Data analysis example

Overview

Here we present an example analysis report from the OCMS. The data are from a published study that were made publically available. The original article can be found here and the raw data can be downloaded from here. The data are 16S rRNA amplicon sequencing data (V4 region sequenced on the MISeq) from wild-type mice and MMP-9 deficient mice in either untreated or DSS treatemd conditions (i.e. inflammatory bowel disease model. Please see the original article for more details.

Pre-processing

The sequencing have been run through the OCMS Dada2 pipeline (please see the example dada2 report) and the analyses performed here are based on the amplicon sequence variant (ASV) table that is output from that pipeline. It should be noted here that the sequence quality in a lot of the samples was poor. This appeared to cause a problem for Dada2 as we were unable to assign taxononmy even at the phylum level for the majority of sequences from 15 samples. These samples have been removed for the purposes of this analysis i.e. as this is simply an example report.

Setup

If you are running the analysis step-by-step, make sure you have read the GitHub page so that the relevant dependencies are installed and the configuration file is set up correctly.

Here we set up the script with colours that we want to use for different groups as well as the required packages and script dependencies.

Set up colours for the different groups

Here we set up colour maps for subsequent plots i.e. group colours

```
group.colours <- c("grey", "purple", "black", "blue")
names(group.colours) <- c("WT:water", "WT:DSS", "MMP-9KO:water", "MMP-9KO:DSS")
```

Attach relevant R packages

Next we attach the packages that we will need for the analyses.

```
library(knitr)
library(gridExtra)
library(phyloseq)
library(ggplot2)
library(data.table)
library(vegan)
library(pheatmap)
library(RColorBrewer)
```

Source the configuration file and scripts from the repositories that this script depends on

The required repositories that contain R scripts for functions used in the analyses should have been downloaded and the locations of these specified in ocms_example_16srRNA.config.R.

```
source("ocms_example_16SrRNA.config.R")
source(paste0(AmpSeqKit.dir, "/R/diversity.R"))
source(paste0(AmpSeqKit.dir, "/R/relab.R"))
source(paste0(AmpSeqKit.dir, "/R/plots.R"))
source(paste0(AmpSeqKit.dir, "/R/utils.R"))
source(paste0(AmpSeqKit.dir, "/R/differential_abundance.R"))
source(paste0(NGSKit.dir, "/R/deseq2_helper.R"))
source(paste0(MIGTranscriptomeExplorer.dir, "/R/MIGTranscriptome_plotting.R"))
source(paste0(MIGTranscriptomeExplorer.dir, "/R/MIGTranscriptome_retrieval.R"))
```

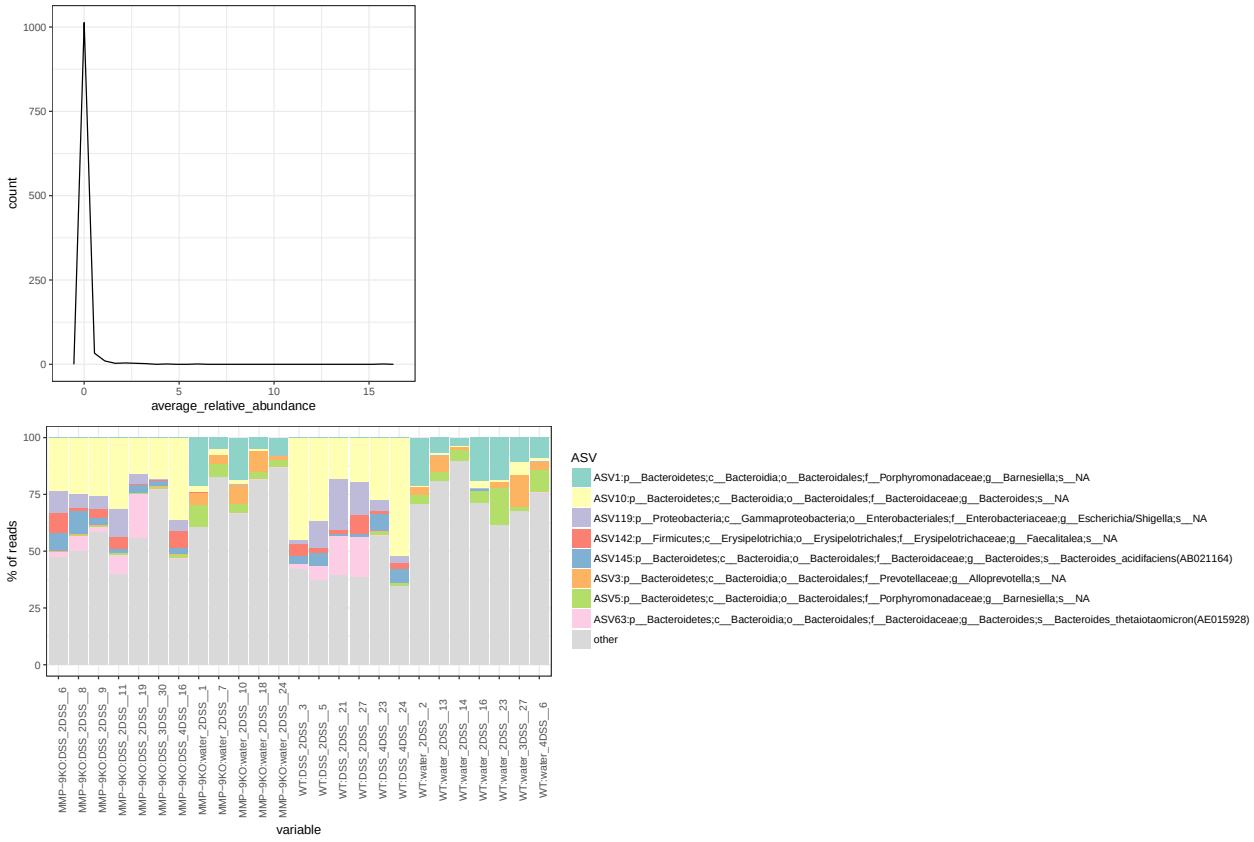
Now we are ready to do the analyses.

General features

First we assess the relative abundance of the ASVs along with their taxonomic assignments. We look at the average (across samples) relative abundance distribution of ASVs and the ASV/taxonomic distribution across individual samples. For the purposes of the second plot we are plotting the ASVs that are present at an abundance pf $> 5\%$ in at least 5 samples. All of the rest of the ASVs are lumped into the "other" category.

```
# Read metadata
metadata <- read.csv(metadata, header=T, stringsAsFactors=F, sep="\t")

# Remove whitespace in the genotype column
metadata$Genotype <- gsub(" ", "", metadata$Genotype)

# Create a new variable that combines genotype and phenotype (phenotype in this case is DSS treatment)
metadata$condition <- paste0(metadata$Genotype, ":", metadata$Phenotype)
rownames(metadata) <- metadata$X.sample_name

# Read in ASV counts data
asvs <- data.frame(data.table::fread(asv.counts, header=T, stringsAsFactors=F, sep="\t"))

# Format the data so that the ASV IDs are the rownames
rownames(asvs) <- asvs$sequence
asvs <- asvs[,2:ncol(asvs)]

# Remove unneccessary bits in the column names and make sure they match the
# metadata sample names
colnames(asvs) <- gsub("stool.", "", colnames(asvs))

# Get rid of the troublesome samples
asvs <- asvs[,colSums(asvs[grep("p__NA", rownames(asvs)),]) < 5000]

# Subset metadata
metadata <- metadata[metadata$X.sample_name %in% colnames(asvs),]

# Order the ASV table the same as metadata
```

```r
asvs <- asvs[,metadata$X.sample_name]

# Calculate relative abundance
asv.relab <- relab(asvs)

# Take the average abundance
ave.relab <- data.frame(average_relative_abundance=rowMeans(asv.relab))

# Plot the distribution of relative abundance
distribution <- ggplot(ave.relab, aes(average_relative_abundance))
distribution <- distribution + geom_freqpoly() + theme_bw()

# plot the ASVs that are > 1% on average
toplot <- filterRows(asv.relab, 5, 5)
toplot["other",] <- unname(100-colSums(toplot))

# Add the condition for visualisation purposes
colnames(toplot) <- paste0(metadata$condition, "_", colnames(toplot))
toplot <- toplot[,mixedsort(colnames(toplot))]
toplot$ASV <- rownames(toplot)

# Get barplot colours
bar.colors <- brewer.pal(nrow(toplot), "Set3")

# Reshape for plotting
toplot <- melt(toplot)
relab.bar <- plotBar(toplot) + scale_fill_manual(values=bar.colors)

# Create layout matrix for the plots
hlay <- rbind(c(1, NA, NA),
              c(2, 2, 2))
grid.arrange(distribution, relab.bar, layout_matrix=hlay)
```



As expected there are many ASVs at low abundance and few that make up the majority of the community. In this dataset it is already fairly clear that there are differences in DSS treated mice in both the mmp-9 KO and WT mice.

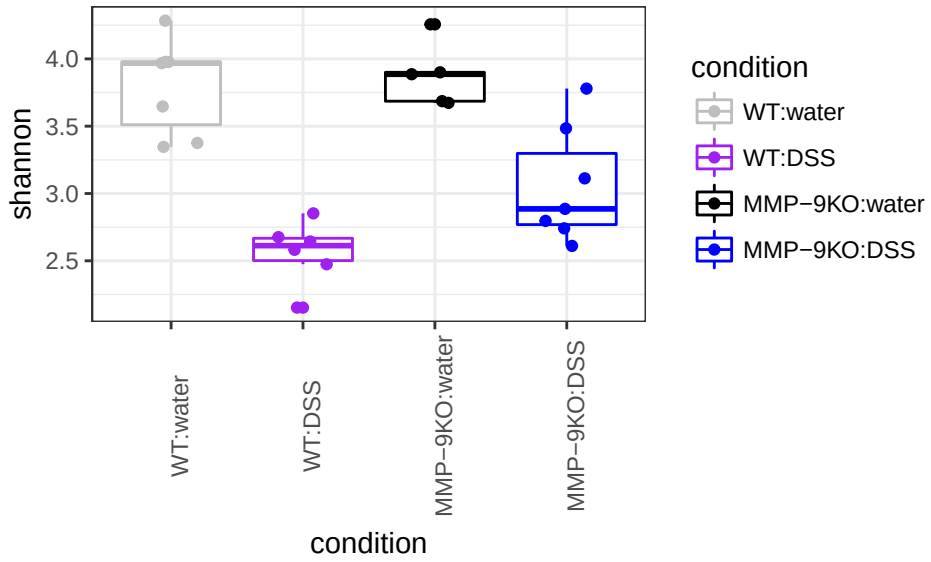
Alpha diversity

Here we asses whether there are any differences between the experimental groups in terms of alpha diversity (within-sample diversity) using the Shannon diversity index. The Kruskal-Wallis test is used to determine statistical significance of any difference.

```
# Here we use phyloseq...
dat <- otu_table(asvs, taxa_are_rows=TRUE)
sample.data <- sample_data(metadata)
dat <- merge_phyloseq(dat, sample.data)

# Shannon diversity
richness <- estimate_richness(dat, measures=c("Shannon"))

# Add Shannon diversity to metadata as is easier to plot this way
metadata$shannon <- richness$Shannon

# Make sure the plot orders the way that we want
metadata$condition <- factor(metadata$condition, levels=c("WT:water", "WT:DSS", "MMP-9KO:water", "MMP-9
```

```
# Plot alpha diversity
p1 <- ggplot(metadata, aes(x=condition, y=shannon, group=condition, colour=condition))
```

```
p2 <- p1 + geom_boxplot()
p3 <- p2 + theme_bw()
p4 <- p3 + geom_jitter(width=0.2)
p5 <- p4 + scale_colour_manual(values=group.colours)
p5 + theme(axis.text.x=element_text(angle=90))
```




```
# Create data frame that can be used as input to the kruskall wallis test
to.test <- data.frame(alpha.diversity=metadata$shannon,
                      condition=metadata$condition)

d <- multiFactorKruskalTest(to.test)
kable(d)
```

	~factor	chi.squared	p.value
2	condition	17.0401758241758	0.000693422612454179

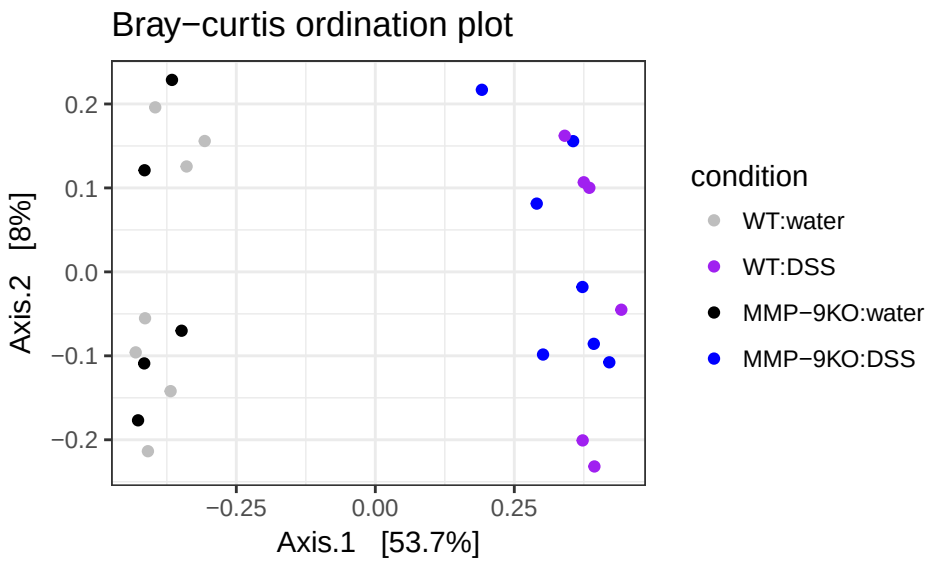
The analyses above indicate that there is a significant difference in Shannon diversity between the groups. It is clear that DSS induces a reduction in Shannon diversity regardless of genotype.

Beta diversity

Next we assess beta-diversity. The ASV counts are converted to relative abundances and bray-curtis dissimilarity is calculated using the phyloseq package.

```
# Remake phyloseq object with relative abundances
dat.relab <- otu_table(asv.relab, taxa_are_rows=TRUE)
sample.data <- sample_data(metadata)
dat.relab <- merge_phyloseq(dat.relab, sample.data)


# Get dissimilarity - Bray-Curtis in this case
dat.dissimilarity <- phyloseq::distance(dat.relab, method="bray")
```

```
# Ordination
dat.mds <- ordinate(dat.relab, "MDS", distance=dat.dissimilarity)

# Plot the ordination
p6 <- plot_ordination(dat.relab, dat.mds, color="condition")
p7 <- p6 + theme_bw()
p8 <- p7 + scale_colour_manual(values=group.colours)
p9 <- p8 + ggtitle("Bray-curtis ordination plot")
p9
```



There is a clear change in microbial composition according to DSS treatment and no visual evidence for an association with genotype. We can formally test this using a PERMANOVA test that is implemented using the adonis function in the R package vegan. Below is a table of the results using adonis with 1000 permutations.

```
# Do the PERMANOVA
perm1 <- adonis(t(asv.relab) ~ Phenotype + Genotype, method="bray", data=metadata, permutations=1000)
kable(data.frame(perm1$aov.tab))
```

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr..F.
Phenotype	1	3.4487108	3.4487108	25.2211716	0.5268242	0.0009990
Genotype	1	0.0892639	0.0892639	0.6528062	0.0136359	0.6073926
Residuals	22	3.0082519	0.1367387	NA	0.4595398	NA
Total	24	6.5462265	NA	NA	1.0000000	NA

There is a significant effect of DSS treatment as suspected and a non-sigificant effect of genotype on microbial composition.

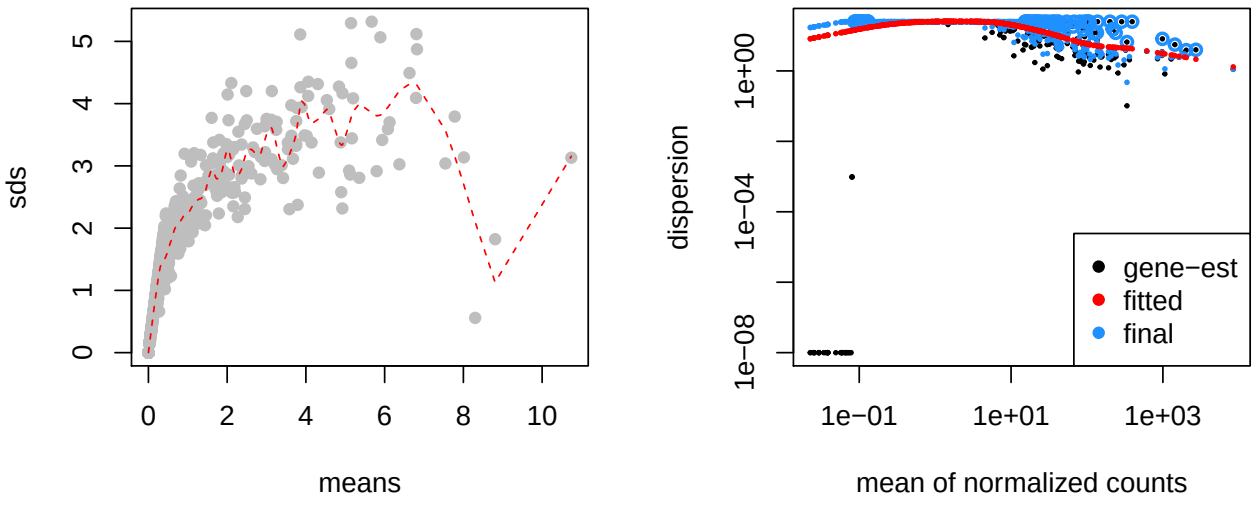
ASV differential abundance

For differential abundance analysis we use the R package DESeq2. The input to this analysis was the ASV count table.

Normalisation

First we perform the normalisation procedure and summarise the mean-variance relationship of abundance estimates across samples.

```
# Create dds object with covariates
rownames(metadata) <- metadata$X.sample_name
dds <- DESeqDataSetFromMatrix(countData = asvs,
                              colData = metadata,
                    design = ~ condition)

dds$condition <- factor(dds$condition, levels=c("WT:water", "MMP-9KO:water", "WT:DSS", "MMP-9KO:DSS"))

# Run DESeq2 analysis
dds.lrt <- DESeq(dds, test="LRT", fitType="local", reduced=~1)
res <- results(dds.lrt)

# get transformed counts
nt <- normTransform(dds)
ntdf <- data.frame(assay(nt))

# Plot mean-variance relationship and dispersion
# estimates

par(mfrow=c(1,2))
plotMeanSd(ntdf)
plotDispEsts(dds.lrt)
```



Differential abundance

Significantly differentially abundant ASVs have been called using the likelihood ratio test (LRT) implementeed in DESeq2. This takes all four levels in the factor condition and looks for differentially abundant features across all groups simultaneously. We plot a heatmap of differentially abundant ASVs in order to ge a feel for the conditions where differences are observed. The number of significantly differentially abundant ASVs at an adjusted p-value < 0.05 are provided in the table at the top of this section.

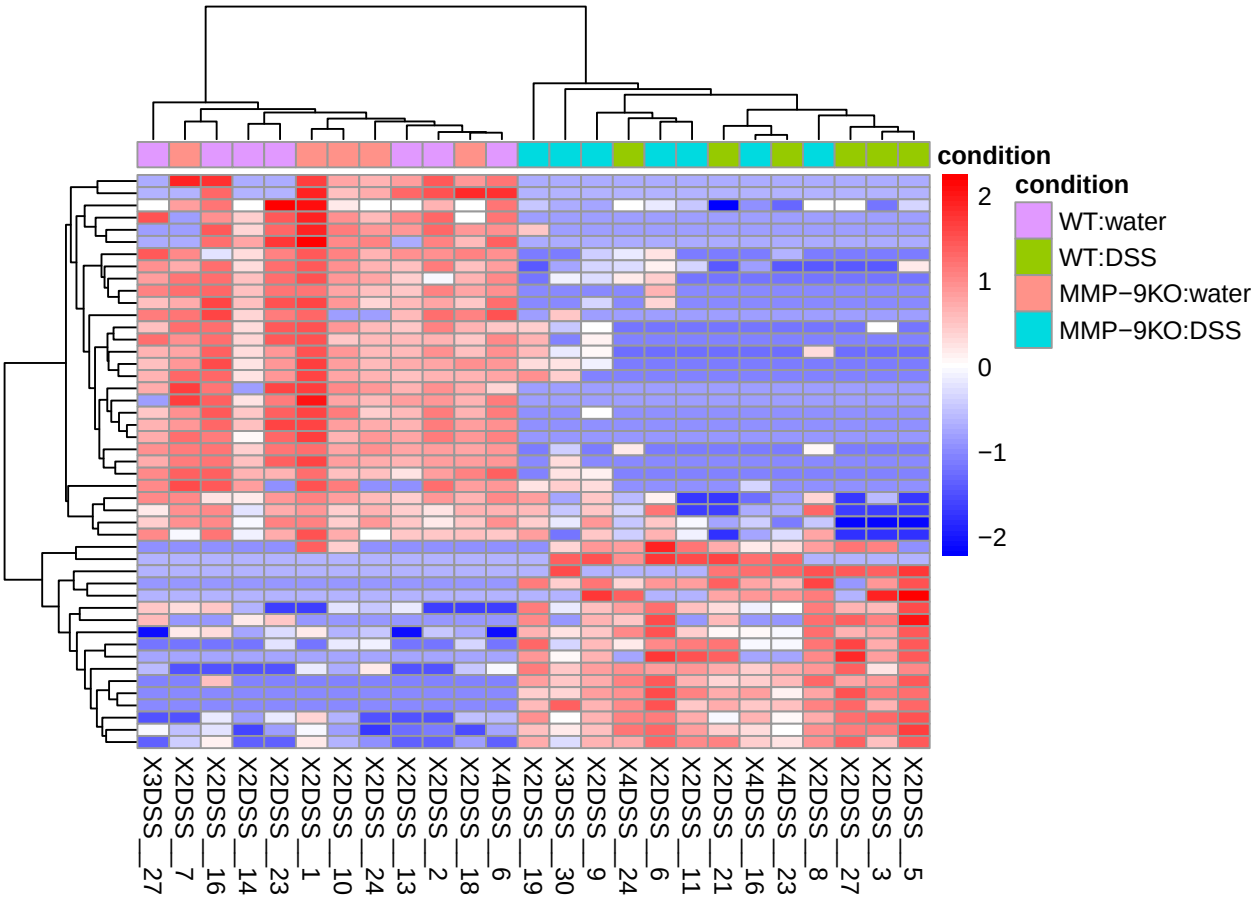
```
# Get differential abundance results
res <- results(dds.lrt)
```

```
res2 <- data.frame(res@listData)
rownames(res2) <- rownames(res)

# Subset for genes that are differentially regulated
res.diff <- res2[res2$padj < 0.05 & !(is.na(res2$padj)),]

# Get number different
ndiff <- nrow(res.diff)
df <- data.frame("Number DE" = ndiff)
kable(df, caption="number of genes differentially expressed at p < 0.05")
```

Table 3: number of genes differentially expressed at p < 0.05

Number.DE
47

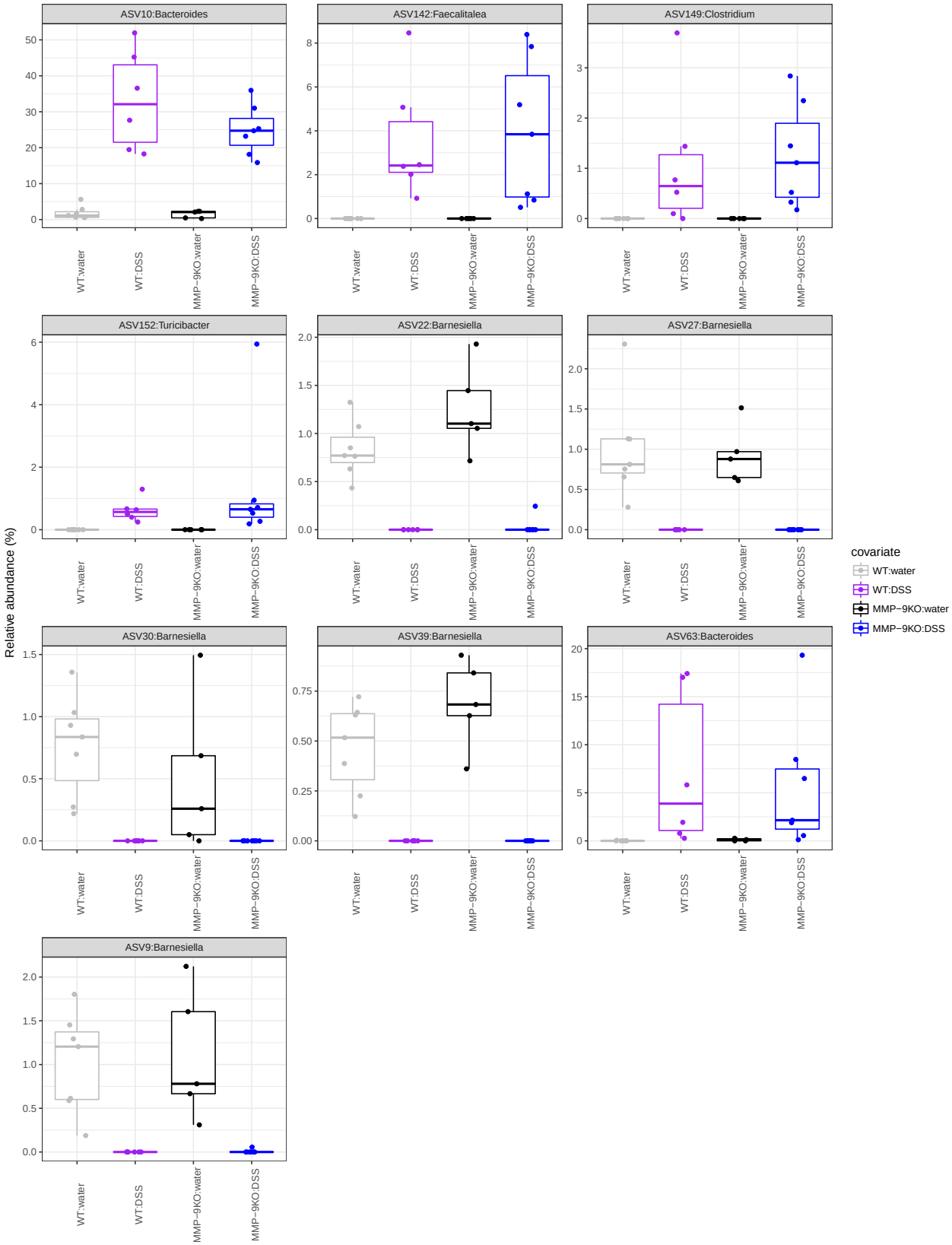
```
# Get matrix with diff ASV's
ntdf.diff <- ntdf[rownames(res.diff),]

# Make annotation
col.annotation <- data.frame(condition=metadata$condition)
rownames(col.annotation) <- paste0("X",rownames(metadata))

# Heatmap to show the differences
heatmapMatrixWithSampleAnnotation(ntdf.diff, labels=FALSE, col.annotation)
```



Differentially abundant ASVs

Below we show the abundances of the significantly differentially abundant ASVs (top 10).

```
# Order by p-value
res.diff <- res.diff[order(res.diff$padj, decreasing=FALSE),]

# Plot the top 10
res.diff.top <- res.diff[1:10,]
asv.relab.diff <- asv.relab[rownames(res.diff.top),]

# Get the names for the genera without all levels in the name
genera <- getShortNames(rownames(asv.relab.diff), type="ASV", level="genus")
asv <- unlist(strsplit(rownames(asv.relab.diff),":"))
asv <- asv[seq(1, length(asv),2)]
rownames(asv.relab.diff) <- paste0(asv, ":", genera)

# Plot the ASVs of interest. Although this says plotGeneOfInterest() it is a genreeic function for any
p <- plotGeneOfInterest("", asv.relab.diff, metadata, variable="condition") + ylab("Relative abundance
p <- p + facet_wrap(~test_id, nrow=4, ncol=3, scale="free")
p + scale_colour_manual(values=group.colours)
```


ASV10:Bacteroides
ASV142:Faecalitalea
ASV149:Clostridium
ASV152:Turicibacter
ASV22:Barnesiella
ASV27:Barnesiella
ASV30:Barnesiella
ASV39:Barnesiella
ASV63:Bacteroides
ASV9:Barnesiella
Relative abundance (%)
WT:water
WT:DSS
MMP−9KO:water
MMP−9KO:DSS
covariate
WT:water
WT:DSS
MMP−9KO:water
MMP−9KO:DSS