DGE_Analysis

2023-04-06

1. Load packages

```
library(tidyverse)
library(DESeq2)
library(EnhancedVolcano)
library(gridExtra)
library(grid)
```

2. Data transformation

2.a. Import data sets The data sets are provided by Dr. Rebecca L. Young.

```
frog_raw_counts_no_outlier <- read.table(file = "/stor/work/Bio321G_RY_Spring2023/StudentDirectories/Ky
                                         row.names = 1,
                                         header = TRUE,
                                         sep = ",")

frog_metadata <- read.delim(file = "/stor/work/Bio321G_RY_Spring2023/StudentDirectories/KyNguyen/MiniPr
```

2.b. DeSeq2 One of the main objective for the analysis of count data from RNA-seq is to identify differentially expressed genes. The count data are presented as a table which reports, for each sample, the number of sequence fragments that have been assigned to each gene. One of the benefit of the DESeq2 package is that it can take in a raw count table (i.e., no need for normalizing gene count)

```
# DESeq2 requires counts to be a matrix not a data.frame
dds_species <- DESeqDataSetFromMatrix(countData = frog_raw_counts_no_outlier,
                                      colData = frog_metadata,
                                      design = ~condition)

dds_species <- DESeq(dds_species)
```

2.c. Pairings We can compare any two species using the contrast. `condition` is the name of the factor we are comparing. We want to compare every combination for the 4 species - 6 total pairings. The species listed first becomes the "numerator", the one listed second is the "denominator" (i.e., if the log2FoldChange > 0, the ratio is > 1, and vice versa)

```
# E_anthonyi and E_tricolor are aposematic, E_boulengeri and E_machalilla are cryptic

# Same skin coloration pairings
E_anthonyi_vs_E_tricolor <- results(dds_species,
                                   contrast = c("condition", "E_anthonyi", "E_tricolor"))

E_boulengeri_vs_E_machalilla <- results(dds_species,
                                       contrast = c("condition", "E_boulengeri", "E_machalilla"))


# Different skin coloration pairings
E_boulengeri_vs_E_anthonyi <- results(dds_species,
                                     contrast = c("condition", "E_boulengeri", "E_anthonyi"))

E_boulengeri_vs_E_tricolor <- results(dds_species,
                                     contrast = c("condition", "E_boulengeri", "E_tricolor"))

E_machalilla_vs_E_tricolor <- results(dds_species,
                                     contrast = c("condition", "E_machalilla", "E_tricolor"))

E_machalilla_vs_E_anthonyi <- results(dds_species,
                                     contrast = c("condition", "E_machalilla", "E_anthonyi"))
```

```
E_anthonyi_vs_E_tricolor <- as.data.frame(E_anthonyi_vs_E_tricolor)
E_boulengeri_vs_E_machalilla <- as.data.frame(E_boulengeri_vs_E_machalilla)
E_boulengeri_vs_E_anthonyi <- as.data.frame(E_boulengeri_vs_E_anthonyi)
E_boulengeri_vs_E_tricolor <- as.data.frame(E_boulengeri_vs_E_tricolor)
E_machalilla_vs_E_tricolor <- as.data.frame(E_machalilla_vs_E_tricolor)
E_machalilla_vs_E_anthonyi <- as.data.frame(E_machalilla_vs_E_anthonyi)

# List of data frame
pairings_table <- list(E_anthonyi_vs_E_tricolor, E_boulengeri_vs_E_machalilla,
                       E_boulengeri_vs_E_anthonyi, E_boulengeri_vs_E_tricolor,
                       E_machalilla_vs_E_tricolor, E_machalilla_vs_E_anthonyi)

# Vector of titles used for subsequent plotting
pairings_title <- c("E_anthonyi (Aposematic) versus E_tricolor (Aposematic)",
                    "E_boulengeri (Cryptic) versus E_machalilla (Cryptic)",
                    "E_boulengeri (Cryptic) versus E_anthonyi (Aposematic)",
                    "E_boulengeri (Cryptic) versus E_tricolor (Aposematic)",
                    "E_machalilla (Cryptic) versus E_tricolor (Aposematic)",
                    "E_machalilla (Cryptic) versus E_anthonyi (Aposematic)")
```

2.d. Convert to data frame and store the variables (and title)

3. Visualization - Volcano Plot

3.a. Default p-value cutoff To visualize the results of differential gene expression analyses, we'll start with plotting the volcano plots using the default p-value cutoff of 10e-6 (that is, 10^{-5})

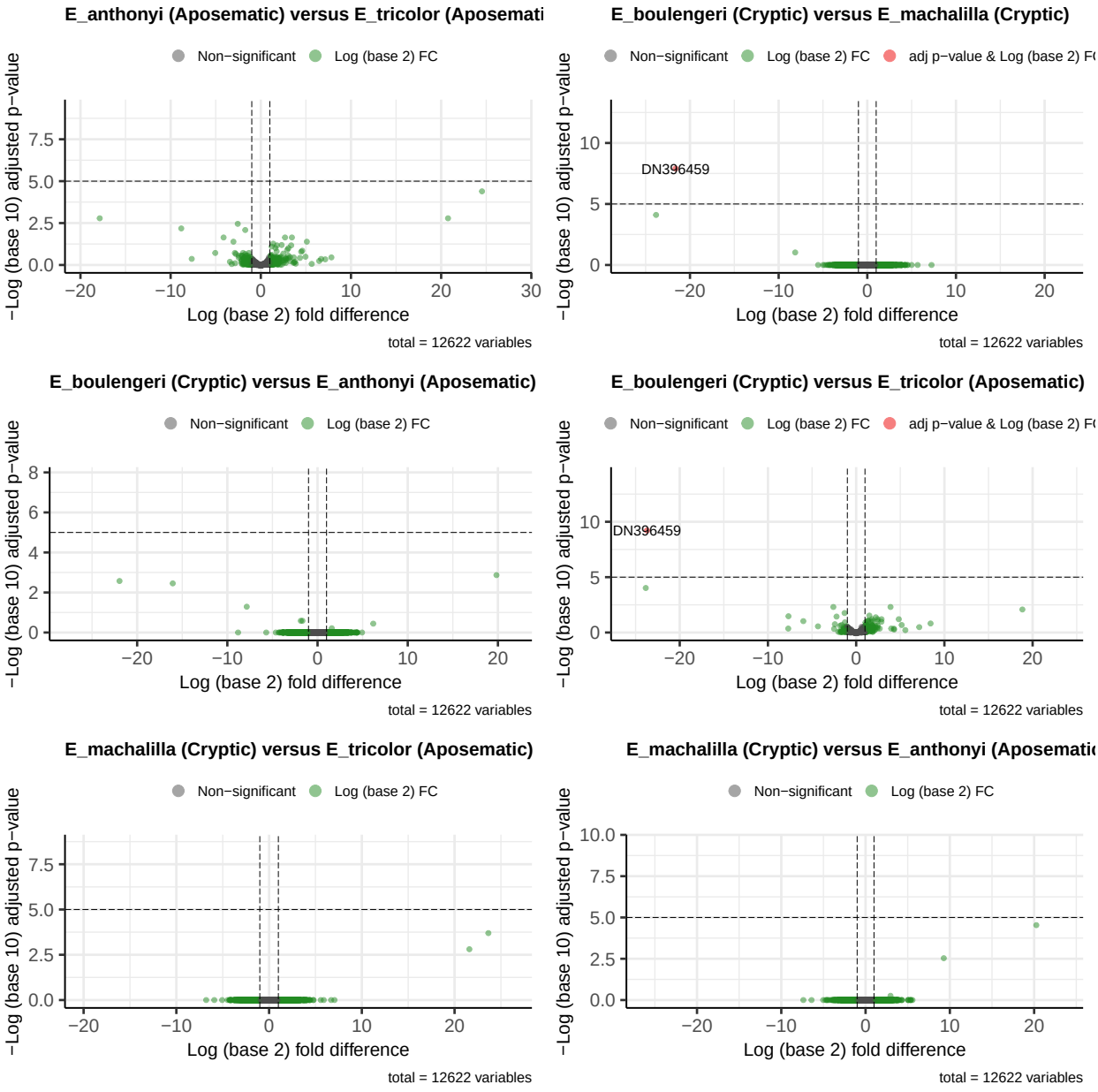
```
# Function to plot Volcano Plot
VolcanoPlot_DefaultCutoff <- function(pairings, title){
  EnhancedVolcano(pairings,
                  lab = rownames(pairings),
                  title = title,
                  subtitle = NULL,
                  legendLabels = c("Non-significant",
                                   "Log (base 2) FC",
                                   "adj p-value",
                                   "adj p-value & Log (base 2) FC"),
                  x = "log2FoldChange",
                  y = "padj",
                  xlab = "Log (base 2) fold difference",
                  ylab = "-Log (base 10) adjusted p-value")
}
```

```
# Display setting: {r, fig.width = 15, fig.height = 15}
# fig.dim doesn't work for some reasons

plot_list = list()

for (i in 1:6){
  plot <- VolcanoPlot_DefaultCutoff(pairings_table[[i]], pairings_title[i])
  plot_list[[i]] <- plot
}

grid.arrange(grobs = plot_list, ncol = 2)
```



3.b. 0.05 p-value cutoff Initially, we're only able to obtain one statistically significant differentiated gene. Therefore, we can lower our p-value cutoff 0.05, which still maintain a high confidence level of 95%.

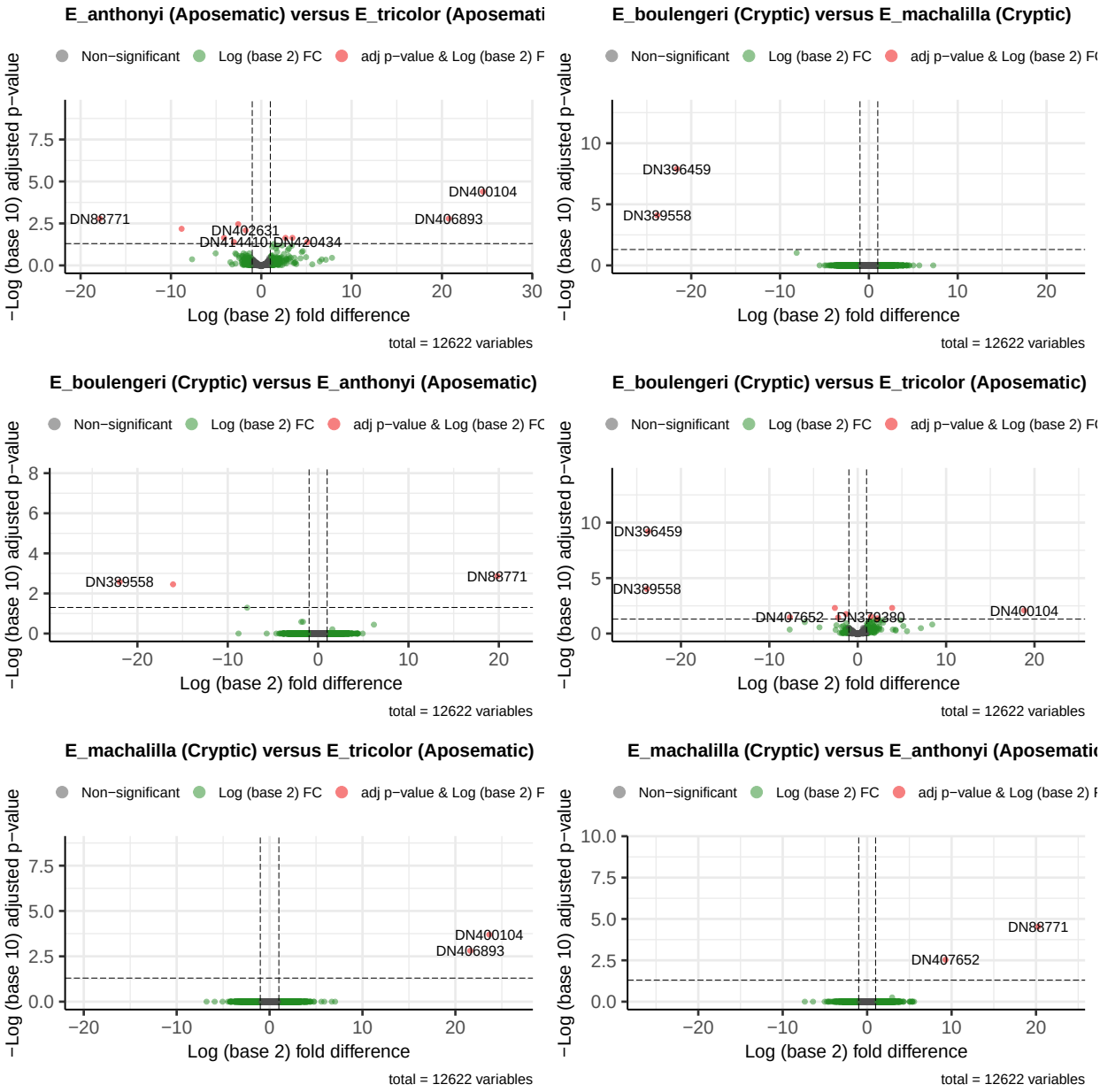
```
VolcanoPlot_0.05Cutoff <- function(pairings, title){
  EnhancedVolcano(pairings,
                  lab = rownames(pairings),
                  title = title,
                  subtitle = NULL,
                  legendLabels = c("Non-significant",
                                   "Log (base 2) FC",
                                   "adj p-value",
```

```
                                    "adj p-value & Log (base 2) FC"),
                  x = "log2FoldChange",
                  y = "padj",
                  pCutoff = 0.05,      # Set p-value cutoff to 0.05
                  xlab = "Log (base 2) fold difference",
                  ylab = "-Log (base 10) adjusted p-value")
}
```

```
plot_list = list()

for (i in 1:6){
  plot <- VolcanoPlot_0.05Cutoff(pairings_table[[i]], pairings_title[i])
  plot_list[[i]] <- plot
}

grid.arrange(grobs = plot_list, ncol = 2)
```



As shown by the volcano plots, the lowered p-value cutoff produces more statistically significant differentiated genes. Moreover, we're able to see some repeating genes, which can be informative going forward with our GO Analysis.

4. Discovering pattern for the differential gene expression analyses.

We want to do some preliminary testing to see if there's any specific pattern for the DGE Analysis using this phylogenetic time tree. Intuitively, we would expect the more closely related the species the less differentially expressed genes they have.

The phylogenetic time tree is constructed manually from information provided by the TimeTree website (Link: http://www.timetree.org/)

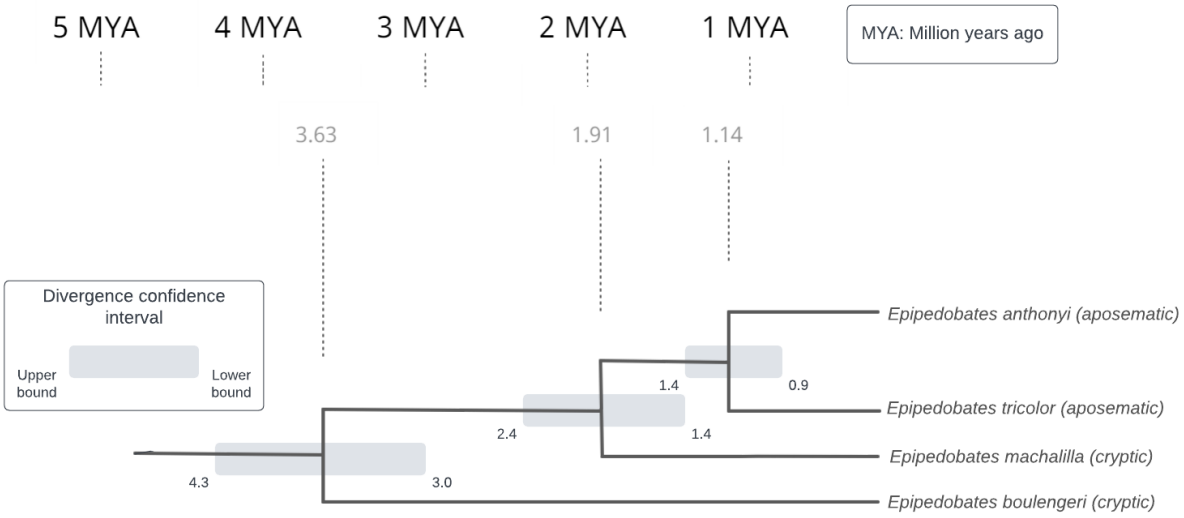


Figure 1: Epipedobates phylogenetic time tree

However, we cannot find any phylogenetic relationship for the DGE Analysis as

1. The two closest species, *E.anthonyi* and *E.tricolor*, have the most differentially expressed genes, while the less closely related *E.boulengeri* and *E.anthonyi* have much fewer.
2. The number of differentially expressed genes doesn't offer any distinction as the pairing of a cryptic and an aposematic (*E.anthonyi vs. E.tricolor*) has 10 differentially expressed genes, while the pairing of two cryptics (*E.anthonyi vs E.tricolor*) has 11 differentially expressed genes, and the remaining pairings has 3, 2, 2, and 2, showing no clear separation in term of the number.
3. Differentially expressed gene is independent of the skin coloration. An example is the gene DN396459, which appeared in all three DGE analysis, two of which are cryptic and aposematic pairings (*E.boulengeri vs. E.anthonyi* and *E.boulengeri vs E.tricolor*) and a pairings of two cryptics (*E.boulengeri vs. E.machalilla*)

Overall, all the analyses does not provide any clear difference between the cryptic and aposematic species using differential gene expression analysis.

5. Male vs. Female comparisons

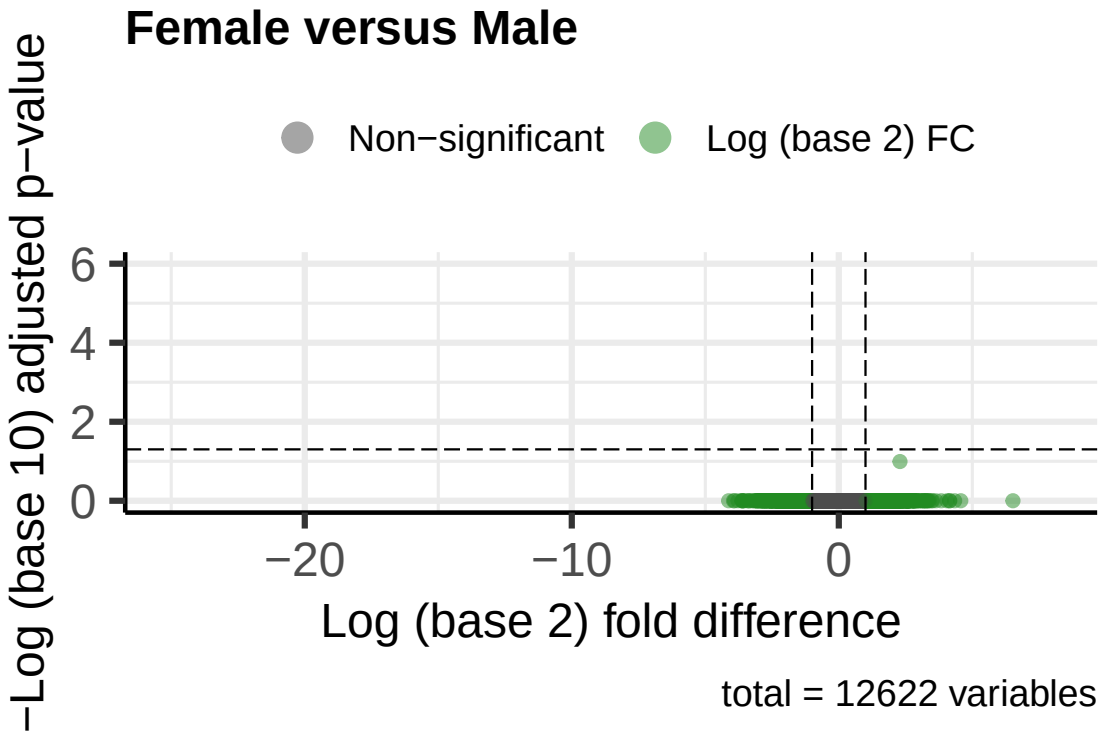
Using our PCA plots, let's discover if there's any differentially expressed genes between the males and females

```
sex <- c("F", "M", "F", "M", "F", "F", "F", "F", "F", "F", "M", "M", "F", "F", "F", "F", "F", "F","M")
frog_metadata <- cbind(frog_metadata, sex)


dds_sex <- DESeqDataSetFromMatrix(countData = frog_raw_counts_no_outlier,
                                  colData = frog_metadata,
                                  design = ~sex)
dds_sex <- DESeq(dds_sex)
```

```
Female_vs_Male <- results(dds_sex,
                          contrast = c("sex", "F", "M"))
Female_vs_Male <- as.data.frame(Female_vs_Male)
```

```
EnhancedVolcano(Female_vs_Male,
    lab = rownames(Female_vs_Male),
    title = 'Female versus Male',
    subtitle = NULL,
    legendLabels = c("Non-significant",
                     "Log (base 2) FC",
                     "adj p-value",
                     "adj p-value & Log (base 2) FC"),
    x = "log2FoldChange",
    y = "padj",
    pCutoff = 0.05,
    xlab = "Log (base 2) fold difference",
    ylab = "-Log (base 10) adjusted p-value")
```



From the volcano plot, we can see that sex has no significance, at least on its own. Therefore, we could perform further analysis with the species' skin coloration with respect to its sex.