Volcano Plot

2023-11-05

This R script generates a volcano plot to visualize differentially expressed genes (DEGs) from RNA sequencing data, highlighting genes based on significance (p-value) and regulation (up- or down-regulated).

Based on the following script: https://github.com/kevinblighe/EnhancedVolcano

Install necessary packages if not already installed

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
```

```
## Bioconductor version '3.16' is out-of-date; the current release version '3.20'
##   is available with R version '4.4'; see https://bioconductor.org/install
```

```
BiocManager::install("EnhancedVolcano")
```

```
## Bioconductor version 3.16 (BiocManager 1.30.22), R 4.2.2 (2022-10-31)
```

```
## Warning: package(s) not installed when version(s) same as or greater than current; use
##   `force = TRUE` to re-install: 'EnhancedVolcano'
```

```
## Old packages: 'ape', 'aplot', 'askpass', 'backports', 'BH', 'BiocManager',
##   'bit', 'bit64', 'bitops', 'boot', 'brew', 'brio', 'broom', 'bslib', 'cachem',
##   'callr', 'cli', 'cluster', 'codetools', 'colorspace', 'commonmark',
##   'cowplot', 'cpp11', 'crayon', 'credentials', 'curl', 'data.table', 'DBI',
##   'dbplyr', 'dendextend', 'desc', 'digest', 'downlit', 'evaluate', 'fansi',
##   'farver', 'fastmap', 'foreign', 'fs', 'gert', 'ggforce', 'ggfun', 'ggh4x',
##   'ggnewscale', 'ggplot2', 'ggraph', 'ggrepel', 'gh', 'glue', 'graphlayouts',
##   'gtable', 'haven', 'highr', 'htmltools', 'htmlwidgets', 'httpuv', 'httr2',
##   'igraph', 'jsonlite', 'KernSmooth', 'knitr', 'later', 'lattice', 'markdown',
##   'mgcv', 'munsell', 'nlme', 'openssl', 'patchwork', 'pkgbuild', 'pkgdown',
##   'pkgload', 'plotly', 'polyclip', 'processx', 'profvis', 'progress',
##   'promises', 'ps', 'ragg', 'Rcpp', 'RcppArmadillo', 'RcppEigen', 'RCurl',
##   'readr', 'remotes', 'reprex', 'rlang', 'rmarkdown', 'roxygen2', 'rpart',
##   'RSQLite', 'rstudioapi', 'RUnit', 'rvest', 'sass', 'scales', 'scatterpie',
##   'seriation', 'shadowtext', 'shiny', 'stringi', 'survival', 'sys',
##   'systemfonts', 'testthat', 'textshaping', 'tidygraph', 'tidyr', 'tidyselect',
##   'tidytree', 'timechange', 'tinytex', 'tweenr', 'usethis', 'uuid', 'vctrs',
##   'vegan', 'viridis', 'vroom', 'waldo', 'withr', 'xfun', 'XML', 'xml2',
##   'xopen', 'yaml', 'yulab.utils', 'zip'
```

Load the data from a CSV file

```
mat<-read.csv("/Users/jeffreyreina/Documents/Salk/RNAseq MDA-MB-231 results/03.Result_X202SC23073852-Z0
```

Convert the log2fc column to numeric format to ensure correct plotting

```
mat$log2fc<-as.numeric(mat$log2FoldChange)
```

Load the EnhancedVolcano library for creating volcano plots

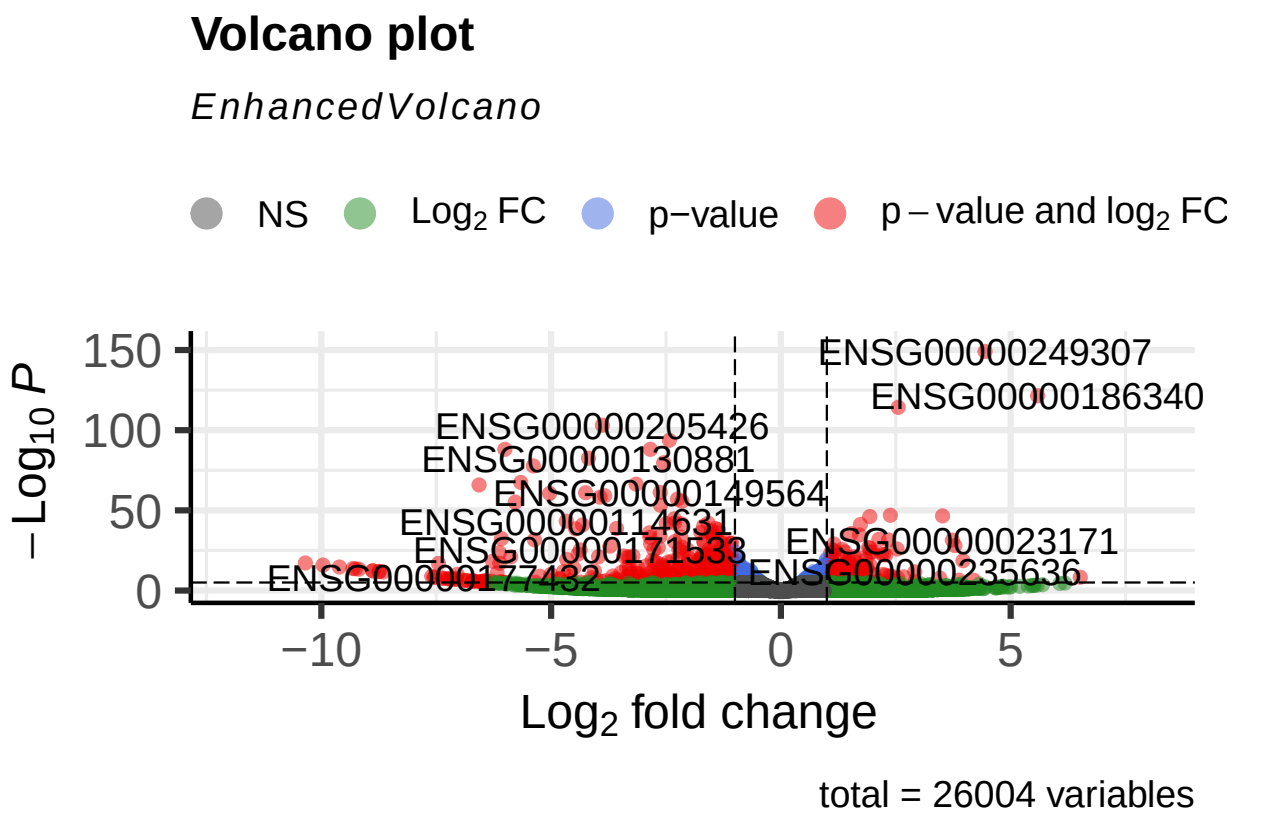
```
library(EnhancedVolcano)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggrepel
```

Create a basic volcano plot with default settings

```
EnhancedVolcano(mat,
                lab = rownames(mat),
                x = "log2FoldChange",
                y = "pvalue")
```



Preview the first few rows of mat to check the data

```
head(mat)
```

```
##                      KO1         KO2         KO3        WT1        WT2
## ENSG00000249307 1986.40039 1437.371843 1596.072302   77.33116   79.94080
## ENSG00000186340  849.20828 1014.942376  908.588551   16.33757   21.03705
## ENSG00000124225 7736.24814 9679.193068 9017.853265 1503.05629 1436.41000
## ENSG00000205426   65.85296   70.034359   58.185474  808.16505 1022.40079
## ENSG00000160932  955.35932 1037.175506  822.653082 4814.13681 4847.77856
## ENSG00000164692   17.69184    6.669939    7.161289  725.38803  658.88051
##                      WT3         KO         WT log2FoldChange    pvalue
## ENSG00000249307   73.10935 1673.28151   76.79377       4.443199 8.02e-150
## ENSG00000186340   19.93891  924.24640   19.10451       5.585994 3.39e-122
## ENSG00000124225 1567.42025 8811.09816 1502.29551       2.552490 5.62e-115
## ENSG00000205426 1025.74640   64.69093  952.10408      -3.884789 7.03e-104
## ENSG00000160932 5462.15495  938.39597 5041.35677      -2.426633  1.97e-94
## ENSG00000164692  649.12245   10.50769  677.79700      -6.006074  8.91e-89
```

```
##                      padj gene_name gene_chr gene_start  gene_end gene_strand
## ENSG00000249307 1.40e-145 LINC01088        4   78971748  79308798           +
## ENSG00000186340 2.95e-118     THBS2        6  169215780 169254044           -
## ENSG00000124225 3.26e-111    PMEPA1       20   57648392  57711536           -
## ENSG00000205426 3.07e-100     KRT81       12   52285913  52291534           -
## ENSG00000160932  6.87e-91      LY6E        8  143017982 143023832           +
## ENSG00000164692  2.26e-85    COL1A2        7   94394561  94431232           +
##                 gene_length   gene_biotype
## ENSG00000249307        3995      antisense
## ENSG00000186340        6412 protein_coding
## ENSG00000124225        5619 protein_coding
## ENSG00000205426        1929 protein_coding
## ENSG00000160932        2640 protein_coding
## ENSG00000164692       11156 protein_coding
##                                                                                gene_descriptio
## ENSG00000249307        long intergenic non-protein coding RNA 1088 [Source:HGNC Symbol;Acc:HGNC:49148
## ENSG00000186340                                    thrombospondin 2 [Source:HGNC Symbol;Acc:HGNC:11780
## ENSG00000124225 prostate transmembrane protein, androgen induced 1 [Source:HGNC Symbol;Acc:HGNC:14107
## ENSG00000205426                                          keratin 81 [Source:HGNC Symbol;Acc:HGNC:6458
## ENSG00000160932                lymphocyte antigen 6 family member E [Source:HGNC Symbol;Acc:HGNC:6727
## ENSG00000164692                       collagen type I alpha 2 chain [Source:HGNC Symbol;Acc:HGNC:2198
##                 tf_family KO1_count KO2_count KO3_count WT1_count WT2_count
## ENSG00000249307         -      2021      1293      1783        71        95
## ENSG00000186340         -       864       913      1015        15        25
## ENSG00000124225         -      7871      8707     10074      1380      1707
## ENSG00000205426         -        67        63        65       742      1215
## ENSG00000160932         -       972       933       919      4420      5761
## ENSG00000164692         -        18         6         8       666       783
##                 WT3_count   KO1_fpkm     KO2_fpkm     KO3_fpkm    WT1_fpkm
## ENSG00000249307        66 32.1807860  23.28415504  25.94788662   1.2546431
## ENSG00000186340        18  8.5717076  10.24368005   9.20322332   0.1651492
## ENSG00000124225      1415 89.1082670 111.47776510 104.23422610  17.3379879
## ENSG00000205426       926  2.2094781   2.34956283   1.95906372  27.1550320
## ENSG00000160932      4931 23.4212169  25.42473635  20.23853241 118.1944229
## ENSG00000164692       586  0.1026387   0.03869203   0.04169163   4.2144836
##                    WT2_fpkm    WT3_fpkm    log2fc
## ENSG00000249307   1.2998601   1.1876881  4.443199
## ENSG00000186340   0.2131259   0.2018154  5.585994
## ENSG00000124225  16.6059715  18.1039220  2.552490
## ENSG00000205426  34.4297484  34.5107200 -3.884789
## ENSG00000160932 119.2844267 134.2784553 -2.426633
## ENSG00000164692   3.8365694   3.7762816 -6.006074
```

Add a significant column to flag genes with p-value less or equal to 0.05

```
mat$significant<-ifelse (mat$pvalue <= 0.05, "true", "false")
```

Calculate -log10(p-value) for better visualization on the y-axis

```
mat$log10pval<--log10(mat$pvalue)
```

Label genes as up or down regulated based on log2FoldChange

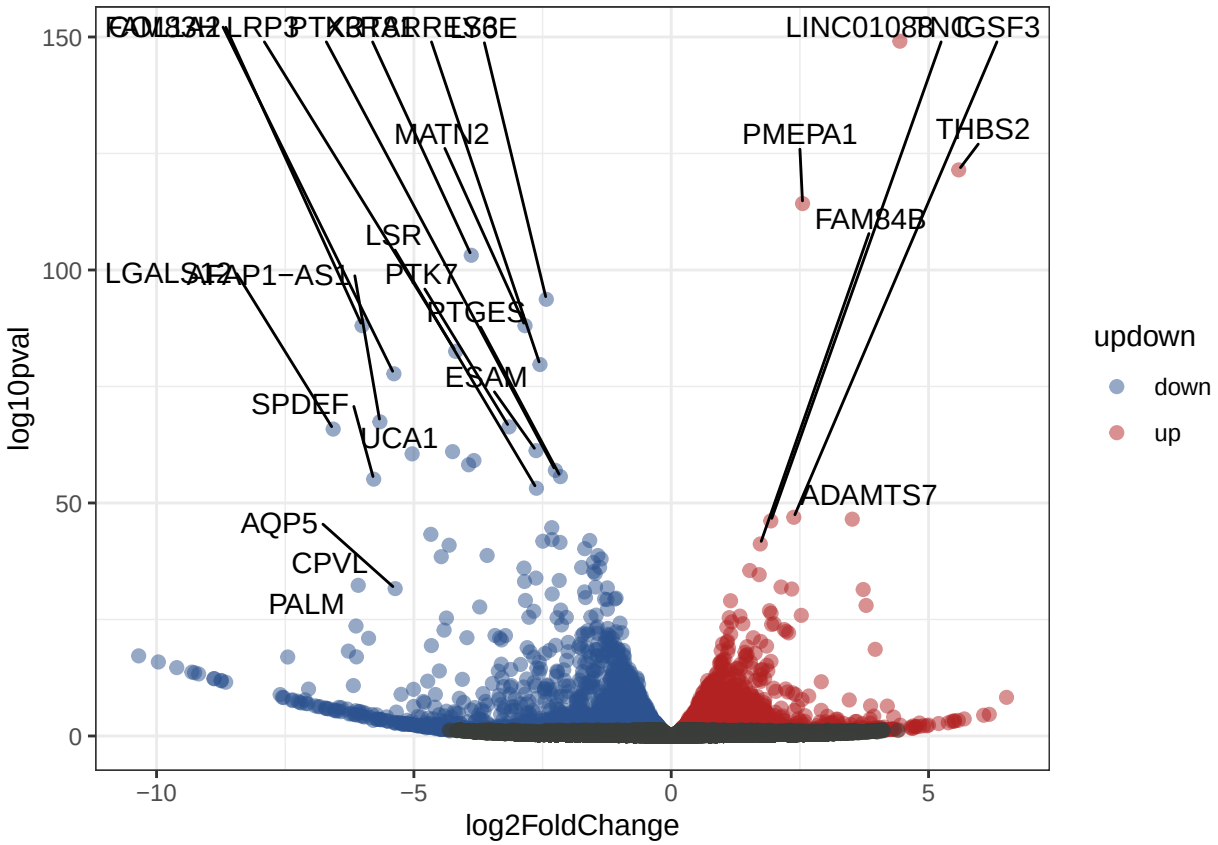
```
mat$updown<-ifelse (mat$log2FoldChange <= 0, "down", "up")
```

Customize volcano plot with colors for significant and non-significant points

```
log10pval<--log(mat$pvalue, 10)

ggplot(data=mat, aes(x = log2FoldChange, y = log10pval)) +
  geom_point(data = subset(mat, significant %in% c("true")), size = 2,
             alpha = 0.5,  aes(color = updown), shape=19) +
  scale_color_manual(values = c("#2c538f", "firebrick")) + geom_point(data = subset(mat, significant %in
               alpha = 0.5,  colour = "#3a3d3a") +
  theme_bw() +
  geom_text_repel(data = subset(mat, significant %in% c("true")), aes(label = gene_name))
```

```
## Warning: ggrepel: 4957 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Save the customized plot as a PDF

```
pdf("volcanoKOvsWT.pdf", width=6, height=4.5)

ggplot(data=mat, aes(x = log2FoldChange, y = log10pval)) +
  geom_point(data = subset(mat, significant %in% c("true")), size = 2,
             alpha = 0.5,  aes(color = updown), shape=19) +
  scale_color_manual(values = c("#2c538f", "firebrick")) + geom_point(data = subset(mat, significant %in
               alpha = 0.5,  colour = "#3a3d3a") +
  theme_bw() +
  geom_text_repel(data = subset(mat, significant %in% c("true")), aes(label = gene_name))
```

```
## Warning: ggrepel: 4958 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```