differential-expression-analysis

Victoria Latynina

2023-03-21

Differential Expression analysis

Dataset: GSE106542 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106542)

Differences in expression between TEMRA IL7-high and TEMRA IL7 low

Pipeline steps:

- Preprocess counts data
- Run VST for PCA, remove outliers, run PCA
- Run differential expression
- Find pathways for DE

```
library(ggplot2)
library(DESeq2)
library(apeglm)
library(ggrepel)
library(dplyr)
library(org.Hs.eg.db)
library(PCAtools)
library(GEOquery)
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=40),tidy=TRUE)
```

Preprocess counts data

```
path <- "~/Code/sys_bio/Part_2/materials/GSE106542_RAW/GSE106542_Bulk_raw_counts.txt"
data <- read.table(path, row.names = 1, sep = "\t",
    header = 1)
data$gene_symbol <- mapIds(org.Hs.eg.db,
    gsub("\\.\\d+", "", rownames(data)),
    column = "SYMBOL", "ENSEMBL")
data$gene_name <- mapIds(org.Hs.eg.db, gsub("\\.\\d+",
    "", rownames(data)), column = "GENENAME",
    "ENSEMBL")

mapping <- data[, c("gene_symbol", "gene_name"),
    drop = TRUE]
counts <- data[, 2:ncol(data) - 2]
colnames(counts) <- gsub("BRNA_", "", colnames(counts))

annotations <- getGEO("GSE106542")[[1]]
pdata <- pData(annotations)[, c("title",
    "description", "organism_ch1", "molecule_ch1",
    "subject #:ch1", "cell subtype surface markers:ch1",
    "cell subtype:ch1", "cell type:ch1",
    "longitudinal visit:ch1")]
rownames(pdata) <- gsub("Bulk_RNA-seq_",
    "", pdata$title)
```

Match columns and rows for all the data

```
col_order <- rownames(pdata)
counts <- counts[, col_order]
identical(colnames(counts), rownames(pdata))
```

```
## [1] TRUE
```

Create a new dataset without “TEM” and “TCM” rows

```
temra_pdata <- subset(pdata, pdata$`cell subtype:ch1` !=
    "TEM" & pdata$`cell subtype:ch1` != "TCM")
col_order <- rownames(temra_pdata)
temra_counts <- counts[, col_order]
identical(colnames(temra_counts), rownames(temra_pdata))
```

```
## [1] TRUE
```

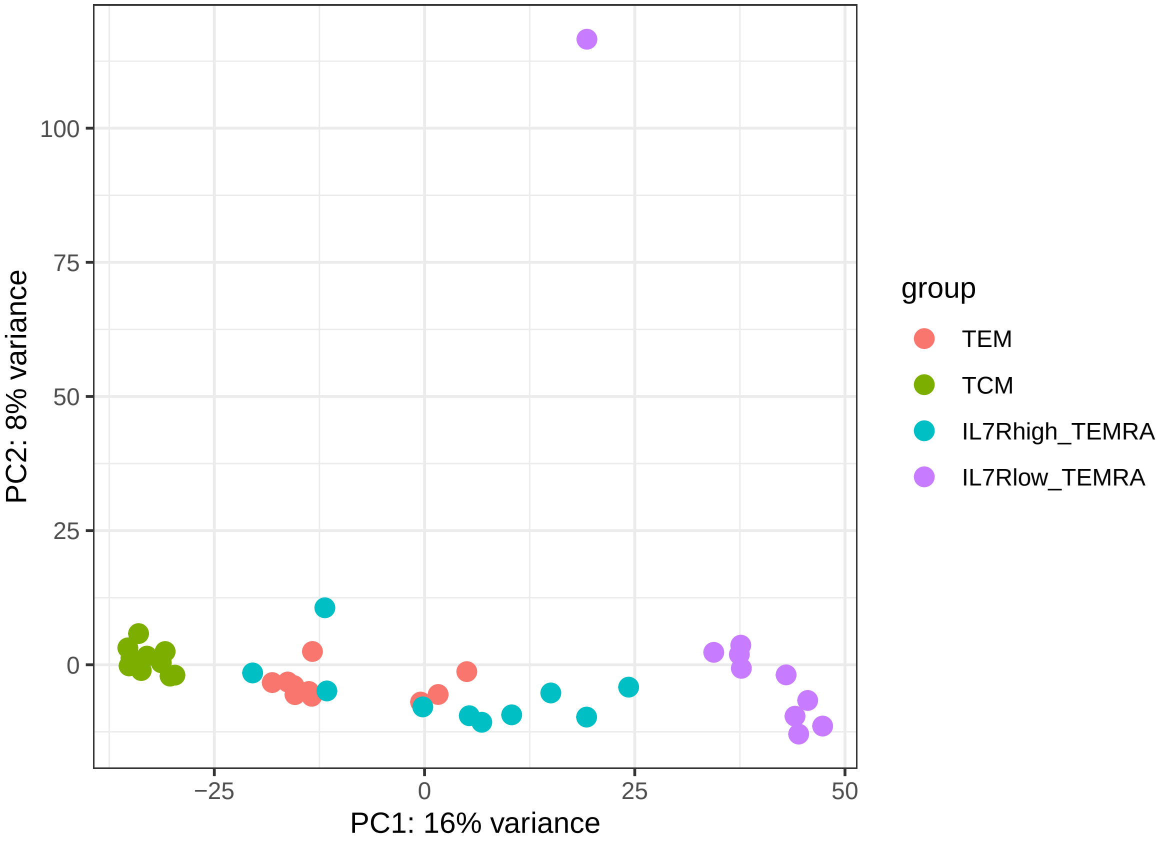
Run differential expression for all cell subtypes

```
pdata$Cell_subtype <- as.character(pdata$`cell subtype:ch1`)
pdata$Cell_subtype <- factor(pdata$Cell_subtype,
    levels = c("TEM", "TCM", "IL7Rhigh_TEMRA",
        "IL7Rlow_TEMRA"))
pdata$Donor <- as.character(pdata$`subject #:ch1`)
pdata$Donor <- factor(pdata$Donor, levels = c("Subject16",
    "Subject20", "Subject21", "Subject22",
    "Subject23"))
dds <- DESeqDataSetFromMatrix(countData = counts,
    colData = pdata, design = ~Cell_subtype +
        Donor)
dds <- DESeq(dds)
resultsNames(dds)
```

```
## [1] "Intercept"                          "Cell_subtype_TCM_vs_TEM"
## [3] "Cell_subtype_IL7Rhigh_TEMRA_vs_TEM" "Cell_subtype_IL7Rlow_TEMRA_vs_TEM"
## [5] "Donor_Subject20_vs_Subject16"       "Donor_Subject21_vs_Subject16"
## [7] "Donor_Subject22_vs_Subject16"       "Donor_Subject23_vs_Subject16"
```

Run VST for PCA, run PCA and notice outliers

```
vst <- varianceStabilizingTransformation(dds)
plotPCA(vst, intgroup = c("Cell_subtype"),
    ntop = nrow(vst)) + theme_bw() + theme(aspect.ratio = 1)
```

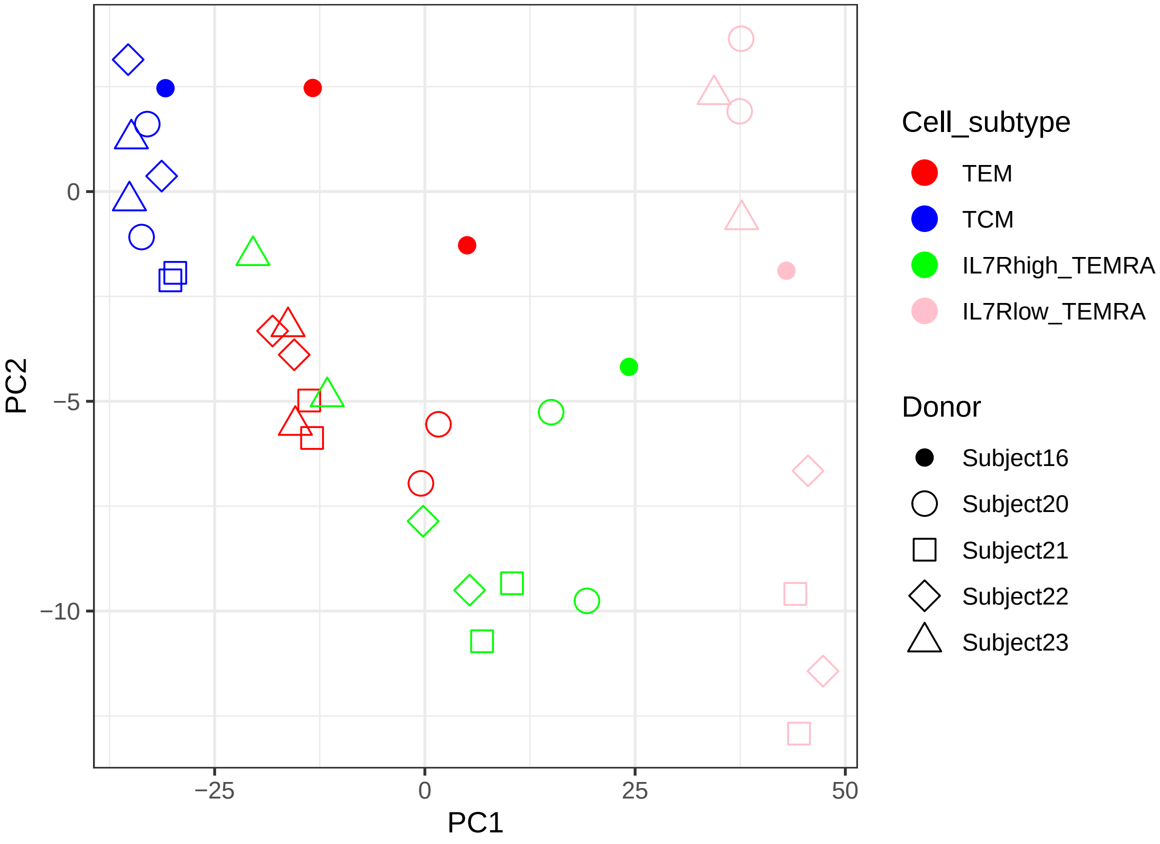


Add thresholds to get rid of outliers, plot PCA by donors and by cell subtypes

```
PCA_data <- plotPCA(vst, intgroup = c("Cell_subtype",
    "Donor"), ntop = nrow(vst), returnData = TRUE)

cell_colors <- c(TEM = "red", TCM = "blue",
    IL7Rhigh_TEMRA = "green", IL7Rlow_TEMRA = "pink")
donor_shapes <- c(Subject16 = 20, Subject20 = 21,
    Subject21 = 22, Subject22 = 23, Subject23 = 24)
outliers <- PCA_data$PC2 > 5
PCA_data <- subset(PCA_data, !outliers)
```

```
ggplot(PCA_data, aes(x = PC1, y = PC2, color = Cell_subtype,
    shape = Donor)) + geom_point(size = 4) +
    theme_bw() + theme(aspect.ratio = 1) +
    scale_color_manual(values = cell_colors) +
    scale_shape_manual(values = donor_shapes)
```



Run differential expression for target cell subtypes (“IL7Rhigh_TEMRA”, “IL7Rlow_TEMRA”)

```
temra_pdata$Cell_subtype <- as.character(temra_pdata$`cell subtype:ch1`)
temra_pdata$Cell_subtype <- factor(temra_pdata$Cell_subtype,
    levels = c("IL7Rhigh_TEMRA", "IL7Rlow_TEMRA"))
temra_pdata$Donor <- as.character(temra_pdata$`subject #:ch1`)
temra_pdata$Donor <- factor(temra_pdata$Donor,
    levels = c("Subject16", "Subject20",
        "Subject21", "Subject22", "Subject23"))
dds <- DESeqDataSetFromMatrix(countData = temra_counts,
    colData = temra_pdata, design = ~Cell_subtype +
        Donor)
dds <- DESeq(dds)
resultsNames(dds)
```

```
## [1] "Intercept"
## [2] "Cell_subtype_IL7Rlow_TEMRA_vs_IL7Rhigh_TEMRA"
## [3] "Donor_Subject20_vs_Subject16"
## [4] "Donor_Subject21_vs_Subject16"
## [5] "Donor_Subject22_vs_Subject16"
## [6] "Donor_Subject23_vs_Subject16"
```

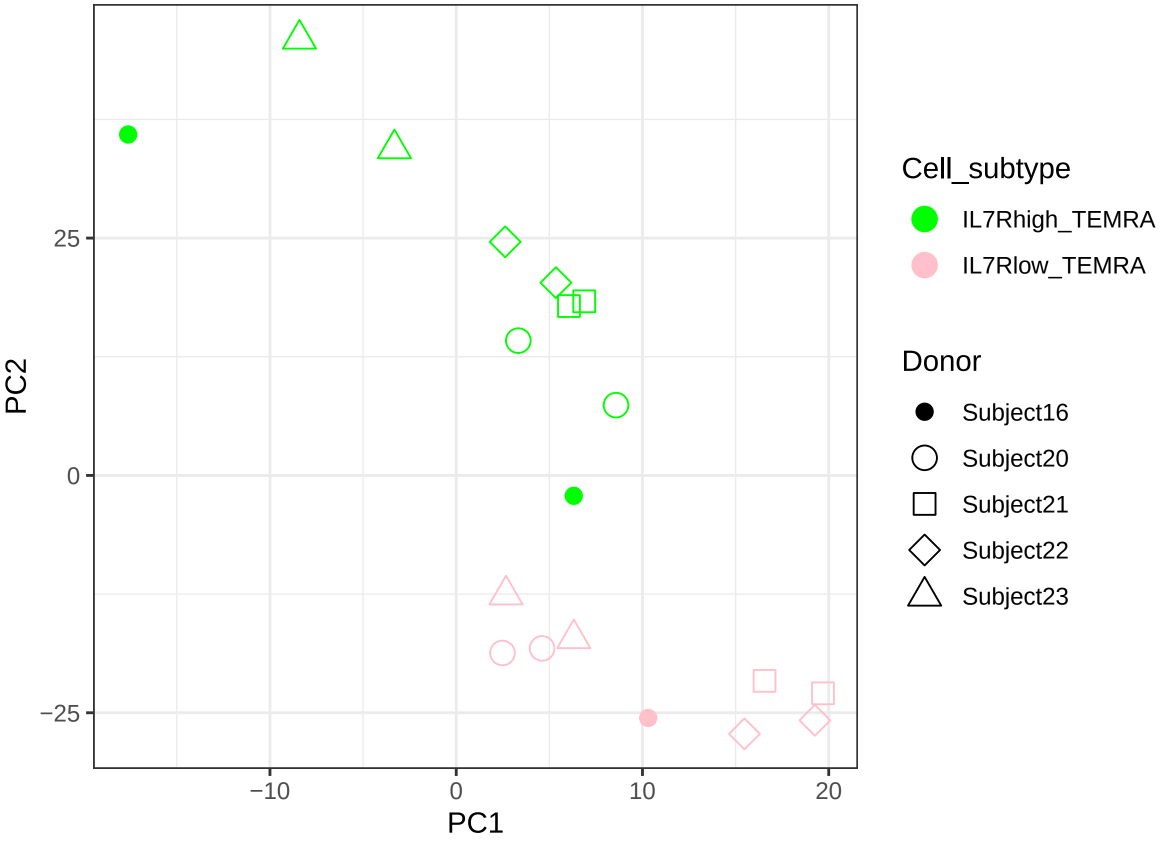
PCA of target cell subtypes

```
vst <- varianceStabilizingTransformation(dds)

pca_data <- prcomp(t(assay(vst)))
pca_df <- data.frame(PC1 = pca_data$x[, 1],
    PC2 = pca_data$x[, 2], Cell_subtype = colData(vst)$Cell_subtype,
    Donor = colData(vst)$Donor)

threshold_PC1 <- 80  #50 * sd(pca_df$PC1)
threshold_PC2 <- 10 * sd(pca_df$PC2)
fpca_df <- pca_df[abs(pca_df$PC1) < threshold_PC1 &
    abs(pca_df$PC2) < threshold_PC2, ]

ggplot(fpca_df, aes(x = PC1, y = PC2, color = Cell_subtype,
    shape = Donor)) + geom_point(size = 4) +
    theme_bw() + theme(aspect.ratio = 1) +
    scale_color_manual(values = cell_colors) +
    scale_shape_manual(values = donor_shapes)
```

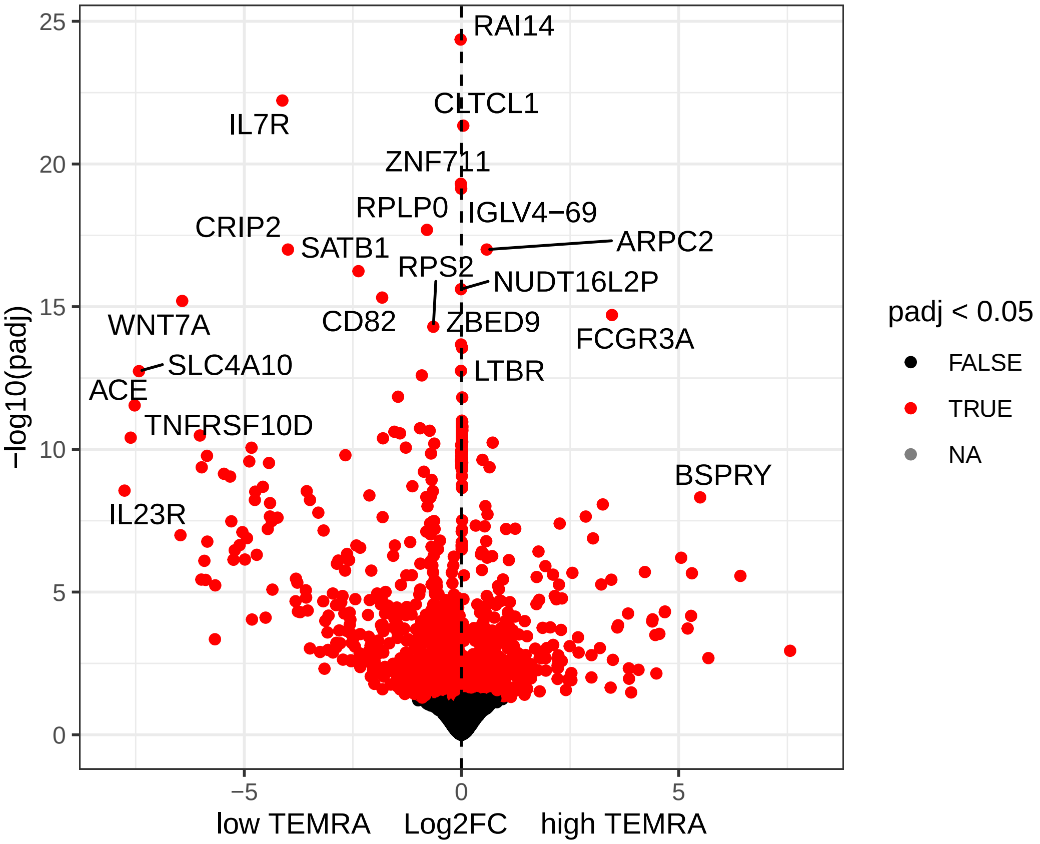


Get volcano plots

```r
res <- lfcShrink(dds, coef = "Cell_subtype_IL7Rlow_TEMRA_vs_IL7Rhigh_TEMRA",
    type = "apeglm")
res$gene_symbol <- mapIds(org.Hs.eg.db, gsub("\\.\\d+",
    "", rownames(res)), column = "SYMBOL",
    "ENSEMBL")
head(res)
```

```
## log2 fold change (MAP): Cell subtype IL7Rlow TEMRA vs IL7Rhigh TEMRA
## Wald test p-value: Cell subtype IL7Rlow TEMRA vs IL7Rhigh TEMRA
## DataFrame with 6 rows and 6 columns
##                      baseMean log2FoldChange     lfcSE      pvalue        padj
##                     <numeric>      <numeric> <numeric>   <numeric>   <numeric>
## ENSG00000000003.10   20.4825     -0.0409977  0.266200 8.23654e-04 1.36952e-02
## ENSG00000000005.5     0.0000             NA        NA          NA          NA
## ENSG00000000419.8   597.6428     -0.1060671  0.166264 4.13039e-01 7.56774e-01
## ENSG00000000457.9   239.5410     -0.0702519  0.176285 5.88313e-01 8.56518e-01
## ENSG00000000460.12  121.3805      0.0259527  0.256977 6.02521e-01 8.64641e-01
## ENSG00000000938.8  2631.8403      2.2597447  0.423539 3.60961e-10 3.97107e-08
##                     gene_symbol
##                     <character>
## ENSG00000000003.10      TSPAN6
## ENSG00000000005.5         TNMD
## ENSG00000000419.8         DPM1
## ENSG00000000457.9        SCYL3
## ENSG00000000460.12    C1orf112
## ENSG00000000938.8          FGR
```

```r
resDF <- as.data.frame(res)
ggplot(resDF, aes(x = log2FoldChange, y = -log10(padj),
    color = padj < 0.05)) + geom_point() +
    theme_bw() + scale_color_manual(values = c("black",
    "red")) + geom_text_repel(data = resDF %>%
    dplyr::filter(padj < 1e-07), aes(label = gene_symbol),
    color = "black") + xlim(c(-8, 8)) + xlab("low TEMRA    Log2FC    high TEMRA") +
    geom_vline(xintercept = 0, lty = 2) +
    theme(aspect.ratio = 1)
```



Pathway analysis

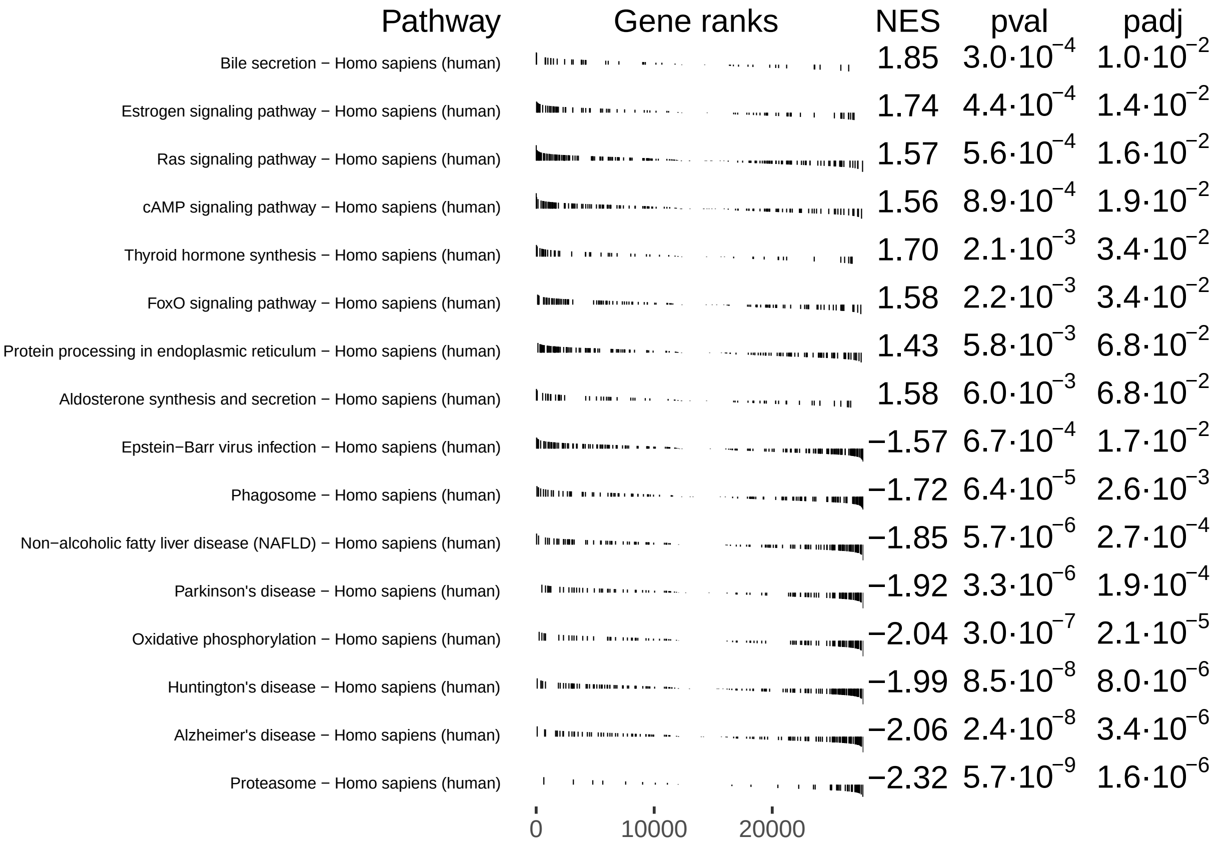
```
library(fgsea)

deResults <- results(dds)
deResults$gene_symbol <- mapIds(org.Hs.eg.db,
    gsub("\\.\\d+", "", rownames(deResults)),
    column = "SYMBOL", "ENSEMBL")
stats <- deResults$stat
names(stats) <- deResults$gene_symbol
complete_cases <- complete.cases(stats)
stats <- stats[complete_cases]
top_genes <- resDF %>%
    dplyr::filter(padj < 1e-07)
```

```
load("~/Code/sys_bio/Part_2/materials/keggSymbolHuman.rdata")
fgseaResults <- fgseaMultilevel(keggSymbolHuman,
    stats, minSize = 15, maxSize = 500)

topPathwaysUp <- fgseaResults[ES > 0, ][head(order(pval),
    n = 8), pathway]
topPathwaysDown <- fgseaResults[ES < 0, ][head(order(pval),
    n = 8), pathway]
topPathways <- c(topPathwaysUp, rev(topPathwaysDown))
```

```
plotGseaTable(keggSymbolHuman[topPathways],
    stats, fgseaResults, gseaParam = 0.5,
    pathwayLabelStyle = list(size = 6))
```



Pathway	Gene ranks	NES	pval	padj
Bile secretion − Homo sapiens (human)		1.85	$3.0\cdot10^{-4}$	$1.0\cdot10^{-2}$
Estrogen signaling pathway − Homo sapiens (human)		1.74	$4.4\cdot10^{-4}$	$1.4\cdot10^{-2}$
Ras signaling pathway − Homo sapiens (human)		1.57	$5.6\cdot10^{-4}$	$1.6\cdot10^{-2}$
cAMP signaling pathway − Homo sapiens (human)		1.56	$8.9\cdot10^{-4}$	$1.9\cdot10^{-2}$
Thyroid hormone synthesis − Homo sapiens (human)		1.70	$2.1\cdot10^{-3}$	$3.4\cdot10^{-2}$
FoxO signaling pathway − Homo sapiens (human)		1.58	$2.2\cdot10^{-3}$	$3.4\cdot10^{-2}$
Protein processing in endoplasmic reticulum − Homo sapiens (human)		1.43	$5.8\cdot10^{-3}$	$6.8\cdot10^{-2}$
Aldosterone synthesis and secretion − Homo sapiens (human)		1.58	$6.0\cdot10^{-3}$	$6.8\cdot10^{-2}$
Epstein−Barr virus infection − Homo sapiens (human)		−1.57	$6.7\cdot10^{-4}$	$1.7\cdot10^{-2}$
Phagosome − Homo sapiens (human)		−1.72	$6.4\cdot10^{-5}$	$2.6\cdot10^{-3}$
Non−alcoholic fatty liver disease (NAFLD) − Homo sapiens (human)		−1.85	$5.7\cdot10^{-6}$	$2.7\cdot10^{-4}$
Parkinson's disease − Homo sapiens (human)		−1.92	$3.3\cdot10^{-6}$	$1.9\cdot10^{-4}$
Oxidative phosphorylation − Homo sapiens (human)		−2.04	$3.0\cdot10^{-7}$	$2.1\cdot10^{-5}$
Huntington's disease − Homo sapiens (human)		−1.99	$8.5\cdot10^{-8}$	$8.0\cdot10^{-6}$
Alzheimer's disease − Homo sapiens (human)		−2.06	$2.4\cdot10^{-8}$	$3.4\cdot10^{-6}$
Proteasome − Homo sapiens (human)		−2.32	$5.7\cdot10^{-9}$	$1.6\cdot10^{-6}$