Esther Olabisi-Adeniyi Student ID: 1238826

Project 4: Analysis of RNAseq count data

Introduction

RNA-sequencing (RNAseq) is a powerful technique used to investigate gene expression by identifying and quantifying RNA molecules in samples. RNAseq begins with generating short reads of DNA, followed by alignment, quantification, and differential expression analysis (Yang & Kim, 2015). In this tutorial, I will explore RNAseq count data derived from luminal and basal cell subpopulations of mice at different physiological states (virgin, 18.5 day pregnant and 2 day lactating mice). The main techniques to be used are principal components analysis (PCA) to reduce dimensionality and visually highlight patterns, as well as a volcano plot to visualize differences in gene expression. We will carry out these techniques on RStudio. With PCA, I would like to observe how clusters form by cell subpopulation and physiological state. The volcano plot on the other hand will explore significant changes in gene expression levels and the function of those genes. The dataset for this tutorial is available at *NCBI Gene Expression Omnibu*s with accession GSE60450.

Tutorial

Path to scratch directory with count data: /scratch/eolabisi/MyProject4

1. **Ensure we have packages tidyverse, Bioconductor package DESeq2**

```
BiocManager::install(c("AnnotationHub", "DESeq2",
"ensembldb", "dblyr", "EnhancedVolcano"))
library(DESeq2)
library(ensembldb)
library(dplyr)
library(AnnotationHub)
library(EnhancedVolcano)
library(tidyverse)
```

2. **Downloading raw counts from NCBI GEO**

We are downloading RNA-seq counts from NCBI GEO (Accession number: GSE60450), under *Supplementary file,* to our RStudio working directory. We will read the dataset as a table.

```
# read RNA-seq count matrix downloaded from
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60450
. The first column contains "EntrezGeneIDs" that will be
our rownames. Remove "Length" column, as it is not needed.
```

```
counts_f <- "GSE60450_Lactation-GenewiseCounts.txt"
raw_counts <- read.table(counts_f, header=T, row.names = 1)
raw_counts <- raw_counts[,-1]
```

3. Organizing metadata for later use

Since metadata did not accompany the downloaded raw counts, we will create a metadata dataframe with the information provided on the accession page. Selecting "Samples" in the "Scope" section atop the page allows us to access full information about the sample IDs.

```
#column names to rownames for new dataframe
ID <- colnames(raw_counts)
col.data <- as.data.frame(ID)
#cell type column
col.data$celltype <- c("basal", "basal", "basal", "basal",
"basal", "basal", "luminal", "luminal", "luminal",
"luminal", "luminal", "luminal")
#status of individual
col.data$status <- c("virgin", "virgin", "pregnant",
"pregnant", "lactating", "lactating", "virgin", "virgin",
"pregnant", "pregnant", "lactating", "lactating")
```

4. Obtaining annotations from *Ensembl*

AnnotationHub allows us to extract annotations from the *Ensembl* database. We will obtain genomic data for our target species and isolate the necessary information.

```
#connect database
annot_hub <- AnnotationHub::AnnotationHub()
#Query the db and view the latest(last on the list) version
mm <- AnnotationHub::query(annot_hub, c("Mus musculus",
"EnsDb"))
mm
#select the latest
mm <- mm[["AH109655"]]
#view annotations and other info. Some columns of interest
include entrezid and description
mm_annots <- genes(mm, return.type = "data.frame")
#select specific columns
mm_annots <- genes(mm, return.type = "data.frame")%>%
  select(gene_name, description, entrezid)
```

5. Creating DESeqDataSet and data transformation

We will create a DESeqDataset which will be transformed so distance matrixes will be available to create PCA plots.

```
#create DESeqDataset first
de.dataset <- DESeqDataSetFromMatrix(countData=raw_counts,
colData=col.data, design= ~celltype + status)
#Transform dataset
?rlog
rlog_data <- rlog(de.dataset)
```

6. Creating PCA plots

We will create PCA plots and visualize clusters by cell type and status of individuals.

```
#PCA by cell type
plotPCA(rlog_data, intgroup="celltype")
#PCA by status
plotPCA(rlog_data, intgroup="status")
```

7. Running differential expression (DE) analysis

We will use the DESeq() function to produce DE results.

```
#run differential expression analysis with DESeq()
de.analysis <- DESeq(de.dataset)

#view results. Set tidy to true to view results as a
dataframe. EntrezIDs are in column 1.
de_results <- results(de.analysis, tidy=T)
```

8. Adding annotations to results

We will add mm_annots from step #3 to our results. Next is a left join because it helps to retains all results, including those rows that are absent in the annotations. I will filter out NA values later, so I prefer to keep all results at this step.

```
#add annotations but first convert entrezid column from
list to character for successful merging.
mm_annots$entrezid <- as.character(mm_annots$entrezid)
annotated_df <- left_join(de_results, mm_annots,
by=join_by(row==entrezid))
#More observations in annotated_df than the original
de_results which suggests we have duplicates. Checking...
yes, we do.
annotated_df %>%
  group_by_all() %>%
  filter(n()>1)
#remove duplicates
annotated_df <- annotated_df[!duplicated(annotated_df$row),
]
```

```
#remove NA adjusted-pvalues
annotated_df <- annotated_df %>%
  filter(!is.na(padj))
```

9. Creating volcano plot visualizing DE

We are a making a volcano plot to highlight some of the most differentially expressed genes among the samples. Our cutoff for p-value is 0.05 and 1.5 for fold change.

```
#extract names of the most significant (both up and down-regulated genes)
most_sig_genes <- top_n(annotated_df, -20, padj)$gene_name
most_sig_genes
#do volcano plot highlighting some of the most significant genes
EnhancedVolcano(annotated_df,
                lab = annotated_df$gene_name,
                 x = 'log2FoldChange',
                 y = 'pvalue',
                pCutoff = 0.05,
                FCcutoff = 1.5,
                selectLab = c("Pigr", "Wap", "Pik3r1", "Glycam1", "Dsc2", "Slc24a3", "Lao1","Fgfr3", "Kit", "Csn1s1","Csn2", "Csn1s2a", "Csn1s2b", "Naaa", "Sh2b2", "Atp6v1b1", "Fcgbp", "Rab11fip1"),
                title = "Differential Expression of luminal and basal mammary gland cells",
                subtitle = ""
                 )
```

Results

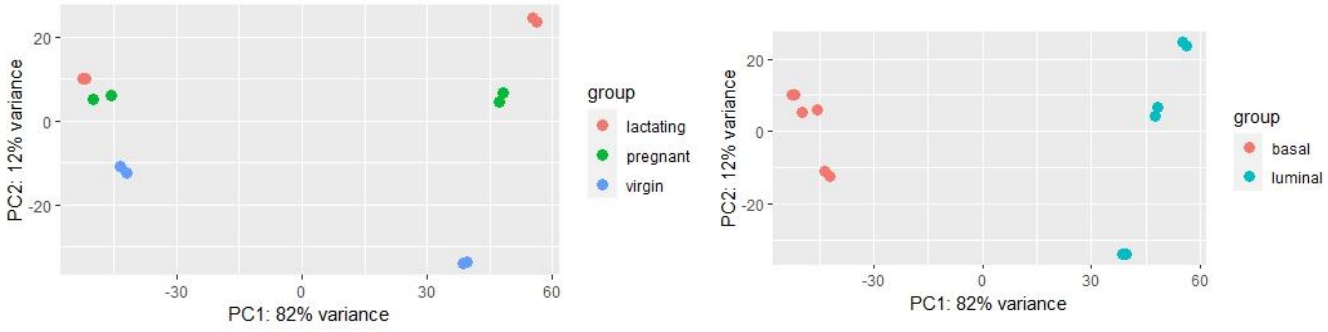

Figure 1: Left= PCA of mammary gland cells by physiological state. Right= PCA of mammary gland cells by subpopulation.

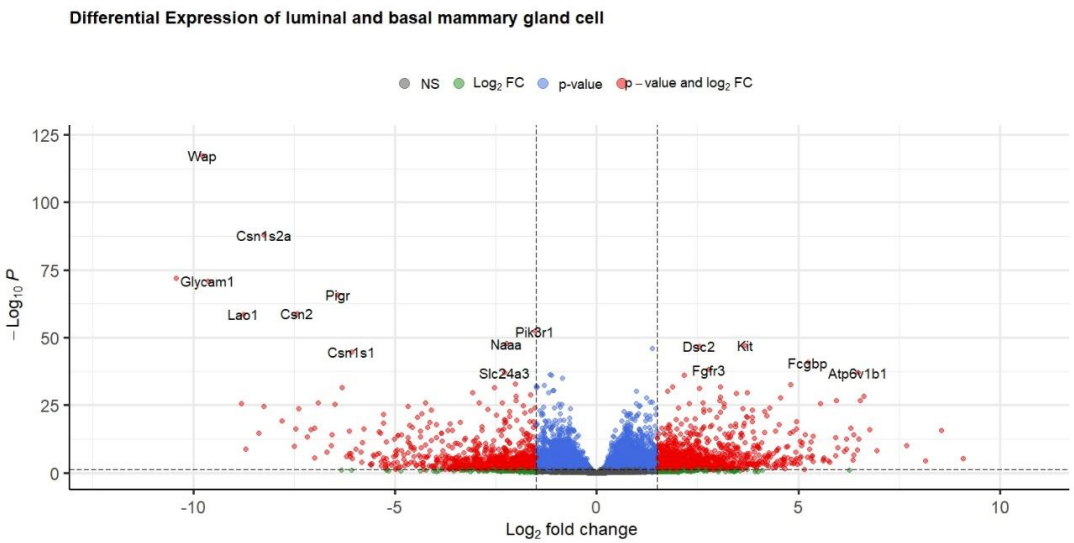

Figure 2: Highlighting genes with the most significantly different expression levels among individuals and cell subpopulations. Left half: Downregulated genes. Right half: Upregulated genes. NS=Non-significant (P<0.05). $Log_2FC=1.5$.

Gene Name	P-value	Description
PIGR	1.28E-66	polymeric immunoglobulin receptor [Source:MGI Symbol;Acc:MGI:103080]
WAP	4.98E-118	whey acidic protein [Source:MGI Symbol;Acc:MGI:98943]
CCNG1	3.29E-37	cyclin G1 [Source:MGI Symbol;Acc:MGI:102890]
PIK3R1	4.07E-53	phosphoinositide-3-kinase regulatory subunit 1 [Source:MGI Symbol;Acc:MGI:97583]
GLYCAM1	1.06E-71	glycosylation dependent cell adhesion molecule 1 [Source:MGI Symbol;Acc:MGI:95759]
DSC2	1.44E-47	desmocollin 2 [Source:MGI Symbol;Acc:MGI:103221]
SLC24A3	7.67E-38	solute carrier family 24 (sodium/potassium/calcium exchanger), member 3 [Source:MGI Symbol;Acc:MGI:2137513]
VAPB	9.42E-37	vesicle-associated membrane protein, associated protein B and C [Source:MGI Symbol;Acc:MGI:1928744]
LAO1	2.39E-59	L-amino acid oxidase 1 [Source:MGI Symbol;Acc:MGI:2140628]
FGFR3	6.57E-39	fibroblast growth factor receptor 3 [Source:MGI Symbol;Acc:MGI:95524]
KIT	6.96E-48	KIT proto-oncogene receptor tyrosine kinase [Source:MGI Symbol;Acc:MGI:96677]
CSN1S1	1.85E-45	casein alpha s1 [Source:MGI Symbol;Acc:MGI:88540]
CSN2	1.10E-59	casein beta [Source:MGI Symbol;Acc:MGI:88541]
CSN1S2A	1.06E-88	casein alpha s2-like A [Source:MGI Symbol;Acc:MGI:88542]
CSN1S2B	1.08E-72	casein alpha s2-like B [Source:MGI Symbol;Acc:MGI:105312]
NAAA	1.45E-48	N-acylethanolamine acid amidase [Source:MGI Symbol;Acc:MGI:1914361]
SH2B2	8.02E-37	SH2B adaptor protein 2 [Source:MGI Symbol;Acc:MGI:1345171]
ATP6V1B1	8.79E-38	ATPase, H+ transporting, lysosomal V1 subunit B1 [Source:MGI Symbol;Acc:MGI:103285]
FCGBP	1.56E-41	Fc fragment of IgG binding protein [Source:MGI Symbol;Acc:MGI:2444336]
RAB11FIP1	8.43E-47	RAB11 family interacting protein 1 (class I) [Source:MGI Symbol;Acc:MGI:1923017]

Table 1: 20 most significant genes having differential expression levels.

Discussion

Through PCA, I was able to cluster transformed count data of all samples. According to PCA by mammary cell subpopulation (Fig 1), principal component 1(PC2) explains that most of the variance (82%) exists between luminal and basal cell subpopulations (Fig 1). PC2 highlights that there is less variance (12%) within each subpopulation (Fig 1). I expected that mammary cells from virgin mice would vary from those in both pregnant and lactating mice since the latter groups are more similar in their physiological states. Fig 1 corroborates my assumption because there is greater variation between virgin individuals and pregnant/lactating than between the latter groups.

For my differential expression analysis, I annotated the results so that we could highlight genes with significant differential expression levels (Fig 2) and their functional properties (Table 1). Results demonstrate that more genes were significantly downregulated than were upregulated (Fig 2). ATP6V1B1 and FCGBP are examples of the upregulated genes while WAP and GLYCAM1 were downregulated (Fig 2). Table 1 highlights the roles of these genes as provided by *Ensembl.*

Although our dataset used only two mice per physiological state to investigate gene expression, results still elucidate distinctions among physiological states and cell types. We were also able to observe differences in gene expression among the groups. In a future project, I would like to build on these skills by exploring RNAseq from the alignment step to the data analysis step.

References

Yang, I. S., & Kim, S. (2015). Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. *Genomics & Informatics*, *13*(4), 119. https://doi.org/10.5808/gi.2015.13.4.119