

Quality controls of sequencing data

Stijn Schreven

23 April 2022

Contents

Load packages	2
Input files	2
1. Negative controls	2
1.1. Prepare data	2
1.2. Table of genus composition	3
2. Positive controls: mock communities	4
2.1. Create correlation matrix	4
2.2. Correlation between sequenced and theoretical mocks	4
2.3. Correlation between mock samples	5
2.4. Conclusion	5
3. DNA isolation replicates	6
4. PCR replicates	7
5. Biofilms and underlying substrate	7
5.1. Prepare data	8
5.2. Plot presets	8
5.3. Heatmap at genus level	8
5.4. Spearman correlation at genus level	10
6. Substrate depth	11
6.1. Prepare data	11
6.2. Heatmap at genus level	11
6.3. Spearman correlations	12

Load packages

```
library(phyloseq)
library(microbiome)
library(microbiomeutilities)
library(plyr)
library(knitr)
library(ggplot2)
library(viridis)
```

Input files

```
# negative controls
pscontr <- readRDS("./phyobjects/ps1.contr.rds")

# positive controls (mocks)
psmock <- readRDS("./phyobjects/ps1.mock.rds")
# theoretical mock community composition
mockT <- read.delim("./input_data/Mocks_composition.txt", sep="\t")

# technical replicates of DNA isolation
psbiol <- readRDS("./phyobjects/ps1.biol.rds")

# technical replicates of PCR
pstech <- readRDS("./phyobjects/ps1.tech.rds")

# biofilms
psfilm <- readRDS("./phyobjects/ps1.film.rds")

# substrate depth
psdepth <- readRDS("./phyobjects/ps1.sdepth.rds")
```

1. Negative controls

We included no-template controls for the PCR. They provide clues on DNA contamination after DNA isolation. We looked at genus-level community composition.

1.1. Prepare data

```
# add OTU column and remove tree
taxcontr <- as.data.frame(pscontr@tax_table)
taxcontr$OTU <- rownames(taxcontr)
tax_table(pscontr) <- tax_table(as.matrix(taxcontr))
pscontr@phy_tree <- NULL

# format to besthit
```

```

pscontr <- format_to_besthit(pscontr)

# aggregate to genus
pscontr.g <- microbiome::aggregate_taxa(pscontr, "Genus")
pscontr.g <- microbiome::transform(pscontr.g, "compositional")
pscontr.g

## phyloseq-class experiment-level object
## otu_table() OTU Table: [ 28 taxa and 5 samples ]
## sample_data() Sample Data: [ 5 samples by 14 sample variables ]
## tax_table() Taxonomy Table: [ 28 taxa by 7 taxonomic ranks ]

```

1.2. Table of genus composition

Supplementary Table S1 in manuscript Chapter 3 in PhD thesis and submission to *Applied and Environmental Microbiology*.

```

# heatmap
p.hm.neg <- plot_heatmap(pscontr.g, method = "MDS", distance = "bray",
                        sample.order = "Description") + labs(x = "Sample", y = "Genus")

# create table from heatmap
hm.neg.df <- reshape2::dcast(p.hm.neg$data[,1:3], OTU ~ Sample, fun.aggregate = sum,
                             margins= "Sample")
colnames(hm.neg.df) <- c("Genus", "NTC 1", "NTC 2", "NTC 3", "NTC 4", "NTC 5", "all")
hm.neg.df <- hm.neg.df[order(hm.neg.df$all, decreasing = T), ]
hm.neg.df[,c(2:7)] <- round(hm.neg.df[,c(2:7)], digits = 3)
hm.neg.df <- hm.neg.df[, -7]
write.csv(hm.neg.df, "./tables/PCR_blanks_composition_gen.csv")

kable(hm.neg.df)

```

	Genus	NTC 1	NTC 2	NTC 3	NTC 4	NTC 5
27	Nesterenkonia	0.276	0.326	0.430	0.420	0.357
24	Caldakalibacillus	0.126	0.113	0.348	0.310	0.329
21	Ralstonia	0.191	0.130	0.055	0.081	0.069
25	Halomonas	0.132	0.133	0.039	0.056	0.079
23	Bacillus	0.075	0.061	0.052	0.045	0.060
22	f__Bacillaceae	0.073	0.058	0.037	0.031	0.032
3	Shewanella	0.035	0.067	0.021	0.031	0.033
2	f__Halomonadaceae	0.025	0.040	0.012	0.020	0.021
13	f__Xanthomonadaceae	0.028	0.029	0.000	0.000	0.004
1	Cupriavidus	0.012	0.013	0.004	0.002	0.009
20	Achromobacter	0.005	0.004	0.000	0.002	0.002
12	Dietzia	0.005	0.005	0.000	0.000	0.000
9	Geobacillus	0.002	0.007	0.000	0.000	0.000
14	Georgenia	0.003	0.003	0.000	0.000	0.000
26	k__NA	0.000	0.000	0.002	0.000	0.003
4	Delftia	0.000	0.002	0.000	0.001	0.001
10	Rhodococcus	0.002	0.002	0.000	0.000	0.000
11	Sphingomonas	0.002	0.002	0.000	0.000	0.000

	Genus	NTC 1	NTC 2	NTC 3	NTC 4	NTC 5
16	Aerococcus	0.003	0.000	0.000	0.000	0.000
15	Mycobacterium	0.003	0.000	0.000	0.000	0.000
28	Lactobacillus	0.000	0.000	0.000	0.000	0.002
5	Staphylococcus	0.000	0.001	0.000	0.000	0.000
8	Dermacoccus	0.000	0.001	0.000	0.000	0.000
19	Rubrobacter	0.001	0.000	0.000	0.000	0.000
6	Tetragenococcus	0.000	0.001	0.000	0.000	0.000
17	f__Comamonadaceae	0.001	0.000	0.000	0.000	0.000
18	Glutamicibacter	0.001	0.000	0.000	0.000	0.000
7	Brevundimonas	0.000	0.001	0.000	0.000	0.000

2. Positive controls: mock communities

The synthetic mock communities provide a quality check of sequencing effort, *i.e.* whether the sequenced composition represents the theoretical composition. In each sequencing library, we included PCR products of two mock communities (community 3 and 4) of different composition. We compared community composition on genus level with relative abundances.

2.1. Create correlation matrix

```
# extract genus composition of mock samples
psmock.g <- microbiome::aggregate_taxa(psmock, "Genus")
psmock.g <- microbiome::transform(psmock.g, "compositional")
psmock.g <- format_to_besthit(psmock.g)

# convert to dataframe, including genus names
psmock.df <- data.frame(abundances(psmock.g))
psmock.df$Genus <- tax_table(psmock.g)[, "Genus"]

# merge dataframes by column Genus
mocks.df <- merge(psmock.df, mockT, by = "Genus", all.x = T)
rownames(mocks.df) <- mocks.df$Genus
mocks.df <- subset(mocks.df, select = -Genus)

# Spearman correlations
mcor <- cor(mocks.df, use = "pairwise.complete.obs", method = "spearman")
```

2.2. Correlation between sequenced and theoretical mocks

```
# assemble results in data frame
mock3 <- c("mock3.A", "mock3.B", "mock3.C", "mock3.D", "mock3.E")
mock4 <- c("mock4.A", "mock4.B", "mock4.C", "mock4.D", "mock4.E")

m3 <- data.frame("Sample" = mock3, "r"=NA)
m4 <- data.frame("Sample" = mock4, "r"=NA)
```

```

# extract data from correlation matrix
for(i in 1:nrow(m3)){
  a = match("MC3",labels(mcor)[[1]])
  b = match(m3$Sample[i],labels(mcor)[[2]])
  m3$r[i] = mcor[a,b]
}
for(i in 1:nrow(m4)){
  a = match("MC4",labels(mcor)[[1]])
  b = match(m4$Sample[i],labels(mcor)[[2]])
  m4$r[i] = mcor[a,b]
}

# summarise results (mean and SD)
m34 <- rbind(m3,m4)
m34$mock <- c(3,3,3,3,3,4,4,4,4,4)
m34$mock <- as.factor(m34$mock)
QC1 <- ddply(m34, .(mock), summarise, mean = mean(r), SD = sd(r))
kable(QC1)

```

mock	mean	SD
3	0.8333041	0.0350161
4	0.7150864	0.0413801

2.3. Correlation between mock samples

```

mcor.m3 <- mcor[c(1,3,5,7,9),c(1,3,5,7,9)]
mcor.m4 <- mcor[c(2,4,6,8,10),c(2,4,6,8,10)]

rho.m3 <- data.frame("mock"=3,"r"=c(lower.tri(mcor.m3, diag=F))*c(mcor.m3))
rho.m3 <- subset(rho.m3, r>0)
rho.m4 <- data.frame("mock"=4,"r"=c(lower.tri(mcor.m4, diag=F))*c(mcor.m4))
rho.m4 <- subset(rho.m4, r>0)

rho.m34 <- rbind(rho.m3, rho.m4)
rho.m34$mock <- as.factor(rho.m34$mock)

m34.sum <- ddply(rho.m34, .(mock), summarise, mean = mean(r), SD = sd(r))
kable(m34.sum)

```

mock	mean	SD
3	0.9529445	0.0218384
4	0.9764101	0.0162427

2.4. Conclusion

PCR replicates within either positive control (mock community 3 or 4) were highly correlated: mock 3 mean(r) = 0.95, mock 4 mean(r) = 0.98. Samples of both mock communities correlated less with their

corresponding theoretical compositions: mock 3 $\text{mean}(r) = 0.83$, mock 4 $\text{mean}(r) = 0.72$.

3. DNA isolation replicates

We collected duplicate substrate samples at different timepoints in a subset of the containers, in order to assess the representability of a sample of the entire substrate.

```
# transform data: relative abundance and genus level
psbiol.g <- microbiome::aggregate_taxa(psbiol, "Genus")
psbiol.g <- microbiome::transform(psbiol.g, "compositional")

# prepare dataframe
biolsam <- subset(meta(psbiol),
                  select = c(Description, ContainerID, Diet, Density,
                             Duplicate, Timepoint))
biolsam$samTime <- biolsam$ContainerID:biolsam$Timepoint
dup1 <- subset(biolsam, Duplicate == "no")
dup2 <- subset(biolsam, Duplicate == "biological")
dfdup <- merge(dup1, dup2, by="samTime")
dfdup <- subset(dfdup, select = -c(6,9:13))
colnames(dfdup) <- c("samTime", "dup1", "ContainerID", "Diet", "Density",
                    "Timepoint", "dup2")

dfdup$r = NA

# correlation matrix at genus level
psbiol.df.g <- data.frame(otu_table(psbiol.g))
colnames(psbiol.df.g) <- sub(pattern = "X", replacement = "", x = colnames(psbiol.df.g))
dcor <- cor(as.matrix(psbiol.df.g), method = "spearman")

# fill correlations in dfdup dataframe
for(i in 1:nrow(dfdup)) {
  a = match(dfdup$dup1[i], labels(dcor)[[1]])
  b = match(dfdup$dup2[i], labels(dcor)[[2]])
  dfdup$r[i] = dcor[a, b]
}

kable(dfdup[,c(3:5,8)])
```

ContainerID	Diet	Density	r
1	CF	0	0.4651143
1	CF	0	0.8463650
13	CF	200	0.9363105
13	CF	200	0.3843646
17	CS	0	0.7859489
17	CS	0	0.5971738
29	CS	200	0.7541647
29	CS	200	0.7279294
33	CM	0	0.9047260
33	CM	0	0.7732997
45	CM	200	0.8678466
45	CM	200	0.8802335

Conclusion

Community composition of bacterial genera is quite consistent across biological duplicates. In containers 1 (day 15), 13 (day 5) and 17 (day 5) correlations were lower. But in the other samples, replicates correlated for $r = 0.73-0.90$. This shows that there can be considerable variation in DNA isolation or heterogeneity within the substrate.

4. PCR replicates

We duplicated one substrate and two larval samples (PCR products) in separate sequence libraries, in order to assess the reproducibility of sequencing runs (in addition to the mock communities).

```
# transform data: relative abundance and OTU or genus level
pstech.g <- microbiome::aggregate_taxa(pstech, "Genus")
pstech.g <- microbiome::transform(pstech.g, "compositional")

# prepare dataframe
tdup1 <- c("6.A.", "7.C.", "13.G.")
tdup2 <- c("6.A.T", "7.C.T", "13.G.T")
dftdup <- data.frame(tdup1, tdup2, r = NA)

# correlation matrix at genus level
pstech.df.g <- data.frame(otu_table(pstech.g))
colnames(pstech.df.g) <- sub(pattern = "X", replacement = "", x = colnames(pstech.df.g))
tcor <- cor(as.matrix(pstech.df.g), method = "spearman")

# fill correlations in dftdup dataframe
for(i in 1:nrow(dftdup)) {
  a = match(dftdup$tdup1[i], labels(tcor)[[1]])
  b = match(dftdup$tdup2[i], labels(tcor)[[2]])
  dftdup$r[i] = tcor[a, b]
}

kable(dftdup)
```

tdup1	tdup2	r
6.A.	6.A.T	0.9115243
7.C.	7.C.T	0.8759624
13.G.	13.G.T	0.9843456

Conclusion

The technical duplicates showed consistency in both read numbers and composition. Correlations were $r = 0.88-0.98$ at genus level. This showed that PCR and sequencing was reproducible with our own sample material (substrate and larvae).

5. Biofilms and underlying substrate

We collected biofilms samples from the top layer of the substrate (day 5) from all diets, as we removed them when collecting the substrate sample (because the biofilm may shift bias of substrate composition to only

the community in the top layer). In order to assess what we have removed and how different that is from the underlying substrate samples, we compared biofilm and substrate below.

5.1. Prepare data

```
# genus level, all genera
psfilm.r <- microbiome::aggregate_taxa(psfilm, "Genus")
psfilm.r <- microbiome::transform(psfilm, "compositional")

# add OTU column and remove tree
biofilm <- psfilm
taxfilm <- as.data.frame(biofilm@tax_table)
taxfilm$OTU <- rownames(taxfilm)
tax_table(biofilm) <- tax_table(as.matrix(taxfilm))
biofilm@phy_tree <- NULL

# add best hit, merge at genus level, without tree
biofilm <- format_to_besthit(biofilm)
biofilm.g25 <- microbiome::aggregate_taxa(biofilm, "Genus", top = 25)
biofilm.g25 <- microbiome::transform(biofilm.g25, "compositional")
```

5.2. Plot presets

```
theme_hm <- theme_classic() +
  theme(text = element_text(size = 15),
        axis.text.y = element_text(face = 'italic'),
        legend.key = element_blank(),
        panel.spacing.x = unit(0, "lines"),
        strip.background = element_rect(colour = "black", fill = NA))

labs_biof <- as_labeller(c(
  "CF" = "chicken\nfeed", "CS" = "camelina", "CM" = "chicken\nmanure",
  "5" = "ID_5", "17" = "ID_17", "21" = "ID_21", "24" = "ID_24", "25" = "ID_25",
  "41" = "ID_41"
))
```

5.3. Heatmap at genus level

Supplementary Figure S4 in manuscript Chapter 3 in PhD thesis and Supplementary Figure S6 in manuscript submitted to *Applied and Environmental Microbiology*.

```
# extract plot data
biof.g.df <- plot_heatmap(biofilm.g25, method = "NMDS", distance = "bray")$data
biof.g.df <- subset(biof.g.df, OTU != "Other")
biof.g.df$Type <- revalue(biof.g.df$Type, c("biofilm"="B", "substrate"="S"))

# 24.H. Type is substrate should be biofilm (probably done because not really a biofilm
## but more the top layer of substrate which was darker; but for graph need to rename
## here):
```



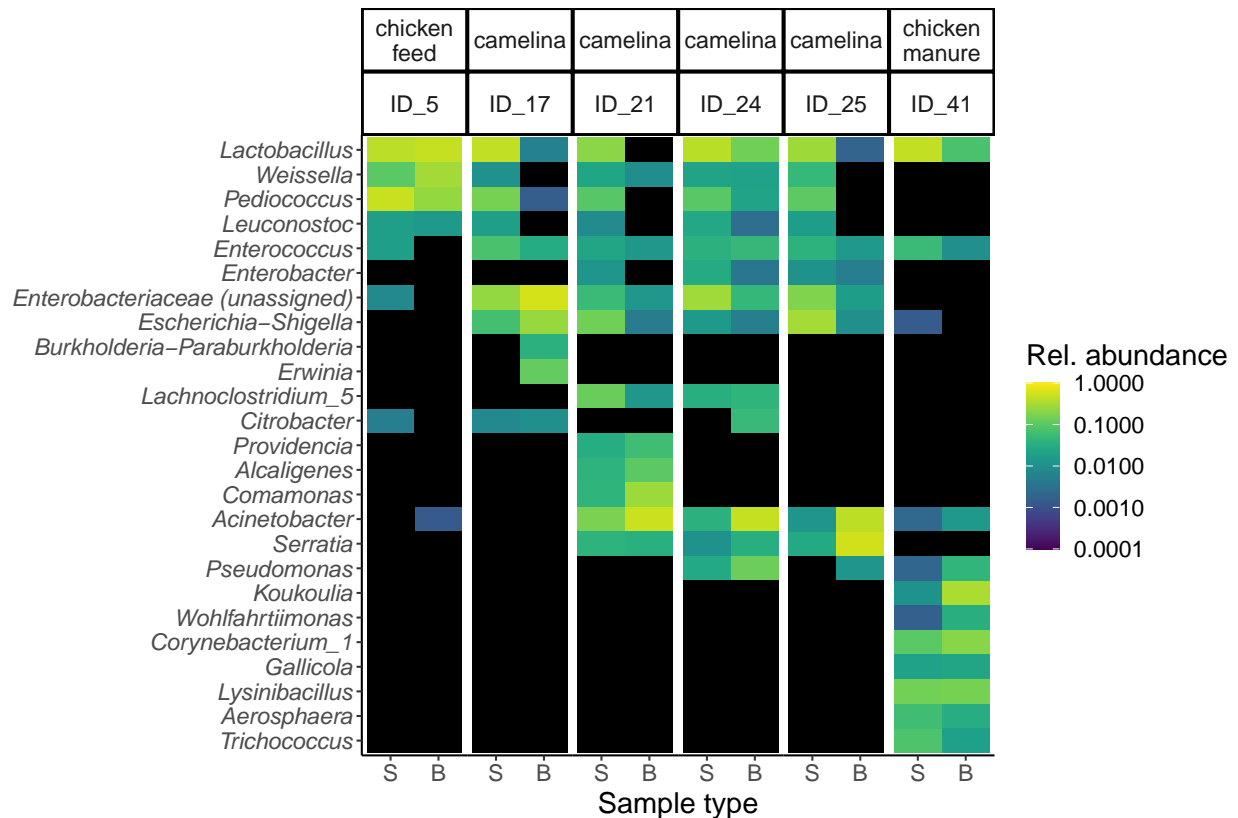
```

biof.g.df$Type2 <- biof.g.df$Type
biof.g.df[biof.g.df$Sample == "24.H.", "Type2"] <- "B"

# rename f__Enterobacteriaceae to Enterobacteriaceae (unassigned)
biof.g.df$OTU <- revalue(biof.g.df$OTU,
                          c("f__Enterobacteriaceae" = "Enterobacteriaceae (unassigned)"))

# heatmap
hm.biof.g <- ggplot(biof.g.df, aes(x = Type2, y = OTU)) +
  geom_tile(aes(fill = Abundance)) +
  scale_fill_viridis("Rel. abundance", option = "D",
                    na.value = "black", trans = "log10",
                    limits = c(.0001,1),
                    labels = function(n){format(n, scientific = F)}) +
  labs(y = NULL, x = "Sample type") +
  facet_grid(~Diet+ContainerID, scales = "free_x", labeller = labs_biof) +
  theme_hm
hm.biof.g

```



```

# export heatmap
ggsave(plot = hm.biof.g, "./figures/Biofilms_heatmap_gen.png", h = 6, w = 9)
ggsave(plot = hm.biof.g, "./figures/Biofilms_heatmap_gen.pdf", h = 150, w = 225, u = "mm")

```

5.4. Spearman correlation at genus level

```
# convert phyloseq to dataframe
psfilm.df <- data.frame(otu_table(psfilm.r))
colnames(psfilm.df) <- sub(pattern = "X", replacement = "", x = colnames(psfilm.df))

# prepare dataframe
fsubstr <- c("5.B.", "17.B.", "21.B.", "24.B.", "25.B.", "41.B.")
fbiofilm <- c("5.H.", "17.H.", "21.H.", "24.H.", "25.H.", "41.H.")
film.rho <- data.frame(fsubstr, fbiofilm, r=NA)

# Spearman correlations
fcor <- cor(as.matrix(psfilm.df), method = "spearman")

# fill correlations in film.rho data frame
for(i in 1:nrow(film.rho)){
  a = match(film.rho$fsubstr[i], labels(fcor)[[1]])
  b = match(film.rho$fbiofilm[i], labels(fcor)[[2]])
  film.rho$r[i] = fcor[a,b]
}

kable(film.rho)
```

fsubstr	fbiofilm	r
5.B.	5.H.	0.4939112
17.B.	17.H.	0.2543273
21.B.	21.H.	0.5589531
24.B.	24.H.	0.5379370
25.B.	25.H.	0.1564925
41.B.	41.H.	0.4652731

Conclusion

In the CS samples, large differences occurred between biofilm and substrate bacterial composition ($r = 0.16$ - 0.56). These mainly illustrated dominance of either aerobic (biofilm) or anaerobic/fermenting bacteria (substrate).

- The fungal biofilm in CF differed from the substrate ($r = 0.49$), with slight enrichment in *Weissella* and *Lactobacillus*, while being reduced in *Pediococcus*.
- The lightbrown slime biofilm in CM showed reduction in *Lactobacillus* and enrichment in Xanthomonadaceae and *Koukoulia*, compared to substrate ($r = 0.47$).
- The biofilms of CS diet had different composition depending on type of biofilm. In general all had much lower *Lactobacillus* abundance compared to substrate.
 - a brown crust (17) was dominated by Enterobacteriaceae (unmatched genus), and reduced in Lactobacillaceae (*Lactobacillus* and *Pediococcus*).
 - the other biofilms (21: yellow biofilm, 24: darker top substrate, 25: red slime) were enriched in *Acinetobacter*. Sample 25 (red slime) was also much enriched in *Serratia*, which explains the red colour of the biofilm as these bacteria are known to produce red pigment. Sample 21 was additionally much enriched in *Comamonas*.

6. Substrate depth

For CS container 28 and CM container 37, we took two additional samples of the substrate on day 15: one from the top half and one from the bottom half. The reason was that there were striking differences in appearance between the two layers: in CS the top was dark gray whereas the bottom was yellowbrown; in CM the substrate was more liquid in the top and more dense in the bottom. We wanted to assess if these differences related to differences in bacterial community composition.

6.1. Prepare data

```
# genus level and relative abundance, with tree
psdepth.r <- microbiome::transform(psdepth, "compositional")
psdepth.r <- microbiome::aggregate_taxa(psdepth.r, "Genus")

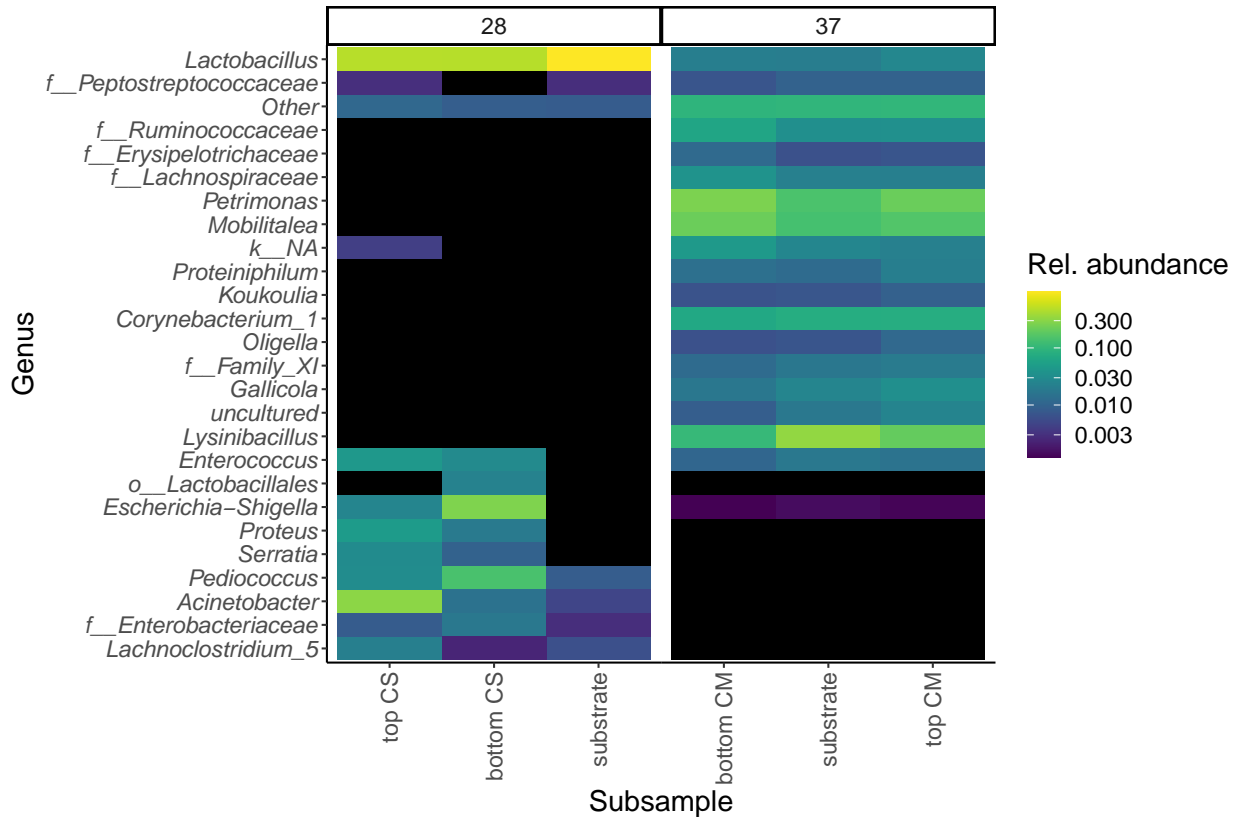
# add OTU column and remove tree
psdepth.com <- psdepth
taxdepth <- as.data.frame(psdepth.com@tax_table)
taxdepth$OTU <- rownames(taxdepth)
tax_table(psdepth.com) <- tax_table(as.matrix(taxdepth))
psdepth.com@phy_tree <- NULL

# add best hit, merge at genus level, without tree
psdepth.com <- format_to_besthit(psdepth.com)
psdepth.g25 <- microbiome::aggregate_taxa(psdepth.com, "Genus", top = 25)
psdepth.g25 <- microbiome::transform(psdepth.g25, "compositional")
```

6.2. Heatmap at genus level

```
# extract data
hm.depth.df <- plot_heatmap(psdepth.g25, method = "NMDS", distance = "bray")$data

# plot heatmap customized
p.hm.depth <- ggplot(hm.depth.df, aes(x=Details, y=OTU)) +
  geom_tile(aes(fill=Abundance)) +
  scale_fill_viridis("Rel. abundance", option = "D", na.value = "black",
                    trans = "log10") +
  facet_grid(~ContainerID, scales = "free_x") +
  scale_x_discrete(labels = c("na" = "substrate",
                             "dark_top_substrate" = "top CS",
                             "light_bottom_substrate" = "bottom CS",
                             "bottom_substrate_dense" = "bottom CM",
                             "top_substrate_liquid" = "top CM")) +
  labs(x = "Subsample", y = "Genus") + theme_hm +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
p.hm.depth
```



6.3. Spearman correlations

```
# convert phyloseq to dataframe
psdepth.df <- data.frame(otu_table(psdepth.r))
colnames(psdepth.df) <- sub(pattern = "X", replacement = "", x = colnames(psdepth.df))

# prepare dataframe
dsubstr <- c("28.F.", "28.F.", "37.F.", "37.F.")
dhalf <- c("28.H.", "29.H.", "37.H.", "38.H.")
depth.rho <- data.frame(dsubstr, dhalf, r=NA)

# Spearman correlation matrix
dcor <- cor(as.matrix(psdepth.df), method = "spearman")

# fill correlations in depth.rho data frame
for(i in 1:nrow(depth.rho)){
  a = match(depth.rho$dsubstr[i], labels(dcor)[[1]])
  b = match(depth.rho$dhalf[i], labels(dcor)[[2]])
  depth.rho$r[i] = dcor[a,b]
}

kable(depth.rho)
```

dsubstr	dhalf	r
28.F.	28.H.	0.6537991
28.F.	29.H.	0.4810911
37.F.	37.H.	0.9498851
37.F.	38.H.	0.9050324

Conclusion

The heatmaps correlations showed that in container 37, the total sample and both top and bottom sample were rather similar in composition ($r = 0.91 - 0.95$, correlation between total and bottom or top). For container 28 however, there were differences between top and bottom sample. Although all three samples were dominated by *Lactobacillus*, the top and bottom differ in the abundance of *Acinetobacter* (more in top) and *Escherichia-Shigella* and *Pediococcus* (more in bottom), and all three genera were lower in the total sample ($r = 0.48 - 0.65$, correlation between total and bottom or top).