Week 3

Hien Le

28/01/2022

EXercise set 3: Differential expression analysis

Bulk RNA-seq: statistics of data and models

In this week exercise, we will work with downstream analysis by learning about a model based methods for differntial expression analysis. DESeq2 R package is used to normalization of the expression of RNA.

Dataset

In this session, gene qunatification files is started to use under csv file.

Tasks

Tasks 1

Contruct a 2*2 table

```
sample_sheet = read.csv("/student_data/BBT.BI.203_2022/data/ex3/sample_sheet.csv", header = TRUE)
base::table(sample_sheet$Collection.method, sample_sheet$RNA.Isolation.method)
```

Questions

Question 1

How many sa ples have been collected using RP method? 5 How many RP samples have been treated with the trizol protocol? 1

Building a count matrix and fitting the DESeq2 model

Tasks

Task1

Generate a data matrix from the set of count files in the data directory

Could you suggest any ways which is better to create the data_matrix?

```
# Loading the data with 4 columns: gene ID, counts for unstranded RNA-seq, counts for the 1st read stra

BPH_659 <- read.table("/student_data/BBT.BI.203_2022/data/ex3/BPH_659.bamReadsPerGene.out.tab",sep="\t"
BPH_671 <- read.table("/student_data/BBT.BI.203_2022/data/ex3/BPH_671.bamReadsPerGene.out.tab",sep="\t"
BPH_701 <- read.table("/student_data/BBT.BI.203_2022/data/ex3/BPH_701.bamReadsPerGene.out.tab",sep="\t"

CRPC_261 <- read.table("/student_data/BBT.BI.203_2022/data/ex3/CRPC_261.bamReadsPerGene.out.tab",sep="\
CRPC_539 <- read.table("/student_data/BBT.BI.203_2022/data/ex3/CRPC_539.bamReadsPerGene.out.tab",sep="\
CRPC_541 <- read.table("/student_data/BBT.BI.203_2022/data/ex3/CRPC_541.bamReadsPerGene.out.tab",sep="\

PC_12517 <- read.table("/student_data/BBT.BI.203_2022/data/ex3/PC_12517.bamReadsPerGene.out.tab",sep="\
PC_15194 <- read.table("/student_data/BBT.BI.203_2022/data/ex3/PC_15194.bamReadsPerGene.out.tab",sep="\
PC_19403 <- read.table("/student_data/BBT.BI.203_2022/data/ex3/PC_19403.bamReadsPerGene.out.tab",sep="\

data_matrix = matrix(, ncol = 9, nrow = 58243)

data_matrix[,1] = BPH_659$V2
data_matrix[,2] = BPH_671$V2
data_matrix[,3] = BPH_701$V2
data_matrix[,4] = CRPC_261$V2
data_matrix[,5] = CRPC_539$V2
data_matrix[,6] = CRPC_541$V2
data_matrix[,7] = PC_12517$V2
data_matrix[,8] = PC_15194$V2
data_matrix[,9] = PC_19403$V2
row.names(data_matrix) = BPH_659$V1
colnames(data_matrix) = c("BPH_659", "BPH_671", "BPH_701", "CRPC_261", "CRPC_539", "CRPC_541", "PC_1251
```

Task2

Load the sample sheet into an R session and gennerate a formula object

```
sample_sheet = read.csv("/student_data/BBT.BI.203_2022/data/ex3/sample_sheet.csv", header = TRUE)
formula_object = as.formula(~ Type + Collection.method + RNA.Isolation.method)
```

Task3

Generate a DESeqDataSet object

```
library(DESeq2)
DeSeqDataSet <- DESeqDataSetFromMatrix(countData=data_matrix,
                                       colData=sample_sheet,
                                       design=formula_object)
```

Questions

Question 1

What are the dimensions of the count matrix? dim(data_matrix) = 58243 rows and 9 columns

Question 2

In the task2, we need to create the formula: ~ Collection.method + Sample.name

Fitting the DESeq2 model

Tasks

Task1

Compute the overall read count for each gene using the rowSums function

```
rowSums(data_matrix)
```

Task2

Remove genes with an overall read count of zero

```
df_row_sum = data.frame(apply(data_matrix, 1, sum))
valid_rows = df_row_sum > 0
data_matrix_filtered = data_matrix[valid_rows,]
dim(data_matrix_filtered)
dim(data_matrix)
```

Task3

From the remining genes, estimate the read count value corresponding to the lower 10% of the distribution using the quantile() function

```
df_row_filtered = apply(data_matrix_filtered, 1, sum)
threshold <- quantile(df_row_filtered, 0.1, na.rm=FALSE, names = TRUE)
threshold
```

Task4

Plot a histogram of the overall read count values

```
hist(df_row_filtered, xlim = c(0,10000), breaks ="fd", xlab = "gene_count", main = "Histograme of gene c
abline(v=0.01, col="red")
```

Task5

Remove genes with an overall read count lowere of equal to the quntile value from Task 3

```
row_sum_filtered <- data.frame(apply(data_matrix_filtered, 1, sum))
valid_rows_filtered <- row_sum_filtered > threshold
data_matrix_filtered_threshold = data_matrix_filtered[valid_rows_filtered,]
dim(data_matrix_filtered_threshold)
```

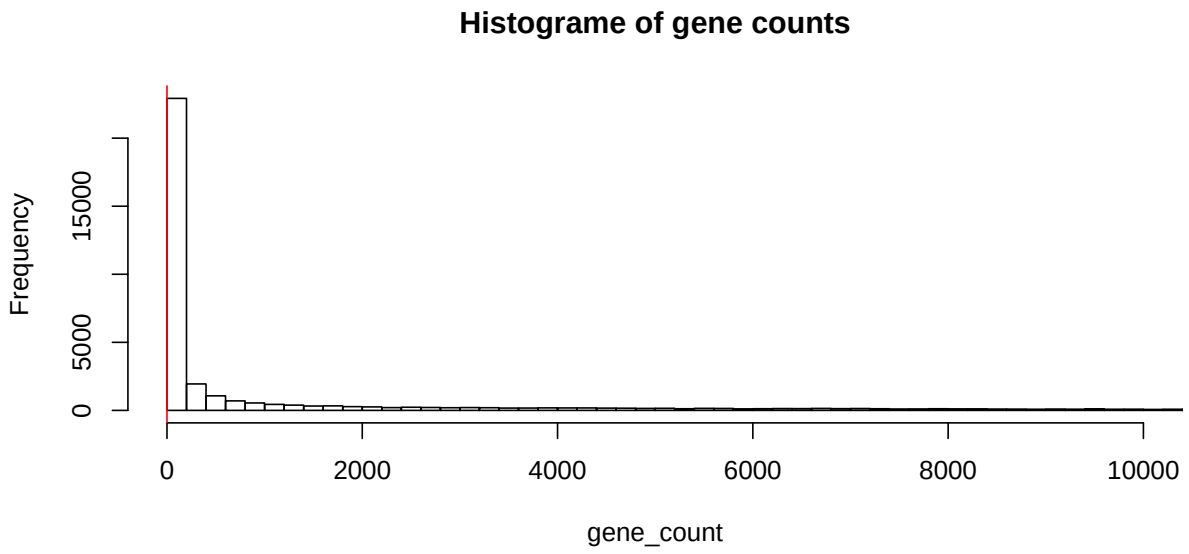

Figure 1: Histograme of gene counts

Task6

Filter the DESeqDataSet object with the genes that have passed all of the filtering above

```
DESeqDataSet_filterd =  DESeqDataSetFromMatrix(data_matrix_filtered_threshold,
                                    colData=sample_sheet,
                                    design=formula_object)
```

Task7

Fit the DESeq2 modle using the DESeq() function.

```
DESeq(DESeqDataSet_filterd)

dds = estimateSizeFactors(DESeqDataSet_filterd)
dds = estimateDispersions(DESeqDataSet_filterd)
```

Task8

Plot the dispersion parameter against the means of normalized counts using the plotDisEsts function

```
plotDispEsts(DESeq(DESeqDataSet_filterd))
```

Questions

Question1

How many genes have a read count of zero in every sample? 17369 genes

What is the percentage with respect to all genes? 29.82%

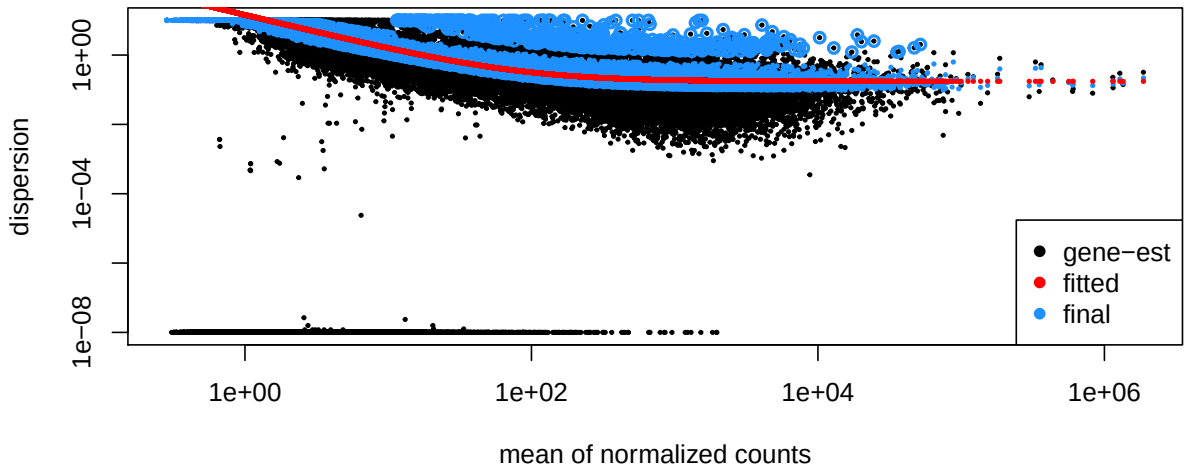

Figure 2: Dispersion plot

Question2

How many genes get discarded because they have an overall count lower of equal to the value from the Task3? 4071

Question3

How many genes remain after all the filtering? 35481 genes

Question4

Extract the model matrix from the DESeqDataSet object

```
dds <- DESeq(DESeqDataSet_filterd)
attr(dds, "modelMatrix")
```

a: How many columns dose it have? 5

b: What do the represent? They represent the information of sample_sheet.csv

c: Why there is no columns describing BPH samples? Because BPH samples are used as the control for the nomalization here. We can see the comparision of CRPC with BPH and PC with BPH.

d. What dose the Intercept column represent? y-intercept in a regression model.

Question5

Inspect the dispersion plot.

Calling differntially expressed genes

Now that we have the DESeq2 model fitted to our data, we can proceed to call differentially expressed genes (DEGs). In this section, we will compute log-fold changes, p-values, and adjusted p-values for two example sample comparisons.

Tasks

Task1

Use the results() function to compute log2 fold changes and p-values for the BPH and PC comparison.

```
a = results(dds, contrast = c("Type","PC", "BPH"))
a
```

Task2

Use the results() function to compute log2 fold changes and p-values for the RP and TURP collection method comparison

```
b = results(dds, contrast = c("Collection.method","RP", "TURP"))
```

Task3

At this point, save R workspace as an RData file so that it can be loaded and reused easily later.

```
save.image(file = "/student_data/BBT.BI.203_2022/students/leh/ex3/week_3.RData")
```

Task4

Use the lfcShrink function to compute shrunken log2-fold changes for BPH and PC comparison from Task 1.

```
lfcShrink(dds, coef=3, res= BPH_PC_results, type = "apeglm") # My R studio has the error with lfcShrink
```

Error in lfsShrink: could not find function "lfcShrink". I notice it quite late so I have not had time to contact and fix it before submision.

Task5

```
DeSeqDataSet # from task3 of building a count matrix and fitting the DESeq2 model
DESeqDataSet_filterd # from task6 in section Fitting the DESeq2 model.

DESeq_dataset = DeSeqDataSet
# top-level function for finding number of DEGs
getDEGs <- function (DESeq_dataset, contrast, IndependentFiltering){
  if (IndependentFiltering==TRUE){
    DESeq_dataset <- DESeq(DESeq_dataset)
    res <- results(DESeq_dataset, contrast = contrast, independentFiltering=TRUE)
    differentially_expressed <- sum(res$padj[!is.na(res$padj)] < .1)
    return(differentially_expressed)
  }
```

```
  if (IndependentFiltering==FALSE){
    DESeq_dataset <- DESeq(DESeq_dataset)
    res <- results(DESeq_dataset, contrast = contrast, independentFiltering=FALSE)
    differentially_expressed <- sum(res$padj[!is.na(res$padj)] < .1)
    return(differentially_expressed)
  }
}

# call the top-level function with independent filtering
DEGs <- getDEGs(DESeq_dataset, c("Type", "BPH", "PC"), TRUE)

# call the top-level function without independent filtering
DEGs <- getDEGs(DESeq_dataset, c("Type", "BPH", "PC"), FALSE)
```

Questions

Question1

Inspect the DESeqResults object from Tas1 using the summary function.

```
summary(a)
```

Results: out of 36803 with nonzero total read count, with adjusted p values < 0.1, 218 DEGs are upregulated (0.61%) while 174 DEGs are downregulated (0.49%). Outliers are reported 0 (0%).

Question2

Inspect the DSSeqResults object from Task2 using the summary function.

```
summary(b)
```

Uisng the differences in RNA collection methods (RP and TURP) with adjusted p-value < 0.1, 9 DEGs are upregulated (0.024%) while 70 DEGs are downregulated (0.21%).

Question3

Inspect the DESeqResults object from Task4 using the summary. lfcShrink function does not work in my R-studio. I have noticed it too late so I have not had time to fix it. However, I guess the number of DEGs is the same in the both analysis methods.

Question4

Look at the analysis from Task5 and anwer the following questions. a. 267

b. 420

c. 392

d. When there is no filtering, DESeq exludes low-count genes to optimize analysis.