Lab7

David Wilson

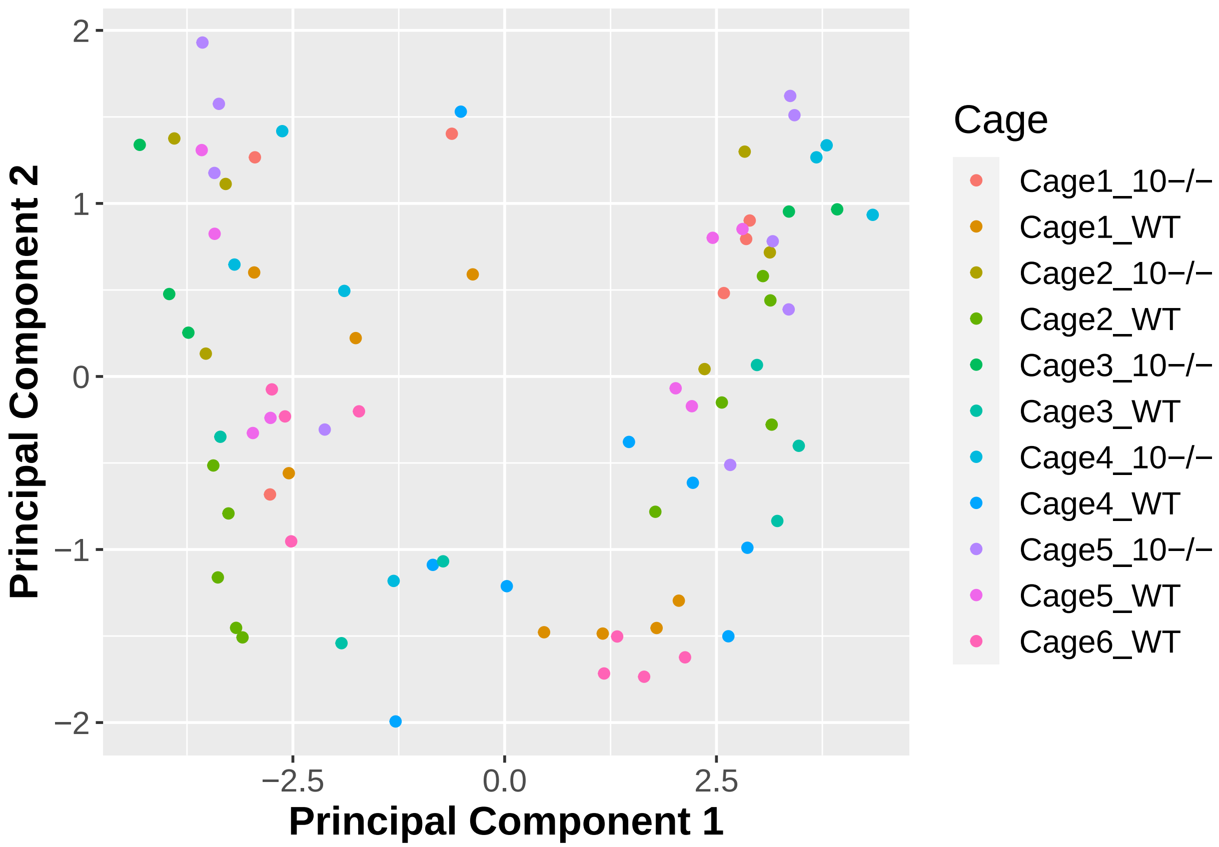
4/23/2022

```
#data preprocessing
setwd('/Users/dwils152/Google Drive/My Drive/spring2022/adv_stats/BINF8310/Lab7')
data <-read.table('prePostPhylum.txt', header=TRUE, sep="\t")
num_cols <- ncol(data)
col_classes <- c(rep("character",4), rep("numeric", num_cols-4))
data <-read.table('prePostPhylum.txt',header=TRUE,sep="\t",colClasses=col_classes) #?
data_spec <- data[,5:10]

#build pca model
pcoa <- princomp(data_spec)
```

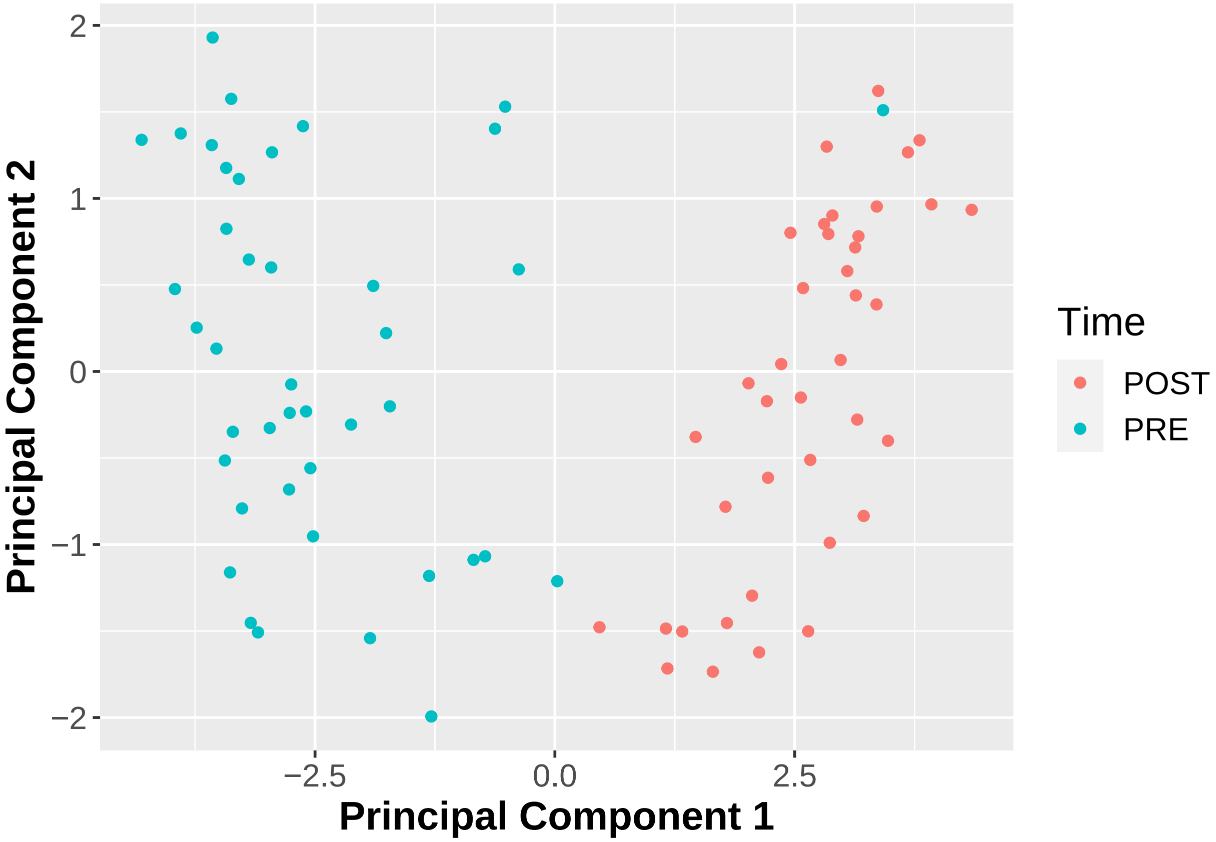
(1) Download the dataset. Perform PCA ordination.

(2) Graph PCA1 vs. PCA2. Make three versions of the graph. One colored by genotype, one colored by cage and one colored by timepoint (pre-vs-post)

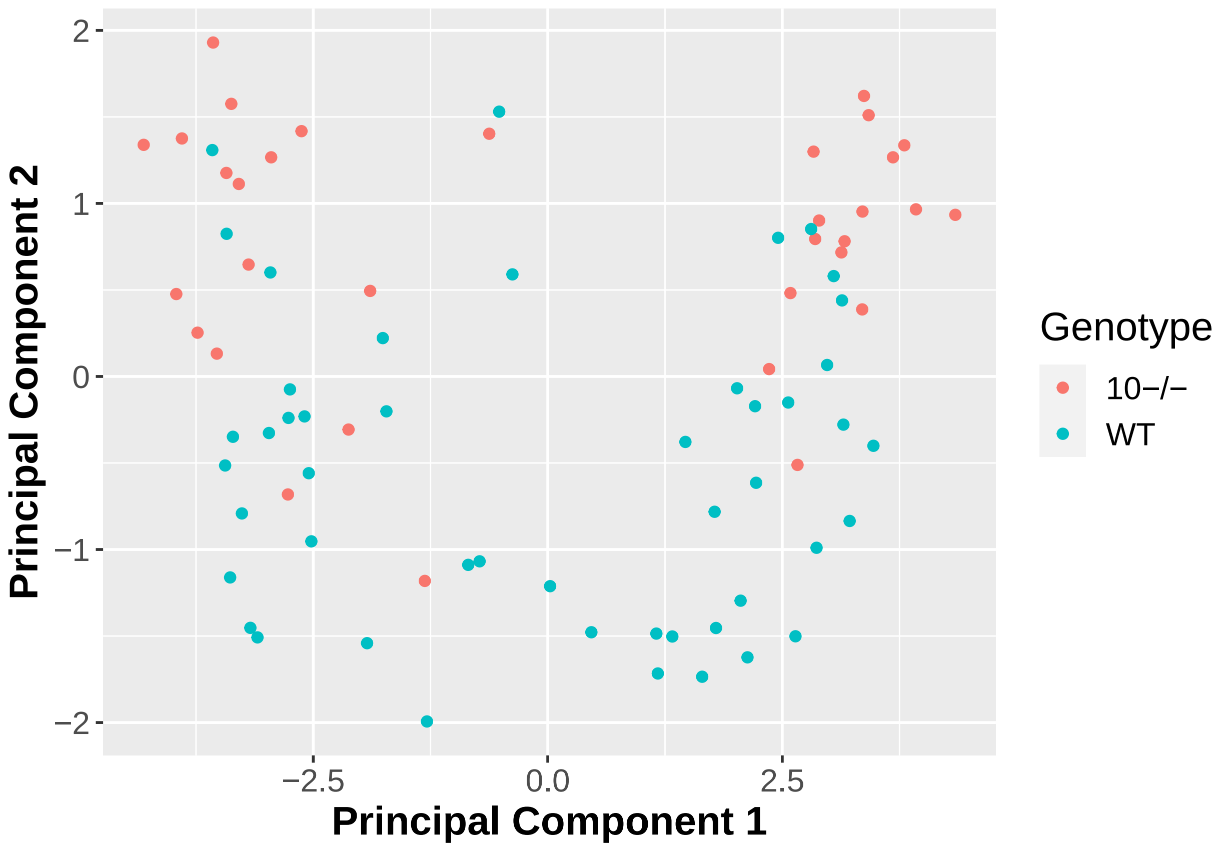
```
#plot PCA
library(ggplot2)

pca_df <- data.frame(data$cage, data$time, data$genotype, pcoa$scores[,1], pcoa$scores[,2])
names(pca_df) <- c("Cage", "Time", "Genotype", "x", "y")

ggplot(pca_df, aes(x=x, y=y, color=Cage)) +
  geom_point() + ylab("Principal Component 2") + xlab("Principal Component 1") + theme(text=element_tex
```



```
ggplot(pca_df, aes(x=x, y=y, color=Time)) + geom_point() + ylab("Principal Component 2") + xlab("Princi
```



```
ggplot(pca_df, aes(x=x, y=y, color=Genotype)) + geom_point() + ylab("Principal Component 2") + xlab("Pri
```



```
library(dplyr)
```

(3) Fill in the following table for p-values testing the null hypothesis for PCA 1 and 2. For cage, use a way one-ANOVA. For genotype and timepoint ("pre" vs "post") use a t-test

```
## Warning: package 'dplyr' was built under R version 4.1.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

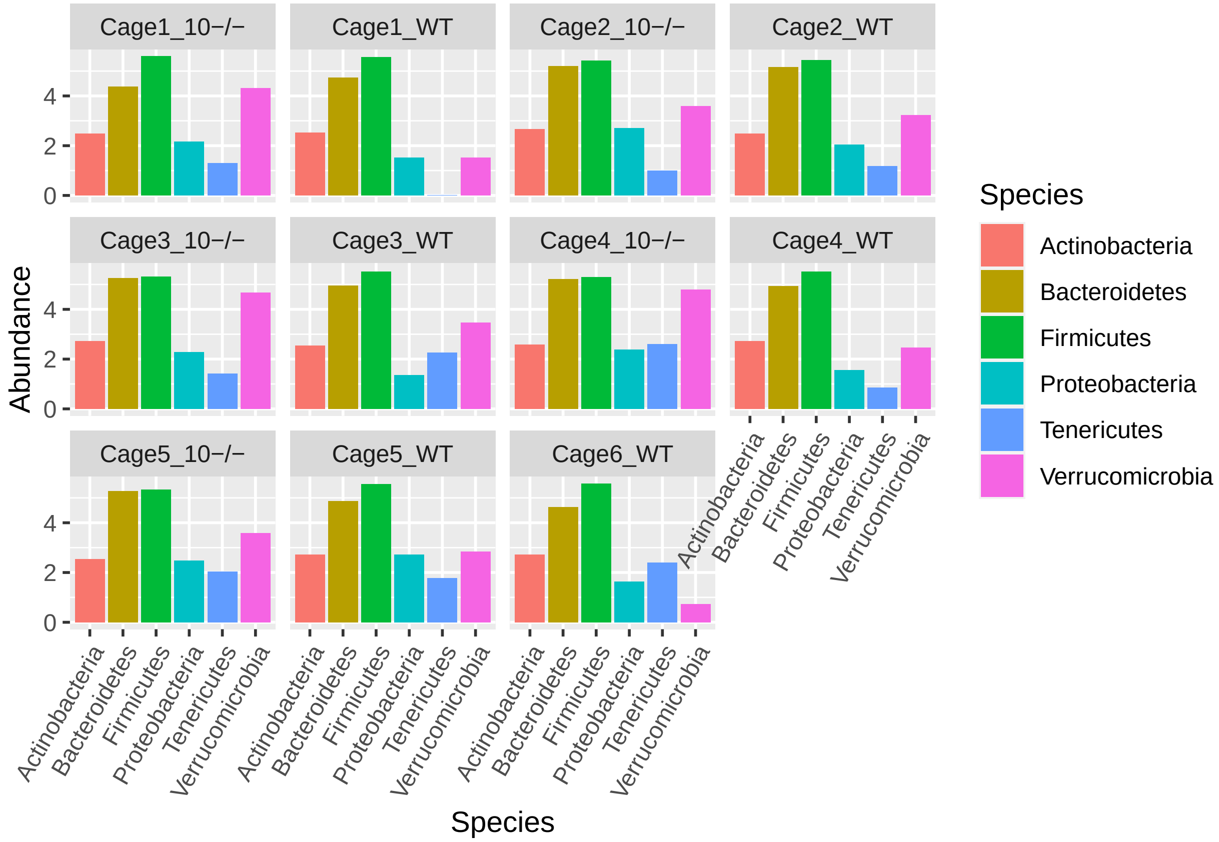
```
#filter data by category
pre <- pca_df %>% filter(pca_df$Time=="PRE")
post <- pca_df %>% filter(pca_df$Time=="POST")
WT <- pca_df %>% filter(pca_df$Genotype=="WT")
knockout <- pca_df %>% filter(pca_df$Genotype=="10-/-")

#perform t-tests and anova
time_pc1 <- t.test(pre$x, post$x)$p.value
time_pc2 <- t.test(pre$y, post$y)$p.value

genotype_pc1 <- t.test(WT$x, knockout$x)$p.value
genotype_pc2 <- t.test(WT$y, knockout$y)$p.value

cage_pc1 <- anova(lm(pca_df$x ~ pca_df$Cage))$'Pr(>F)'[1]
cage_pc2 <- anova(lm(pca_df$y ~ pca_df$Cage))$'Pr(>F)'[1]

pc_df <- data.frame(c(time_pc1, genotype_pc1, cage_pc1), c(time_pc2, genotype_pc2, cage_pc2))
names(pc_df) <- c("PC1", "PC2")
row.names(pc_df) <- c("Time", "Genotype", "Cage")

#library(kableExtra)
pc_df
```

```
##                   PC1          PC2
## Time     2.519974e-29 4.268188e-01
## Genotype 9.297010e-01 1.274344e-10
## Cage     9.920581e-01 1.629589e-07
```

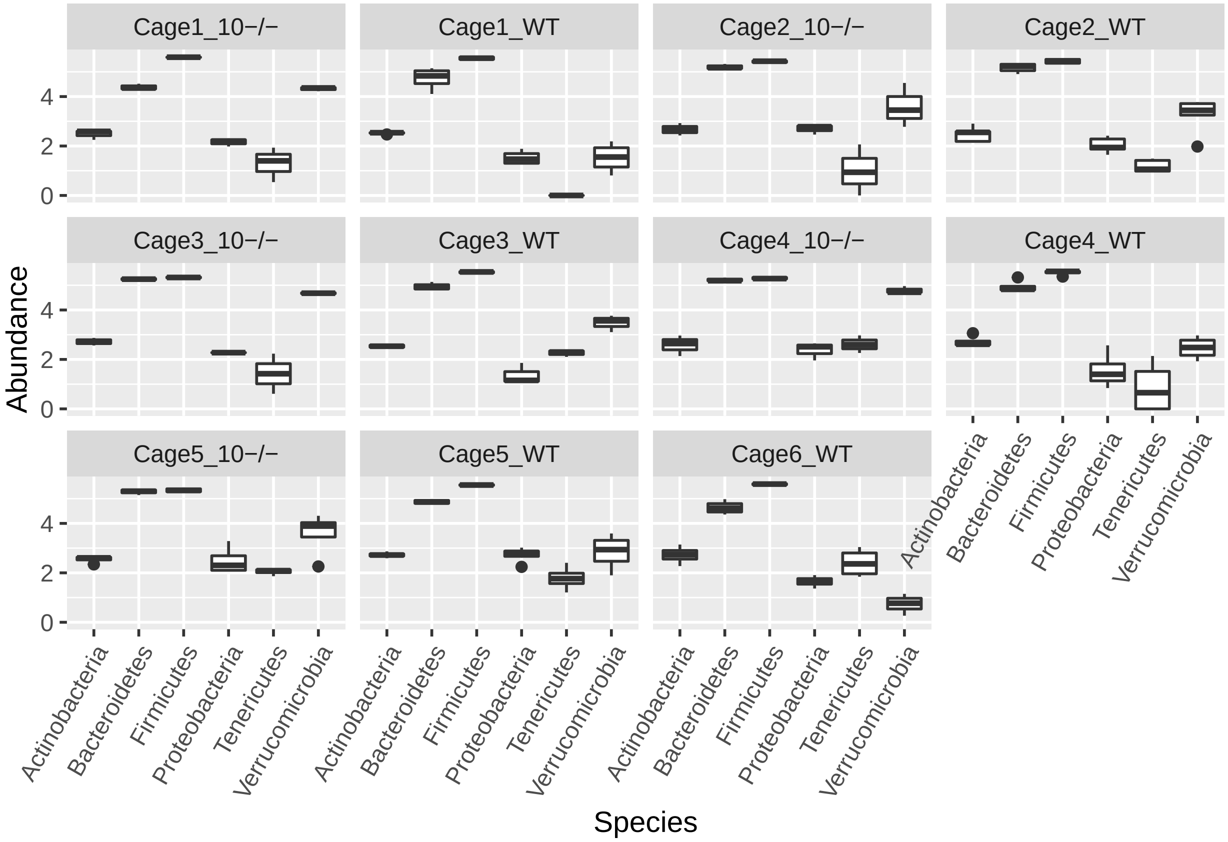
Time has the largest affect on PC1, Genotype has the largest affect on PC. Cage does not seem to have a significant affect

(4)For the POST timepoints only: A. For each phyla, graph the relative abundance of that phyla vs. cage. Does there appear to be a cage effect across different phyla?

```
post_cage_avg <- data.frame(data$cage, data$genotype, data$time, data_spec) %>% filter(data.time=="POST

flatten <- pivot_longer(post_cage_avg, cols=c(Tenericutes, Verrucomicrobia, Bacteroidetes, Actinobacteri

ggplot(flatten, aes(x=Species, y=value, group=data.cage, fill=Species)) + geom_bar(stat='identity') + fa
```



```
#--------------------------

post_cage <- data.frame(data$cage, data$genotype, data$time, data_spec) %>% filter(data.time=="POST") %

flatten <- pivot_longer(post_cage, cols=c(Tenericutes, Verrucomicrobia, Bacteroidetes, Actinobacteria, 

ggplot(flatten, aes(x=Species, y=value, group=Species)) + geom_boxplot() + facet_wrap(~data.cage) + yla
```



From visual inspection it is difficult to tell if there is a significant cage effect. The species' abundances follow the same general distribution… however, we can see that Bacteriodetes, Firmicutes, and Verrucomicrobia have the highest variance among cages. Also, Tenericutes is absent in Cage_1_WT.

B. For each phyla build a mixed linear model with genotype as the fixed variable and cage as a random variable. Report the intraclass correlation coefficient for each phyla. Are there any phyla that are significantly different for genotype in the mixed model at a 10% false discovery rate?

```
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 4.1.2
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.1.2
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(lmerTest)
```

```
##
## Attaching package: ’lmerTest’

## The following object is masked from ’package:lme4’:
##
##     lmer

## The following object is masked from ’package:stats’:
##
##     step
```

```
post_cage <- data.frame(post_cage)

actinobacteria <- lmer(data=post_cage, Actinobacteria ~ data.genotype + (1 | data.cage) )
```

```
## boundary (singular) fit: see help(’isSingular’)
```

```
bacteroidetes <- lmer(data=post_cage, Bacteroidetes ~ data.genotype + (1 | data.cage) )
firmicutes <- lmer(data=post_cage, Firmicutes ~ data.genotype + (1 | data.cage) )
proteobacteria <- lmer(data=post_cage, Proteobacteria ~ data.genotype + (1 | data.cage) )
tenericutes <- lmer(data=post_cage, Tenericutes ~ data.genotype + (1 | data.cage) )
verrucomicrobia <- lmer(data=post_cage, Verrucomicrobia ~ data.genotype + (1 | data.cage))

me_models <- c(actinobacteria, bacteroidetes, firmicutes, proteobacteria, tenericutes, verrucomicrobia)

me_names <- c('actinobacteria', 'bacteroidetes', 'firmicutes', 'proteobacteria', 'tenericutes', 'verruc
me_pvals <- vector(length=length(me_models))
me_corr <- vector(length=length(me_models), mode='numeric')

#  -- looping over the models was problematic and i was losing my mind
#
# for ( i in 1:length(me_models) ) {
#
#   me_pvals[i] <- as.data.frame(summary(me_models[i], ddf="Kenward-Roger")$coefficients)$`Pr(>|t|)`[2]
#
#   #extracting correlation data
#   cor_df <- as.data.frame(VarCorr(me_models[i]))
#   psi <- cor_df$sd_cor[1]
#   sigma <- cor_df$sd_cor[2]
#
#   me_corr[i] <- (psi^2 / (psi^s + sigma^2))
#
# }

#extract p-values of the slopes of the me-models -- df estimated by Kenward-Rogers method
me_pvals[1] <- as.data.frame(summary(actinobacteria, ddf="Kenward-Roger")$coefficients)$`Pr(>|t|)`[2]
me_pvals[2] <- as.data.frame(summary(bacteroidetes, ddf="Kenward-Roger")$coefficients)$`Pr(>|t|)`[2]
me_pvals[3] <- as.data.frame(summary(firmicutes, ddf="Kenward-Roger")$coefficients)$`Pr(>|t|)`[2]
me_pvals[4] <- as.data.frame(summary(proteobacteria, ddf="Kenward-Roger")$coefficients)$`Pr(>|t|)`[2]
```

```r
me_pvals[5] <- as.data.frame(summary(tenericutes, ddf="Kenward-Roger")$coefficients)$`Pr(>|t|)`[2]
me_pvals[6] <- as.data.frame(summary(verrucomicrobia, ddf="Kenward-Roger")$coefficients)$`Pr(>|t|)`[2]

#calculate the intraclass correlation
cor_df <- as.data.frame(VarCorr(actinobacteria))
psi <- cor_df$sdcor[1]
sigma <- cor_df$sdcor[2]
me_corr[1] <- (psi^2 / (psi^2 + sigma^2))

cor_df <- as.data.frame(VarCorr(bacteroidetes))
psi <- cor_df$sdcor[1]
sigma <- cor_df$sdcor[2]
me_corr[2] <- (psi^2 / (psi^2 + sigma^2))

cor_df <- as.data.frame(VarCorr(firmicutes))
psi <- cor_df$sdcor[1]
sigma <- cor_df$sdcor[2]
me_corr[3] <- (psi^2 / (psi^2 + sigma^2))

cor_df <- as.data.frame(VarCorr(proteobacteria))
psi <- cor_df$sdcor[1]
sigma <- cor_df$sdcor[2]
me_corr[4] <- (psi^2 / (psi^2 + sigma^2))

cor_df <- as.data.frame(VarCorr(tenericutes))
psi <- cor_df$sdcor[1]
sigma <- cor_df$sdcor[2]
me_corr[5] <- (psi^2 / (psi^2 + sigma^2))

cor_df <- as.data.frame(VarCorr(verrucomicrobia))
psi <- cor_df$sdcor[1]
sigma <- cor_df$sdcor[2]
me_corr[6] <- (psi^2 / (psi^2 + sigma^2))

sig_df <- data.frame(me_names, me_pvals, me_corr)
```

```r
adj_pvals <- p.adjust(as.vector(sig_df$me_pvals), method="BH")

print(adj_pvals)
```

```
## [1] 0.74962874 0.48353969 0.08312252 0.08312252 0.71240958 0.04970349
```

There are three species that are significantly different for genotype at 10% FDR according to the mixed-effect models 1) firmicutes, proteobacteria, and verrucomicrobia