DESeq analysis, Volcano plot and PCA plot for common patients

21267455 Ella Zhang

2024-11-25

```r
library(DESeq2)
```

```
##      S4Vectors
```

```
##      stats4
```

```
##      BiocGenerics
```

```
## 
##    'BiocGenerics'
```

```
## 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, table,
##     tapply, union, unique, unsplit, which.max, which.min
```

```
## 
##    'S4Vectors'
```

```
## The following object is masked from 'package:utils':
## 
##     findMatches
```

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

```
##      IRanges
```

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

##      GenomicRanges

##      GenomeInfoDb

##      SummarizedExperiment

##      MatrixGenerics

##      matrixStats

##
##    '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

##      Biobase

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

##
##    'Biobase'

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

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

```r
library(ggplot2)
```

```
## Warning:   'ggplot2' R 4.4.2
```

```r
library(pheatmap)
```

```
## Warning:   'pheatmap' R 4.4.2
```

```r
library(clusterProfiler)
```

```
##
```

```
## clusterProfiler v4.12.6 Learn more at https://yulab-smu.top/contribution-knowledge-mining/
##
## Please cite:
##
## G Yu. Thirteen years of clusterProfiler. The Innovation. 2024,
## 5(6):100722
```

```
##
##    'clusterProfiler'
```

```
## The following object is masked from 'package:IRanges':
##
##     slice
```

```
## The following object is masked from 'package:S4Vectors':
##
##     rename
```

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

```r
library(org.Hs.eg.db)
```

```
##      AnnotationDbi
```

```
##
##    'AnnotationDbi'
```

```
## The following object is masked from 'package:clusterProfiler':
##
##     select
```

```
##
```

```r
library(data.table)
```

```
##
##    'data.table'
```

```
## The following object is masked from 'package:SummarizedExperiment':
##
##     shift
```

```
## The following object is masked from 'package:GenomicRanges':
##
##     shift
```

```
## The following object is masked from 'package:IRanges':
##
##     shift
```

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

```r
library(readr)
library(dplyr)
```

```
##
##    'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following object is masked from 'package:AnnotationDbi':
##
##     select
```

```
## The following object is masked from 'package:Biobase':
##
##     combine
```

```
## The following object is masked from 'package:matrixStats':
##
##     count
```

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

```
## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect
## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

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

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

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

```
clinical.patients <- read.table("data_clinical_patient.txt", sep = "\t", header = TRUE)
data.mutations <- read.table("data_mutations.txt", sep = "\t", header = TRUE)
data.RNAseq <- read.csv("RNAseq_BRCA.csv")
```

```
library(stringr)

# Rename columns to match the desired format
colnames(data.RNAseq) <- sapply(colnames(data.RNAseq), function(name) {

  segments <- strsplit(name, "\\.")[[1]][1:3]

  paste(segments, collapse = "-")
})

colnames(data.RNAseq)[1] <- "Transcript_ID"
```

```
unique.clinical <- as.data.frame(unique(clinical.patients$PATIENT_ID))
unique.mutations <- as.data.frame(unique(data.mutations$Tumor_Sample_Barcode))
unique.RNA <- as.data.frame(colnames(data.RNAseq[,2:1232]))

colnames(unique.clinical) <- "Patient_ID"
colnames(unique.mutations) <- "Patient_ID"
colnames(unique.RNA) <- "Patient_ID"

unique.mutations$Patient_ID <- substr(unique.mutations$Patient_ID, 1, 12)

# Find common patient IDs across all three data frames
common_patient_ids <- Reduce(intersect, list(unique.clinical$Patient_ID, unique.mutations$Patient_ID, u

filtered.clinical <- clinical.patients[clinical.patients$PATIENT_ID %in% common_patient_ids, ]
filtered.mutations <- data.mutations[substr(data.mutations$Tumor_Sample_Barcode, 1, 12) %in% common_pat
filtered.RNAseq <- data.RNAseq[, c("Transcript_ID", common_patient_ids)]  #Keep only the columns of com
```

```
RNAseq_numeric <- as.matrix(filtered.RNAseq[, -1])
rownames(RNAseq_numeric) <- filtered.RNAseq$Transcript_ID

filtered.clinical$SurvivalStatus <- ifelse(filtered.clinical$OS_MONTHS > 36, "HighSurvival", "LowSurviva
filtered.clinical$SurvivalStatus <- as.factor(filtered.clinical$SurvivalStatus)

dds <- DESeqDataSetFromMatrix(countData = RNAseq_numeric,
                              colData = filtered.clinical,
                              design = ~ SurvivalStatus)
```

```
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 12189 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)
```

```
## estimating dispersions
```

```
## fitting model and testing
```

```
res <- results(dds, contrast = c("SurvivalStatus", "HighSurvival", "LowSurvival"))

res <- lfcShrink(dds, coef = 2, type = "apeglm")
```

```
## using 'apeglm' for LFC shrinkage. If used in published research, please cite:
##     Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior distributions for
##     sequence count data: removing the noise and preserving large differences.
##     Bioinformatics. https://doi.org/10.1093/bioinformatics/bty895
```

```
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, possibly due to insufficient
## numeric precision
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, possibly due to insufficient
## numeric precision
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, possibly due to insufficient
## numeric precision
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, possibly due to insufficient
## numeric precision
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, possibly due to insufficient
## numeric precision

## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, unable to sufficiently decrease the
## function value
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, unable to sufficiently decrease the
## function value

## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, possibly due to insufficient
## numeric precision
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, possibly due to insufficient
## numeric precision
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, possibly due to insufficient
## numeric precision
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, possibly due to insufficient
## numeric precision
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, possibly due to insufficient
## numeric precision
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, possibly due to insufficient
## numeric precision

## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, unable to sufficiently decrease the
## function value
## Warning in nbinomGLM(x = x, Y = YNZ, size = size, weights = weightsNZ, offset =
## offsetNZ, : the line search routine failed, unable to sufficiently decrease the
## function value
```

```r
summary(res)
```

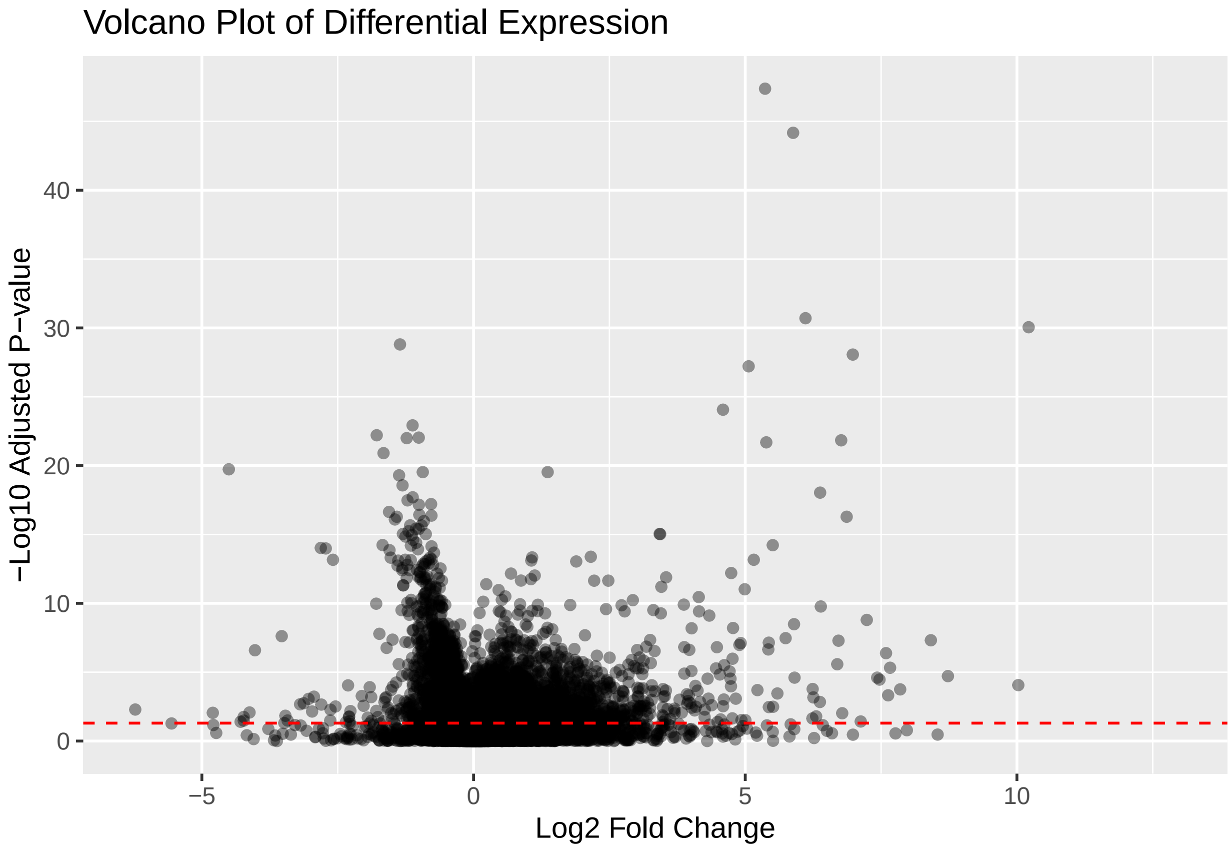
```
##
## out of 57944 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 6686, 12%
## LFC < 0 (down)     : 3644, 6.3%
## outliers [1]       : 0, 0%
## low counts [2]     : 17930, 31%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
sig_res <- res[which(res$padj < 0.05), ]

write.csv(as.data.frame(sig_res), "DEGs_results.csv")
```

There were 57,944 genes included in the analysis, genes with zero counts in all samples were excluded. the p value is less than 0.1, which means this data is considered statistical significant.

```
res_df <- as.data.frame(res)

ggplot(res_df, aes(x = log2FoldChange, y = -log10(padj))) +
  geom_point(alpha = 0.4) +
  geom_hline(yintercept = -log10(0.05), linetype = "dashed", color = "red") +
  xlab("Log2 Fold Change") +
  ylab("-Log10 Adjusted P-value") +
  ggtitle("Volcano Plot of Differential Expression")
```

```
## Warning: Removed 20646 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

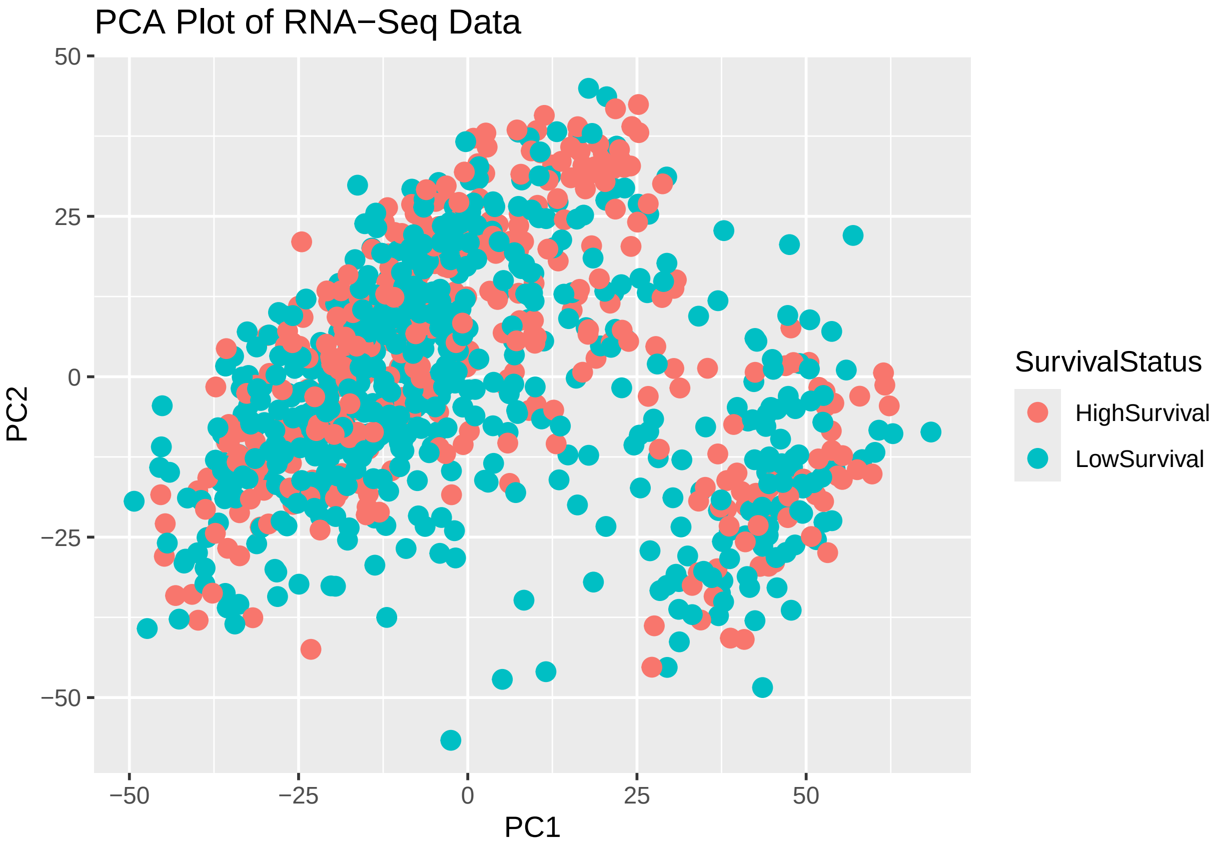


This plot identifies a small subset of genes with significant differential expression, their biological roles can be investigated in survival outcomes

```r
# Perform PCA
vsd <- vst(dds, blind = FALSE)
pca_data <- plotPCA(vsd, intgroup = "SurvivalStatus", returnData = TRUE)
```

```
## using ntop=500 top features by variance
```

```r
# Visualize PCA
ggplot(pca_data, aes(PC1, PC2, color = SurvivalStatus)) +
  geom_point(size = 3) +
  ggtitle("PCA Plot of RNA-Seq Data") +
  xlab("PC1") +
  ylab("PC2")
```



This plot shows the partial separation between survival groups, suggesting that survival status can be one of the factors that can influence the gene expression.