Gene Expression Analysis

We are going to analyse genes in the Drosophila Melanogaster genome for differential expression in two different cell types: Neurons and neuroblasts.

We are going to use data published in 2012 by Berger et al, which was downloaded from <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38764>

The file *gene\_expression.txt* contains the normalised FPKM expression values from both groups with 3 replicates each.

We are going to use R for this workshop, but the same results can be achieved using python or even Excel.

1. Load the data as a table in R (data=read.table…) and familiarise yourself with the format.
2. Filter the data for genes that have a decent level of expression. This helps to avoid unnecessary testing of genes that are not going to produce any meaningful results anyway and then over-correct with multiple hypothesis correction.  
   Only keep rows in the table where at least 3 samples have an FPKM value of 0.5 or higher. This cutoff is somewhat arbitrary, and you are welcome to play around with other methods. This procedure should retain 8000-9000 genes only. These are the genes expressed in one or both of the cell types.  
   keep <- rowSums(data[,2:7]>0.5) >= 3 #this is an EdgeR function, so load the library   
   data = data[keep,]
3. The data is already normalised for library size (FPKM values), but not for composition bias. Let’s see whether the latter is an issue by calculating the average logFC for all genes other than the top 20% highest log fold changes.  
   data$logFC = log2(…)  
   background=data[order(abs(data$logFC), decreasing=F),][1:(dim(data)[1]\*0.8),]  
   summary(background$logFC)  
   Do you think this distribution of log fold changes is worth correcting for? You would do this by scaling FPKM values in one sample to adjust for the overall shift in logFCs.  
   data[,2:4] = data[,2:4]/correction\_factor  
   However, since values go both ways (higher and lower than 0), it might also be that the general assumption for TMM (most genes not DE) is violated here.
4. Perform the Welch T-test on all of the rows of the matrix. Store the P-value from the test in a new column.  
   data$Pval = -1  
   for(i in 1:dim(data)[1]){data[i, “Pval”] = t.test(x=…,y=…, var.equal=F)$p.value}  
   Have a look at the distribution of P-values. How many are significant? What if we apply Bonferroni correction?
5. Read the set of published DE genes from this study as a table as well.  
   How many of the genes that have a raw/corrected P-value smaller than 0.05 are reported by the authors?  
   merge(data, published, by=1)  
   What kind of multiple hypothesis correction was applied to the published results, If any?