Enrichment analysis for breast cancer

Loading required libraries

```
library(DESeq2)
library(clusterProfiler)
library(org.Hs.eg.db)
library(pathview)
library(biomaRt)
library(ReactomePA)
library(ggplot2)
library(RColorBrewer)
library(ggsci)
```

```
## Warning: package 'ggsci' was built under R version 4.4.1
```

Read count matrix into CSV

```
read_count_data <- function(file_path){
  counts_data <- read.csv(file_path, row.names = 1)
  expression_data <- round(counts_data)
  return(expression_data)
}

expression_matrix <- read_count_data("../Enrichment analysis for breast cancer/Data/GSE183947_fpkm.csv"

head(expression_matrix,2)
```

```
##        tumor.rep1 tumor.rep2 tumor.rep3 tumor.rep4 tumor.rep5 tumor.rep6
## TSPAN6          1          2          0          5          5          5
## TNMD            0          0          0          0          0          0
##        tumor.rep7 tumor.rep8 tumor.rep9 tumor.rep10 tumor.rep11 tumor.rep12
## TSPAN6          4          4          6          12           6           4
## TNMD            0          0          0           0           0           0
##        tumor.rep13 tumor.rep14 tumor.rep15 tumor.rep16 tumor.rep17 tumor.rep18
## TSPAN6           8          11           4          14          10           7
## TNMD             0           1           0           0           0           0
##        tumor.rep19 tumor.rep20 tumor.rep21 tumor.rep22 tumor.rep23 tumor.rep24
## TSPAN6           6           7           9           7          10           7
## TNMD             0           0           0           0           0           0
##        tumor.rep25 tumor.rep26 tumor.rep27 tumor.rep28 tumor.rep29 tumor.rep30
## TSPAN6           5           2           5           5           9           2
## TNMD             0           0           0           1           0           1
##        normal.rep1 normal.rep2 normal.rep3 normal.rep4 normal.rep5 normal.rep6
## TSPAN6          12           3          13          15           7           0
## TNMD             6           2           0           2           0           0
##        normal.rep7 normal.rep8 normal.rep9 normal.rep10 normal.rep11
## TSPAN6          10           7           5            6            6
## TNMD             0           0          11            0            3
##        normal.rep12 normal.rep13 normal.rep14 normal.rep15 normal.rep16
## TSPAN6            7           11           16           12           12
## TNMD              0            0            0            0            1
##        normal.rep17 normal.rep18 normal.rep19 normal.rep20 normal.rep21
## TSPAN6           10            9            7            9            7
## TNMD              1            0            0            0            0
##        normal.rep22 normal.rep23 normal.rep24 normal.rep25 normal.rep26
## TSPAN6            8            9            6            6            6
## TNMD              0            1            0            0            0
##        normal.rep27 normal.rep28 normal.rep29 normal.rep30
## TSPAN6            4            5           10            5
## TNMD              9            1            0            0
```

Read metadata into CSV

```r
read_metadata <- function(file_path){
  coldata <- read.csv(file_path, row.names = 1)
  return (coldata)
}

meta_data <- read_metadata("../Enrichment analysis for breast cancer/Data/metadata.csv")
head(meta_data)
```

```
##            condition description
## tumor rep1     tumor   CA.102548
## tumor rep2     tumor   CA.104338
## tumor rep3     tumor   CA.105094
## tumor rep4     tumor   CA.109745
## tumor rep5     tumor  CA.1906415
## tumor rep6     tumor  CA.1912627
```

Convert condition column in metadata to factor

```r
meta_data$condition <- as.factor(meta_data$condition)
meta_data$description <- as.factor(meta_data$description)
```

Make sure the row names in metadata matches to the column names in expression matrix

```r
all(rownames(meta_data) %in% colnames(expression_matrix))
```

```
## [1] FALSE
```

Match the row names in metadata to the column names in expression matrix

```r
rownames(meta_data) = colnames(expression_matrix)
```

Construct a DESeqDataSet.

```r
deseqdataset <- function(){
  deseqdataset <- DESeqDataSetFromMatrix(countData = expression_matrix,
                                         colData = meta_data,
                                         design = ~ condition)

  return(deseqdataset)
}

deseqdataset_object <- deseqdataset()
```

```
## converting counts to integer mode
```

```r
deseqdataset_object
```

```
## class: DESeqDataSet
## dim: 20246 60
## metadata(1): version
## assays(1): counts
## rownames(20246): TSPAN6 TNMD ... RP11-474G23.1 AC005358.1
## rowData names(0):
## colnames(60): tumor.rep1 tumor.rep2 ... normal.rep29 normal.rep30
## colData names(2): condition description
```

Pre-filtering: removing rows with low gene counts

keep rows that have at least 10 reads total

```r
pre_filter <- function(){
  keep <- rowSums(counts(deseqdataset_object)) >= 10
  deseqdataset <- deseqdataset_object[keep,]
  return (deseqdataset)
}

filtered_expression_counts <- pre_filter()
head(filtered_expression_counts)
```

```
## class: DESeqDataSet
## dim: 6 60
## metadata(1): version
## assays(1): counts
## rownames(6): TSPAN6 TNMD ... C1orf112 FGR
## rowData names(0):
## colnames(60): tumor.rep1 tumor.rep2 ... normal.rep29 normal.rep30
## colData names(2): condition description
```

Differential expression analysis

```
diff_expr_analysis <- function(){
  deseq_analysis <- DESeq(deseqdataset_object)
  result <- results(deseq_analysis)
  return (result)
}

deseq_result <- diff_expr_analysis()
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 944 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

```
deseq_result
```

```
## log2 fold change (MLE): condition tumor vs normal
## Wald test p-value: condition tumor vs normal
## DataFrame with 20246 rows and 6 columns
##               baseMean log2FoldChange     lfcSE      stat     pvalue
##              <numeric>      <numeric> <numeric> <numeric>  <numeric>
## TSPAN6        7.021979      -0.392290  0.198452  -1.97675 0.04806938
## TNMD          0.741177      -2.612570  1.097660  -2.38013 0.01730672
## DPM1          6.355119       0.404755  0.288440   1.40326 0.16054029
## SCYL3         6.478091       0.533397  0.177903   2.99825 0.00271533
## C1orf112      8.698340       0.275094  0.208244   1.32102 0.18649502
## ...                 ...            ...      ...        ...        ...
## RP11-1084J3.4  0.184986      0.0967793  1.193944  0.0810585  0.9353954
## RP11-944L7.5   0.000000             NA        NA         NA         NA
## FLJ00388       0.214750     -0.4802983  1.840875 -0.2609077  0.7941637
## RP11-474G23.1  0.203916     -0.1917588  0.996792 -0.1923760  0.8474477
## AC005358.1     0.642168     -2.8028670  1.101281 -2.5450981  0.0109247
##                    padj
##               <numeric>
## TSPAN6        0.1125225
## TNMD          0.0506295
## DPM1          0.2776822
## SCYL3         0.0115777
## C1orf112      0.3092731
## ...                 ...
## RP11-1084J3.4        NA
## RP11-944L7.5         NA
## FLJ00388             NA
## RP11-474G23.1        NA
## AC005358.1    0.0350874
```

Convert DESeq result into DataFrame

```
df_deseq_result <- as.data.frame(deseq_result)
```

Extract differentially expressed genes that have padj <= 0.01

```
sig_genes <- function(){
significant_genes <- df_deseq_result[df_deseq_result$padj <= 0.01,] %>% na.omit(significant_genes)
ordered_sig_genes <- significant_genes[order(significant_genes$padj, decreasing = FALSE), ]
return(ordered_sig_genes)
}

sign_genes <- sig_genes()
head(sign_genes)
```

```
##         baseMean log2FoldChange     lfcSE      stat       pvalue         padj
## DEFB130 13.34575      -4.632775 0.2687001 -17.24143 1.297799e-66 2.448818e-62
## LCN6    35.01702      -3.524994 0.2095224 -16.82395 1.629301e-63 1.537164e-59
## CCDC177 12.20217      -4.314318 0.2674657 -16.13036 1.561155e-58 9.819141e-55
## SOX7    85.13424      -2.776410 0.1959286 -14.17052 1.394592e-45 6.578637e-42
## MDGA2   17.26324      -2.858934 0.2021341 -14.14375 2.041162e-45 7.702936e-42
## KLK9    15.08985      -3.979599 0.2944088 -13.51725 1.237090e-41 3.890441e-38
```

Write significant genes into CSV file

```r
write_sig_genes <- function(out_path){
  write.csv(sign_genes, file = out_path )
}

write_sig_genes("../Enrichment analysis for breast cancer/outputs/significant_genes.csv")
```

Convert Gene SYMBOLs to ENTREZ IDs

```r
entrez_ids <- function(){
# copy the rownames of significant genes and store it in gene_names
gene_names <- rownames(sign_genes)
# convert gene names into ENTREZID
entrez_ids <- mapIds(org.Hs.eg.db,
                     keys = gene_names,
                     column = "ENTREZID",
                     keytype = "SYMBOL",
                     multiVals = "first")

# create column named ENTREZID in sign_genes that contain ENTREZID
# of significant genes
sign_genes$ENTREZID <- entrez_ids

# remove ENTREZID that contain NA
sign_genes <- sign_genes[!is.na(sign_genes$ENTREZID), ]
return(sign_genes)
}

signficant <- entrez_ids()
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```r
head(signficant)
```

```
##          baseMean log2FoldChange     lfcSE      stat       pvalue         padj
## LCN6     35.01702      -3.524994 0.2095224 -16.82395 1.629301e-63 1.537164e-59
## CCDC177  12.20217      -4.314318 0.2674657 -16.13036 1.561155e-58 9.819141e-55
## SOX7     85.13424      -2.776410 0.1959286 -14.17052 1.394592e-45 6.578637e-42
## MDGA2    17.26324      -2.858934 0.2021341 -14.14375 2.041162e-45 7.702936e-42
## KLK9     15.08985      -3.979599 0.2944088 -13.51725 1.237090e-41 3.890441e-38
## UGT2A1   10.74854      -2.658282 0.2033343 -13.07345 4.669720e-39 1.258756e-35
##          ENTREZID
## LCN6       158062
## CCDC177     56936
## SOX7        83595
## MDGA2      161357
## KLK9       284366
## UGT2A1      10941
```

Up-regulated significant genes

```
up_sig_genes <- function(){
  upregulated_genes <- signficant[signficant$log2FoldChange>0 , ]
  return(upregulated_genes)
}

upregulated_sig_genes <- up_sig_genes()
head(upregulated_sig_genes,5)
```

```
##          baseMean log2FoldChange     lfcSE     stat       pvalue         padj
## MYBL2   12.758600       3.205627 0.3210231 9.985659 1.761263e-23 1.145975e-20
## E2F1     8.894549       2.548089 0.2597684 9.809079 1.029033e-22 6.472275e-20
## MMP11   96.083267       3.434537 0.3703246 9.274396 1.786286e-20 8.426358e-18
## MTHFD2  33.815814       1.868244 0.2042300 9.147746 5.813359e-20 2.562632e-17
## CXCL10  11.179072       3.759932 0.4118765 9.128784 6.927270e-20 2.895567e-17
##         ENTREZID
## MYBL2       4605
## E2F1        1869
## MMP11       4320
## MTHFD2     10797
## CXCL10      3627
```

ENTREZIDs of up-regulated significant genes

```
up_entrez_ids <- upregulated_sig_genes$ENTREZID
head(up_entrez_ids)
```

```
## [1] "4605"  "1869"  "4320"  "10797" "3627"  "51203"
```

Down-regulated significant genes

```
down_sig_genes <- function(){
  downregulated_genes <- signficant[signficant$log2FoldChange < 0 , ]
  return(downregulated_genes)
}

downregulated_sig_genes <- down_sig_genes()
head(downregulated_sig_genes,5)
```

```
##          baseMean log2FoldChange     lfcSE      stat       pvalue         padj
## LCN6     35.01702      -3.524994 0.2095224 -16.82395 1.629301e-63 1.537164e-59
## CCDC177 12.20217      -4.314318 0.2674657 -16.13036 1.561155e-58 9.819141e-55
## SOX7     85.13424      -2.776410 0.1959286 -14.17052 1.394592e-45 6.578637e-42
## MDGA2    17.26324      -2.858934 0.2021341 -14.14375 2.041162e-45 7.702936e-42
## KLK9     15.08985      -3.979599 0.2944088 -13.51725 1.237090e-41 3.890441e-38
##         ENTREZID
## LCN6      158062
## CCDC177    56936
## SOX7       83595
## MDGA2     161357
## KLK9      284366
```

ENTREZIDs of down-regulated significant genes

```
down_entrez_ids <- downregulated_sig_genes$ENTREZID
head(down_entrez_ids)
```

```
## [1] "158062" "56936"  "83595"  "161357" "284366" "10941"
```

Gene Ontology

Group up-regulated significant genes that have similar BP GO terms

```
go_up <- function(){
    go <- groupGO( gene = up_entrez_ids,
                   OrgDb = org.Hs.eg.db,
                   ont = "BP",  # Biological Process
                   readable = TRUE)
    return(go)

}

go_terms_up <- go_up()
head(go_terms_up)
```

```
##                      ID           Description Count GeneRatio
## GO:0000003 GO:0000003          reproduction   132  132/1608
## GO:0002376 GO:0002376 immune system process   262  262/1608
## GO:0008152 GO:0008152     metabolic process  1061 1061/1608
## GO:0009987 GO:0009987      cellular process  1425 1425/1608
## GO:0016032 GO:0016032         viral process    72   72/1608
## GO:0022414 GO:0022414  reproductive process   131  131/1608
##
## GO:0000003
## GO:0002376
## GO:0008152
## GO:0009987 MYBL2/E2F1/MMP11/MTHFD2/CXCL10/NUSAP1/TK1/IDH2/CCNB1/COL10A1/PYCR1/PITX1/FN1/LMNB1/TROAP/
## GO:0016032
## GO:0022414
```

Convert GO terms of up-regulated significant genes to DataFrame

```
dataframe_go_up <- function(){
  df_go_terms_up <- as.data.frame(go_terms_up)
  return(df_go_terms_up)
}

df_go_group_up <- dataframe_go_up()
```

```
df_go_group_up_top7 <- head(df_go_group_up,7)
```

Barplot of up-regulated BP GO terms

```
bar_plot_up <- function(){
p <- ggplot(df_go_group_up_top7, aes(x = reorder(Description, - Count),
                                     y = Count, fill = Description)) +
                                     geom_bar(stat = "identity") +
                                     ggtitle("BP of up-regulated GO terms") +
                                     coord_flip() +
                                     theme_bw() +
                                     scale_fill_jama()+
                                     theme(plot.title = element_text(size = 12,
                                                                face = "bold",
                                                                hjust = 0.5))+
                                     xlab("Description")


  jpeg("../Enrichment analysis for breast cancer/outputs/BPgroup_up_barplot.jpeg")
  print(p)
  dev.off()

}

bar_plot_up()
```

```
## pdf
##   2
```

Group down-regulated significant genes that have similar BP GO terms

```
go_down <- function(){
  go <- groupGO(gene = down_entrez_ids,
                OrgDb = org.Hs.eg.db,
                ont = "BP",
                readable = TRUE)
  return(go)
}
```

```
go_terms_down <- go_down()
head(go_terms_down)
```

```
##                    ID           Description Count GeneRatio
## GO:0000003 GO:0000003          reproduction   214  214/2267
## GO:0002376 GO:0002376 immune system process   264  264/2267
## GO:0008152 GO:0008152     metabolic process  1196 1196/2267
## GO:0009987 GO:0009987      cellular process  1886 1886/2267
## GO:0016032 GO:0016032         viral process    28   28/2267
## GO:0022414 GO:0022414  reproductive process   213  213/2267
##
## GO:0000003
## GO:0002376
## GO:0008152
## GO:0009987 SOX7/MDGA2/UGT2A1/OC90/GOLGA7B/SYNPO2/ZNF709/SIGLEC5/LTB4R2/DES/MYH11/MRPS17/CLN3/ABCA8/M
## GO:0016032
## GO:0022414
```

Convert GO terms of down-regulated significant genes to DataFrame

```
dataframe_go_down <- function(){
  df_go_terms_down <- as.data.frame(go_terms_down)
  return(df_go_terms_down)
}

df_go_group_down <- dataframe_go_down()
```

```
df_go_group_down_top7 <- head(df_go_group_down,7)
```

Barplot of down-regulated BP GO terms

```
bar_plot_down <- function(){

p <- ggplot(df_go_group_down_top7, aes(x = reorder(Description, - Count),
                                       y = Count, fill = Description)) +
                                       geom_bar(stat = "identity") +
                                       ggtitle("BP of down-regulated GO terms")+
                                       coord_flip() +
                                       theme_bw() +
                                       scale_fill_jama()+
                                      theme(plot.title = element_text(size = 12,
                                                          face = "bold",
                                                          hjust = 0.5))+
                                       xlab("Description")


  jpeg("../Enrichment analysis for breast cancer/outputs/BPgroup_down_barplot.jpeg")
  print(p)
  dev.off()

}

bar_plot_down()
```

```
## pdf
##   2
```

Over-representation analysis

Go enrichment analysis

Enriched GO terms among up-regulated significant genes

```
enrich_go_up <- function(){
        ego_up <- enrichGO( gene          = up_entrez_ids,
                            universe      = signficant$ENTREZID,
                            OrgDb         = org.Hs.eg.db,
                            keyType       = "ENTREZID",
                            ont           = "BP",
                            pvalueCutoff  = 0.05,
                            qvalueCutoff  = 0.01,
                            pAdjustMethod = "BH",
                            readable      = TRUE)
        return(ego_up)

}

enrichment_go_up <- enrich_go_up()
head(enrichment_go_up)
```

```
##                      ID                    Description GeneRatio  BgRatio
## GO:0051276 GO:0051276        chromosome organization  122/1506 145/3507
## GO:0006259 GO:0006259          DNA metabolic process  168/1506 236/3507
## GO:0007059 GO:0007059         chromosome segregation   85/1506 102/3507
## GO:0007049 GO:0007049                     cell cycle  259/1506 419/3507
## GO:0098813 GO:0098813 nuclear chromosome segregation   72/1506  84/3507
## GO:0000819 GO:0000819   sister chromatid segregation   57/1506  63/3507
##                  pvalue     p.adjust       qvalue
## GO:0051276 1.715341e-25 5.290112e-22 4.526695e-22
## GO:0006259 1.015358e-19 1.565681e-16 1.339738e-16
## GO:0007059 1.843292e-17 1.894904e-14 1.621450e-14
## GO:0007049 9.862727e-17 7.604162e-14 6.506804e-14
## GO:0098813 2.646066e-16 1.632094e-13 1.396566e-13
## GO:0000819 1.786669e-15 9.183477e-13 7.858208e-13
##
## GO:0051276
## GO:0006259
## GO:0007059
## GO:0007049 MYBL2/E2F1/NUSAP1/TK1/CCNB1/PLK1/AURKB/FANCA/MCM2/CDCA8/STMN1/MCM4/FOXM1/KIFC1/KIF2C/NCAP
## GO:0098813
## GO:0000819
##            Count
## GO:0051276   122
## GO:0006259   168
## GO:0007059    85
## GO:0007049   259
## GO:0098813    72
## GO:0000819    57
```

Dataframe of enriched GO terms among up-regulated significant genes

```
df_go_terms_up <- function(){
  df_go_term_up <- as.data.frame(enrichment_go_up)
  return(df_go_term_up)
}

df_go_up <- df_go_terms_up()
```

```
df_go_up_top7 <- head(df_go_up, 7)
```

BarPlot of up-regulated BP enriched GO terms

```
bar_plot_enriched_up <- function() {

  jama_colors <- pal_jama("default")(7)
  p <- ggplot(df_go_up_top7, aes(x = reorder(Description, -Count),
                                 y = Count,
                                 fill = p.adjust)) +
       geom_bar(stat = "identity") +
       coord_flip() +
       scale_fill_gradientn(name = "p.adjust", colors = jama_colors) +
       labs(title = "Bar plot of BP of Enriched Up-regulated GO terms",
            x = "GO Term",
            y = "Gene Count") +
       theme_bw() +
       theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))

  jpeg("../Enrichment analysis for breast cancer/outputs/BP_enriched_up_barplot.jpeg")
  print(p)
  dev.off()
}

bar_plot_enriched_up()
```

```
## pdf
##   2
```

Dotplot of up-regulated BP enriched GO terms

```r
dot_plot_enriched_up <- function(){
  jama_colors <- pal_jama("default")(7)
  p <- ggplot(df_go_up_top7, aes(x = reorder(Description, -Count),
                                  y = Count, size = Count, color = p.adjust)) +
    geom_point(alpha = 0.6) +
    coord_flip() +
    scale_size_continuous(range = c(3, 8), name = "Gene Count") +
    scale_color_gradientn(name = "p.adjust", colors = jama_colors) +
    labs(title = "Dot plot of BP of enriched up-regulated genes",
         x = "GO Term",
         y = "Gene Count") +
    theme_bw() +
    theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))


  jpeg("../Enrichment analysis for breast cancer/outputs/BP_enriched_up_dotplot.jpeg")
  print(p)
  dev.off()
}

dot_plot_enriched_up()
```

```
## pdf
##   2
```

```r
jpeg("../Enrichment analysis for breast cancer/outputs/Network_plot_up.jpeg")
cnet_plot_up <- cnetplot(enrichment_go_up, showCategory = 2, vertex.label.cex = 1.2)
print(cnet_plot_up)
```

```
## Warning: ggrepel: 127 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

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

```
## pdf
##   2
```

```r
jpeg("../Enrichment analysis for breast cancer/outputs/GO_graph_up.jpeg")
go_graph_up <- plotGOgraph(enrichment_go_up)
```

```
##
## groupGOTerms:    GOBPTerm, GOMFTerm, GOCCTerm environments built.
```

```
##
## Building most specific GOs .....
```

```
##  ( 8231 GO terms found. )
```

```
##
## Build GO DAG topology ..........

##  ( 8231 GO terms and 18270 relations. )

##
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
##
##     backsolve

##
## Annotating nodes ...............

##  ( 3507 genes annotated to the GO terms. )

## Loading required package: Rgraphviz

## Loading required package: graph

## Loading required package: grid

##
## Attaching package: 'Rgraphviz'

## The following objects are masked from 'package:IRanges':
##
##     from, to

## The following objects are masked from 'package:S4Vectors':
##
##     from, to
```

```r
print(go_graph_up)
```

```
## $dag
## A graphNEL graph with directed edges
## Number of Nodes = 30
## Number of Edges = 42
##
## $complete.dag
## [1] "A graph with 30 nodes."
```

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

```
## pdf
##   2
```

Enriched GO terms among down-regulated significant genes

```
enrich_go_down <- function(){
        ego_down <- enrichGO( gene          = down_entrez_ids,
                              universe      = signficant$ENTREZID,
                              OrgDb         = org.Hs.eg.db,
                              keyType       = "ENTREZID",
                              ont           = "BP",
                              pvalueCutoff  = 0.05,
                              qvalueCutoff  = 0.01,
                              pAdjustMethod = "BH",
                              readable      = TRUE)
        return(ego_down)

}

enrichment_go_down <- enrich_go_down()
head(enrichment_go_down)
```

```
##                     ID                  Description GeneRatio  BgRatio
## GO:0003008 GO:0003008               system process  353/2001 463/3507
## GO:0007267 GO:0007267          cell-cell signaling  317/2001 425/3507
## GO:0044057 GO:0044057 regulation of system process  129/2001 154/3507
## GO:0050877 GO:0050877       nervous system process  191/2001 246/3507
## GO:0008015 GO:0008015            blood circulation  131/2001 160/3507
## GO:0099537 GO:0099537     trans-synaptic signaling  148/2001 185/3507
##                  pvalue     p.adjust       qvalue
## GO:0003008 2.627147e-20 8.149409e-17 6.327275e-17
## GO:0007267 1.075283e-15 1.667764e-12 1.294867e-12
## GO:0044057 4.676197e-13 4.835188e-10 3.754084e-10
## GO:0050877 2.137514e-12 1.657642e-09 1.287009e-09
## GO:0008015 9.131271e-12 5.665040e-09 4.398389e-09
## GO:0099537 1.290363e-11 6.671177e-09 5.179563e-09
##
## GO:0003008 UGT2A1/DES/MYH11/CLN3/MFRP/FGF10/ANK2/GSTM2/CNN1/NTRK2/TLR9/CRYBG3/AKAP12/CACNA1G/PDE2A/L
## GO:0007267
## GO:0044057
## GO:0050877
## GO:0008015
## GO:0099537
##            Count
## GO:0003008   353
## GO:0007267   317
## GO:0044057   129
## GO:0050877   191
## GO:0008015   131
## GO:0099537   148
```

Dataframe of enriched GO terms among down-regulated significant genes

```
df_go_terms_down <- function(){
  df_go_term_down <- as.data.frame(enrichment_go_down)
  return(df_go_term_down)
}

df_go_down <- df_go_terms_down()
```

```
df_go_down_top7 <- head(df_go_down, 7)
```

BarPlot of down-regulated BP enriched GO terms

```
bar_plot_enriched_down<- function(){
  jama_colors <- pal_jama("default")(7)
  p <- ggplot(df_go_down_top7, aes(x = reorder(Description, - Count) ,
                                    y = Count, fill = p.adjust)) +
    geom_bar(stat = "identity") +
    coord_flip() +
    scale_fill_gradientn(name = "p.adjust", colors = jama_colors) +
    labs(title = "Bar plot of BP of Enriched down-regulated Genes",
         x = "GO Term",
         y = "Gene Count") +
    theme_bw() +
    theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))

  jpeg("../Enrichment analysis for breast cancer/outputs/BP_enriched_down_barplot.jpeg")
  print(p)
  dev.off()
}

bar_plot_enriched_down()
```

```
## pdf
##   2
```

Dotplot of down-regulated BP enriched GO terms

```
dot_plot_enriched_down<- function(){

  jama_colors <- pal_jama("default")(7)
  p <- ggplot(df_go_down_top7, aes(x = reorder(Description, -Count),
                                      y = Count, size = Count,
                                      color = p.adjust)) +
        geom_point(alpha = 0.6) +
        coord_flip() +
        scale_size_continuous(range = c(3, 8), name = "Gene Count") +
        scale_color_gradientn(name = "p.adjust", colors = jama_colors) +
        labs(title = "Dot plot of BP of enriched down-regulated genes",
             x = "GO Term",
             y = "Gene Count") +
        theme_bw() +
        theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))
  jpeg("../Enrichment analysis for breast cancer/outputs/BP_enriched_down_dotplot.jpeg",
       width = 700, height = 800)
  print(p)
  dev.off()
}

dot_plot_enriched_down()
```

```
## pdf
##   2
```

```
jpeg("../Enrichment analysis for breast cancer/outputs/GO_graph_down.jpeg")
go_graph_down <- plotGOgraph(enrichment_go_down)
```

```
##
## groupGOTerms:    GOBPTerm, GOMFTerm, GOCCTerm environments built.

##
## Building most specific GOs .....

##  ( 9320 GO terms found. )

##
## Build GO DAG topology ..........

##  ( 9320 GO terms and 20700 relations. )

##
## Annotating nodes ...............

##  ( 3507 genes annotated to the GO terms. )
```

```
print(go_graph_down)
```

```
## $dag
## A graphNEL graph with directed edges
## Number of Nodes = 23
## Number of Edges = 28
##
## $complete.dag
## [1] "A graph with 23 nodes."
```

```
dev.off()
```

```
## pdf
##   2
```

Pathway Enrichment Analysis

KEGG pathway enrichment analysis among up-regulated significant genes

```r
kegg_enrichment_up <- function(){

kegg_up <- enrichKEGG(gene = up_entrez_ids,
                      universe = signficant$ENTREZID,
                      organism = "hsa",
                      pvalueCutoff = 0.05,
                      qvalueCutoff = 0.01,
                      pAdjustMethod = "BH")
  return(kegg_up)

}

kegg_enrich_up <- kegg_enrichment_up()
```

```
## Reading KEGG annotation online: "https://rest.kegg.jp/link/hsa/pathway"...
```

```
## Reading KEGG annotation online: "https://rest.kegg.jp/list/pathway/hsa"...
```

```r
head(kegg_enrich_up)
```

```
##                                category                      subcategory
## hsa04110             Cellular Processes             Cell growth and death
## hsa05169                Human Diseases        Infectious disease: viral
## hsa03013 Genetic Information Processing                      Translation
## hsa04141 Genetic Information Processing Folding, sorting and degradation
## hsa04612             Organismal Systems                    Immune system
## hsa05014                Human Diseases        Neurodegenerative disease
##                ID                                 Description GeneRatio BgRatio
## hsa04110 hsa04110                                  Cell cycle    48/819 57/1827
## hsa05169 hsa05169                    Epstein-Barr virus infection    46/819 57/1827
## hsa03013 hsa03013                    Nucleocytoplasmic transport    24/819 25/1827
## hsa04141 hsa04141 Protein processing in endoplasmic reticulum    36/819 43/1827
## hsa04612 hsa04612          Antigen processing and presentation    23/819 25/1827
## hsa05014 hsa05014                Amyotrophic lateral sclerosis    57/819 85/1827
##                pvalue     p.adjust       qvalue
## hsa04110 5.325247e-10 1.379239e-07 1.132316e-07
## hsa05169 1.909713e-08 2.473078e-06 2.030326e-06
## hsa03013 5.177231e-08 4.469676e-06 3.669476e-06
## hsa04141 1.234732e-07 7.994892e-06 6.563577e-06
## hsa04612 8.136862e-07 4.214895e-05 3.460308e-05
## hsa05014 1.958636e-05 8.454779e-04 6.941131e-04
##
## hsa04110                                                                1869/891/5347/9212/4171/4173/
## hsa05169                                                                               1869/3627/6890
## hsa03013
## hsa04141
## hsa04612
## hsa05014 56893/23225/581/7186/203068/10376/637/4728/5710/4708/79139/9631/10762/5690/7388/5688/4704/84
##          Count
## hsa04110    48
## hsa05169    46
## hsa03013    24
## hsa04141    36
## hsa04612    23
## hsa05014    57
```

DataFrame of KEGG enriched among up-regulated significant genes

```r
dataframe_kegg_up <- function(){
  df_kegg_up_genes <- as.data.frame(kegg_enrich_up)
  return(df_kegg_up_genes)
}

df_kegg_up <- dataframe_kegg_up()
```

```r
df_kegg_up_top7 <- head(df_kegg_up, 7)
```

Barplot for KEGG enriched up-regulated significant gene

```r
bar_plot_kegg_enriched_up<- function(){

  jama_colors <- pal_jama("default")(7)
  p <- ggplot(df_kegg_up_top7, aes(x = reorder(Description, - Count),
                                   y = Count, fill = p.adjust)) +
         geom_bar(stat = "identity", width = 0.8) +
         coord_flip() +
         scale_fill_gradientn(name = "p.adjust", colors = jama_colors) +
         labs(title = "Bar plot of Enriched KEGG pathway of up-regulated Genes",
             x = "Enriched pathway",
             y = "Gene Count") +
         theme_bw() +
         theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))

  jpeg("../Enrichment analysis for breast cancer/outputs/kegg_enriched_up_barplot.jpeg",
  width = 700, height = 800)
  print(p)
  dev.off()

}

bar_plot_kegg_enriched_up()
```

```
## pdf
##   2
```

Dotplot for KEGG enriched up-regulated significant gene

```r
dot_plot_kegg_enriched_up<- function(){
  jama_colors <- pal_jama("default")(7)
  p  <- ggplot(df_kegg_up_top7, aes(x = reorder(Description, -Count),
                                   y = Count, size = Count, color = p.adjust))+
        geom_point(alpha = 0.6) +
        coord_flip() +
        scale_size_continuous(range = c(3, 8), name = "Gene Count") +
        scale_color_gradientn(name = "p.adjust", colors = jama_colors) +
        labs(title = "Dot plot of BP of enriched up-regulated genes",
             x = "GO Term",
             y = "Gene Count") +
        theme_bw() +
        theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))


  jpeg("../Enrichment analysis for breast cancer/outputs/kegg_enriched_up_dotplot.jpeg",
  width = 700, height = 800)
  print(p)
  dev.off()
}

dot_plot_kegg_enriched_up()
```

```
## pdf
##   2
```

Visualize the top enriched pathway that have smallest qval

```
visualize_top_path_up <- function(){
  pathview(gene.data = up_entrez_ids,
           pathway.id = "hsa04110",
           species = "hsa",
           kegg.dir = "../Enrichment analysis for breast cancer/outputs/")
}

visualize_top_path_up()
```

Browse the top enriched pathway for up-regulated sig genes

```
browseKEGG(kegg_enrich_up, 'hsa04110')
```

KEGG pathway enrichment analysis among down-regulated significant genes

```
kegg_enrichment_down <- function(){
kegg_down <- enrichKEGG(gene = down_entrez_ids,
                        universe = signficant$ENTREZID,
                        organism = "hsa",
                        pvalueCutoff = 0.05,
                        qvalueCutoff = 0.01,
                        pAdjustMethod = "BH")
return(kegg_down)
}

kegg_enrich_down <- kegg_enrichment_down()
head(kegg_enrich_down)
```

```
##                                       category
## hsa04080 Environmental Information Processing
## hsa04020 Environmental Information Processing
## hsa04740                   Organismal Systems
## hsa04024 Environmental Information Processing
## hsa04014 Environmental Information Processing
## hsa00590                           Metabolism
##                                  subcategory       ID
## hsa04080 Signaling molecules and interaction hsa04080
## hsa04020                 Signal transduction hsa04020
## hsa04740                      Sensory system hsa04740
## hsa04024                 Signal transduction hsa04024
## hsa04014                 Signal transduction hsa04014
## hsa00590                    Lipid metabolism hsa00590
##                                      Description GeneRatio BgRatio       pvalue
## hsa04080 Neuroactive ligand-receptor interaction   72/1008 84/1827 1.104773e-09
## hsa04020               Calcium signaling pathway   64/1008 76/1827 4.439362e-08
## hsa04740                  Olfactory transduction   26/1008 26/1827 1.664055e-07
## hsa04024                  cAMP signaling pathway   52/1008 65/1827 1.885335e-05
## hsa04014                   Ras signaling pathway   50/1008 63/1827 4.140199e-05
## hsa00590             Arachidonic acid metabolism   19/1008 20/1827 1.101413e-04
##              p.adjust       qvalue
## hsa04080 2.883456e-07 2.383983e-07
## hsa04020 5.793368e-06 4.789838e-06
## hsa04740 1.447728e-05 1.196952e-05
## hsa04024 1.230181e-03 1.017089e-03
## hsa04014 2.161184e-03 1.786823e-03
## hsa00590 4.404481e-03 3.641535e-03
##
## hsa04080 56413/3953/7068/2899/2901/117/5179/1511/10800/6863/185/5745/1910/6865/5733/2893/2925/154/13
## hsa04020                                             56413/2255/4915/8913/2252/10800/845/185/1956/19
## hsa04740
## hsa04024
## hsa04014
## hsa00590
##          Count
## hsa04080    72
## hsa04020    64
## hsa04740    26
## hsa04024    52
## hsa04014    50
## hsa00590    19
```

DataFrame of KEGG enriched among down-regulated significant genes

```
dataframe_kegg_down <- function(){
  df_kegg_down_genes <- as.data.frame(kegg_enrich_down)
  return(df_kegg_down_genes)
}

df_kegg_down <- dataframe_kegg_down()
```

```
df_kegg_down_top7 <- head(df_kegg_down, 7)
```

Barplot for KEGG enriched down-regulated significant gene

```
bar_plot_kegg_enriched_down<- function(){

  jama_colors <- pal_jama("default")(7)
  p <- ggplot(df_kegg_down_top7, aes(x = reorder(Description, - Count),
                                     y = Count, fill = p.adjust)) +
       geom_bar(stat = "identity", width = 0.8) +
       coord_flip() +
       scale_fill_gradientn(name = "p.adjust", colors = jama_colors) +
       labs(title = "Bar plot of Enriched KEGG pathway of down-regulated Genes",
           x = "Enriched pathway",
           y = "Gene Count") +
       theme_bw() +
       theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))

  jpeg("../Enrichment analysis for breast cancer/outputs/kegg_enriched_down_barplot.jpeg",
  width = 700, height = 800)
  print(p)
  dev.off()

}

bar_plot_kegg_enriched_down()
```

```
## pdf
##   2
```

Dotplot for KEGG enriched down-regulated significant gene

```
dot_plot_kegg_enriched_down <- function(){
  jama_colors <- pal_jama("default")(7)
  p <- ggplot(df_kegg_down_top7, aes(x = reorder(Description, -Count),
                                     y = Count, size = Count,
                                     color = p.adjust)) +
       geom_point(alpha = 0.6) +
       coord_flip() +
       scale_size_continuous(range = c(3, 8), name = "Gene Count") +
       scale_color_gradientn(name = "p.adjust", colors = jama_colors) +
       labs(title = "Dot plot of BP of enriched down-regulated genes",
           x = "GO Term",
           y = "Gene Count") +
       theme_bw() +
       theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))


  jpeg("../Enrichment analysis for breast cancer/outputs/kegg_enriched_down_dotplot.jpeg",
  width = 700, height = 800)
  print(p)
  dev.off()
}

dot_plot_kegg_enriched_down()
```

```
## pdf
##   2
```

Visualize the top enriched pathway that have smallest qval

```
visualize_top_path_down <- function(){
  pathview(gene.data = down_entrez_ids,
           pathway.id = "hsa04080",
           species = "hsa",
           kegg.dir = "../Enrichment analysis for breast cancer/outputs/")
}

visualize_top_path_down()
```

Browse the top enriched pathway for down-regulated sig genes

```
browseKEGG(kegg_enrich_down, 'hsa04080')
```

Reactome pathway enrichment analysis among up-regulated significant genes

```
reactome_up_genes <- function(){
  reactome_enrichment <- enrichPathway(gene          = up_entrez_ids,
                                       universe      = signficant$ENTREZID,
                                       organism      = "human",
                                       pvalueCutoff  = 0.05,
                                       qvalueCutoff  = 0.01,
                                       pAdjustMethod = "BH")

  return(reactome_enrichment)
}

reactome_enriched_path_up <- reactome_up_genes()
head(reactome_enriched_path_up)
```

```
##                          ID              Description GeneRatio  BgRatio
## R-HSA-1640170 R-HSA-1640170               Cell Cycle  144/1102 176/2432
## R-HSA-69278     R-HSA-69278      Cell Cycle, Mitotic  129/1102 154/2432
## R-HSA-68886     R-HSA-68886                  M Phase   81/1102  93/2432
## R-HSA-69620     R-HSA-69620   Cell Cycle Checkpoints   69/1102  78/2432
## R-HSA-5663205 R-HSA-5663205       Infectious disease  152/1102 217/2432
## R-HSA-9824446 R-HSA-9824446 Viral Infection Pathways  123/1102 167/2432
##                     pvalue     p.adjust       qvalue
## R-HSA-1640170 5.315486e-25 3.407226e-22 1.885599e-22
## R-HSA-69278   2.699667e-24 8.652434e-22 4.788358e-22
## R-HSA-68886   1.326233e-17 2.833718e-15 1.568213e-15
## R-HSA-69620   6.490027e-16 1.040027e-13 5.755629e-14
## R-HSA-5663205 1.134875e-14 1.451721e-12 8.033991e-13
## R-HSA-9824446 1.358865e-14 1.451721e-12 8.033991e-13
##
## R-HSA-1640170                                        4605/1869/7083/891/4001/5347/9212/4171/55143/41
## R-HSA-69278
## R-HSA-68886
## R-HSA-69620
## R-HSA-5663205 142/3654/9636/3159/23225/6772/3838/7428/6184/2214/4939/5230/1104/1174/10095/203068/593
## R-HSA-9824446
##               Count
## R-HSA-1640170   144
## R-HSA-69278     129
## R-HSA-68886      81
## R-HSA-69620      69
## R-HSA-5663205   152
## R-HSA-9824446   123
```

DataFrame of Reactome enriched pathways of up-regulated significant genes

```
dataframe_reactome_up <- function(){
  df_reactome_up_genes <- as.data.frame(reactome_enriched_path_up)
  return(df_reactome_up_genes)
}

df_reactome_up <- dataframe_reactome_up()
```

Visualize Reactome Pathway Enrichment Results

```
df_reactome_up_top7 <- head(df_reactome_up, 7)
```

Barplot for Reactome enriched up-regulated significant gene

```
bar_plot_reactome_enriched_up<- function(){
  jama_colors <- pal_jama("default")(7)
  p <- ggplot(df_reactome_up_top7, aes(x = reorder(Description, - Count),
                                       y = Count, fill = p.adjust)) +
       geom_bar(stat = "identity", width = 0.8) +
       coord_flip() +
       scale_fill_gradientn(name = "p.adjust", colors = jama_colors) +
       labs(title = "Bar plot of Enriched Reactome pathway of up-regulated Genes",
            x = "Enriched pathway",
            y = "Gene Count") +
       theme_bw() +
       theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))

  jpeg("../Enrichment analysis for breast cancer/outputs/reactome_enriched_up_barplot.jpeg",
  width = 700, height = 800)
  print(p)
  dev.off()

}

bar_plot_reactome_enriched_up()
```

```
## pdf
##   2
```

Dotplot for Reactome enriched up-regulated significant gene

```
dot_plot_reactome_enriched_up<- function(){
 jama_colors <- pal_jama("default")(7)
 p <- ggplot(df_reactome_up_top7, aes(x = reorder(Description, -Count),
                                      y = Count, size = Count,
                                      color = p.adjust)) +
      geom_point(alpha = 0.6) +
      coord_flip() +
      scale_size_continuous(range = c(3, 8), name = "Gene Count") +
      scale_color_gradientn(name = "p.adjust", colors = jama_colors) +
      labs(title = "Dot plot of BP of enriched up-regulated genes",
           x = "GO Term",
           y = "Gene Count") +
      theme_bw() +
      theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))


jpeg("../Enrichment analysis for breast cancer/outputs/reactome_enriched_up_dotplot.jpeg",
width = 700, height = 800)
print(p)
dev.off()
}
```

```
dot_plot_reactome_enriched_up()
```

```
## pdf
##   2
```

visualize the top enriched Reactome pathway for up-regulated sig genes

Take the generated URL and browse it

```
visualize_reactome_path_up <- function(){
  # Convert the Reactome ID to a URL for visualization
  reactome_url <- paste0("https://reactome.org/PathwayBrowser/#/", "R-HSA-1640170")
  # Print the URL for manual review
  print(reactome_url)

}

visualize_reactome_path_up()
```

```
## [1] "https://reactome.org/PathwayBrowser/#/R-HSA-1640170"
```

Reactome pathway enrichment analysis for down-regulated significant genes

```
reactome_down_genes <- function(){
  reactome_enrichment <- enrichPathway(gene = down_entrez_ids,
                                       universe = signficant$ENTREZID,
                                       organism = "human",
                                       pvalueCutoff = 0.05,
                                       qvalueCutoff = 0.01,
                                       pAdjustMethod = "BH")

  return(reactome_enrichment)
}

reactome_enriched_path_down <- reactome_down_genes()
head(reactome_enriched_path_down)
```

```
##                             ID                              Description GeneRatio
## R-HSA-397014   R-HSA-397014                        Muscle contraction   54/1330
## R-HSA-500792   R-HSA-500792                       GPCR ligand binding   76/1330
## R-HSA-372790   R-HSA-372790                         Signaling by GPCR  116/1330
## R-HSA-211945   R-HSA-211945 Phase I - Functionalization of compounds   25/1330
## R-HSA-9709957 R-HSA-9709957                        Sensory Perception   52/1330
## R-HSA-5576891 R-HSA-5576891                        Cardiac conduction   38/1330
##                 BgRatio       pvalue     p.adjust       qvalue
## R-HSA-397014    59/2432 5.060975e-10 3.239024e-07 2.930038e-07
## R-HSA-500792    95/2432 1.362333e-07 2.914955e-05 2.636883e-05
## R-HSA-372790  156/2432 1.366385e-07 2.914955e-05 2.636883e-05
## R-HSA-211945   25/2432 2.525729e-07 3.907865e-05 3.535075e-05
## R-HSA-9709957  61/2432 3.053019e-07 3.907865e-05 3.535075e-05
## R-HSA-5576891  42/2432 4.873757e-07 5.198674e-05 4.702748e-05
##
## R-HSA-397014
## R-HSA-500792
## R-HSA-372790  56413/5138/111/5296/115557/117/5179/2840/10800/6863/185/10850/5745/1956/1910/7225/6865/
## R-HSA-211945
## R-HSA-9709957
## R-HSA-5576891
##              Count
## R-HSA-397014     54
## R-HSA-500792     76
## R-HSA-372790    116
## R-HSA-211945     25
## R-HSA-9709957    52
## R-HSA-5576891    38
```

DataFrame of Reactome enriched pathways of down-regulated significant genes

```
dataframe_reactome_down <- function(){
  df_reactome_down_genes <- as.data.frame(reactome_enriched_path_down)
  return(df_reactome_down_genes)
}

df_reactome_down <- dataframe_reactome_down()
```

Visualize Reactome Pathway Enrichment Results

Select the top 7 enriched reactome pathways

```
df_reactome_down_top7 <- head(df_reactome_down, 7)
```

Barplot for Reactome enriched down-regulated significant gene

```
bar_plot_reactome_enriched_down<- function(){
  jama_colors <- pal_jama("default")(7)
  p <- ggplot(df_reactome_down_top7, aes(x = reorder(Description, - Count) ,
                                          y = Count, fill = p.adjust)) +
         geom_bar(stat = "identity", width = 0.8) +
         coord_flip() +
         scale_fill_gradientn(name = "p.adjust", colors = jama_colors) +
         labs(title = "Bar plot of Enriched Reactome pathway of down-regulated Genes",
              x = "Enriched pathway",
              y = "Gene Count") +
         theme_bw() +
         theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))

  jpeg("../Enrichment analysis for breast cancer/outputs/reactome_enriched_down_barplot.jpeg",
  width = 700, height = 800)
  print(p)
  dev.off()

}

bar_plot_reactome_enriched_down()
```

```
## pdf
##   2
```

Dotplot for Reactome enriched downregulated significant gene

```
dot_plot_reactome_enriched_down<- function(){
  jama_colors <- pal_jama("default")(7)
  p <- ggplot(df_reactome_down_top7, aes(x = reorder(Description, -Count),
                                         y = Count, size = Count,
                                         color = p.adjust)) +
       geom_point(alpha = 0.6) +
       coord_flip() +
       scale_size_continuous(range = c(3, 8), name = "Gene Count") +
       scale_color_gradientn(name = "p.adjust", colors = jama_colors) +
       labs(title = "Dot plot of BP of enriched down-regulated genes",
             x = "GO Term",
             y = "Gene Count") +
       theme_bw() +
       theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))


jpeg("../Enrichment analysis for breast cancer/outputs/reactome_enriched_down_dotplot.jpeg",
width = 700, height = 800)
print(p)
dev.off()
}

dot_plot_reactome_enriched_down()
```

```
## pdf
##   2
```

visualize the top enriched Reactome pathway for down-regulated sig genes

Take the generated URL and browse it

```
visualize_reactome_path_down <- function(){
  # Convert the Reactome ID to a URL for visualization
  reactome_url <- paste0("https://reactome.org/PathwayBrowser/#/", "R-HSA-397014")
  # Print the URL for manual review
  print(reactome_url)
}

visualize_reactome_path_down()
```

```
## [1] "https://reactome.org/PathwayBrowser/#/R-HSA-397014"
```