Intro to Transcriptomics-Assignment3Brukt

Nicole Black, Wade Boohar, Kayla Xu

07/17/22

Deliverables -Upload this R Notebook to your GitHub and submit the link to your Repo on Brightspace. -Include any graphs or figures created in this assignment in the folder with your R notebook with descriptive file names.

Since this is an optional partner activity, it is okay if your answers are the same as your partner's as long as everyone understands it and could explain it in their own words if asked. Each person must individually push their code to Github. *At the top of your R Notebook, write the name of you and your partner(s) as a comment.*

Complete the following coding activity and answer any following questions as comments in your R Notebook

In SummarizedExperiment Tutorial, you learned how to manipulate the SummarizedExperiment data structure and turn it into more readable dataframes, saving them as rna_counts, rna_clinical, and rna_genes. In this semi-guided assignment, you will use these dataframes to perform differential expression analysis based on tumor status.

Pre-Assignment

```r
knitr::opts_knit$set(root.dir = normalizePath("/home1/tefera/analysis_data"))

# Silence all messages/warnings across the whole notebook
knitr::opts_chunk$set(message = FALSE, warning = FALSE)

# Install if missing
if (!require("DESeq2", quietly = TRUE))
  BiocManager::install("DESeq2")
if (!require("EnhancedVolcano", quietly = TRUE))
  BiocManager::install("EnhancedVolcano")
if (!require("dplyr", quietly = TRUE))
  install.packages("dplyr")

# Load packages quietly
suppressPackageStartupMessages({
  library(DESeq2)
  library(EnhancedVolcano)
  library(dplyr)
})
```

Load in all necessary packages

```r
if (!require("DESeq2", quietly = TRUE))
  BiocManager::install("DESeq2")
if (!require("EnhancedVolcano", quietly = TRUE))
  BiocManager::install("EnhancedVolcano")
```

```r
library(DESeq2)
library(EnhancedVolcano)
library(dplyr)
```

1 Read in the rna_clinical, rna_genes, and rna_counts dataframes which you made in the “SummarizedExperiment Guided Tutorial” R Notebook

```r
rna_clinical <- read.csv("/home1/tefera/analysis_data/BRCA_rna_clinical.csv")
rna_genes <- read.csv("/home1/tefera/analysis_data/BRCA_rna_genes.csv")
rna_counts <- read.csv("/home1/tefera/analysis_data/BRCA_rna_counts.csv", row.names = 1)

# quick check
dim(rna_clinical)
```

```
## [1] 1231   86
```

```r
dim(rna_genes)
```

```
## [1] 60660    11
```

```r
dim(rna_counts)
```

```
## [1] 60660  1231
```

2 In this assignment, you will run differential expression analysis comparing patient samples by whether the sample is from a tumor or normal tissue (this is the definition column in rna_clinical). You will need to choose a variable to control for covariance of: age and/or PAM50 subtype (paper_BRCA_Subtype_PAM50).

Manipulate those columns so that they are ready for differential expression analysis (hint: what kind of variables are they? what data type are they by default? do you need to handle unknown values?) Filter out genes with a total expression across all patients less than 1000.

```r
rna_clinical$definition <- as.factor(rna_clinical$definition)
rna_clinical$paper_BRCA_Subtype_PAM50 <- as.factor(rna_clinical$paper_BRCA_Subtype_PAM50)

rna_clinical <- rna_clinical %>%
  filter(!is.na(definition), !is.na(paper_BRCA_Subtype_PAM50))

keep_genes <- rowSums(rna_counts) >= 1000
rna_counts_filtered <- rna_counts[keep_genes, ]
```

3 Perform the differential expression analysis, All you need to do is fill in the appropriate # terms

```r
# Fix sample names
colnames(rna_counts_filtered) <- gsub("\\.", "-", colnames(rna_counts_filtered))
colnames(rna_counts_filtered) <- substr(colnames(rna_counts_filtered), 1, 12)

# Get patient IDs and clean clinical data
rna_clinical$patient_id <- substr(rna_clinical$patient, 1, 12)
rna_clinical <- rna_clinical[!duplicated(rna_clinical$patient_id), ]
rna_clinical <- rna_clinical[!is.na(rna_clinical$paper_BRCA_Subtype_PAM50), ]
rownames(rna_clinical) <- rna_clinical$patient_id

# Match samples between both datasets
common_samples <- intersect(colnames(rna_counts_filtered), rownames(rna_clinical))
rna_counts_filtered <- rna_counts_filtered[, common_samples]
rna_clinical <- rna_clinical[common_samples, ]
```

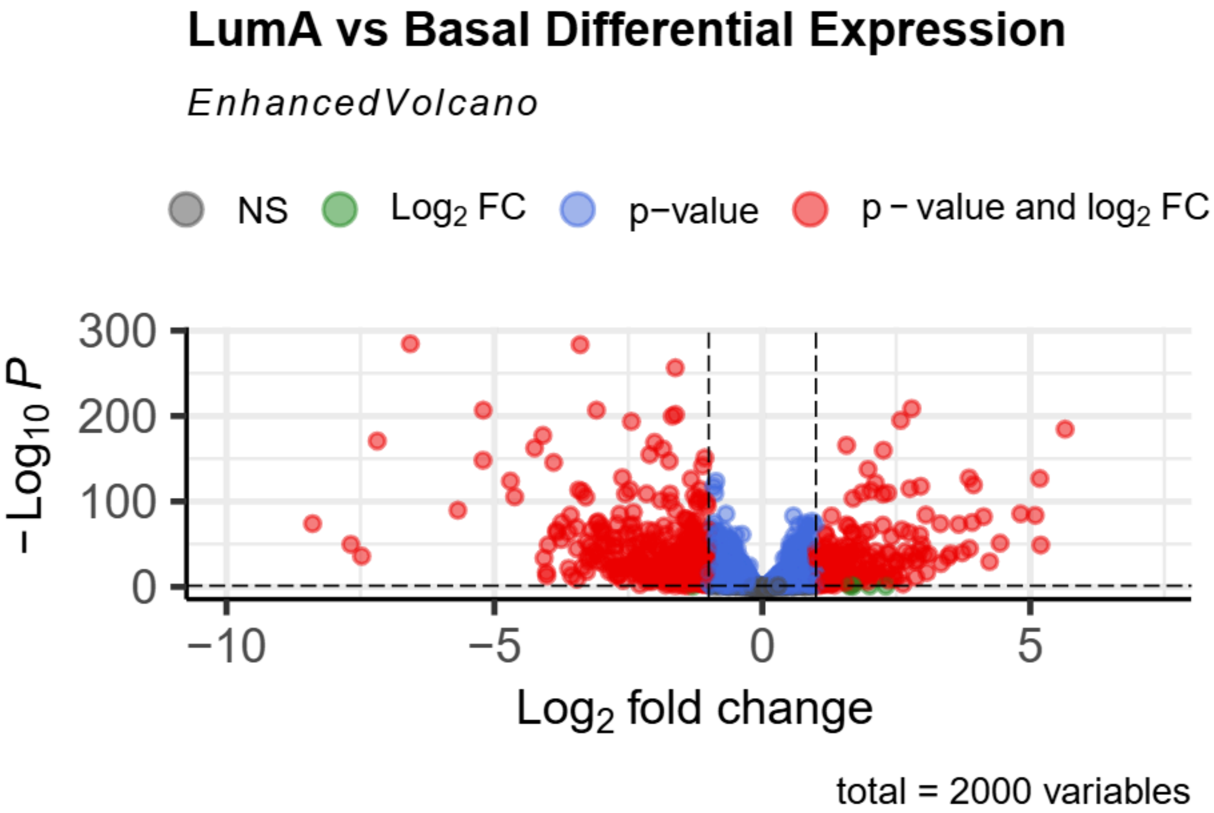
```r
# Make subtype a factor
rna_clinical$paper_BRCA_Subtype_PAM50 <- as.factor(rna_clinical$paper_BRCA_Subtype_PAM50)

# Pick 2,000 random genes so R does not crash
set.seed(123)
subset_genes <- sample(rownames(rna_counts_filtered), 2000)
rna_counts_subset <- rna_counts_filtered[subset_genes, ]

# Run DESeq2 to compare LumA vs Basal
dds <- DESeqDataSetFromMatrix(
  countData = rna_counts_subset,
  colData = rna_clinical,
  design = ~ paper_BRCA_Subtype_PAM50
)

dds <- DESeq(dds)
res <- results(dds, contrast = c("paper_BRCA_Subtype_PAM50", "LumA", "Basal"))
res_df <- as.data.frame(res)
```

Prepare results dataframe for EnhancedVolcano plotting. Add two columns, “-log10(padj)” and “gene_name”. Fill in these columns appropriately.

```r
res_df$gene_name <- rna_genes$gene_name[match(rownames(res_df), rna_genes$X)]
res_df$negLog10Padj <- -log10(res_df$padj)
```

4 Now we will use the EnhancedVolcano package to plot our results. The code is already completed and should run without adjustment if all code up to here is correct.

```r
EnhancedVolcano(res_df,
                lab = res_df$gene_name,
                x = 'log2FoldChange',
                y = 'padj',
                title = 'LumA vs Basal Differential Expression',
                pCutoff = 0.05,
                FCcutoff = 1.0,
                pointSize = 2.0,
                labSize = 3.0)
```



```r
ggsave("LumA_vs_Basal_volcano.png", width = 8, height = 6, dpi = 300)
```

5 # Explain what genes from each part of the Volcano Plot mean in terms of their significance and up/down regulation. top-right genes: These genes are much higher in LumA samples and are statistically significant. bottom-right genes: These genes are a little higher in LumA, but not significant. top-left genes: These genes are much higher in Basal samples and are statistically significant. bottom-left genes: These genes are a little higher in Basal, but not significant. top-middle genes: These genes have small changes, but are very significant. bottom-middle genes: These genes do not change much between LumA and Basal and are not significant.

Save the picture of the volcano plot (using either ggsave() or right clicking and manually downloading the image and push this .Rmd and the image to GitHub)