Assignment 4.3

Mahdi

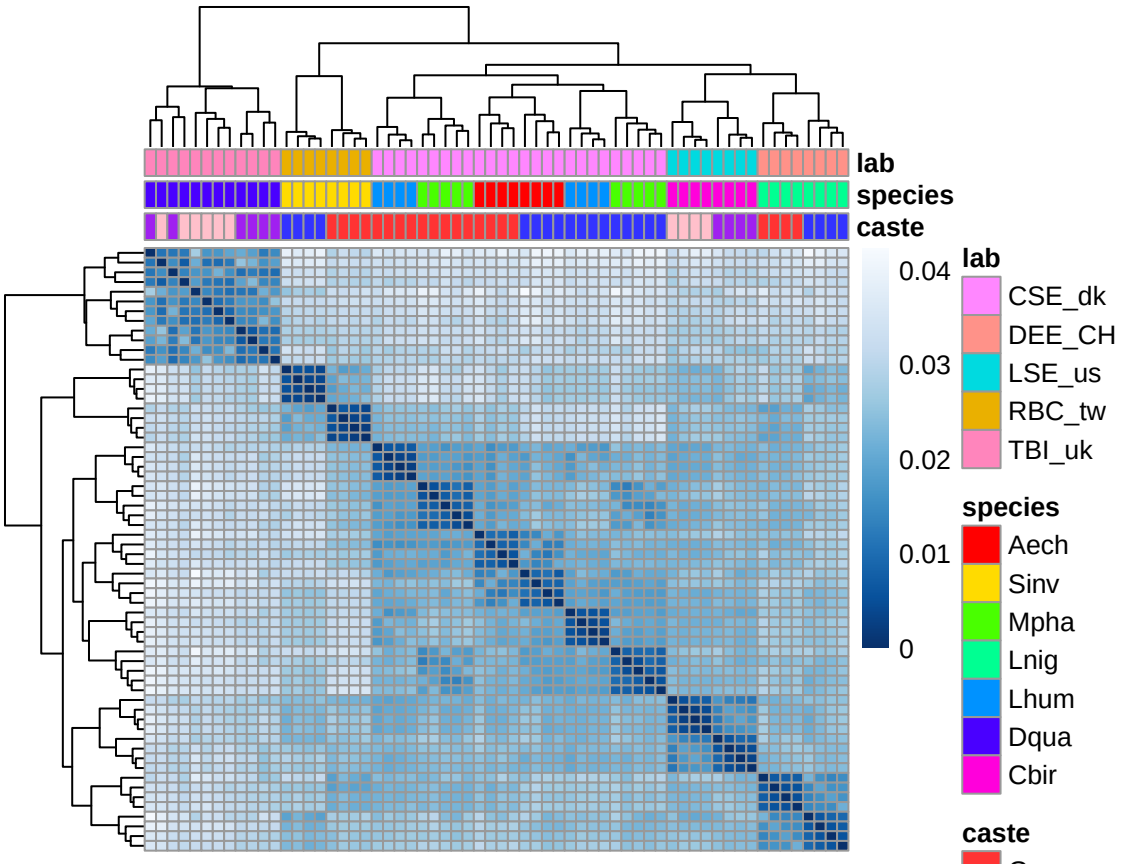
11/3/2020

We performed batch/species/colony effect correction for the five ant species (A.ech, M.pha, S.inv, L.hum, and L.nig). Try to include also the two queenless ant species in colony effect correction, and report the similarity matrix (Heatmap+Hclustering that we used in the classroom) and PCA result. (Code and Plots, and a briefly discussion about what you see in the plot) (15 points)

```
# Learn how to correct for batch effect.
# Learn how to identify differentially expressed genes across batch
load('inputData.Rdata')
source("shared_functions.R")

# Section 2: Using Combat to normalize batch effect.

sampleTable$caste[which(sampleTable$caste == 'Minor_worker')] = 'Worker'  # For simplity, we treat mino
normal_ant = which(sampleTable$species %in% c("Aech",'Mpha',"Lhum",'Sinv',"Lnig", "Dqua","Cbir"))
ortholog_exp.ant = ortholog_counts[,normal_ant]
sampleTable.ant = droplevels(sampleTable[normal_ant,])
ortholog_exp.ant = ortholog_exp.ant[!apply(ortholog_exp.ant, 1, anyNA),]  #Removed genes showing NA (e.
ortholog_exp.ant.norm = log2(normalize.quantiles(ortholog_exp.ant)+1)
colnames(ortholog_exp.ant.norm) = colnames(ortholog_exp.ant)
rownames(ortholog_exp.ant.norm) = rownames(ortholog_exp.ant)
ortholog_exp.ant.norm = ortholog_exp.ant.norm[apply(ortholog_exp.ant.norm, 1,
                                                    FUN = function(x) return(var(x, na.rm = T) > 0)),]

batch = droplevels(sampleTable.ant$colony) # Normalization for species identity.
modcombat = model.matrix(~1, data = sampleTable.ant)
combat.ortholog_exp.ant = ComBat(dat=ortholog_exp.ant.norm, batch=batch, mod=modcombat,
                             mean.only = F, par.prior=TRUE, prior.plots=FALSE)
```

```
## Found 318 genes with uniform expression within a single batch (all zeros); these will not be adjuste
```

```
sampleDists.combat = as.dist(1 - cor(combat.ortholog_exp.ant,method = 's'))
pheatmap(sampleDists.combat,annotation_col = sampleTable.ant[,c(1:3)],
         annotation_colors = ann_colors,
         color = colors)
```

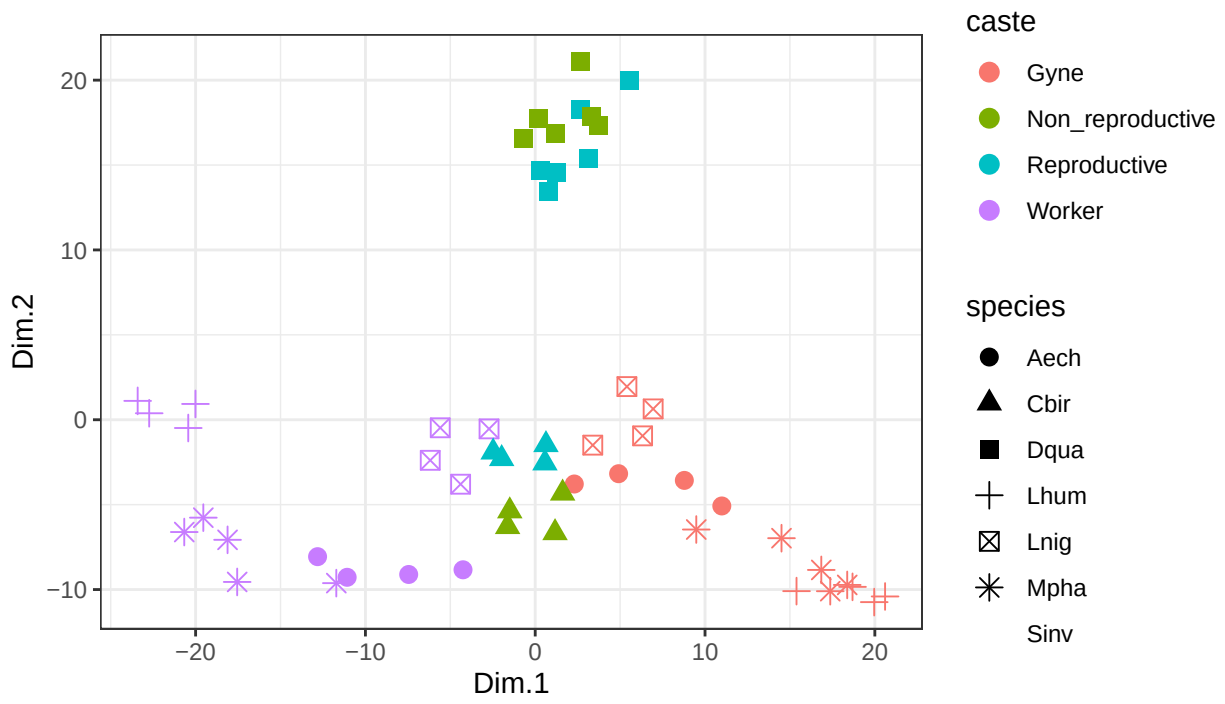


In the heatmap we can see that the ants are mainly clustered by species and somewhat clustered by lab. There is also some clusterihng by caste. This means even after correcting for species, colony and lab, some effect of these factors still remain.

```
var.gene = order(apply(combat.ortholog_exp.ant,1,var),decreasing = T)[c(1:1000)]
ortholog_exp.combat.pca <- PCA(t(combat.ortholog_exp.ant[var.gene,]),ncp = 4, graph = FALSE)

# Take a look at the amount of variations explained by each PC.
#fviz_eig(ortholog_exp.combat.pca, addlabels = TRUE,main = 'Explained variance for each PC')

pca.combat.var = ortholog_exp.combat.pca$eig
pca.combat.data = cbind(ortholog_exp.combat.pca$ind$coord,sampleTable.ant)
ggplot(pca.combat.data, aes(x = Dim.1, y = Dim.2, color = caste, shape = species)) +
  geom_point(size=3) +
  coord_fixed()+
  theme_bw()
```

```
## Warning: The shape palette can deal with a maximum of 6 discrete values because
## more than 6 becomes difficult to discriminate; you have 7. Consider
## specifying shapes manually if you must have them.
```

```
## Warning: Removed 8 rows containing missing values (geom_point).
```

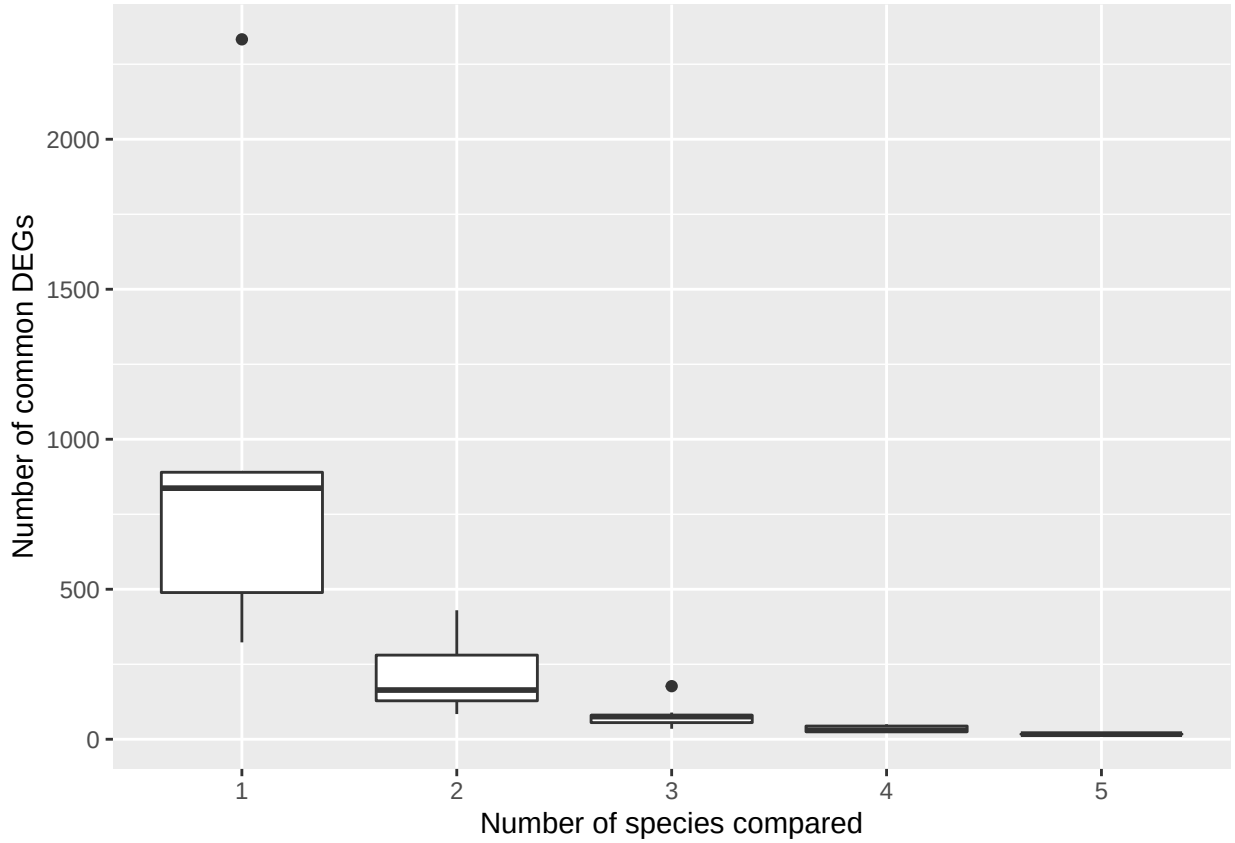


The first pincipal component of the PCA plot seperates according to caste, and it can be seen that both castes of the Cbir ants are found between the gyne caste and the worker caste. The second principal component separates according to species and shows that the Dqua species (queenless) is quite different from all the other ant species including Cbir.

In the class, we tested the number of differentially expressed genes between gyne and worker (castes) in one of the species using DESeq2. Can you report the number of overlapping DEGs in two, three, four, and all five typical ant species, i.e. Aech Mpha Lhum Lnig and Sinv, similar to the last slide of the class lecture? And how many DEGs have you found if you use the model: Exp ~ Caste + Species + Caste:Species ? Is this number different from the number of overlapping DEGs? Why? (Number of DEGs for the two situations, and a brief discussion about the difference) (15 points)

```
# Section 4: Identification of caste differentially expressed genes
normal_ant = which(sampleTable$species %in% c("Aech",'Mpha',"Lhum",'Sinv',"Lnig"))
ortholog_exp.ant = ortholog_counts[,normal_ant]

ortholog_counts.ant = ortholog_counts[,normal_ant]
ortholog_counts.ant = ortholog_counts.ant[!apply(ortholog_counts.ant, 1, anyNA),]
ortholog_counts.ant.norm = matrix(as.integer(ortholog_counts.ant), ncol = dim(ortholog_counts.ant)[2],
                                  dimnames = list(rownames(ortholog_counts.ant),colnames(ortholog_counts

get_dif_exp_genes <- function(target_species, model=1){
  if(model == 1){
    dds <- DESeqDataSetFromMatrix(
      ortholog_counts.ant.norm[,which(sampleTable.ant$species %in% target_species)], sampleTable.ant[wh
```

```
      design = ~caste)
  } else{
    dds <- DESeqDataSetFromMatrix(
      ortholog_counts.ant.norm[,which(sampleTable.ant$species %in% target_species)], sampleTable.ant[wh
      design = ~caste +species+caste:species)
  }
  dds = DESeq(dds)
  res.aech = results(dds, contrast = c("caste",c("Gyne",'Worker')),alpha = 0.05)
  return(res.aech)
}

ants_names <- c("Mpha", "Aech", "Lhum", "Lnig", "Sinv")
dseq_results <- list()
filtered_genes <- list()

for(name in ants_names){
  dseq_results[[name]] <- get_dif_exp_genes(name)
    filtered_genes[[name]] <- dseq_results[[name]] %>%
      as.data.frame %>%
      rownames_to_column("Gene") %>%
      filter(padj < 0.05) %>%
      select(Gene)
}

combinations <- list()
gene_count <- list()

for(i in 1:5){
  combinations[[i]] <- combn(ants_names, i) # list of species names combinations
  columns <- dim(combinations[[i]])[2]
  for(c in 1:columns){
    species <- combinations[[i]][,c] #vector of species names
    group <- paste0(species, collapse = "-")
    gene_list <- list() #list of vectors of gene names
    for(s in species){
      gene_list[[s]] <- filtered_genes[[s]]$Gene
    }
    gene_count[[group]] <- length(Reduce(intersect, gene_list))
  }
}

model_1 <- gene_count %>% as.data.frame() %>% t %>% as.data.frame()
model_1$V2 <- as.factor(str_count(rownames(model_1), "[.]") + 1)

model_1 %>%
  ggplot(aes(x=V2, y=V1)) + geom_boxplot() +
  labs(x="Number of species compared", y="Number of common DEGs")
```




```
model_2 <- get_dif_exp_genes(c("Mpha", "Aech", "Lhum", "Lnig", "Sinv"), 2)
m2_dim <- model_2 %>% as.data.frame() %>% filter(padj < 0.05) %>% dim

print(paste("DEGs in Model 1:", model_1 %>% filter(V2 == 5) %>% select(V1)))
```

```
## [1] "DEGs in Model 1: 17"
```

```
print(paste("DEGs in Model 2:", m2_dim[1]))
```

```
## [1] "DEGs in Model 2: 434"
```

The first model results in only 17 genes that overlap for all 5 species, but the second model shows 434 genes overlap. Even after trying to correct for the effect of species on the gene transcription, some effect still remains. This means different genes will be expressed for different species, so some genes may be differentially expresssed in some species while not in others. These genes will be missed by the first model but captured by the second model since the second model takes the effect of species into account.

Besides of using distance measuring method, we can also use factor analysis approach to identify gene modules. Try to extract the weight of genes from PC1 and PC2 of the five ant species (normalized for species identify), and apply that to transform the gene expression data on queenless ant species (Hint: Take a look at http://www.iro.umontreal.ca/~pift6080/H09/documents/papers/pca_tutorial.pdf (Links to an external site.)). [Note: This is similar to train a gene network from the training data set (the five ant species) and test it on the target data set (the queenless ant).] (Code and Plot with the PC data of both the five typical ants and the reconstructed PC data of the two queenless ants, and a brief discussion about the result.) (3.4 points)