

Beta diversity: phylum-level stacked barplot and genus-level heatmaps

Stijn Schreven

1 June 2021

Contents

Introduction	1
Load packages	2
Input files	2
1. Prepare data	2
2. Plot presets	3
3. Barplot Phylum	3
3.1. Prepare data	4
3.2. Plot	4
4. Heatmaps at genus level	5
4.1. Top genera - input	5
4.2. Top genera - plot	6
4.3. All genera - data	7
4.4. All genera - plot	7
4.5. All genera - export as table	10
5. Taxa without annotation	10

Introduction

We selected the five most abundant phyla by using `aggregate_taxa(..., top = 5)`; the most abundant genera were selected based on thresholds of maximum relative abundance $> 10\%$ and present in $> 10\%$ of samples, yielding a list of 38 genera.

Load packages

```
library(phyloseq)
library(microbiome)
library(microbiomeutilities)
library(plyr)
library(magrittr)
library(emmeans)
library(sciplot)
library(reshape2)
library(vegan)
library(knitr)
library(ggplot2)
library(viridis)
```

Input files

```
ps1.exp <- readRDS("./phyobjects/ps1.exp.rds")
print(ps1.exp)
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 2424 taxa and 284 samples ]
## sample_data() Sample Data:      [ 284 samples by 14 sample variables ]
## tax_table()   Taxonomy Table:    [ 2424 taxa by 6 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 2424 tips and 2423 internal nodes ]
```

1. Prepare data

```
# add OTU column, remove tree
ps1.exp.com <- ps1.exp
taxic <- as.data.frame(ps1.exp.com@tax_table)
taxic$OTU <- rownames(taxic)
tax_table(ps1.exp.com) <- tax_table(as.matrix(taxic))
ps1.exp.com@phy_tree <- NULL

# aggregate top 5 phyla, without tree
ps.phylum <- aggregate_taxa(ps1.exp.com, "Phylum", top = 5)
ps.phylum <- microbiome::transform(ps.phylum, "compositional")

# aggregate to genus, with phy_tree
totg <- microbiome::aggregate_taxa(ps1.exp, "Genus")
# remove taxa with 0 abundance
totg <- prune_taxa(taxa_sums(otu_table(totg)) > 0, totg)
# add OTU column
totg.tax <- as.data.frame(totg@tax_table)
```

```

totg.tax$OTU <- rownames(totg.tax)
tax_table(totg) <- tax_table(as.matrix(totg.tax))
# relative abundance
totg.r <- microbiome::transform(totg, "compositional")

# selection of most abundant and prevalent genera
## max(abundance) > .1 (at least 10% in a sample)
## and prevalence > .1 (10% of samples)
totg.otu <- as.data.frame(t(abundances(totg.r)))
m.totg.otu <- reshape2::melt(totg.otu)
colnames(m.totg.otu) <- c("OTU", "abund")
sum.totg <- ddply(m.totg.otu, .(OTU), summarise,
  max = max(abund),
  prev = sum(abund > 0)/length(abund))
top.totg <- subset(sum.totg, max > .1 & prev > .1) # 38 genera
totg.top <- prune_taxa(taxa_names(totg.r) %in% droplevels(top.totg$OTU), totg.r)
totg.top <- prune_samples(sample_sums(otu_table(totg.top)) > 0, totg.top)

```

2. Plot presets

```

labs_stack <- as_labeller(c(
  "0" = "day 0", "5" = "day 5", "10" = "day 10", "15" = "day 15",
  "CF" = "chicken feed", "CS" = "camelina", "CM" = "chicken manure",
  "larvae" = "larvae", "substrate" = "substrate"))

labs_unmatch <- as_labeller(c(
  "CF" = "chicken feed", "CS" = "camelina", "CM" = "chicken manure",
  "larvae" = "larvae", "substrate" = "substrate"))

theme_stack <- theme_bw() + theme(
  axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),
  text = element_text(size = 15),
  strip.text = element_text(size = 10),
  panel.spacing.x = unit(0, "lines"),
  strip.background = element_rect(colour = "black", fill = "white"))

theme_hm <- theme_classic() + theme(
  axis.text.y = element_text(colour = 'black', face = 'italic'),
  legend.key = element_blank(), text = element_text(size = 20),
  panel.spacing.x = unit(0, "lines"),
  strip.background = element_rect(colour = "black", fill = "white"))

```

3. Barplot Phylum

Supplementary Figure S1 in manuscript Chapter 3 in PhD thesis and Supplementary Figure S2 in manuscript submitted to *Applied and Environmental Microbiology*. Stacked errorbar plot.

3.1. Prepare data

```
# extract plot data
plot.phyl.df <- plot_composition(ps.phylum)$data
names(plot.phyl.df)[names(plot.phyl.df) == "Sample"] <- "Description"
plot.phyl.df <- merge(meta(ps.phylum), plot.phyl.df, by = "Description")

# summarise, mean and SE
phyl <- ddply(plot.phyl.df, .(Diet, Timepoint, Density, Type, OTU),
              summarise, mean = mean(Abundance), se = se(Abundance))

# edit and reorder phyla
levels(phyl$OTU)[1] <- "Unassigned taxa"
phyl$OTU <- sub(pattern = "[a-z]__", replacement = "", phyl$OTU)
phyl$OTU <- as.factor(phyl$OTU)
phyl$OTU <- factor(phyl$OTU, levels(phyl$OTU)[c(1:3,5,6,4)])

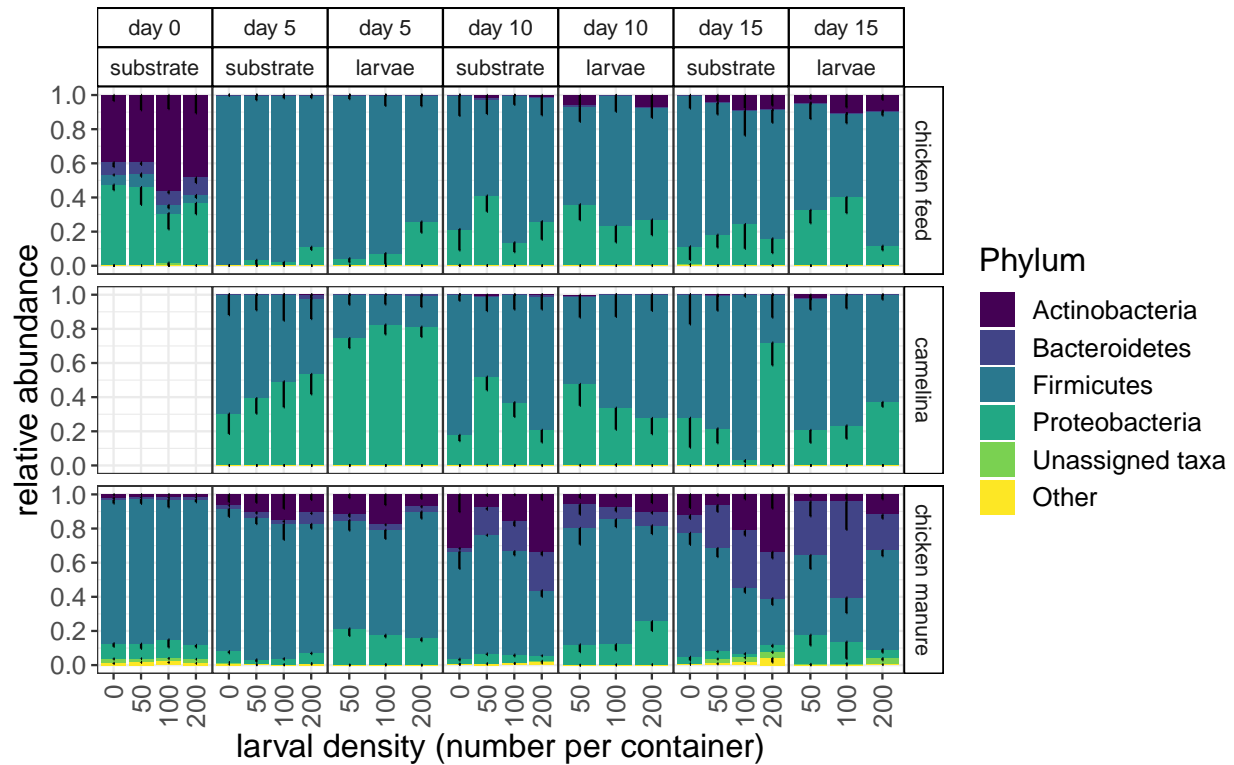
# calculate confidence intervals (mean +/- SE)
# lower is mean - se, upper is only mean (so that in plot only lower limit is displayed)
phyl <- phyl %>% mutate(lower = mean - se, upper = mean)

# calculate stacked confidence interval values
#(level 2 = mean(level 1) + CI(level 2) ; level 3 = mean(1) + mean(2) + CI(level 3))

phyl[phyl$OTU=='Unassigned taxa',8:9] <- phyl[phyl$OTU=='Other',6] +
  phyl[phyl$OTU=='Unassigned taxa',8:9]
phyl[phyl$OTU=='Proteobacteria',8:9] <- phyl[phyl$OTU=='Other',6] +
  phyl[phyl$OTU=='Unassigned taxa',6] + phyl[phyl$OTU=='Proteobacteria',8:9]
phyl[phyl$OTU=='Firmicutes',8:9] <- phyl[phyl$OTU=='Other',6] +
  phyl[phyl$OTU=='Unassigned taxa',6] + phyl[phyl$OTU=='Proteobacteria',6] +
  phyl[phyl$OTU=='Firmicutes',8:9]
phyl[phyl$OTU=='Bacteroidetes',8:9] <- phyl[phyl$OTU=='Other',6] +
  phyl[phyl$OTU=='Unassigned taxa',6] + phyl[phyl$OTU=='Proteobacteria',6] +
  phyl[phyl$OTU=='Firmicutes',6] + phyl[phyl$OTU=='Bacteroidetes',8:9]
phyl[phyl$OTU=='Actinobacteria',8:9] <- phyl[phyl$OTU=='Other',6] +
  phyl[phyl$OTU=='Unassigned taxa',6] + phyl[phyl$OTU=='Proteobacteria',6] +
  phyl[phyl$OTU=='Firmicutes',6] + phyl[phyl$OTU=='Bacteroidetes',6] +
  phyl[phyl$OTU=='Actinobacteria',8:9]
```

3.2. Plot

```
pbar3 <- ggplot(phyl, aes(x = Density, y = mean, fill = OTU)) +
  geom_bar(stat = "identity", position = "stack") +
  scale_fill_viridis(discrete = T, option = "D") +
  labs(y = "relative abundance", x = "larval density (number per container)",
       fill = "Phylum") +
  facet_grid(Diet ~ Timepoint + Type, scales = "free_x", labeller = labs_stack) +
  geom_errorbar(data = phyl, aes(ymin = lower, ymax = upper), width = 0) +
  scale_y_continuous(limits = c(0,1), n.breaks = 6) +
  theme_stack
pbar3
```



```
ggsave(plot=pbar3, "./figures/Phylum_stack5_SE.png", h = 5, w = 8)
ggsave(plot=pbar3, "./figures/Phylum_stack5_SE.pdf", h = 125, w = 200, u = "mm")
```

4. Heatmaps at genus level

Median relative abundances.

4.1. Top genera - input

```
# extract heatmap plot data
ps.hm.df <- plot_heatmap(totg.top, method = "CAP", distance = "bray",
                        formula = ~ Diet * Density * Timepoint * Type)$data

# rename levels
ps.hm.df$Diet <- revalue(ps.hm.df$Diet, c(
  "CF" = "chicken feed", "CS" = "camelina", "CM" = "chicken manure"))

# clean up taxonomic names
ps.hm.df$taxname <- ifelse(ps.hm.df$Genus == "g__",
  yes = paste(ps.hm.df$Family, "(unassigned)"),
  no = paste(ps.hm.df$Genus))
ps.hm.df$taxname <- gsub(ps.hm.df$taxname, pattern = "[a-z]__", replacement = "")
ps.hm.df$taxname <- as.factor(ps.hm.df$taxname)
#levels(ps.hm.df$taxname)
levels(ps.hm.df$taxname)[1] <- "Lactobacillales (unassigned)"
levels(ps.hm.df$taxname)[37] <- "Lachnospiraceae (uncultured)"
```

```

# summarise per treatment (Diet*Time*Type*Density)
ps.hm.sum <- ddpby(ps.hm.df, .(Timepoint, Diet, Type, Density, taxname, OTU),
  summarise, median = median(Abundance), mean = mean(Abundance))
ps.hm.sum$group <- interaction(ps.hm.sum$Diet, ps.hm.sum$Density, ps.hm.sum$Timepoint,
  ps.hm.sum$Type, drop = TRUE)

# distance matrix Bray-Curtis for taxa clustering
ps.hm.cast <- reshape2::dcast(ps.hm.sum[,c(5,7,9)], taxname ~ group, value.var = "median")
rownames(ps.hm.cast) <- ps.hm.cast[,1]
ps.hm.mat <- as.matrix(ps.hm.cast[,c(2:72)])
bray30.sp <- vegdist(ps.hm.mat, method = "bray")

# reorder levels by clustering
sp.order <- hclust(bray30.sp)$order
ps.hm.sum$taxname <- factor(ps.hm.sum$taxname, levels(ps.hm.sum$taxname)[sp.order])

```

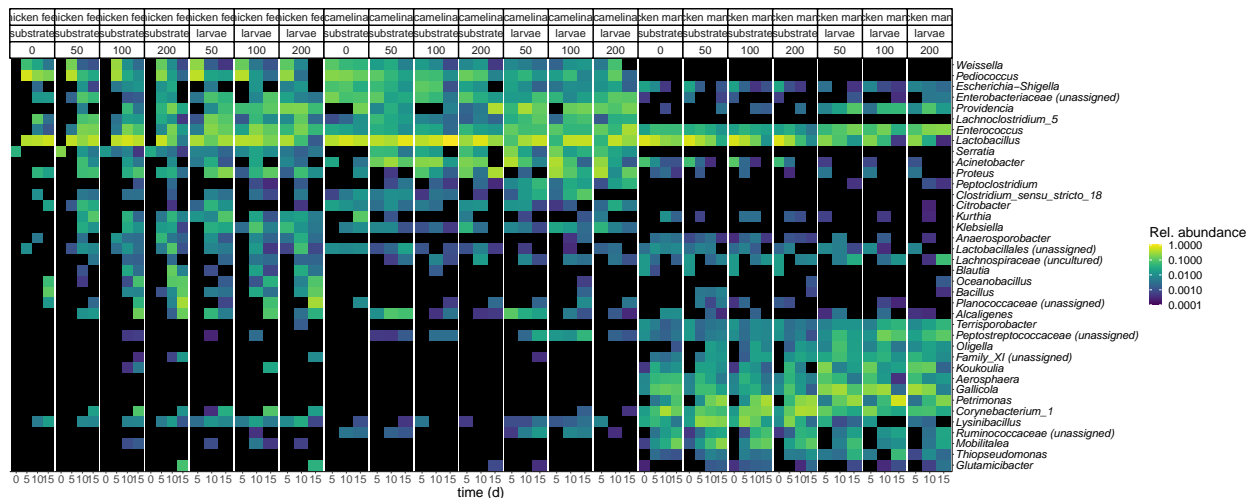
4.2. Top genera - plot

Figure 5 in manuscript Chapter 3 in PhD thesis and Figure 4 in manuscript submitted to *Applied and Environmental Microbiology*. With Bray-Curtis complete-linkage clustering of taxa.

```

p.ps.hm <- ggplot(ps.hm.sum, aes(x=Timepoint, y=taxname)) +
  geom_tile(aes(fill=mean)) +
  scale_fill_viridis("Rel. abundance", option = "D",
    na.value = "black", trans = "log10",
    limits = c(.0001, 1),
    labels = function(n){format(n, scientific = F)}) +
  facet_grid(~ Diet + Type + Density, scales = "free") +
  labs(x = "time (d)", y = NULL) +
  scale_y_discrete(position = "right") +
  theme_hm
p.ps.hm

```



```

ggsave(plot = p.ps.hm, "./figures/Genus_heatmapTop.png", h = 10, w = 25)
ggsave(plot = p.ps.hm, "./figures/Genus_heatmapTop.pdf", h = 250, w = 625, u = "mm")

#dat <- subset(p.ps.hm$data, mean > 0)
#min(dat$mean) # min is .00025, so choose limits c(0.0001, 1) for heatmap.

```

4.3. All genera - data

```

# extract heatmap plot data
hm.all.df <- plot_heatmap(totg.r, method = "CAP", distance = "bray",
                          formula = ~ Timepoint * Diet * Type * Density)$data
# make 1 group factor
hm.all.df$group <- interaction(hm.all.df$Type, hm.all.df$Timepoint, hm.all.df$Density,
                               hm.all.df$Diet, drop = TRUE)
# make 1 factor for tax names
hm.all.df$taxname <- interaction(hm.all.df$Domain, hm.all.df$Phylum, hm.all.df$Class,
                                 hm.all.df$Order, hm.all.df$Family, hm.all.df$Genus,
                                 sep = ";", drop = TRUE)
hm.all.df$taxname <- as.factor(hm.all.df$taxname)

# summarise per group
hm.all.sum <- ddply(hm.all.df, .(Diet, Density, Timepoint, Type, group, taxname),
                   summarise, median = median(Abundance), mean = mean(Abundance))

# distance matrix Bray-Curtis for taxa clustering
hm.all.cast <- reshape2::dcast(hm.all.sum[,5:7], taxname ~ group, value.var = "median")
rownames(hm.all.cast) <- hm.all.cast[,1]
# remove empty rows
hm.all.cast$sum <- apply(hm.all.cast[2:72], 1, sum)
hm.all.cast <- subset(hm.all.cast, sum != 0)
hm.all.cast$taxname <- droplevels(hm.all.cast$taxname)
# matrix
hm.all.mat <- as.matrix(hm.all.cast[,2:72])
bray.all.sp <- vegdist(hm.all.mat, method = "bray")
sp.all.order <- hclust(bray.all.sp)$order
# subset data to exclude taxa with 0 abundance
hm.all.sum <- subset(hm.all.sum, taxname %in% hm.all.cast$taxname)

# reorder genera by clustering
hm.all.sum$taxname <- factor(hm.all.sum$taxname, levels(hm.all.sum$taxname)[sp.all.order])

```

4.4. All genera - plot

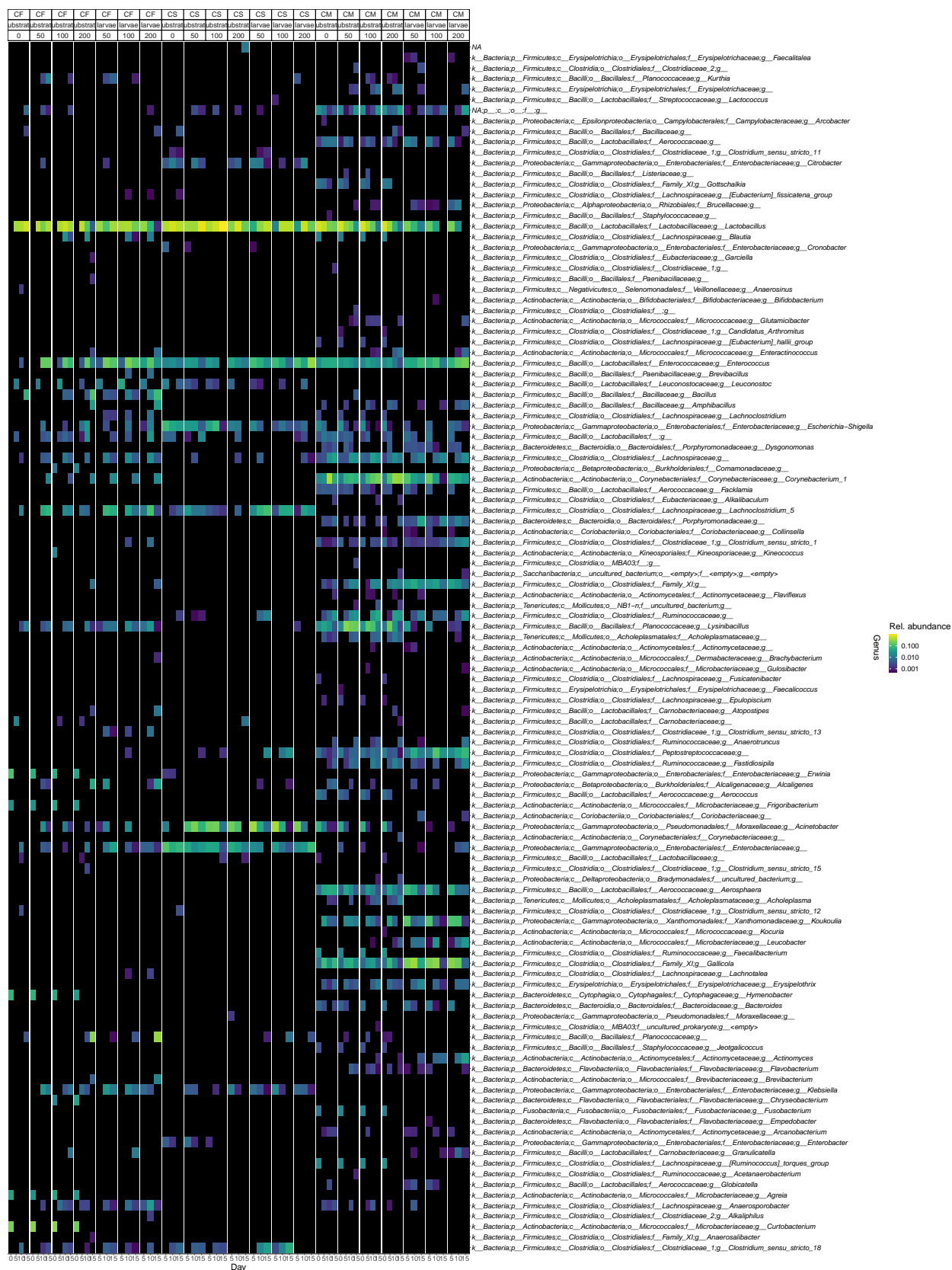
With Bray-Curtis complete-linkage clustering of taxa.

```

hm.all <- ggplot(hm.all.sum, aes(x = Timepoint, y = taxname)) +
  geom_tile(aes(fill = median)) +
  scale_fill_viridis("Rel. abundance", option = "D", na.value = "black",
                    trans = "log10") +
  facet_grid(~ Diet + Type + Density, scales = "free") +

```

```
scale_y_discrete(position = "right") +  
labs(x = "Day", y = "Genus") +  
theme_hm  
hm.all
```

```
ggsave(plot = hm.all, "./figures/Genus_heatmap_all.png", h = 40, w = 30)
```

4.5. All genera - export as table

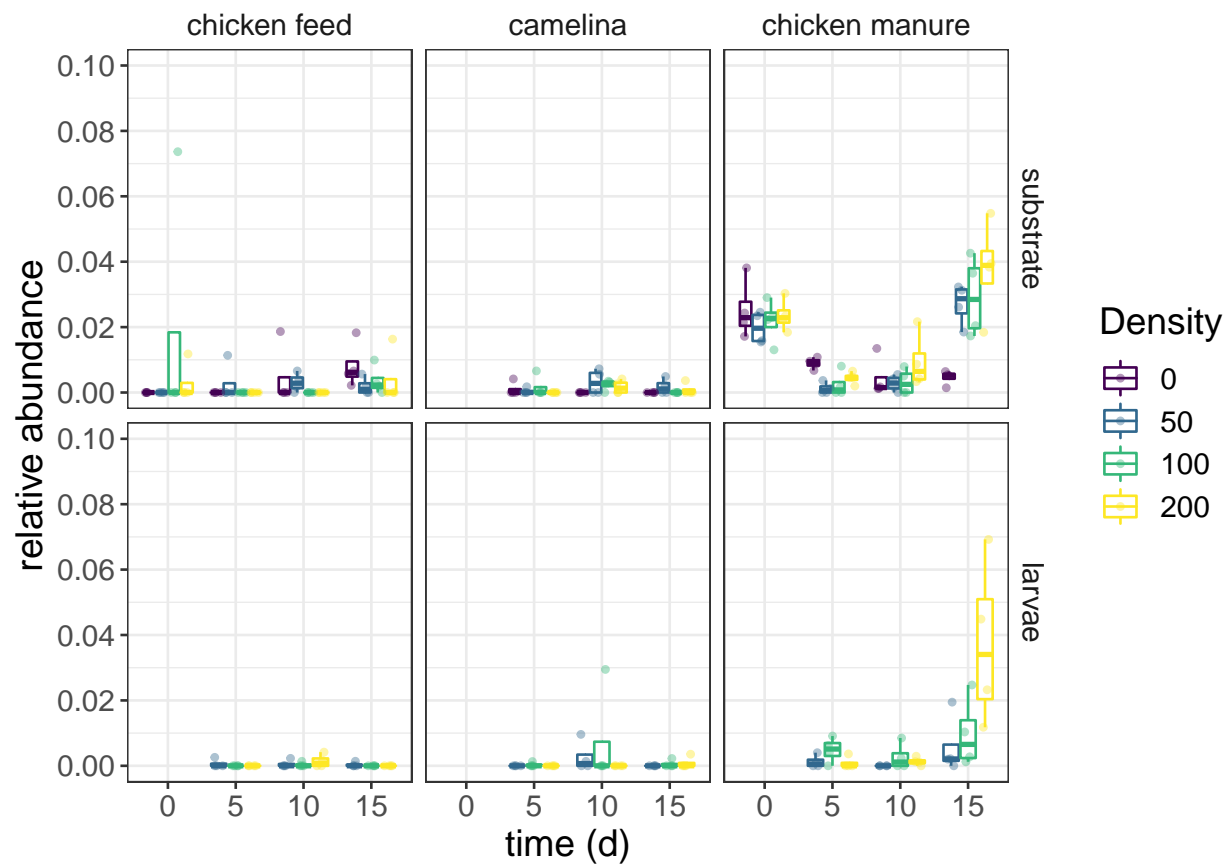
Create a dataframe with treatment by genus and median relative abundance.

```
# export dataframe to txt-file
write.table(hm.all.cast[, -c(1,73)], "./tables/Ch2_Genus_all_median.txt", sep="\t")
```

5. Taxa without annotation

```
unmatch.df <- hm.all.df[hm.all.df$taxname == "NA;p__;c__;o__;f__;g__",]

# plot
pbox.unmatch <- ggplot(unmatch.df, aes(x = Timepoint, y = Abundance,
                                         colour = Density)) +
  geom_boxplot(position = position_dodge2(),
               outlier.size = 0, outlier.alpha = 0) +
  geom_point(alpha = 0.4, size = 1, position = position_jitterdodge()) +
  scale_color_viridis(discrete=T, option="D") +
  scale_y_continuous(limits = c(0,.1), n.breaks = 6) +
  labs(y = "relative abundance", x = "time (d)") +
  facet_grid(Type~Diet, labeller = labs_unmatch) +
  theme_bw() + theme(text = element_text(size = 15),
                     strip.background = element_blank())
pbox.unmatch
```



```
ggsave("./figures/UnmatchedTaxa_plot.png", h = 5, w = 7)
```