

Introduction to Bulk RNAseq data analysis

Annotation and Visualisation of Differential Expression Results

Last modified: 01 Mar 2022

Contents

Overview	1
Adding annotation to the DESeq2 results	1
Query the database	2
One we prepared earlier and one-to-many relationships	4
Visualisation	5
P-value histogram	5
Shrinking the log2FoldChange	5
MA plots	6
Volcano Plots	6
Exercise 1 - Volcano plot for 33 days	8
Exercise 2 - MA plot for day 33 with ggplot2	9
Venn Diagram	9
Heatmap	11

```
library(AnnotationHub)
library(AnnotationDbi)
library(ensemldb)
library(DESeq2)
library(tidyverse)
```

Before starting this section, we will make sure we have all the relevant objects from the Differential Expression analysis.

```
ddsObj.interaction <- readRDS("RObjects/DESeqDataSet.interaction.rds")
results.interaction.11 <- readRDS("RObjects/DESeqResults.interaction_d11.rds")
results.interaction.33 <- readRDS("RObjects/DESeqResults.interaction_d33.rds")
```

Overview

- Getting annotation
- Visualizing DE results

Adding annotation to the DESeq2 results

We have a list of significantly differentially expressed genes, but the only annotation we can see is the Ensembl Gene ID, which is not very informative.

There are a number of ways to add annotation. One method is to do this using a Bioconductor annotation package. These packages which are re-built every periodically with the latest annotations. These packages are listed on the annotation section of the Bioconductor, and are installed in the same way as regular Bioconductor packages.

Another approach is to use `biomaRt`, an interface to the BioMart resource. Using BioMart ensures that you are able to get the latest annotations for the GeneIDs, and can match the version of the gene annotation that was used for read counting.

A third method is to use `AnnotationHub`, this is like the bioconductor packages but in an online database like `bioMaRt`. They keep them slightly more up to date than the standard bioconductor packages and each time you use them the results are cached on your machine.

Today we will use the `AnnotationHub` method. A workflow for annotation with `biomaRt` is included in the extended materials section accessible on the course website.

Query the database

First we need to get the correct database from `AnnotationHub`. We make the instance (the first time we do this it will create a local cache on your machine so that repeat queries are very quick).

As you can see `ah` contains huge amounts of information and it is constantly changing. This is why it gives us the snapshot date so we know when our cached version is from. The `ah` object actually online contains pointers to where all the information is online and we don't want to download all of them as it would take a very long time and we don't need all of it.

This object is a vector and you can get information about a single resource by indexing with a single bracket `[` or download a resource with a double bracket `[[`.

```
# create an annotationhub instance
ah <- AnnotationHub()
ah

## AnnotationHub with 62386 records
## # snapshotDate(): 2021-10-20
## # $dataprotider: Ensembl, BroadInstitute, UCSC, ftp://ftp.ncbi.nlm.nih.gov/g...
## # $species: Homo sapiens, Mus musculus, Drosophila melanogaster, Bos taurus,...
## # $rdataclass: GRanges, TwoBitFile, BigWigFile, EnsDb, Rle, OrgDb, ChainFile...
## # additional mcols(): taxonomyid, genome, description,
## #   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
## #   rdatapath, sourceurl, sourcetype
## # retrieve records with, e.g., 'object[["AH5012"]]'
##
##           title
## AH5012    | Chromosome Band
## AH5013    | STS Markers
## AH5014    | FISH Clones
## AH5015    | Recomb Rate
## AH5016    | ENCODE Pilot
## ...      ...
## AH100299 | Zonotrichia_albicollis.Zonotrichia_albicollis-1.0.1.ncrna.2bit
## AH100300 | Zosterops_lateralis_melanops.ASM128173v1.cdna.all.2bit
## AH100301 | Zosterops_lateralis_melanops.ASM128173v1.dna_rm.toplevel.2bit
## AH100302 | Zosterops_lateralis_melanops.ASM128173v1.dna_sm.toplevel.2bit
## AH100303 | Zosterops_lateralis_melanops.ASM128173v1.ncrna.2bit
```

```
ah[1]
```

```
## AnnotationHub with 1 record
## # snapshotDate(): 2021-10-20
## # names(): AH5012
## # $dataprovder: UCSC
## # $species: Homo sapiens
## # $rdataclass: GRanges
## # $rdatadateadded: 2013-03-26
## # $title: Chromosome Band
## # $description: GRanges object from UCSC track 'Chromosome Band'
## # $taxonomyid: 9606
## # $genome: hg19
## # $sourcetype: UCSC track
## # $sourceurl: rtracklayer://hgdownload.cse.ucsc.edu/goldenpath/hg19/database...
## # $sourcesize: NA
## # $tags: c("cytoBand", "UCSC", "track", "Gene", "Transcript",
## #       "Annotation")
## # retrieve record with 'object[["AH5012"]]'
```

```
# Download the database we want to use
```

```
MouseEnsDb <- query(ah, c("EnsDb", "Mus musculus", "102"))[[1]]
```

This database contains the entire gene annotation from Ensembl release 102 for mouse. It includes descriptions of genes, transcripts, exons, UTRs etc.

We can turn the whole gene-level annotation table into a data frame so we can work with it using the tidyverse suite of tools.

```
annotations <- genes(MouseEnsDb, return.type = "data.frame")
```

```
# lets see what information we have
```

```
colnames(annotations)
```

```
## [1] "gene_id"          "gene_name"        "gene_biotype"
## [4] "gene_seq_start"  "gene_seq_end"     "seq_name"
## [7] "seq_strand"      "seq_coord_system" "description"
## [10] "gene_id_version" "canonical_transcript" "symbol"
## [13] "entrezid"
```

```
annot <- annotations %>%
  select(gene_id, gene_name, entrezid) %>%
  filter(gene_id %in% rownames(results.interaction.11))
```

NOTE: You may get an error with this command that looks like:

```
Error in (function (classes, fdef, mtable) :
  unable to find an inherited method for function 'select' for signature '"data.frame"'
```

This is due to the `select` function from `dplyr` (part of `tidyverse`) being masked by the `select` function from one of the annotation packages. This will have happened because the annotation package was loaded after the `tidyverse`. You can either restart your R session and reload the required packages, this time being sure to load `tidyverse` last, or just use `dplyr::select` to explicitly use the `select` function from `dplyr`:

```
annot <- annotations %>%
  dplyr::select(gene_id, gene_name, entrezid) %>%
  dplyr::filter(gene_id %in% rownames(results.interaction.11))
```

Missing annotations

Let's inspect the annotation.

```
head(annot)
```

```
##           gene_id gene_name entrezid
## 1 ENSMUSG00000051951      Xkr4  497097
## 2 ENSMUSG00000025900        Rp1   19888
## 3 ENSMUSG00000025902      Sox17   20671
## 4 ENSMUSG00000102269     Gm7357     NA
## 5 ENSMUSG00000103922     Gm6123     NA
## 6 ENSMUSG00000033845     Mrpl15   27395
```

```
length(annot$entrezid)
```

```
## [1] 20091
```

```
length(unique(annot$entrezid))
```

```
## [1] 17278
```

```
sum(is.na(annot$entrezid)) # Why are there NAs in the ENTREZID column?
```

```
## [1] 2782
```

Gene/transcript/protein IDs mapping between different databases not always perfect. Although majority of IDs map between databases, small subset may not have matching ID or may have more than one match. This is because feature identification algorithms, naming methodologies and versions may differ among databases. For instance NCBI and HGNC give same ID for different gene versions, whereas Ensembl assigned separate IDs for gene versions. Read interesting discussion on biostars.

There are some Ensembl IDs with no EntrezID. These gene ids has no corresponding Entrez ID in the `EnsDb` database package. The Ensembl and Entrez databases don't match on a 1:1 level although they have started taking steps towards consolidating in recent years.

One we prepared earlier and one-to-many relationships

To ensure everyone is working with same annotation, we have created an annotation table.

In this case we used the `biomaRt` package to download annotations directly from Ensembl. In this cases we can get additional columns, but will also sometimes get one-to-many relationships, where one Ensembl ID maps to multiple Entrez IDs. This sort of problem is common when mapping between annotation sources, but they have already been dealt with for us in AnnotationHub. If we wanted more control over this we would need to manually curate these one-to-many relationships ourselves.

In annotation table below we have modified the column names and dealt with the one-to-many/missing issues for Entrez IDs. The code we used for doing this is available in the extended materials section.

We will load out pre-created annotation table, and then combine it with our results table.

```
ensemblAnnot <- readRDS("RObjects/Ensembl_annotations.rds")
colnames(ensemblAnnot)
```

```
## [1] "GeneID"      "Entrez"      "Symbol"      "Description" "Biotype"
## [6] "Chr"        "Start"       "End"         "Strand"
```

```
annot.interaction.11 <- as.data.frame(results.interaction.11) %>%
  rownames_to_column("GeneID") %>%
```

```
left_join(ensemblAnnot, "GeneID") %>%
  rename(logFC=log2FoldChange, FDR=padj)
```

NOTE: If you had the issue earlier with `select` being masked and you used the `dplyr::select` solution, you will also have an issue `rename` and will need to use `dplyr::rename`.

Finally we can output the annotation DE results using `write_tsv`.

```
write_tsv(annot.interaction.11, "results/Interaction.11_Results_Annotated.txt")
```

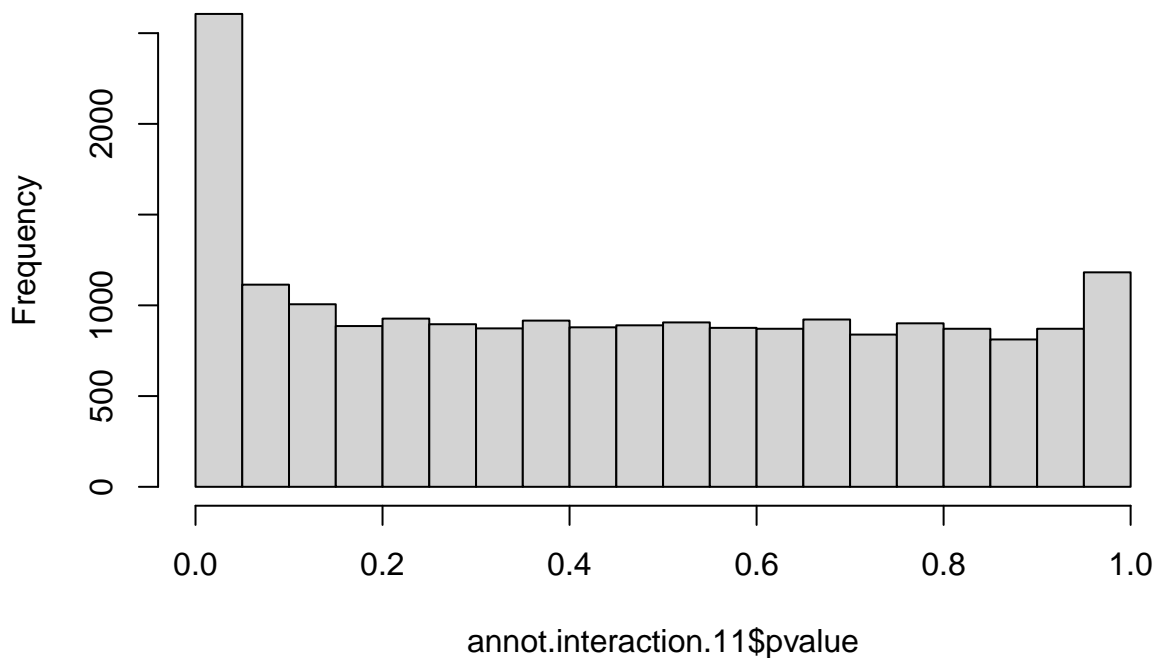
Visualisation

P-value histogram

A quick and easy “sanity check” for our DE results is to generate a p-value histogram. What we should see is a high bar at 0 - 0.05 and then a roughly uniform tail to the right of this. There is a nice explanation of other possible patterns in the histogram and what to do when you see them in this post.

```
hist(annot.interaction.11$pvalue)
```

Histogram of annot.interaction.11\$pvalue



Shrinking the log2FoldChange

DESeq2 provides a function called `lfcShrink` that shrinks log-Fold Change (LFC) estimates towards zero using an empirical Bayes procedure. The reason for doing this is that there is high variance in the LFC estimates when counts are low and this results in lowly expressed genes appearing to show greater differences between groups than highly expressed genes. The `lfcShrink` method compensates for this and allows better visualisation and ranking of genes. We will use it for our visualisation of the data.

```
ddsShrink.11 <- lfcShrink(ddsObj.interaction,
                        res = results.interaction.11,
                        type = "ashr")
```

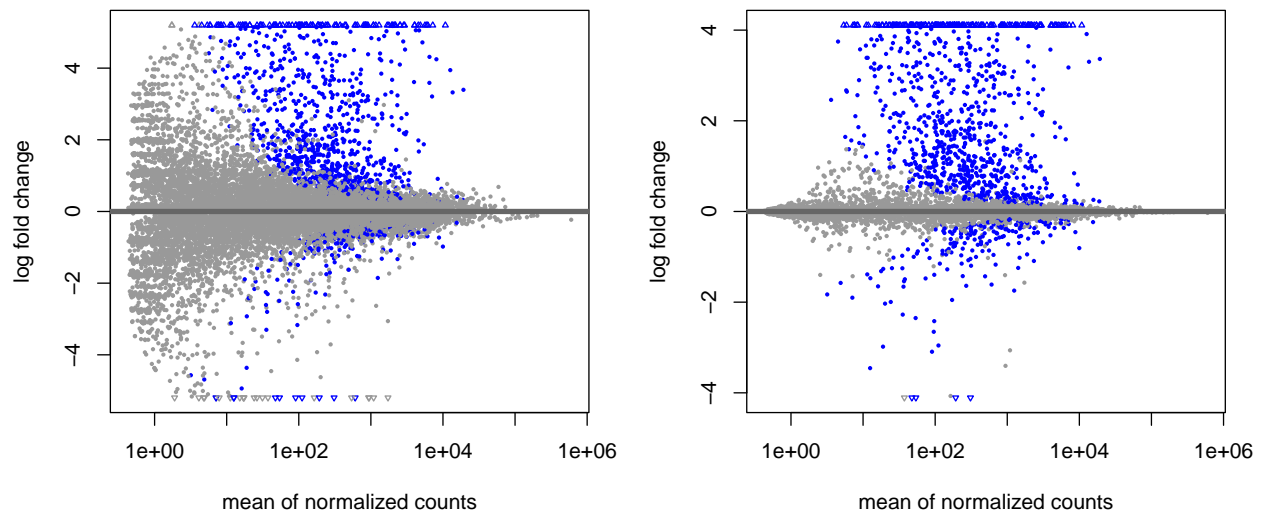
```
## using 'ashr' for LFC shrinkage. If used in published research, please cite:
##     Stephens, M. (2016) False discovery rates: a new deal. Biostatistics, 18:2.
##     https://doi.org/10.1093/biostatistics/kxw041
```

```
shrinkTab.11 <- as.data.frame(ddsShrink.11) %>%
  rownames_to_column("GeneID") %>%
  left_join(ensemblAnnot, "GeneID") %>%
  rename(logFC=log2FoldChange, FDR=padj)
```

MA plots

MA plots are a common way to visualize the results of a differential analysis. We met them briefly towards the end of the DESeq2 session. This plot shows the log-Fold Change for each gene against its average expression across all samples in the two conditions being contrasted. DESeq2 has a handy function for plotting this. Let's use it too compare the shrunk and un-shrunk fold changes.

```
par(mfrow=c(1,2))
plotMA(results.interaction.11, alpha=0.05)
plotMA(ddsShrink.11, alpha=0.05)
```



The DESeq2 in `plotMA` function is fine for a quick look, but these inbuilt functions aren't easy to customise, make changes to the way it looks or add things such as gene labels. For this we would recommend using the `ggplot` package.

Volcano Plots

Another common visualisation is the *volcano plot* which displays a measure of significance on the y-axis and fold-change on the x-axis. We will use `ggplot` to create this.

A Brief Introduction to ggplot2

The `ggplot2` package has emerged as an attractive alternative to the traditional plots provided by base R. A full overview of all capabilities of the package is available from the cheatsheet.

In brief:-

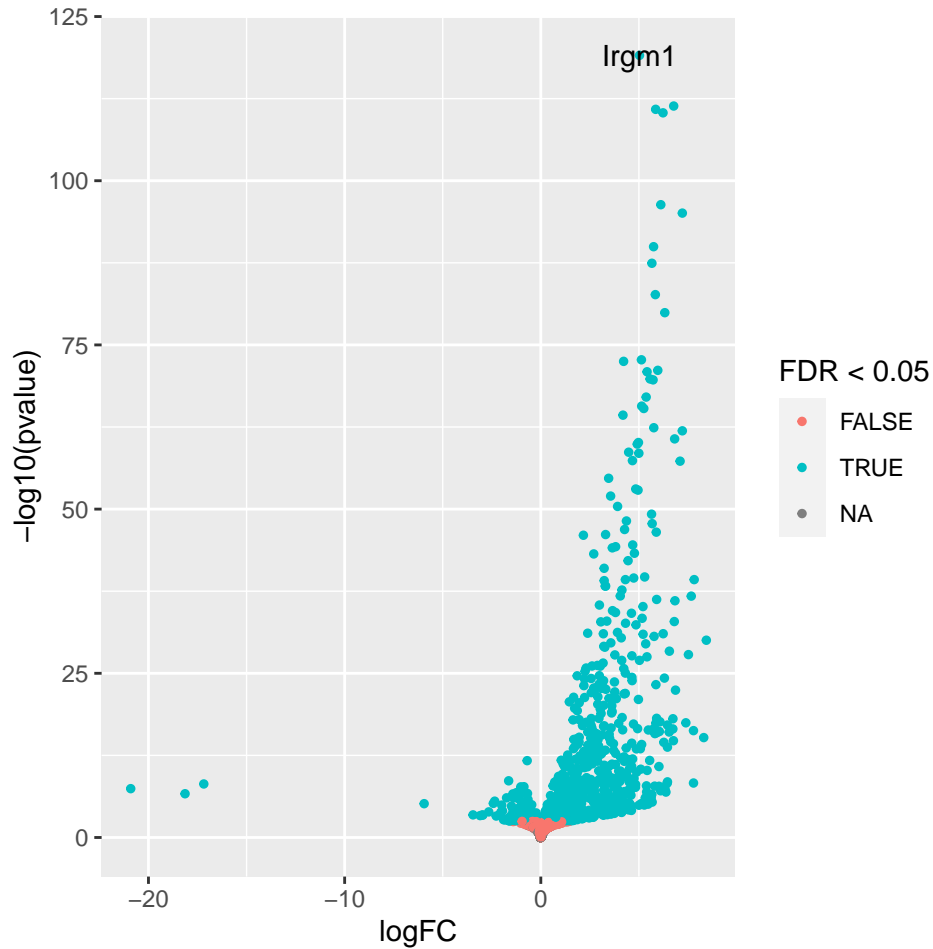
- `shrinkTab.11` is our data frame containing the variables we wish to plot
- `aes` creates a mapping between the variables in our data frame to the *aesthetic* properties of the plot:
 - the x-axis will be mapped to `logFC`
 - the y-axis will be mapped to the `-log10(pvalue)`
- `geom_point` specifies the particular type of plot we want (in this case a scatter plot)
- `geom_text` allows us to add labels to some or all of the points
 - see the cheatsheet for other plot types

The real advantage of `ggplot2` is the ability to change the appearance of our plot by mapping other variables to aspects of the plot. For example, we could colour the points based on the sample group. To do this we can add metadata from the `sampleinfo` table to the data. The colours are automatically chosen by `ggplot2`, but we can specify particular values. For the volcano plot we will colour according whether the gene has a pvalue below 0.05. We use a `-log10` transformation for the y-axis; it's commonly used for p-values as it means that more significant genes have a higher scale.

```
volcanoTab.11 <- shrinkTab.11 %>%
  mutate(`-log10(pvalue)` = -log10(pvalue))

ggplot(volcanoTab.11, aes(x = logFC, y=`-log10(pvalue)`)) +
  geom_point(aes(colour=FDR < 0.05), size=1) +
  geom_text(data=-top_n(.x, 1, wt=-FDR), aes(label=Symbol))
```

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



Exercise 1 - Volcano plot for 33 days

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.

```
## 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
```

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

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

(d) Compare these two volcano plots, what differences can you see between the two contrasts?


```
## 5 ENSMUSG00000~ FALSE          FALSE          FALSE          TRUE
## 6 ENSMUSG00000~ FALSE          FALSE          FALSE          FALSE
## 7 ENSMUSG00000~ TRUE           FALSE          TRUE           FALSE
## 8 ENSMUSG00000~ FALSE          FALSE          FALSE          FALSE
## 9 ENSMUSG00000~ FALSE          FALSE          FALSE          FALSE
## 10 ENSMUSG00000~ FALSE         FALSE          FALSE          FALSE
## # ... with 20,081 more rows
```

The first column contains the gene ids. The remaining columns are the four sets that we want to plot on the venn diagram:

- Significantly up-regulated on day 11
- Significantly down-regulated on day 11
- Significantly up-regulated on day 33
- Significantly down-regulated on day 33

These columns contain logical vectors (TRUE/FALSE) that indicate when or not the gene is included in that set.

```
vennDat <- tibble(Geneid=rownames(results.interaction.11)) %>%

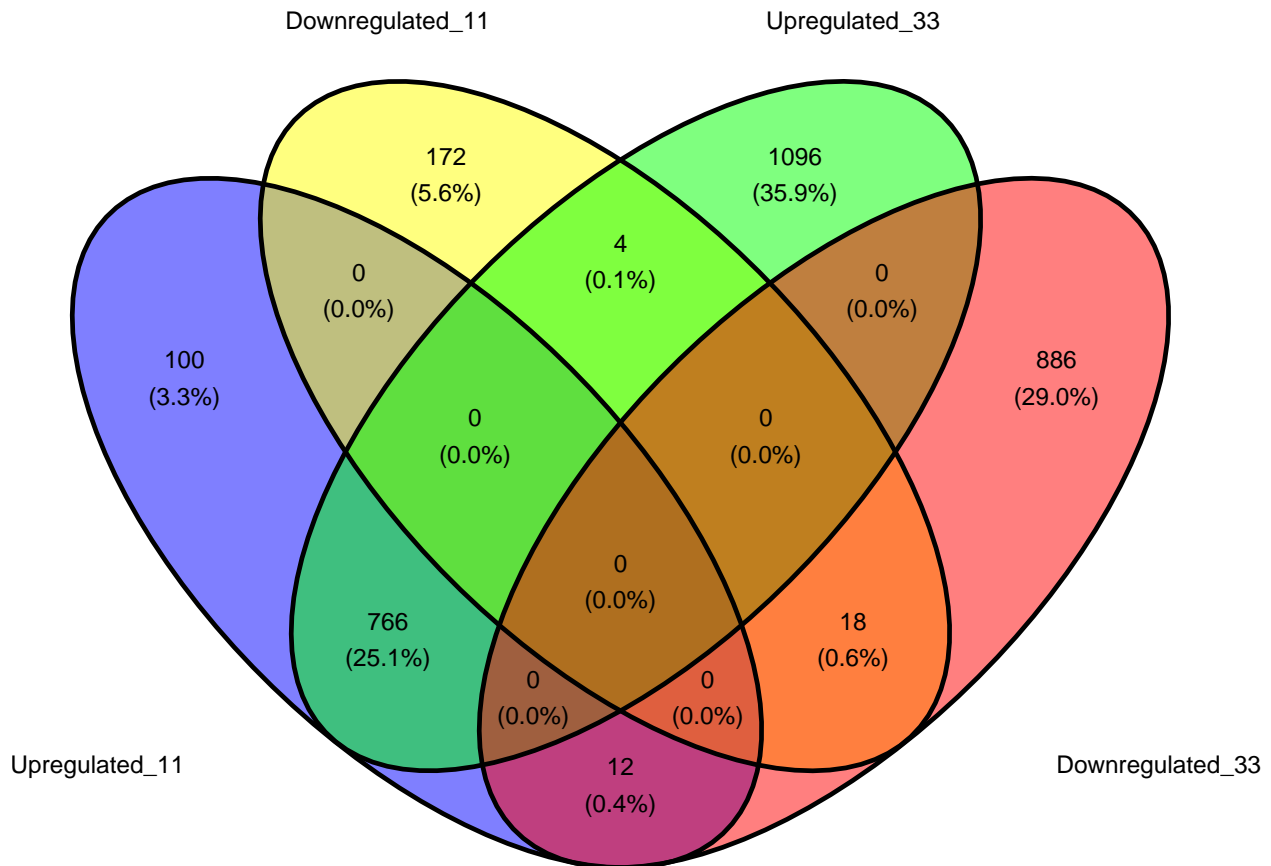
mutate(Upregulated_11 = results.interaction.11$padj < 0.05 &
       !is.na(results.interaction.11$padj) &
       results.interaction.11$log2FoldChange > 0) %>%

mutate(Downregulated_11 = results.interaction.11$padj < 0.05 &
       !is.na(results.interaction.11$padj) &
       results.interaction.11$log2FoldChange < 0) %>%

mutate(Upregulated_33 = results.interaction.33$padj < 0.05 &
       !is.na(results.interaction.33$padj) &
       results.interaction.33$log2FoldChange > 0) %>%

mutate(Downregulated_33 = results.interaction.33$padj < 0.05 &
       !is.na(results.interaction.33$padj) &
       results.interaction.33$log2FoldChange < 0)

ggvenn(vennDat, set_name_size = 4)
```



Heatmap

We're going to use the package `ComplexHeatmap` (Gu, Eils, and Schlesner 2016). We'll also use `circlize` to generate a colour scale (Gu et al. 2014).

```
library(ComplexHeatmap)
library(circlize)
```

We can't plot the entire data set, let's just select the top 300 by FDR. We'll want to use normalised expression values, so we'll use the `vst` function.

```
# get the top genes
sigGenes <- shrinkTab.11 %>%
  top_n(300, wt=-FDR) %>%
  pull("GeneID")

# filter the data for the top 300 by padj
plotDat <- vst(ddsObj.interaction)[sigGenes,] %>%
  assay()
```

The range expression values for different genes can vary widely. Some genes will have very high expression. Our heatmap is going to be coloured according to gene expression. If we used a colour scale from 0 (no expression) to the maximum expression, the scale will be dominated by our most extreme genes and it will be difficult to discern any difference between most of the genes.

To overcome this we will z-scale the counts. This scaling method results in values for each that show the number of standard deviations the gene expression is from the mean for that gene across all the sample - the

mean will be '0', '1' means 1 standard deviation higher than the mean, '-1' means 1 standard deviation lower than the mean.

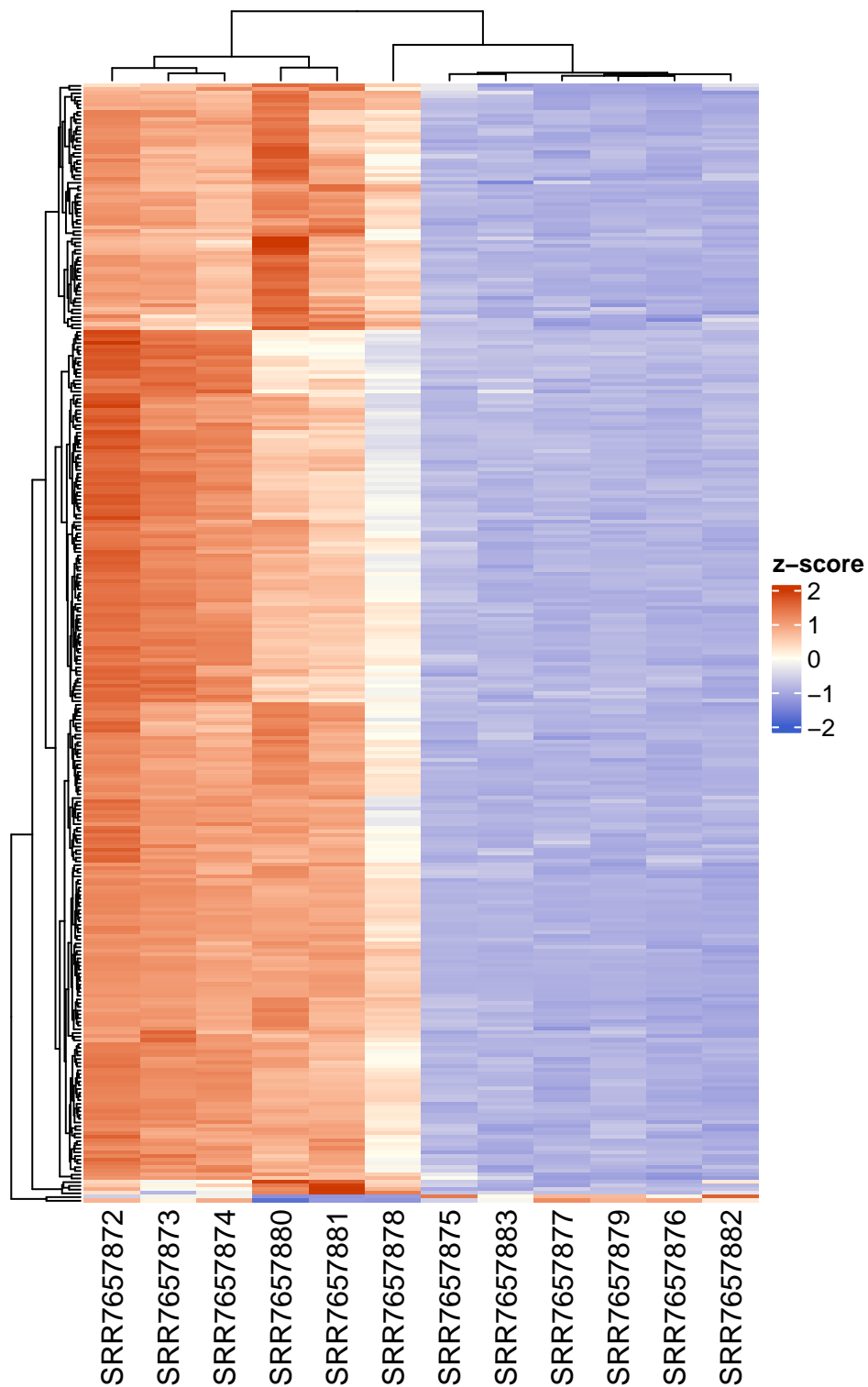
```
z.mat <- t(scale(t(plotDat), center=TRUE, scale=TRUE))
```

```
# colour palette
```

```
myPalette <- c("royalblue3", "ivory", "orangered3")
```

```
myRamp <- colorRamp2(c(-2, 0, 2), myPalette)
```

```
Heatmap(z.mat, name = "z-score",  
        col = myRamp,  
        show_row_names = FALSE)
```

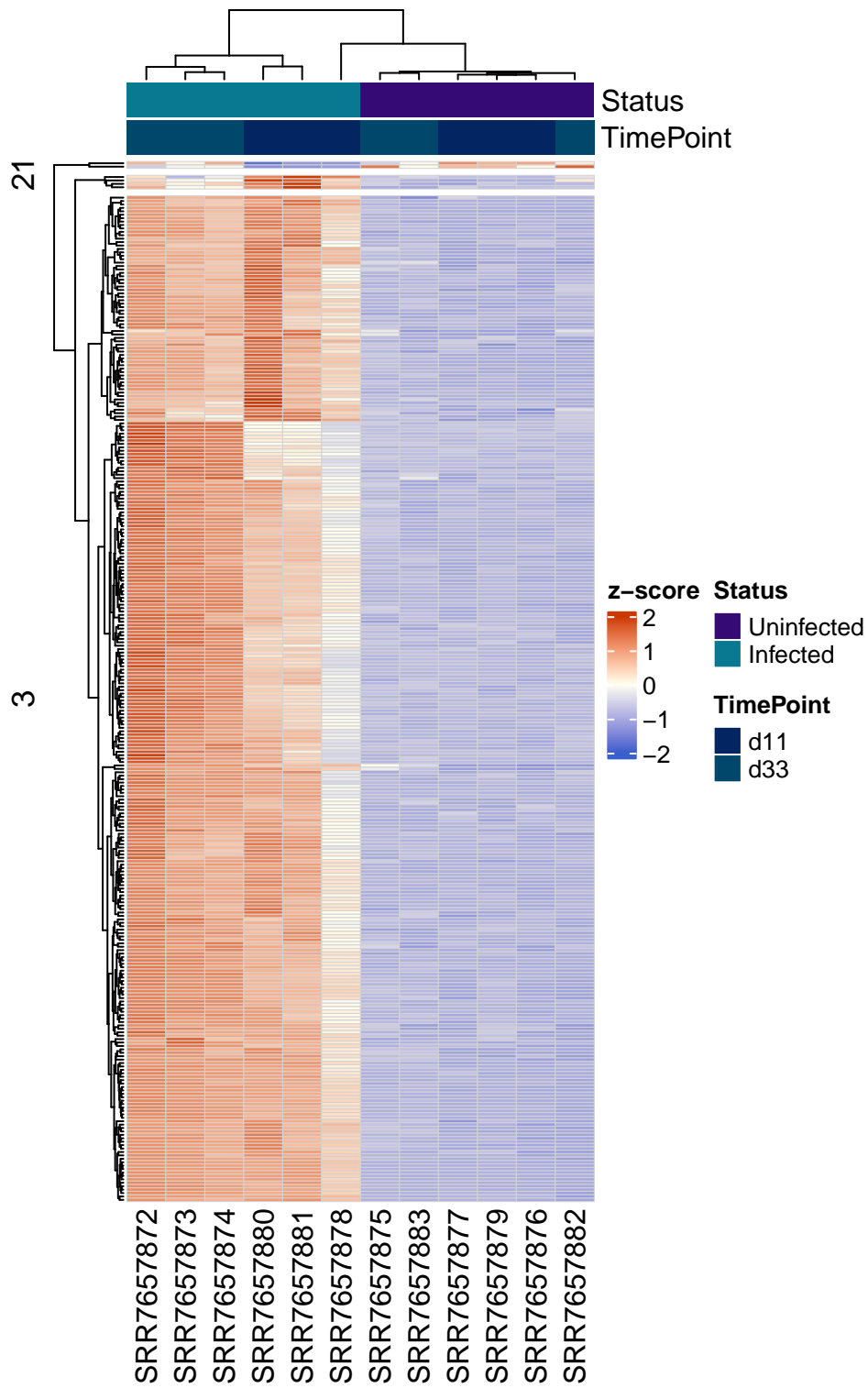


we can also split the heat map into clusters and add some annotation.

```
ha1 = HeatmapAnnotation(df = colData(ddsObj.interaction)[,c("Status", "TimePoint")])

Heatmap(z.mat, name = "z-score",
        col = myRamp,
        show_row_name = FALSE,
```

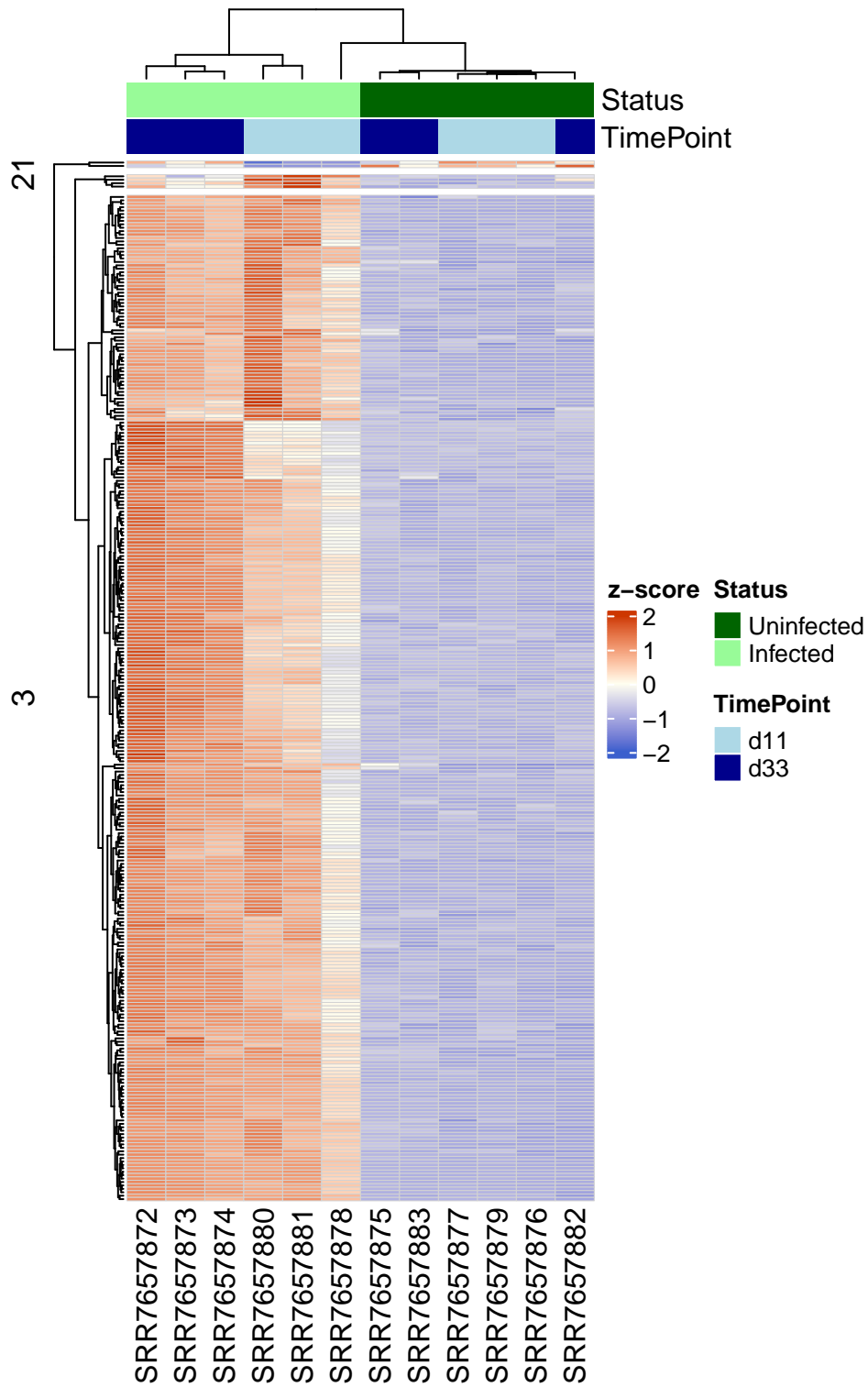
```
split=3,
rect_gp = gpar(col = "lightgrey", lwd=0.3),
top_annotation = ha1)
```



Whenever we teach this session several student always ask how to set the colours of the bars at the top of the heatmap. This is shown below.

```
ha1 = HeatmapAnnotation(df = colData(ddsObj.interaction)[,c("Status", "TimePoint")],
                        col = list(Status = c("Uninfected" = "darkgreen",
                                              "Infected" = "palegreen"),
                                   TimePoint = c("d11" = "lightblue",
                                                  "d33" = "darkblue")))

Heatmap(z.mat, name = "z-score",
        col = myRamp,
        show_row_name = FALSE,
        split=3,
        rect_gp = gpar(col = "lightgrey", lwd=0.3),
        top_annotation = ha1)
```



```

saveRDS(annot.interaction.11, file="results/Annotated_Results.d11.rds")
saveRDS(shrinkTab.11, file="results/Shrunk_Results.d11.rds")
saveRDS(annot.interaction.33, file="results/Annotated_Results.d33.rds")
saveRDS(shrinkTab.33, file="results/Shrunk_Results.d33.rds")

```


- Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. “Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data.” *Bioinformatics*.
- Gu, Zuguang, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. 2014. “Circlize Implements and Enhances Circular Visualization in r.” *Bioinformatics* 30: 2811–12.