Introduction to Bulk RNAseq data analysis

Annotation and Visualisation of Differential Expression Results - Solutions

Abbi Edwards

Last modified: 19 Apr 2021

Exercise 1 - Retrieve the full annotation

So far we have retrieved the annotation for just 1000 genes, but we need annotations for the entire results table.

A reminder of the code we have used so far:

```
# lets set it up
ourCols <- c("GENEID", "SYMBOL", "ENTREZID")
ourKeys <- rownames(results.interaction.11)[1:1000]

# run the query
annot <- AnnotationDbi::select(EnsDb.Mmusculus.v79,
                keys=ourKeys,
                columns=ourCols,
                keytype="GENEID")
```

(a) Run the same query using all of the genes in our results table (`results.interaction.33`), and this time include the biotype of the genes too. Hint: You can find the name of the column for this by running `columns(EnsDb.Mmusculus.v79)`

(b) How many Ensembl genes have multipe Entrez IDs associated with them?

(c) Are all of the Ensembl gene IDs annotated? If not, why do you think this is?

```
# (a)
ourKeys <- rownames(results.interaction.11)

# (b)
ourCols <- c("SYMBOL", "GENEID", "ENTREZID", "GENEBIOTYPE")

# run the query
annot <- AnnotationDbi::select(EnsDb.Mmusculus.v79,
                keys=ourKeys,
                columns=ourCols,
                keytype="GENEID")

# (c)
annot %>%
    add_count(GENEID) %>%
    dplyr::filter(n>1) %>%
    distinct(GENEID) %>%
    count()
```

```
# (d)
length(unique(annot$GENEID))
length(ourKeys)
```

Exercise 2 - Volcano plot for 33 days

Now it's your turn! We just made the volcano plot for the 11 days contrast, you will make the one for the 33 days contrast.

If you haven't already make sure you load in our data and annotation. You can copy and paste the code below.

```
# First load data and annotations
results.interaction.33 <- readRDS("RObjects/DESeqResults.interaction_d33.rds")
ensemblAnnot <- readRDS("RObjects/Ensembl_annotations.rds")
```

(a) Shrink the results for the 33 days contrast.

```
#Shrink our values
ddsShrink.33 <- lfcShrink(ddsObj.interaction,
                          res = results.interaction.33,
                          type = "ashr")
```

```
## using 'ashr' for LFC shrinkage. If used in published research, please cite:
##     Stephens, M. (2016) False discovery rates: a new deal. Biostatistics, 18:2.
##     https://doi.org/10.1093/biostatistics/kxw041
```

```
shrinkTab.33 <- as.data.frame(ddsShrink.33) %>%
    rownames_to_column("GeneID") %>%
    left_join(ensemblAnnot, "GeneID") %>%
    rename(logFC=log2FoldChange, FDR=padj)
```

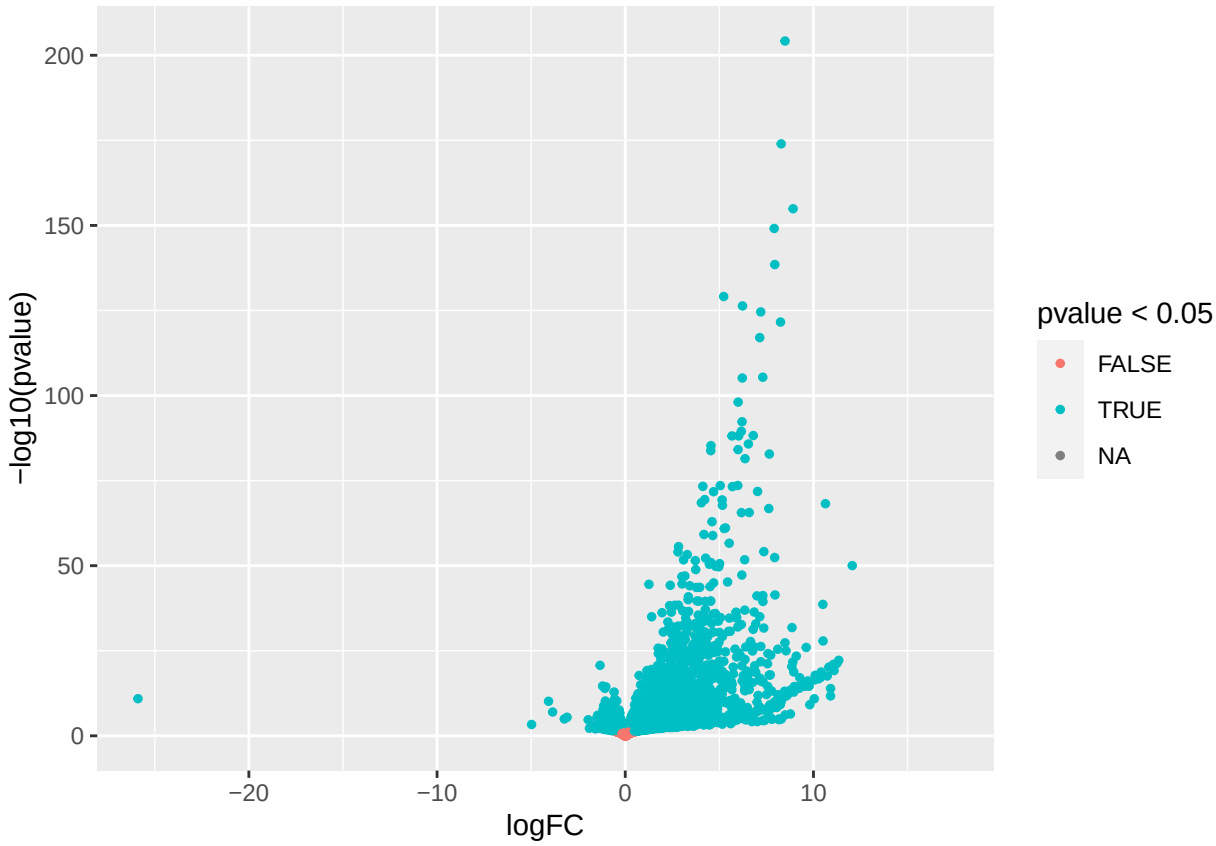
(b) Create a new column of -log10(pvalue) values in your shrinkTab for 33 days.

(c) Create a plot with points coloured by P-value < 0.05 similar to how we did in the first volcano plot

```
volcanoTab.33 <- shrinkTab.33 %>%
    mutate(`-log10(pvalue)` = -log10(pvalue))

ggplot(volcanoTab.33, aes(x = logFC, y=`-log10(pvalue)`)) +
    geom_point(aes(colour=pvalue < 0.05), size=1)
```

```
## Warning: Removed 47 rows containing missing values (geom_point).
```


200
150
100
50
0
−log10(pvalue)
−20
−10
0
10
logFC
pvalue < 0.05
FALSE
TRUE
NA