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/HD"
```

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.

```
##                              SRR8866867 SRR8866869 SRR8866870 SRR8866871
## ENSG00000223972.5|DDX11L1             5          9         22         15
## ENSG00000237613.2|FAM138A             0          0          8         13
## ENSG00000240361.2|OR4G11P             0          0          0          0
## ENSG00000186092.6|OR4F5               0          0          0          0
## ENSG00000238009.6|AL627309.1         42         66         81         68
##                              SRR8866872 SRR8866873 SRR8866874 SRR8866875
## ENSG00000223972.5|DDX11L1            74         30          7          3
## ENSG00000237613.2|FAM138A            22          6          0          0
## ENSG00000240361.2|OR4G11P             0          0          0          0
## ENSG00000186092.6|OR4F5               0          0          0          0
## ENSG00000238009.6|AL627309.1         47         56         26         13
##                              SRR8866876 SRR8866877 SRR8866878 SRR8866879
## ENSG00000223972.5|DDX11L1            10         11          4         22
## ENSG00000237613.2|FAM138A             2          5          4          4
## ENSG00000240361.2|OR4G11P             0          0          0          0
## ENSG00000186092.6|OR4F5               0          0          0          0
## ENSG00000238009.6|AL627309.1         17         12         20         40
##                              SRR8866881
## ENSG00000223972.5|DDX11L1            22
## ENSG00000237613.2|FAM138A             7
## ENSG00000240361.2|OR4G11P             0
## ENSG00000186092.6|OR4F5               0
## ENSG00000238009.6|AL627309.1         32
```

```
## [1] 58721    13
```

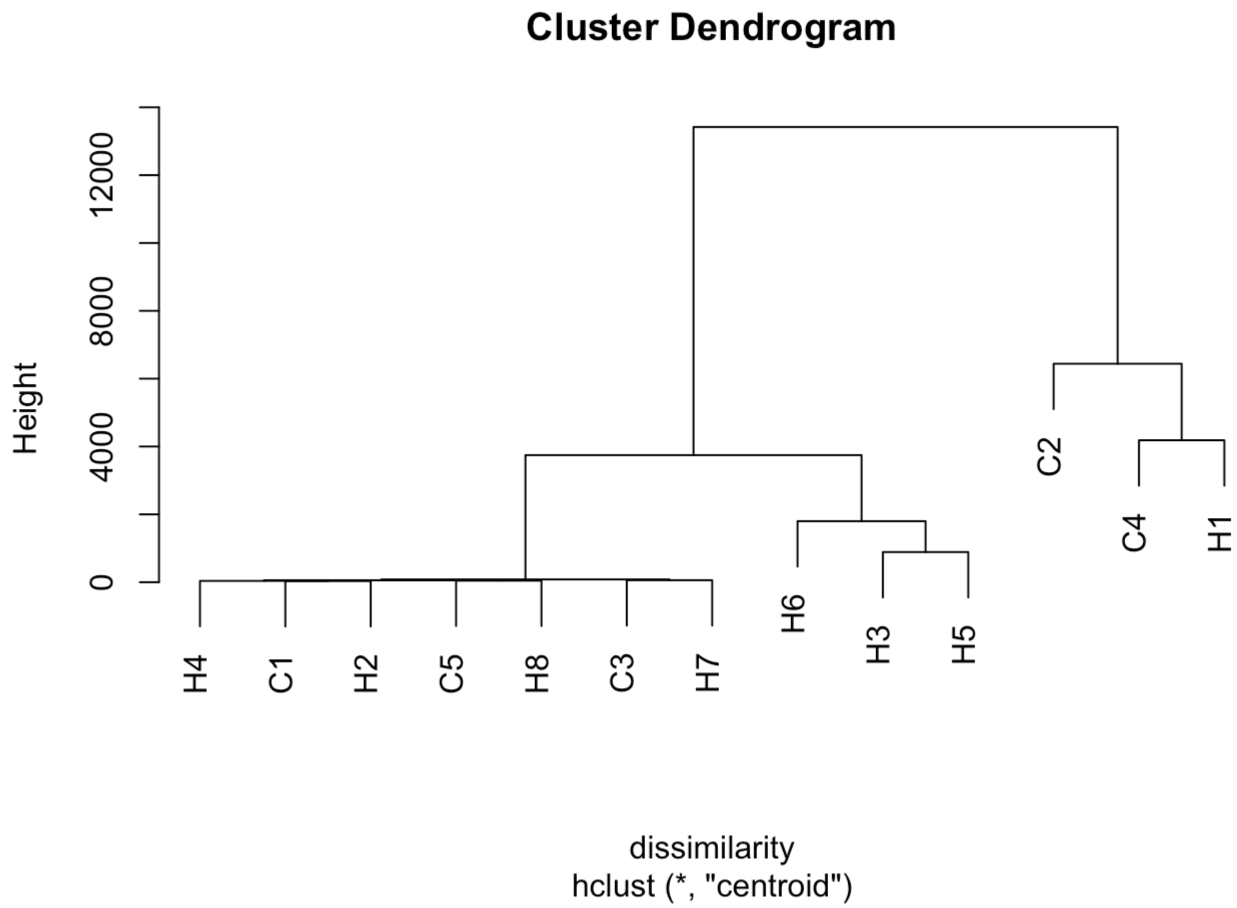
Read in sample metadata. This contains basic information about each sample.

The read.delim() command is similar to the read.csv() command but is more flexible because we can read in files which use tabs ’ which separate the columns. Here we replace the sample accession number with a more informative sample ID.

```
##           run condition
## 1  SRR8866867        C1
## 2  SRR8866869        C2
## 3  SRR8866870        C3
## 4  SRR8866871        C4
## 5  SRR8866872        C5
## 6  SRR8866873        H1
## 7  SRR8866874        H2
## 8  SRR8866875        H3
## 9  SRR8866876        H4
## 10 SRR8866877        H5
## 11 SRR8866878        H6
## 12 SRR8866879        H7
## 13 SRR8866881        H8
##                               C1 C2 C3 C4 C5 H1 H2 H3 H4 H5 H6 H7 H8
## ENSG00000223972.5|DDX11L1      5  9 22 15 74 30  7  3 10 11  4 22 22
## ENSG00000237613.2|FAM138A      0  0  8 13 22  6  0  0  2  5  4  4  7
## ENSG00000240361.2|OR4G11P      0  0  0  0  0  0  0  0  0  0  0  0  0
## ENSG00000186092.6|OR4F5        0  0  0  0  0  0  0  0  0  0  0  0  0
## ENSG00000238009.6|AL627309.1 42 66 81 68 47 56 26 13 17 12 20 40 32
```

Adding biological sex to the metadata file

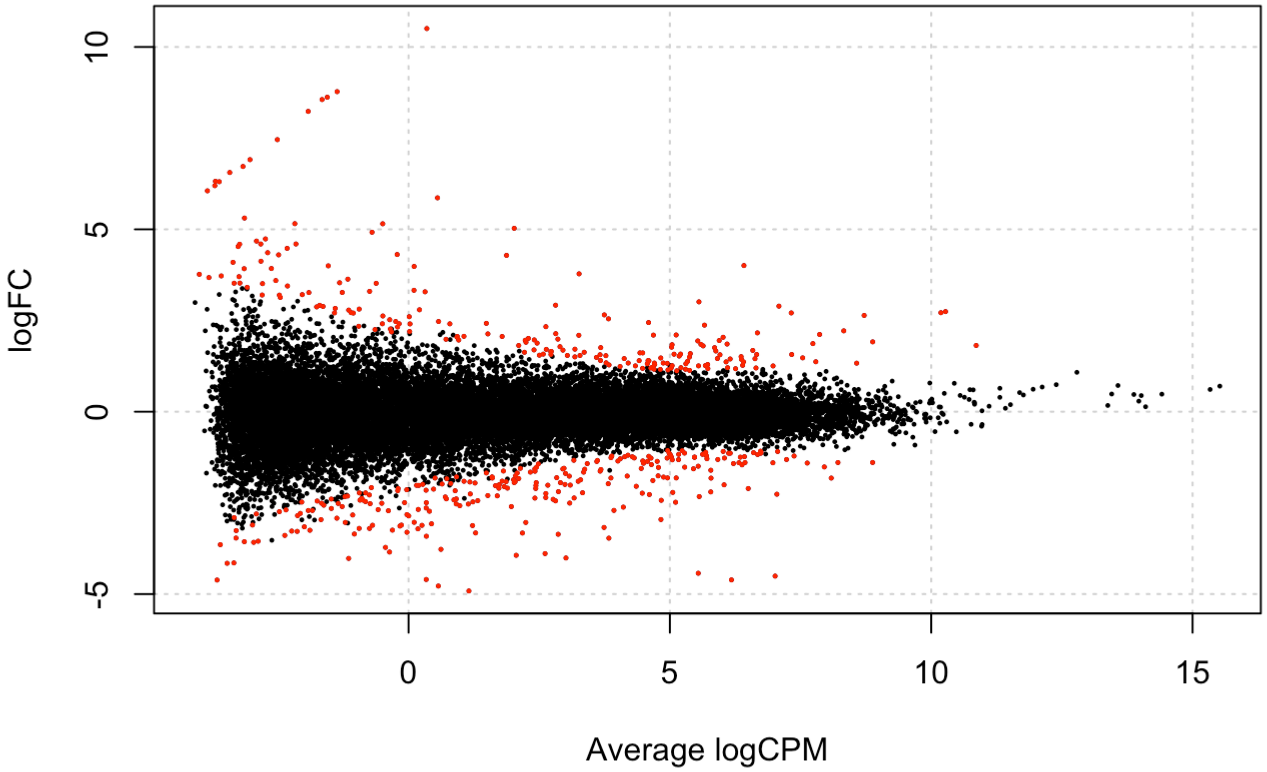
You may have noticed that information about the individuals age and sex is missing from the metadata. By looking at expression of genes on the X and Y chromosomes we can determine the biological sex of these samples. The method to infer sex from gene expression is a little complicated but we can return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



```
##     C1     C2     C3     C4     C5     H1     H2     H3     H4     H5     H6 
##  86598      0 132479      3  47346      6  40326      4  11020      0      0 
##     H7     H8 
##  63411  47923 
```

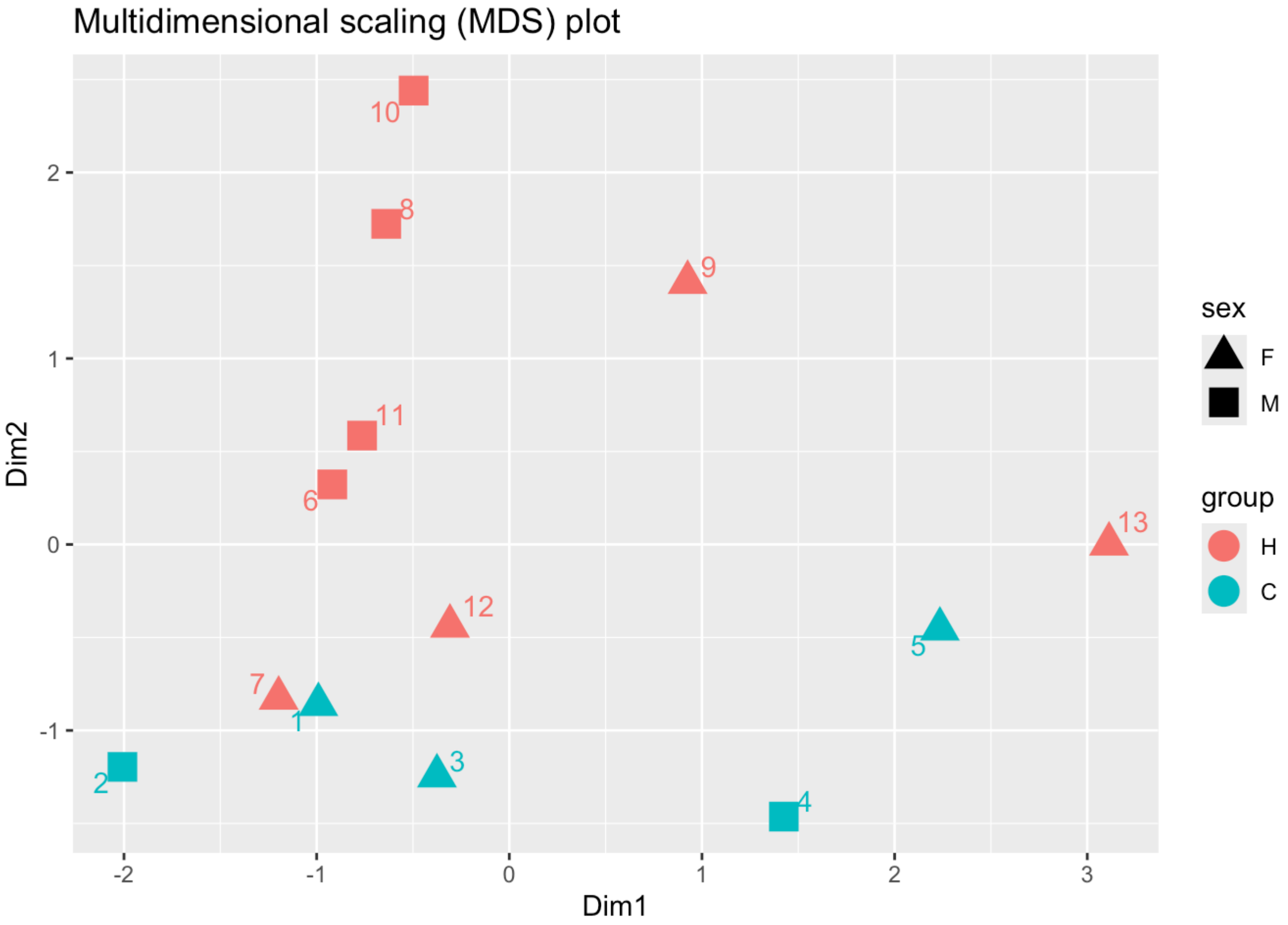
Differential expression analysis with edgeR likelihood ratio test

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
```

```
##        groupH
## Down      261
## NotSig  25786
## Up        211
##                                    gene     logFC    logCPM        LR
## ENSG00000181195.10|PENK            PENK -4.611021 6.1780918 134.07683
## ENSG00000183379.8|SYNDIG1L     SYNDIG1L -4.506622 7.0140572 102.97756
## ENSG00000280064.1|AC130304.1 AC130304.1 10.505464 0.3493158  99.61570
## ENSG00000173110.7|HSPA6           HSPA6  4.007187 6.4170369  92.51951
## ENSG00000197261.11|C6orf141    C6orf141 -4.009607 3.0114464  88.24125
## ENSG00000147246.9|HTR2C           HTR2C -3.466872 3.8321329  75.03502
## ENSG00000115155.17|OTOF            OTOF -3.173695 3.7402248  72.62305
## ENSG00000135245.9|HILPDA         HILPDA  3.012459 5.5558962  68.10858
## ENSG00000159167.11|STC1            STC1  3.780388 3.2625576  67.88557
## ENSG00000285238.2|AC006064.6 AC006064.6  5.028363 2.0209652  67.30227
##                                    PValue          FDR
## ENSG00000181195.10|PENK      5.256171e-31 1.364444e-26
## ENSG00000183379.8|SYNDIG1L   3.389663e-24 4.399594e-20
## ENSG00000280064.1|AC130304.1 1.850314e-23 1.601070e-19
## ENSG00000173110.7|HSPA6      6.666423e-22 4.326324e-18
## ENSG00000197261.11|C6orf141  5.793894e-21 3.008061e-17
## ENSG00000147246.9|HTR2C      4.624386e-18 2.000732e-14
## ENSG00000115155.17|OTOF      1.569351e-17 5.819803e-14
## ENSG00000135245.9|HILPDA     1.547363e-16 4.997499e-13
## ENSG00000159167.11|STC1      1.732643e-16 4.997499e-13
## ENSG00000285238.2|AC006064.6 2.329136e-16 6.046178e-13
```


Multidimensional scaling (MDS) plot

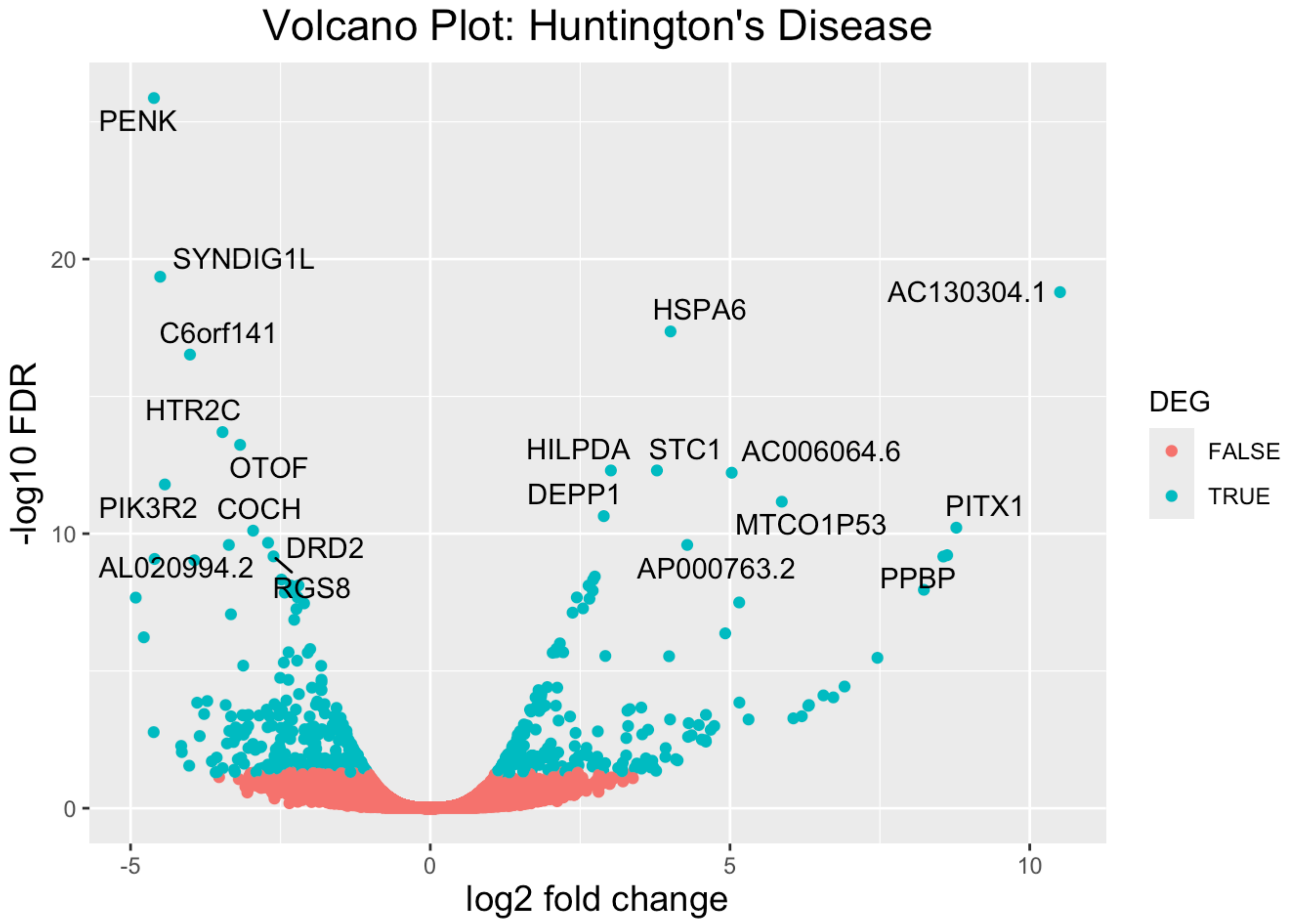
```
## quartz_off_screen
##                 2
```

Save result file to working directory

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

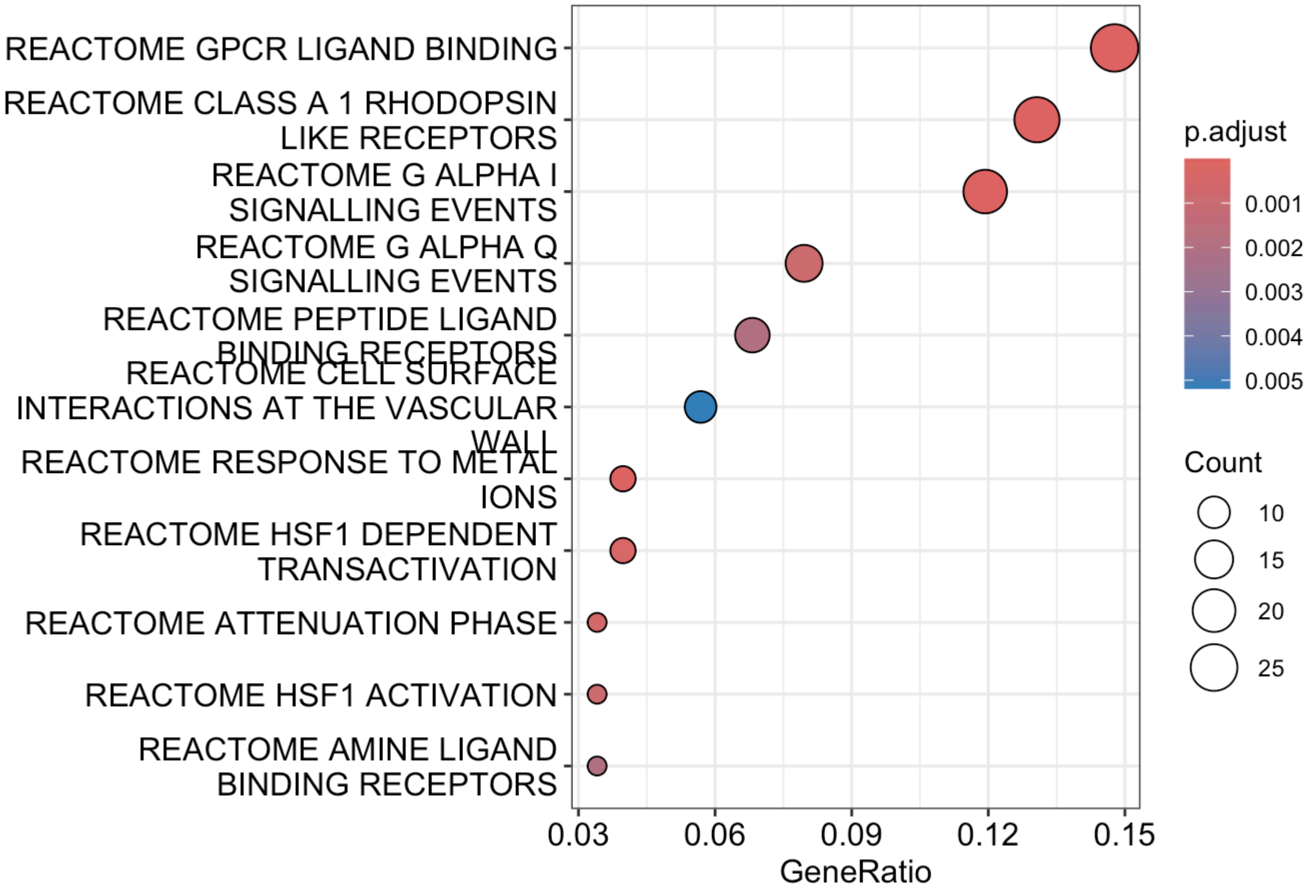
Displaying results in volcano plot

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.

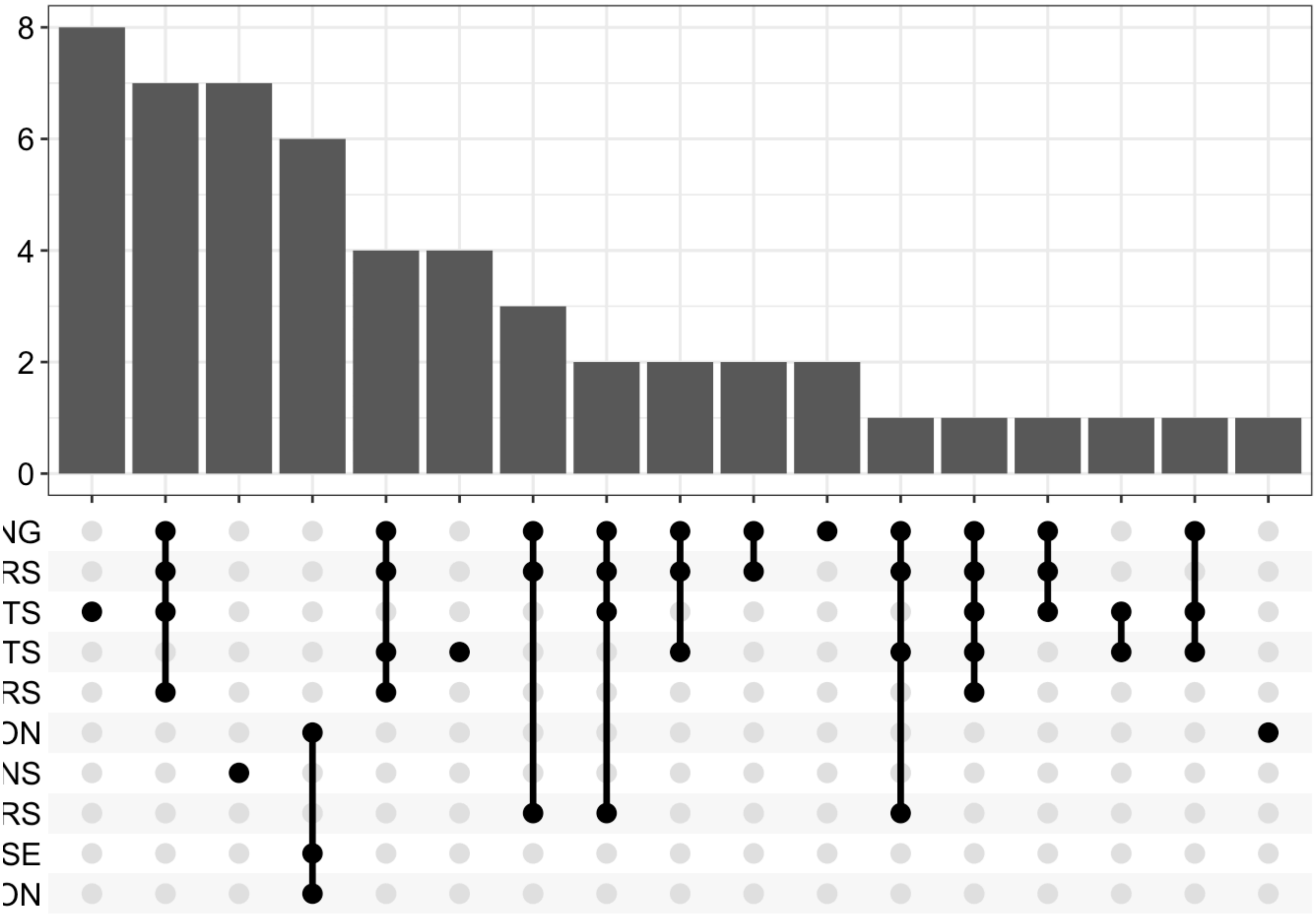


```
## quartz_off_screen
##                  2
```

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,]  15.39436   835.6056
## [2,] 459.60564 24947.3944
```

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

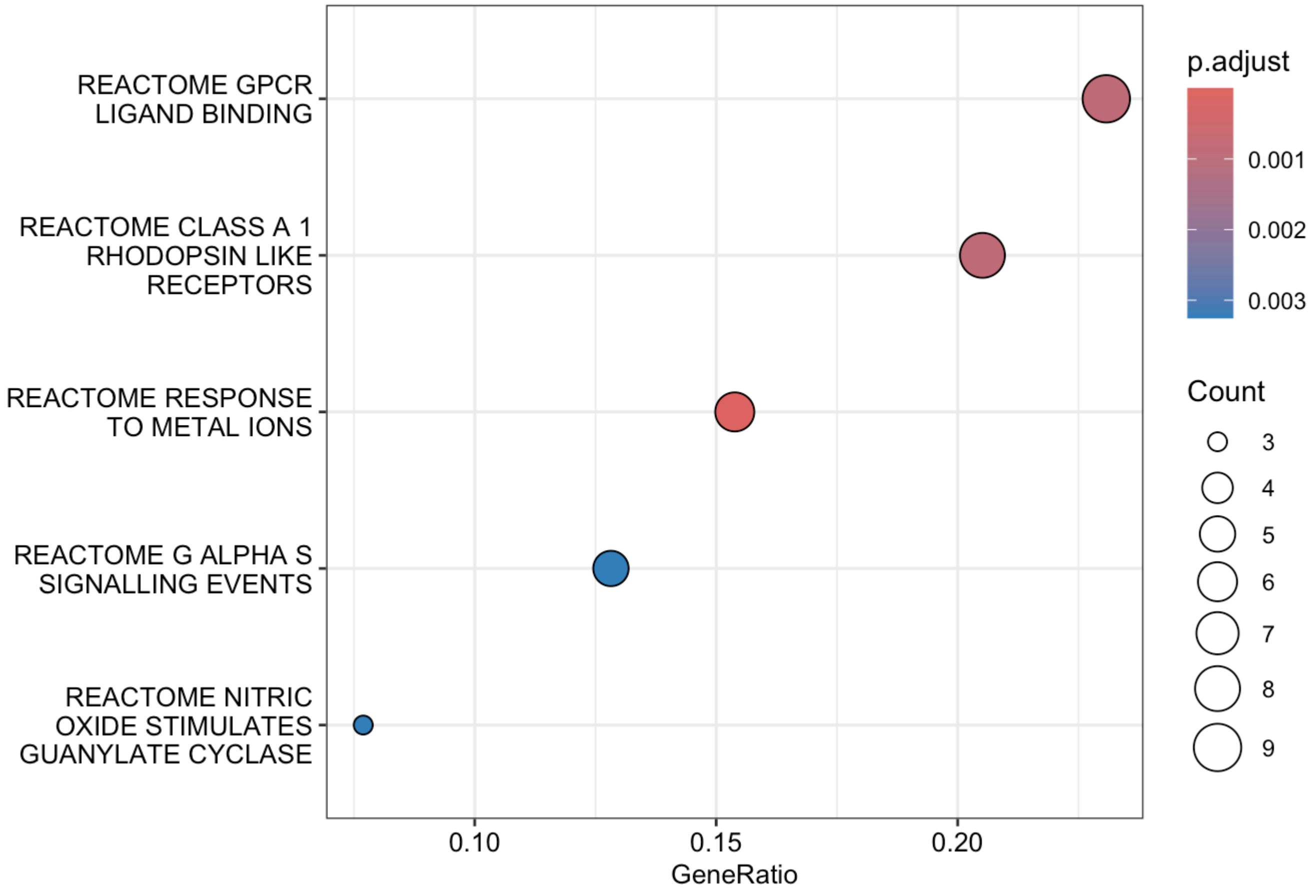
```
##      [,1]  [,2]
## [1,]   50   801
## [2,]  425 24982
```

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

```
##           [,1]       [,2]
## [1,]  8.819951 -1.1971436
## [2,] -1.614189  0.2190959
```

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

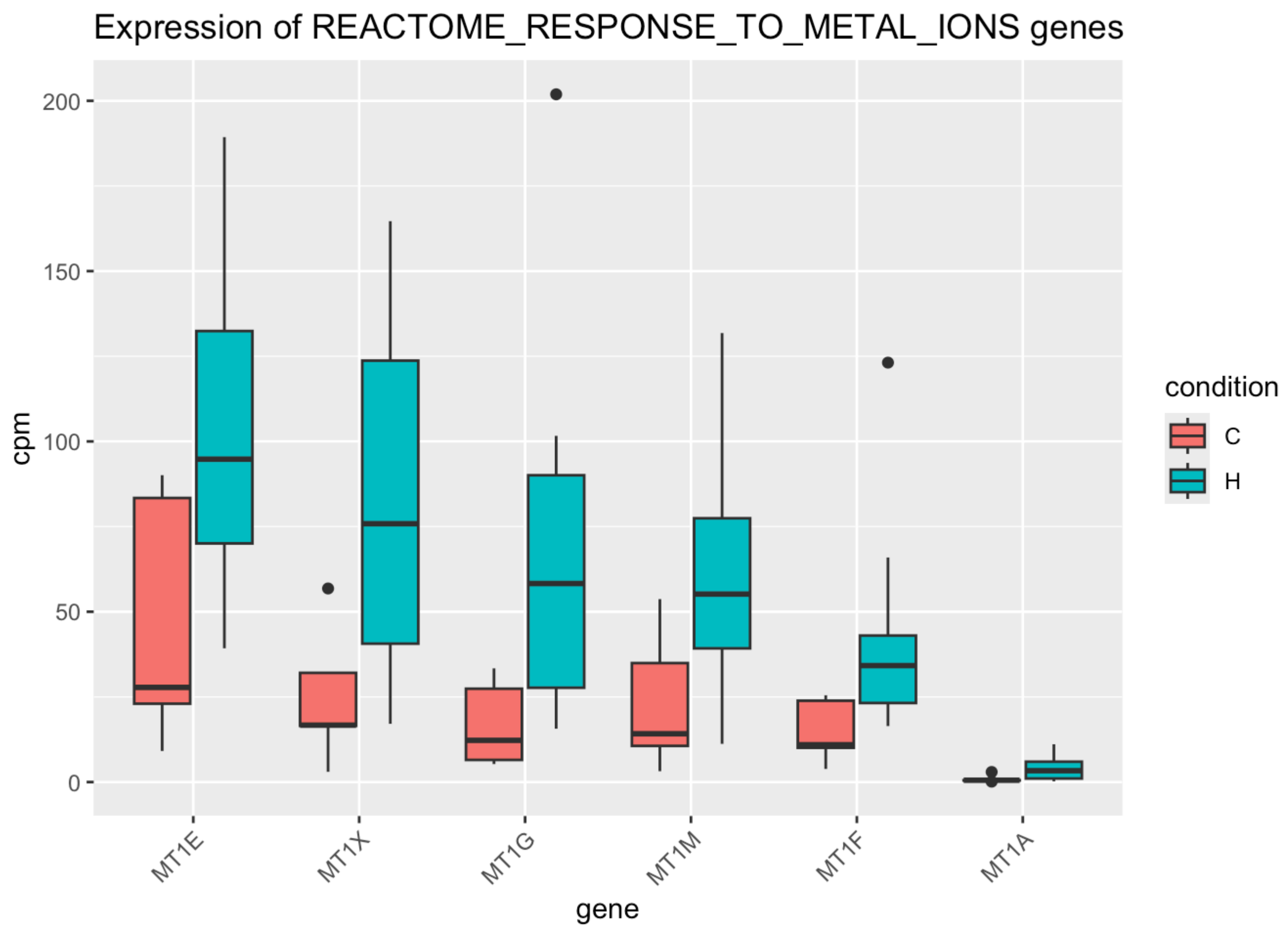
```
## [1] 4.751146e-19
```



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

Extract genes from interesting pathway. Select pathway with

pathway variable



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

```
##                           gene    logFC   logCPM       LR       PValue
## ENSG00000125144.13|MT1G MT1G 2.373315 5.661814 41.43463 1.218792e-10
## ENSG00000169715.14|MT1E MT1E 1.435172 6.369217 16.23232 5.603014e-05
## ENSG00000205364.3|MT1M  MT1M 1.846728 5.593146 26.52880 2.596389e-07
## ENSG00000205362.11|MT1A MT1A 2.422530 1.488122 19.13133 1.220251e-05
## ENSG00000198417.6|MT1F  MT1F 1.769107 5.032884 24.56387 7.188713e-07
## ENSG00000187193.8|MT1X  MT1X 1.951544 5.944082 28.36844 1.002870e-07
##                                  FDR threshold
## ENSG00000125144.13|MT1G 7.532972e-08      TRUE
## ENSG00000169715.14|MT1E 6.323827e-03      TRUE
## ENSG00000205364.3|MT1M  8.753165e-05      TRUE
## ENSG00000205362.11|MT1A 1.810077e-03      TRUE
## ENSG00000198417.6|MT1F  2.006570e-04      TRUE
## ENSG00000187193.8|MT1X  3.885582e-05      TRUE
```