SciX: Genies

Helen King/Lachlan Gray

7/06/2024

Introduction

This file is an Rmarkdown file. Upon successfully installed R and RStudio you should be able to follow the instructions below. If you run into errors please flag on the GitHub page or search the error in Google.

To run each chunk of code press the green arrow next to the command or select the code and run with 'command + enter'.

Opening this Rmarkdown file will place us into the required working directory. A directory is another name for folders on the computer. All of the plots and files we generate will be saved to the working directory. We can find which directory we are in with the getwd() command:

```
## [1] "/Users/hellyk/Desktop/Weatheritt_Lab_Y3/SciX/PD"
```

For MacOS and Linux getwd() should return the following path: "/Users/USER/Desktop/SciX-main/PD"

Setting up R

We can then see which files are available in this directory with the dir() command:

Install packages

If asked to update all/some/none just enter 'a' in the console below.

Load the packages

Set working directory

We need to tell R which folder (known as directory) our data is located and where we want to store our results. This directory will be the unzipped SciX-main directory we just downloaded.
As naming paths are different in Mac and PC please make sure you run the appropriate line of code.

Read in RNA sequencing count matrix. This is the data that we will be using for this experiment

The read.csv command allows us to load a comma separated file into R. This file contains data in the form of a matrix (a grid of numbers). The "header" option is set to "T" which means that the first row of the file contains the names of the columns. The "row.names" option is set to "1" which means that the first column of the file containing the gene names is used to name each row. We then print the first five rows of the matrix (which includes all of the columns) to the screen. It also prints the dimensions of the matrix, which tells us the number of rows (genes) and columns (individuals) in the matrix.

```
##                                GSM2843848 GSM2843849 GSM2843850 GSM2843851
## ENSG00000223972.5|DDX11L1              37        140         55         96
## ENSG00000243485.5|MIR1302-2HG           0          6          6          5
## ENSG00000268020.3|OR4G4P                0          0          0          0
## ENSG00000240361.2|OR4G11P               0          0          0          0
## ENSG00000186092.6|OR4F5                 5          0          0          2
##                                GSM2843852 GSM2843853 GSM2843854 GSM2843855
## ENSG00000223972.5|DDX11L1              74         48         61         89
## ENSG00000243485.5|MIR1302-2HG           6          2          0          0
## ENSG00000268020.3|OR4G4P                0          0          0          0
## ENSG00000240361.2|OR4G11P               0          0          0          0
## ENSG00000186092.6|OR4F5                 0          0          0          0
##                                GSM2843856 GSM2843857 GSM2843858 GSM2843859
## ENSG00000223972.5|DDX11L1              37        136         88         36
## ENSG00000243485.5|MIR1302-2HG           0          8          0          0
## ENSG00000268020.3|OR4G4P                0          0          0          0
## ENSG00000240361.2|OR4G11P               0          0          0          0
## ENSG00000186092.6|OR4F5                 0          0          0          1
##                                GSM2843860 GSM2843861 GSM2843862 GSM2843863
## ENSG00000223972.5|DDX11L1              21         90         53         76
## ENSG00000243485.5|MIR1302-2HG           4          7          0         10
## ENSG00000268020.3|OR4G4P                0          0          0          0
## ENSG00000240361.2|OR4G11P               0          0          0          0
## ENSG00000186092.6|OR4F5                 5         51          0          0
```

```
## [1] 58721    16
```

Read in sample metadata

```
##    age condition    sex individual
## 1   77   control female GSM2843849
## 3   72   control female GSM2843852
## 5   64   control female GSM2843855
## 7   66   control   male GSM2843848
## 9   78   control   male GSM2843850
## 11  81   control   male GSM2843851
## 13  73   control   male GSM2843853
## 15  62   control   male GSM2843854
## 17  63   control   male GSM2843856
## 19  55   disease female GSM2843862
## 21  82   disease female GSM2843863
## 23  88   disease   male GSM2843857
## 25  78   disease   male GSM2843858
## 27  79   disease   male GSM2843859
## 29  91   disease   male GSM2843860
## 31  76   disease   male GSM2843861
```

To make the column names more informative we replace with

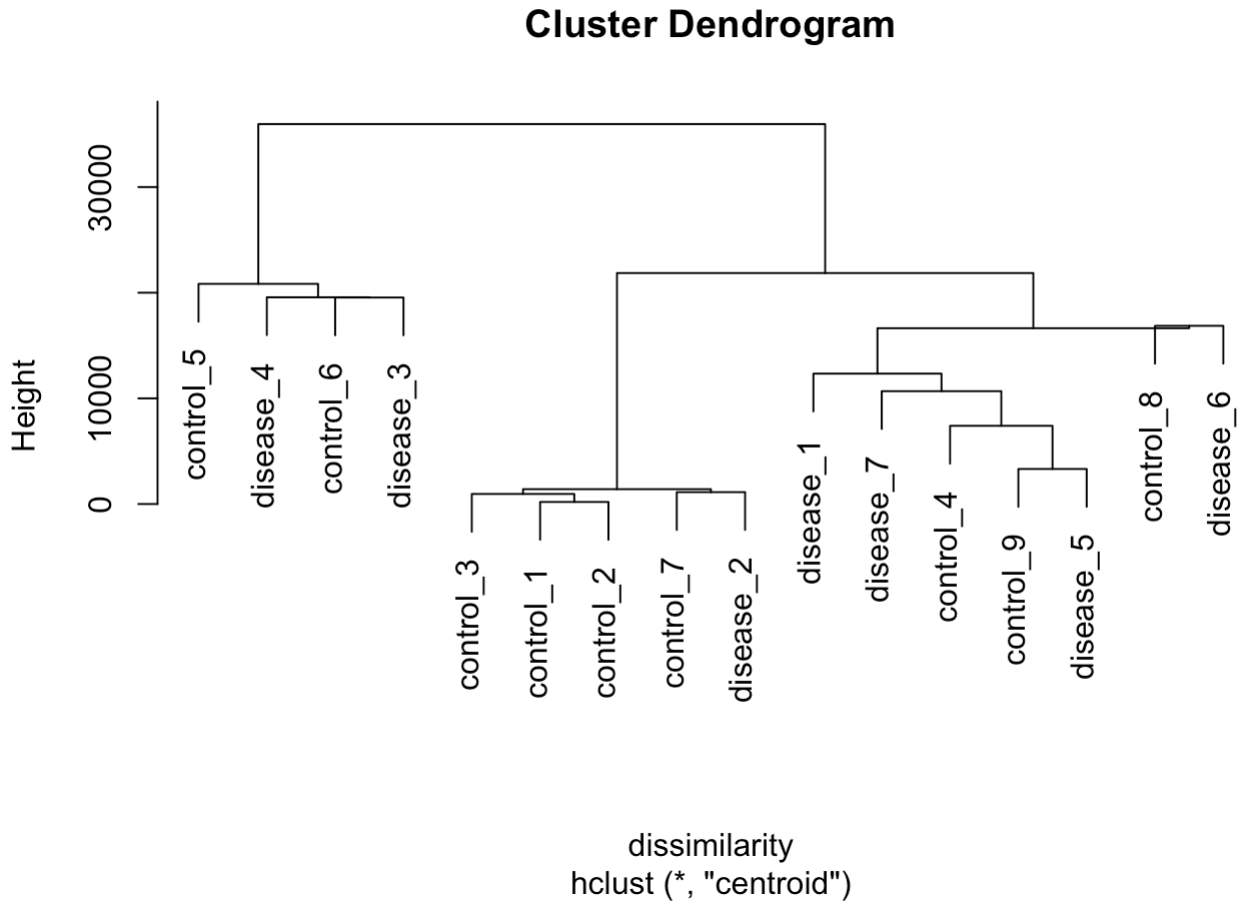
metadata$condition column

```
##                               control_1 control_2 control_3 control_4 control_5
## ENSG00000223972.5|DDX11L1           140        74        89        37        55
## ENSG00000243485.5|MIR1302-2HG         6         6         0         0         6
## ENSG00000268020.3|OR4G4P              0         0         0         0         0
## ENSG00000240361.2|OR4G11P             0         0         0         0         0
## ENSG00000186092.6|OR4F5               0         0         0         5         0
##                               control_6 control_7 control_8 control_9 disease_1
## ENSG00000223972.5|DDX11L1            96        48        61        37        53
## ENSG00000243485.5|MIR1302-2HG         5         2         0         0         0
## ENSG00000268020.3|OR4G4P              0         0         0         0         0
## ENSG00000240361.2|OR4G11P             0         0         0         0         0
## ENSG00000186092.6|OR4F5               2         0         0         0         0
##                               disease_2 disease_3 disease_4 disease_5 disease_6
## ENSG00000223972.5|DDX11L1            76       136        88        36        21
## ENSG00000243485.5|MIR1302-2HG        10         8         0         0         4
## ENSG00000268020.3|OR4G4P              0         0         0         0         0
## ENSG00000240361.2|OR4G11P             0         0         0         0         0
## ENSG00000186092.6|OR4F5               0         0         0         1         5
##                               disease_7
## ENSG00000223972.5|DDX11L1            90
## ENSG00000243485.5|MIR1302-2HG         7
## ENSG00000268020.3|OR4G4P              0
## ENSG00000240361.2|OR4G11P             0
## ENSG00000186092.6|OR4F5              51
```

Adding biological sex to the metadata file

You may have noticed that information about the individuals age and sex is missing from the metadata. By looking at expression of genes on the X and Y chromosomes we can determine the biological sex of these samples. The method to infer sex from gene expression is a little complicated but we can return to this later if

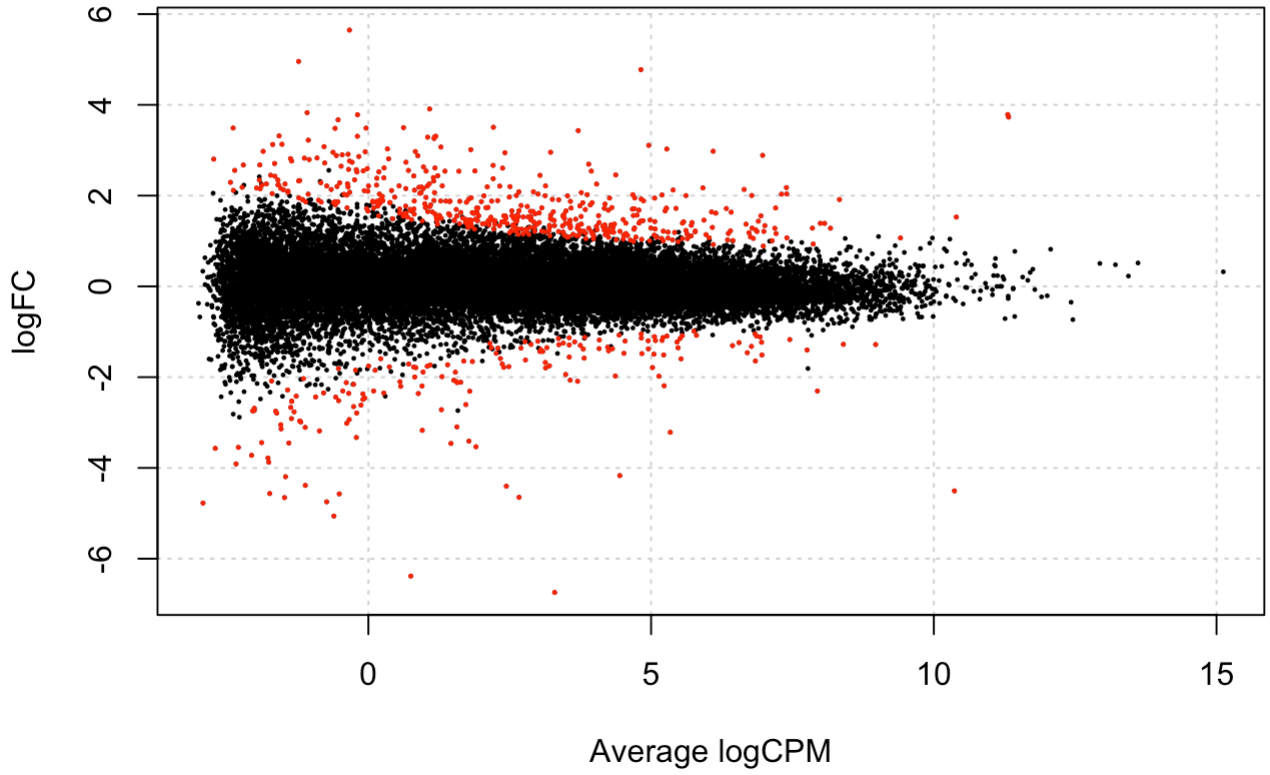
you'd like.




```
## control_1 control_2 control_3 control_4 control_5 control_6 control_7 control_8
##    400756    378854    197262        29       131       136    154709       102
## control_9 disease_1 disease_2 disease_3 disease_4 disease_5 disease_6 disease_7
##       572       209    313827        13        23        11        84       192
```

Differential expression analysis with edgeR likelihood ratio test

We will perform a statistical test to determine which genes are different between our conditions. For this, we will use the likelihood ratio test which takes models from each condition and compares them. We then make our disease samples the reference group. This tells us the difference in gene expression in relation to our disease group. For example, a gene with a positive (+) logFC is upregulated in disease and a negative (-) logFC is downregulated in disease. We then filter out lowly expressed genes, normalise the expression values and perform the test. To visualise our results, we create plots to show differentially expressed genes.




```
## quartz_off_screen
##                 2
```

```
##        groupdisease
## Down            177
## NotSig        21927
## Up              505
```

```
##                                gene     logFC    logCPM        LR       PValue
## ENSG00000173110.7|HSPA6      HSPA6  4.776892  4.819443 200.04667 2.040080e-45
## ENSG00000142319.17|SLC6A3   SLC6A3 -6.744363  3.294244 178.24357 1.172006e-40
## ENSG00000229807.11|XIST       XIST -4.505627 10.364959 123.79354 9.348158e-29
## ENSG00000204389.9|HSPA1A    HSPA1A  3.786322 11.307760 101.55625 6.946281e-24
## ENSG00000204388.6|HSPA1B    HSPA1B  3.730131 11.322352  99.99425 1.528400e-23
## ENSG00000108691.9|CCL2        CCL2  3.107621  4.958159  93.37760 4.321137e-22
## ENSG00000165646.13|SLC18A2 SLC18A2 -4.402911  2.440210  91.07677 1.382050e-21
## ENSG00000140379.7|BCL2A1    BCL2A1  3.432536  3.711322  89.85868 2.557939e-21
## ENSG00000132002.7|DNAJB1    DNAJB1  2.977453  6.097688  86.59917 1.329030e-20
## ENSG00000106211.8|HSPB1      HSPB1  2.887431  6.972562  82.80741 9.045033e-20
##                                   FDR
## ENSG00000173110.7|HSPA6    4.354155e-41
## ENSG00000142319.17|SLC6A3  1.250709e-36
## ENSG00000229807.11|XIST    6.650611e-25
## ENSG00000204389.9|HSPA1A   3.706373e-20
## ENSG00000204388.6|HSPA1B   6.524148e-20
## ENSG00000108691.9|CCL2     1.537105e-18
## ENSG00000165646.13|SLC18A2 4.213884e-18
## ENSG00000140379.7|BCL2A1   6.824281e-18
## ENSG00000132002.7|DNAJB1   3.151730e-17
## ENSG00000106211.8|HSPB1    1.930487e-16
```

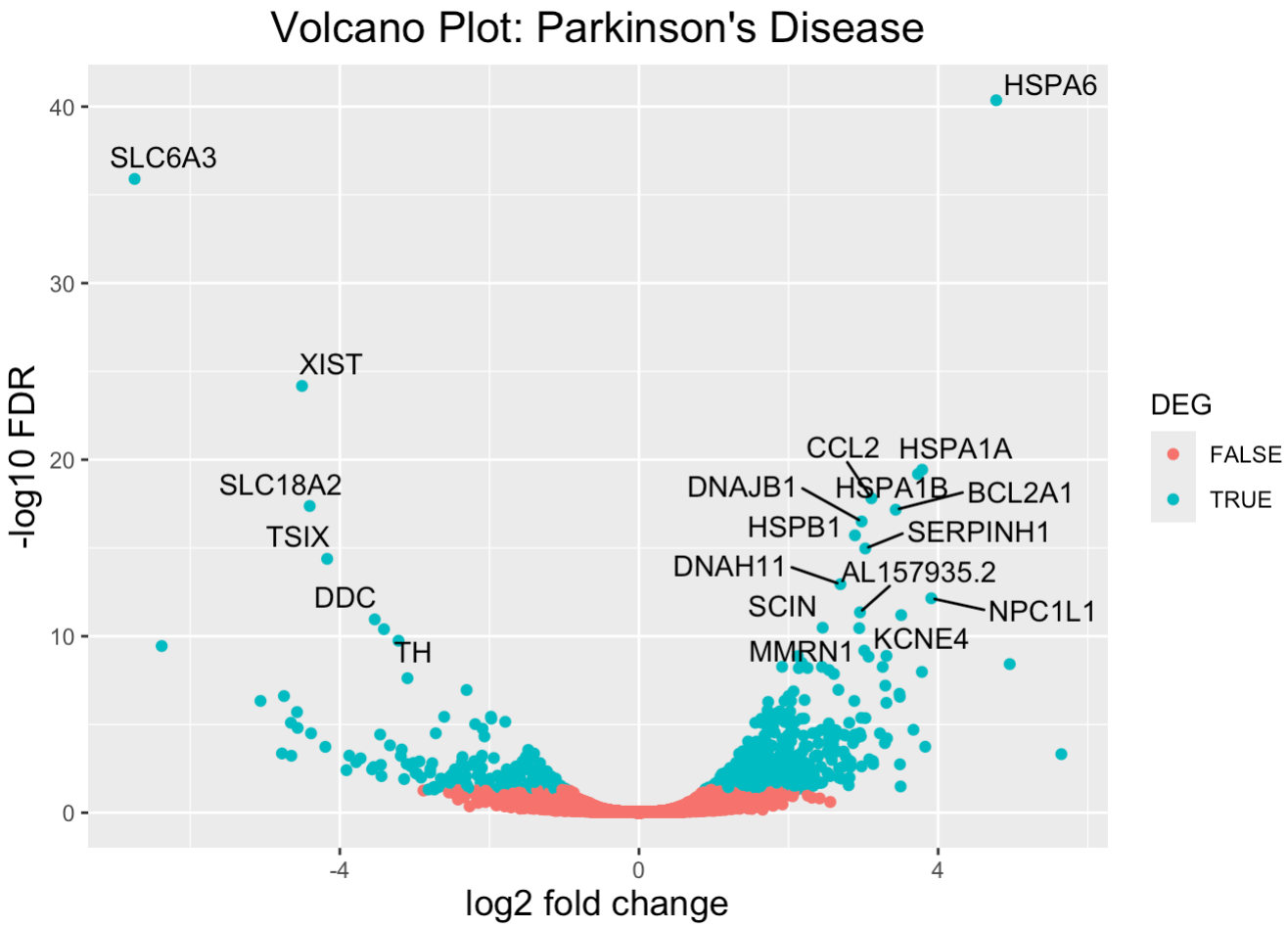
```
## quartz_off_screen
##                 2
```

Save result file to working directory

```
write.table(lrt, row.names = F, sep = "\t", 'edgeR-LRT.PD.txt')
```

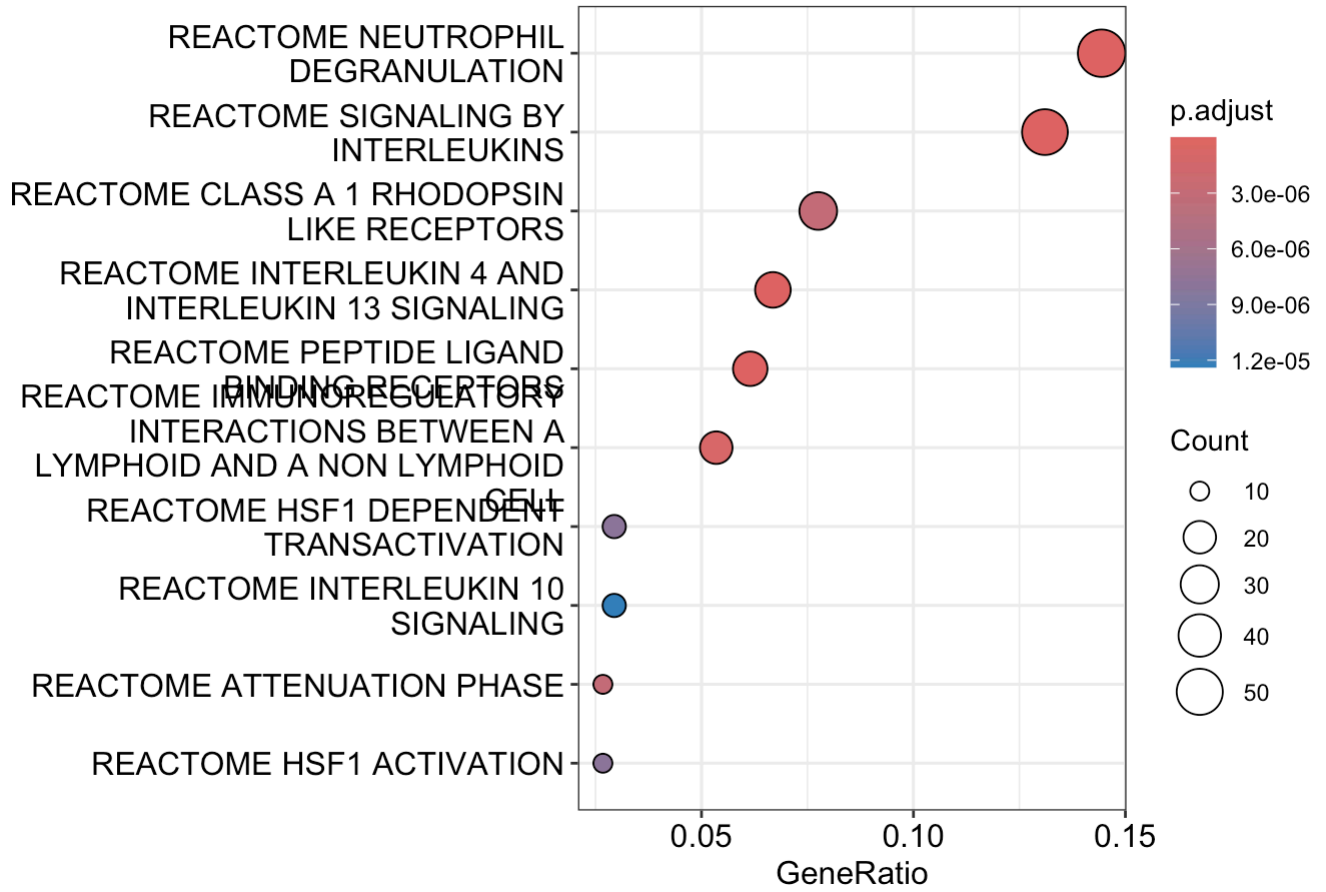
Displaying results in volcano plot

This plot displays the log fold-change and false dicovery rate for each gene. You can select the number of genes to label with the **n.genes** variable below.

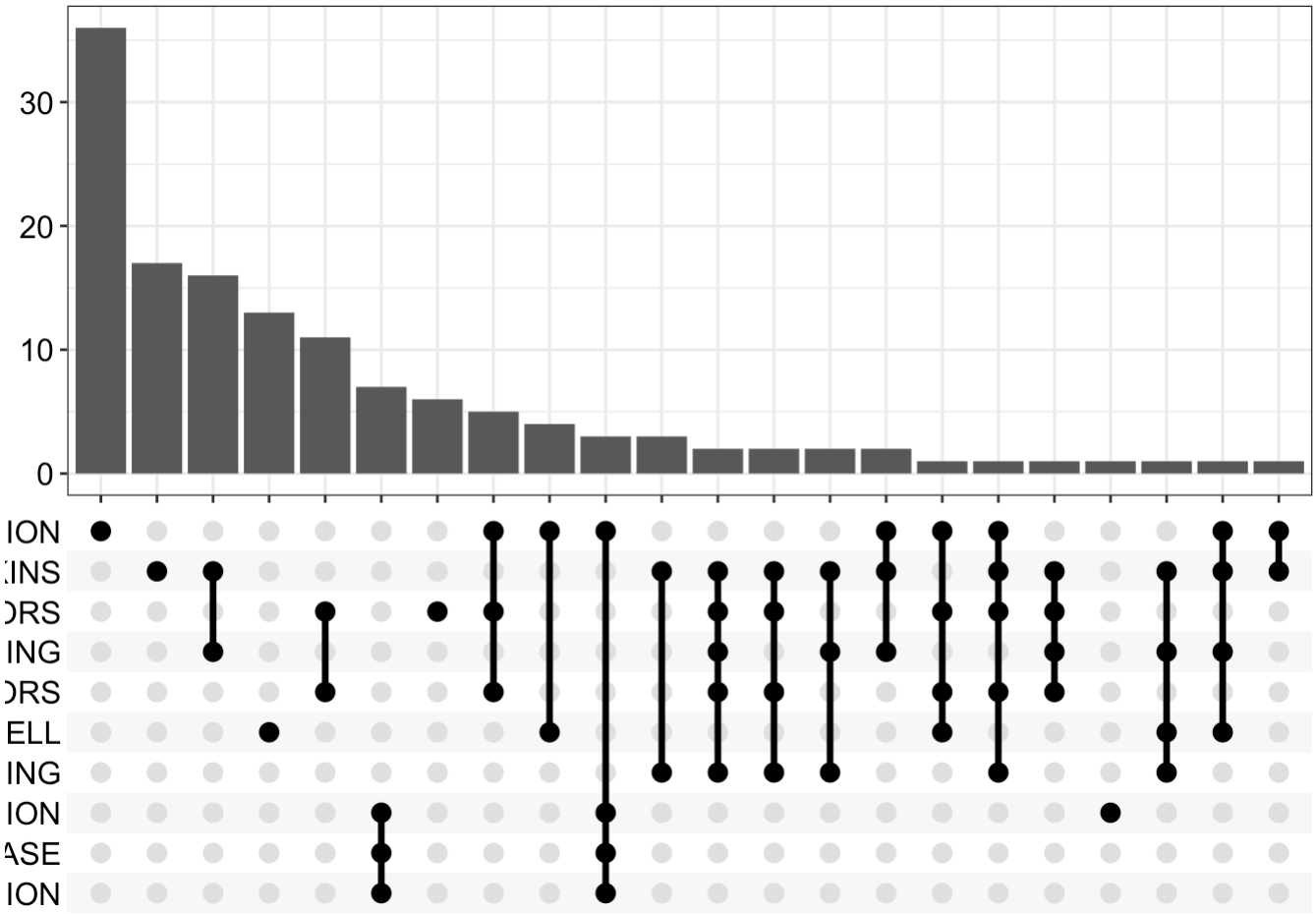



```
## quartz_off_screen
##                 2
```

Over Representation Analysis (ORA)




```
## quartz_off_screen
##                 2
```



```
## quartz_off_screen
##                 2
```

Match genes to DisGeneNet and perform chi-squared test

```
## [1] "Expected values"
```

```
##           [,1]      [,2]
## [1,]  53.64081  1681.359
## [2,] 645.35919 20228.641
```

```
## [1] "Observed values"
```

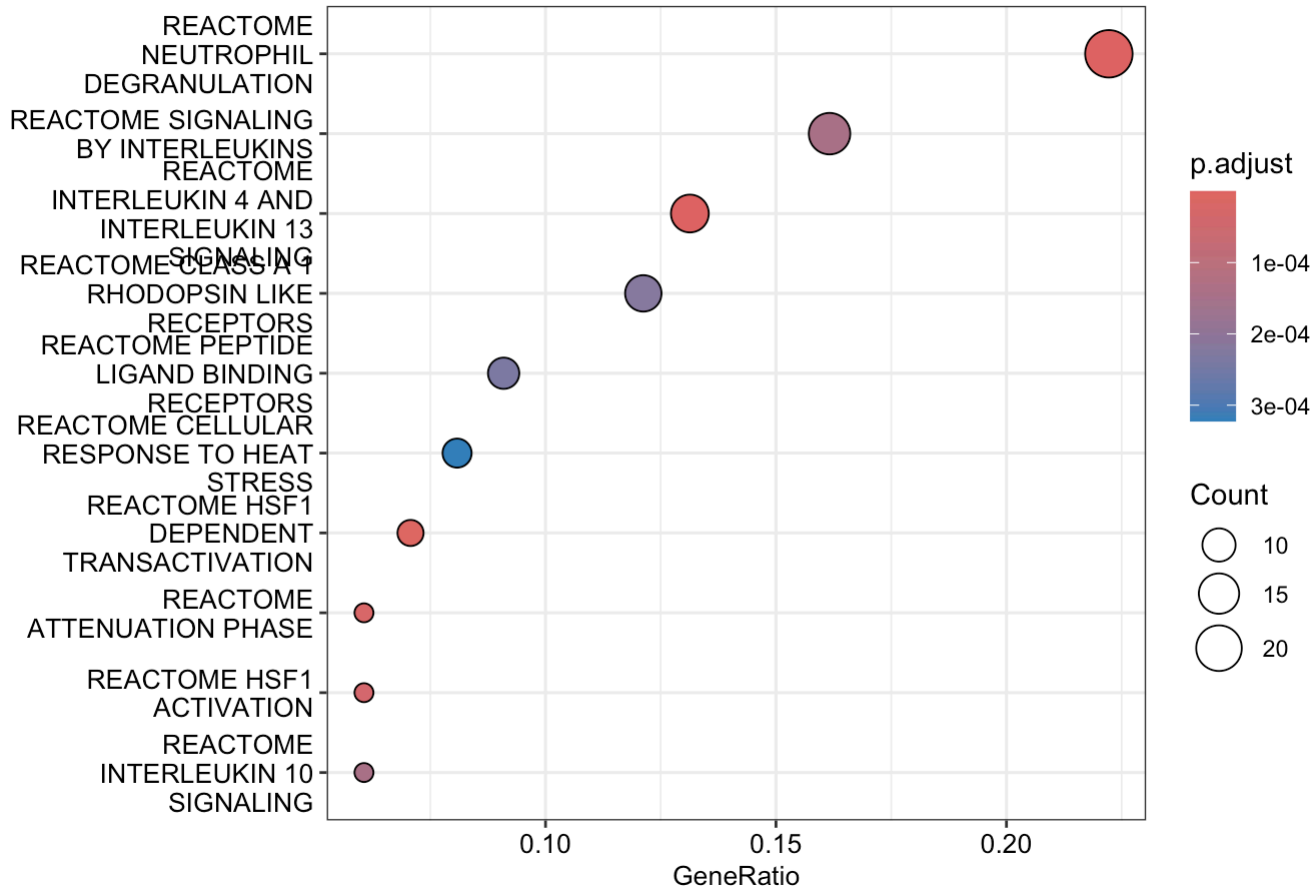
```
##      [,1]  [,2]
## [1,]  121  1614
## [2,]  578 20296
```

```
## [1] "Pearson residuals"
```

```
##           [,1]      [,2]
## [1,]  9.197064 -1.642732
## [2,] -2.651527  0.473602
```

```
## [1] "chi.squared p.value"
```

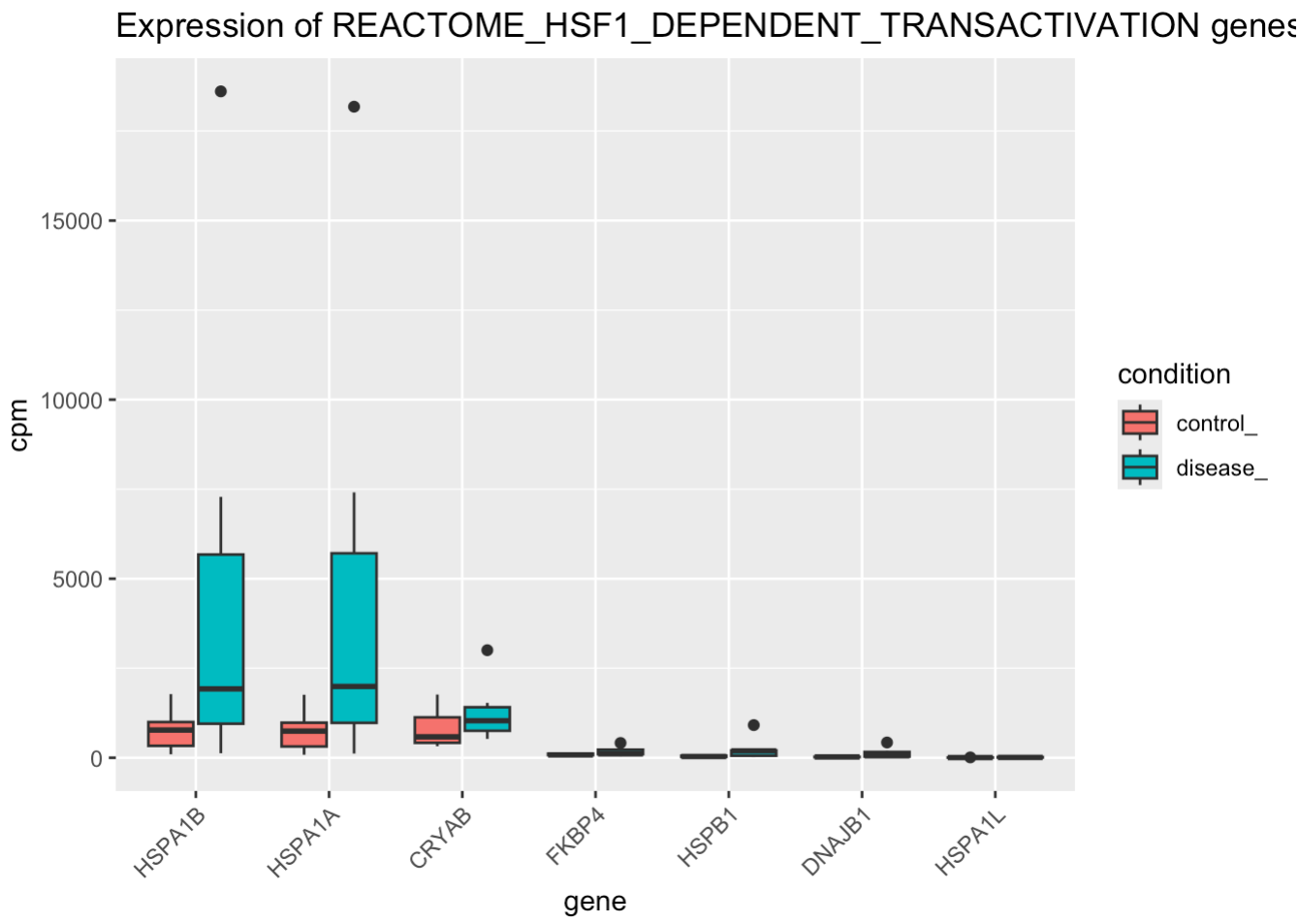
```
## [1] 4.869461e-22
```




```
## quartz_off_screen
##                 2
```

Extract genes from interesting pathway. Select pathway with

pathway variable




```
## quartz_off_screen
##                 2
```

```
##                             gene    logFC    logCPM        LR       PValue
## ENSG00000204390.9|HSPA1L HSPA1L 1.497812  3.220834  14.71860 1.248091e-04
## ENSG00000204389.9|HSPA1A HSPA1A 3.786322 11.307760 101.55625 6.946281e-24
## ENSG00000204388.6|HSPA1B HSPA1B 3.730131 11.322352  99.99425 1.528400e-23
## ENSG00000106211.8|HSPB1   HSPB1 2.887431  6.972562  82.80741 9.045033e-20
## ENSG00000109846.7|CRYAB   CRYAB 1.067847  9.974296  10.05083 1.522792e-03
## ENSG00000004478.7|FKBP4   FKBP4 1.211059  6.993762  13.49997 2.385671e-04
## ENSG00000132002.7|DNAJB1 DNAJB1 2.977453  6.097688  86.59917 1.329030e-20
##                                    FDR threshold
## ENSG00000204390.9|HSPA1L 7.721186e-03      TRUE
## ENSG00000204389.9|HSPA1A 3.706373e-20      TRUE
## ENSG00000204388.6|HSPA1B 6.524148e-20      TRUE
## ENSG00000106211.8|HSPB1  1.930487e-16      TRUE
## ENSG00000109846.7|CRYAB  4.749481e-02      TRUE
## ENSG00000004478.7|FKBP4  1.254126e-02      TRUE
## ENSG00000132002.7|DNAJB1 3.151730e-17      TRUE
```