## June 2018

# An Introduction to Deep-Sequencing Data Analysis

## Exercise #5: Learn how to detect and cluster differentially expressed genes from RNA-Seq data using DESeq2

***Dena Leshkowitz & Ester Feldmesser***

### Introduction

In previous exercises we performed QC, and mapping of RNA-Seq data from an experiment performed in Arabidopsis. For the purpose of the exercise, we chose 4 time points for which we have biological duplicates. In this exercise we will run the DESeq2 package in R in order to detect genes that are differentially expressed between day 8 and the other days.

Answers for exercise 3 are found at- http://dors.weizmann.ac.il/course/course2018/AnswersForExercise3.docx

### Instructions

### Part 1

We will run R on WEXAC using the Rstudio application.

In a browser type the following URL and enter your class user and password.

Class 1-20 use <https://appsrv.wexac.weizmann.ac.il/rstudio/auth-sign-in>

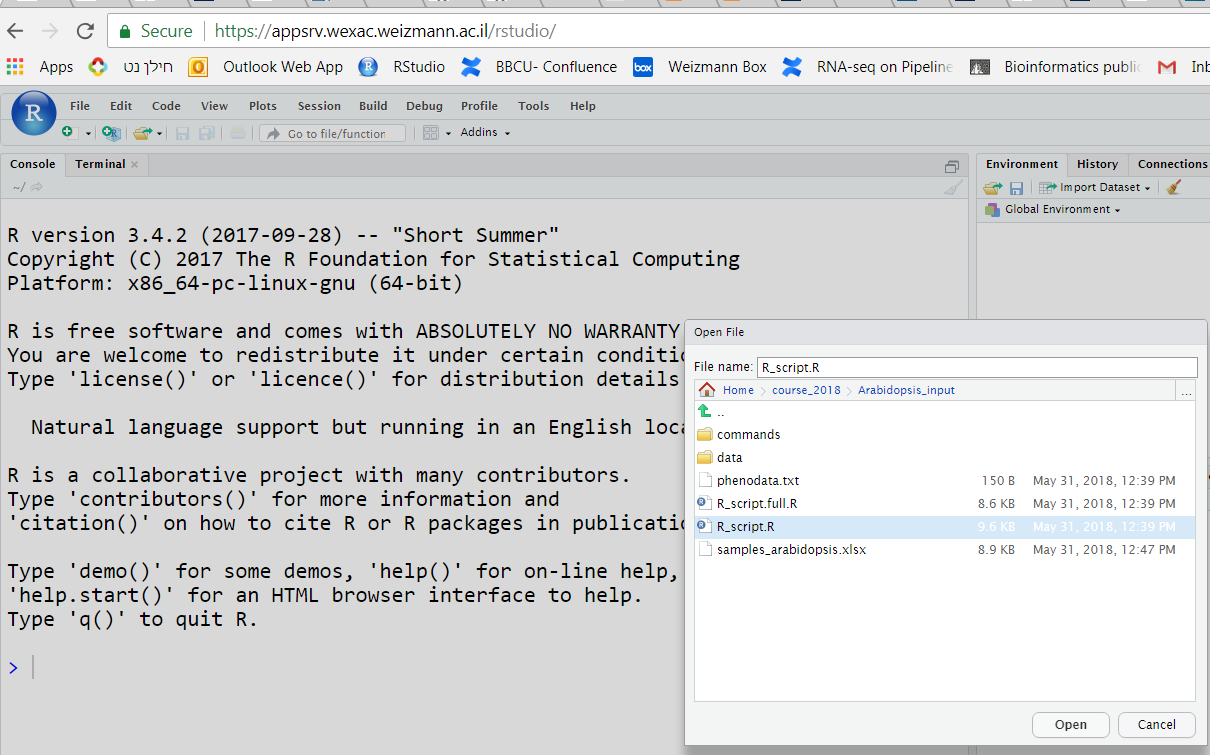
Class 20-40 use <https://appsrv.wexac.weizmann.ac.il/biostudio/auth-sign-in>

In the terminal run the following commands:

*mkdir ~/RNA-Seq-Arabodopsis-Deseq*

*cp ~/course\_2018/Arabidopsis\_input/Arabidopsis\_script.Rmd ~/RNA-Seq-Arabodopsis-Deseq*

In the Rstudio application, open the File menu and click Open File… to load the Rmd script (named: Arabidopsis\_script.Rmd) found under ~/RNA-Seq-Arabodopsis-Deseq, as shown below (orange arrow).



Rmd is a file format with chunks of embedded R code. For more information on this format see <https://rmarkdown.rstudio.com/articles_intro.html>.

Following is an example of a chunk-

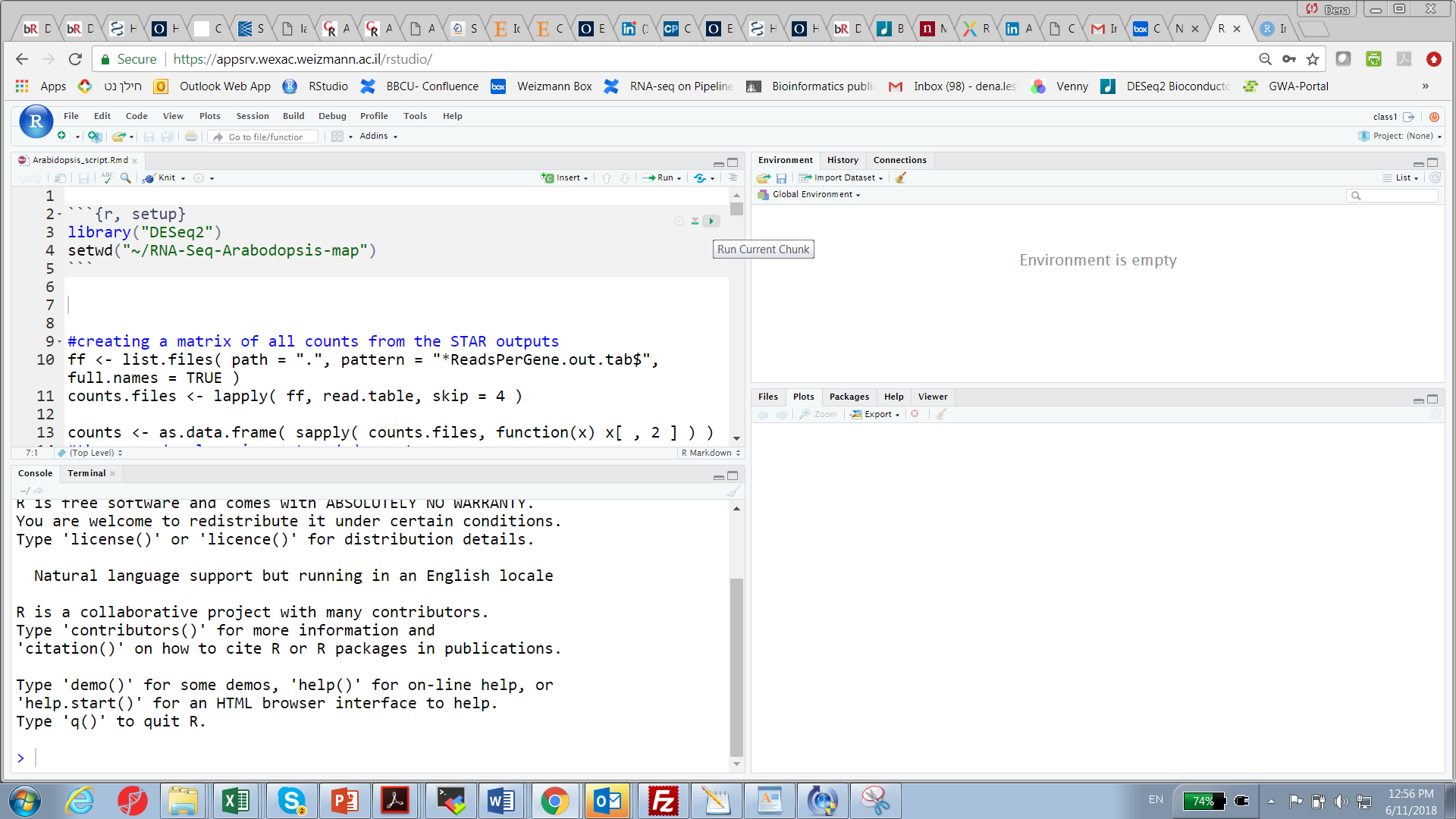
```{r}

summary(cars)

```

summary(cars) is the command, the other lines delimit the chunk.

You can run within the file each command separately or a whole chunk as below.

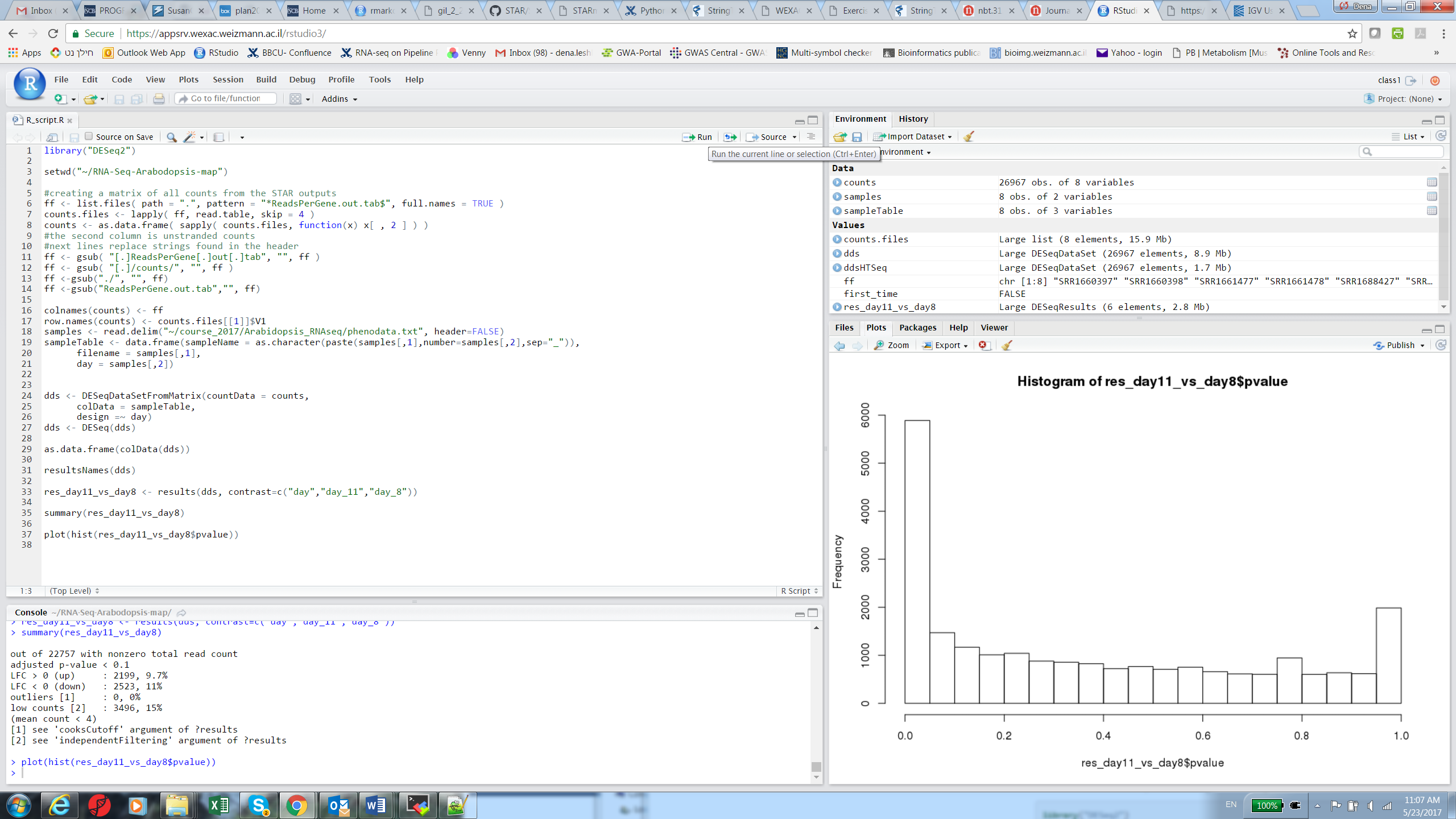


In addition you can make an html file with the commands and their outputs such as plots. Within this exercise you will create such a file.

Between the chunks there are lines that are not commands.

They are explanations on the commands within the chunks.

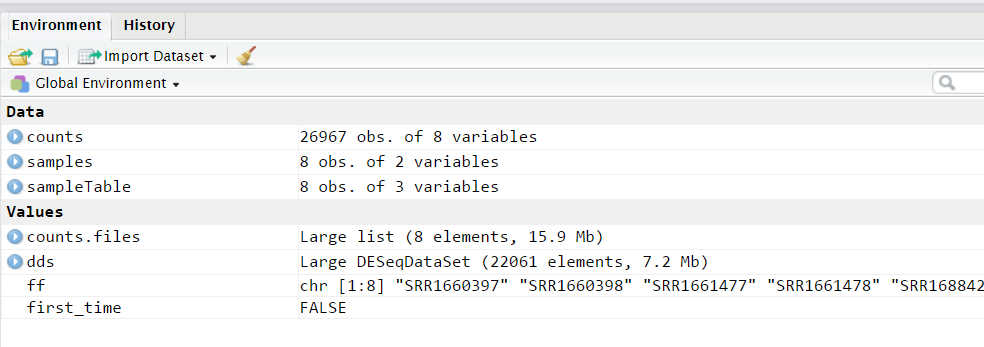
The lines that start with # are comments – not a command.



Go to Rstudio, run the first four chunks and return here. Be sure that you understand the commands that you are running, otherwise ask.

In the right panel you can see the objects that you will create (as below)

Please notice that the questions in the Rmd file are also listed below for you convenience.



To see the counts matrix you can click on counts.

**Question 1: Look at the dimensions of the counts object and answer: for how many genes do we have counts?**

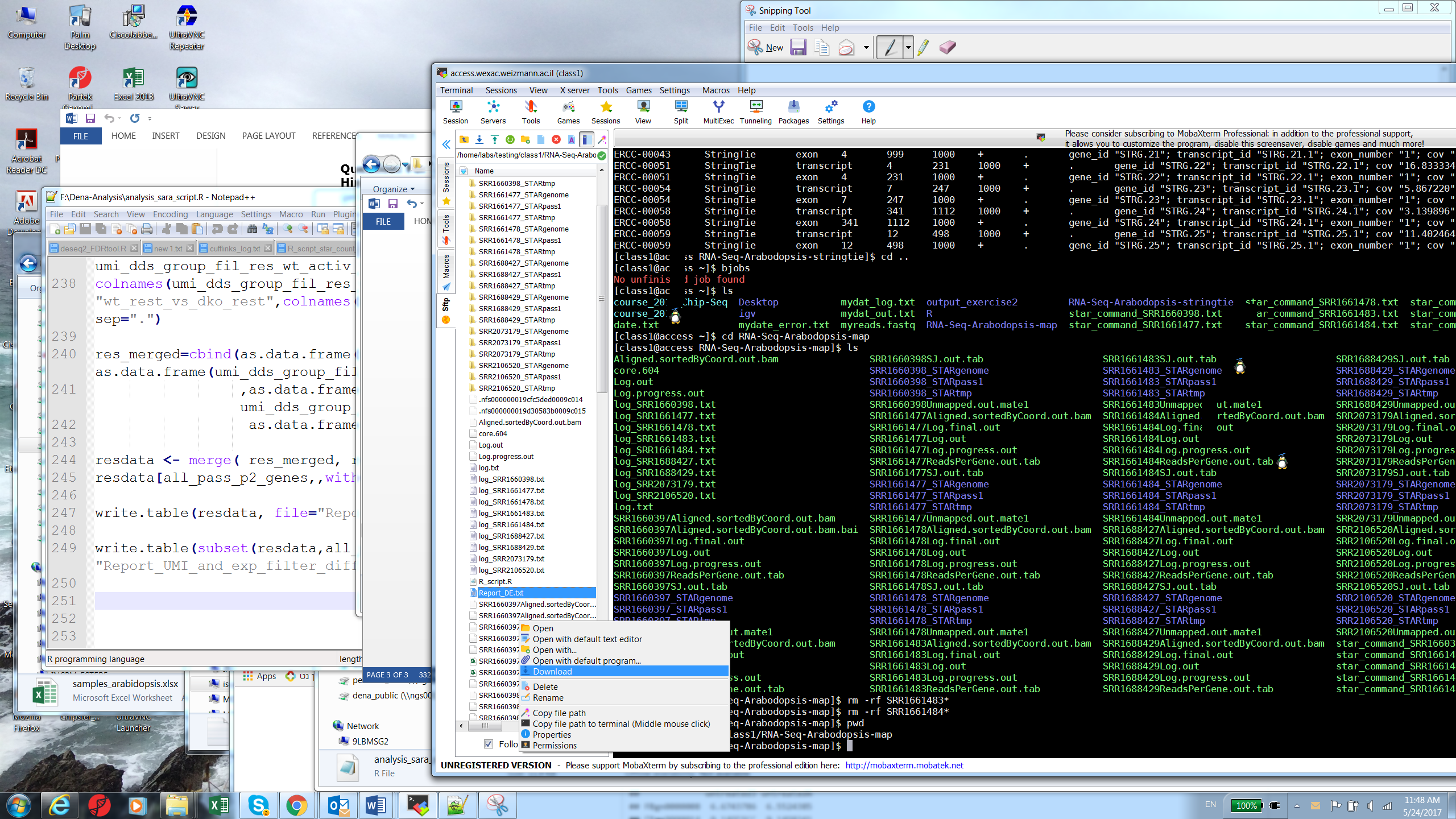
**Question 2: How many genes were removed from the dds object?**

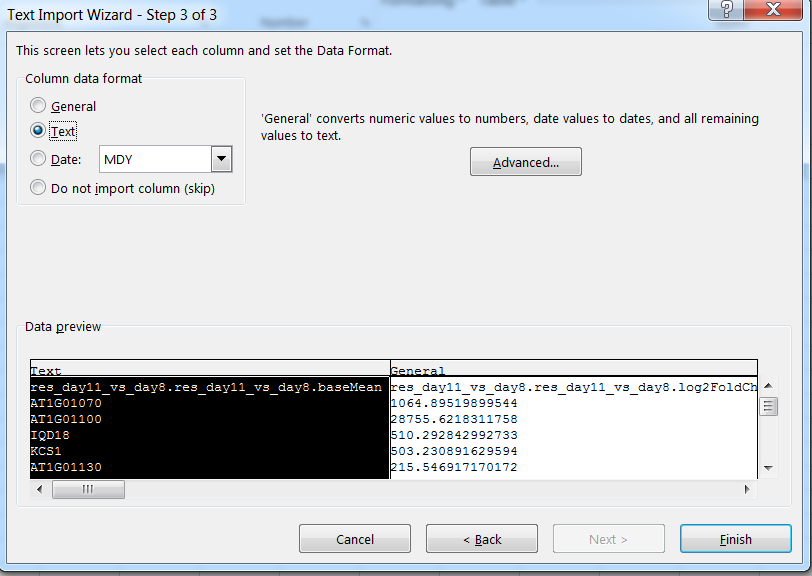
**Question 3:**

**Observing the size factors, which sample has the highest number of counts? Hint for normalization, counts are divided by size factor.**

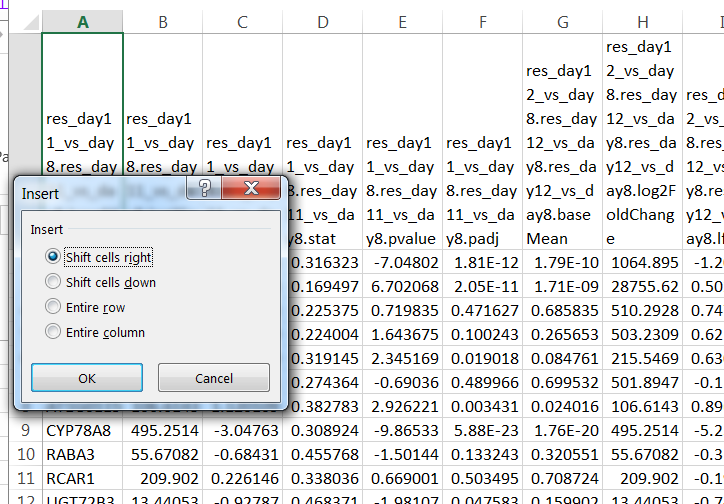
AFTER FINISHING PART 1 –

To open the report file you created, return to MobaXterm window. On the left there is a sftp pane, set the location to the required path to reach the report file (see below red arrow), right click on the report file and select the option “Download”. Open the file with Excel. Make sure to define the first column that contains the gene symbol names as text, see below. This step is important since some gene names are otherwise interpreted as dates and are changed by Excel.





There is a need to shift the first cell of the header to the right.



You are now ready for the second part of the exercise. You can return now to the Rstudio.

### Part2

The second part of the exercise will include exploratory analysis, to learn about the relationship between the samples (replicates and different conditions) and clustering of the differentially expressed genes that will be used later for functional analysis.

### Exploratory analysis

### In this part you will continue performing the commands that appear in the script, for the exploratory analysis, we will build a matrix of correlations between the samples, note that the correlations are clustered, and then we will perform hierarchical clustering of the samples using Pearson’s dissimilarity. Last, we will plot a principal component analysis (PCA).

### Question 4: Which samples are close to each other? Are these the replicated samples?

If you don’t remember the names of the samples, you can click on the sampleTable as you did before for counts.

### Clustering of differentially expressed genes

The differentially expressed genes will be clustered using k-means. The values used as input for the clustering are normalized log values for DE genes. The optimal number of clusters to be used in k-means needs to be determined and this task is not trivial, if we choose too little clusters, they will be very heterogeneous, if we choose too much, they will over fit the data.

From <https://www.r-bloggers.com/finding-optimal-number-of-clusters/>:

“The elbow method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn’t give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the “elbow criterion”. This “elbow” cannot always be unambiguously identified.”

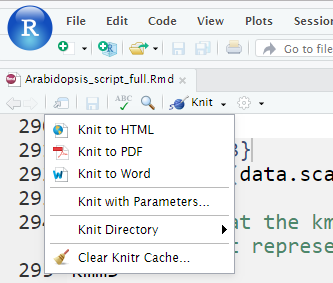
The parameter we will look at in the plot of the exercise is the within clusters sum of squares (an estimate of the variance inside the clusters) and it should be smaller as we add clusters. This estimate is negatively correlated to the variance explained and can be used instead.

**Question 5: Can you identify clearly the optimal number of clusters? Which number/s do you suggest to try?**

**Question 6: What part of the variance is explained with 3 clusters?**

**Question 7: Do 4 clusters look better (more homogeneity inside clusters) than 3? Answer just by looking by eye at the k-means with 3 or 4 clusters.**

Your script is now ready, save it as Arabidopsis\_script\_full.Rmd and run it to make an html report, by selecting the Knit button and knit to HTML (see below).



### Supplementary

### DESeq2 vignettes (manual)

<https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#summarizedexperiment-input>

DESeq2 example of workflow <http://www.bioconductor.org/help/workflows/rnaseqGene/>

Elbow method to determine the number of clusters in k-means

<https://www.r-bloggers.com/finding-optimal-number-of-clusters/>