

Quality checks of Illumina sequencing

Stijn Schreven

4 March 2021

Contents

Load packages	1
Input files	2
1. Negative controls	2
1. Prepare data	2
2. Heatmap	2
2. Positive controls: mock communities	3
2.1. Prepare data	4
2.2. Correlation between theoretical mocks and mock samples	4
2.3. Correlation between mock samples	5
3. Technical replicates	5
3.1. Consistency of DNA extraction replicates	5
3.2. Consistency of PCR replicates	7
4. Export table	8

Load packages

```
library(phyloseq)
library(microbiome)
library(microbiomeutilities)
library(reshape2)
library(plyr)
library(dplyr)
library(sciplot)
library(ggplot2)
library(viridis)
library(knitr)
```

Input files

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

# mock samples
psmock <- readRDS("./phyobjects/ps1.mock.rds")
# file with theoretical mock community composition
mockT <- read.delim("./input_data/Schreven_Ch4_mocks_composition.txt")

# technical replicates of DNA extraction and PCR
psbiol <- readRDS("./phyobjects/ps1.biol.rds") # DNA extraction replicates
pstech <- readRDS("./phyobjects/ps1.tech.rds") # PCR replicates
```

1. Negative controls

The negative controls are controls for the DNA isolation and PCR. We looked at composition at genus level.

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, use top 30 for heatmap
pscontr.g30 <- microbiome::aggregate_taxa(pscontr, "Genus", top = 30)
pscontr.g30 <- microbiome::transform(pscontr.g30, "compositional")

# plot presets
theme_neg <- theme_bw() +
  theme(axis.text.y = element_text(colour = 'black', face = 'italic'),
        legend.key = element_blank(),
        text = element_text(size=10),
        strip.background = element_blank(),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
labs_neg <- as_labeller(c("25" = "25 cycles", "30" = "30 cycles"))
```

2. Heatmap

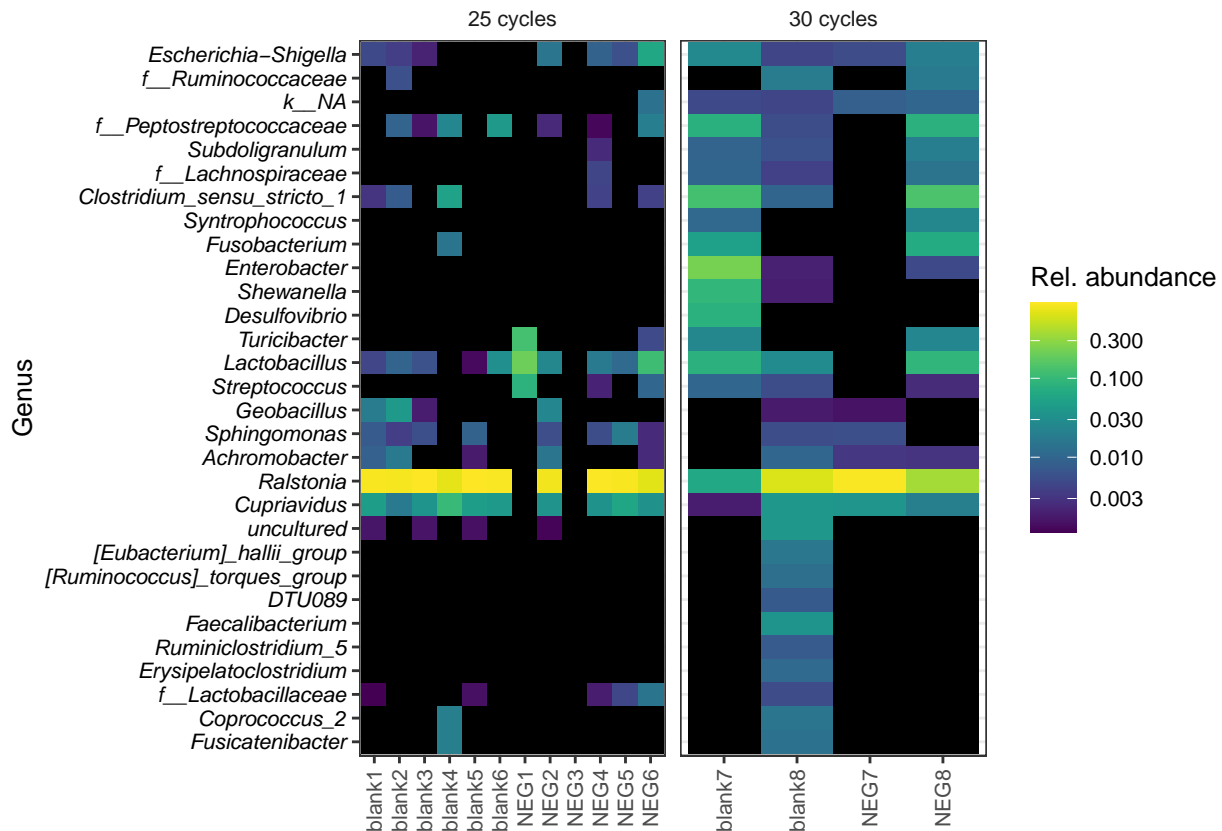
```
# extract plot data
hm.neg.df <- plot_heatmap(pscontr.g30, method = "MDS", distance = "bray",
                          sample.order = "Description")$data
```

```

hm.neg.df1 <- subset(hm.neg.df, OTU != "Other")
hm.neg.df1$Sample <- revalue(hm.neg.df1$Sample, c(
  "NEG.7" = "NEG7", "NEG.8" = "NEG8"))
hm.neg.df1 <- with(hm.neg.df1, hm.neg.df1[order(hm.neg.df1$Sample),])

# plot heatmap
p.hm.neg <- ggplot(hm.neg.df1, aes(x = Sample, y = OTU)) +
  geom_tile(aes(fill = Abundance)) +
  scale_fill_viridis("Rel. abundance", option = "D", na.value = "black",
    trans="log10") +
  labs(x=NULL,y="Genus") +
  facet_grid(~nCycles, scales = "free_x", labeller = labs_neg) +
  theme_neg
p.hm.neg

```



Note: NEG1-8 are the negative controls of DNA extraction, blank1-8 are the negative controls of PCR.

2. Positive controls: mock communities

The mock communities provide a quality check of sequencing effort (whether the composition represents a true composition). In each sequencing library, we included 2 mock communities (community 3 and 4) to compare to the given community composition.

2.1. Prepare data

```
# add OTU column, remove tree
taxmock <- as.data.frame(psmock@tax_table)
taxmock$OTU <- rownames(taxmock)
tax_table(psmock) <- tax_table(as.matrix(taxmock))
psmock.bh <- format_to_besthit(psmock)

# transform to relative abundance
psmock.g <- aggregate_taxa(psmock.bh, "Genus")
psmock.g.r <- microbiome::transform(psmock.g, "compositional")
```

2.2. Correlation between theoretical mocks and mock samples

Create correlation matrix:

```
# convert to dataframe, including genus names
psmock.df <- data.frame(otu_table(psmock.g.r))
psmock.df$Genus <- tax_table(psmock.g.r)[, "Genus"]
# merge dataframes by column Genus
mocks.df <- merge(psmock.df, mockT, by = "Genus", all.x = T)
rownames(mocks.df) <- mocks.df$Genus
mocks.df <- subset(mocks.df, select = -Genus)

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

Summarize results:

```
# assemble results in data frame
mock3 <- c("Fmock3", "Gmock3", "Hmock3", "Hmock3.Tdup")
mock4 <- c("Fmock4", "Gmock4", "Hmock4", "Hmock4.Tdup")

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

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 <- bind_rows(m3,m4)
m34$mock <- c(3,3,3,3,4,4,4,4)
m34$mock <- as.factor(m34$mock)
```

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

mock	mean	SD
3	0.7904497	0.0383641
4	0.7310493	0.0204626

2.3. Correlation between mock samples

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

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)
QC2 <- ddply(rho.m34, .(mock), summarise, mean = mean(r), SD = sd(r))
kable(QC2)
```

mock	mean	SD
3	0.9464095	0.0490304
4	0.9524091	0.0346208

3. Technical replicates

Assessing the reproducibility of technical replicates of DNA extraction and PCR, by Spearman rank correlations.

3.1. Consistency of DNA extraction replicates

3.1.1. Prepare data

```
# add OTU column
taxbiol <- as.data.frame(psbiol@tax_table)
taxbiol$OTU <- rownames(taxbiol)
tax_table(psbiol) <- tax_table(as.matrix(taxbiol))

# rel. abundance, with tree
psbiol.g <- microbiome::aggregate_taxa(psbiol, "Genus")
psbiol.g.r <- microbiome::transform(psbiol.g, "compositional")
psbiol.df <- as.data.frame(abundances(psbiol.g.r))
```

3.1.2. Spearman correlation at OTU level

```
# prepare data frame to insert correlation metrics in next step
biolsam <- subset(meta(psbiol.g.r), select = c(Description, ContainerID,
      Treatment, Diet, Type))

# Spearman correlations
dcor <- cor(as.matrix(psbiol.df), method = "spearman")
mcor <- reshape2::melt(dcor)

# select only the relevant comparisons
mcors <- mcor[mcor$Var1 != mcor$Var2, ] # exclude self-comparisons
colnames(mcors) <- c("Description", "Var2", "r") # rename col1 Description
mcors <- merge(mcors, biolsam, by = "Description") # import metadata sample1
colnames(mcors)[c(1:2)] <- c("Var1", "Description") # rename col2 Description
mcors <- merge(mcors, biolsam, by = "Description") # import metadata sample2
mcors <- mcors[mcors$ContainerID.x == mcors$ContainerID.y &
      mcor$Type.x == mcor$Type.y, ] # subset duplicates
rownames(mcors) <- NULL # reset rownames
mcors2 <- mcors[-c(4,8:12, 16, 20:24, 30, 31, 33:36, 41,
      44:48, 52, 55, 57:60, 66, 68:72, 76:78),] # remove double comparisons
mcors2 <- mcors2[, -c(8:11)] # remove double columns
mcors2$typecode <- ifelse(mcors2$Type.x == "substrate", "K", "N")
mcors2$Sample <- interaction(mcors2$ContainerID.x, mcor$typecode)
colnames(mcors2) <- c("Dup1", "Dup2", "r", "ContainerID",
      "Treatment", "Diet", "Type", "typecode", "Sample")
mcors3 <- mcor2[, c(3, 5:7, 9)]
mcor3$Sample <- droplevels(mcor3$Sample)

# summarise
mcor.sum <- ddply(mcor3, .(Sample, Diet, Type, Treatment), summarise,
      mean = mean(r), SD = sd(r))
kable(mcor.sum)
```

Sample	Diet	Type	Treatment	mean	SD
15.K	CF	substrate	S/E	1.0000000	0.0000000
18.K	CM	substrate	Si/Es	0.8276126	0.1035742
23.K	CM	substrate	Ss/E	0.2084898	0.1125500
32.K	CM	substrate	S/E	0.9399580	0.0073766
4.K	CF	substrate	Si/Es	0.6458666	0.0797464
15.N	CF	larvae	S/E	0.9681109	0.0349204
30.N	CM	larvae	S/E	0.9581304	0.0083875

Correlations are low among replicates of 4.K ($r = 0.65$) and 23.K (0.21). 4.K is inoculated CF at t0, 23.K is autoclaved CM at t0.

3.1.3. Conclusion

Non-sterile larva and substrate samples were very reproducible, *i.e.* with isolation replicates having high Spearman correlations and showing smaller variation. The inoculated substrates and especially sterile sub-

strates at day 0 turned out unreproducible, as correlations were low for sterile CM.

3.2. Consistency of PCR replicates

3.2.1. Load data

```
# add OTU column and remove tree
taxtech <- as.data.frame(pstech@tax_table)
taxtech$OTU <- rownames(taxtech)
tax_table(pstech) <- tax_table(as.matrix(taxtech))

# rel. abundance, with tree
pstech.g <- microbiome::aggregate_taxa(pstech, "Genus")
pstech.g <- microbiome::transform(pstech.g, "compositional")
pstech.df <- as.data.frame(abundances(pstech.g))
```

3.2.2. Spearman correlation

```
# prepare data frame to insert correlation metrics in next step
techsam <- subset(meta(pstech.g),
                  select = c(Description, ContainerID, Treatment, Diet, Type))

# Spearman correlations, using rcorr function (returns r and P-values)
dcor3 <- cor(as.matrix(pstech.df), method = "spearman")
mcor3 <- reshape2::melt(dcor3)

# select only the relevant comparisons
mcor3s <- mcor3[mcor3$Var1 != mcor3$Var2, ] # exclude self-comparisons
colnames(mcor3s) <- c("Description", "Var2", "r") # rename col1 Description
mcor3s <- merge(mcor3s, techsam, by = "Description") # import metadata sample1
colnames(mcor3s)[c(1:2)] <- c("Var1", "Description") # rename col2 Description
mcor3s <- merge(mcor3s, techsam, by = "Description") # import metadata sample2
mcor3s <- mcor3s[mcor3s$ContainerID.x == mcor3s$ContainerID.y &
                 mcor3s$Type.x == mcor3s$Type.y, ] # subset duplicates
rownames(mcor3s) <- NULL # reset rownames
mcor3s2 <- mcor3s[-c(2, 4, 6, 8, 11, 13:17, 19:26, 28:31,
                   33:34, 36:49, 51:56), ] # remove double comparisons
mcor3s2 <- mcor3s2[, -c(8:11)] # remove double columns
rownames(mcor3s2) <- NULL
mcor3s2$Sample <- c("15.N", "16.K", "18.N", "33.M", "4.K1", "4.K1", "4.K1",
                  "L18", "L3", "L3", "L3", "L7")

colnames(mcor3s2) <- c("Dup1", "Dup2", "r", "ContainerID",
                    "Treatment", "Diet", "Type", "Sample")

# summarise
mcors.sum3 <- ddply(mcor3s2, .(Sample, Diet, Type, Treatment), summarise,
                   mean = mean(r), SD = sd(r))
kable(mcors.sum3)
```

Sample	Diet	Type	Treatment	mean	SD
15.N	CF	larvae	S/E	1.0000000	NA
16.K	CF	substrate	S/E	1.0000000	NA
18.N	CM	larvae	Si/Es	0.9724949	NA
33.M	CM	substrate	Si/Es	0.9539984	NA
4.K1	CF	substrate	Si/Es	0.6588479	0.0881887
L18	na	eggs	Es	0.1215969	NA
L3	na	eggs	E	0.3373065	0.0584845
L7	na	eggs	E	0.2169765	NA

PCR replicates of egg samples show very low correlations ($r = 0.12 - 0.34$), t0 inoculated CF (4.K, $r = 0.66$) shows low correlation, rest of samples high ($r = 0.95 - 1.00$).

3.2.3. Conclusion

Non-sterile samples of both larvae and substrates show high correlations, *i.e.* high reproducibility. However, inoculated CF t0 substrates have low correlation (0.66), and eggs have very low correlations (0.12 - 0.34), suggesting that these community data are very variable and not reproducible.

4. Export table

Supplementary Table S5.

```
mcors.sum$Replication <- "DNA extraction"
mcors.sum3$Replication <- "PCR"
replicates.sum <- rbind(mcors.sum, mcors.sum3)
write.csv(replicates.sum, "./tables/Supplementary_Table_S5.csv")
```