December 2019

An Introduction to Deep-Sequencing Data Analysis
Exercise 5: Clustering of gene expression data

Ester Feldmesser and Bareket Dassa

Please read all the instructions before you start the exercise.

Create a folder called “exercise5” in your WEXAC-mounted classNN folder (testing\classNN\exercise5). Create a file with your answers called “Exercise_5_2019_answers.docx”.

In this assignment, you will perform an exploratory analysis on RNA-Seq data. The data was originated during a time course flower development experiment in Arabidopsis. It was downloaded from the public domain, see the following link for more details: https://www.ncbi.nlm.nih.gov/pubmed/26084880. We chose four time points of development, for which we have biological duplicates.

We performed an RNA-seq analysis, which includes mapping of the reads on Arabidopsis genome using STAR and quantifying the reads on gene exons. We then ran the DESeq2 package in R in order to detect genes that are differentially expressed between day 8 and the other days.

The exploratory analysis helps to learn about the relationship between the samples (replicates and different conditions), and clustering of the differentially expressed genes shows different gene expression patterns that will be used later for functional analysis.

Exploratory analysis

Download the gene expression table from:

http://dors.weizmann.ac.il/course/exercise_5_2019/arabidopsis_rld.csv

The values displayed in columns in this file are log2 normalized values, and all the genes in the experiment are included.

In order to explore the relationship between the samples (replicates and different conditions), you will use the **iDEP.90** web-tool.

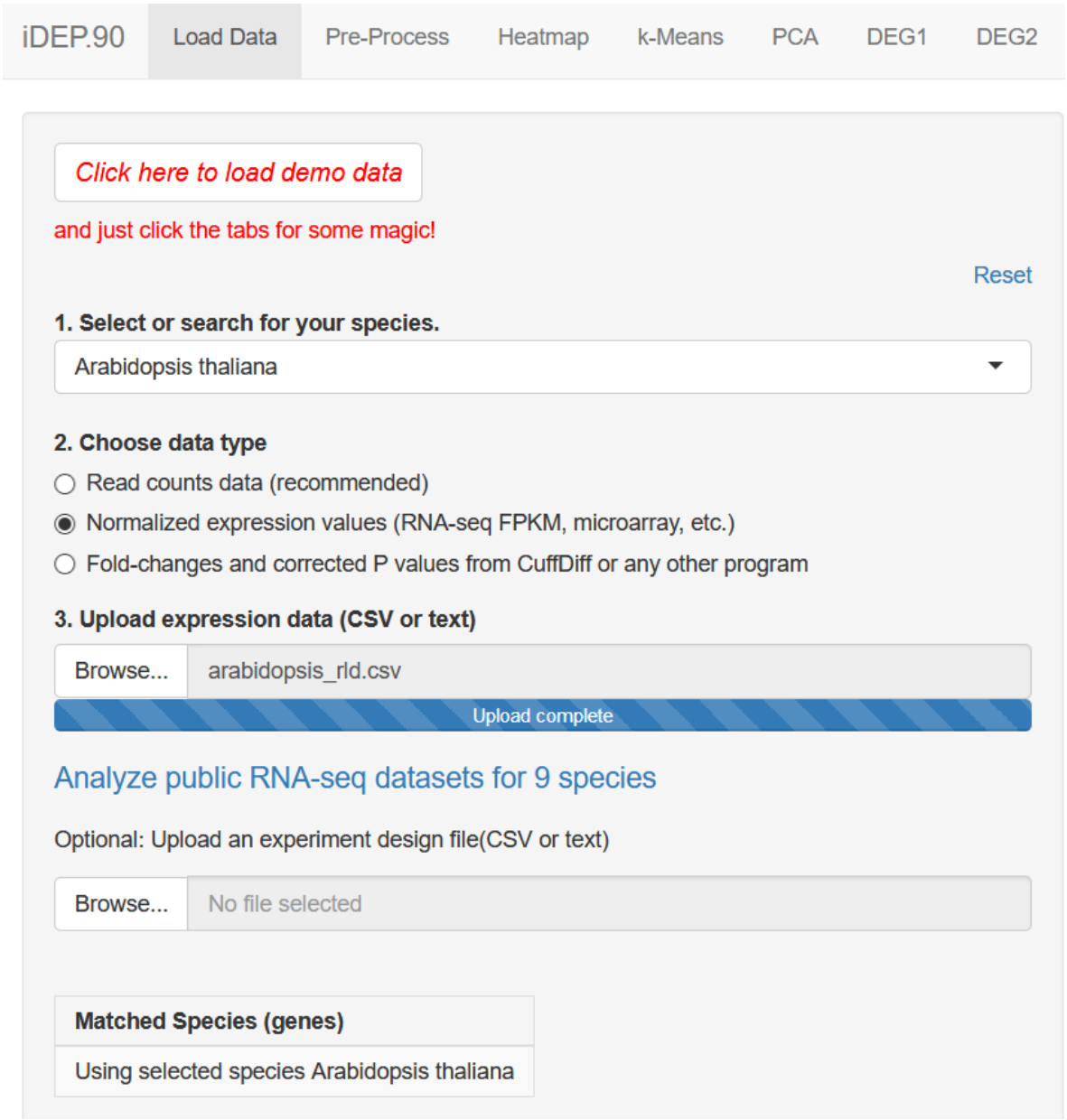
Open the url: http://bioinformatics.sdstate.edu/idep/ on your browser.

Select “Load Data” from the top menu.

Select *Arabidopsis thaliana* for your species.

Choose data type: Normalized expression values (RNA-seq FPKM, microarray, etc.)
Our data is already normalized an log2 transformed

Upload the Arabidopsis gene expression table “arabidopsis_rld.csv”.
Wait patiently until the upload is complete and a table of values appears on the right side of the page.

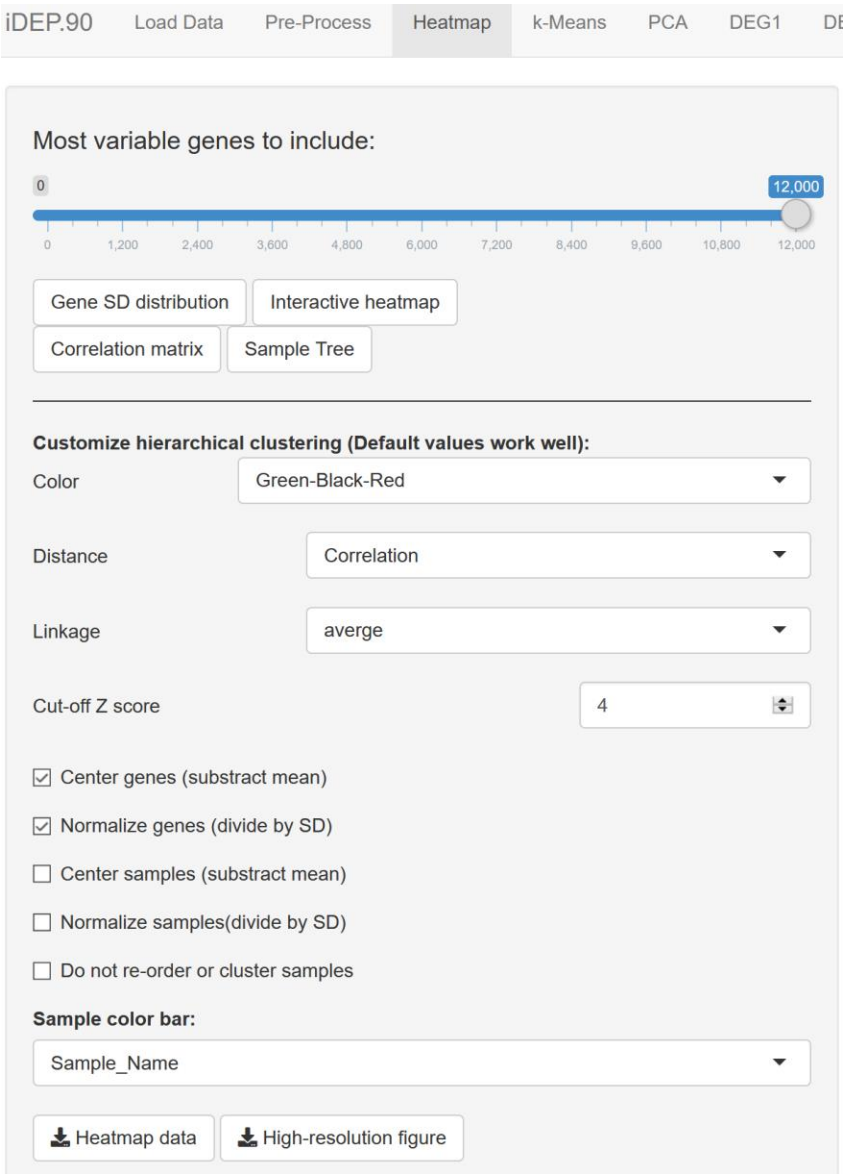


Click on “PCA” from the top menu. Look at the graph and answer:

1. A. What part of the variance is explained by the first PC?
   B. Which two days are the most similar?
   C. Save the image in your answers document.

Select “Heatmap” to visualize hierarchical clustering. Make sure you include ALL the allowed variable genes in the “Most variable genes to include”, and check the box for “Normalized genes” (see below). The later will standardize the gene values.

Although the heatmap shows 12,000 genes, we are interested for the exploratory analysis in clustering only the samples and not the genes.



To see the samples clustering, click on “Sample Tree”.

2. A. Save the image in your answers document.
   B. Use the tree or the “correlation matrix” to answer:
      Are the replicates of the same day clustered together?
   C. The upper part of the dendrogram divides the samples into two clusters.
      Name the samples that are assigned to each cluster.

Close the “Sample Tree” window.

3. Copy the heatmap (hierarchical clustering of the 12.000 most variable genes) and save it in your answers file.

Clustering of differentially expressed (DE) genes

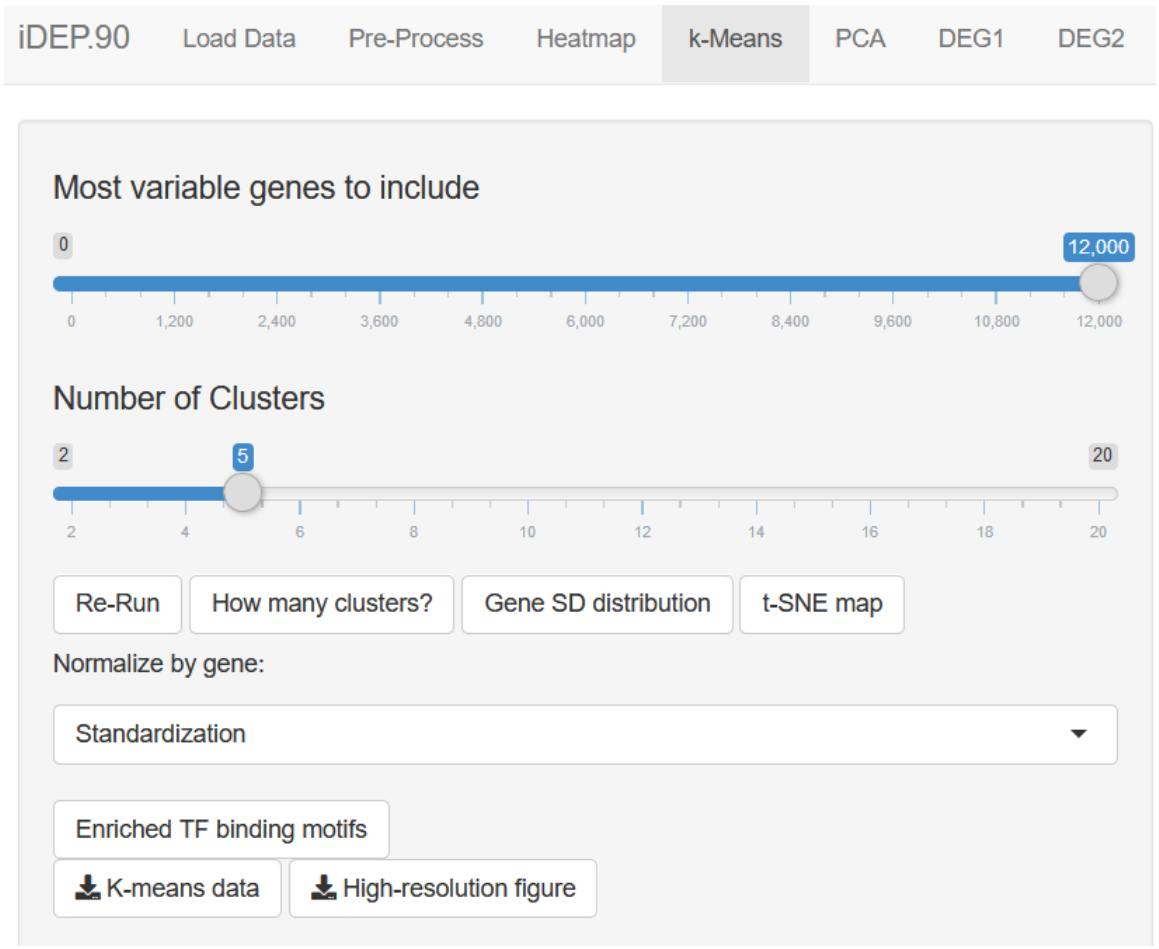
The differentially expressed genes will be clustered using k-means. The values used as input for the clustering are normalized log values for DE genes. The optimal number of clusters to be used in k-means needs to be determined and this task is not trivial. If we choose too little clusters, the clusters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

Download the gene expression table from:
http://dors.weizmann.ac.il/course/exercise_5_2019/arabidopsis_DE_rld.csv

Go back to the “Load Data” tab, and Reset it.

Upload the file using the same parameters as before and then select the “k-Means” analysis tab from the top menu.

Set the “Most variable genes to include” to the maximum in order to include all the differentially expressed genes in our clustering.
Set the “Normalize by gene” to Standardization.



Next you will evaluate the number of clusters that represent different gene expression patterns. Use the “How many clusters” function to estimate the number of clusters from the elbow graph.

“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 should be chosen at this point, hence the “elbow criterion”. This “elbow” cannot always be unambiguously identified.” (https://www.r-bloggers.com/finding-optimal-number-of-clusters/).

The parameter we will look at in the Elbow plot 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.

4. Can you estimate what is the optimal number of clusters that will capture the differences between days? Explain your answer.
5. Set the number of clusters to 2.
   Copy and save the image of the clustered heatmap that you created.
6. Set the number of clusters to 5.
   Copy and save the image of the clustered heatmap that you created.
7. In which of the two heatmaps (using either 2 or 5 clusters) are the clusters more homogenous?
8. Choose the cluster (from the 5 clusters heatmap) that shows upregulation in day 16. How many genes are in this cluster?
9. Find a cluster (from the 5 clusters heatmap) that shows upregulated genes only in early flower development. What is the name of this cluster? How many genes are in this cluster?
10. Bonus question: look at the heatmap that you created in question 3, using the 12.000 most variable genes. Are you able to detect from this heatmap which are the differentially expressed genes? Please explain.

The end!