RNA Quantification

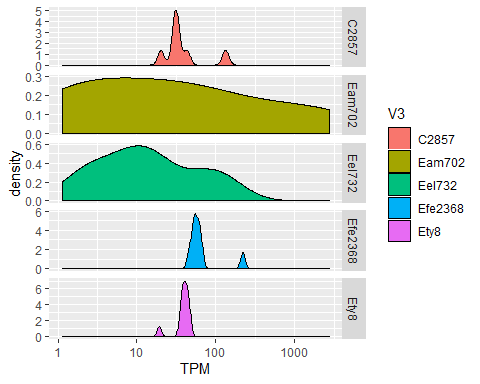
Imogen Dumville

21/02/2020

Document displaying statistical analyses carried out on expression data from RNAseq analysed strains. These icluded amarillans702, elymi, festucae strains 2368 and Fl1 and typhina E8.

Function produced to iterate over quant.sf files in single directory (called quantfiles). geom\_density produced for each strain TPM variability. Note that amarillans and elymi distributions are not as expected. May be cause for concern for later results - TBC

subsetCP <- read.table("subsetCP.tsv", quote="\"", comment.char="", stringsAsFactors=FALSE)  
  
files <- list.files(path="./RNA/quantfiles/", pattern="\*.sf", full.names=TRUE, recursive=FALSE)  
  
process\_one\_file <- function(fname){  
 quant <- read.delim(fname, stringsAsFactors = FALSE)  
 quant <- quant %>% arrange(desc(TPM))  
 leng <- nrow(quant)  
 rows <- as.integer(attr(quant, "row.names"))  
 quant$rank <- rows/leng  
 quantsubsetCP <- subsetCP[c(which(subsetCP$V2 %in% quant$Name)),]  
 merged <- merge(quant, quantsubsetCP, by.x = "Name", by.y = "V2")  
 return(merged)  
}  
  
res <- lapply(files, process\_one\_file)  
actual\_res <- bind\_rows(res)  
ggplot(actual\_res, aes(TPM, fill=V3)) + geom\_density() + scale\_x\_log10() + facet\_grid(V3 ~ ., scales="free")



GG facet bar plot produced for each ortholog associated with cyclic peptides, where ‘proportion’ is rank of TPM in that strain. Plot produced for each ortholog presence and rank in each strain. Note variation in expression presence and relative amount. Caution to be made when comparing ‘rank’ of expression amongst strains.

names(actual\_res)[6] <- "Proportion"  
  
plot <- ggplot(actual\_res, aes(V1, Proportion, fill=V3))  
plot <- plot + geom\_bar(stat = "identity", position = position\_dodge2(width = 0.9, preserve = "single"))  
plot + facet\_grid(~V1, scales = "free\_x", space = "free\_x", switch = "x") +   
 theme(axis.text.x = element\_blank(),  
 axis.ticks.x = element\_blank(),  
 strip.background = element\_blank())

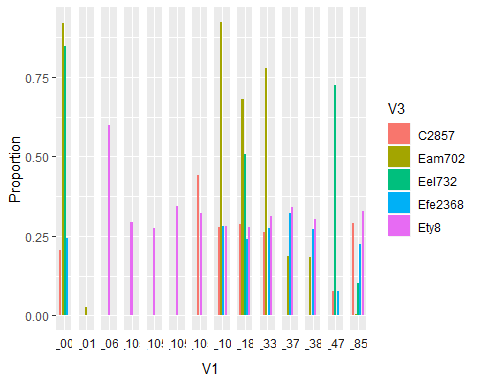


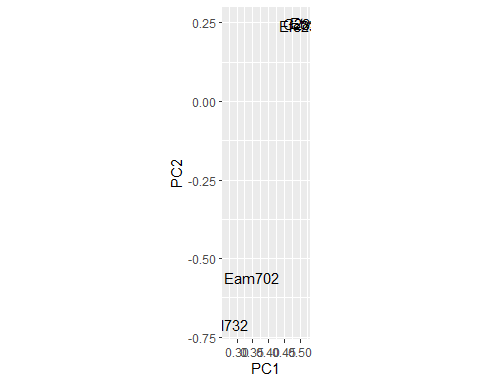
Table is produced of each ortholog associated with strain name and associated TPM and rank.

ortho\_long <- read.delim("~/CoxExtension/ortho\_long.tsv", header=FALSE, stringsAsFactors=FALSE)  
  
process\_one\_file\_all\_og <- function(fname){  
 quant <- read.delim(fname, stringsAsFactors = FALSE)  
 quant <- quant %>% arrange(desc(TPM))  
 leng <- nrow(quant)  
 rows <- as.integer(attr(quant, "row.names"))  
 quant$rank <- rows/leng  
 quantsubsetall <- ortho\_long[c(which(ortho\_long$V2 %in% quant$Name)),]  
 merged <- merge(quant, quantsubsetall, by.x = "Name", by.y = "V2")  
 return(merged)  
}  
  
allres <- lapply(files, process\_one\_file\_all\_og)  
all\_actual\_res <- bind\_rows(allres)  
  
all\_actual\_res <- all\_actual\_res[,-c(1,2,3,5)]  
  
mydata <- melt(all\_actual\_res, TPM=c("TPM","rank"))  
write.table(mydata, "TMPRankallorthologs.tsv", quote=FALSE, sep='\t', row.names = FALSE)  
head(mydata)

## V1 V3 variable value  
## 1 og\_6853 Eam702 TPM 0.519836  
## 2 og\_0644 Eam702 TPM 62.836155  
## 3 og\_0157 Eam702 TPM 468.333379  
## 4 og\_7959 Eam702 TPM 34.479049  
## 5 og\_7542 Eam702 TPM 88.851458  
## 6 og\_3582 Eam702 TPM 75.403795

PCA plots were produced with input of all orthologs. Note amarillans and elymi are varied on the first two Principle Components. This needs further investigation due to earlier TPM variability.

TPM <- subset(mydata, variable == "TPM")  
TPM\_wider <- pivot\_wider(TPM, names\_from="V3", values\_from="value", values\_fn = list(value = mean))  
  
  
TPM\_wider <- TPM\_wider[complete.cases(TPM\_wider),]  
row.names(TPM\_wider) <- TPM\_wider$V1  
  
PCA = prcomp(TPM\_wider[3:7], scale = TRUE, center = TRUE)  
PCA\_df = as.data.frame(PCA$rotation)  
PCA\_df$strain <- row.names(PCA\_df)  
ggplot(PCA\_df, aes(PC1, PC2, label = strain)) + geom\_text() + coord\_equal()



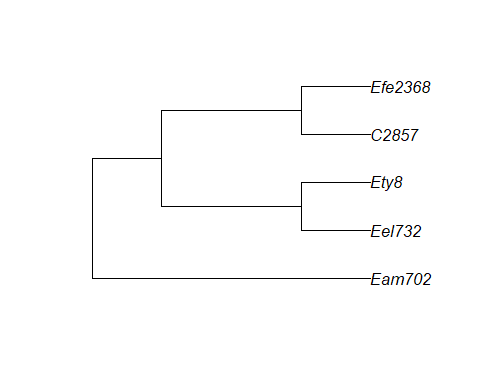
Distance between PCA1 and PCA2 for each ortholog was counted and aids determination of strength of impact on strain differences

pythagorean <- function(a, b){  
 hypotenuse <- sqrt(a^2 + b^2)  
 return(hypotenuse)  
}  
  
PCA\_values <- PCA$x  
row.names(PCA\_values) <- TPM\_wider$V1  
  
PCA\_values <- PCA\_values[,c(1,2)]  
hypo <- pythagorean(PCA\_values[,1], PCA\_values[,2])  
hypo <- as.data.frame(hypo)  
distance <- hypo$hypo  
PCA\_values <- as.data.frame(PCA\_values)  
PCA\_values$distance <- distance  
  
morethan10 <- which(PCA\_values$distance >= 10)  
subsetPCA\_values <- PCA\_values[c(morethan10),] #56  
  
dist\_orthologs <- row.names(subsetPCA\_values)  
  
distrows <- lapply(dist\_orthologs, function(x) which(ortho\_long$V1 == (x)))  
rownums <- unlist(distrows)  
subsetorthos <- ortho\_long[(rownums),]  
write.table(subsetorthos, file = "distancegenesmorethan10.tsv", row.names = FALSE, col.names = FALSE, quote = FALSE)  
head(subsetorthos)

## V1 V2 V3  
## 150 og\_1555 CCE31242 CCE27021  
## 9781 og\_1555 EamaE4668\_000594-T1 EamaE4668  
## 17449 og\_1555 EamaE57\_000771-T1 EamaE57  
## 25202 og\_1555 Ebac200745\_001417-T1 Ebac200745  
## 32971 og\_1555 EbraE4804\_005838-T1 EbraE4804  
## 40991 og\_1555 EbroAL0426\_000826-T1 EbroAL0426

Using strain tree produced in alternative code(EpiAllInclCpur.R) and allstrainframe (RNAquant.R) Strain tree is produced with only strains of interest

strain\_tree <- read.tree("~/Summer Scholarship 2019/EpichloeAll/FinalDocuments/straintree")  
CPtree <- keep.tip(strain\_tree, c("C2857", "Eam702", "Eel732", "Efe2368", "Ety8"))  
x <- plot(CPtree)

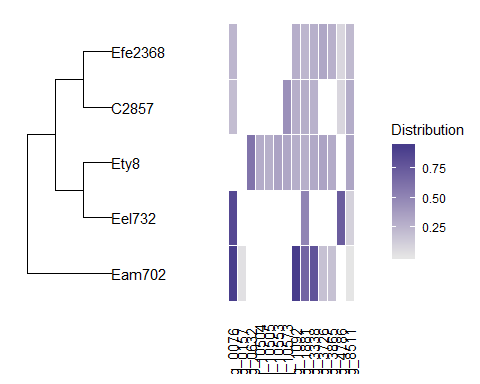


Heatmaps are produced of various compenents of expression. Rank heatmap (below) display high variation amongst strains.

ggstraintree <- ggtree(CPtree)

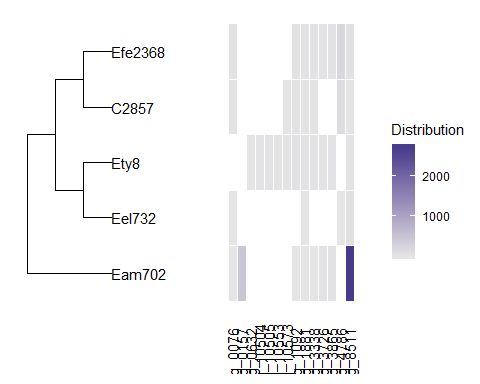
## Warning in fortify.phylo(data, ...): 'edge.length' contains NA values...  
## ## setting 'edge.length' to NULL automatically when plotting the tree...

joined <- ggstraintree %<+% actual\_res  
y <- joined +  
 geom\_tiplab(aes()) +  
 theme(legend.position = "right") +   
 scale\_color\_viridis()  
  
  
rankres <- actual\_res  
  
framespread <- spread(rankres, V1, Proportion)  
framespread <- framespread[,6:20]  
  
  
compress <- function(x) c(na.omit(x), NA)[1]  
compressed <- aggregate(framespread[2:15], framespread[1], compress)  
rownames(compressed) <- compressed$V3  
compressed$V3 <- NULL  
  
gheatmap(y, compressed, offset = 4, colnames\_angle = 90,low = "grey90", high = "slateblue4", colnames\_offset\_y = -0.5, width = 1.5) + theme(legend.title = element\_text()) + labs(fill = "Distribution")



TPM heatmaps are unclear due to large scale variation. Note again difference in amarillans are likely to be artifacts.

TPMres <- actual\_res  
  
framespread <- spread(TPMres, V1, TPM)  
framespread <- framespread[,6:20]  
compress <- function(x) c(na.omit(x), NA)[1]  
compressed <- aggregate(framespread[2:15], framespread[1], compress)  
rownames(compressed) <- compressed$V3  
compressed$V3 <- NULL  
gheatmap(y, compressed, offset = 4, colnames\_angle = 90,low = "grey90", high = "slateblue4", colnames\_offset\_y = -0.5, width = 1.5) + theme(legend.title = element\_text()) + labs(fill = "Distribution")



log(TPM) was calculated to scale differences. Again, limited variabilty amongst strains excluding elymi and amarillans is shown.

logTPMres <- actual\_res  
  
logTPMres$logTPM <- log(logTPMres$TPM)  
framespread <- spread(logTPMres, V1, logTPM)  
framespread <- framespread[,7:21]  
compress <- function(x) c(na.omit(x), NA)[1]  
compressed <- aggregate(framespread[2:15], framespread[1], compress)  
rownames(compressed) <- compressed$V3  
compressed$V3 <- NULL  
gheatmap(y, compressed, offset = 4, colnames\_angle = 90,low = "grey90", high = "slateblue4", colnames\_offset\_y = -0.5, width = 1.5) + theme(legend.title = element\_text()) + labs(fill = "Distribution")

