Statistical Analysis

Harlan Gillespie

27/09/2021

##Dependencies

```
library(SummarizedExperiment)
library(genefilter)
library(limma)
library(dbplyr)
library(ggplot2)
```

Load the data

Data is loaded and the edata is transformed and filtered as it was in "Exploratory Data Analysis.Rmd".

Linear Regression

The null hypothesis is that there is no change in gene expression between fetal and adult (control) groups.

Model matrix is made to include both the variable of interest, Group, and the covariate RIN. RIN was shown to have some impact on gene expression in the PCA analysis so adjusting for this covariate is appropriate

```
mod1 = model.matrix( ~ pdata$RIN + pdata$Group)
fit1 = lmFit(filt_edata, mod1)
fit1 = eBayes(fit1)
```

```
## Warning: Zero sample variances detected, have been offset away from zero
```

```
tt = topTable(fit1, number = Inf, coef = 3)
genes_logFC = subset(tt, select = c(logFC, P.Value, adj.P.Val))
```

Volcano Plot

Create labels for whether genes are up-regulated, down-regulated or neither. (+-)1.5 was chosen as a cut-off for logFC, a relatively low cut-off value. P-values already take into account logFC so filtering by a high logFC is somewhat redundant. Also, a gene with an extremely high logFC does not necessarily suggest a greater effect on a biological function. A small change in expression of a gene on which many processes depend upon could have a greater effect on phenotype than a large change in expression of a gene involved in an island, isolated from significant networks. For this reason, filtering out too many genes with a low logFC value may exclude key findings.

```
genes_logFC$diffexp = "NO"
genes_logFC$diffexp[genes_logFC$logFC > 1.5 & genes_logFC$adj.P.Val < 0.05] = "UP"
genes_logFC$diffexp[genes_logFC$logFC < -1.5 & genes_logFC$adj.P.Val < 0.05] = "DOWN"
```

Set colour scheme for differential expression.

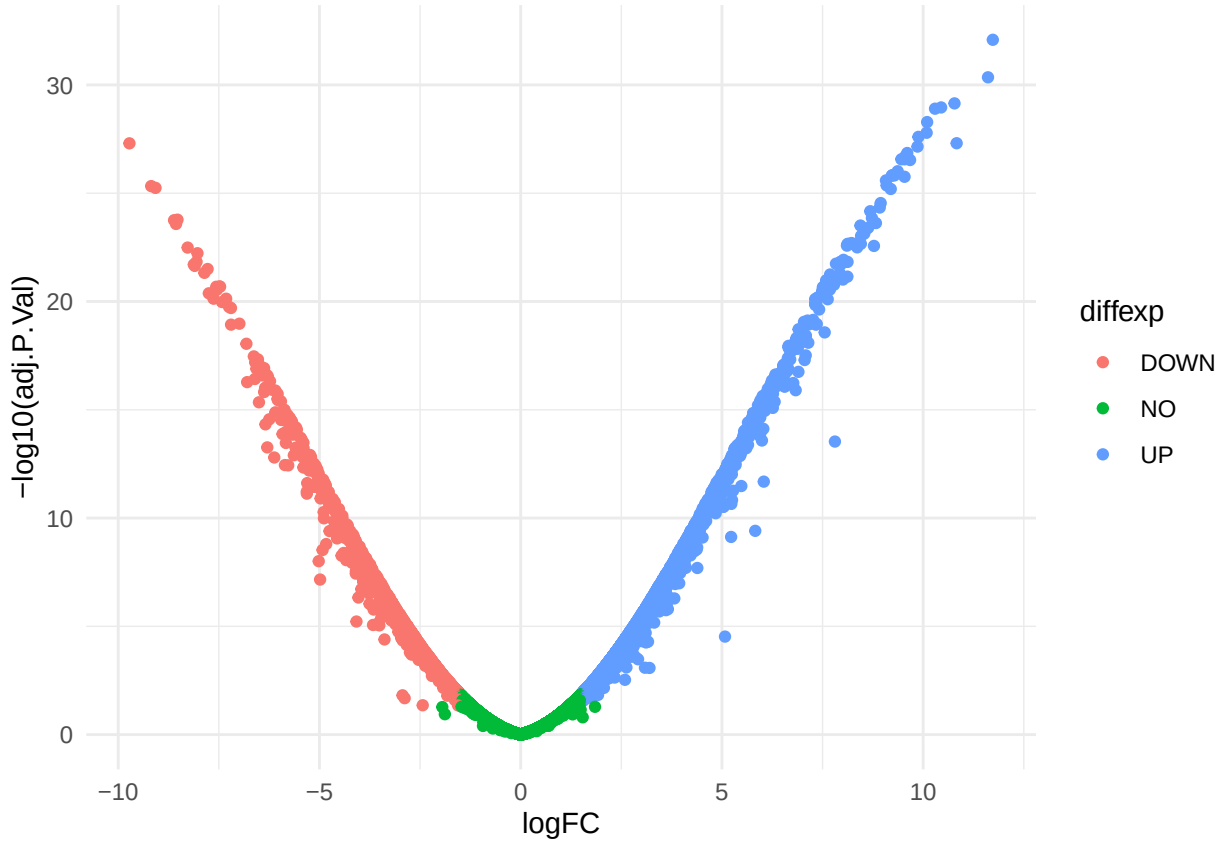
```
mycolors = c("blue", "red", "black")
names(mycolors) = c("DOWN", "UP", "NO")
```

Create new column which labels deferentially expressed genes

```
genes_logFC$diffexplab = NA
genes_logFC$diffexplab[genes_logFC$diffexp != "NO"] = rownames(genes_logFC)[genes_logFC$diffexp != "NO"]
```

Create the volcano plot with colour coding and labels

```
ggplot(data=genes_logFC, aes(x=logFC, y=-log10(adj.P.Val), col=diffexp)) +
    geom_point() +
    theme_minimal()
```



Create TSV file

Lastly, we can export the logFC, P.value and adjusted (FDR) P.value of each gene in the form of a TSV file.

```
glFC_tsv = genes_logFC[,c(1,2,3)]
glFC_tsv = tibble::rownames_to_column(glFC_tsv, "Entrez.ID")
write.table(glFC_tsv, file = "genes_logFC.tsv", sep = "\t", col.names = colnames(glFC_tsv), row.names =
```

Session Info

```
sessionInfo()
```

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
```

```
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] ggplot2_3.3.5               dbplyr_2.1.1
##  [3] limma_3.46.0                genefilter_1.72.1
##  [5] SummarizedExperiment_1.20.0 Biobase_2.50.0
##  [7] GenomicRanges_1.42.0        GenomeInfoDb_1.26.7
##  [9] IRanges_2.24.1              S4Vectors_0.28.1
## [11] BiocGenerics_0.36.1         MatrixGenerics_1.2.1
## [13] matrixStats_0.59.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.7             lattice_0.20-44        assertthat_0.2.1
##  [4] digest_0.6.27          utf8_1.2.1             R6_2.5.0
##  [7] RSQLite_2.2.7          evaluate_0.14          highr_0.9
## [10] httr_1.4.2             pillar_1.6.1           zlibbioc_1.36.0
## [13] rlang_0.4.11           annotate_1.68.0        blob_1.2.2
## [16] Matrix_1.3-4           rmarkdown_2.9          labeling_0.4.2
## [19] splines_4.0.5          stringr_1.4.0          RCurl_1.98-1.3
## [22] bit_4.0.4              munsell_0.5.0          DelayedArray_0.16.3
## [25] compiler_4.0.5         xfun_0.23              pkgconfig_2.0.3
## [28] htmltools_0.5.1.1      tidyselect_1.1.1       tibble_3.1.2
## [31] GenomeInfoDbData_1.2.4 XML_3.99-0.6           fansi_0.5.0
## [34] crayon_1.4.1           dplyr_1.0.7            withr_2.4.2
## [37] bitops_1.0-7           grid_4.0.5             xtable_1.8-4
## [40] gtable_0.3.0           lifecycle_1.0.0        DBI_1.1.1
## [43] magrittr_2.0.1         scales_1.1.1           stringi_1.6.2
## [46] cachem_1.0.5           farver_2.1.0           XVector_0.30.0
## [49] ellipsis_0.3.2         generics_0.1.0         vctrs_0.3.8
## [52] tools_4.0.5            bit64_4.0.5            glue_1.4.2
## [55] purrr_0.3.4            fastmap_1.1.0          survival_3.2-11
## [58] yaml_2.2.1             AnnotationDbi_1.52.0   colorspace_2.0-2
## [61] memoise_2.0.0          knitr_1.33
```