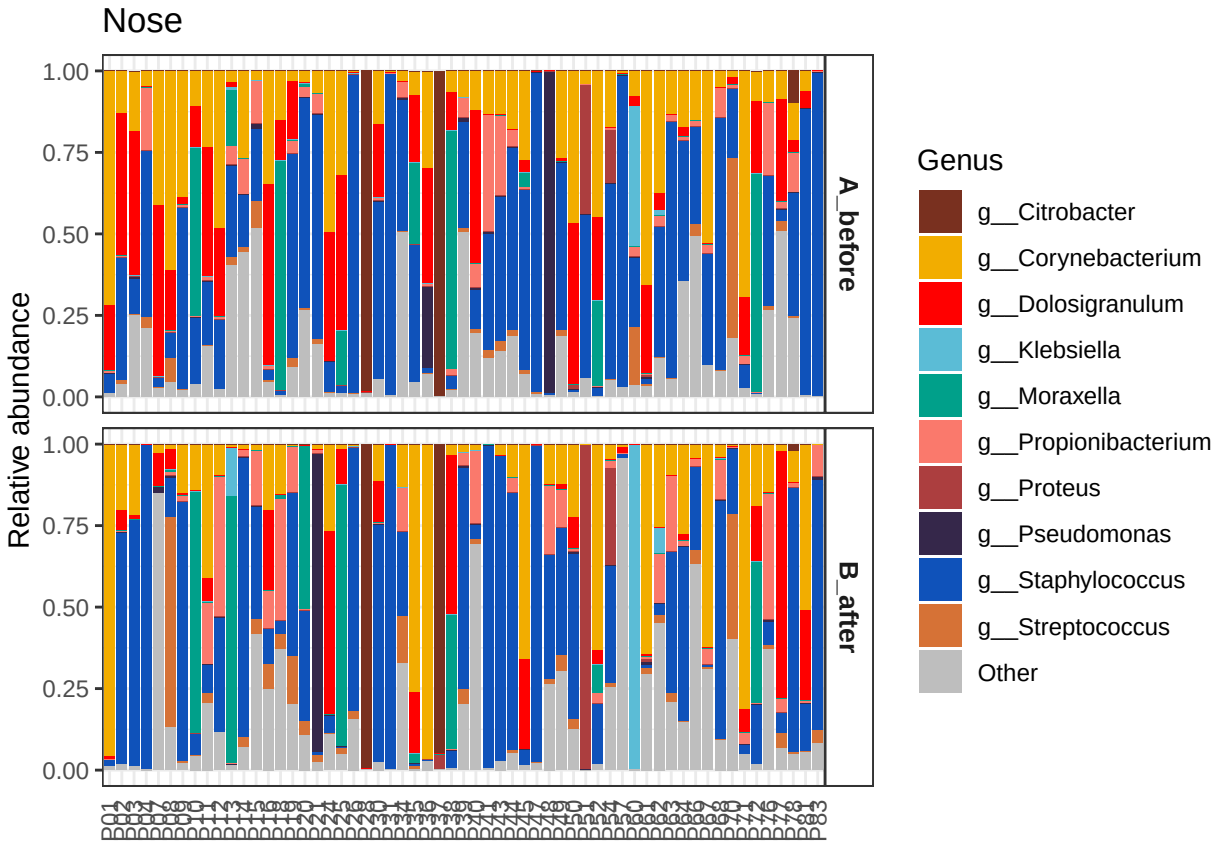
to data frame

```
p_df_o <- psmelt(ps1_clean_nose_tu_gs_rel)
p_df_o$Genus <- as.character(p_df_o$Genus)
p_df_o$Genus[!(p_df_o$OTU %in% Genus10)] <- "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)

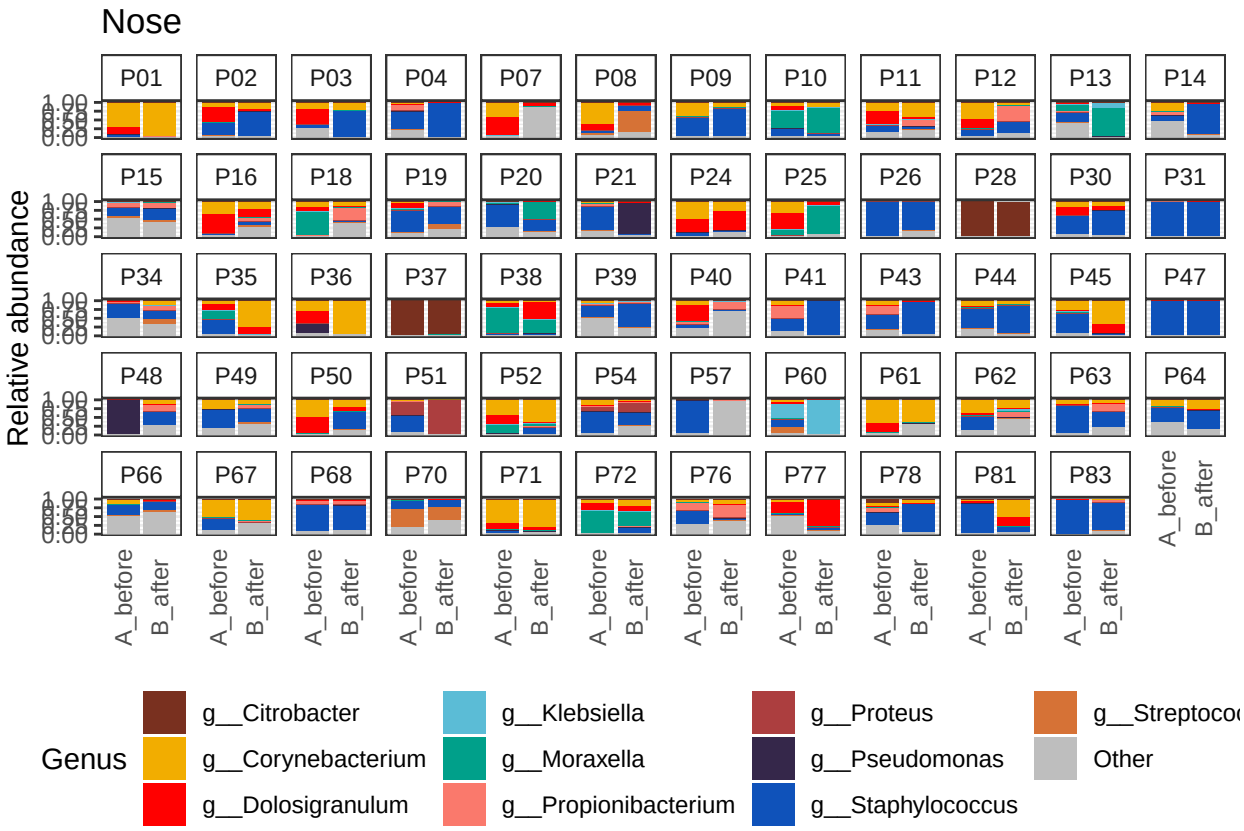
```
mycols <- c(g__Citrobacter = "#79301F", g__Corynebacterium = "#F2AD00",
    g__Dolosigranulum = "#FF0000", g__Escherichia = "#F98400",
    g__Klebsiella = "#5BBCD6", g__Moraxella = "#00A08A",
    g__Propionibacterium = "#FA796C", g__Proteus = "#AC3E3F",
    g__Staphylococcus = "#0F52BA", g__Streptococcus = "#D67236",
    Other = "grey", g__Anaerococcus = "#79402E", g__Enterococcus = "#9986A5",
    g__Finegoldia = "#CCBA72", g__Morganella = "#0F0D0E",
    g__Porphyromonas = "#0B775E", g__Pseudomonas = "#35274A")

a <- ggplot(p_df_o, aes(x = Patient_ID, y = Abundance, fill = Genus)) +
    geom_bar(stat = "identity", width = 0.9) + facet_grid(time_point ~
    ., scales = "free") + scale_fill_manual(values = mycols) +
    theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 90, vjust = 0.5),
        strip.background = element_rect(fill = "white"),
        strip.text.y = element_text(size = 10, face = "bold")) +
    ylab("Relative abundance") + ggtitle("Nose")
a
```



Patient-wise plots

```
ggplot(p_df_o, aes(x = time_point, y = Abundance, fill = Genus)) +
    geom_bar(stat = "identity", width = 0.9) + facet_wrap(. ~
    Patient_ID, nrow = 5) + scale_fill_manual(values = mycols) +
```

```
    theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 90, 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("Nose")
```



Subset the top 10 genera (without other)

```
ps1_clean_nose_tu_gs_rel <- prune_taxa(taxa_names(ps1_clean_nose_tu_gs_rel) %in%
    Genus10, ps1_clean_nose_tu_gs_rel)
ps1_clean_nose_tu_gs_rel
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 10 taxa and 118 samples ]
## sample_data() Sample Data:       [ 118 samples by 9 sample variables ]
## tax_table()   Taxonomy Table:    [ 10 taxa by 7 taxonomic ranks ]
```

to data frame

```
p_df <- psmelt(ps1_clean_nose_tu_gs_rel)
p_df_d <- dcast(p_df, Patient_ID + Genus ~ 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 ~ Patient_ID,
    value.var = "Percent_point_change")
```

Visualize in a heatmap

```
pdf(file = "plots/Nose_heatmap.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 = 1.5, cexCol = 1, adjCol = 1, key.xlab = "Relative abundance change \nin percent p
    keysize = 0.7, key.title = NA, main = "NOSE")

dev.off()
```

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

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

(Paired Wilcoxon test)

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

```
## `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                wilcox_p_value BH_adjusted_wilcox_p_value
##    <chr>                         <dbl>                      <dbl>
##  1 g__Citrobacter             0.315                      0.450
##  2 g__Corynebacterium         0.125                      0.281
##  3 g__Dolosigranulum          0.000247                   0.00247
##  4 g__Klebsiella              0.173                      0.288
##  5 g__Moraxella               0.483                      0.593
##  6 g__Propionibacterium       0.131                      0.281
##  7 g__Proteus                 0.141                      0.281
##  8 g__Pseudomonas             0.718                      0.718
##  9 g__Staphylococcus          0.533                      0.593
## 10 g__Streptococcus           0.0334                     0.167
```

Dolosigranulum and Streptococcus have a significant *overall* change in the nose. After multiple testing correction (Benjamini-Hochberg), only Dolosigranulum is still significant.

Do they *overall* decrease or increase?

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

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

```
## # A tibble: 10 x 2
##    Genus                Mean_percent_point_change
##    <chr>                                   <dbl>
##  1 g__Citrobacter                       -0.00196
##  2 g__Corynebacterium                   -0.0112
##  3 g__Dolosigranulum                    -0.0669
##  4 g__Klebsiella                         0.0131
##  5 g__Moraxella                          0.00540
##  6 g__Propionibacterium                  0.0223
##  7 g__Proteus                            0.0130
##  8 g__Pseudomonas                       -0.00560
##  9 g__Staphylococcus                    -0.0186
## 10 g__Streptococcus                      0.0135
```

Dolosigranulum decreases and Streptococcus increases *overall* in the nose.

Combine with tuf data:

Does change in the genus Staphylococcus correlate with change in individual Staph species?

```
p_df_d_STAPH <- p_df_d %>% select(Patient_ID, Genus, Percent_point_change) %>%
    filter(Genus == "g__Staphylococcus") %>% select(Patient_ID,
    Percent_point_change) %>% dplyr::rename(g__Staphylococcus_percent_point_change = Percent_point_chan
dim(p_df_d_STAPH)
```

```
## [1] 59  2
```

```
p_df_d_tuf_nose <- read.table(file = "tables/p_df_d_tuf_nose.csv",
    sep = ";", header = TRUE)
p_df_d_tuf_nose <- p_df_d_tuf_nose %>% rename_at(vars(Staphylococcus_aureus:Staphylococcus_warneri),
    function(x) {
        paste0(x, "_percent_point_change")
    })
dim(p_df_d_tuf_nose)
```

```
## [1] 65 11
```

```
## Subset to the same patients for which we have 16S data
p_df_d_tuf_nose <- p_df_d_tuf_nose[p_df_d_tuf_nose$Patient_ID %in%
    p_df_d_STAPH$Patient_ID, ]
dim(p_df_d_tuf_nose)
```

```
## [1] 51 11
```

```
p_df_d_STAPH <- p_df_d_STAPH[p_df_d_STAPH$Patient_ID %in%
    p_df_d_tuf_nose$Patient_ID, ]

p_df_d_STAPH1 <- left_join(p_df_d_STAPH, p_df_d_tuf_nose,
    by = "Patient_ID")
dim(p_df_d_STAPH1)
```

```
## [1] 51 12
```

For those patients that both have 16S and tuf data:

Test Staph genus correlation with all Staph species:

```
p_df_d_STAPH2 <- p_df_d_STAPH1 %>% pivot_longer(cols = starts_with("Staphylococcus"),
    names_to = "Species", values_to = "Percent_point_change")

test_res <- p_df_d_STAPH2 %>% group_by(Species) %>% group_modify(~broom::tidy(cor.test(~g__Staphylococcu
    Percent_point_change, data = .x)))

test_res$BH_adjusted_p_value <- p.adjust(test_res$p.value,
    method = "BH")
test_res
```

```
## # A tibble: 10 x 10
## # Groups:   Species [10]
##    Species estimate statistic p.value parameter conf.low conf.high method
##    <chr>      <dbl>     <dbl>   <dbl>     <int>    <dbl>     <dbl> <chr>
##  1 Staphy~   0.366      2.75  0.00825        49   0.101     0.583  Pears~
##  2 Staphy~  -0.395     -3.01  0.00417        49  -0.604    -0.133  Pears~
##  3 Staphy~  -0.312     -2.30  0.0259         49  -0.541    -0.0398 Pears~
##  4 Staphy~  -0.0790    -0.555 0.582          49  -0.347     0.201  Pears~
##  5 Staphy~  -0.0627    -0.440 0.662          49  -0.333     0.217  Pears~
```

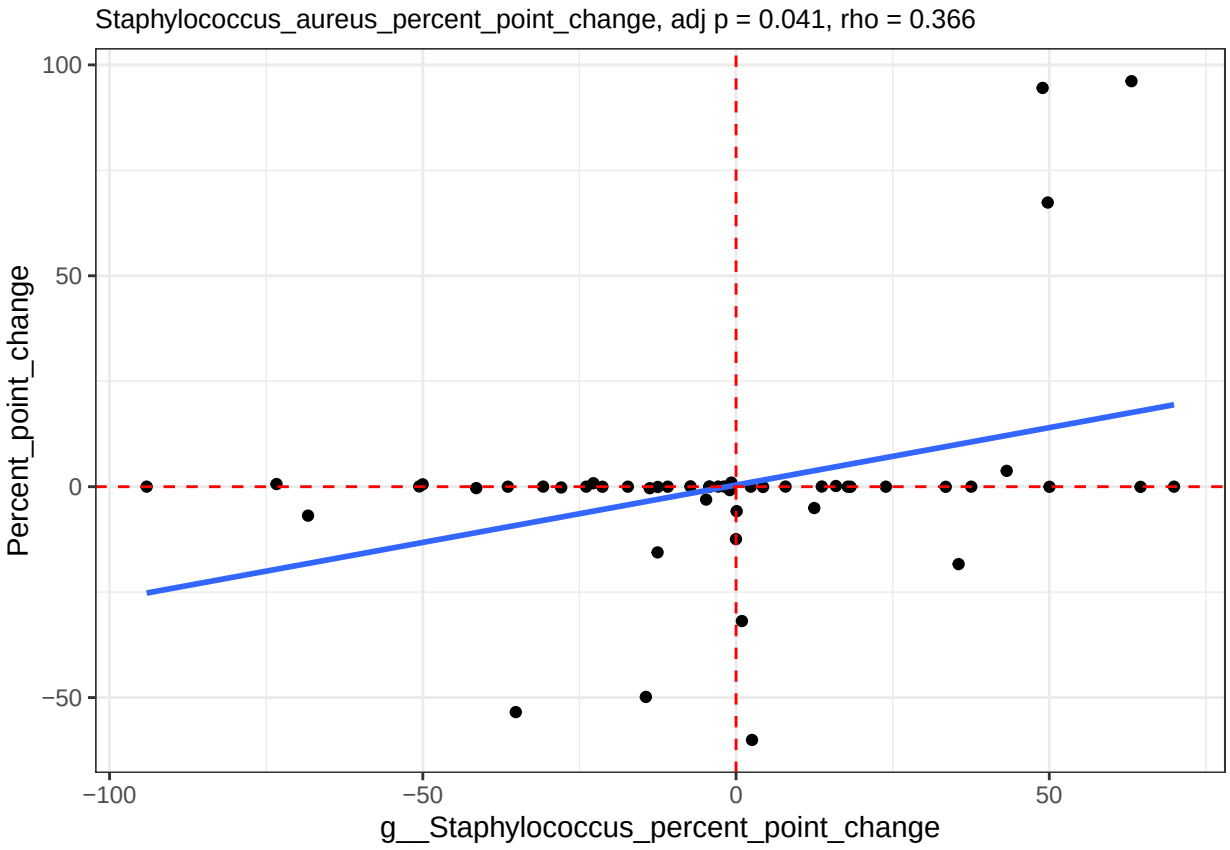
```
##  6 Staphy~  -0.167     -1.18  0.242          49  -0.423     0.114  Pears~
##  7 Staphy~  -0.0541    -0.379 0.706          49  -0.325     0.225  Pears~
##  8 Staphy~   0.142      1.00  0.322          49  -0.139     0.401  Pears~
##  9 Staphy~  -0.0288    -0.202 0.841          49  -0.302     0.249  Pears~
## 10 Staphy~   0.310      2.28  0.0271         49   0.0371    0.539  Pears~
## # ... with 2 more variables: alternative <chr>, BH_adjusted_p_value <dbl>
```

Estimate is the Pearson's correlation coefficient.

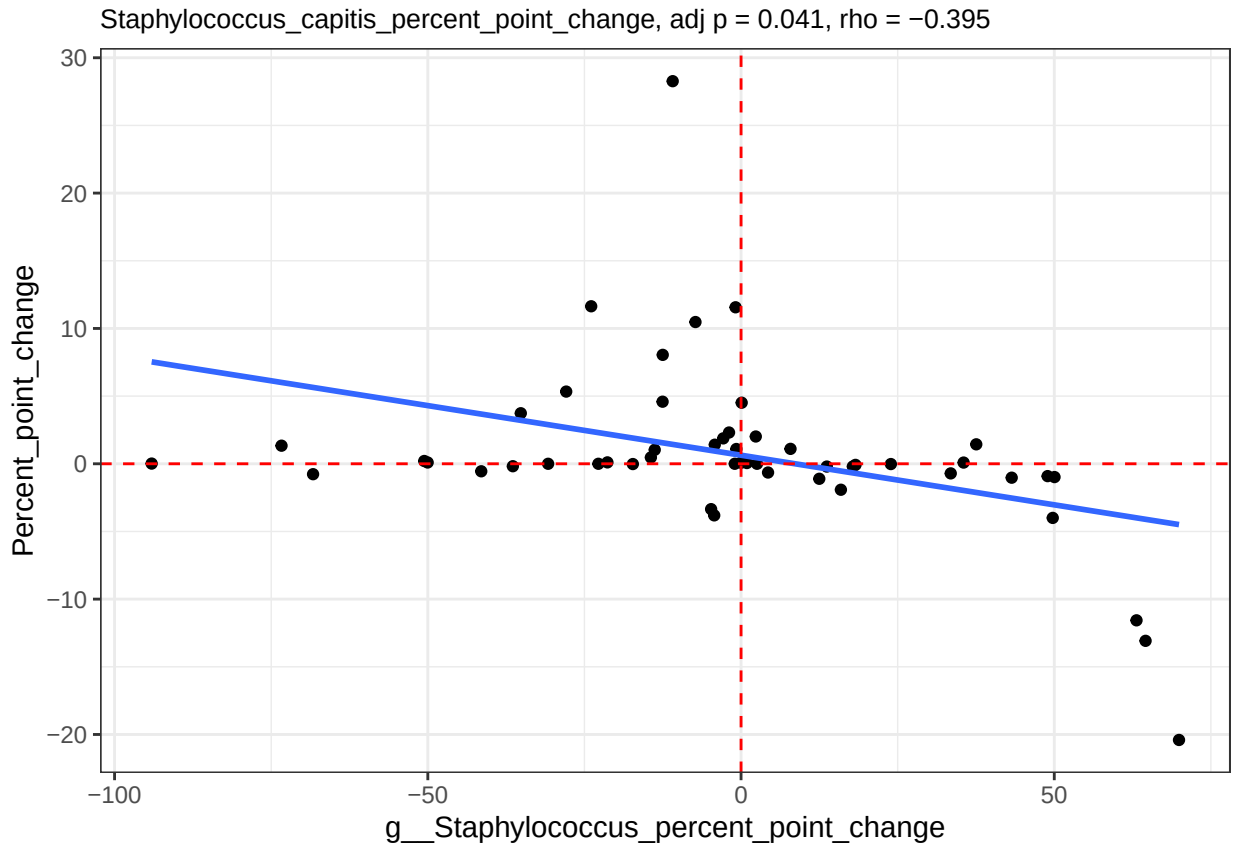
```
test_res1 <- as.data.frame(test_res[, c("Species", "estimate",
    "BH_adjusted_p_value")])

p_df_d_STAPH2 <- dplyr::left_join(p_df_d_STAPH2, test_res,
    by = "Species")
p_df_d_STAPH2 <- as.data.frame(p_df_d_STAPH2)


p_df_d_STAPH2 <- p_df_d_STAPH2 %>% mutate_at(vars(BH_adjusted_p_value,
    estimate), round, 3)

plots <- p_df_d_STAPH2 %>% group_by(Species) %>% do(plots = ggplot(data = .) +
    aes(x = g__Staphylococcus_percent_point_change, y = Percent_point_change) +
    ggtitle(paste0(unique(.$Species), ", adj p = ", unique(.$BH_adjusted_p_value),
        ", rho = ", unique(.$estimate))) + geom_point() +
    geom_smooth(method = "lm", se = FALSE) + geom_vline(xintercept = 0,
    color = "red", linetype = "dashed") + theme(plot.title = element_text(size = 10)) +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed"))
plots$plots
```

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

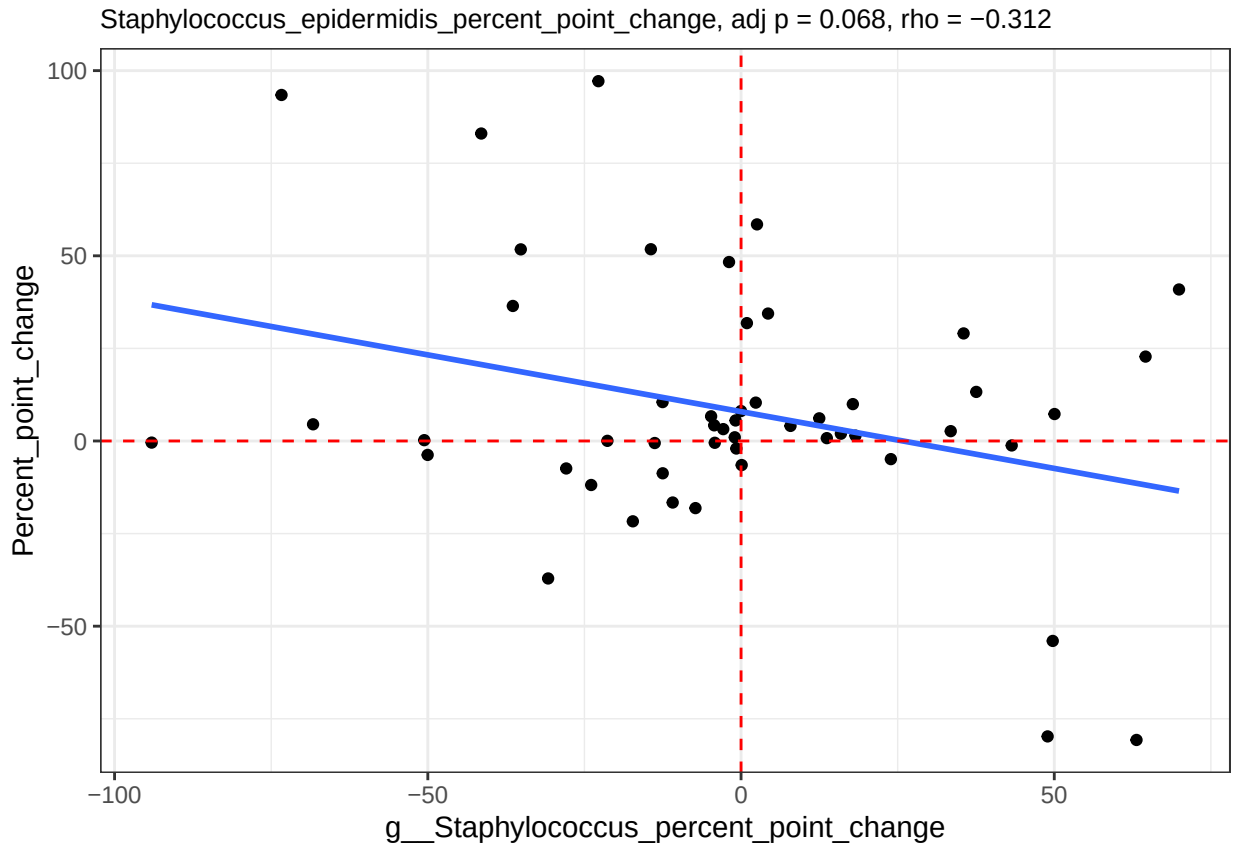
```
## `geom_smooth()` using formula 'y ~ x'
```



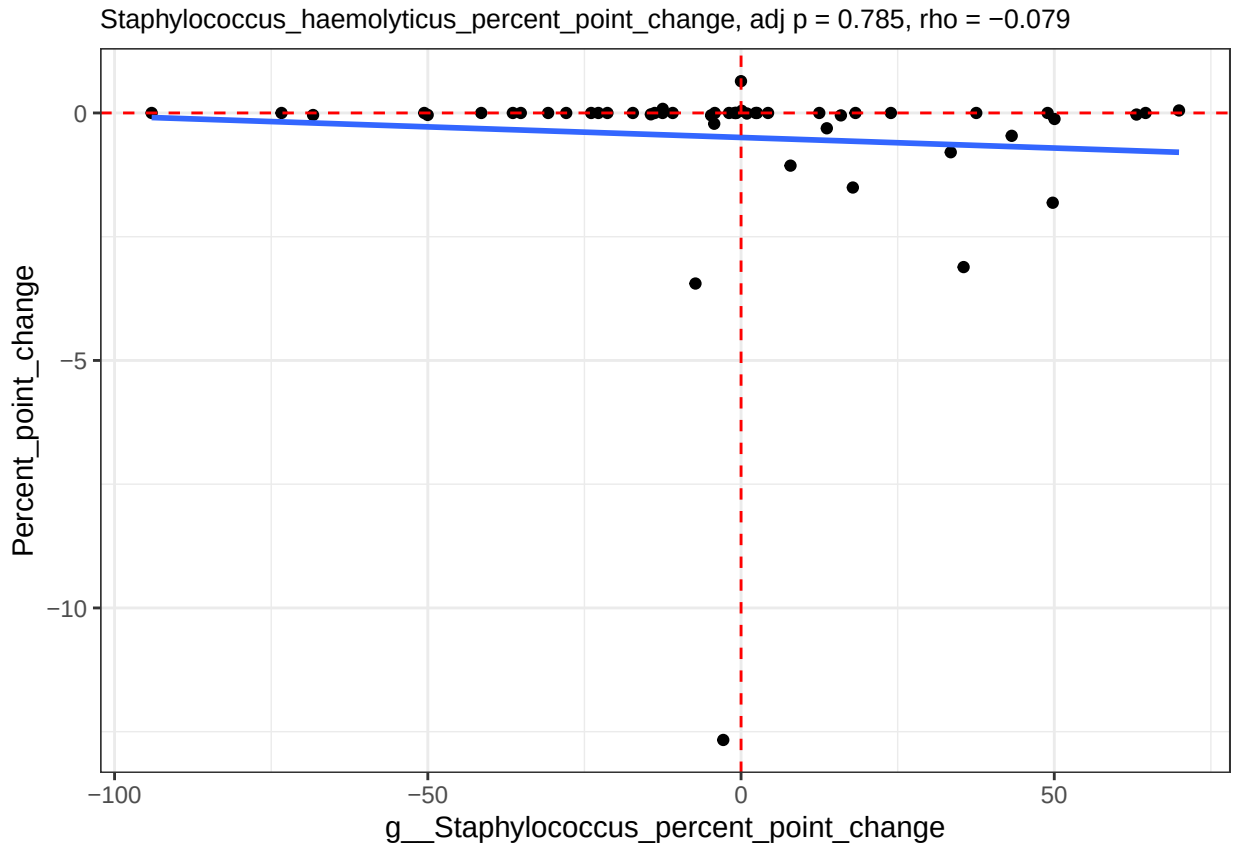

```
##
## [[2]]

## `geom_smooth()` using formula 'y ~ x'
```



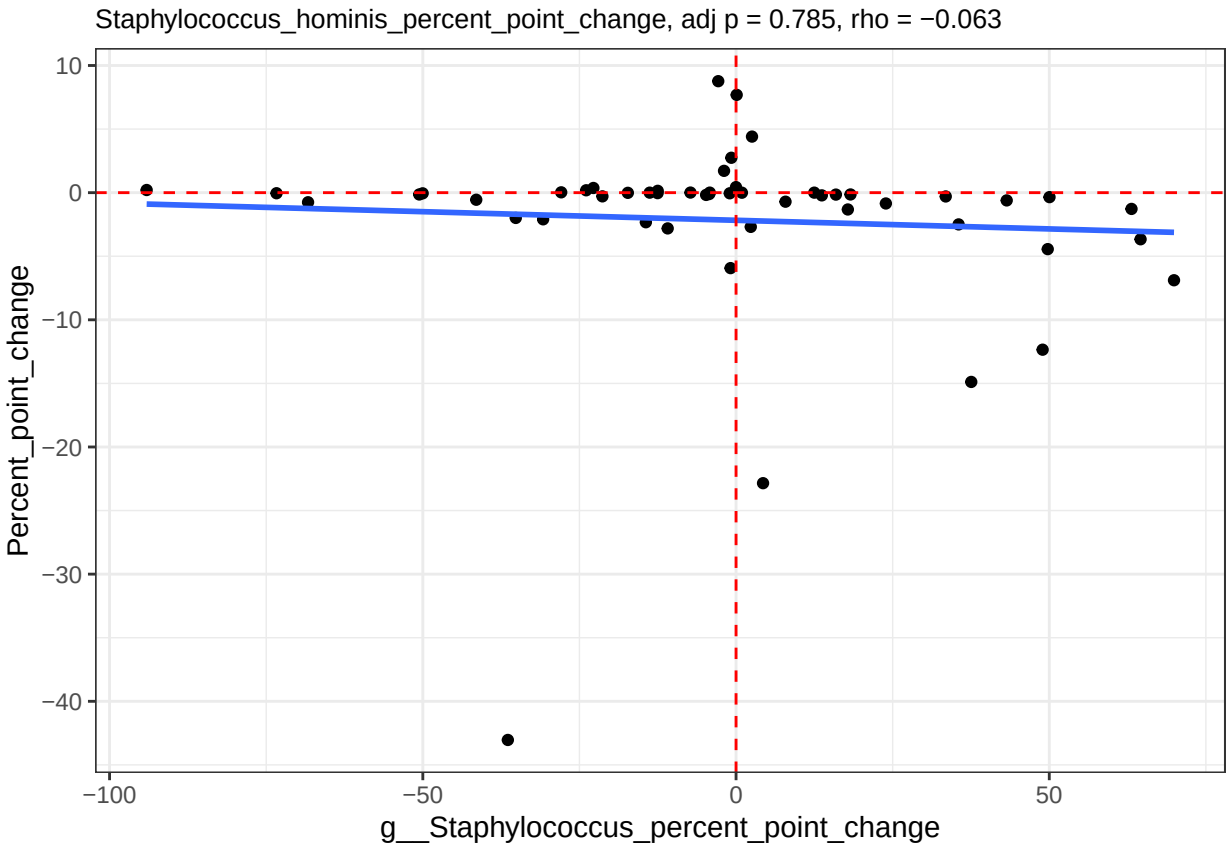

```
##
## [[3]]

## `geom_smooth()` using formula 'y ~ x'
```



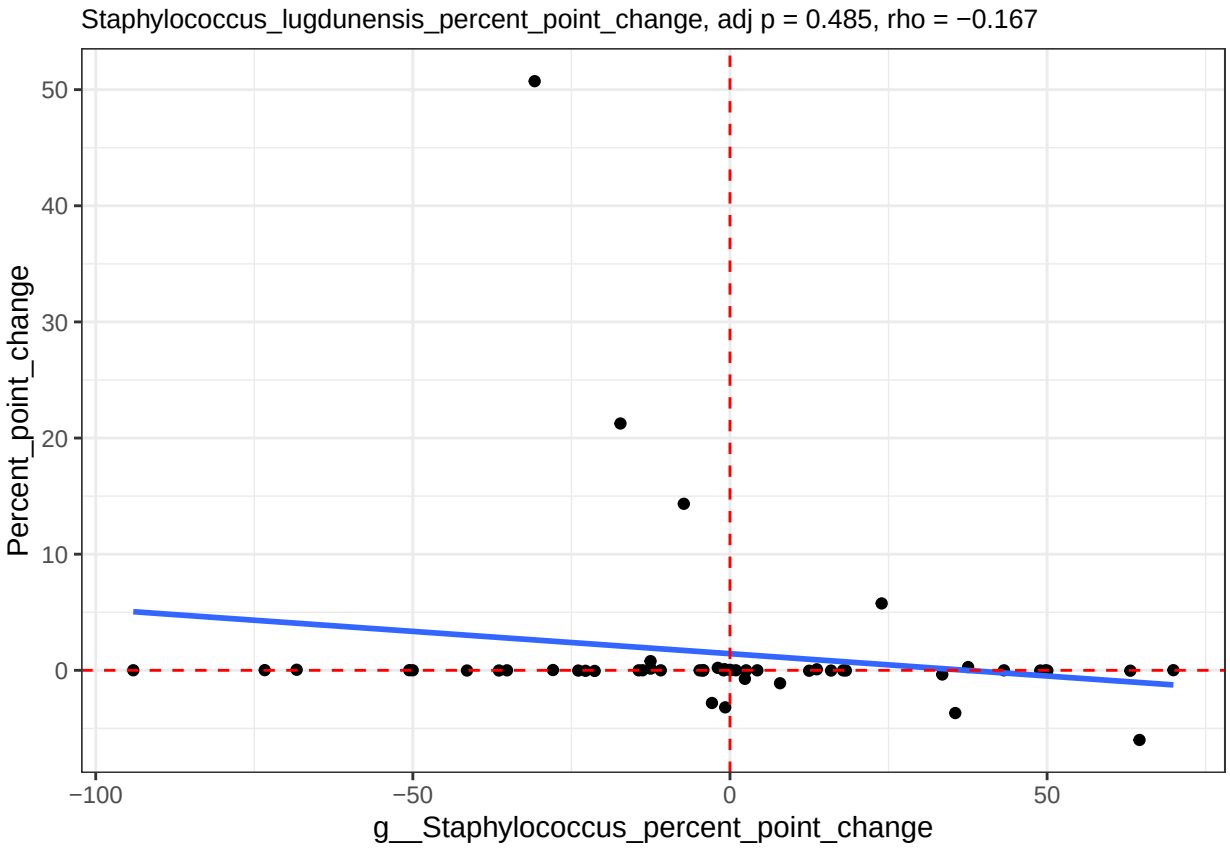

```
##
## [[4]]

## `geom_smooth()` using formula 'y ~ x'
```



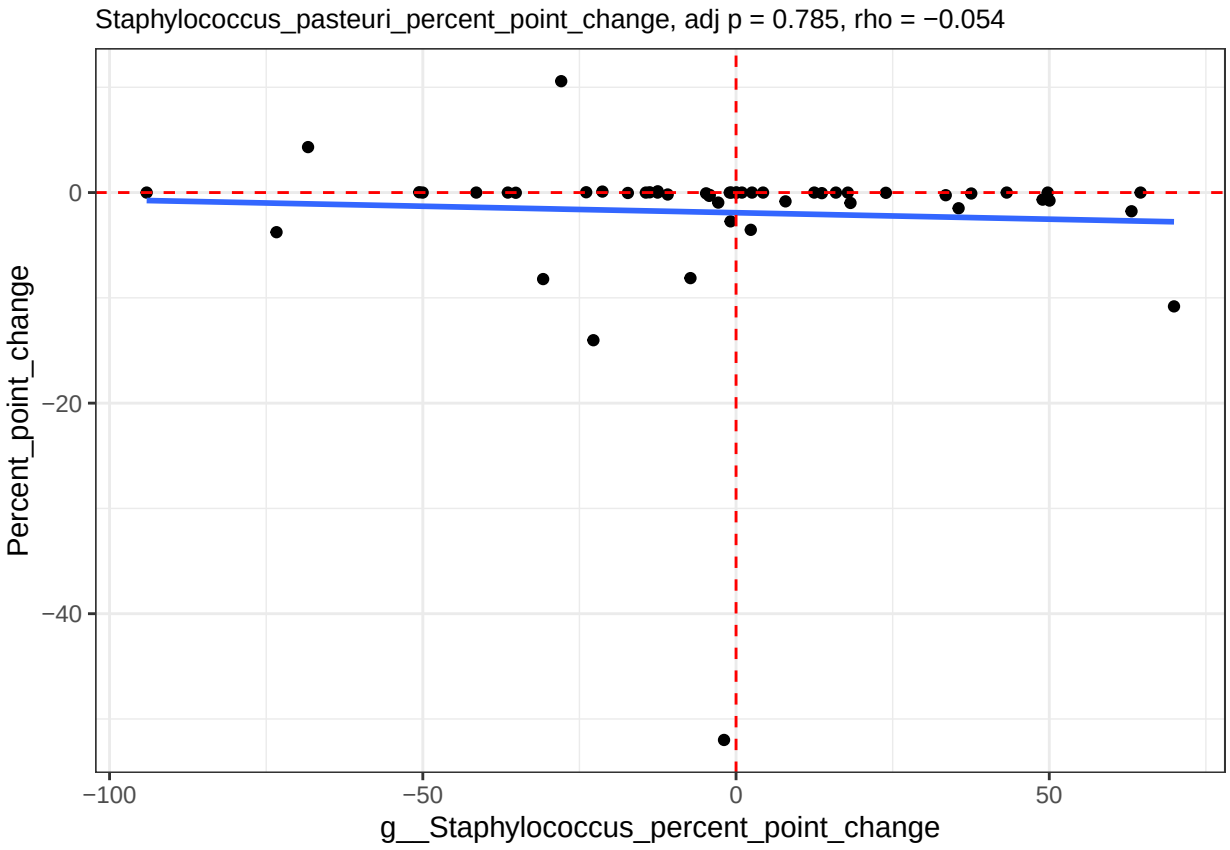

```
##
## [[5]]

## `geom_smooth()` using formula 'y ~ x'
```



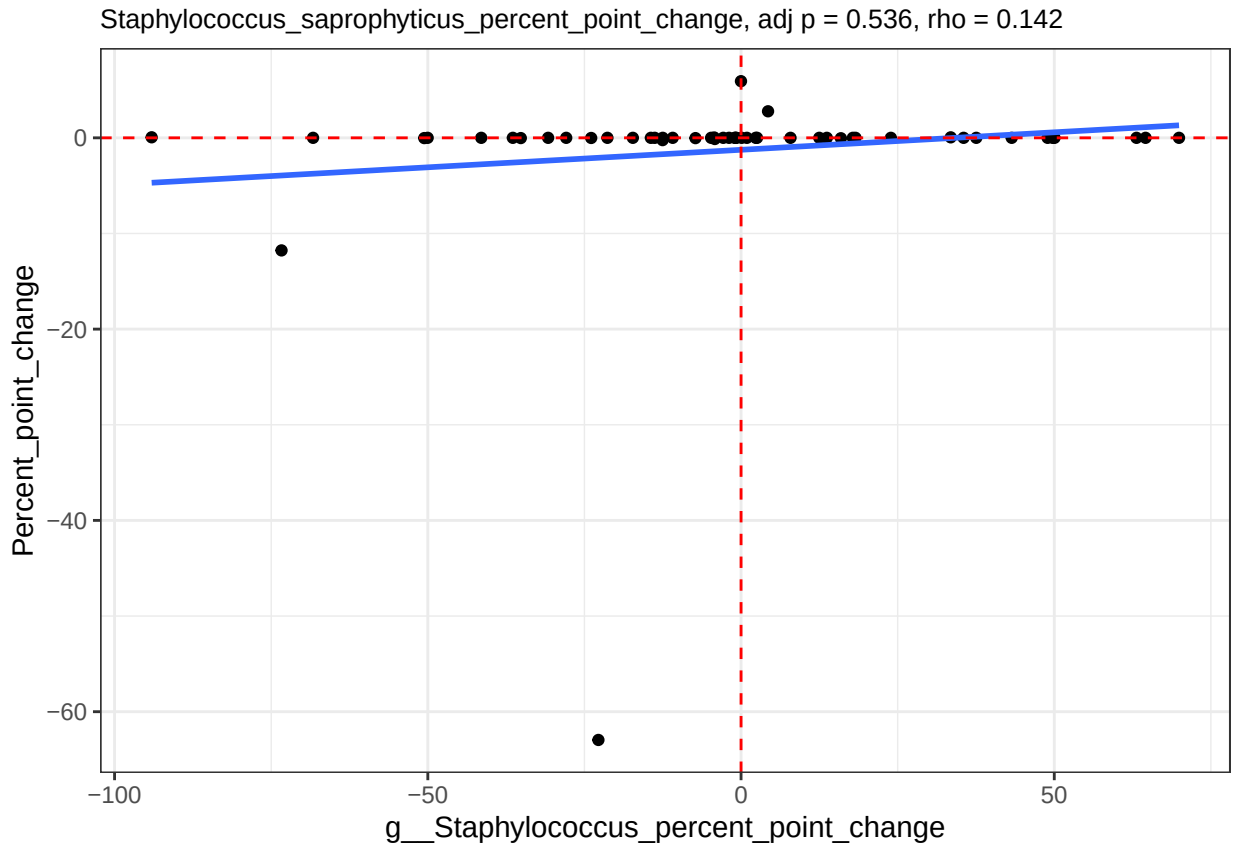

```
##
## [[6]]

## `geom_smooth()` using formula 'y ~ x'
```



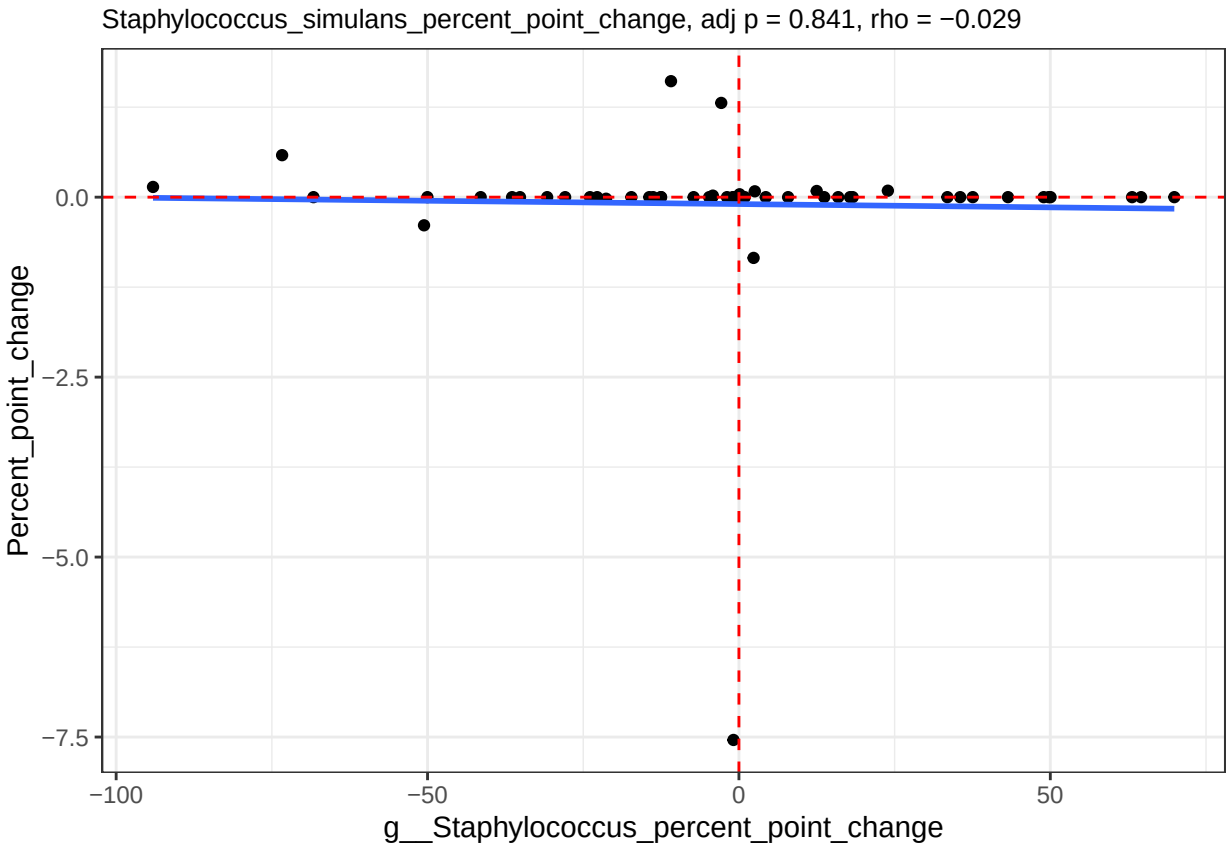

```
##
## [[7]]

## `geom_smooth()` using formula 'y ~ x'
```



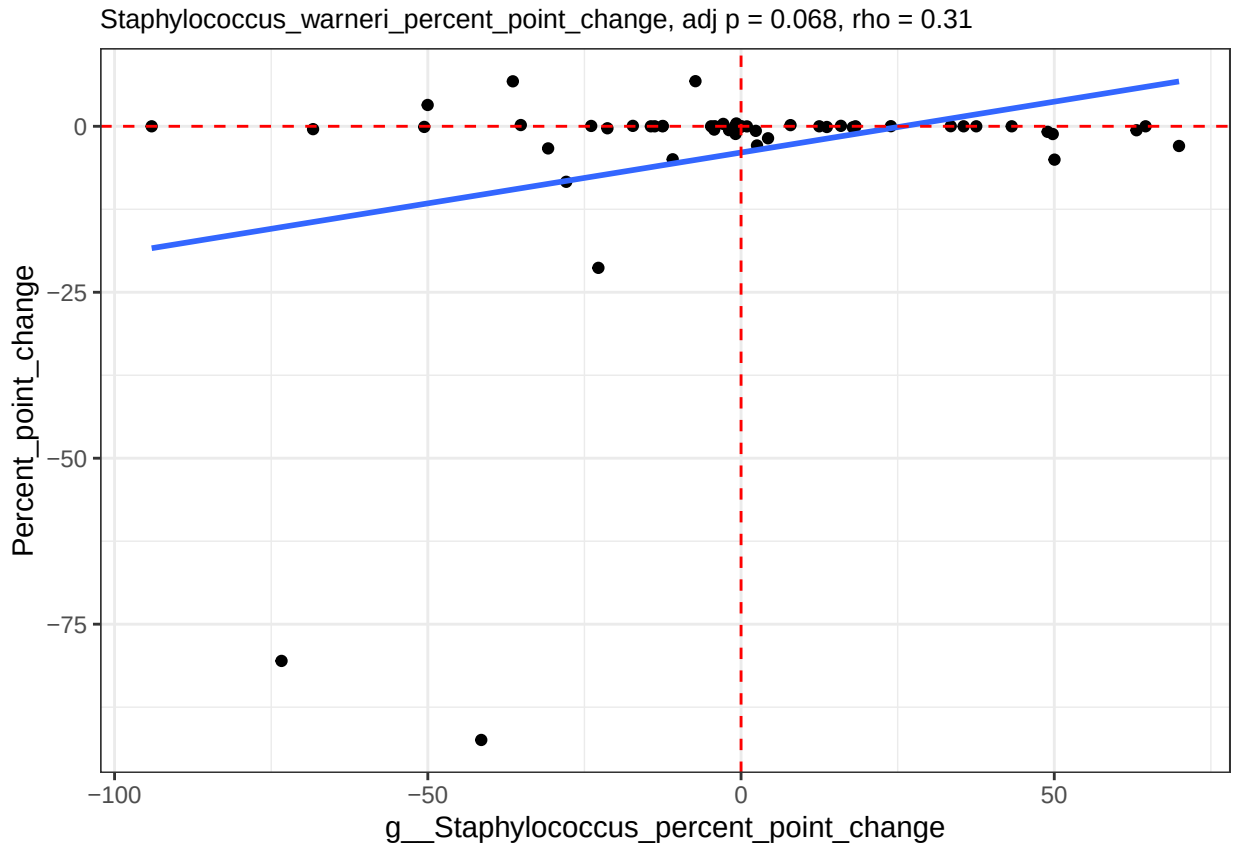

```
##
## [[8]]

## `geom_smooth()` using formula 'y ~ x'
```




```
##
## [[9]]

## `geom_smooth()` using formula 'y ~ x'
```




```
##
## [[10]]

## `geom_smooth()` using formula 'y ~ x'
```




```
pdf("plots/Staph_correlations_nose.pdf")
for (i in 1:10) {
    print(plots$plots[[i]])
}
```

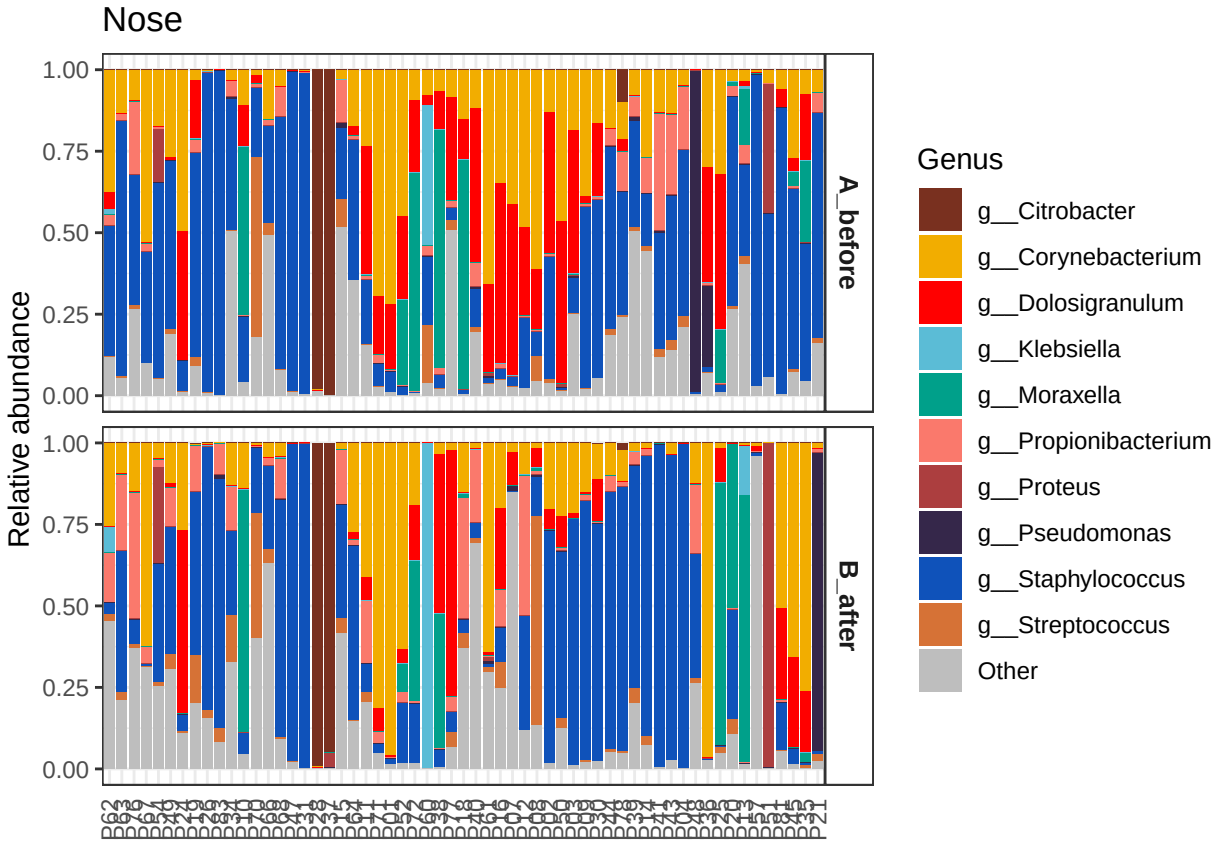
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
dev.off()
```

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

Make a version of the barplots that has the same order of patients as the heatmap

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

a <- ggplot(p_df_o, aes(x = Patient_ID, y = Abundance, fill = Genus)) +
    geom_bar(stat = "identity", width = 0.9) + facet_grid(time_point ~
    ., scales = "free") + scale_fill_manual(values = mycols) +
    theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 90, vjust = 0.5),
        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)
a
```




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

Groin

```
ps1_clean_groin <- prune_samples(sample_data(ps1_clean)$Sample_type ==
    "Groin", ps1_clean)
ps1_clean_groin <- prune_taxa(taxa_sums(ps1_clean_groin) !=
    0, ps1_clean_groin)
ps1_clean_groin
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 763 taxa and 126 samples ]
## sample_data() Sample Data:       [ 126 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 763 taxa by 7 taxonomic ranks ]
```

```
sample_data(ps1_clean_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"
```

Number of patients with groin samples overall

```
length(unique(sample_data(ps1_clean_groin)$Patient_ID))
```

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

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

```
table(sample_data(ps1_clean_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
```

Exclude 4 patients with only one time point

```
ps1_clean_groin <- prune_samples(!sample_data(ps1_clean_groin)$Patient_ID %in%
    c("P12", "P16", "P19", "P32"), ps1_clean_groin)
ps1_clean_groin
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 763 taxa and 122 samples ]
## sample_data() Sample Data:       [ 122 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 763 taxa by 7 taxonomic ranks ]
```

```
table(sample_data(ps1_clean_groin)$Patient_ID)
```

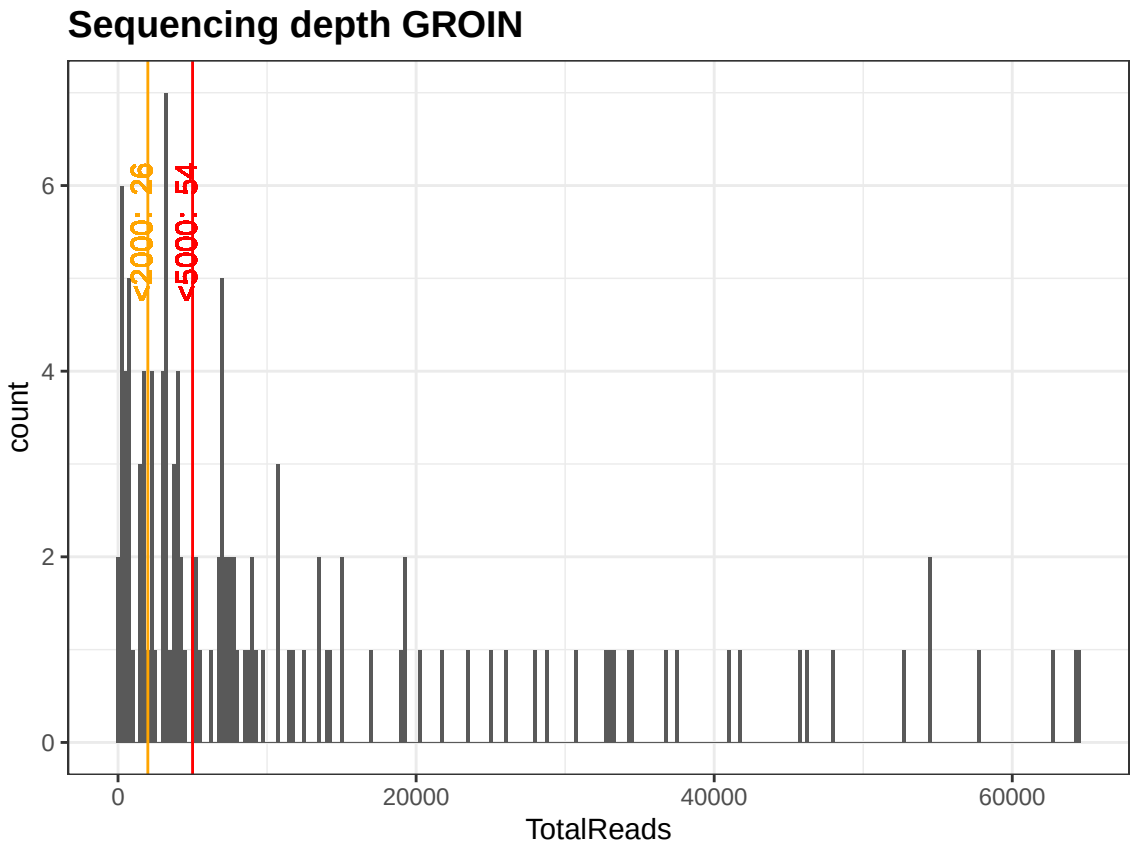
```
##
## P01 P02 P03 P04 P05 P06 P07 P09 P10 P11 P13 P15 P17 P18 P20 P21 P22 P23 P24 P25
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## P26 P27 P28 P33 P35 P36 P37 P39 P43 P45 P47 P48 P49 P50 P51 P53 P54 P55 P61 P62
##   2   2   2   2   2   2   2   2   2   2   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(ps1_clean_groin)$Patient_ID))
```

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

Seq depth

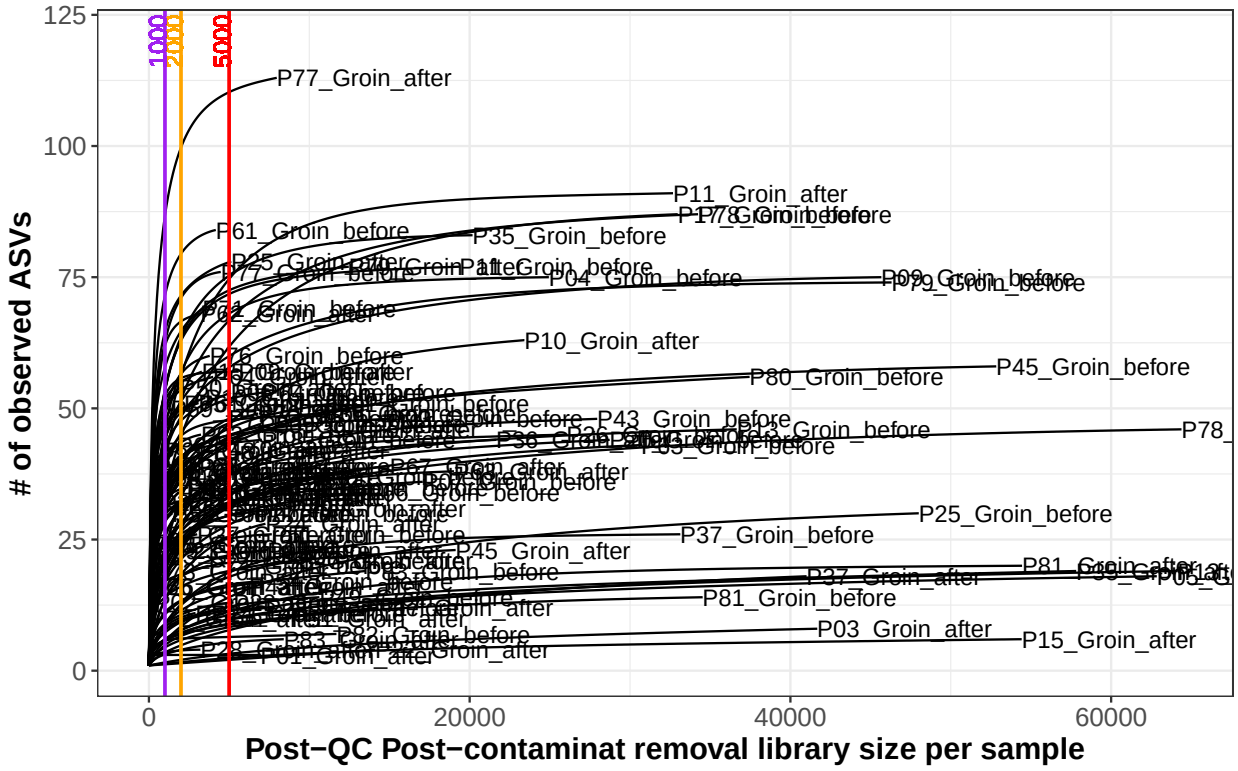
```
sdt = data.table::data.table(as(sample_data(ps1_clean_groin),
    "data.frame"), TotalReads = sample_sums(ps1_clean_groin),
    keep.rownames = TRUE)
data.table::setnames(sdt, "rn", "SampleID")
pSeqDepth = ggplot(sdt, aes(TotalReads)) + geom_histogram(binwidth = 250) +
    geom_vline(xintercept = 5000, color = "red") + geom_text(aes(x = 4550,
    label = paste("<5000: ", nrow(sdt[sdt$TotalReads < 5000])),
    y = 5.5), colour = "red", angle = 90) + geom_text(aes(x = 1550,
    label = paste("<2000: ", nrow(sdt[sdt$TotalReads < 2000])),
    y = 5.5), colour = "orange", angle = 90) + geom_vline(xintercept = 2000,
    color = "orange") + ggtitle("Sequencing depth GROIN") +
    theme(plot.title = element_text(size = 14, face = "bold"))

pSeqDepth
```



Do the rarefaction curves justify that we remove samples with reads <1000 / <2000?

Rarefaction curves

```
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 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 = 120), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 120), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 120), colour = "purple",
        angle = 90, size = 4)
p3
```

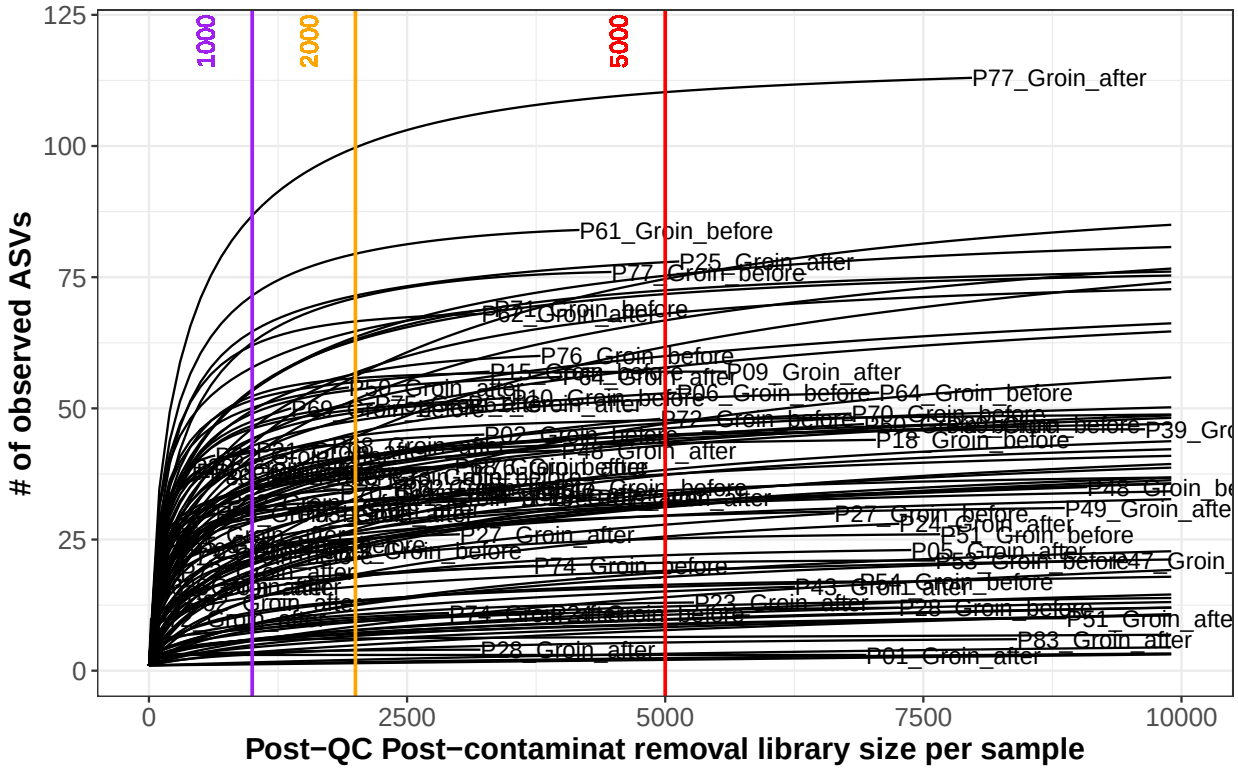


Zoom

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

```
## Warning: Removed 43 rows containing missing values (geom_text).
```

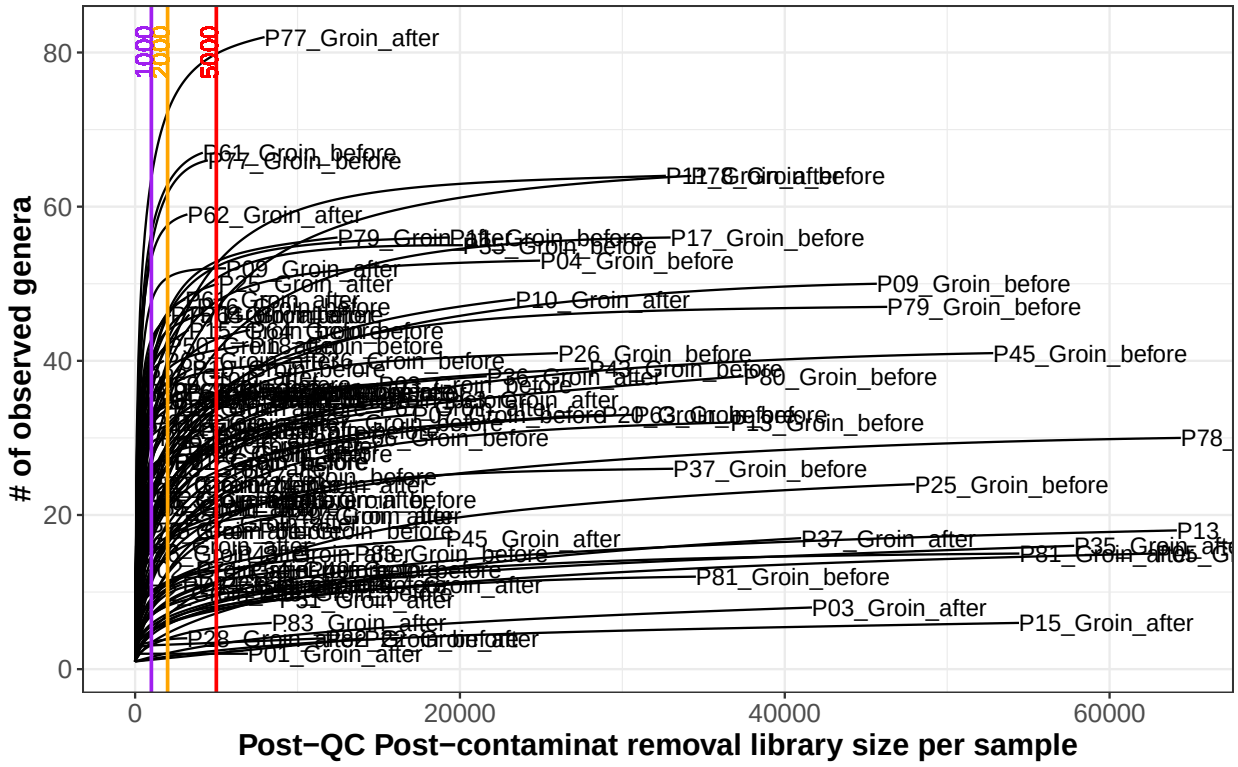
```
## Warning: Removed 8836 row(s) containing missing values (geom_path).
```



How do the rarefaction curves look on genus level?

Rarefaction curves

```
p4 <- p4 + 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 genera") + 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 = 80), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 80), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 80), colour = "purple",
        angle = 90, size = 4)
p4
```

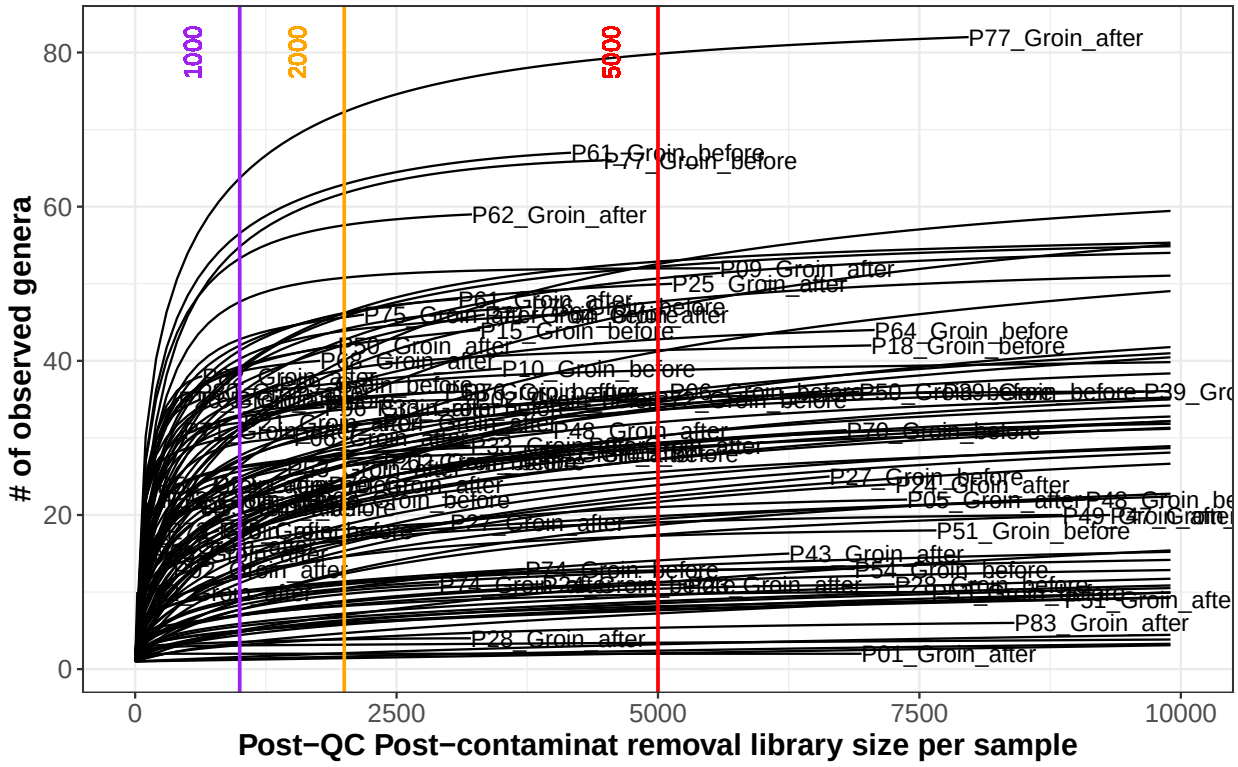


Zoom

```
p4 + xlim(0, 10000)
```

```
## Warning: Removed 43 rows containing missing values (geom_text).
```

```
## Warning: Removed 8836 row(s) containing missing values (geom_path).
```



Exclude samples with <2000 reads

```
summary(sample_sums(ps1_clean_groin))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      21    2564    6842   13164   16560   64387
```

```
ps1_clean_groin_tu <- prune_samples(!sample_sums(ps1_clean_groin) <
    2000, ps1_clean_groin)
ps1_clean_groin_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 763 taxa and 96 samples ]
## sample_data() Sample Data:       [ 96 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 763 taxa by 7 taxonomic ranks ]
```

```
summary(sample_sums(ps1_clean_groin_tu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2188    4128    8638   16505   23794   64387
```

Now how many patients still have both time points left after excluding samples <2000?

```
table(sample_data(ps1_clean_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   1   2   2   2   1   2   2   2   2   2   2   2   2   1   1   1   1   2   2
## P26 P27 P28 P33 P35 P36 P37 P39 P43 P45 P47 P48 P49 P50 P51 P53 P54 P61 P62 P63
##   1   2   2   2   2   2   2   2   2   2   2   2   2   1   2   1   1   2   2   2
## P64 P66 P67 P68 P70 P71 P72 P74 P75 P76 P77 P78 P79 P80 P81 P82 P83
##   2   1   2   1   1   1   1   2   1   2   2   2   2   1   2   1   2
```

```
length(unique(sample_data(ps1_clean_groin_tu)$Patient_ID))
```

```
## [1] 57
```

```
ps1_clean_groin_tu <- prune_samples(!sample_data(ps1_clean_groin_tu)$Patient_ID %in%
    c("P02", "P06", "P20", "P21", "P22", "P23", "P26", "P50",
        "P53", "P54", "P66", "P68", "P70", "P71", "P72",
        "P75", "P80", "P82"), ps1_clean_groin_tu)
ps1_clean_groin_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 763 taxa and 78 samples ]
## sample_data() Sample Data:       [ 78 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 763 taxa by 7 taxonomic ranks ]
```

```
length(unique(sample_data(ps1_clean_groin_tu)$Patient_ID))
```

```
## [1] 39
```

39 patients left with 2 time points for groin

Read counts in the remaining patients with 2 samples >2000 reads

```
summary(sample_sums(ps1_clean_groin_tu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2315    4211    9206   17887   27200   64387
```

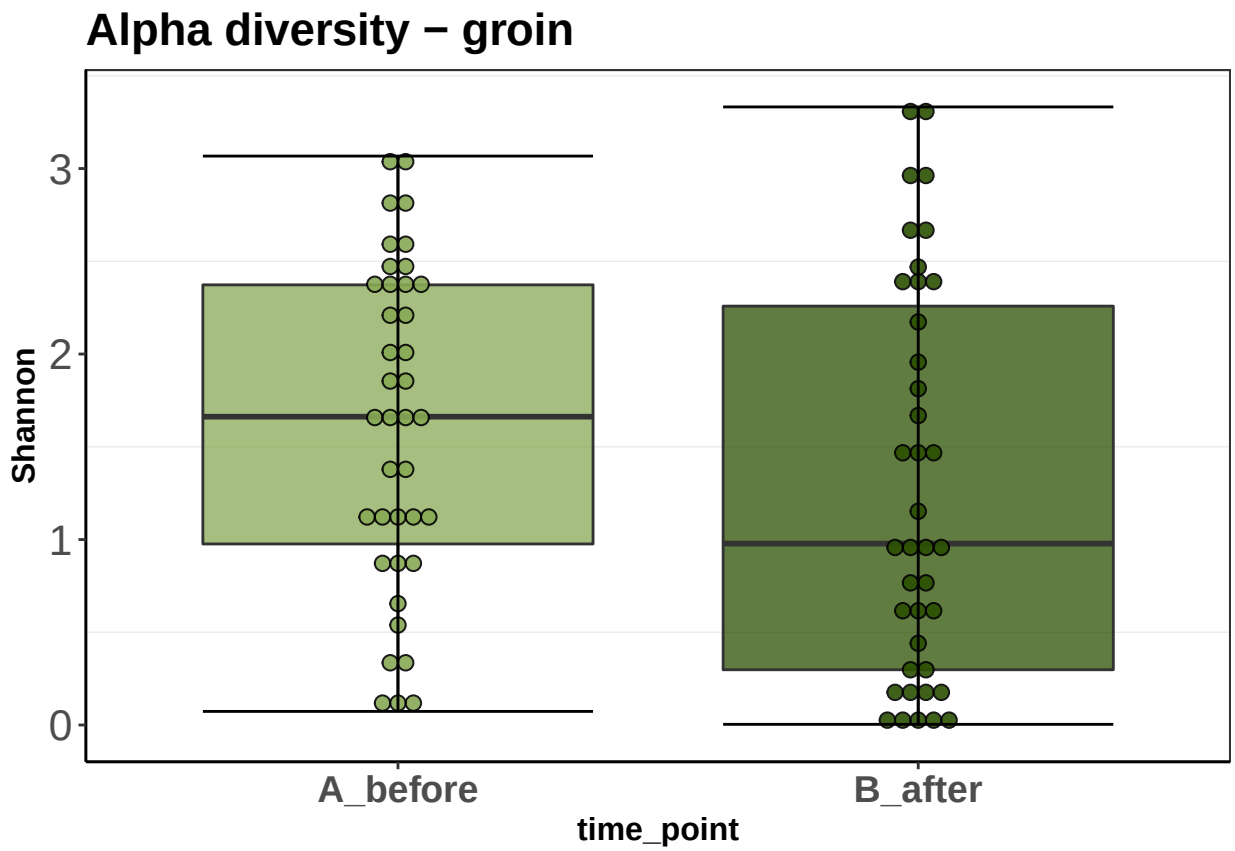
Alpha diversity

```
#### Add diversity measures to the phyloseq object as
#### variables
alpha_div_raw <- estimate_richness(ps1_clean_groin_tu, measures = c("Observed",
    "Chao1", "Shannon", "InvSimpson"))
rownames(alpha_div_raw) <- gsub("X", "", rownames(alpha_div_raw))
ps1_clean_groin_tu <- merge_phyloseq(ps1_clean_groin_tu,
    sample_data(alpha_div_raw))
df_ps1_clean_groin_tu <- as(sample_data(ps1_clean_groin_tu),
    "data.frame")
```

Shannon diversity over time:

```
plot_g_Shannon <- ggplot(df_ps1_clean_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("Alpha diversity - groin")
plot_g_Shannon
```

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




```
ggsave(filename = "plots/Groin_alpha_div_16S.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_ps1_clean_groin_tu_c <- dcast(df_ps1_clean_groin_tu,
    Patient_ID ~ time_point, value.var = "Shannon", drop = FALSE)
wilcox.test(df_ps1_clean_groin_tu_c$A_before, df_ps1_clean_groin_tu_c$B_after,
    paired = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  df_ps1_clean_groin_tu_c$A_before and df_ps1_clean_groin_tu_c$B_after
## V = 534, p-value = 0.04436
## alternative hypothesis: true location shift is not equal to 0
```

Significant decrease in alpha diversity in the groin.

Agglomerate on Genus level

```
rank_names(ps1_clean_groin_tu)
```

```
## [1] "Kingdom" "Phylum"  "Class"   "Order"   "Family"  "Genus"   "Species"
```

```
ps1_clean_groin_tu_gs <- tax_glom(ps1_clean_groin_tu, taxrank = "Genus")
ps1_clean_groin_tu_gs
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 327 taxa and 78 samples ]
## sample_data() Sample Data:       [ 78 samples by 9 sample variables ]
## tax_table()   Taxonomy Table:    [ 327 taxa by 7 taxonomic ranks ]
```

```
ps1_clean_groin_tu_gs <- prune_taxa(taxa_sums(ps1_clean_groin_tu_gs) !=
    0, ps1_clean_groin_tu_gs)
ps1_clean_groin_tu_gs
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 297 taxa and 78 samples ]
## sample_data() Sample Data:       [ 78 samples by 9 sample variables ]
## tax_table()   Taxonomy Table:    [ 297 taxa by 7 taxonomic ranks ]
```

```
rank_names(ps1_clean_groin_tu_gs)
```

```
## [1] "Kingdom" "Phylum"  "Class"   "Order"   "Family"  "Genus"   "Species"
```

Convert to relative abundance

```
ps1_clean_groin_tu_gs_rel <- transform_sample_counts(ps1_clean_groin_tu_gs,
    function(x) x/sum(x))
summary(sample_sums(ps1_clean_groin_tu_gs_rel))
```

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

Subset top 10 genera

```
Genus10 = names(sort(taxa_sums(ps1_clean_groin_tu_gs_rel),
    TRUE)[1:10])
```

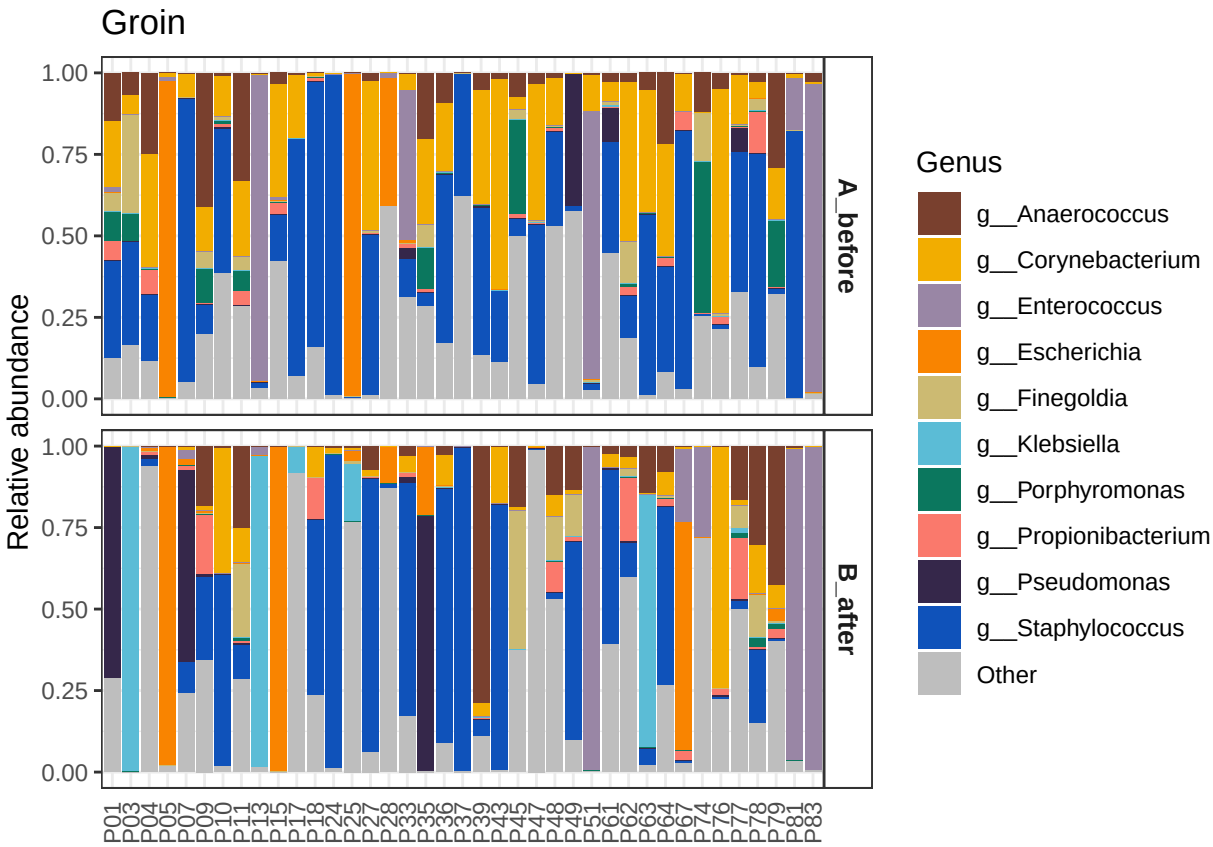
to data frame

```
p_df_o_groin <- psmelt(ps1_clean_groin_tu_gs_rel)
p_df_o_groin$Genus <- as.character(p_df_o_groin$Genus)
p_df_o_groin$Genus[!(p_df_o_groin$OTU %in% Genus10)] <- "Other"
```

Barplots of relative abundance

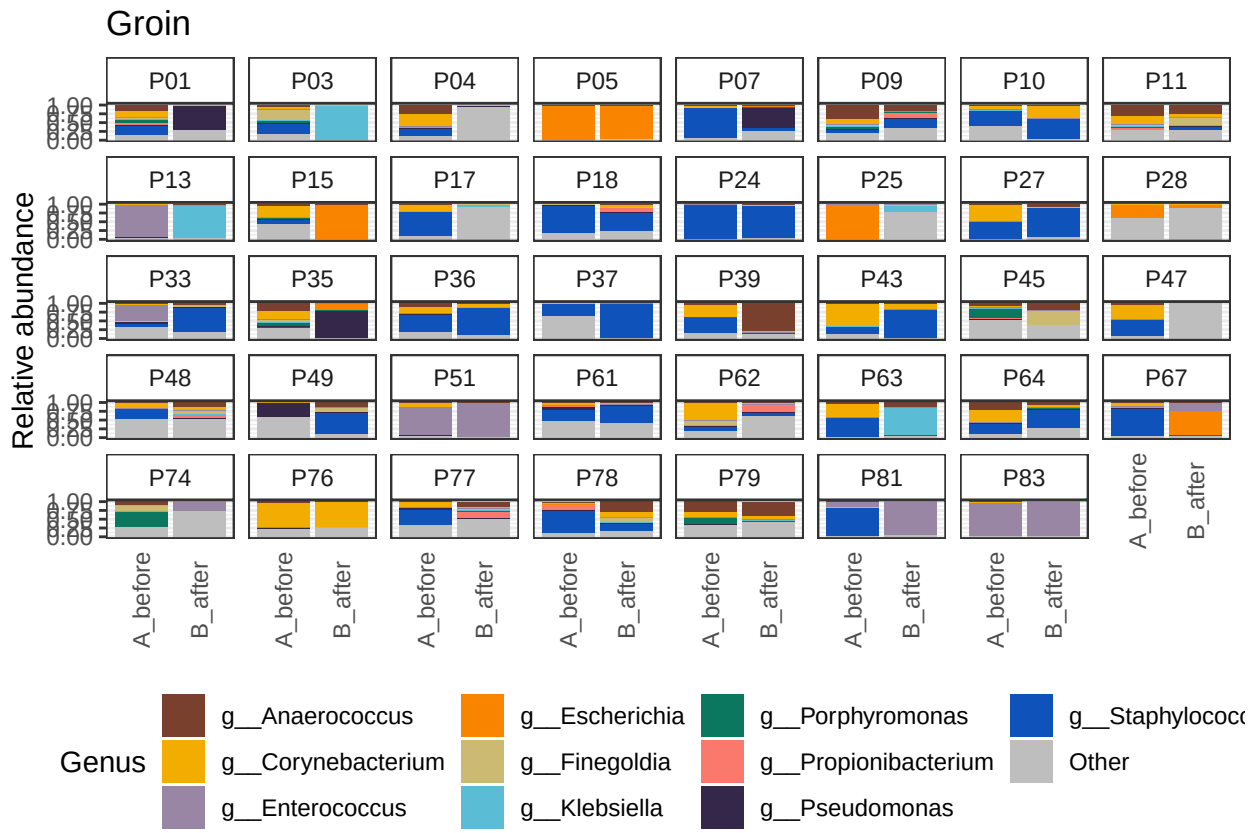
The patients are not in the same order here as in the heatmap, because in the heatmap they are ordered by clustering and here just by number (see ordered version below)

```
b <- ggplot(p_df_o_groin, aes(x = Patient_ID, y = Abundance,
    fill = Genus)) + geom_bar(stat = "identity", width = 0.9) +
    facet_grid(time_point ~ ., scales = "free") + scale_fill_manual(values = mycols) +
    theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 90, vjust = 0.5),
        strip.background = element_rect(fill = "white"),
        strip.text.y = element_text(size = 10, face = "bold")) +
    ylab("Relative abundance") + ggtitle("Groin")

b
```



Patient-wise plots

```
ggplot(p_df_o_groin, aes(x = time_point, y = Abundance,
    fill = Genus)) + geom_bar(stat = "identity", width = 0.9) +
    facet_wrap(. ~ Patient_ID, nrow = 5) + scale_fill_manual(values = mycols) +
    theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 90, 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("Groin")
```



Subset the top 10 genera (without other)

```
ps1_clean_groin_tu_gs_rel <- prune_taxa(taxa_names(ps1_clean_groin_tu_gs_rel) %in%
    Genus10, ps1_clean_groin_tu_gs_rel)
ps1_clean_groin_tu_gs_rel
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 10 taxa and 78 samples ]
## sample_data() Sample Data:       [ 78 samples by 9 sample variables ]
## tax_table()   Taxonomy Table:    [ 10 taxa by 7 taxonomic ranks ]
```

to data frame

```
p_df <- psmelt(ps1_clean_groin_tu_gs_rel)
p_df_d <- dcast(p_df, Patient_ID + Genus ~ 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 ~ Patient_ID,
    value.var = "Percent_point_change")
```

Visualize in a heatmap

```
pdf(file = "plots/Groin_heatmap.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 = 1.5, cexCol = 1, 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 genera do significantly change from before to after?

(Paired Wilcoxon test)

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

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

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

wilc_dfG
```

```
## # A tibble: 10 x 3
##    Genus                wilcox_p_value BH_adjusted_wilcox_p_value
##    <chr>                         <dbl>                      <dbl>
##  1 g__Anaerococcus            0.542                      0.775
##  2 g__Corynebacterium         0.000135                   0.00135
##  3 g__Enterococcus            0.751                      0.939
##  4 g__Escherichia             0.155                      0.310
##  5 g__Finegoldia              0.451                      0.751
##  6 g__Klebsiella              0.126                      0.310
##  7 g__Porphyromonas           0.0341                     0.171
##  8 g__Propionibacterium       0.964                      0.964
##  9 g__Pseudomonas             0.945                      0.964
## 10 g__Staphylococcus          0.108                      0.310
```

Corynebacterium and Porphyromonas have a significant *overall* change in the groin. After multiple testing correction, only Coryne is significant (Yay, that actually fits with Thor's finding back in the days)

Do they *overall* decrease or increase?

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

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

```
## # A tibble: 10 x 2
##    Genus                Mean_percent_point_change
##    <chr>                                   <dbl>
##  1 g__Anaerococcus                       0.0105
##  2 g__Corynebacterium                   -0.119
##  3 g__Enterococcus                       0.00313
##  4 g__Escherichia                        0.0192
##  5 g__Finegoldia                         0.00542
##  6 g__Klebsiella                         0.0768
##  7 g__Porphyromonas                     -0.0355
##  8 g__Propionibacterium                  0.0107
##  9 g__Pseudomonas                        0.0388
## 10 g__Staphylococcus                    -0.0812
```

Corynebacterium and Porphyromonas decrease *overall* in the groin.

Combine with tuf data:

Does change in the genus Staphylococcus correlate with change in individual Staph species?

```
p_df_d_STAPH <- p_df_d %>% select(Patient_ID, Genus, Percent_point_change) %>%
    filter(Genus == "g__Staphylococcus") %>% select(Patient_ID,
    Percent_point_change) %>% dplyr::rename(g__Staphylococcus_percent_point_change = Percent_point_chan
dim(p_df_d_STAPH)
```

```
## [1] 39  2
```

```
p_df_d_tuf_groin <- read.table(file = "tables/p_df_d_tuf_groin.csv",
    sep = ";", header = TRUE)
p_df_d_tuf_groin <- p_df_d_tuf_groin %>% rename_at(vars(Staphylococcus_aureus:Staphylococcus_sciuri),
    function(x) {
        paste0(x, "_percent_point_change")
    })
dim(p_df_d_tuf_groin)
```

```
## [1] 41 11
```

```
## Subset to the same patients for which we have 16S data
p_df_d_tuf_groin <- p_df_d_tuf_groin[p_df_d_tuf_groin$Patient_ID %in%
    p_df_d_STAPH$Patient_ID, ]
dim(p_df_d_tuf_groin)
```

```
## [1] 28 11
```

```
p_df_d_STAPH <- p_df_d_STAPH[p_df_d_STAPH$Patient_ID %in%
    p_df_d_tuf_groin$Patient_ID, ]

p_df_d_STAPH1 <- left_join(p_df_d_STAPH, p_df_d_tuf_groin,
    by = "Patient_ID")
dim(p_df_d_STAPH1)
```

```
## [1] 28 12
```

For those patients that both have 16S and tuf data:

Test Staph genus correlation with all Staph species:

```
p_df_d_STAPH2 <- p_df_d_STAPH1 %>% pivot_longer(cols = starts_with("Staphylococcus"),
    names_to = "Species", values_to = "Percent_point_change")

test_res <- p_df_d_STAPH2 %>% group_by(Species) %>% group_modify(~broom::tidy(cor.test(~g__Staphylococcu
    Percent_point_change, data = .x)))

test_res$BH_adjusted_p_value <- p.adjust(test_res$p.value,
    method = "BH")
test_res
```

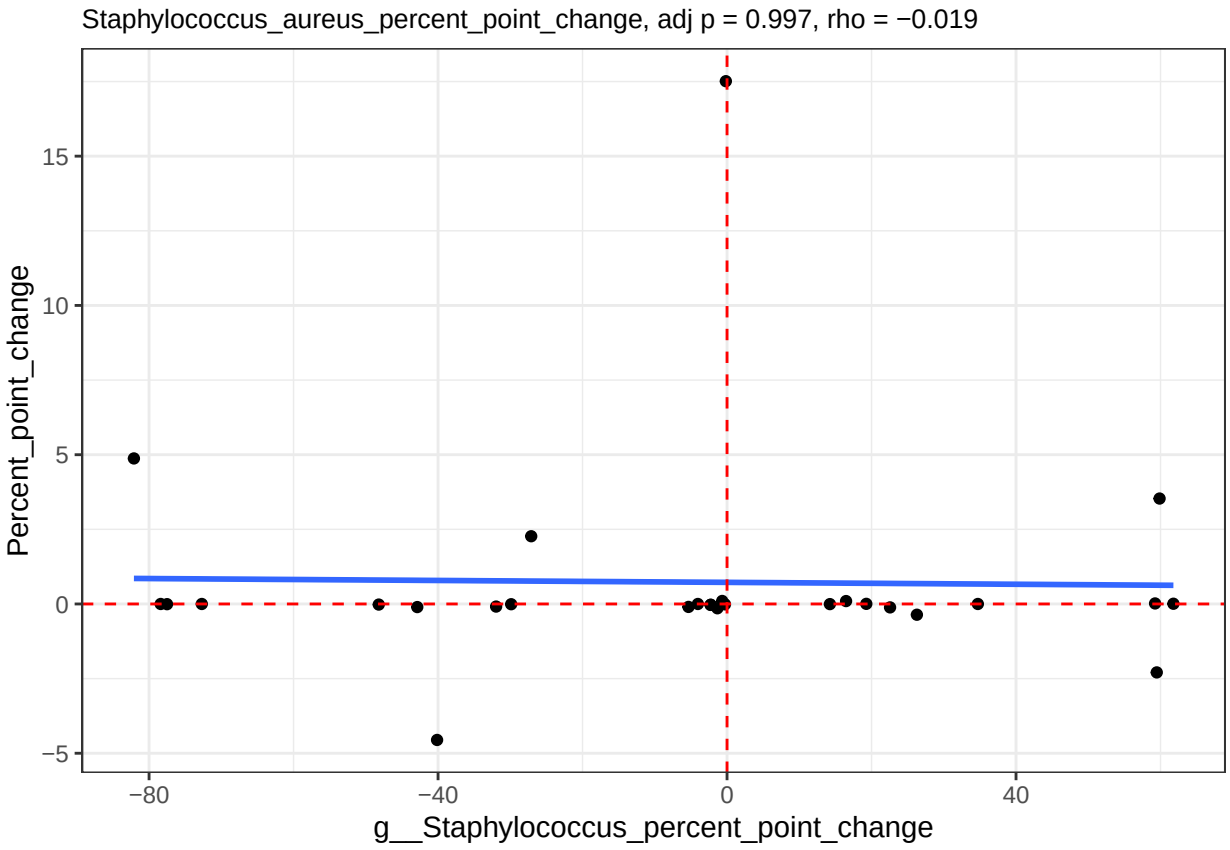
```
## # A tibble: 10 x 10
## # Groups:   Species [10]
##    Species estimate statistic p.value parameter conf.low conf.high method
##    <chr>      <dbl>     <dbl>   <dbl>     <int>    <dbl>     <dbl> <chr>
##  1 Staphy~ -1.87e-2  -0.0952   0.925         26  -0.389     0.357  Pears~
##  2 Staphy~  8.38e-4   0.00427  0.997         26  -0.372     0.374  Pears~
##  3 Staphy~ -4.51e-2  -0.230    0.820         26  -0.411     0.334  Pears~
##  4 Staphy~ -1.28e-1  -0.658    0.516         26  -0.478     0.257  Pears~
##  5 Staphy~  3.52e-1   1.92     0.0665        26  -0.0247    0.641  Pears~
```

```
##  6 Staphy~  7.09e-2   0.362    0.720         26  -0.310     0.433  Pears~
##  7 Staphy~ -1.96e-1  -1.02     0.317         26  -0.530     0.191  Pears~
##  8 Staphy~  3.47e-1   1.88     0.0707        26  -0.0303    0.637  Pears~
##  9 Staphy~ -2.94e-1  -1.57     0.129         26  -0.601     0.0892 Pears~
## 10 Staphy~ -6.25e-2  -0.320    0.752         26  -0.426     0.318  Pears~
## # ... with 2 more variables: alternative <chr>, BH_adjusted_p_value <dbl>
```

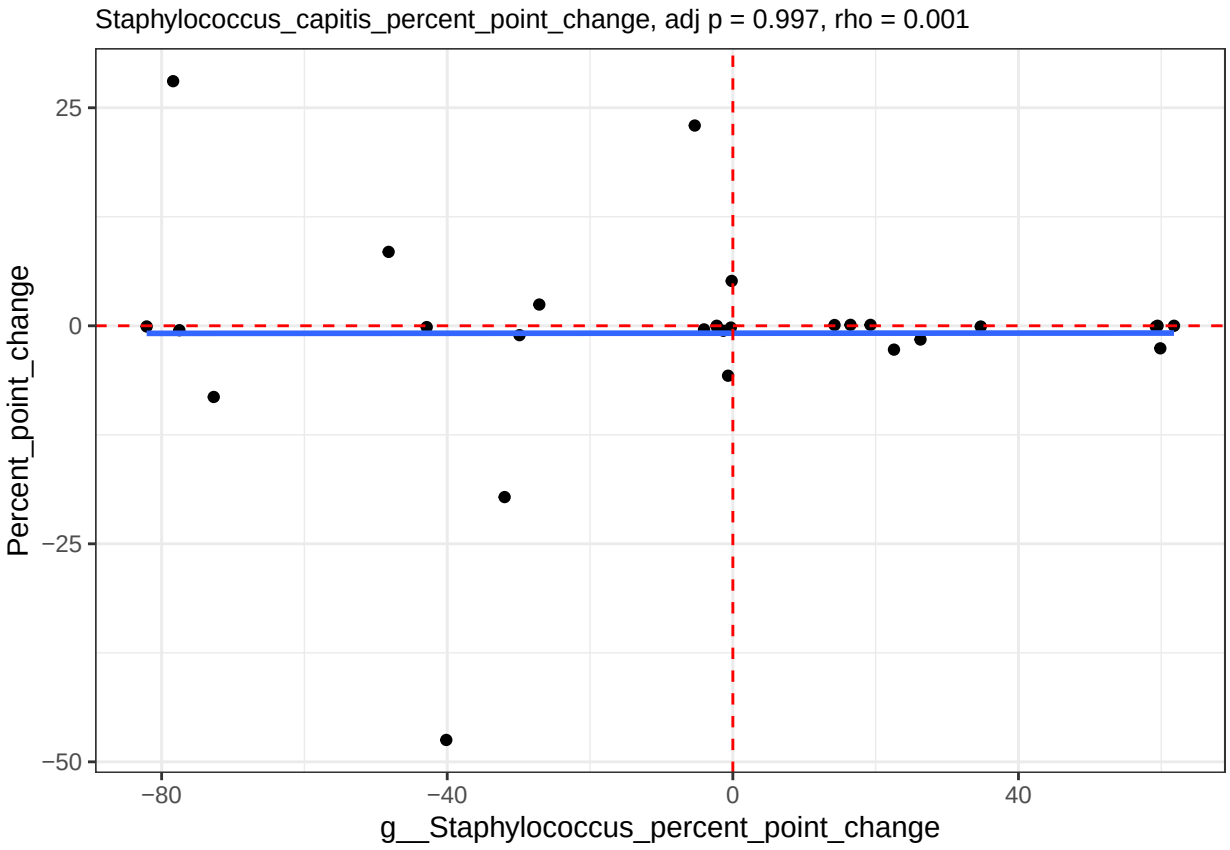
Estimate is the Pearson's correlation coefficient.

```
test_res1 <- as.data.frame(test_res[, c("Species", "estimate",
    "BH_adjusted_p_value")])

p_df_d_STAPH2 <- dplyr::left_join(p_df_d_STAPH2, test_res,
    by = "Species")
p_df_d_STAPH2 <- as.data.frame(p_df_d_STAPH2)


p_df_d_STAPH2 <- p_df_d_STAPH2 %>% mutate_at(vars(BH_adjusted_p_value,
    estimate), round, 3)

plots <- p_df_d_STAPH2 %>% group_by(Species) %>% do(plots = ggplot(data = .) +
    aes(x = g__Staphylococcus_percent_point_change, y = Percent_point_change) +
    ggtitle(paste0(unique(.$Species), ", adj p = ", unique(.$BH_adjusted_p_value),
        ", rho = ", unique(.$estimate))) + geom_point() +
    geom_smooth(method = "lm", se = FALSE) + geom_vline(xintercept = 0,
    color = "red", linetype = "dashed") + theme(plot.title = element_text(size = 10)) +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed"))
plots$plots
```

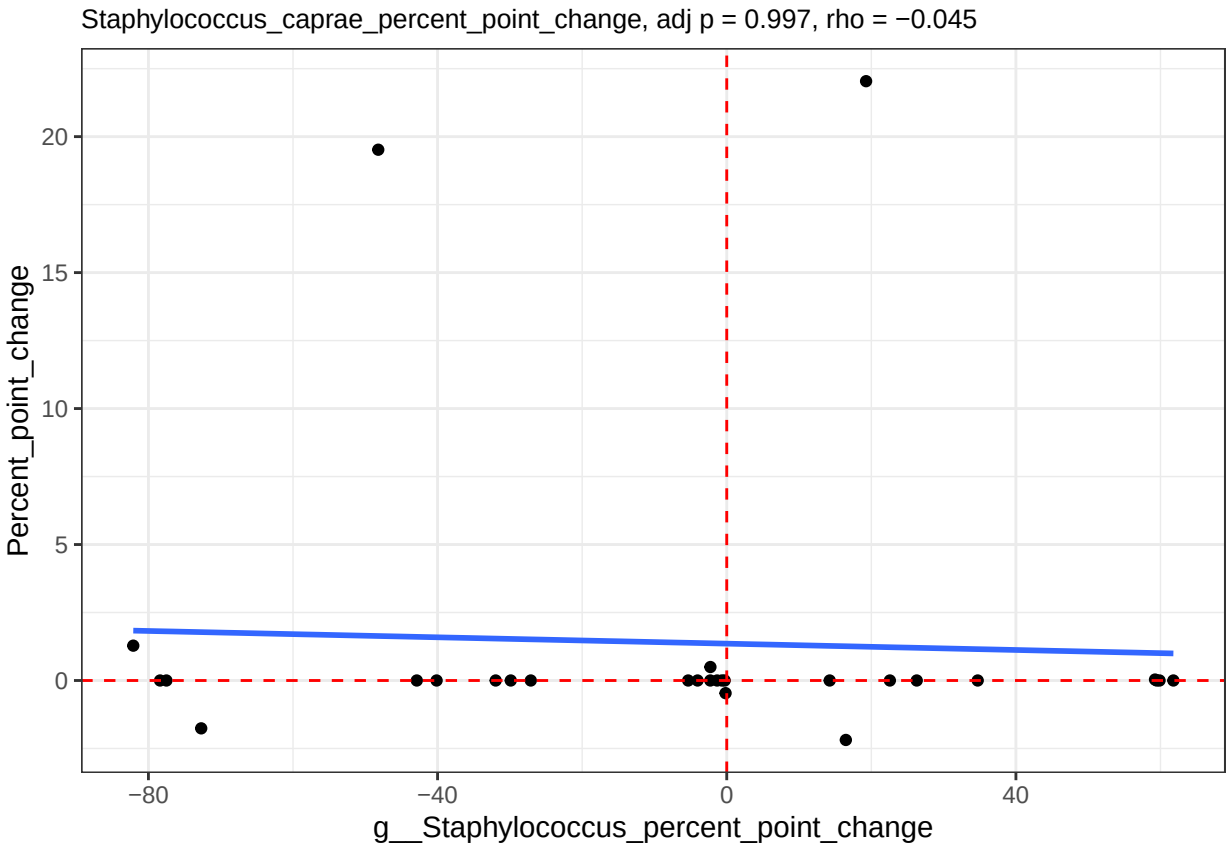
```
## [[1]]

## `geom_smooth()` using formula 'y ~ x'
```



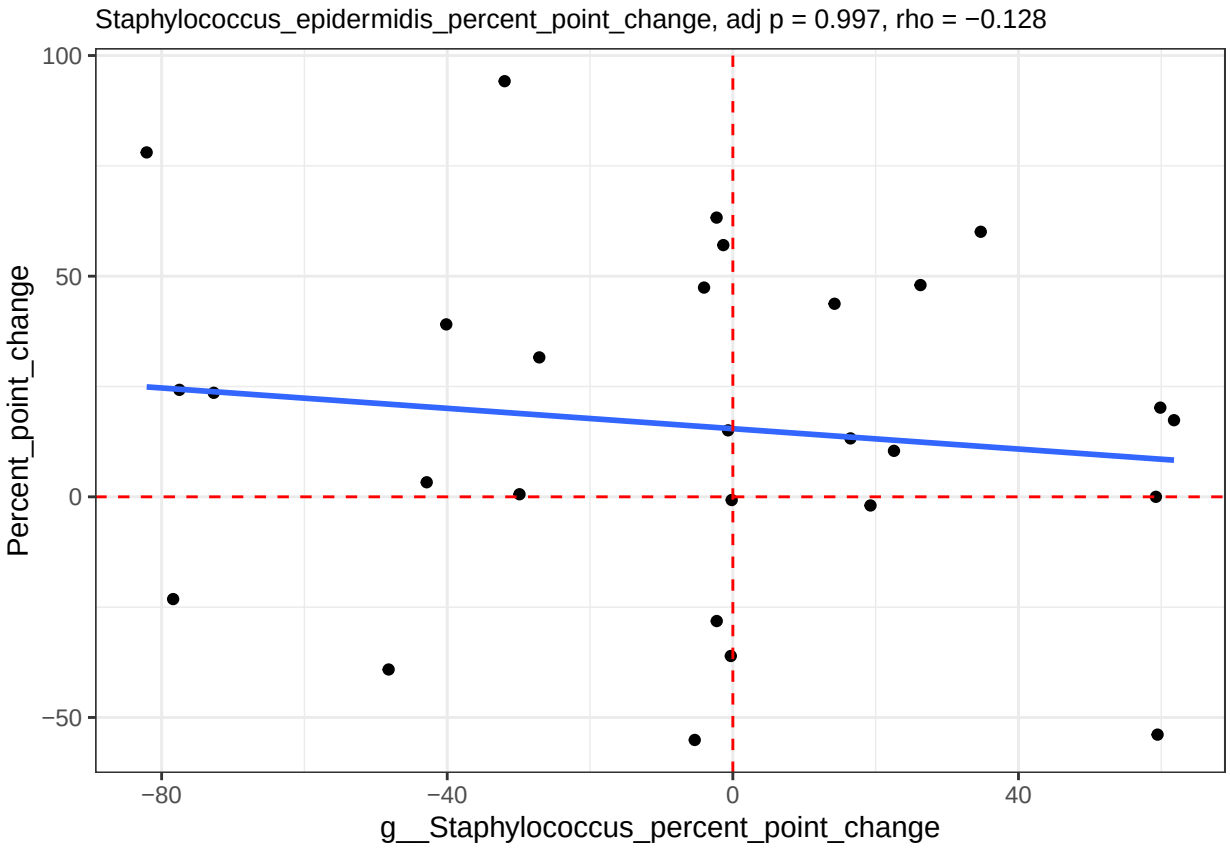

```
##
## [[2]]

## `geom_smooth()` using formula 'y ~ x'
```



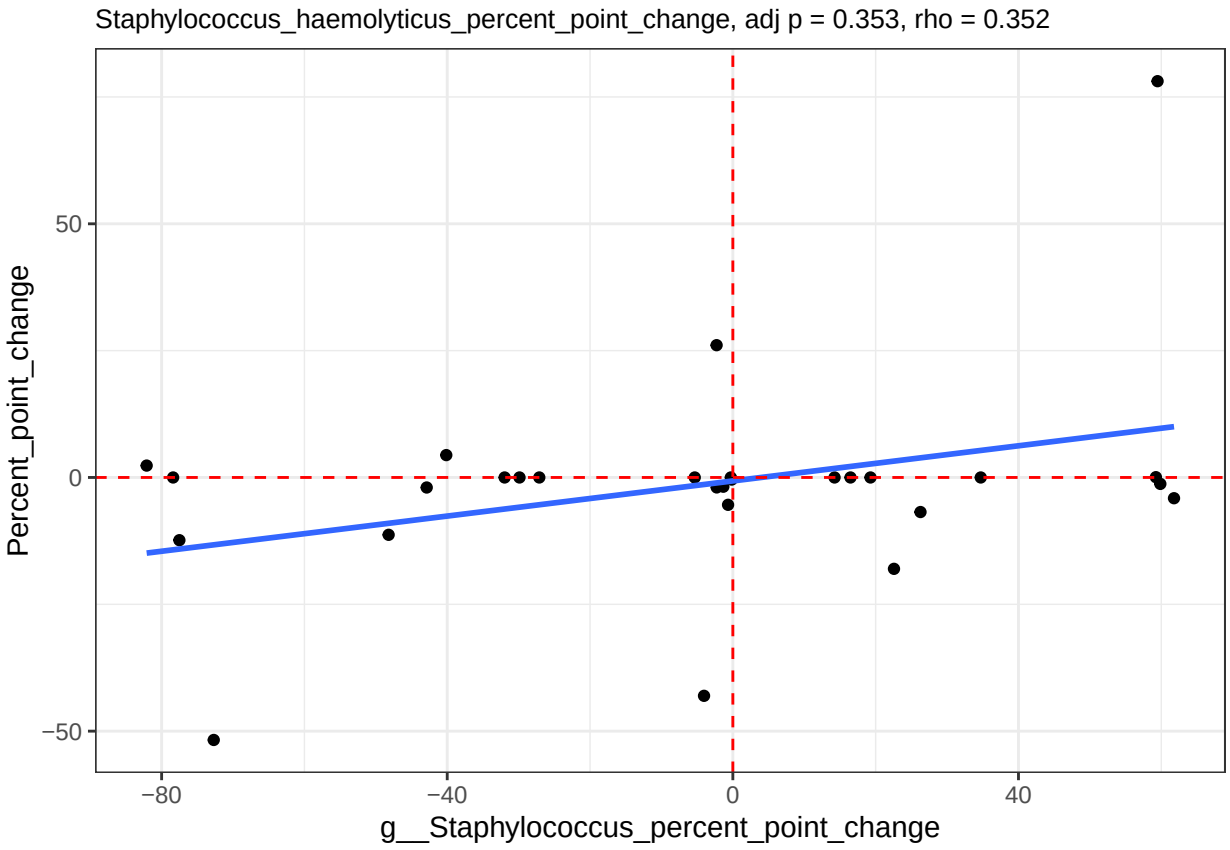

```
##
## [[3]]

## `geom_smooth()` using formula 'y ~ x'
```



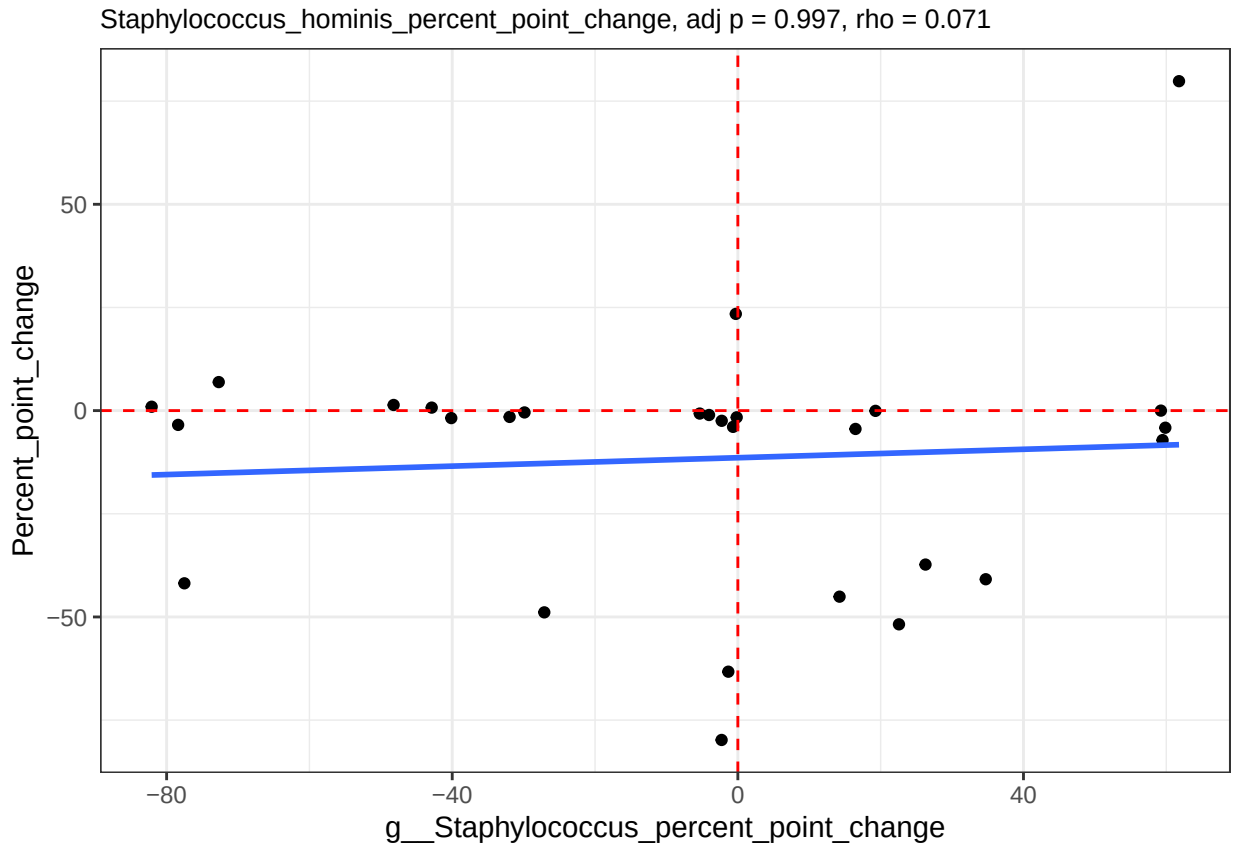

```
##
## [[4]]

## `geom_smooth()` using formula 'y ~ x'
```



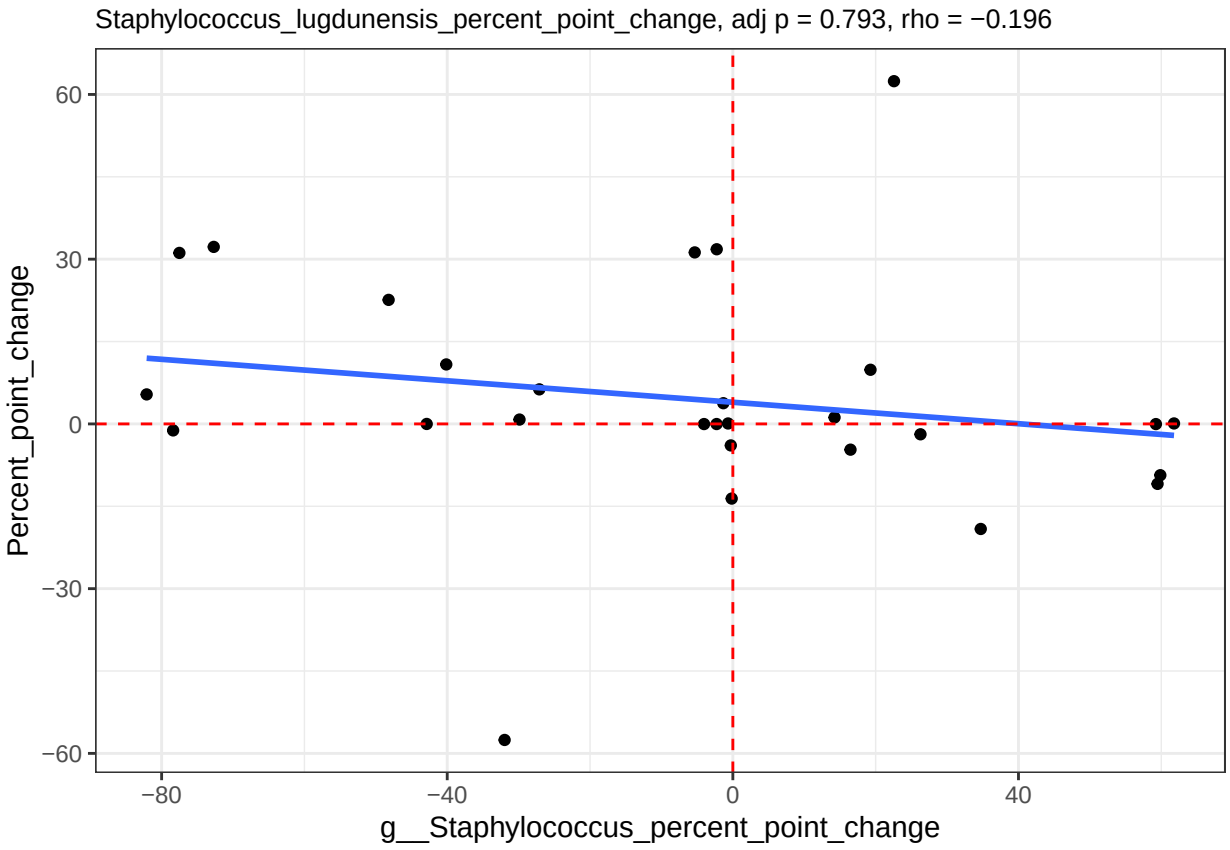

```
##
## [[5]]

## `geom_smooth()` using formula 'y ~ x'
```



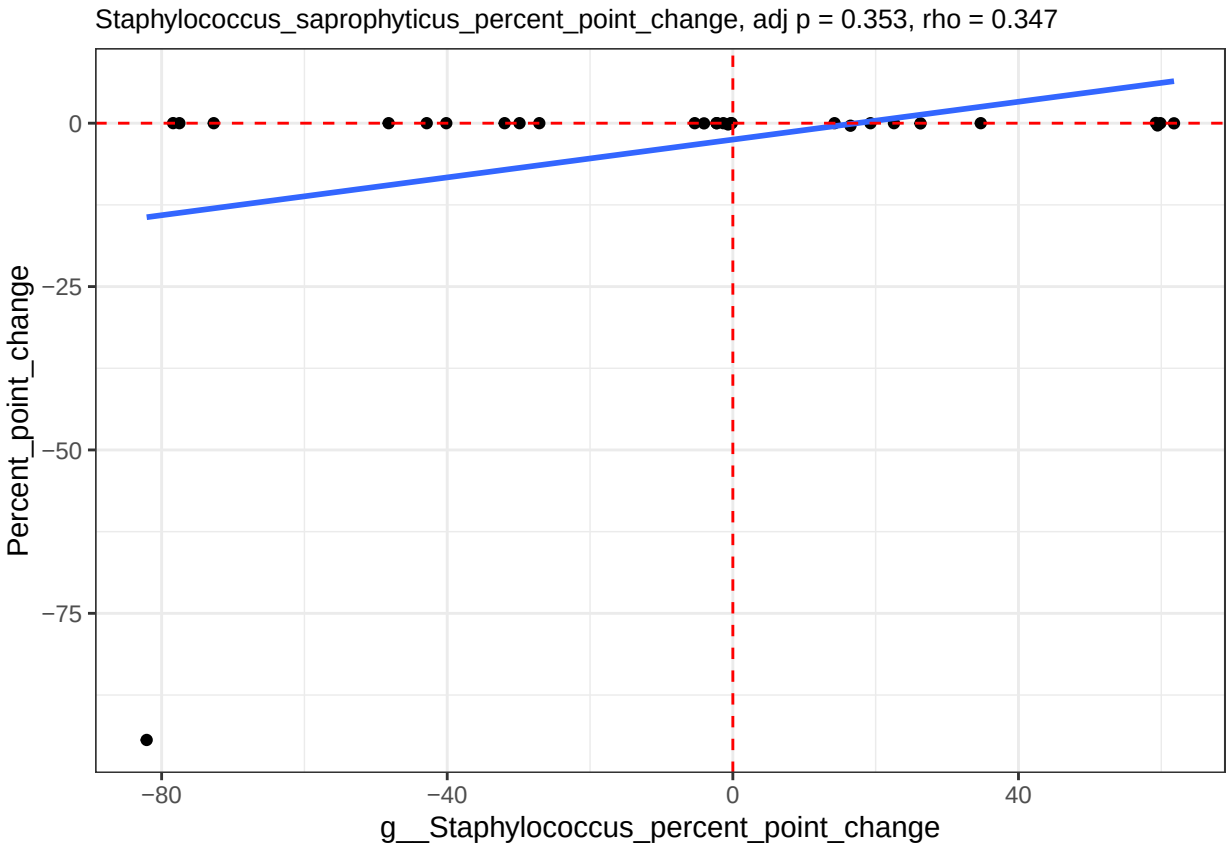

```
##
## [[6]]

## `geom_smooth()` using formula 'y ~ x'
```



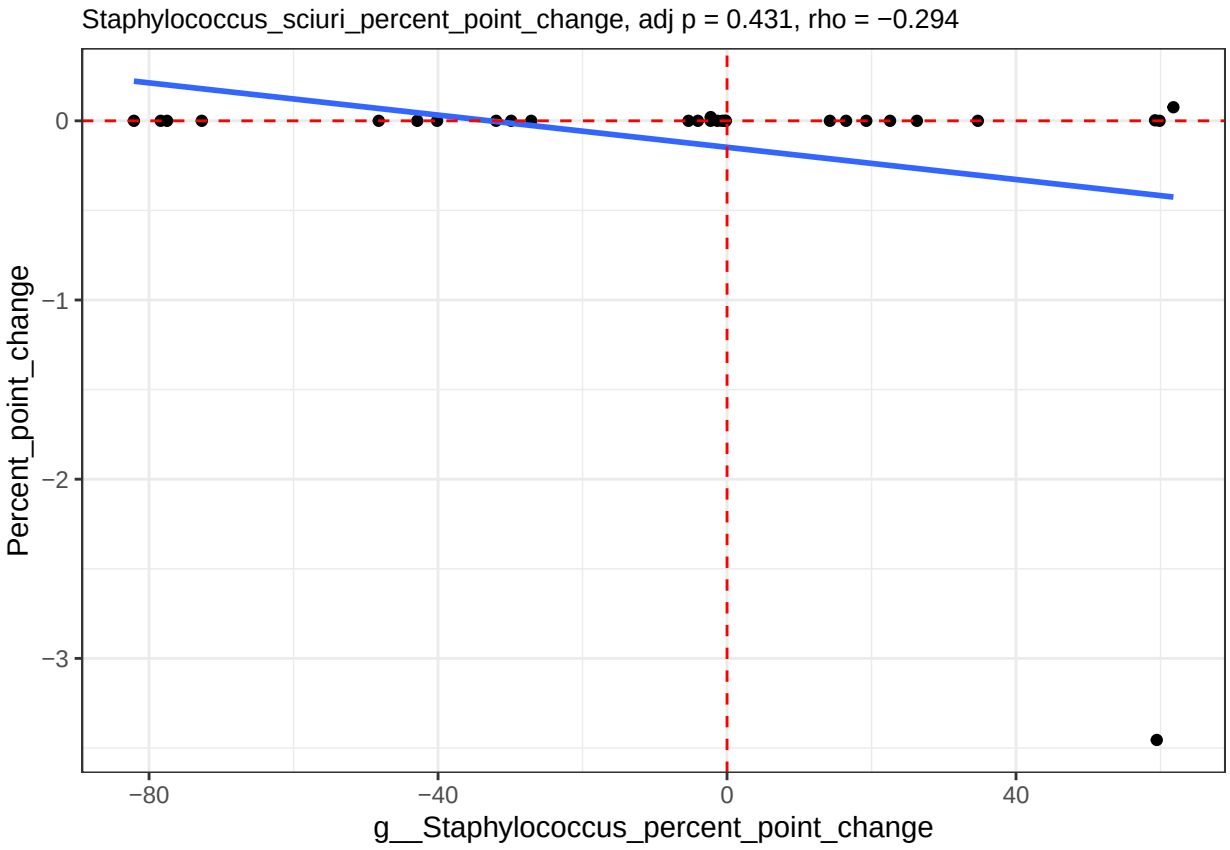

```
##
## [[7]]

## `geom_smooth()` using formula 'y ~ x'
```



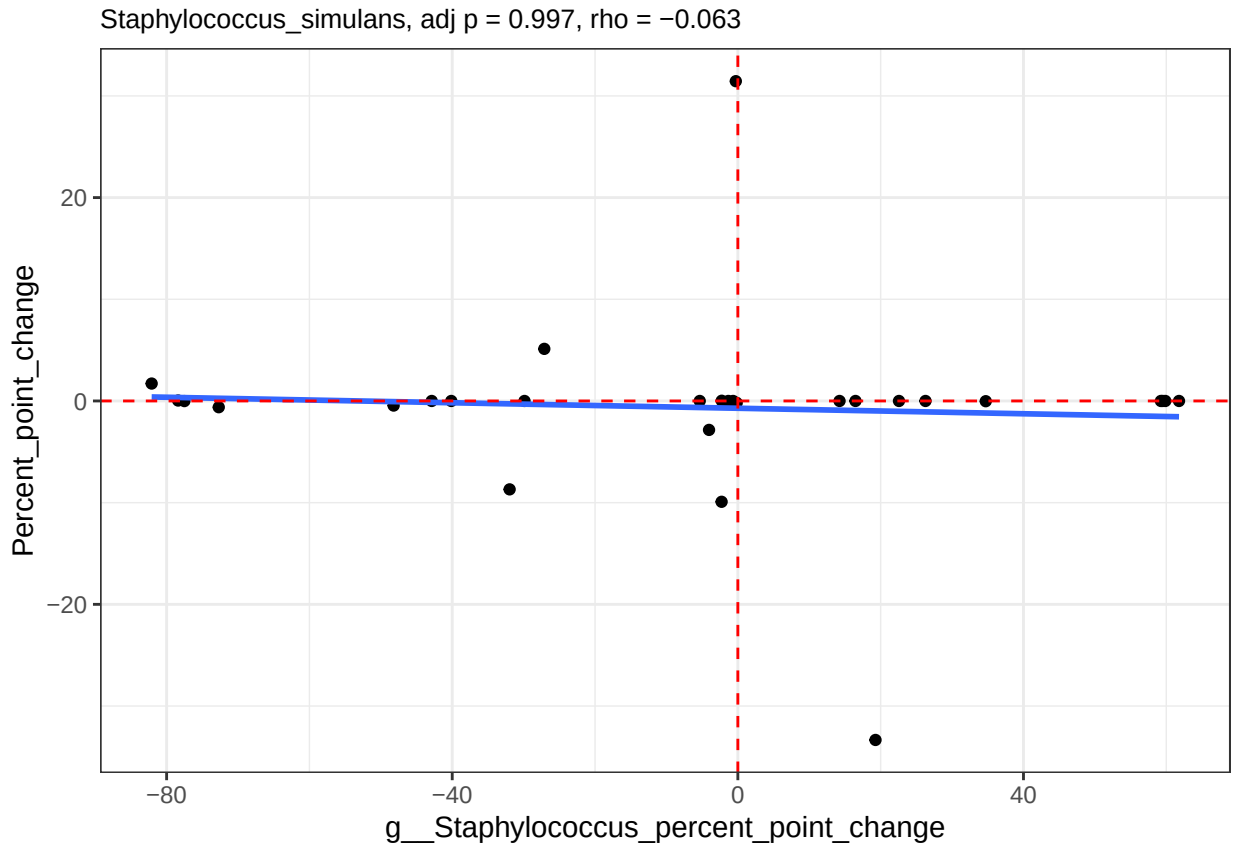

```
##
## [[8]]

## `geom_smooth()` using formula 'y ~ x'
```




```
##
## [[9]]

## `geom_smooth()` using formula 'y ~ x'
```




```
##
## [[10]]

## `geom_smooth()` using formula 'y ~ x'
```




```
pdf("plots/Staph_correlations_groin.pdf")
for (i in 1:10) {
    print(plots$plots[[i]])
}
```

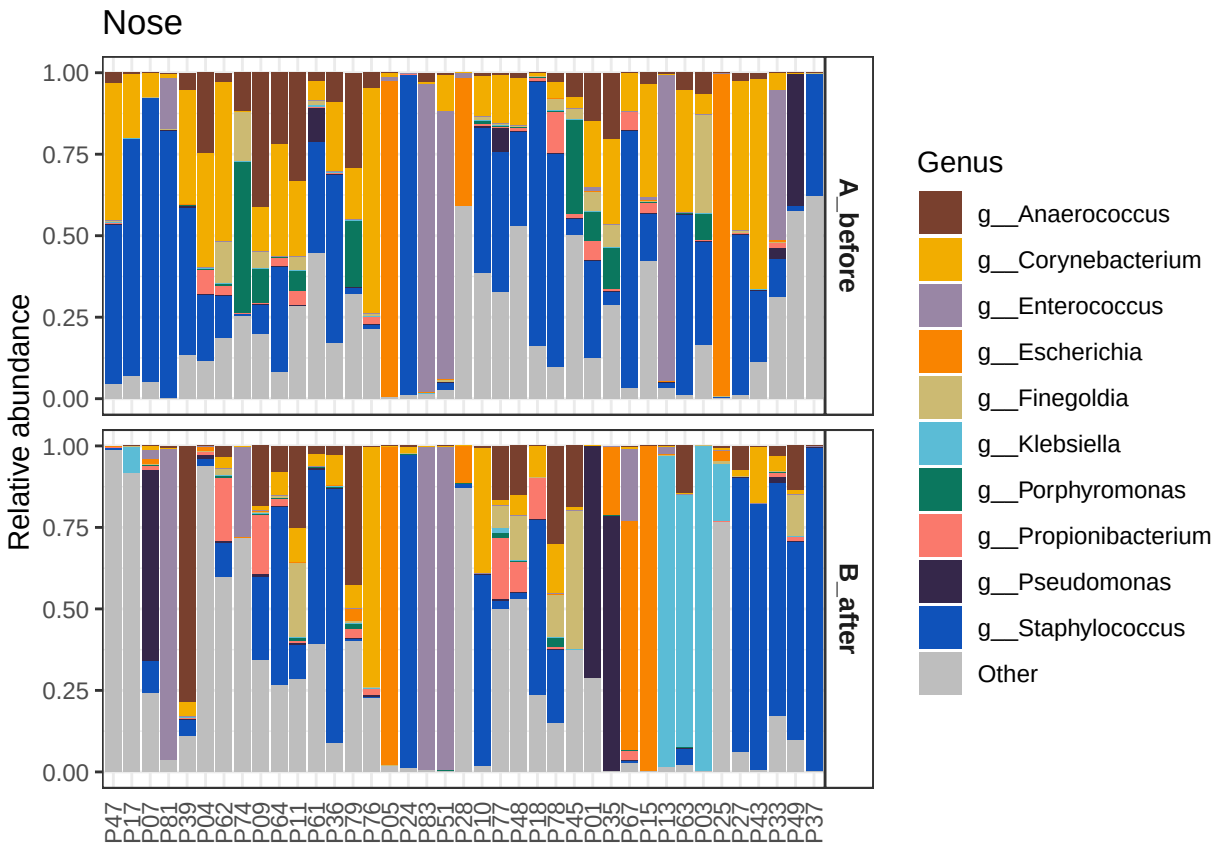
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
dev.off()
```

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

Make a version of the barplots that has the same order of patients as the heatmap

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

a <- ggplot(p_df_o_groin, aes(x = Patient_ID, y = Abundance,
    fill = Genus)) + geom_bar(stat = "identity", width = 0.9) +
    facet_grid(time_point ~ ., scales = "free") + scale_fill_manual(values = mycols) +
    theme(axis.title.x = element_blank(), axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 90, vjust = 0.5),
        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)
a
```




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

Operation_site

```
ps1_clean_operation_site <- prune_samples(sample_data(ps1_clean)$Sample_type ==
    "Operation_site", ps1_clean)
ps1_clean_operation_site <- prune_taxa(taxa_sums(ps1_clean_operation_site) !=
    0, ps1_clean_operation_site)
ps1_clean_operation_site
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 720 taxa and 74 samples ]
## sample_data() Sample Data:       [ 74 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 720 taxa by 7 taxonomic ranks ]
```

```
sample_data(ps1_clean_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"
```

Number of patients with operation site samples overall

```
length(unique(sample_data(ps1_clean_operation_site)$Patient_ID))
```

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

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

```
table(sample_data(ps1_clean_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
```

```
ps1_clean_operation_site
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 720 taxa and 74 samples ]
## sample_data() Sample Data:       [ 74 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 720 taxa by 7 taxonomic ranks ]
```

```
table(sample_data(ps1_clean_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
```

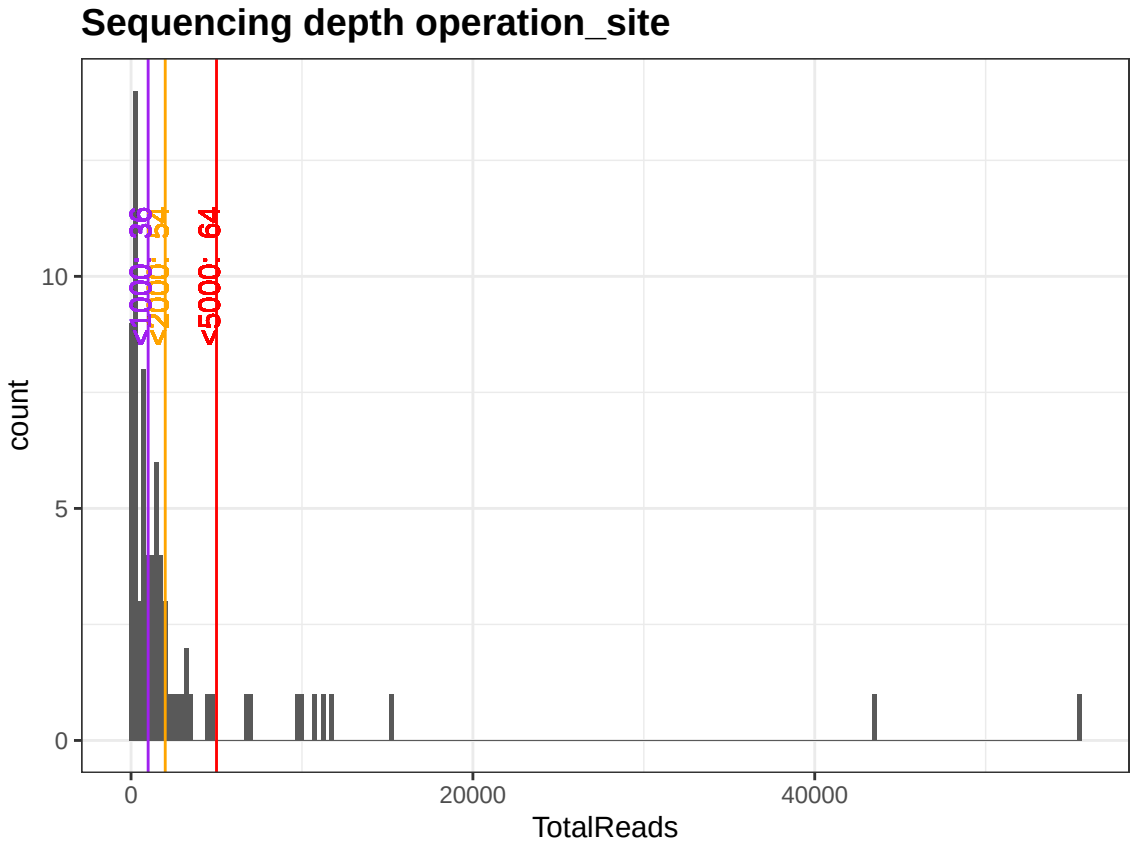
```
length(unique(sample_data(ps1_clean_operation_site)$Patient_ID))
```

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

Seq depth

```
sdt = data.table::data.table(as(sample_data(ps1_clean_operation_site),
    "data.frame"), TotalReads = sample_sums(ps1_clean_operation_site),
    keep.rownames = TRUE)
data.table::setnames(sdt, "rn", "SampleID")
pSeqDepth_OP = ggplot(sdt, aes(TotalReads)) + geom_histogram(binwidth = 250) +
    geom_vline(xintercept = 5000, color = "red") + geom_vline(xintercept = 2000,
    color = "orange") + geom_vline(xintercept = 1000, color = "purple") +
    geom_text(aes(x = 4550, label = paste("<5000: ", nrow(sdt[sdt$TotalReads <
        5000])), y = 10), colour = "red", angle = 90) +
    geom_text(aes(x = 1550, label = paste("<2000: ", nrow(sdt[sdt$TotalReads <
        2000])), y = 10), colour = "orange", angle = 90) +
    geom_text(aes(x = 550, label = paste("<1000: ", nrow(sdt[sdt$TotalReads <
        1000])), y = 10), colour = "purple", angle = 90) +
    ggtitle("Sequencing depth operation_site") + theme(plot.title = element_text(size = 14,
    face = "bold"))
```

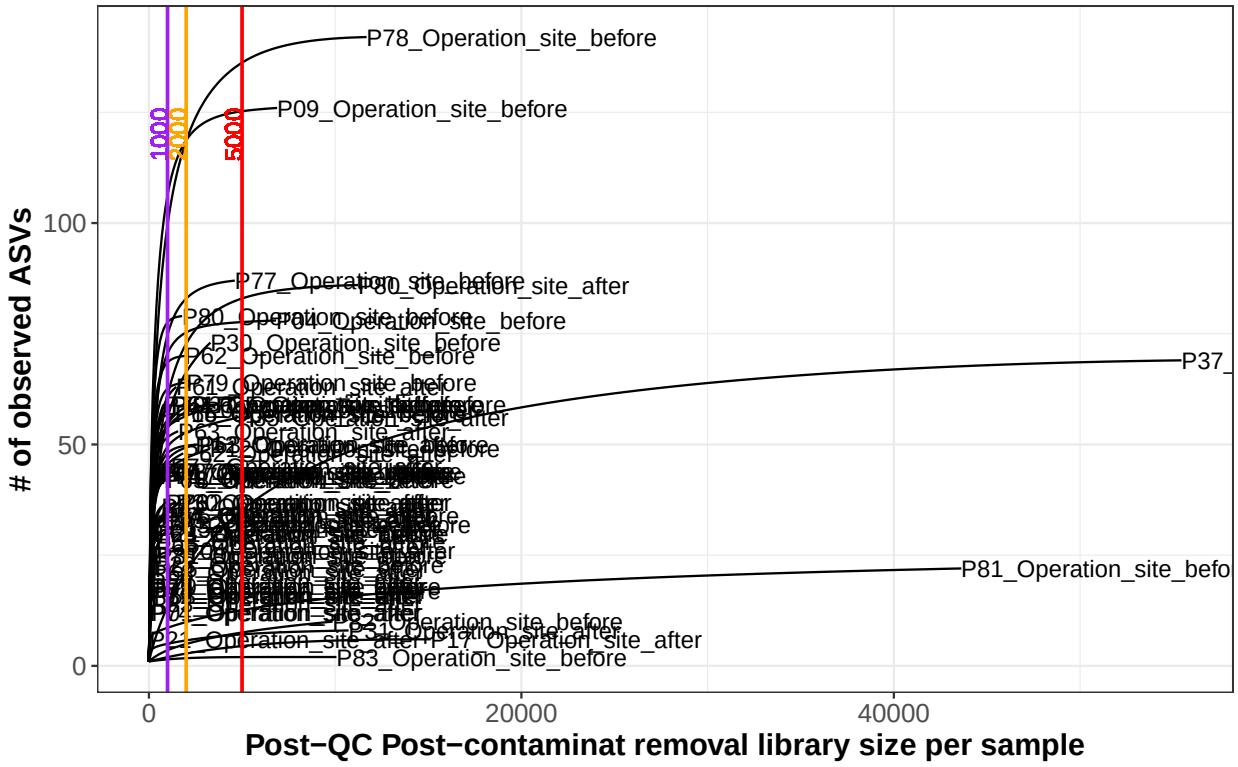
```
pSeqDepth_OP
```



Do the rarefaction curves justify that we remove samples with reads <1000 / <2000?

Rarefaction curves

```
p7 <- p7 + 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 = 120), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 120), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 120), colour = "purple",
        angle = 90, size = 4)
p7
```

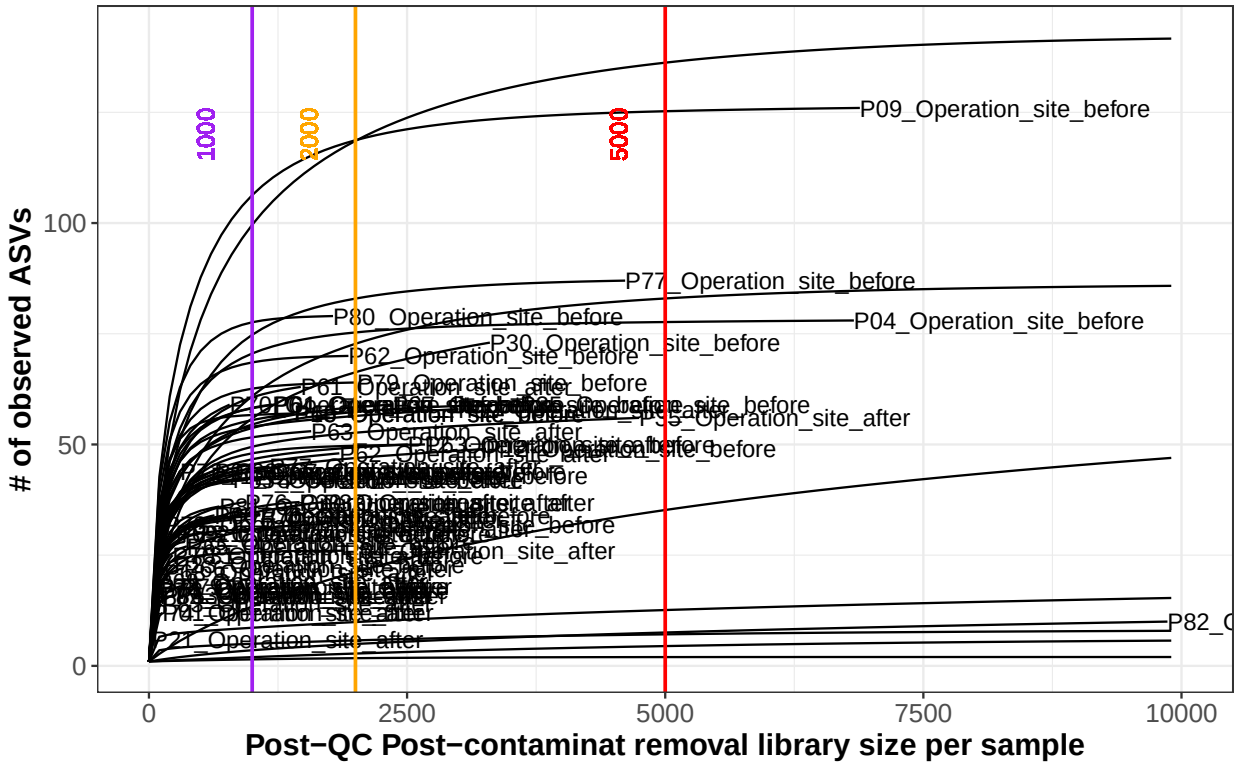


Zoom

```
p7 + xlim(0, 10000)
```

```
## Warning: Removed 7 rows containing missing values (geom_text).
```

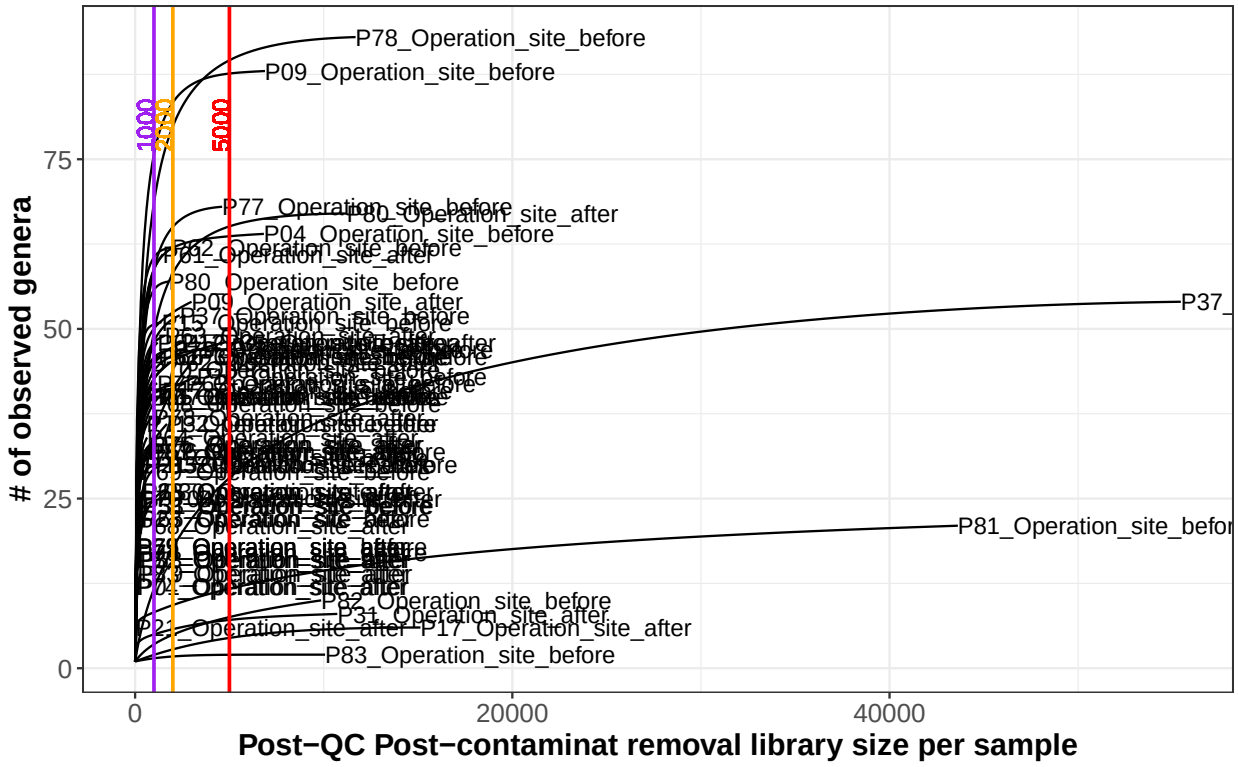
```
## Warning: Removed 890 row(s) containing missing values (geom_path).
```



How do the rarefaction curves look on genus level?

Rarefaction curves

```
p8 <- p8 + 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 genera") + 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 = 80), colour = "red", angle = 90,
    size = 4) + geom_text(aes(x = 1550, label = "2000",
    y = 80), colour = "orange", angle = 90, size = 4) +
    geom_text(aes(x = 550, label = "1000", y = 80), colour = "purple",
        angle = 90, size = 4)
p8
```

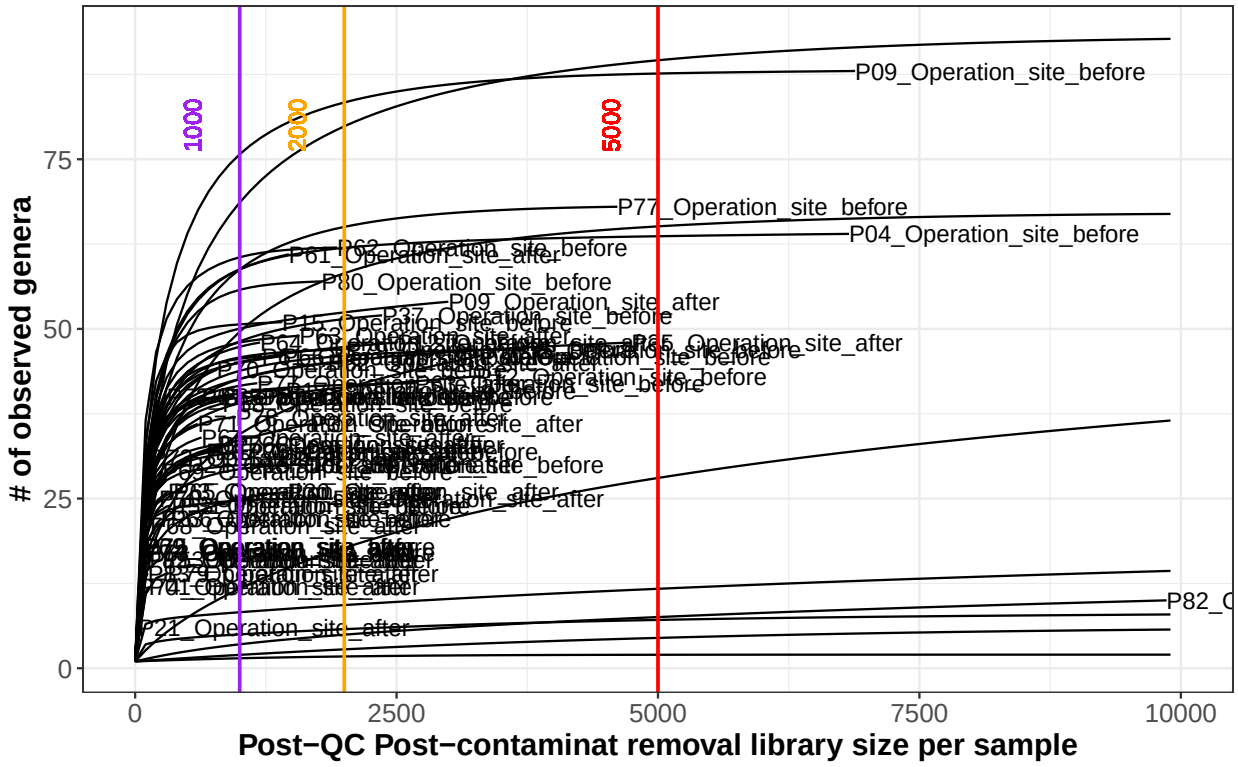


Zoom

```
p8 + xlim(0, 10000)
```

```
## Warning: Removed 7 rows containing missing values (geom_text).
```

```
## Warning: Removed 890 row(s) containing missing values (geom_path).
```



Exclude samples with <2000 reads

```
summary(sample_sums(ps1_clean_operation_site))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      31     311    1080    3429    2276   55445
```

```
ps1_clean_operation_site_tu <- prune_samples(!sample_sums(ps1_clean_operation_site) <
    2000, ps1_clean_operation_site)
ps1_clean_operation_site_tu
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 720 taxa and 20 samples ]
## sample_data() Sample Data:       [ 20 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 720 taxa by 7 taxonomic ranks ]
```

```
summary(sample_sums(ps1_clean_operation_site_tu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2016    3197    5787   10678   10850   55445
```

Now how many patients still have both time points left?

```
table(sample_data(ps1_clean_operation_site_tu)$Patient_ID)
```

```
##
## P04 P09 P12 P17 P30 P31 P35 P37 P63 P77 P78 P79 P80 P81 P82 P83
##   1   2   2   1   1   1   2   2   1   1   1   1   1   1   1   1
```

```
length(unique(sample_data(ps1_clean_operation_site_tu)$Patient_ID))
```

```
## [1] 16
```

```
ps1_clean_operation_site_tu <- prune_samples(!sample_data(ps1_clean_operation_site_tu)$Patient_ID %in%
    c("P04", "P17", "P30", "P31", "P63", "P77", "P78", "P79",
        "P80", "P81", "P82", "P83"), ps1_clean_operation_site_tu)
ps1_clean_operation_site_tu
length(unique(sample_data(ps1_clean_operation_site_tu)$Patient_ID))
```

4 patients left with 2 time points

Read counts in the remaining patients with 2 samples >2000 reads

```
summary(sample_sums(ps1_clean_operation_site_tu))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2016    3197    5787   10678   10850   55445
```

PCoA of Nose and Groin, before vs after surgery

```
ps_n_g <- merge_phyloseq(ps1_clean_nose_tu, ps1_clean_groin_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

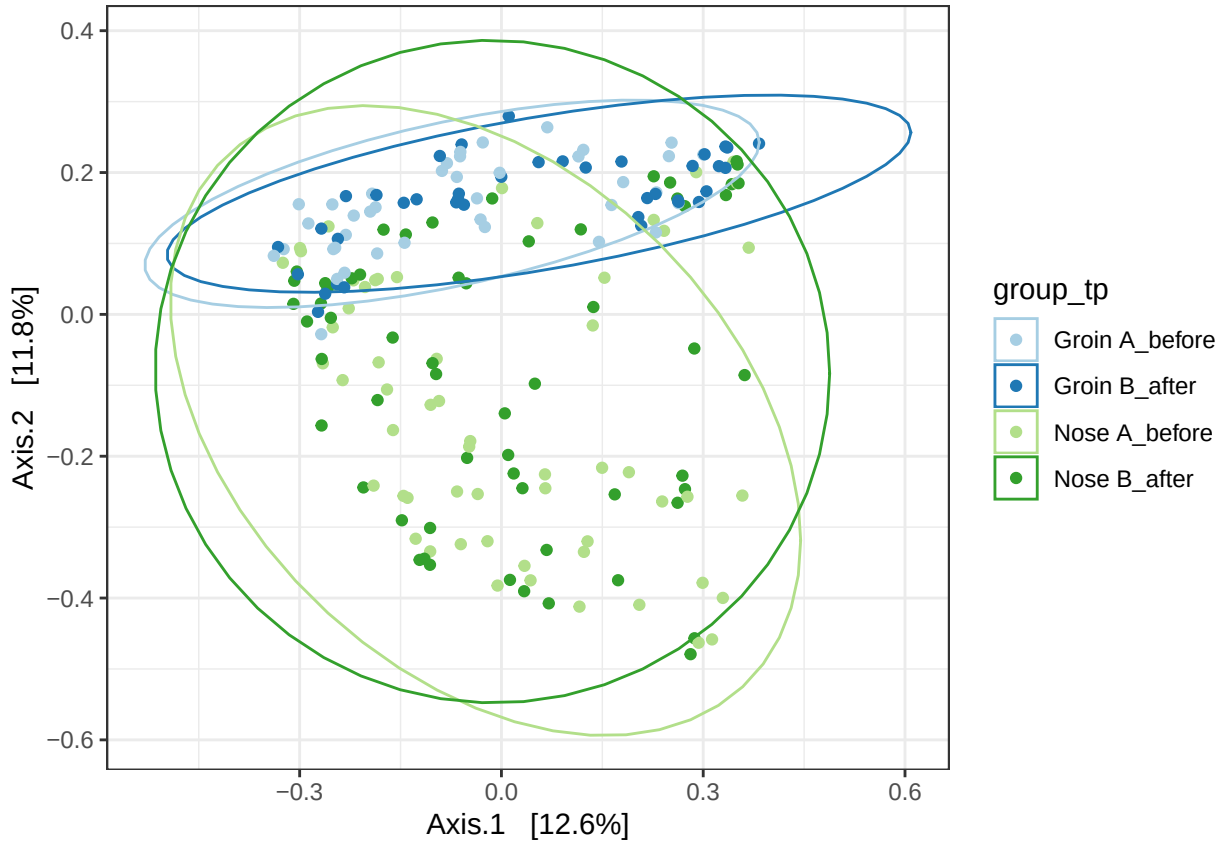
```
ps_n_g_hell <- transform_sample_counts(ps_n_g, function(x) sqrt(x/sum(x)))
```

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

PCoA

PCoA - Axis 1,2

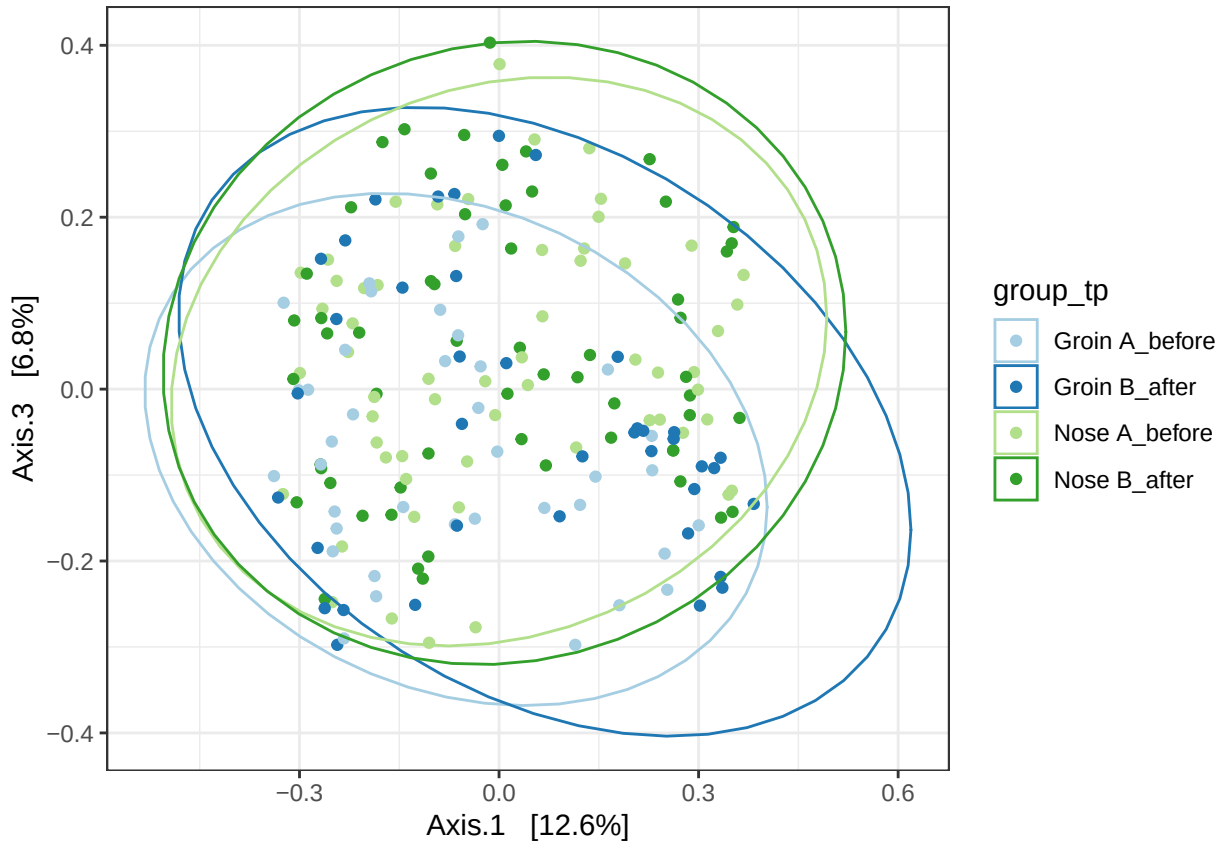
```
ord_plot <- plot_ordination(ps_n_g_hell, ps_n_g_ord, type = "samples",
    color = "group_tp", axes = 1:2)
ord_plot <- ord_plot + stat_ellipse(geom = "polygon", type = "t",
    alpha = 0, aes(fill = group_tp)) + scale_color_brewer(palette = "Paired",
    type = "div") + scale_fill_brewer(palette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




```
ggsave(filename = "plots/PCoA_Axis_1_2_16S.pdf", plot = ord_plot,
    device = cairo_pdf, width = 148, height = 105, units = "mm")
```

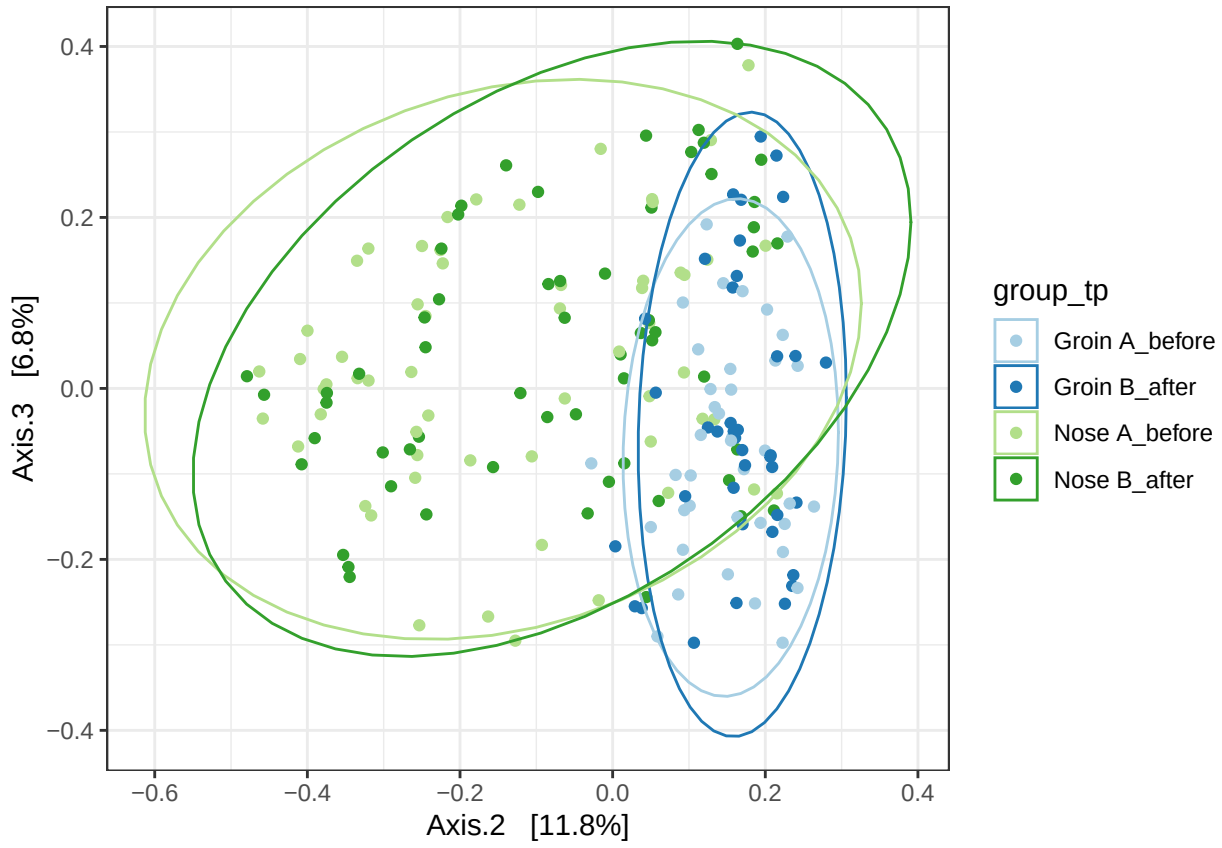
PCoA - Axis 1,3

```
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")  #, k=10, trymax=40
```

```
## Run 0 stress 0.2472374
## Run 1 stress 0.2444938
## ... New best solution
## ... Procrustes: rmse 0.04283883  max resid 0.2318548
## Run 2 stress 0.2369197
## ... New best solution
## ... Procrustes: rmse 0.0595878  max resid 0.2847011
## Run 3 stress 0.2411663
## Run 4 stress 0.2498321
## Run 5 stress 0.2428177
## Run 6 stress 0.2362356
## ... New best solution
## ... Procrustes: rmse 0.04012471  max resid 0.2071571
## Run 7 stress 0.2467225
## Run 8 stress 0.2375731
## Run 9 stress 0.2402737
## Run 10 stress 0.2452953
## Run 11 stress 0.2369924
## Run 12 stress 0.2432567
## Run 13 stress 0.2535058
```

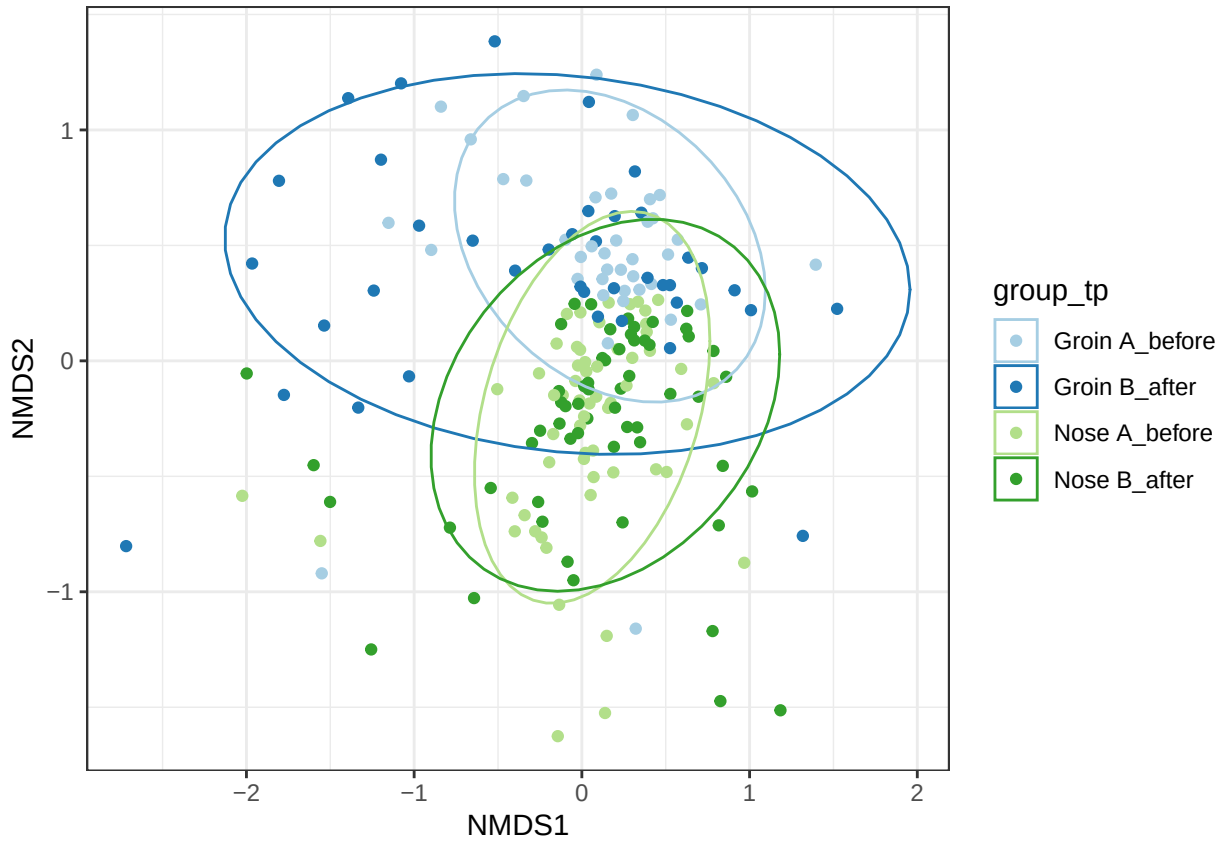
```
## Run 14 stress 0.2387541
## Run 15 stress 0.2365595
## ... Procrustes: rmse 0.02887982  max resid 0.1398078
## Run 16 stress 0.244904
## Run 17 stress 0.2435105
## Run 18 stress 0.2389652
## Run 19 stress 0.2445757
## Run 20 stress 0.2531879
## *** No convergence -- monoMDS stopping criteria:
##      3: no. of iterations >= maxit
##     17: stress ratio > sratmax
```

```
ps_n_g_ord
```

```
##
## Call:
## metaMDS(comm = veganifyOTU(physeq), distance = distance)
##
## global Multidimensional Scaling using monoMDS
##
## Data:     veganifyOTU(physeq)
## Distance: bray
##
## Dimensions: 2
## Stress:     0.2362356
## Stress type 1, weak ties
## No convergent solutions - best solution after 20 tries
## Scaling: centring, PC rotation, halfchange scaling
## Species: expanded scores based on 'veganifyOTU(physeq)'
```

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      3 0.17001 0.056671  4.4073 0.005043 **
## Residuals 192 2.46879 0.012858
## ---
## 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    3     5.638  1.8793  5.6401 0.08099  0.001 ***
## Residuals 192    63.975  0.3332         0.91901
## Total     195    69.612                 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               1    4.510 0.06479 13.5359  0.001 ***
## time_point                1    0.704 0.01011  2.1127  0.004 **
## Sample_type:time_point    1    0.424 0.00609  1.2717  0.147
## Residual                192   63.975 0.91901
## Total                   195   69.612 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 0.01975 0.019755  1.3866 0.2414
## Residuals 116 1.65262 0.014247
```

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     0.398 0.39837  1.2667 0.0108  0.201
## Residuals  116    36.481 0.31449         0.9892
## Total      117    36.880                1.0000
```

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 0.07314 0.073138  6.6942 0.01158 *
## Residuals 76 0.83034 0.010926
## ---
## 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    0.7293 0.72931   2.016 0.02584  0.006 **
## Residuals  76   27.4932 0.36175         0.97416
## Total      77   28.2225                 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)