March 2016

Exercise 3: Clustering of RNA-Seq data

Gilgi Friedlander

In this exercise you will cluster Arabidopsis RNA-Seq data. The data was taken from the paper of Klepikova et al., 2015:
RNA-seq analysis of an apical meristem time series reveals a critical point in Arabidopsis thaliana flower initiation
http://www.ncbi.nlm.nih.gov/pubmed/?term=26084880

For the purpose of the exercise, we chose 4 time points, each time point in duplicate.

The data was analyzed using the following workflow:
The fastq files were mapped to the Arabidopsis genome using TopHat.
HTSeq count was used for counting reads on Arabidopsis genes (TAIR10).
The DESeq2 package was used for normalizing the data and for differential expression analysis.

Today you will cluster the genes in this data. We will use the EXPANDER package in order to cluster these genes. EXPANDER (EXpression Analyzer and DisplayER) is a java-based tool for analysis of gene expression data from the lab of Prof. Ron Shamir, in Tel Aviv University (http://acgt.cs.tau.ac.il/expander/). It is a free package, and anyone can register and download it.

You will find all files for this exercise in the following box link:

https://weizmann.box.com/s/rjjoohm89c66d3fvd4x20mu7h8hq07zt

For the purpose of the exercise, download Expander from that folder (Expander7.1Win).

Save the file and unzip in a folder in Drive D.
(For using Expander for other purposes, please register and download from the Expander site: http://acgt.cs.tau.ac.il/expander/).

Under the “Expander” folder, create a folder named data.
Download the following two files to the data directory:
norm_counts_plus1_all_genes_max_count_gt_20_for_HierarchicalClustering.txt
norm_counts_plus1_fold_abv_4_padj_max_count_20_for_clustering_genes.txt

Create also a directory named results.

The Expander package has many tools. We will use today their clustering tools.

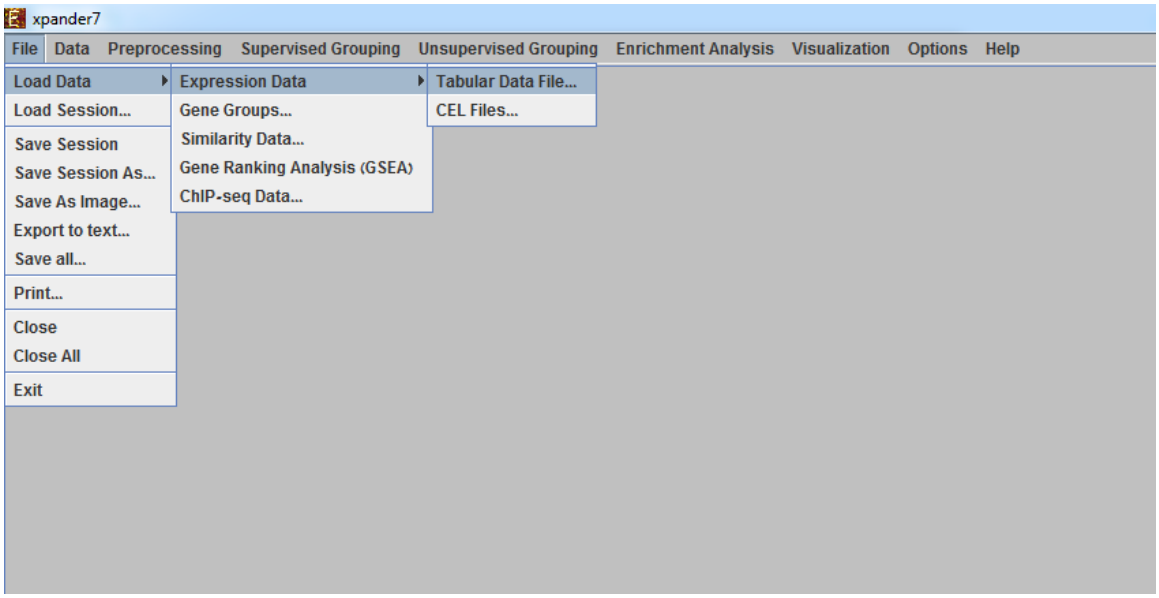
Go to the Expander directory. In this directory double click on **Expander.bat**

Part I: Clustering the samples

The data file you downloaded:
norm_counts_plus1_all_genes_max_count_gt_20_for_HierarchicalClustering.txt
includes all genes that had a count of at least 20 at least in one sample. The genes are on rows. The file includes 18,818 genes. The first column is the gene ID, the second column is the gene symbol, and the next columns are the 8 conditions. The values are DESeq2 normalized counts. 1 was added to all values, so we will be able to log transform the values. Remark: DESeq2 has transformations (such as rld) that can also be used for the clustering.

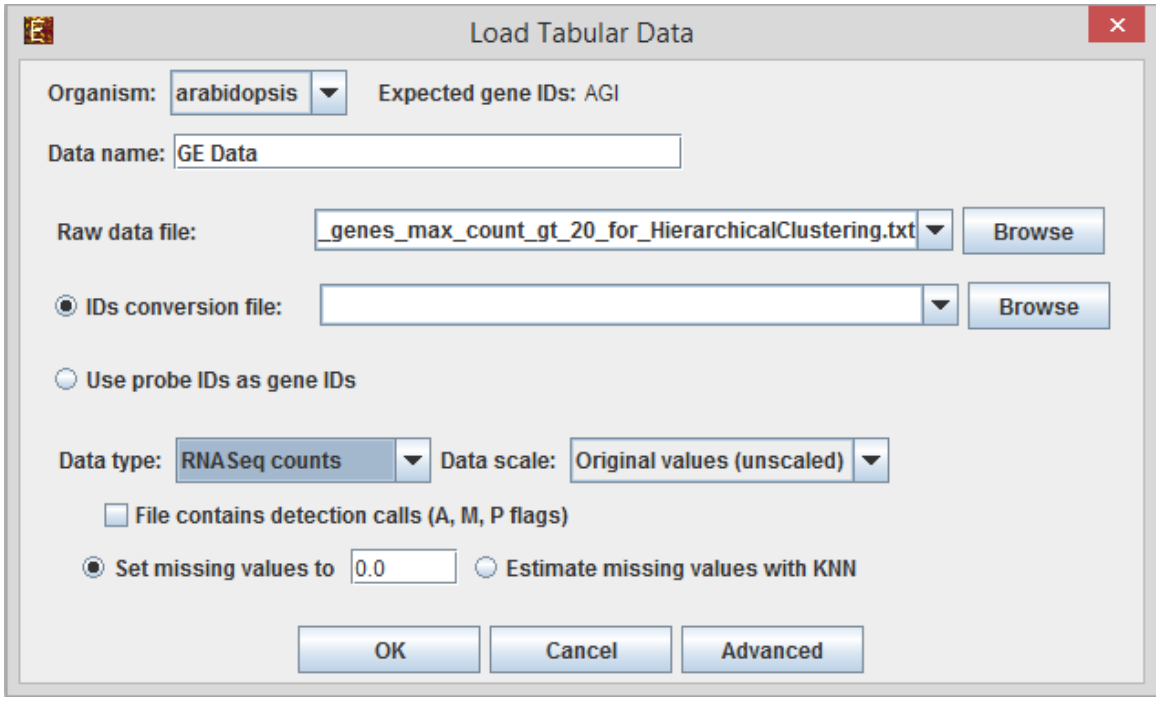
Q1. Why do you think the data was filtered to keep only genes with count of 20 at least in one condition?

Upload that file to Expander:

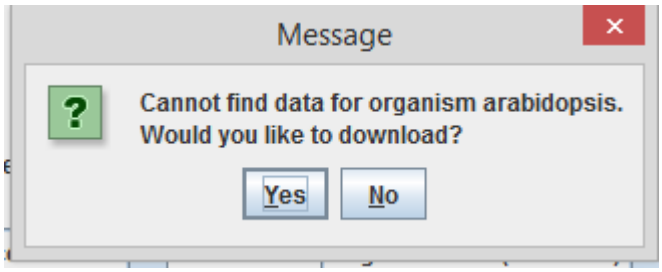


Organism is Arabidopsis (for current exercise it is not important to fill that, since we will only cluster the data, and this information is not required for the clustering).
Choose the file norm_counts_plus1_all_genes_max_count_gt_20_for_HierarchicalClustering.txt you downloaded.

Fill the fields as in the following image:



You will get the following message:



Click No. For the current exercise this data is not required.

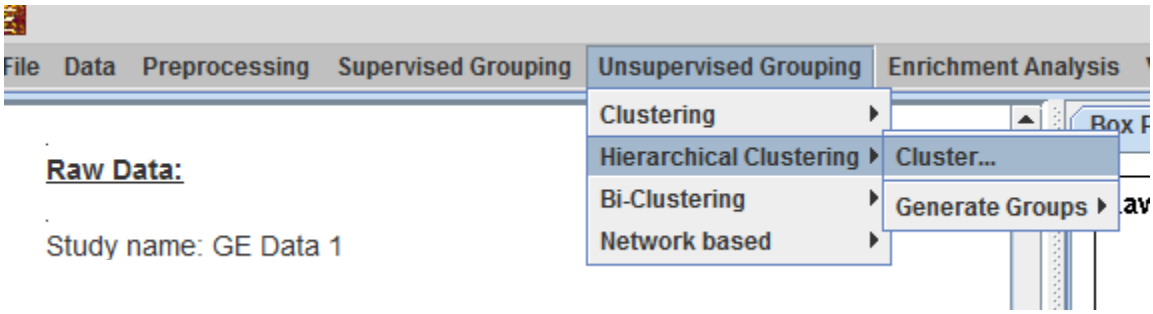
You will also get:

Click OK.

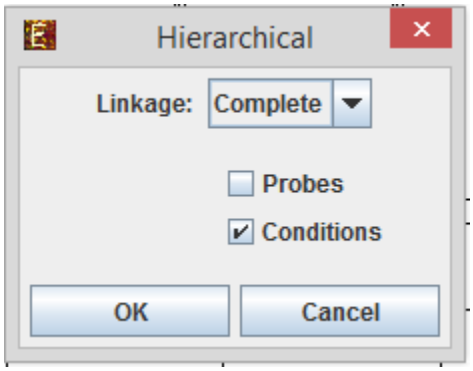
We will cluster the log counts, so first transform the counts to log2base by:

Preprocessing -> Log data

We will now perform hierarchical clustering on the samples:



We will cluster only the conditions, and we will use complete linkage:



click OK.

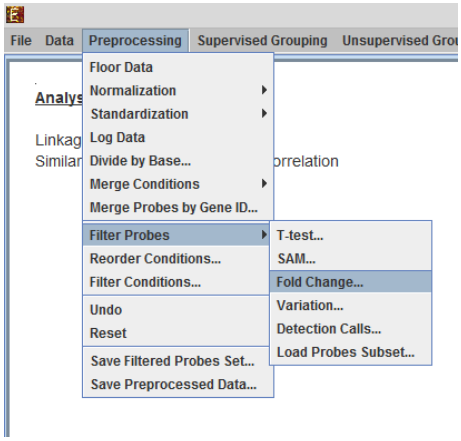
On the right pane you will see the heatmap and on top the dendrogram of the clustering. The heatmap is not informative, so we will focus only on the dendrogram.

Q2. Do the replicates cluster together?

When clustering samples, we can reduce noise by selecting the variable genes. One option is to filter the genes, to keep only genes that their expression changed between any two conditions above 1.5 fold.
Important note: this filtration is NOT based on the significance of the differential expression!!

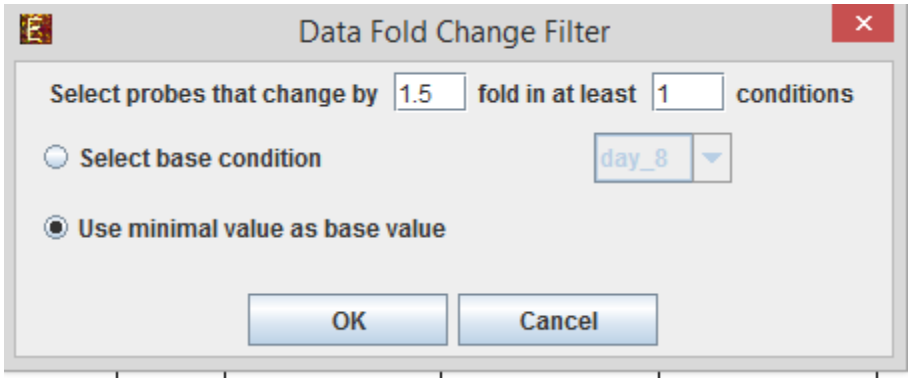
When we cluster the samples, we are asking whether the replicates are clustered together, so we do not want to take in account genes based on their significance!!

To filter the genes:



To keep also previous analysis, choose “Open an additional datasheet and continue”.

For each gene, we take the minimal value, and we will keep a gene if at least in 1 sample it changed above 1.5 fold:



16,919 genes are left after that filtration. Perform the hierarchical clustering on the samples, as before.

Q3. Do the duplicates cluster together after filtering the genes?

Part II: Clustering the genes

We will now cluster the genes in the Arabidopsis data set.
The genes in this data set were filtered with the following criteria:

1. Maximum fold change between at least one of the days versus day 8 is above 4 (Usually one could use a much lower fold change, for example 2, but here we have lots of genes, and for simplifying the exercise, we will work with a smaller data set).
2. Maximal count for the gene>20.

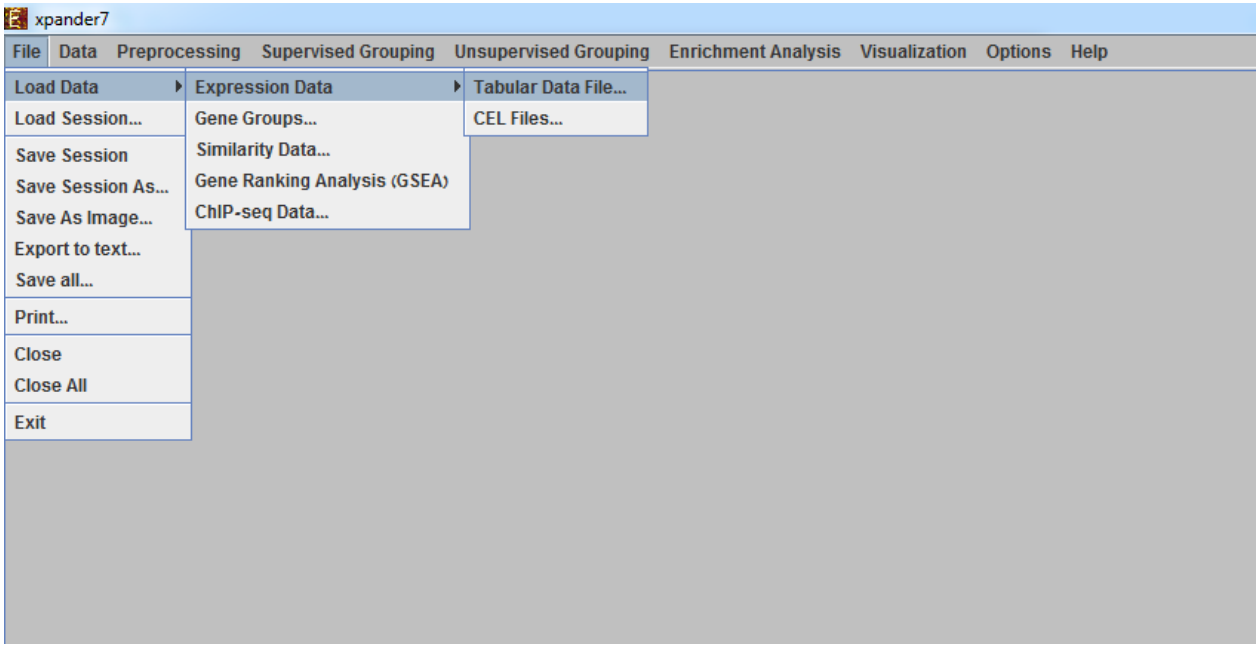
There are 2322 such genes. The DESeq2 normalized counts of these genes are in the file:

norm_counts_plus1_fold_abv_4_padj_max_count_20_for_clustering_genes.txt .

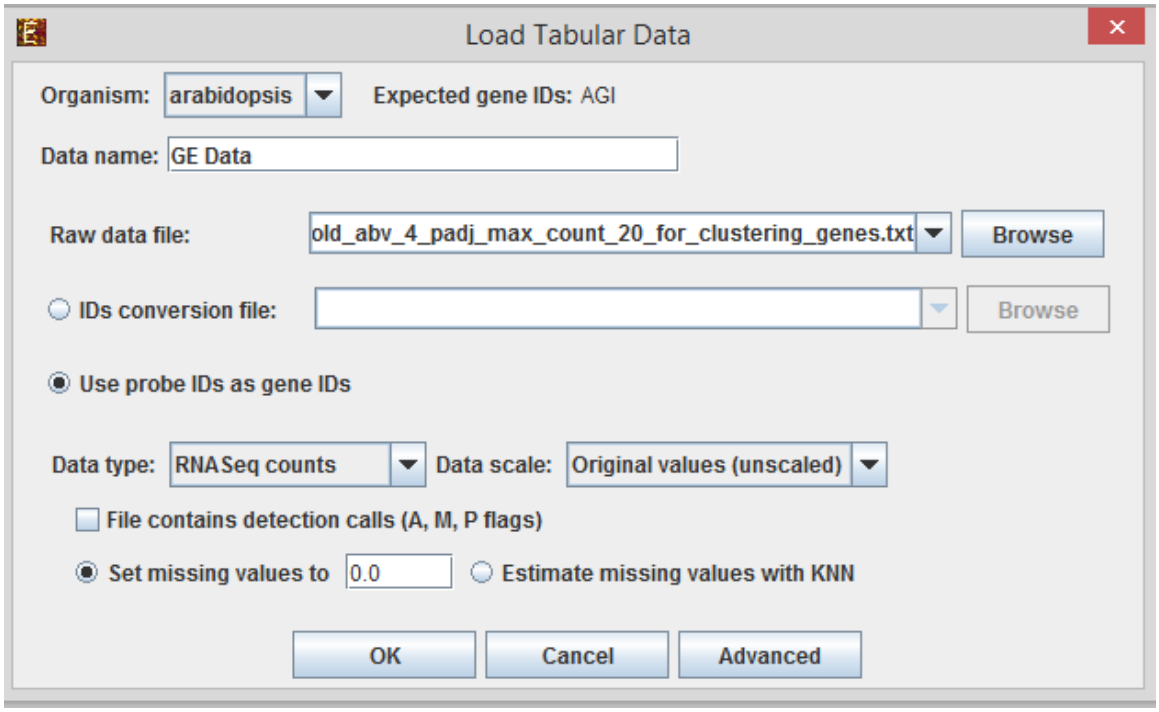
The data in the file are counts. 1 was added to all values, so we will be able to log transform the values. Remark: DESeq2 has transformations that can also be used for the clustering.

One must filter genes before the clustering. It is essential to filter the genes according to the question we are asking. In the current example, we chose genes that were significantly differentially expressed at least in one pair-wise comparison.

Import the 2322 genes to expander: Choose File -> Load data -> Expression Data -> Tabular Data File:



Fill the fields as following:



In the Organism field choose Arabidopsis (though if we are doing only clustering in Expander, the organism does not matter).

Choose the file:

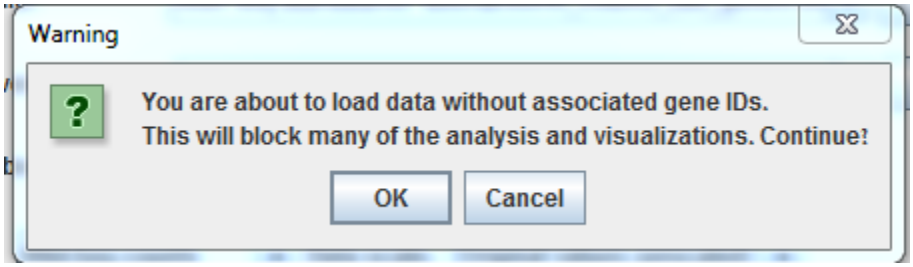
data/norm_counts_plus1_fold_abv_4_padj_max_count_20_for_clustering_genes.txt

in the “Raw data file” field.

Data type is “RNA Seq counts”.

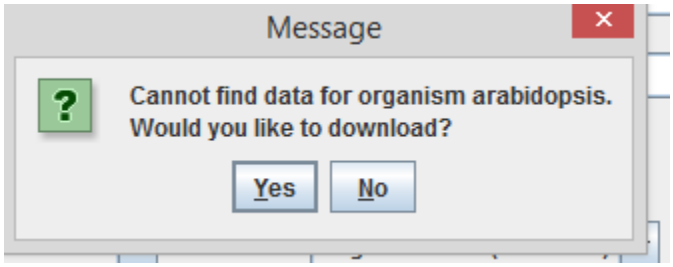
The Data scale is: “Original values (unscaled)”.

You will get the following warning:



Click the OK button (since we will do only clustering; In Expander there are other tools that needs further data on genes).

You will also get the following message:



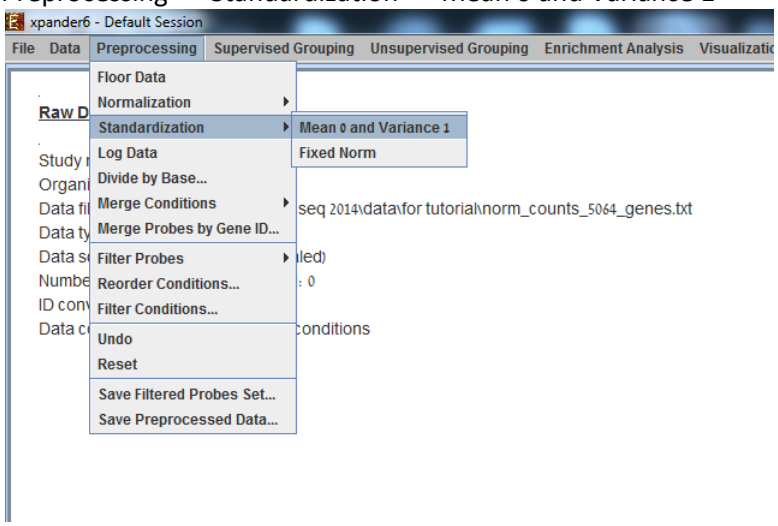
Click No

Since we will cluster the log counts, transform the counts to log2base by:

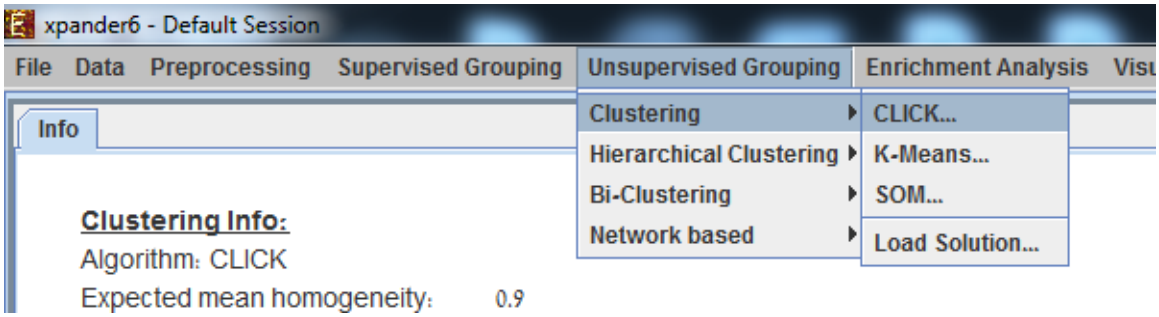
Preprocessing -> Log data

Standardize the data (so the mean of each gene will be 0 and standard deviation 1) by:

Preprocessing => Standardization => Mean 0 and Variance 1

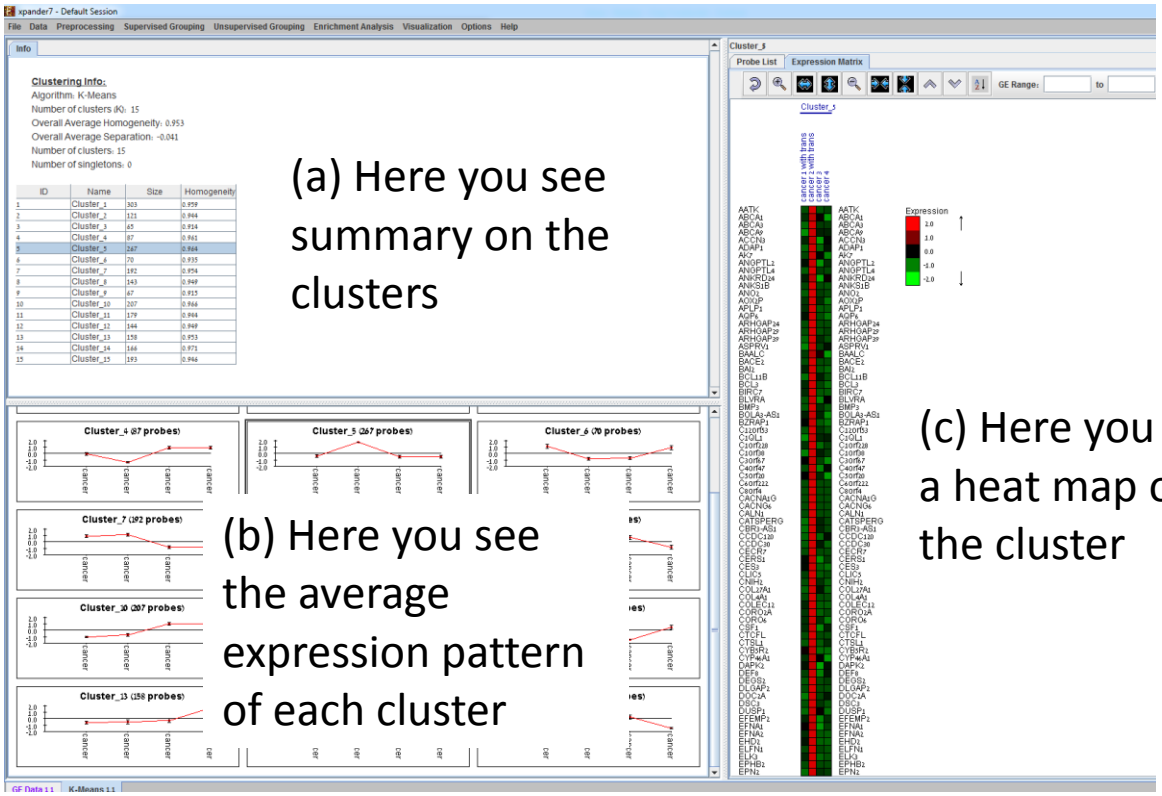


Cluster the data with the CLICK algorithm. The CLICK algorithm was developed in the lab of Prof. Ron Shamir. It utilizes graph theoretic and statistical techniques to identify groups of highly similar elements.
The advantage is that the user chooses homogeneity of the cluster; and does not need to choose the number of clusters.



Start with default homogeneity. The clustering takes some time…

You will get the following screen with panes (a) (b) and (c):



Look on the right pane (c) at the Expression Matrix tab.

When clicking on a cluster in pane (b) or in (a)–see image, you will see a heat map of the chosen cluster in pane (c). In pane c choose the Expression Matrix tab.

Q4. How many clusters did you get? Check in pane a how many singletons did you get? (singletons are single genes that were not classified to any of the clusters).

Look at the different patterns you got.

At the next step, one can study and get biological insight on each cluster. This is above the scope of the current exercise, and will be learned later in this course. Please save the output of this clustering, you will use it later on that course for pathway analysis.
To save:

File -> Export to text

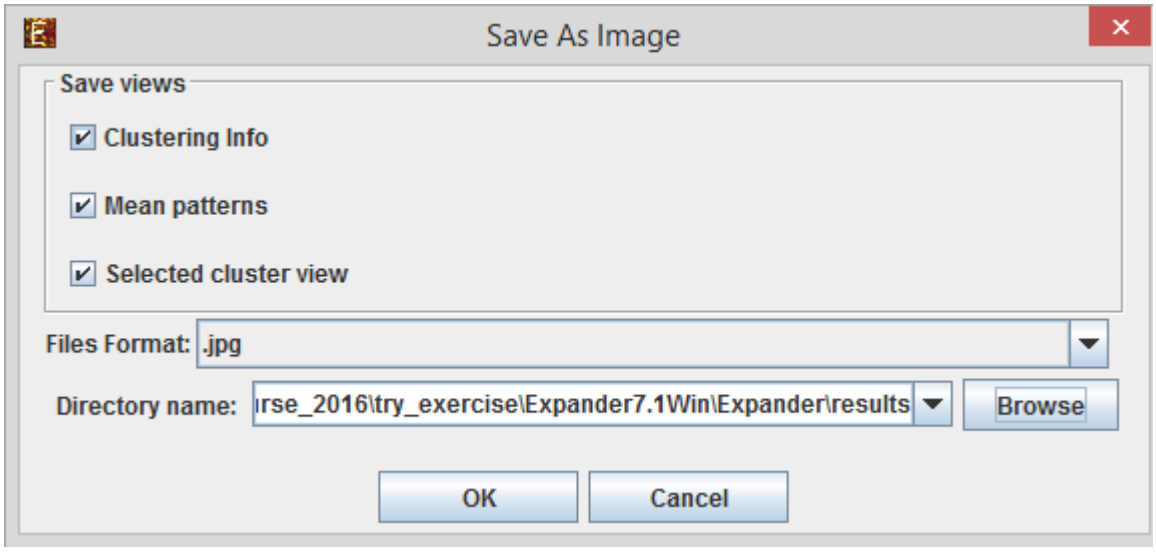
Save the file under the results folder with the name gene_cluster_click.txt

<u>You will use this file later in the course. Make sure you remember where you saved it.</u>

In this file you will have in the first column the gene ID, and in the second column the cluster number.

Save also the images:

File-> Save As Image



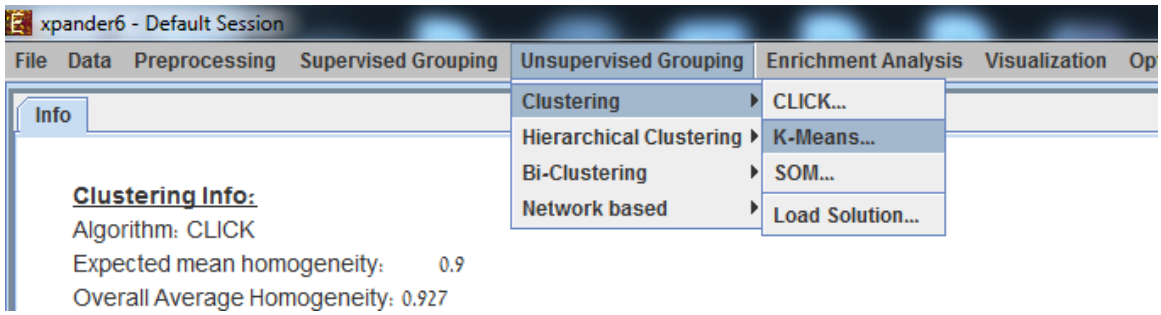
It is also possible to save the session:
File-> Save Session

This will allow you to open the session in the future.

You can try to cluster the data with CLICK, with a higher homogeneity value.

For each clustering trial you perform you will get a different tab on the bottom of the screen. You can go back to previous clustering you did.

Now try to cluster the data with the K-means algorithm:



You need to choose the number of clusters.

---------- **THE END** ----------