# EMMIX-gene

*Andrew Jones*

*2018-03-28*

## EMMIX-gene

The aim of this document is to reproduce the results of the original EMMIX-gene paper (McLachlan et al. (2002)) while demonstrating the EMMIX-gene R-package workflow.

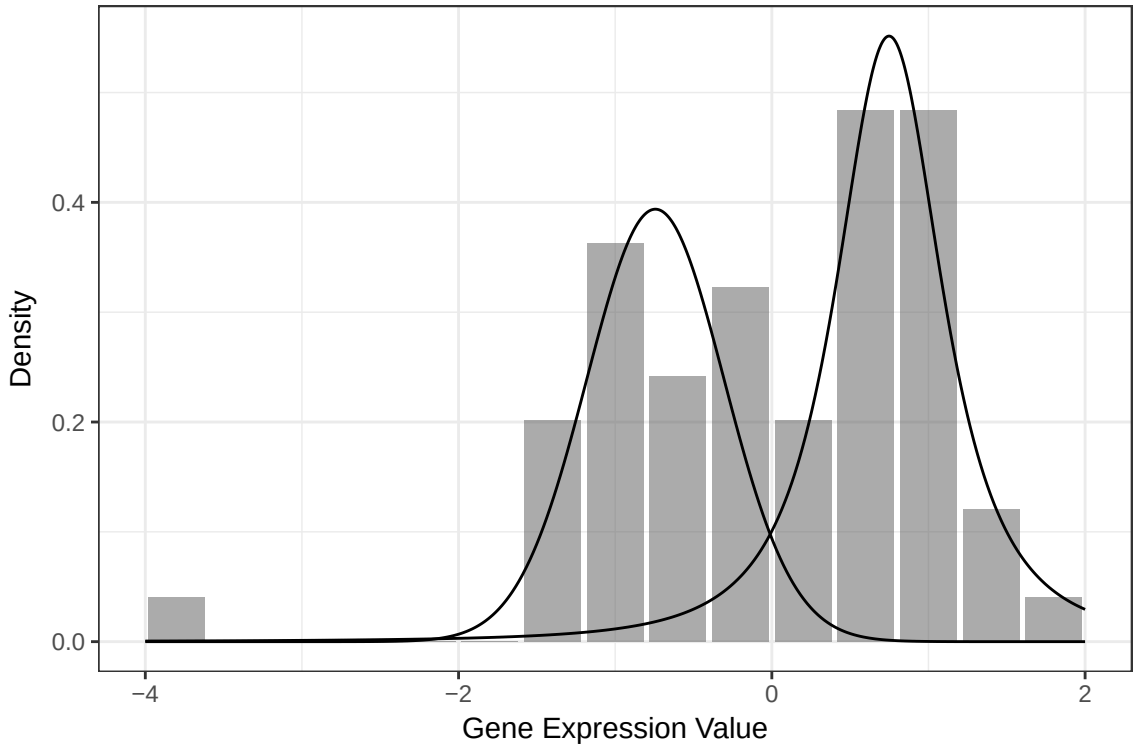
### Data

The original paper analysed two data sets. The Golub Data was a dataset containing the centred and normalised values of the logged expression values of a subset of 3731 genes taken from Golub et al. (1999). The Alon data was a dataset containing the centred and normalised values of the logged expression values of a subset of 2000 genes taken from Alon et al. (1999). The method of subset selection for both datasets follows that of their original papers and is described in McLachlan et al. (2002). The raw Alon and Golub data are also available in the 'colonCA' and 'golubEsets' Bioconductor packages respectively.

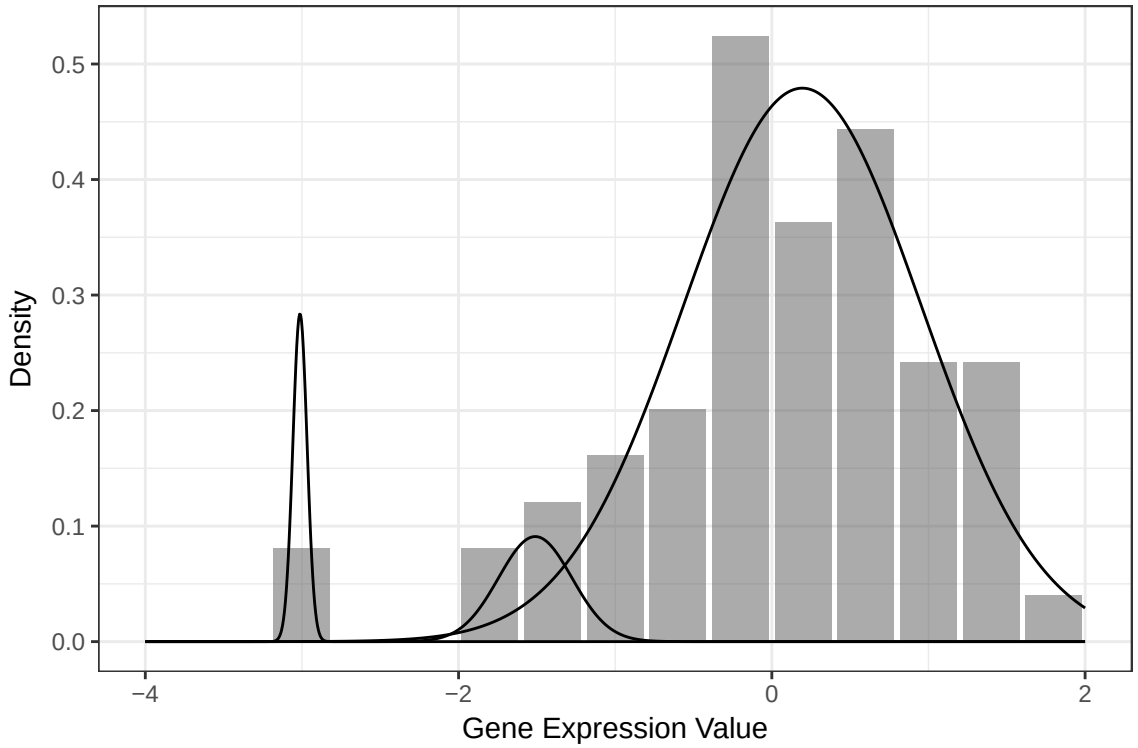
### Single Genes

Figures 1 to 3 in (McLachlan et al. (2002)) are reproduced here using the function plot_single_gene().

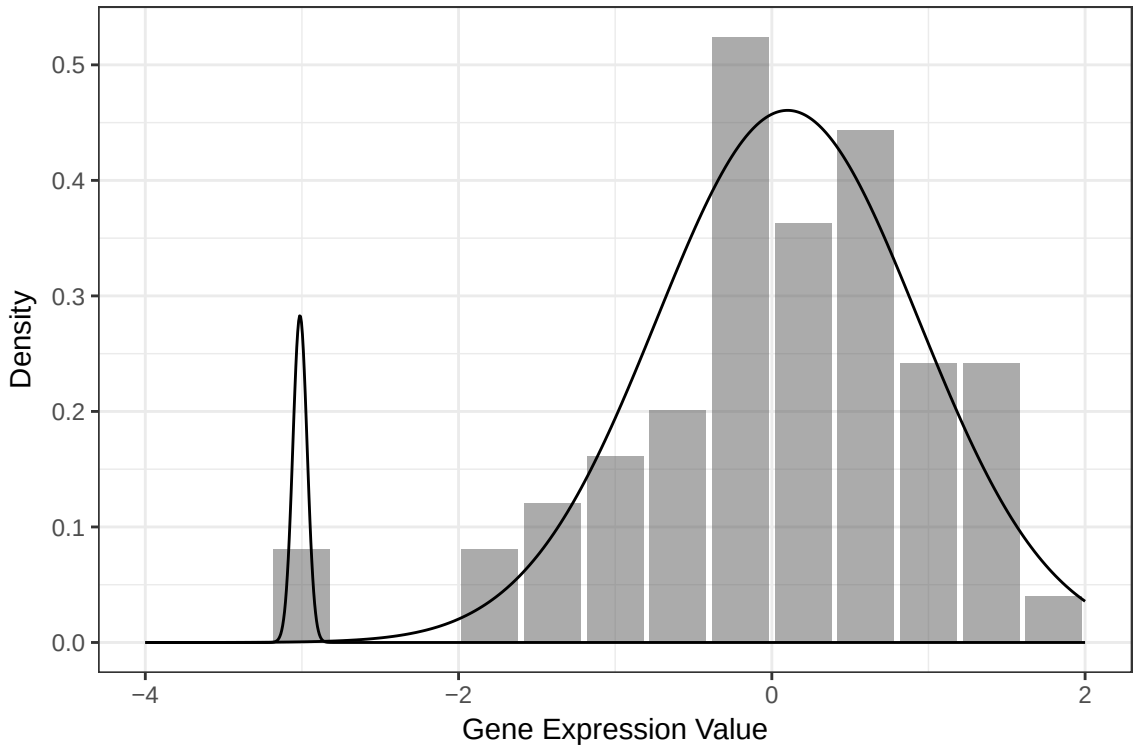
```
plot_single_gene(alon_data,1758)
```



```
plot_single_gene(alon_data,474,g=3)
```



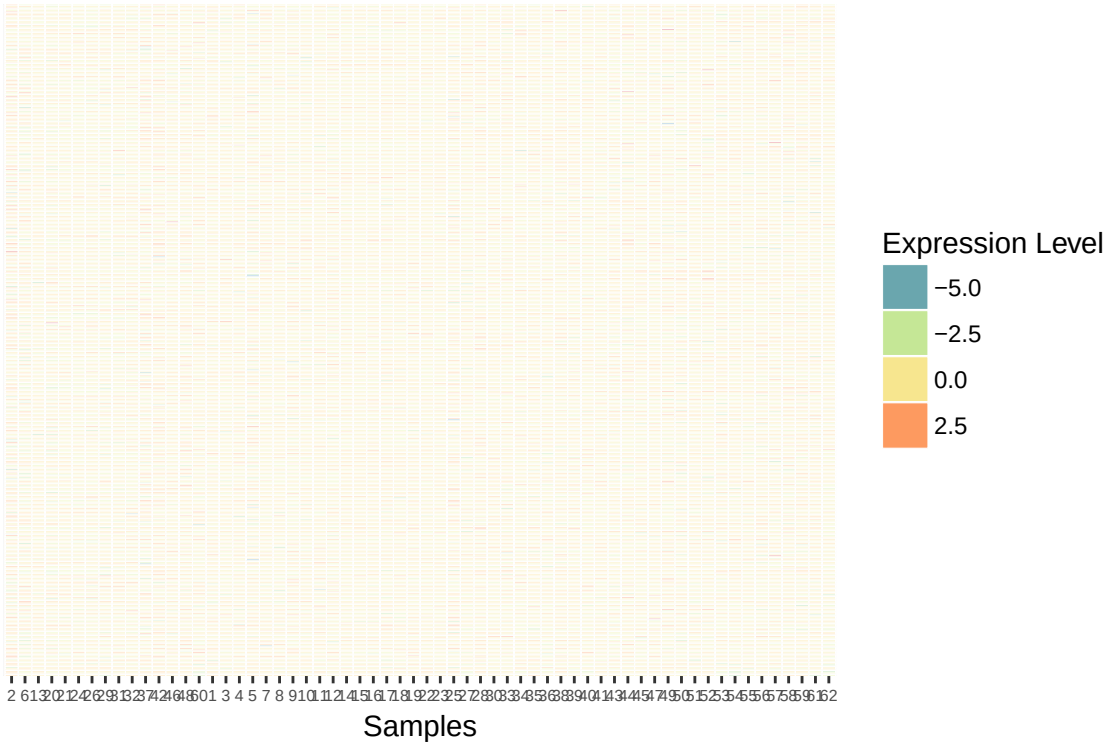

```
plot_single_gene(alon_data, 474, g=2)
```

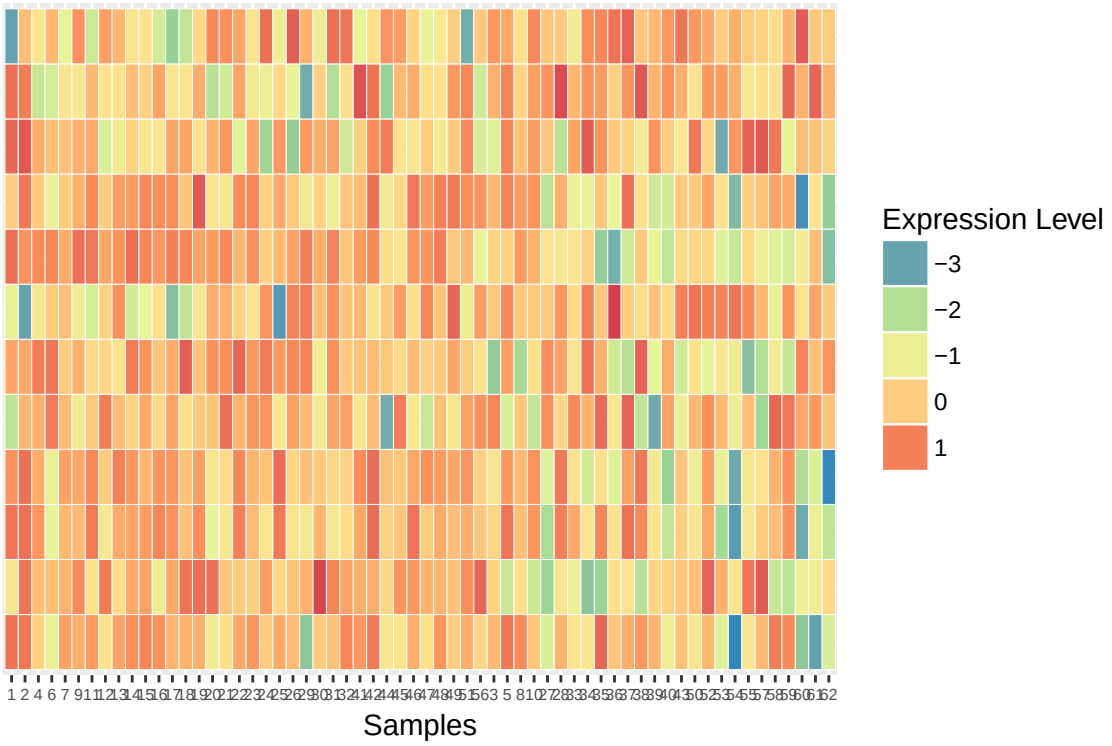


## Select Genes

The method of selecting genes isas in section 5.1 in (McLachlan et al. (2002)) is performed using the function select_genes(). We find that approximately 500 genes were selected, which compares well to 446 genes in the original analysis.

The top genes can then be clustered using mixtures of common factor analysers.






#5.1.2 The tissues can then also be clustered using either mixtures of t distributions or mixtures of common factor analysers.

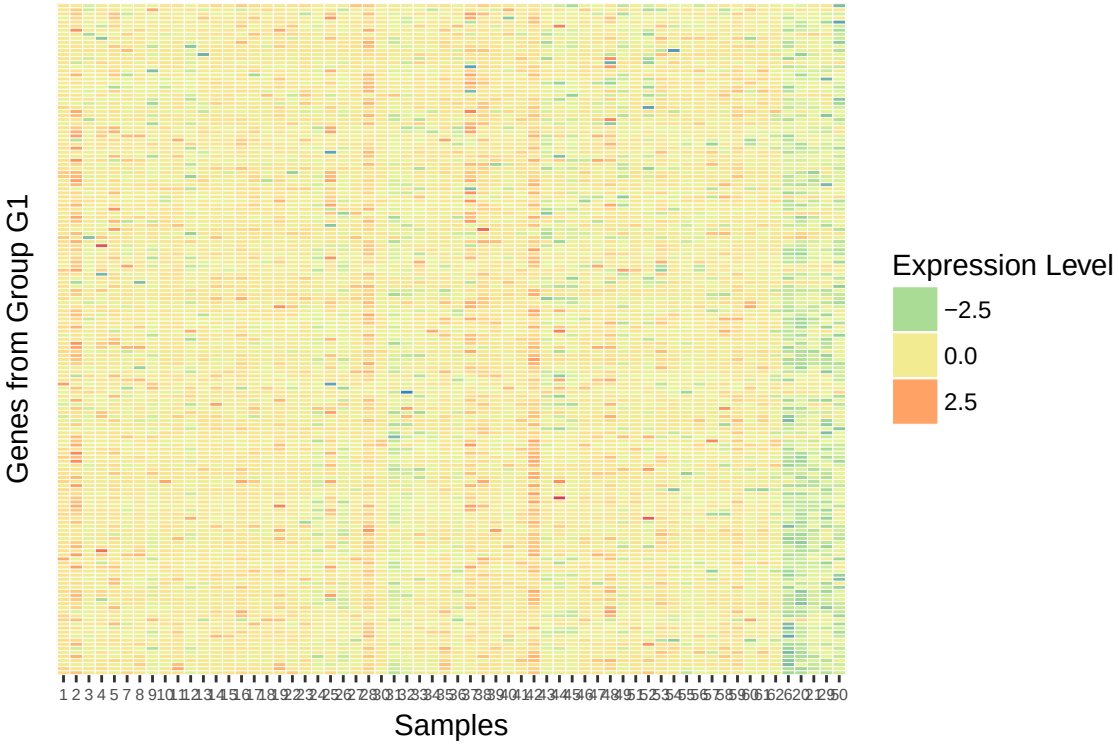
```
## C1:  1 2 3 4 5 7 8 9 10 11 13 14 15 16 17 18 19 22 23 25 27 30 33 38 41 42
## 44 45 46 47 48 49 51 52 56 57 58 61
```

```
## C2:  6 12 20 21 24 26 28 29 31 32 34 35 36 37 39 40 43 50 53 54 55 59 60
## 62
```

Heat maps similar to those in McLachlan et al. (2002) can also be produced for each of the groups.

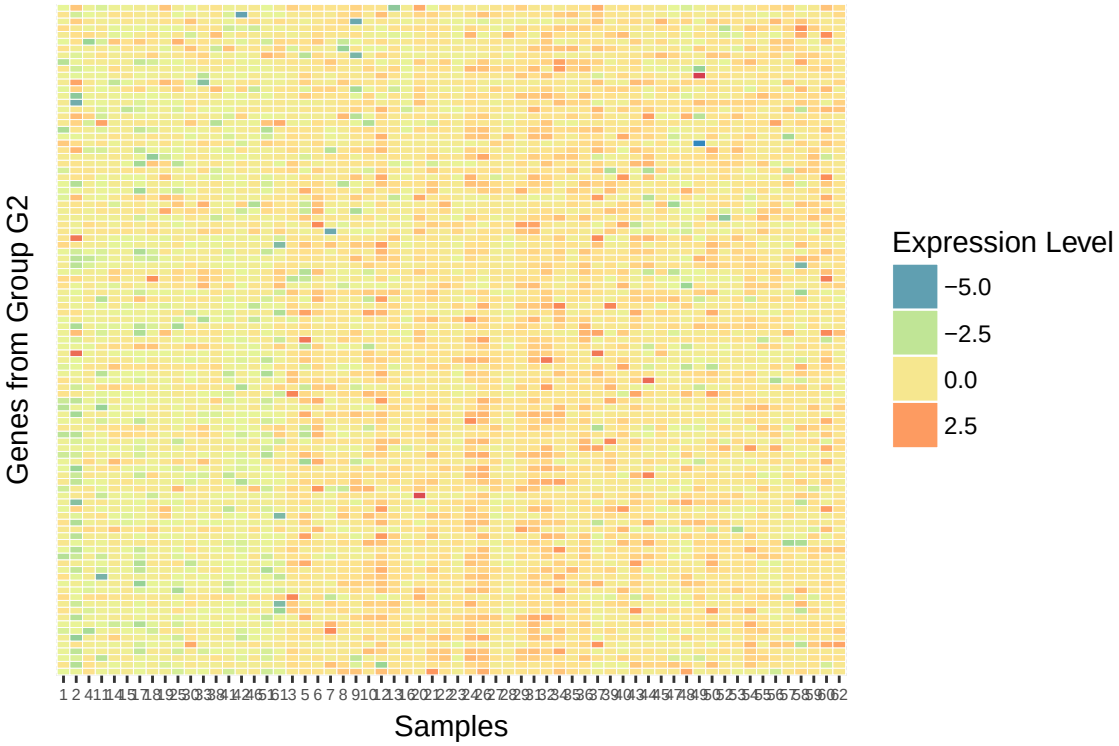
```
## C1:  6 20 21 29 50
```

```
## C2:  1 2 3 4 5 7 8 9 10 11 12 13 14 15 16 17 18 19 22 23 24 25 26 27 28 30
## 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 51 52 53 54 55 56
## 57 58 59 60 61 62
```



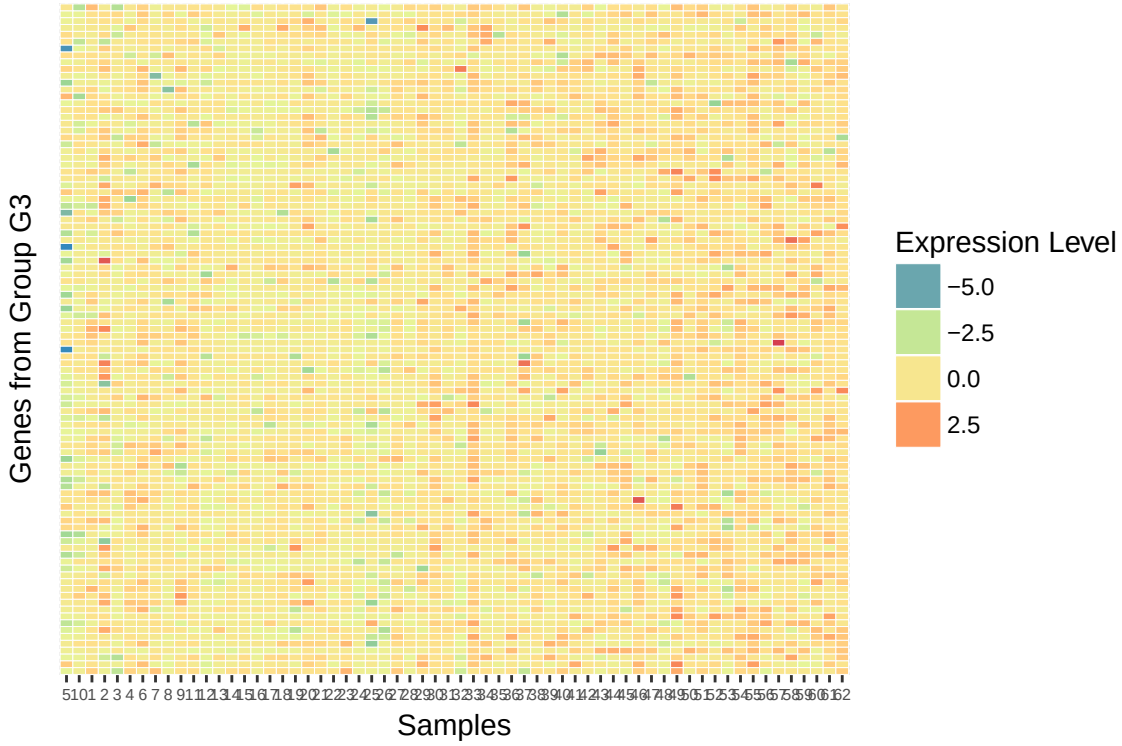

```
## C1:  3 5 6 7 8 9 10 12 13 16 20 21 22 23 24 26 27 28 29 31 32 34 35 36 37
## 39 40 43 44 45 47 48 49 50 52 53 54 55 56 57 58 59 60 62
```

```
## C2:  1 2 4 11 14 15 17 18 19 25 30 33 38 41 42 46 51 61
```

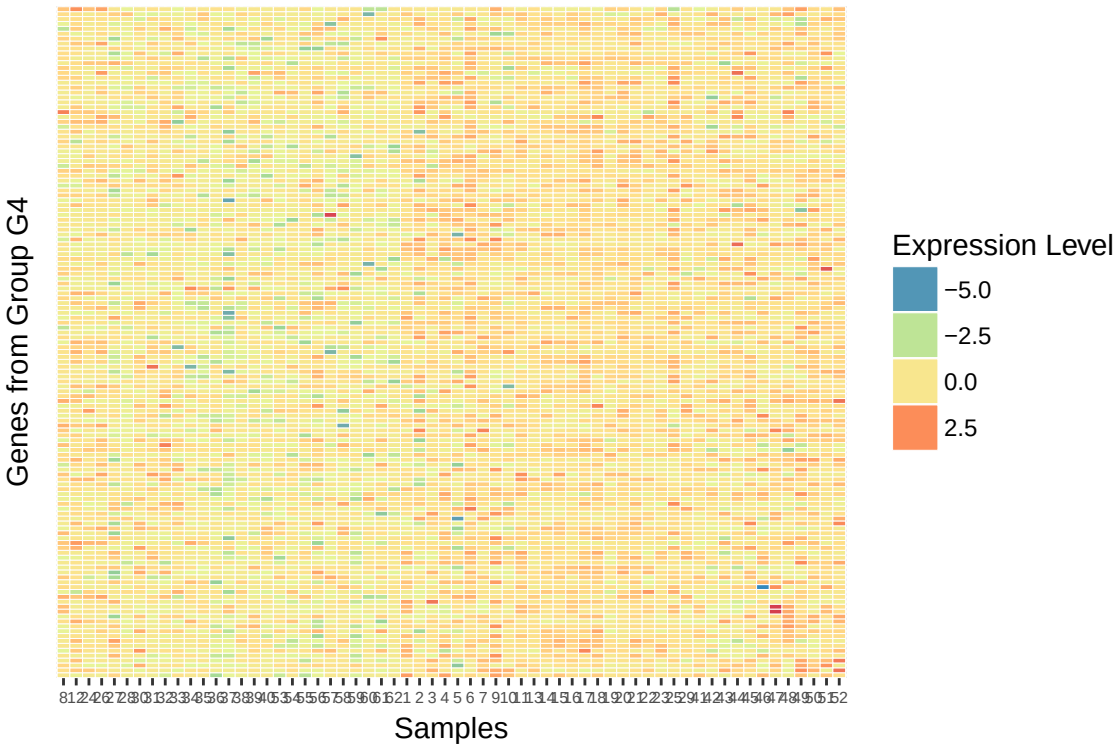



```
## C1:  1 2 3 4 6 7 8 9 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
## 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53
```

```
## 54 55 56 57 58 59 60 61 62
## C2:  5 10
```




```
## C1:  1 2 3 4 5 6 7 9 10 11 13 14 15 16 17 18 19 20 21 22 23 25 29 41 42 43
## 44 45 46 47 48 49 50 51 52
## C2:  8 12 24 26 27 28 30 31 32 33 34 35 36 37 38 39 40 53 54 55 56 57 58
## 59 60 61 62
```




#5.1.3

```
## C1:  1 2 3 4 5 7 8 9 10 11 13 14 15 16 17 18 19 22 23 25 27 30 33 38 41 42
## 44 45 46 47 48 49 51 52 56 57 58 61
```

```
## C2:  6 12 20 21 24 26 28 29 31 32 34 35 36 37 39 40 43 50 53 54 55 59 60
## 62
```

# References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96 (12), 6745–6750.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., and others, 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *science*, 286 (5439), 531–537.

McLachlan, G. J., Bean, R., and Peel, D., 2002. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18 (3), 413–422.