DEA on TCGA Breast Cancer

Parissa Amin

2023-12-22

Loading Data

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(DESeq2)
```

```
## Warning: package 'DESeq2' was built under R version 4.2.2
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
```

```
## Warning: package 'BiocGenerics' was built under R version 4.2.1
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:dplyr':
##
##     combine, intersect, setdiff, union
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:dplyr':
##
##     first, rename

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice

## The following object is masked from 'package:grDevices':
##
##     windows

## Loading required package: GenomicRanges

## Loading required package: GenomeInfoDb

## Warning: package 'GenomeInfoDb' was built under R version 4.2.2

## Loading required package: SummarizedExperiment

## Warning: package 'SummarizedExperiment' was built under R version 4.2.1

## Loading required package: MatrixGenerics

## Warning: package 'MatrixGenerics' was built under R version 4.2.1

## Loading required package: matrixStats

## Warning: package 'matrixStats' was built under R version 4.2.3

##
## Attaching package: 'matrixStats'

## The following object is masked from 'package:dplyr':
##
##     count

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##
## Attaching package: 'Biobase'
```

```
## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians
```

```
## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```
## Warning: multiple methods tables found for 'aperm'
```

```
## Warning: replacing previous import 'BiocGenerics::aperm' by
## 'DelayedArray::aperm' when loading 'SummarizedExperiment'
```

```r
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 4.2.3
```

```r
exp <- read.delim("D:/BulkRNA/Data/data_mrna_seq_v2_rsem.txt")
exp<- na.omit(exp)
exp<- data.frame(t(exp))

colnames(exp) <- exp[ 1,]
exp<- exp[-1:-2, ]
```

Loading Patiant Info Data

```r
PatientInfo <- read.delim("D:/BulkRNA/Data/tcga_clinical_data.tsv")
PatientInfo$Sample.ID <- gsub("-" ,"." , PatientInfo$Sample.ID)

## Defining a new column named Er.status
exp$ER.status <- PatientInfo$ER.Status.By.IHC[match(rownames(exp), PatientInfo$Sample.ID)]
exp<- na.omit(exp)
exp<- exp[exp$ER.status!="Indeterminate",]

remove(PatientInfo)
```

Creating Metadata table with information of our samples

```r
metadata<- as.data.frame(cbind(rownames(exp), exp[,"ER.status"]))
colnames(metadata)<- c("Sample.ID", "ER")
```

```
#####Deleting last column(the genes' names) of exp
expression_data<-  exp[,-20531]
expression_data <- data.frame(lapply(expression_data, as.numeric))
expression_data<- round(expression_data)
expression_data<- as.data.frame(t(expression_data))

remove(exp)
```

Creating dds object by DESeqDataSet function

```
dds<-DESeqDataSetFromMatrix(countData=expression_data,colData=metadata,design=~ER)
```

```
## converting counts to integer mode

## Warning in DESeqDataSet(se, design = design, ignoreRank): some variables in
## design formula are characters, converting to factors
```

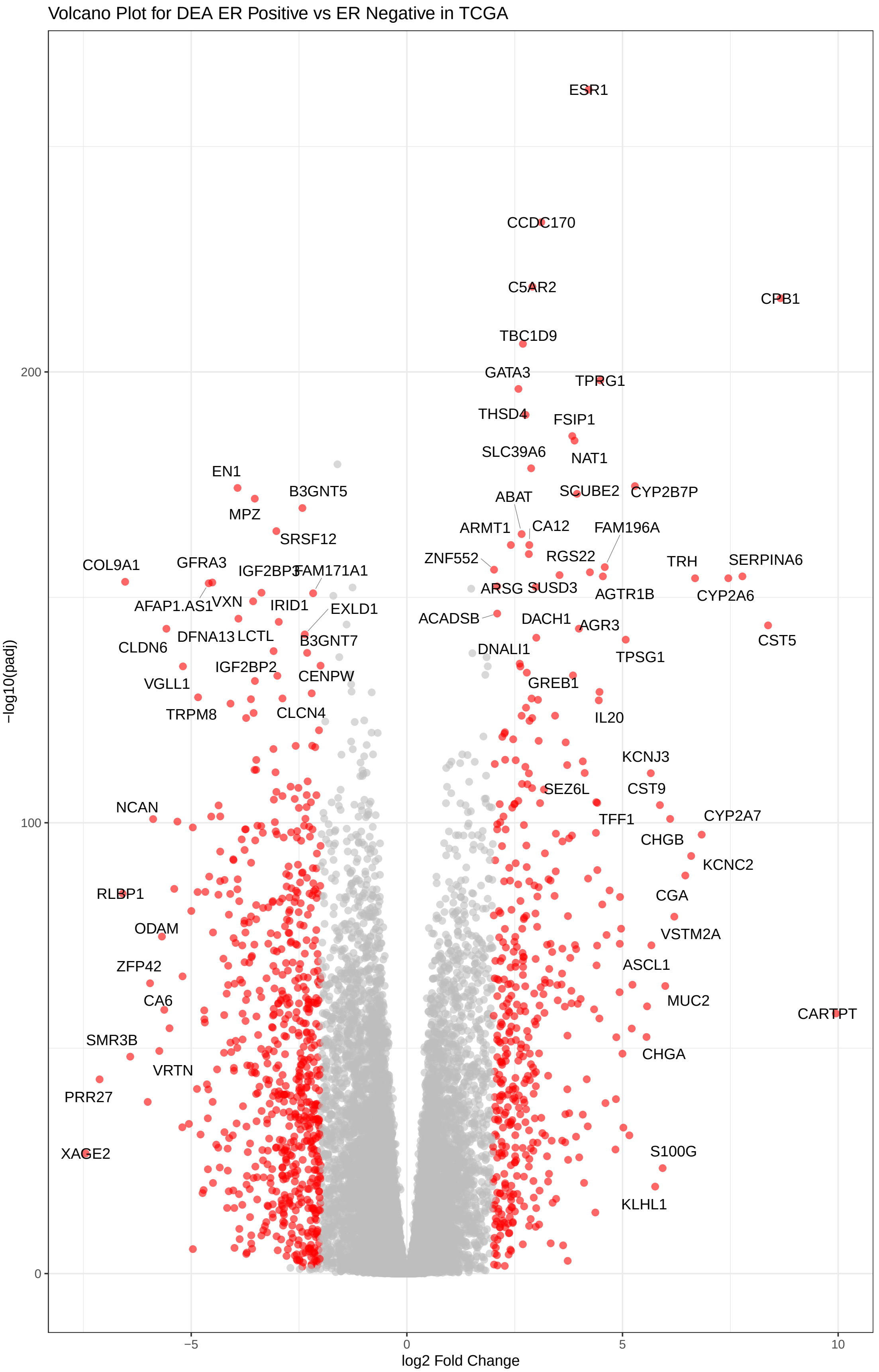
```
dds <- DESeq(dds)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 2911 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

```
My_Results<- results(dds)
My_Results_df<- as.data.frame(My_Results)

My_Results_df$GeneName <- rownames(My_Results_df)
```

```
padj_threshold <- 0.05
logFC_threshold <- 2


ggplot(My_Results_df, aes(x = log2FoldChange, y = -log10(padj))) +
  geom_point(size=2, aes(color = ifelse(abs(log2FoldChange) > logFC_threshold & padj < padj_threshold,
  geom_text_repel(
    aes(label = ifelse(abs(log2FoldChange) > logFC_threshold & padj < padj_threshold, GeneName, "")),
    box.padding = 0.5,
    point.padding = 0.3,
    segment.size = 0.2,
    segment.color = "grey50",
    box.color = "grey50"
  ) +
  scale_color_identity() +
  theme_bw() +
  labs(
    title = "Volcano Plot for DEA ER Positive vs ER Negative in TCGA",
    x = "log2 Fold Change",
    y = "-log10(padj)",
    color = "Significant"
  )
```

```
## Warning in geom_text_repel(aes(label = ifelse(abs(log2FoldChange) >
## logFC_threshold & : Ignoring unknown parameters: `box.colour`

## Warning: Removed 356 rows containing missing values (`geom_point()`).

## Warning: Removed 356 rows containing missing values (`geom_text_repel()`).

## Warning: ggrepel: 1008 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```


Volcano Plot for DEA ER Positive vs ER Negative in TCGA
ESR1
CCDC170
C5AR2
CPB1
TBC1D9
GATA3
TPRG1
THSD4
FSIP1
SLC39A6
NAT1
EN1
B3GNT5
MPZ
ABAT
SCUBE2
CYP2B7P
ARMT1
CA12
FAM196A
SRSF12
ZNF552
RGS22
TRH
SERPINA6
COL9A1
GFRA3
IGF2BP3
FAM171A1
ARSG
SUSD3
AGTR1B
CYP2A6
AFAP1.AS1
VXN
IRID1
EXLD1
ACADSB
DACH1
AGR3
CST5
DFNA13
LCTL
B3GNT7
CLDN6
DNALI1
TPSG1
IGF2BP2
CENPW
VGLL1
GREB1
TRPM8
CLCN4
IL20
KCNJ3
CST9
SEZ6L
NCAN
TFF1
CYP2A7
CHGB
KCNC2
RLBP1
CGA
ODAM
VSTM2A
ZFP42
ASCL1
CA6
MUC2
CARTPT
SMR3B
CHGA
VRTN
PRR27
XAGE2
S100G
KLHL1
200
100
0
−5
0
5
10
−log10(padj)
log2 Fold Change

```r
dev.off()
```

```
## null device 
##           1
```