pathways Enrichment Analysis on TCGA Breast Cancer

Parissa Amin

2023-12-31

DEA

We first perform DEA as before for ER+ vs ER- subtypes of breast tumors

```
library(fgsea)
library(clusterProfiler)
```

```
##
```

```
## Registered S3 methods overwritten by 'treeio':
##   method               from
##   MRCA.phylo           tidytree
##   MRCA.treedata        tidytree
##   Nnode.treedata       tidytree
##   Ntip.treedata        tidytree
##   ancestor.phylo       tidytree
##   ancestor.treedata    tidytree
##   child.phylo          tidytree
##   child.treedata       tidytree
##   full_join.phylo      tidytree
##   full_join.treedata   tidytree
##   groupClade.phylo     tidytree
##   groupClade.treedata tidytree
##   groupOTU.phylo       tidytree
##   groupOTU.treedata    tidytree
##   is.rooted.treedata   tidytree
##   nodeid.phylo         tidytree
##   nodeid.treedata      tidytree
##   nodelab.phylo        tidytree
##   nodelab.treedata     tidytree
##   offspring.phylo      tidytree
##   offspring.treedata   tidytree
##   parent.phylo         tidytree
##   parent.treedata      tidytree
##   root.treedata        tidytree
##   rootnode.phylo       tidytree
##   sibling.phylo        tidytree
```

```
## clusterProfiler v4.4.4  For help: https://yulab-smu.top/biomedical-knowledge-mining-book/
##
## If you use clusterProfiler in published research, please cite:
## T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng, L Zhou, W Tang, L Zhan, X Fu, S Liu, X Bo, and G Yu.
```

```
##
```

```
## Attaching package: 'clusterProfiler'

## The following object is masked from 'package:stats':
##
##     filter
```

```
library(ggplot2)
```

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

```
library(ggpubr)
```

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

```
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
```

```
library(dplyr)
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 object is masked from 'package:clusterProfiler':
##
##     rename
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
## Loading required package: IRanges
## Warning: package 'IRanges' was built under R version 4.2.1
##
## Attaching package: 'IRanges'
## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice
## The following object is masked from 'package:clusterProfiler':
##
##     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'

library(enrichplot)

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

##
## Attaching package: 'enrichplot'

## The following object is masked from 'package:ggpubr':
##
##     color_palette

library(ggupset)

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

library(tidyverse)

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

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

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

```
## Warning: package 'readr' was built under R version 4.2.3
## Warning: package 'purrr' was built under R version 4.2.3
## Warning: package 'stringr' was built under R version 4.2.3
## Warning: package 'forcats' was built under R version 4.2.3
## Warning: package 'lubridate' was built under R version 4.2.3
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## v readr     2.1.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x lubridate::%within%()    masks IRanges::%within%()
## x IRanges::collapse()      masks dplyr::collapse()
## x Biobase::combine()       masks BiocGenerics::combine(), dplyr::combine()
## x matrixStats::count()     masks dplyr::count()
## x IRanges::desc()          masks dplyr::desc()
## x tidyr::expand()          masks S4Vectors::expand()
## x dplyr::filter()          masks clusterProfiler::filter(), stats::filter()
## x S4Vectors::first()       masks dplyr::first()
## x dplyr::lag()             masks stats::lag()
## x BiocGenerics::Position() masks ggplot2::Position(), base::Position()
## x purrr::reduce()          masks GenomicRanges::reduce(), IRanges::reduce()
## x S4Vectors::rename()      masks dplyr::rename(), clusterProfiler::rename()
## x lubridate::second()      masks S4Vectors::second()
## x lubridate::second<-()    masks S4Vectors::second<-()
## x purrr::simplify()        masks clusterProfiler::simplify()
## x IRanges::slice()         masks dplyr::slice(), clusterProfiler::slice()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
library(org.Hs.eg.db)
## Loading required package: AnnotationDbi
## Warning: package 'AnnotationDbi' was built under R version 4.2.2
##
## Attaching package: 'AnnotationDbi'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
## The following object is masked from 'package:clusterProfiler':
##
##     select
library(DOSE)
## Warning: package 'DOSE' was built under R version 4.2.1
## DOSE v3.22.1  For help: https://yulab-smu.top/biomedical-knowledge-mining-book/
##
## If you use DOSE in published research, please cite:
## Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, Qing-Yu He. DOSE: an R/Bioconductor package for Disease
```

```
##############################
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, ]

################PatientInfo
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",]


#############Creating a metadata table with information of our samples
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<-as.data.frame(log2(expression_data+1)) # Adding 1 to avoid log(0)
expression_data<- round(expression_data)
expression_data<- as.data.frame(t(expression_data))

#########################################Create 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
## -- note: fitType='parametric', but the dispersion trend was not well captured by the
##    function: y = a/x + b, and a local regression fit was automatically substituted.
##    specify fitType='local' or 'mean' to avoid this message next time.
## final dispersion estimates
## fitting model and testing
```

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

Selecting DEGs

```r
Genes_in_data <- My_Results_df$GeneName
My_Results_df <- na.omit(My_Results_df)

My_Results_df <- My_Results_df %>% mutate(diffexpressed = case_when(
  log2FoldChange       > 0 & padj < 0.05 ~ 'Up',
  log2FoldChange       < 0 & padj < 0.05 ~ 'Down',
  padj > 0.05 ~ 'NO'))

My_Results_df<-My_Results_df[My_Results_df$diffexpressed !='NO',]
# Split the dataframe into a list of sub-dataframes: upregulated, downregulated genes
deg_results_list <- split(My_Results_df, My_Results_df$diffexpressed)
```

GSEA

Load MSigDB Hallmark gene sets

```r
hallmark_gene_sets <- read.gmt("h.all.v2023.2.Hs.symbols.gmt")
hallmark_gene_sets2 <- hallmark_gene_sets[hallmark_gene_sets$gene %in% Genes_in_data,]
```

Run clusterProfiler

```r
name_of_comparison <- 'ER'
background_genes <- 'hallmark_gene_sets' # for our filename
bg_genes <- hallmark_gene_sets2
padj_cutoff <- 0.05 # p-adjusted threshold, used to filter out pathways
genecount_cutoff <- 5 # minimum number of genes in the pathway, used to filter out pathways

# Run clusterProfiler on each sub-dataframe
res <- lapply(names(deg_results_list),
              function(x) enricher(gene = deg_results_list[[x]]$GeneName,
                                   TERM2GENE = bg_genes))
names(res) <- names(deg_results_list)
```

Convert the list of enrichResults for each sample_pattern to a dataframe with the pathways

```r
res_df <- lapply(names(res), function(x) rbind(res[[x]]@result))
names(res_df) <- names(res)
res_df <- do.call(rbind, res_df)
head(res_df)
```

```
##                                                                                                ID
## Down.HALLMARK_E2F_TARGETS                                     HALLMARK_E2F_TARGETS
## Down.HALLMARK_G2M_CHECKPOINT                                 HALLMARK_G2M_CHECKPOINT
## Down.HALLMARK_ALLOGRAFT_REJECTION                       HALLMARK_ALLOGRAFT_REJECTION
## Down.HALLMARK_KRAS_SIGNALING_DN                           HALLMARK_KRAS_SIGNALING_DN
## Down.HALLMARK_INFLAMMATORY_RESPONSE                 HALLMARK_INFLAMMATORY_RESPONSE
## Down.HALLMARK_INTERFERON_GAMMA_RESPONSE HALLMARK_INTERFERON_GAMMA_RESPONSE
##                                                                                       Description
## Down.HALLMARK_E2F_TARGETS                                     HALLMARK_E2F_TARGETS
## Down.HALLMARK_G2M_CHECKPOINT                                 HALLMARK_G2M_CHECKPOINT
## Down.HALLMARK_ALLOGRAFT_REJECTION                       HALLMARK_ALLOGRAFT_REJECTION
## Down.HALLMARK_KRAS_SIGNALING_DN                           HALLMARK_KRAS_SIGNALING_DN
## Down.HALLMARK_INFLAMMATORY_RESPONSE                 HALLMARK_INFLAMMATORY_RESPONSE
## Down.HALLMARK_INTERFERON_GAMMA_RESPONSE HALLMARK_INTERFERON_GAMMA_RESPONSE
##                                         GeneRatio  BgRatio       pvalue
```

```
## Down.HALLMARK_E2F_TARGETS                   96/894 188/4003 8.190602e-19
## Down.HALLMARK_G2M_CHECKPOINT                90/894 179/4003 4.100105e-17
## Down.HALLMARK_ALLOGRAFT_REJECTION           82/894 178/4003 6.093042e-13
## Down.HALLMARK_KRAS_SIGNALING_DN             75/894 177/4003 8.652047e-10
## Down.HALLMARK_INFLAMMATORY_RESPONSE         75/894 186/4003 1.241787e-08
## Down.HALLMARK_INTERFERON_GAMMA_RESPONSE     66/894 183/4003 1.050654e-05
##                                              p.adjust       qvalue
## Down.HALLMARK_E2F_TARGETS                4.095301e-17 3.103807e-17
## Down.HALLMARK_G2M_CHECKPOINT             1.025026e-15 7.768620e-16
## Down.HALLMARK_ALLOGRAFT_REJECTION        1.015507e-11 7.696475e-12
## Down.HALLMARK_KRAS_SIGNALING_DN          1.081506e-08 8.196676e-09
## Down.HALLMARK_INFLAMMATORY_RESPONSE      1.241787e-07 9.411437e-08
## Down.HALLMARK_INTERFERON_GAMMA_RESPONSE  8.755449e-05 6.635709e-05
##
## Down.HALLMARK_E2F_TARGETS                ANP32E/ASF1B/AURKA/AURKB/BIRC5/BRCA2/BUB1B/MMS22L/CCNB2/CCNE
## Down.HALLMARK_G2M_CHECKPOINT                                                                      AMD1/AU
## Down.HALLMARK_ALLOGRAFT_REJECTION
## Down.HALLMARK_KRAS_SIGNALING_DN
## Down.HALLMARK_INFLAMMATORY_RESPONSE
## Down.HALLMARK_INTERFERON_GAMMA_RESPONSE
##                                          Count
## Down.HALLMARK_E2F_TARGETS                   96
## Down.HALLMARK_G2M_CHECKPOINT                90
## Down.HALLMARK_ALLOGRAFT_REJECTION           82
## Down.HALLMARK_KRAS_SIGNALING_DN             75
## Down.HALLMARK_INFLAMMATORY_RESPONSE         75
## Down.HALLMARK_INTERFERON_GAMMA_RESPONSE     66
```

```
res_df <- res_df %>% mutate(minuslog10padj = -log10(p.adjust) )
```

Subset to those pathways that have p adj < cutoff and gene count > cutoff (you can also do this in the enricher function)

```
target_pws <- unique(res_df$ID[res_df$p.adjust < padj_cutoff & res_df$Count > genecount_cutoff]) # sele
res_df <- res_df[res_df$ID %in% target_pws, ]
```

PLOT

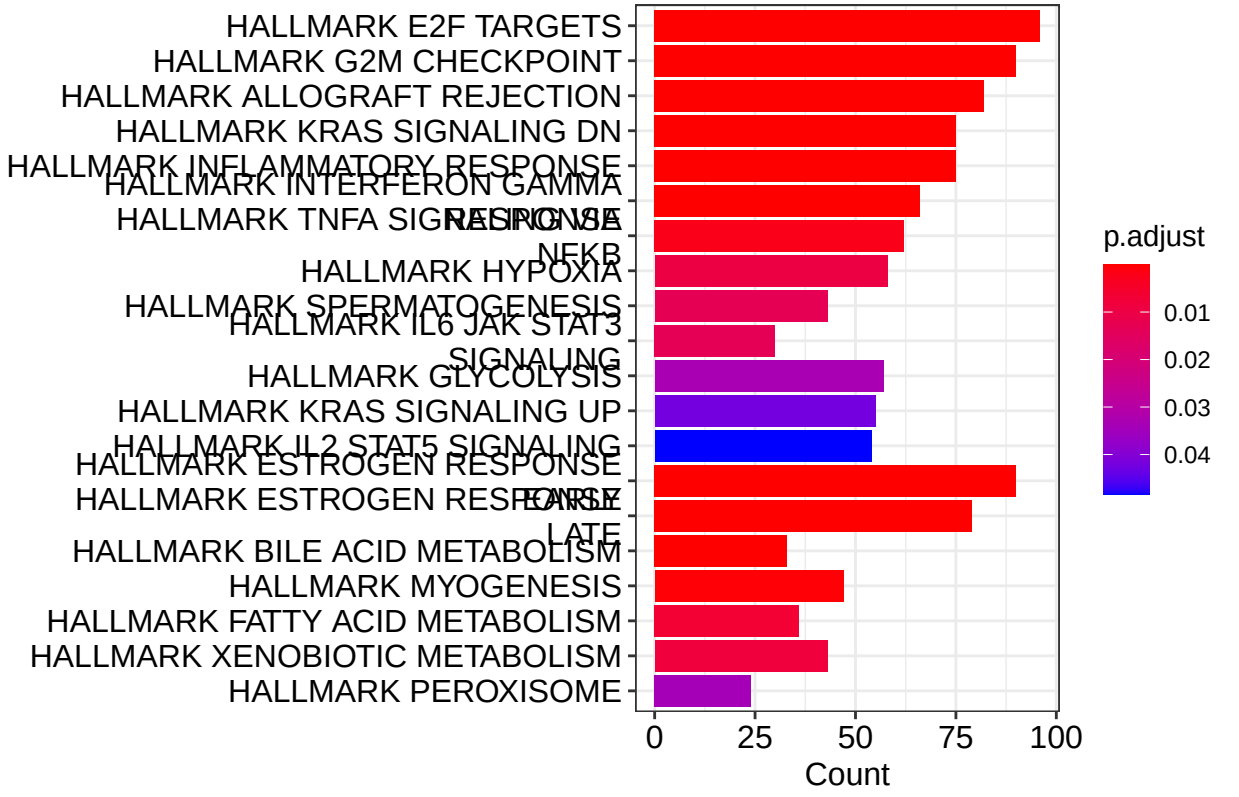
enrichres object

```
enrichres <- new("enrichResult",
                 readable = FALSE,
                 result = res_df,
                 pvalueCutoff = 0.05,
                 pAdjustMethod = "BH",
                 qvalueCutoff = 0.2,
                 organism = "human",
                 ontology = "UNKNOWN",
                 gene =Genes_in_data,
                 keytype = "UNKNOWN",
                 universe = unique(bg_genes$gene),
                 gene2Symbol = character(0),
                 geneSets = bg_genes)
class(enrichres)
```

```
## [1] "enrichResult"
## attr(,"package")
## [1] "DOSE"
```

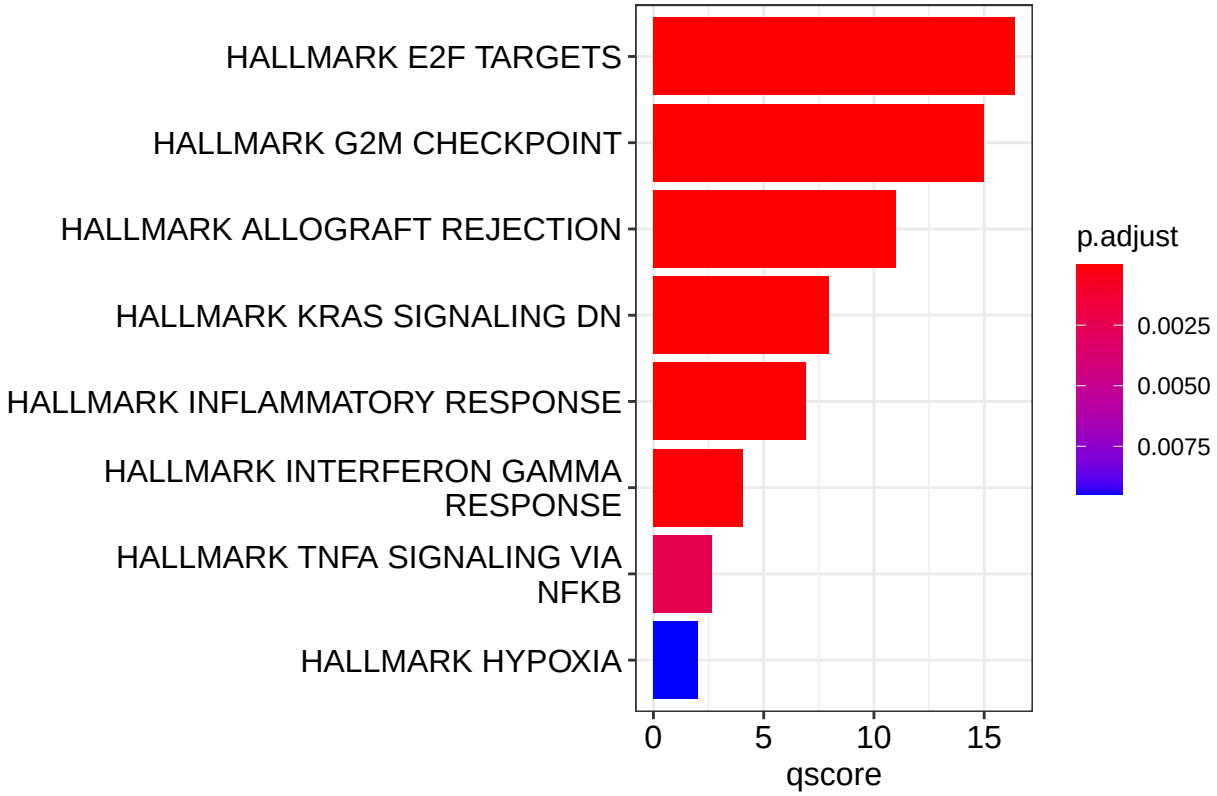
Barplot

```
barplot(enrichres, showCategory = 20)
```



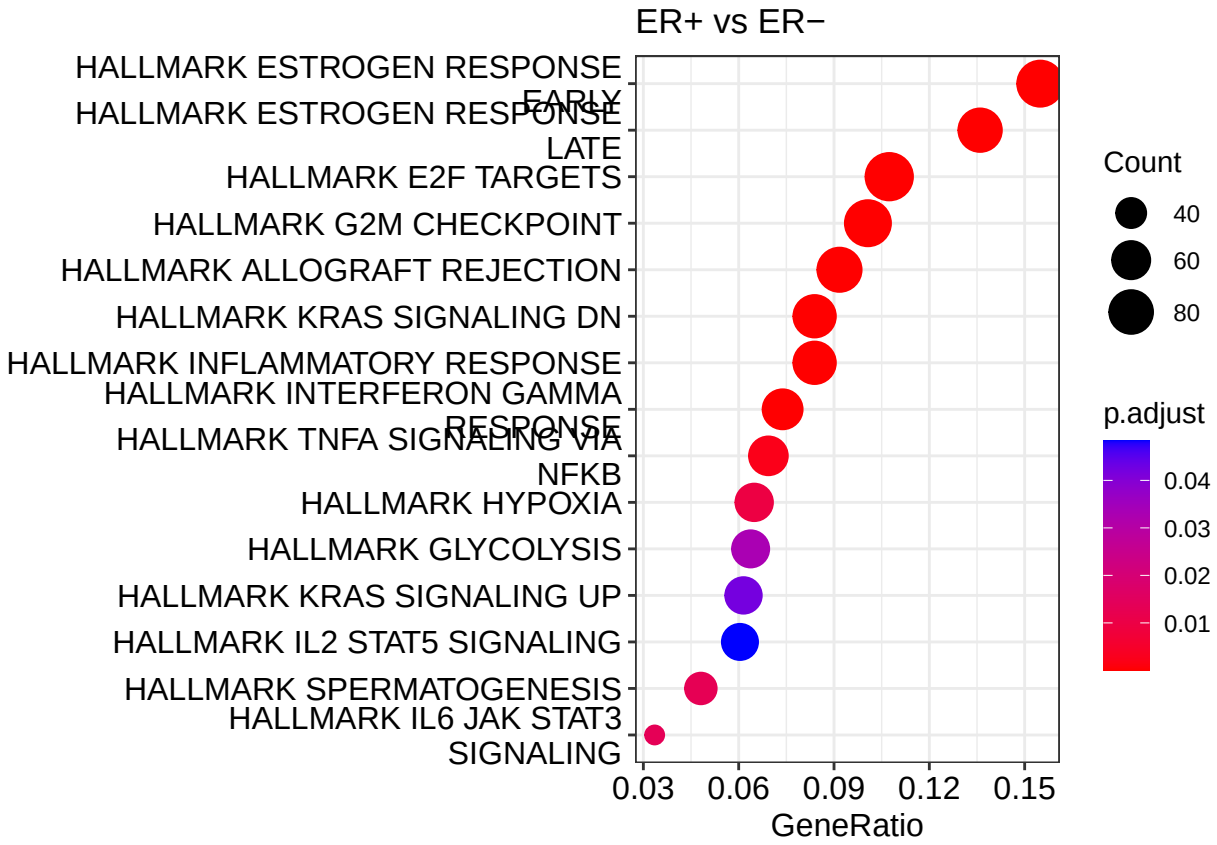
Sorted Barplot

```
mutate(enrichres, qscore = -log(p.adjust, base = 10)) %>%
  barplot(x = "qscore")
```



dot plot

```
dotplot(enrichres, showCategory = 15) + ggtitle("ER+ vs ER-")
```



Conclusion

We ran GSEA to compare ER+ and ER- tumors in breast cancer. As expected, alongside many other pathways, Estrogen Receptor pathways were significantly enriched in ER+ tumors.