RNA-seq Differential Expression

Vaibhav

2025-12-20

Introduction

This analysis investigates hypoxia-induced transcriptional changes in prostate cancer cell lines LNCaP (androgen-sensitive) and PC3 (androgen-independent). Hypoxia is a major driver of tumor aggressiveness and lineage plasticity.

The goal is to identify:

[1]Differentially expressed genes (DEGs) under hypoxia

[2]Global expression patterns using PCA

[3]Hypoxia-regulated biological pathways

[4]Hallmark biological programs altered in LNCaP cells

Loading required libraries

These packages implement:

Differential expression statistics (DESeq2)

Data manipulation (tidyverse)

Visualization (ggplot2, pheatmap)

Pathway analysis (clusterProfiler, ReactomePA, fgsea)

Packages must be installed before knitting

```
library(DESeq2)
library(tidyverse)
library(pheatmap)
library(RColorBrewer)
library(matrixStats)
library(clusterProfiler)
library(ReactomePA)
library(org.Hs.eg.db)
library(fgsea)
```

Loading raw count matrix

```
# Read count matrix

counts <- read.csv("GSE106305_counts_matrix.csv",
row.names = "Geneid",
```

```
stringsAsFactors = FALSE)


counts <- counts[, sort(colnames(counts))]
# View the first few rows
head(counts)
```

```
##                 LNCAP_Hypoxia_S1 LNCAP_Hypoxia_S2 LNCAP_Normoxia_S1
## ENSG00000000003              604              691               367
## ENSG00000000005                0                0                 0
## ENSG00000000419             1995             2302              2160
## ENSG00000000457              554              607               433
## ENSG00000000460              275              350               379
## ENSG00000000938                2                2                 2
##                 LNCAP_Normoxia_S2 PC3_Hypoxia_S1 PC3_Hypoxia_S2 PC3_Normoxia_S1
## ENSG00000000003               380           1059             332              352
## ENSG00000000005                 0              0               0                0
## ENSG00000000419              2454           1974             693              747
## ENSG00000000457               518             88              26               29
## ENSG00000000460               349            390             155              189
## ENSG00000000938                 1              0               0                0
##                 PC3_Normoxia_S2
## ENSG00000000003             971
## ENSG00000000005               0
## ENSG00000000419            1761
## ENSG00000000457              83
## ENSG00000000460             438
## ENSG00000000938               1
```

Creating sample metadata

```
condition <- factor(c(
rep("LNCAP_Hypoxia", 2),
rep("LNCAP_Normoxia", 2),
rep("PC3_Hypoxia", 2),
rep("PC3_Normoxia", 2)
))


colData <- data.frame(condition)
rownames(colData) <- colnames(counts)
head(colData)
```

```
##                       condition
## LNCAP_Hypoxia_S1   LNCAP_Hypoxia
## LNCAP_Hypoxia_S2   LNCAP_Hypoxia
## LNCAP_Normoxia_S1 LNCAP_Normoxia
## LNCAP_Normoxia_S2 LNCAP_Normoxia
## PC3_Hypoxia_S1       PC3_Hypoxia
## PC3_Hypoxia_S2       PC3_Hypoxia
```

DESeq2 dataset

```
dds <- DESeqDataSetFromMatrix(
countData = counts,
colData = colData,
design = ~ condition
)

# Keep genes with at least 10 reads in at least 2 samples
keep <- rowSums(counts(dds) >= 10) >= 2
dds <- dds[keep, ]
```

Differential expression analysis

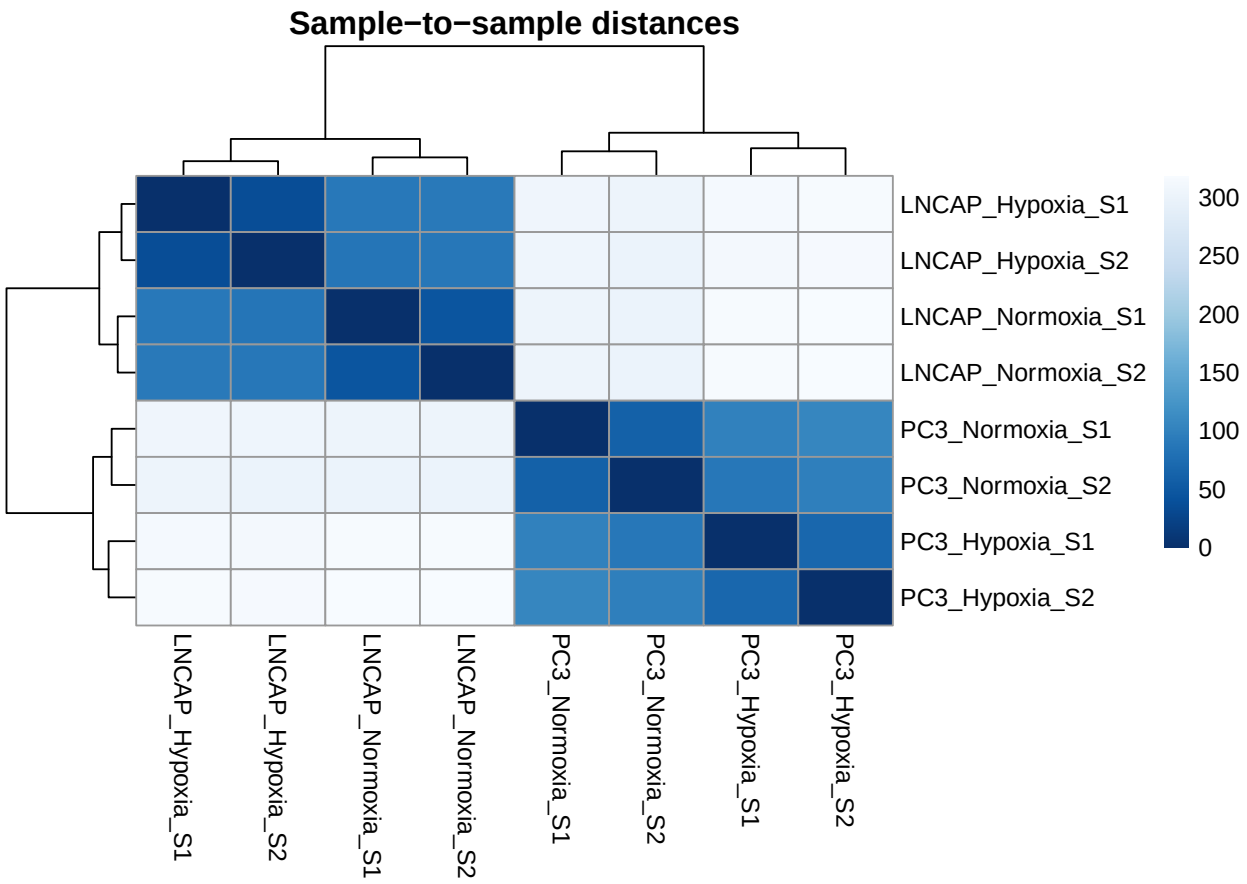
```
dds <- DESeq(dds)

#Normalized counts
norm_counts <- counts(dds, normalized = TRUE)
write.csv(norm_counts, "Normalized_counts.csv")
```

Exploratory analysis: PCA

```
#Variance-stabilizing transformation
vsd <- vst(dds, blind = TRUE)

pca_data <- plotPCA(vsd, intgroup = "condition", returnData = TRUE)
percentVar <- round(100 * attr(pca_data, "percentVar"))


p_pca <- ggplot(pca_data, aes(PC1, PC2, color = condition)) +
geom_point(size = 3) +
labs(x = paste0("PC1: ", percentVar[1], "%"),
y = paste0("PC2: ", percentVar[2], "%"),
title = "PCA (VST transformed)") +
theme_minimal()


ggsave("PCA_plot.png", p_pca, width = 6, height = 5, dpi = 300)
```

Sample-to-sample distance heatmap

```
sample_dists <- dist(t(assay(vsd)))
dist_matrix <- as.matrix(sample_dists)


png("sample_distance_heatmap.png", width = 1200, height = 1000, res = 300)
pheatmap(dist_matrix,
clustering_distance_rows = sample_dists,
clustering_distance_cols = sample_dists,
col = colorRampPalette(rev(brewer.pal(9, "Blues")))(255),
main = "Sample-to-sample distances")
```




```
dev.off()
```

```
## png
##   3
```

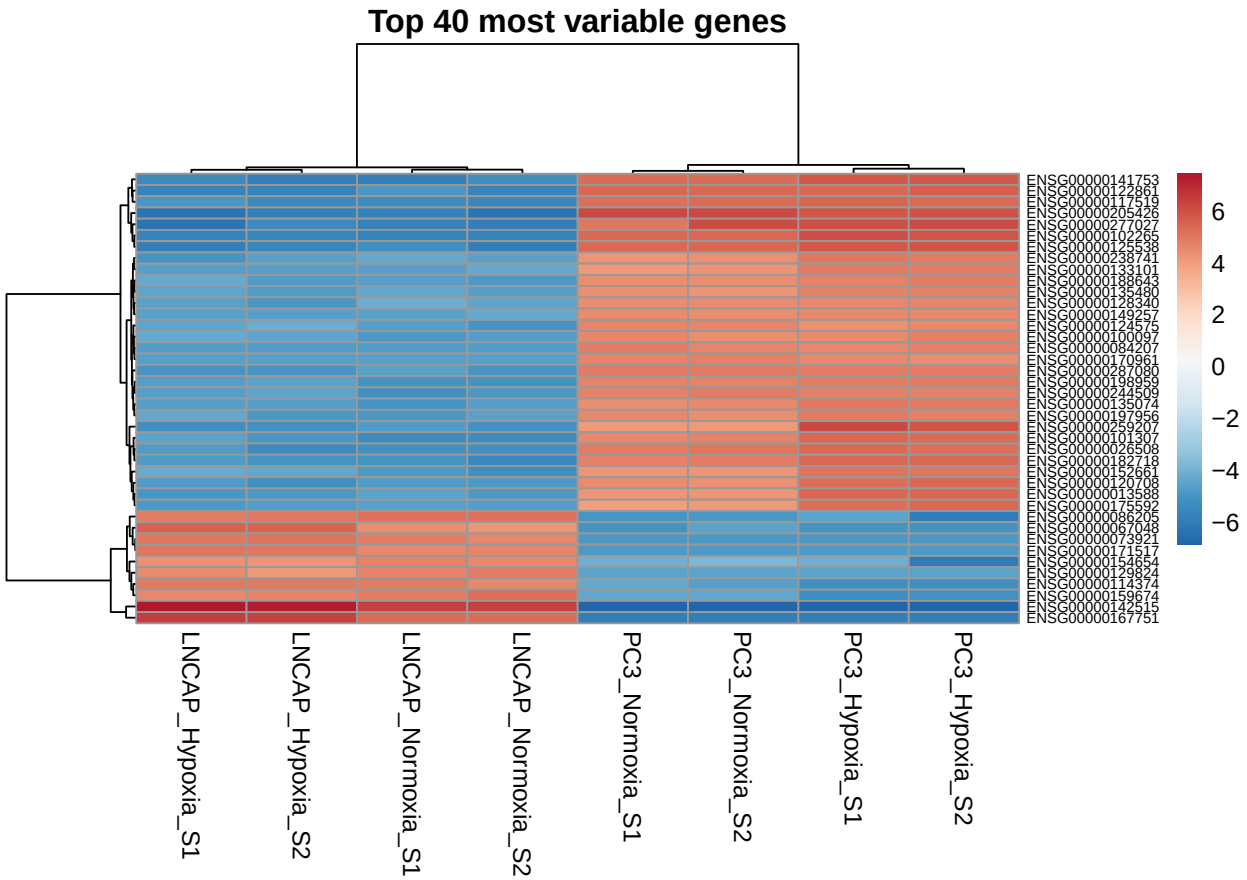
Expression distribution diagnostics

```
png("density_raw_vs_vst.png", width = 2000, height = 2000, res = 300)
par(mfrow = c(2, 2))


plot(density(counts(dds)[,1]), main = "Raw counts", xlab = "Expression")
plot(density(assay(vsd)[,1]), main = "VST counts", xlab = "Expression")


dev.off()
```

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

Highly variable gene heatmap

```
rv <- rowVars(assay(vsd))
top_genes <- order(rv, decreasing = TRUE)[1:40]


mat <- assay(vsd)[top_genes, ]
mat <- mat - rowMeans(mat)


png("top_variable_genes_heatmap.png", width = 1200, height = 1200, res = 300)
pheatmap(mat,
color = colorRampPalette(rev(brewer.pal(9, "RdBu")))(255),
fontsize_row = 6,
main = "Top 40 most variable genes")
```




```
dev.off()
```

```
## png
##   3
```

```
dds_lncap <- dds[, grepl("LNCAP", colnames(dds))]
dds_lncap$condition <- droplevels(dds_lncap$condition)
dds_lncap$condition <- relevel(dds_lncap$condition, ref = "LNCAP_Normoxia")
```

Extract differential expression results

```r
dds_lncap <- DESeq(dds_lncap)


res_lncap <- results(dds_lncap,
contrast = c("condition",
"LNCAP_Hypoxia",
"LNCAP_Normoxia"))
write.csv(as.data.frame(res_lncap), "DEGs_LNCAP.csv")
head(res_lncap)
```

```
## log2 fold change (MLE): condition LNCAP_Hypoxia vs LNCAP_Normoxia
## Wald test p-value: condition LNCAP Hypoxia vs LNCAP Normoxia
## DataFrame with 6 rows and 6 columns
##                   baseMean log2FoldChange     lfcSE      stat      pvalue
##                  <numeric>      <numeric> <numeric> <numeric>   <numeric>
## ENSG00000000003   257.560     0.88827897  0.201949  4.398534 1.08985e-05
## ENSG00000000419  1110.185    -0.00297024  0.128404 -0.023132 9.81545e-01
## ENSG00000000457   264.313     0.39379117  0.192937  2.041031 4.12478e-02
## ENSG00000000460   169.022    -0.13896323  0.248008 -0.560318 5.75262e-01
## ENSG00000001036   146.122    -0.84216624  0.248853 -3.384187 7.13893e-04
## ENSG00000001084   612.416     0.78838940  0.165656  4.759190 1.94372e-06
##                        padj
##                   <numeric>
## ENSG00000000003 1.15282e-04
## ENSG00000000419 9.91507e-01
## ENSG00000000457 1.21213e-01
## ENSG00000000460 7.45255e-01
## ENSG00000001036 4.30967e-03
## ENSG00000001084 2.46578e-05
```

MA plot

```r
png("MA_plot_LNCAP.png", width = 800, height = 600, res = 150)
plotMA(res_lncap, ylim = c(-5, 5))
dev.off()
```

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

Volcano plot

```r
res_df <- as.data.frame(res_lncap) %>%
na.omit() %>%
mutate(regulation = case_when(
padj < 0.05 & log2FoldChange > 1 ~ "Upregulated",
padj < 0.05 & log2FoldChange < -1 ~ "Downregulated",
TRUE ~ "Not significant"
))


p_volcano <- ggplot(res_df, aes(log2FoldChange, -log10(padj), color = regulation)) +
geom_point(alpha = 0.6) +
```

```
theme_minimal() +
labs(title = "Volcano plot: LNCaP hypoxia")


ggsave("Volcano_plot_LNCAP.png", p_volcano, width = 6, height = 5, dpi = 300)
```

Pathway Analysis using GSEA

```
#Converting gene IDs (ENSEMBL → ENTREZ)

res_lncap_df <- as.data.frame(res_lncap)
res_lncap_df$ENSEMBL <- rownames(res_lncap)

id_map <- bitr(
res_lncap_df$ENSEMBL,
fromType = "ENSEMBL",
toType = "ENTREZID",
OrgDb = org.Hs.eg.db
)

head(res_lncap_df)
```

```
##                  baseMean log2FoldChange     lfcSE        stat       pvalue
## ENSG00000000003  257.5595    0.888278972 0.2019489  4.39853403 1.089845e-05
## ENSG00000000419 1110.1847   -0.002970238 0.1284037 -0.02313203 9.815450e-01
## ENSG00000000457  264.3134    0.393791169 0.1929374  2.04103059 4.124779e-02
## ENSG00000000460  169.0223   -0.138963228 0.2480076 -0.56031842 5.752623e-01
## ENSG00000001036  146.1224   -0.842166237 0.2488533 -3.38418723 7.138930e-04
## ENSG00000001084  612.4156    0.788389402 0.1656562  4.75918976 1.943716e-06
##                         padj         ENSEMBL
## ENSG00000000003 1.152817e-04 ENSG00000000003
## ENSG00000000419 9.915074e-01 ENSG00000000419
## ENSG00000000457 1.212133e-01 ENSG00000000457
## ENSG00000000460 7.452553e-01 ENSG00000000460
## ENSG00000001036 4.309675e-03 ENSG00000001036
## ENSG00000001084 2.465782e-05 ENSG00000001084
```

```
#Merge mapping and removing duplicates
res_mapped <- res_lncap_df %>% left_join(id_map, by = "ENSEMBL") %>%
filter(!is.na(ENTREZID)) %>%
distinct(ENTREZID, .keep_all = TRUE)
head(res_mapped)
```

```
##    baseMean log2FoldChange     lfcSE        stat       pvalue         padj
## 1  257.5595    0.888278972 0.2019489  4.39853403 1.089845e-05 1.152817e-04
## 2 1110.1847   -0.002970238 0.1284037 -0.02313203 9.815450e-01 9.915074e-01
## 3  264.3134    0.393791169 0.1929374  2.04103059 4.124779e-02 1.212133e-01
## 4  169.0223   -0.138963228 0.2480076 -0.56031842 5.752623e-01 7.452553e-01
## 5  146.1224   -0.842166237 0.2488533 -3.38418723 7.138930e-04 4.309675e-03
## 6  612.4156    0.788389402 0.1656562  4.75918976 1.943716e-06 2.465782e-05
##           ENSEMBL ENTREZID
## 1 ENSG00000000003     7105
```

```
## 2 ENSG00000000419     8813
## 3 ENSG00000000457    57147
## 4 ENSG00000000460    55732
## 5 ENSG00000001036     2519
## 6 ENSG00000001084     2729
```

Creating ranked gene list

```
gene_ranking <- res_mapped$log2FoldChange
names(gene_ranking) <- res_mapped$ENTREZID
gene_ranking <- sort(gene_ranking, decreasing = TRUE)

head(gene_ranking)
```

```
##      7040    441932      2681     79656 100128264     59350
##  7.680917  7.418113  7.264110  7.082984  6.712309  6.679098
```

Run GSEA using Reactome

```
gsea_reactome <- gsePathway(
geneList = gene_ranking,
organism = "human",
pvalueCutoff = 0.05,
verbose = FALSE
)

head(gsea_reactome@result)
```

```
##                           ID
## R-HSA-156842   R-HSA-156842
## R-HSA-9633012 R-HSA-9633012
## R-HSA-2408557 R-HSA-2408557
## R-HSA-9954716 R-HSA-9954716
## R-HSA-156827   R-HSA-156827
## R-HSA-9954714 R-HSA-9954714
##
## R-HSA-156842                                                                    Eukaryoti
## R-HSA-9633012                                                 Response of EIF2AK4 (GCN2) t
## R-HSA-2408557                                                                          S
## R-HSA-9954716 ZNF598 and the Ribosome-associated Quality Trigger (RQT) complex dissociate a ribosome
## R-HSA-156827                                          L13a-mediated translational silencing of
## R-HSA-9954714                                             PELO:HBS1L and ABCE1 dissociate a ribo
##               setSize enrichmentScore       NES       pvalue    p.adjust
## R-HSA-156842       90      -0.6450714 -2.436321 1.134338e-10 9.68191e-09
## R-HSA-9633012      98      -0.6327030 -2.423390 1.000000e-10 9.68191e-09
## R-HSA-2408557      92      -0.6375252 -2.420458 1.000000e-10 9.68191e-09
## R-HSA-9954716      95      -0.6324646 -2.420341 1.000000e-10 9.68191e-09
## R-HSA-156827      106      -0.6149729 -2.410724 1.000000e-10 9.68191e-09
## R-HSA-9954714      89      -0.6424837 -2.408874 1.000000e-10 9.68191e-09
##                     qvalue rank                  leading_edge
## R-HSA-156842  8.962325e-09 5778 tags=89%, list=32%, signal=61%
## R-HSA-9633012 8.962325e-09 5778 tags=86%, list=32%, signal=59%
```

```
## R-HSA-2408557 8.962325e-09 5778 tags=86%, list=32%, signal=59%
## R-HSA-9954716 8.962325e-09 5778 tags=83%, list=32%, signal=57%
## R-HSA-156827  8.962325e-09 5778 tags=83%, list=32%, signal=57%
## R-HSA-9954714 8.962325e-09 5778 tags=88%, list=32%, signal=60%
##
## R-HSA-156842                                                  51121/6154/6230/6139/6168/6205/6144/9045/61
## R-HSA-9633012                          51121/6154/6230/1054/6139/8894/6168/6205/6144/1649/9045/6167/6
## R-HSA-2408557                                                   51121/6154/6230/6139/6168/6205/6144/9045
## R-HSA-9954716                                                    51121/6154/6230/6139/6168/6205/6144/904
## R-HSA-156827  51121/6154/6230/3646/10480/6139/51386/8894/6168/6205/6144/1975/8664/9045/6167/8667/622
## R-HSA-9954714                                                          51121/6154/6230/6139/6168/6205/61
```

Over-Representation Analysis (ORA)

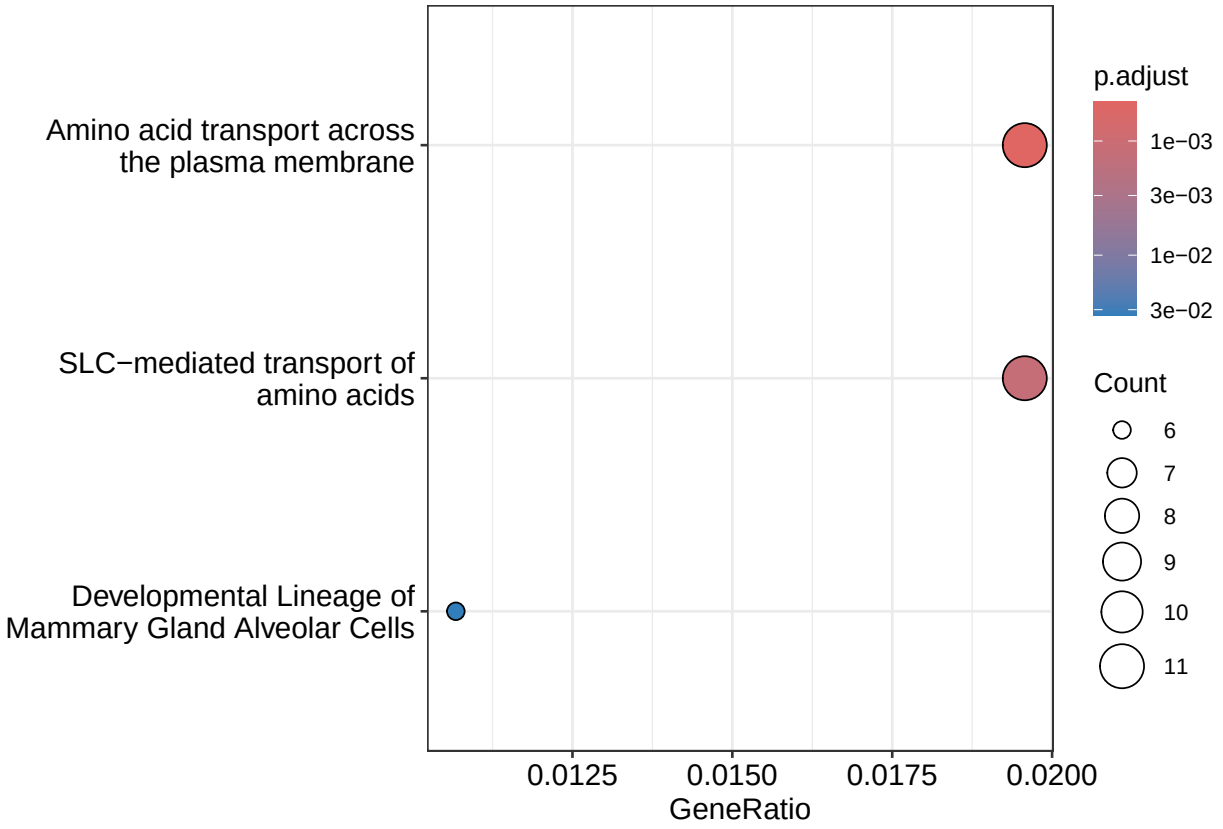
```
#extracted significant genes
sig_genes <- res_mapped %>%
filter(padj < 0.05 & abs(log2FoldChange) > 1) %>%
pull(ENTREZID)
```

Run ORA using Reactome

```
ora_reactome <- enrichPathway(
gene = sig_genes,
organism = "human",
pvalueCutoff = 0.05
)
```

Dotpot

```
d <-dotplot(ora_reactome, showCategory = 20)

ggsave("ora_reactome_dotplot.png", plot = d, width = 10, height = 8, dpi = 300)

d
```



fgsea + Hallmark

Loaded fgsea and Hallmark gene sets

```
library(fgsea)


hallmark_sets <- gmtPathways("h.all.v7.0.symbols.gmt.txt")


# Convert ENSEMBL to SYMBOL
symbol_map <- bitr(
res_lncap$ENSEMBL,
fromType = "ENSEMBL",
toType = "SYMBOL",
OrgDb = org.Hs.eg.db
)


res_symbol <- res_lncap_df %>%
left_join(symbol_map, by = c("ENSEMBL" = "ENSEMBL")) %>%
filter(!is.na(SYMBOL)) %>%
distinct(SYMBOL, .keep_all = TRUE)


ranked_symbols <- res_symbol$log2FoldChange
```

```r
names(ranked_symbols) <- res_symbol$SYMBOL


ranked_symbols <- sort(ranked_symbols, decreasing = TRUE)
```

Run fgsea

```r
fgsea_res <- fgsea(
pathways = hallmark_sets,
stats = ranked_symbols,
minSize = 15,
maxSize = 500,
nperm = 1000
)


fgsea_res[order(padj), .(pathway, NES, padj)][1:6]
```

```
##    pathway   NES  padj
##     <char> <num> <num>
## 1:    <NA>    NA    NA
## 2:    <NA>    NA    NA
## 3:    <NA>    NA    NA
## 4:    <NA>    NA    NA
## 5:    <NA>    NA    NA
## 6:    <NA>    NA    NA
```

Visualized Hallmark enrichment

```r
waterfall_plot <- function(fgsea_res, graph_title) {

  fgsea_df <- as.data.frame(fgsea_res)

  p <- fgsea_df %>%
    mutate(short_name = str_split_fixed(pathway, "_", 2)[,2]) %>%
    ggplot(aes(x = reorder(short_name, NES), y = NES)) +
      geom_bar(stat = "identity", aes(fill = padj < 0.05)) +
      coord_flip() +
      labs(
        x = "Hallmark Pathway",
        y = "Normalized Enrichment Score",
        title = graph_title
      ) +
      theme(
        axis.text.y = element_text(size = 7),
        plot.title = element_text(hjust = 0.5)
      )

  return(p)
}
```

```r
p <- waterfall_plot(
  fgsea_res,
  "Hallmark pathways altered by hypoxia in LNCaP cells"
)

ggsave(
  filename = "Hallmark_fgsea_waterfall_LNCaP.png",
  plot = p,
  width = 8,
  height = 6,
  dpi = 300
)
```