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/AD"
```

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

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

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.

```
##                           SRR12850830 SRR12850831 SRR12850832 SRR12850833
## ENSG00000223972.5|DDX11L1           7          19           6          14
## ENSG00000237613.2|FAM138A           2           2           1           3
## ENSG00000268020.3|OR4G4P            0           1           1           0
## ENSG00000240361.2|OR4G11P           0           1           1           0
## ENSG00000186092.6|OR4F5             1           1           0           0
##                           SRR12850834 SRR12850835 SRR12850836 SRR12850837
## ENSG00000223972.5|DDX11L1           9           9           8           5
## ENSG00000237613.2|FAM138A           0           2           2           0
## ENSG00000268020.3|OR4G4P            0           0           0           0
## ENSG00000240361.2|OR4G11P           0           0           0           0
## ENSG00000186092.6|OR4F5             0           0           1           0
##                           SRR12850838 SRR12850839 SRR12850840 SRR12850841
## ENSG00000223972.5|DDX11L1          11           8           9           4
## ENSG00000237613.2|FAM138A           1           0           2           0
## ENSG00000268020.3|OR4G4P            0           0           0           0
## ENSG00000240361.2|OR4G11P           1           0           0           0
## ENSG00000186092.6|OR4F5             2           0           0           0
##                           SRR12850842 SRR12850843 SRR12850844 SRR12850845
## ENSG00000223972.5|DDX11L1           9           8          12          14
## ENSG00000237613.2|FAM138A           0           1           3           0
## ENSG00000268020.3|OR4G4P            0           0           0           0
## ENSG00000240361.2|OR4G11P           0           0           0           1
## ENSG00000186092.6|OR4F5             1           0           0           0
##                           SRR12850846 SRR12850847 SRR12850848 SRR12850849
## ENSG00000223972.5|DDX11L1          12           6           9           9
## ENSG00000237613.2|FAM138A           2           1           0           0
## ENSG00000268020.3|OR4G4P            0           0           0           0
## ENSG00000240361.2|OR4G11P           0           0           0           0
## ENSG00000186092.6|OR4F5             0           0           0           1
##                           SRR12850850 SRR12850851 SRR12850852 SRR12850853
## ENSG00000223972.5|DDX11L1          12          13          19           9
## ENSG00000237613.2|FAM138A           1           1           2           1
## ENSG00000268020.3|OR4G4P            0           0           0           0
## ENSG00000240361.2|OR4G11P           0           0           0           0
## ENSG00000186092.6|OR4F5             0           0           0           0
##                           SRR12850854 SRR12850855 SRR12850856 SRR12850857
## ENSG00000223972.5|DDX11L1           6          19           5          18
## ENSG00000237613.2|FAM138A           1           2           2           0
## ENSG00000268020.3|OR4G4P            0           2           0           1
## ENSG00000240361.2|OR4G11P           1           1           0           0
## ENSG00000186092.6|OR4F5             0           1           0           1
##                           SRR12850858 SRR12850859
## ENSG00000223972.5|DDX11L1           9           7
## ENSG00000237613.2|FAM138A           0           2
## ENSG00000268020.3|OR4G4P            1           1
## ENSG00000240361.2|OR4G11P           0           0
## ENSG00000186092.6|OR4F5             0           0
```

```
## [1] 58721    30
```

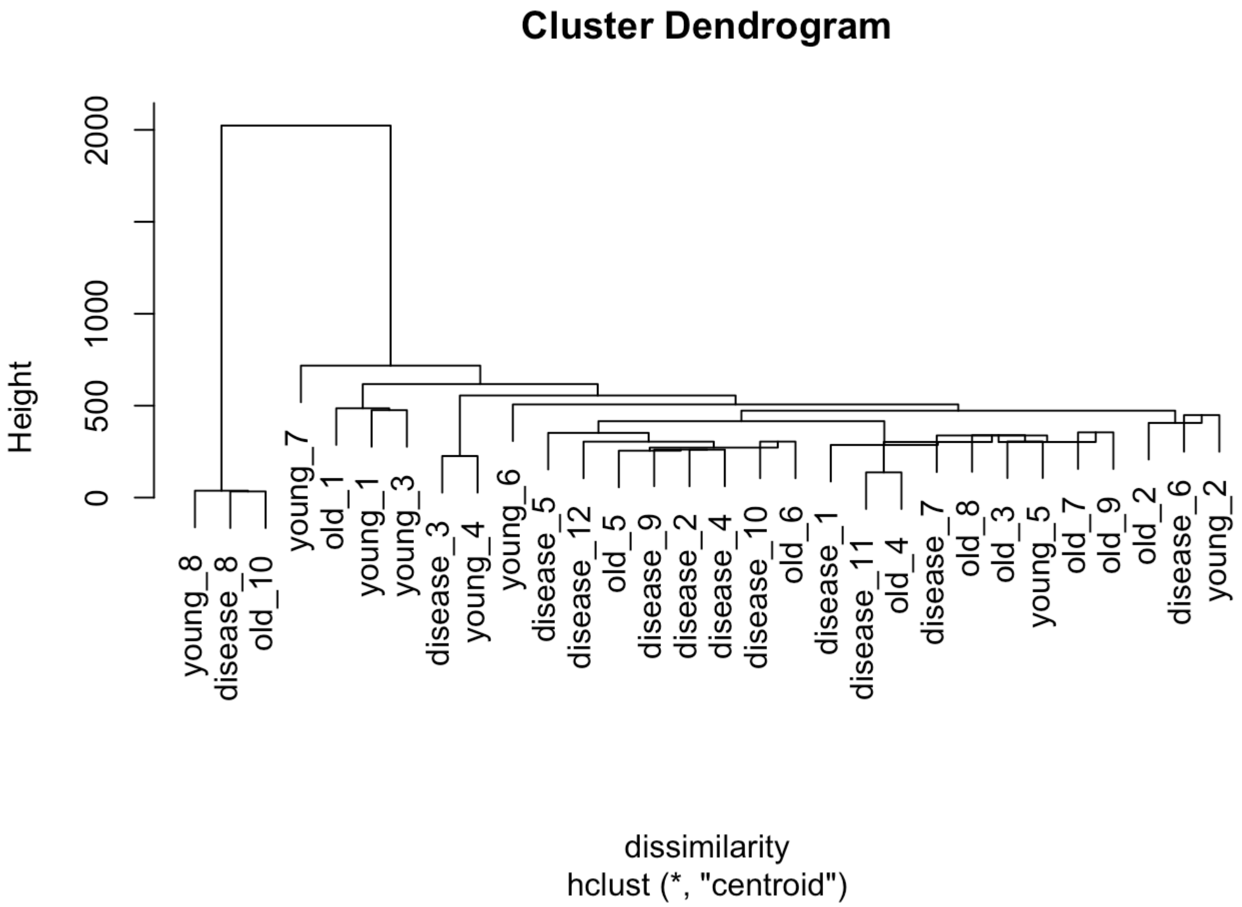
Read in sample metadata

```
##
## disease     old   young
##      12      10       8
```

To make the column names more informative we replace with metadata$condition column

Adding biological sex to the metadata file

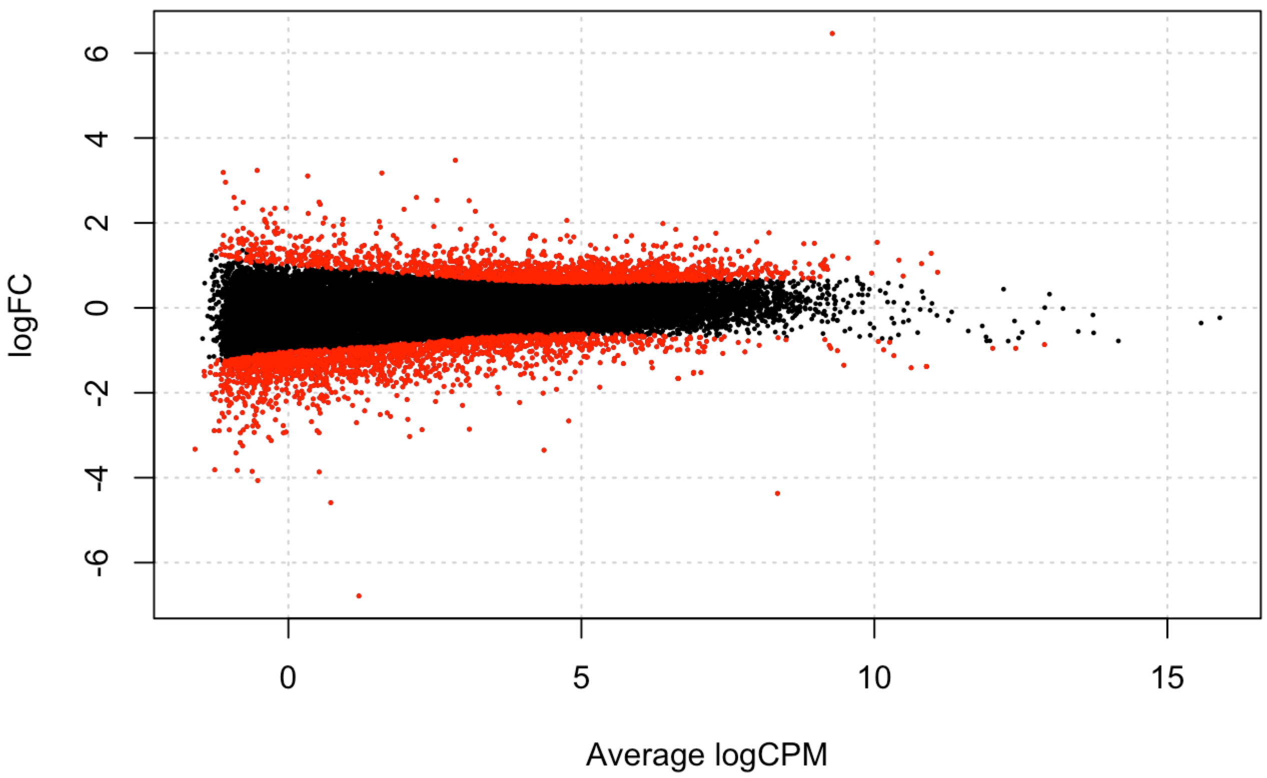
You may have noticed that information about the individuals age and sex is missing from the metadata. By looking at expression of genes on the X and Y chromosomes we can determine the biological sex of these samples. The method to infer sex from gene expression is a little complicated but we can return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



```
##  disease_1  disease_2  disease_3  disease_4  disease_5  disease_6  disease_7 
##          3          8          6          5          5          2          4 
##  disease_8  disease_9 disease_10 disease_11 disease_12      old_1      old_2 
##       8933          2          3          7          6          3         14 
##      old_3      old_4      old_5      old_6      old_7      old_8      old_9 
##          5          6          6          4          4          6          3 
##     old_10    young_1    young_2    young_3    young_4    young_5    young_6 
##      15427          2          2          5          3          5          2 
##    young_7    young_8 
##          5      11001
```

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
```

```
##        groupyoung
## Down         1849
## NotSig      21971
## Up           1681
##                                     gene     logFC    logCPM        LR
## ENSG00000283029.1|AL139099.4 AL139099.4  6.460340 9.2818270 425.09250
## ENSG00000211899.10|IGHM            IGHM -6.785327 1.2031781 156.76072
## ENSG00000257524.6|AL157935.2 AL157935.2 -3.351860 4.3621752 153.62791
## ENSG00000256148.1|AP000763.2 AP000763.2  3.474920 2.8493771 134.63680
## ENSG00000259001.3|AL355075.4 AL355075.4 -4.369389 8.3469720 127.08700
## ENSG00000086570.12|FAT2            FAT2 -2.663770 4.7834306 111.24513
## ENSG00000015520.14|NPC1L1        NPC1L1 -4.588257 0.7233283  96.28757
## ENSG00000274978.1|RNU11           RNU11 -2.857399 3.0883783  94.59137
## ENSG00000157005.3|SST               SST  2.523552 3.0832023  90.98477
## ENSG00000102317.17|RBM3            RBM3  2.057400 4.7517874  85.55373
##                                     PValue          FDR
## ENSG00000283029.1|AL139099.4 1.901163e-94 2.772494e-90
## ENSG00000211899.10|IGHM      5.773461e-36 4.209762e-32
## ENSG00000257524.6|AL157935.2 2.792837e-35 1.357612e-31
## ENSG00000256148.1|AP000763.2 3.964426e-31 1.445345e-27
## ENSG00000259001.3|AL355075.4 1.778056e-29 5.185930e-26
## ENSG00000086570.12|FAT2      5.228860e-26 1.270887e-22
## ENSG00000015520.14|NPC1L1    9.935187e-23 2.069804e-19
## ENSG00000274978.1|RNU11      2.340364e-22 4.266234e-19
## ENSG00000157005.3|SST        1.447820e-21 2.345976e-18
## ENSG00000102317.17|RBM3      2.254907e-20 3.288365e-17
```

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

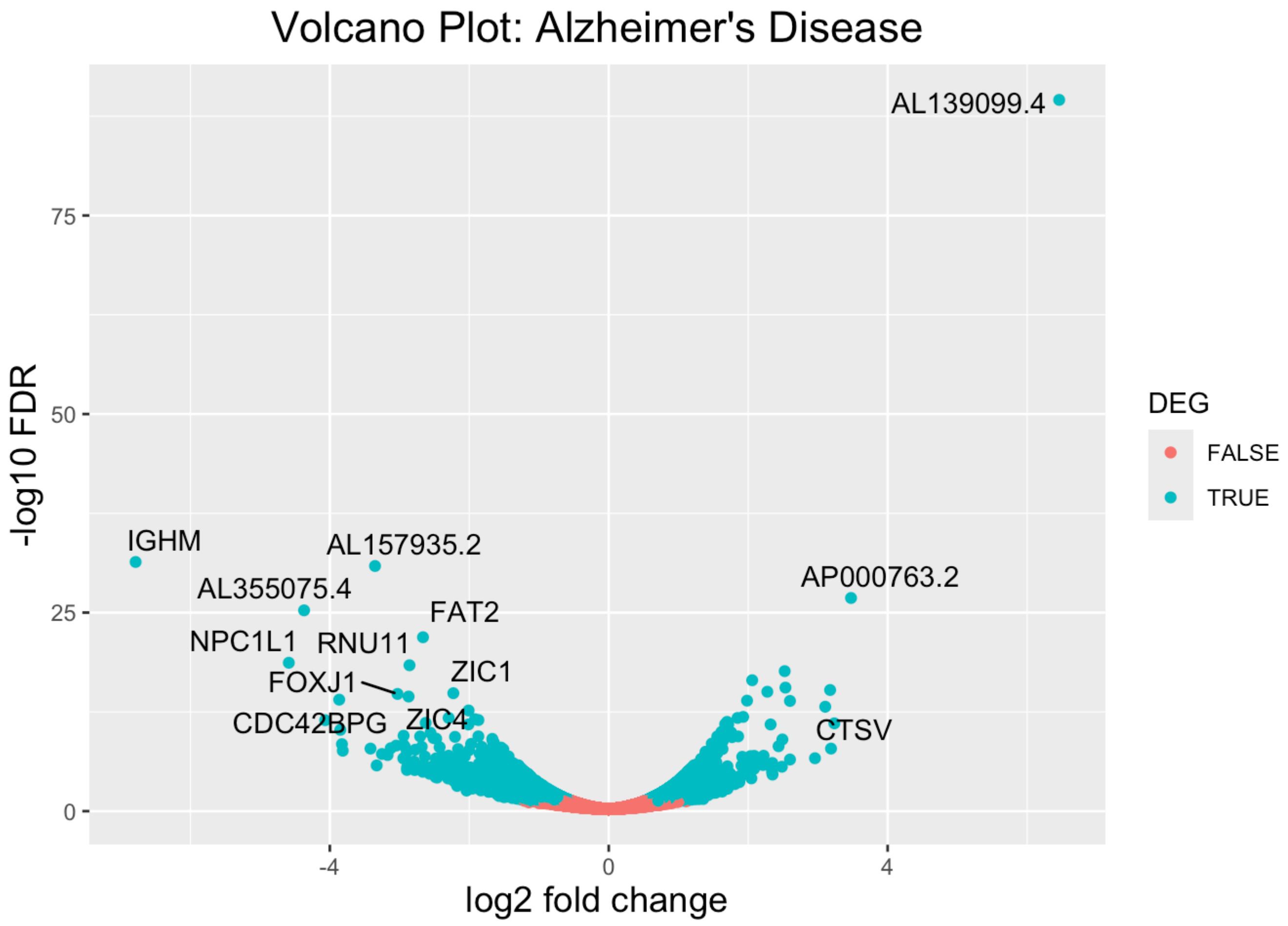
Save result file to working directory

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

Displaying results in volcano plot

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

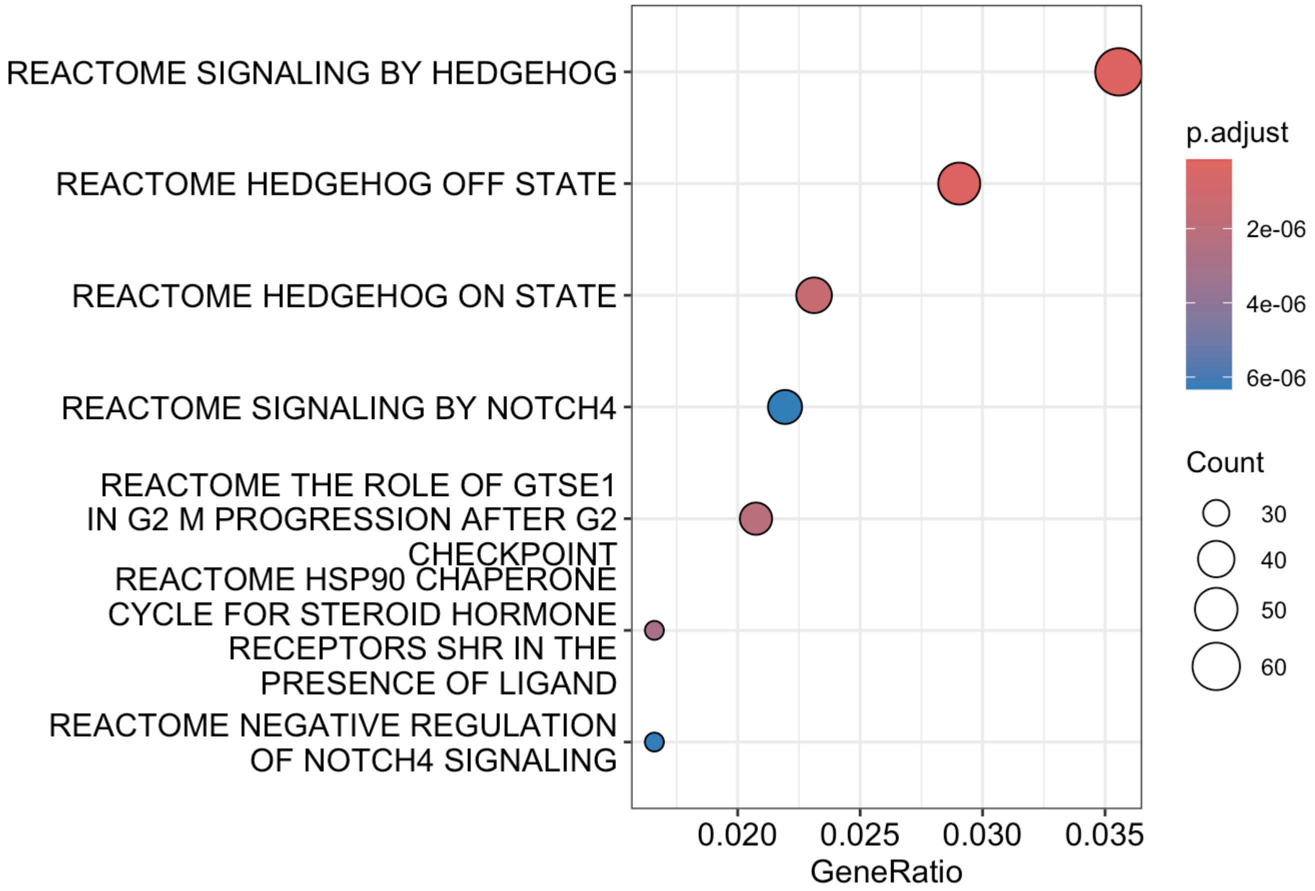
```
## Warning: ggrepel: 7 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



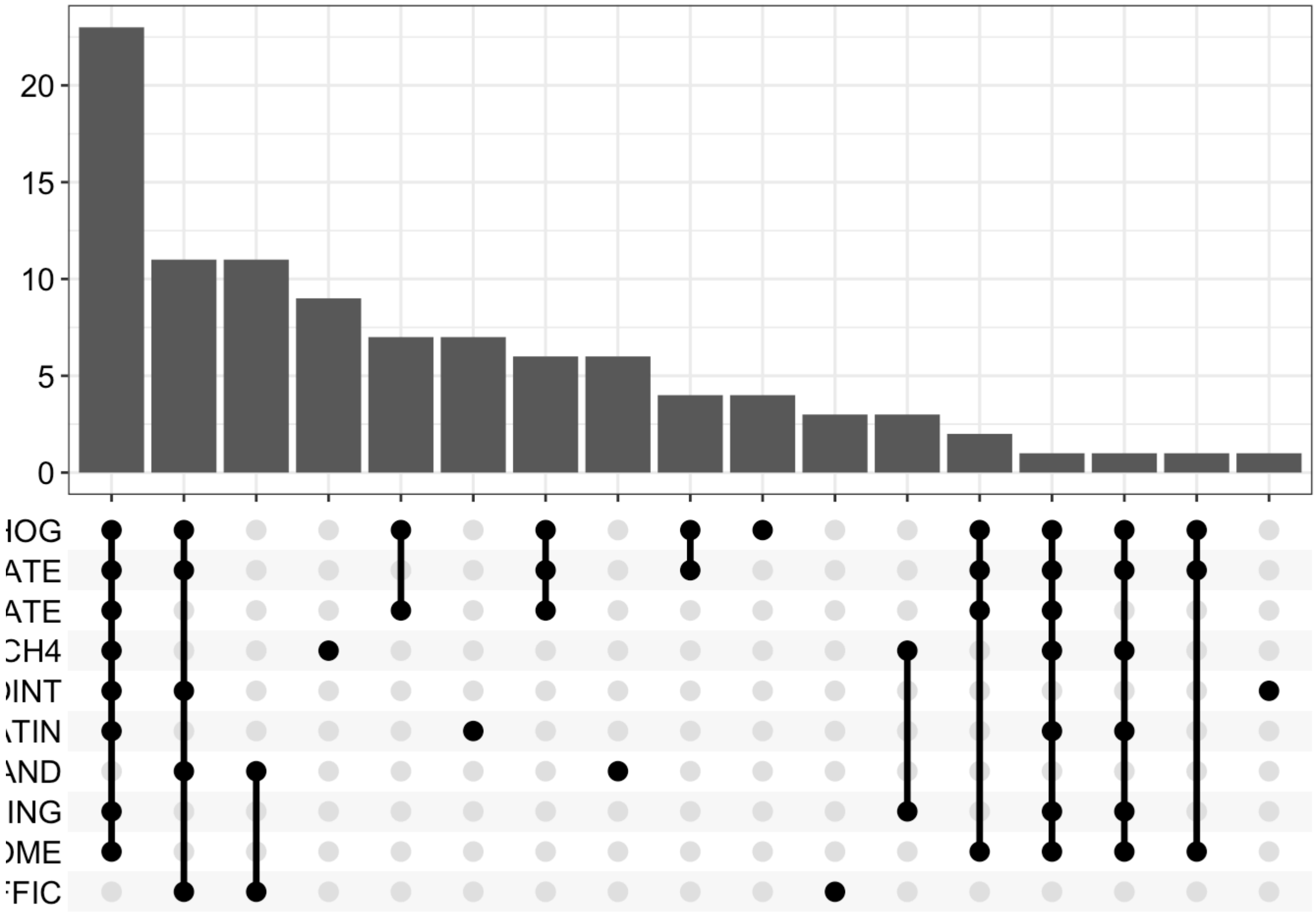
```
## Warning: ggrepel: 5 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

```
## 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,]  511.5594  2263.441
## [2,] 4189.4406 18536.559
```

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

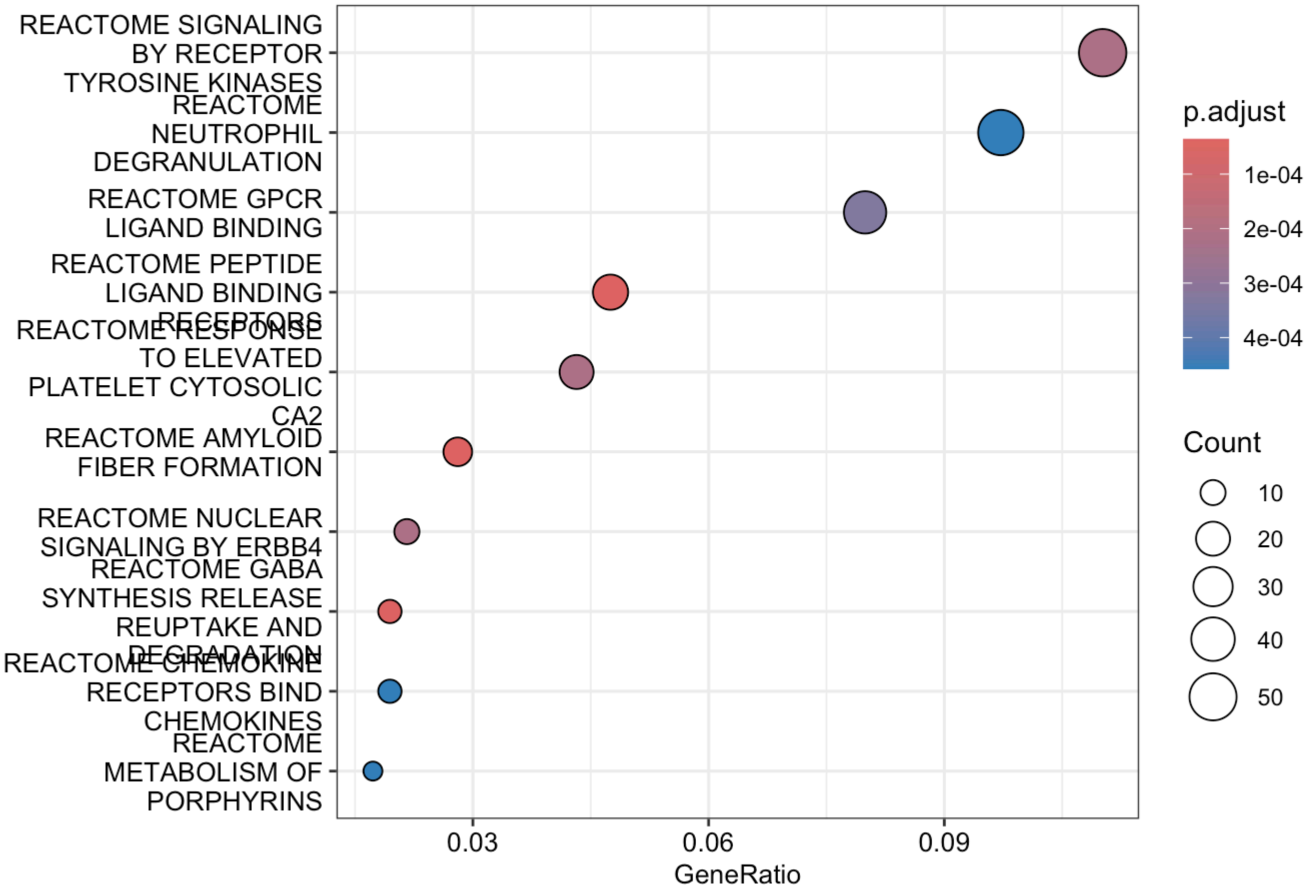
```
##      [,1]  [,2]
## [1,]  590  2185
## [2,] 4111 18615
```

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

```
##           [,1]       [,2]
## [1,]  3.468112 -1.6487569
## [2,] -1.211890  0.5761381
```

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

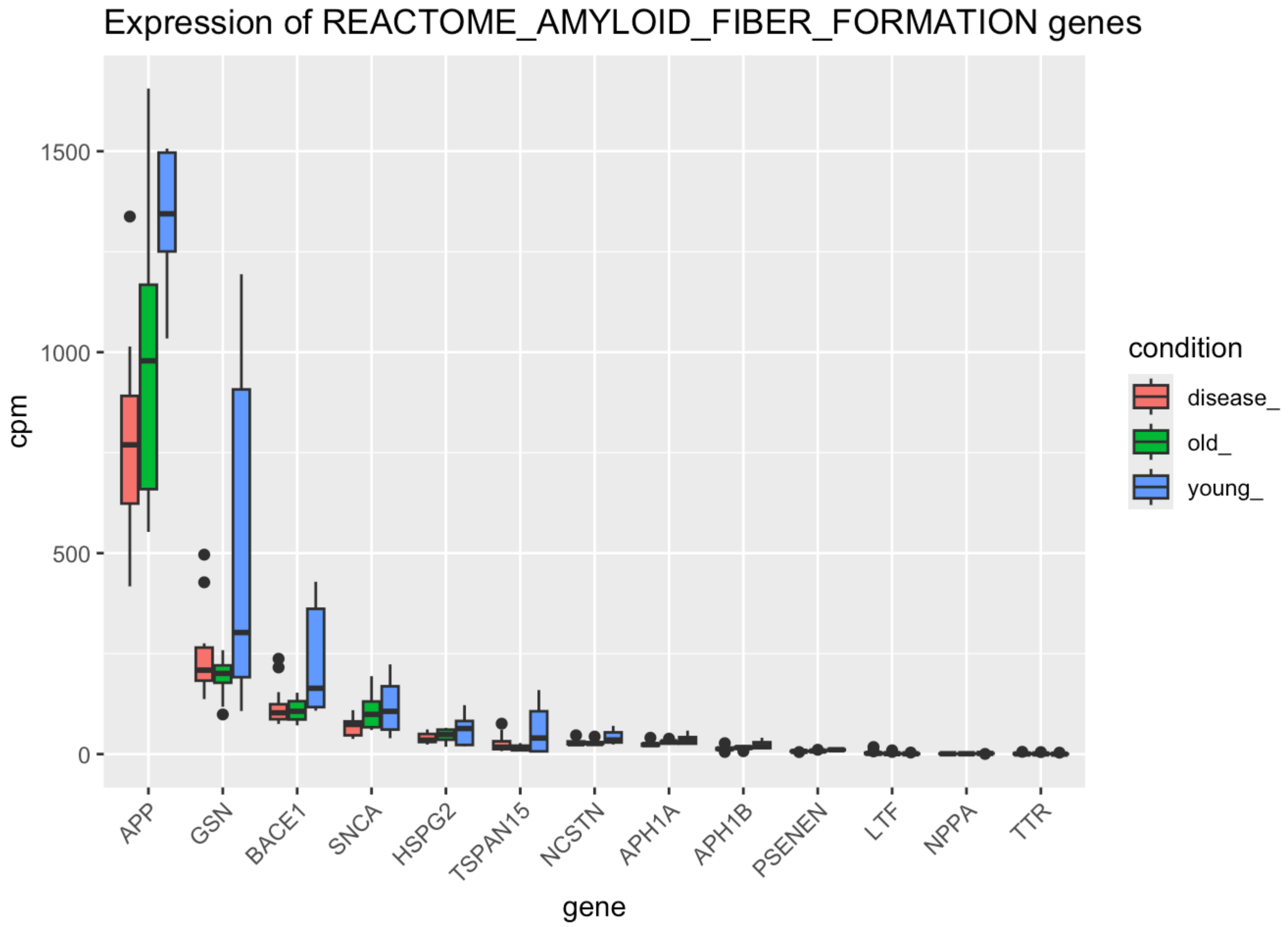
```
## [1] 5.303145e-05
```



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

Extract genes from interesting pathway. Select pathway with

pathway variable



```
## quartz_off_screen
##                 2
##                                 gene      logFC    logCPM        LR       PValue
## ENSG00000175206.10|NPPA      NPPA  0.9787114 0.3252886  7.835562 5.122821e-03
## ENSG00000142798.19|HSPG2    HSPG2  0.5780872 5.5701948  6.857265 8.828160e-03
## ENSG00000117362.12|APH1A    APH1A  0.5368878 4.8879038  5.987256 1.440960e-02
## ENSG00000162736.16|NCSTN    NCSTN  0.5342157 4.9973022  5.966429 1.458078e-02
## ENSG00000012223.12|LTF        LTF -1.7250087 1.3458818 24.103923 9.127418e-07
## ENSG00000145335.15|SNCA      SNCA  0.8413118 6.5459351 13.014926 3.090180e-04
## ENSG00000148180.19|GSN        GSN  1.0308845 8.2201263 17.012365 3.713719e-05
## ENSG00000099282.9|TSPAN15 TSPAN15  1.1242726 5.0017975 26.605707 2.495057e-07
## ENSG00000186318.16|BACE1    BACE1  0.9089784 7.1922119 14.427769 1.456389e-04
## ENSG00000138613.13|APH1B    APH1B  0.7061185 4.0497281  9.718081 1.824640e-03
## ENSG00000118271.10|TTR        TTR -1.1972130 0.2736753  9.360627 2.216963e-03
## ENSG00000205155.7|PSENEN   PSENEN  0.6991133 3.0468730  7.448005 6.350682e-03
## ENSG00000142192.20|APP        APP  0.8178337 9.9527894  9.387935 2.184181e-03
##                                    FDR threshold
## ENSG00000175206.10|NPPA   2.317586e-02      TRUE
## ENSG00000142798.19|HSPG2  3.370649e-02      TRUE
## ENSG00000117362.12|APH1A  4.655179e-02      TRUE
## ENSG00000162736.16|NCSTN  4.691828e-02      TRUE
## ENSG00000012223.12|LTF    3.770722e-05      TRUE
## ENSG00000145335.15|SNCA   3.277422e-03      TRUE
## ENSG00000148180.19|GSN    6.822812e-04      TRUE
## ENSG00000099282.9|TSPAN15 1.237612e-05      TRUE
## ENSG00000186318.16|BACE1  1.889567e-03      TRUE
## ENSG00000138613.13|APH1B  1.127500e-02      TRUE
## ENSG00000118271.10|TTR    1.299531e-02      TRUE
## ENSG00000205155.7|PSENEN  2.700873e-02      TRUE
## ENSG00000142192.20|APP    1.286124e-02      TRUE
```