June 2015

Hands-on for the lecture:

RNA-seq gene level differential expression and clustering

Gilgi Friedlander
The Nancy and Stephen Grand National Center for Personalized Medicine

Exercise for the lecture: RNA-seq gene level differential expression and clustering

In this exercise you will work on the RNA-seq data of Leukemia cell lines.

This experiment includes 4 samples (human):

1. C1 - Leukemia cell line 1 with a specific translocation
2. C2 - Leukemia cell line 2 with a specific translocation
3. C3 - Leukemia cell line 3 (without the translocation)
4. C4 - Leukemia cell line 4 (without the translocation)

RNA was isolated from each cell line, libraries were prepared, and sequences were obtained from the Illumina HiSeq machine. Illumina's pipeline generated the fastq files. The data was analyzed by Dr. Dena Leshkowitz using the following workflow:

The fastq files were mapped to the genome using TopHat
HTSeq count was used for counting reads on genes
The DESeq package was used for normalizing the data and for differential expression analysis

Today you will cluster the genes in this data.

Today, 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.

For the purpose of the exercise, download Expander from:
\\ngs001\Open_Data\Course2015-exercise1\RNA_seq_cluster\Expander

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. Go to the following directory:

\\ngs001\Open_Data\Course2015-exercise1\RNA_seq_cluster\data\

And download the following two files to the data directory:

norm_counts_288_genes.txt
norm_counts_2362_genes.txt

Create also a directory named results.

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

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

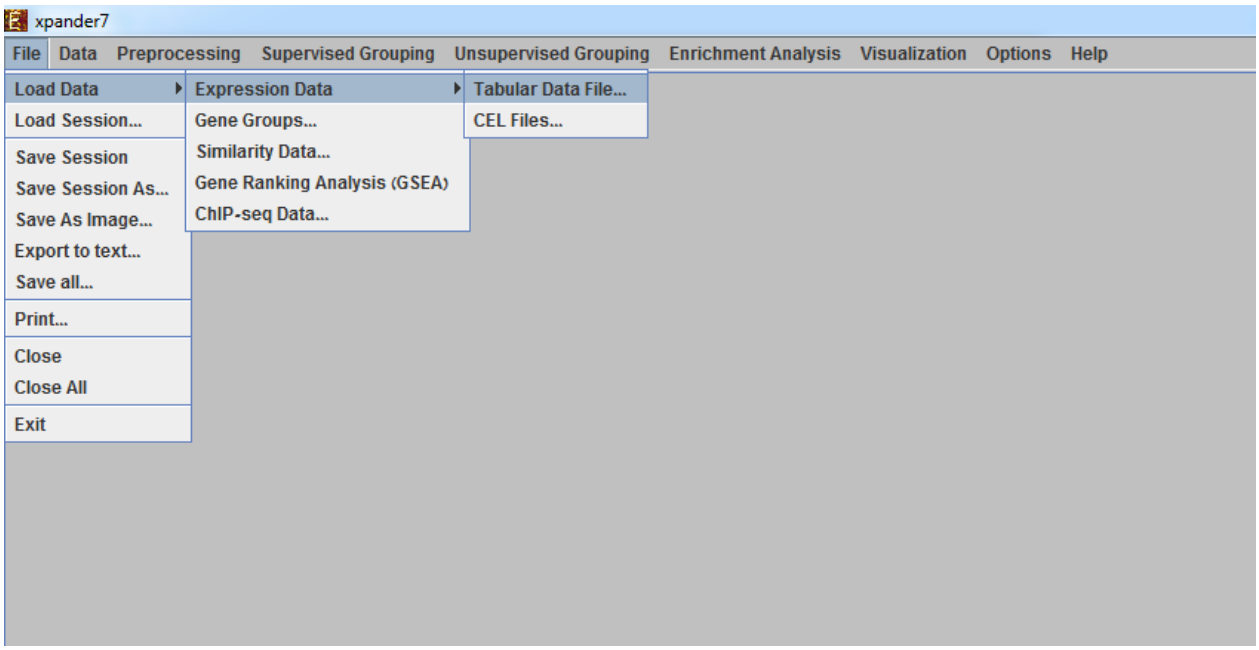
We will now cluster the data in the Leukemia data set.
The genes in this data set were filtered with the following criteria:

1. Maximum fold change between any samples is above 8 (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>200 (here also, one could choose a lower threshold).

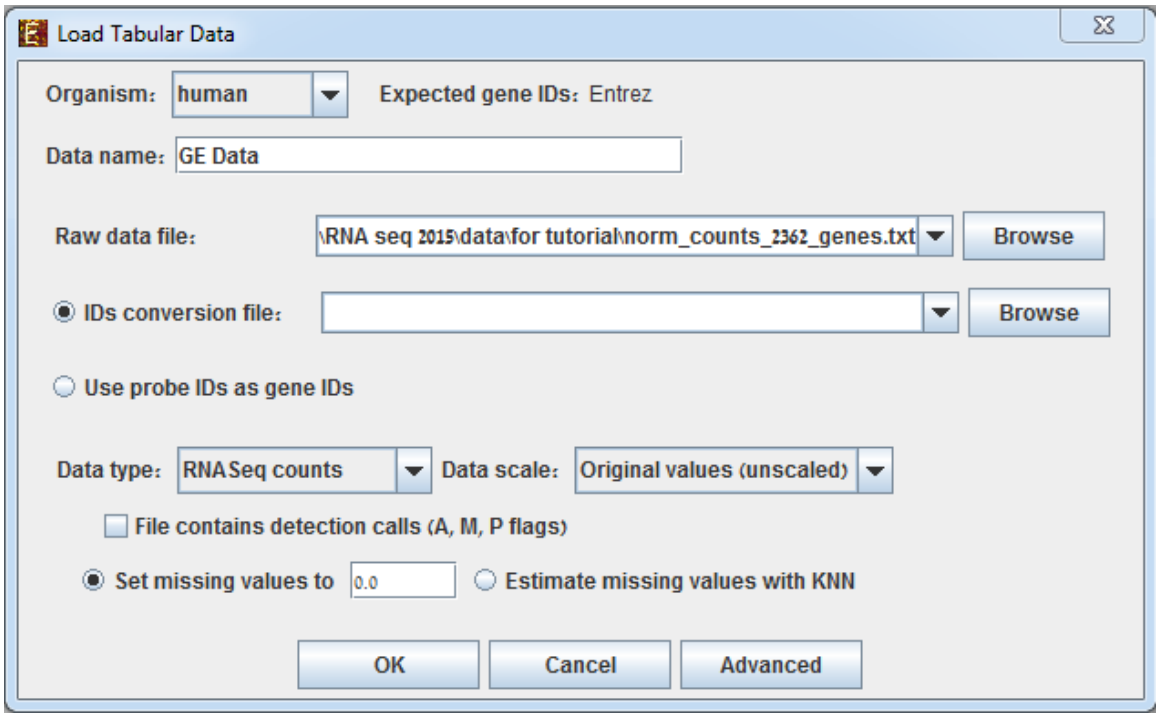
There are 2362 such genes. The DESeq normalized counts of these genes are in the file:

norm_counts_2362_genes.txt (you saved this file under the Expander/data director).

Import the 2362 genes to expander: Choose File -> Load data -> Expression Data -> Tabular Data File:



Fill the fields as following:



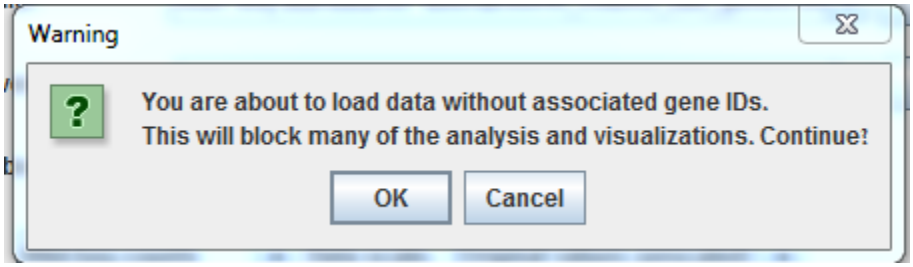
Organism is human (though if we are doing only clustering in Expander, the organism does not matter).

Choose the file data/norm_counts_2362_genes.txt in the “Raw data file”.

Data type is “RNA Seq counts”.

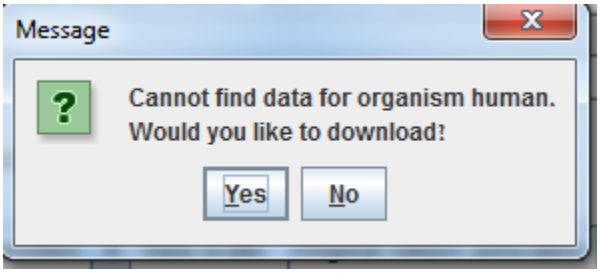
The Data scale is: “Original values (unscaled)”.

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:



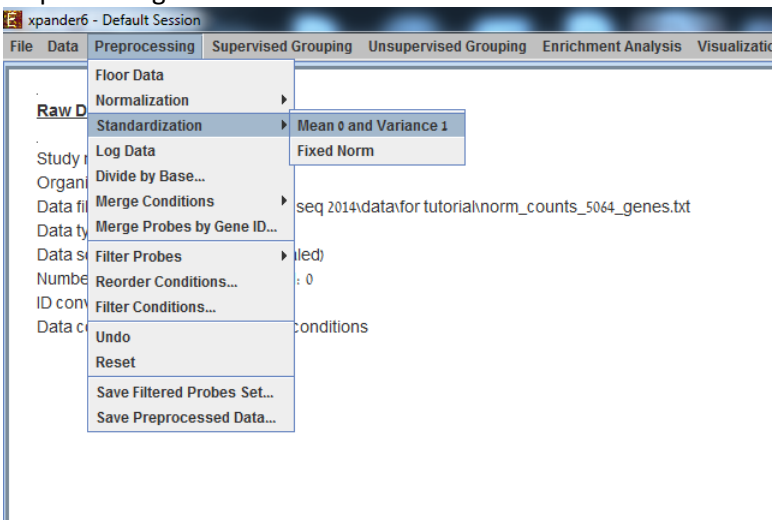
Click No

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

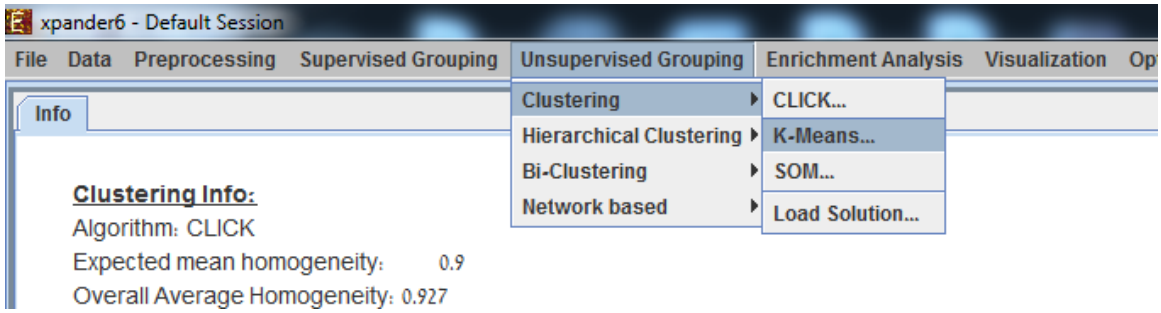
Preprocessing -> Log data

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

Preprocessing => Standardization => Mean 0 and Variance 1

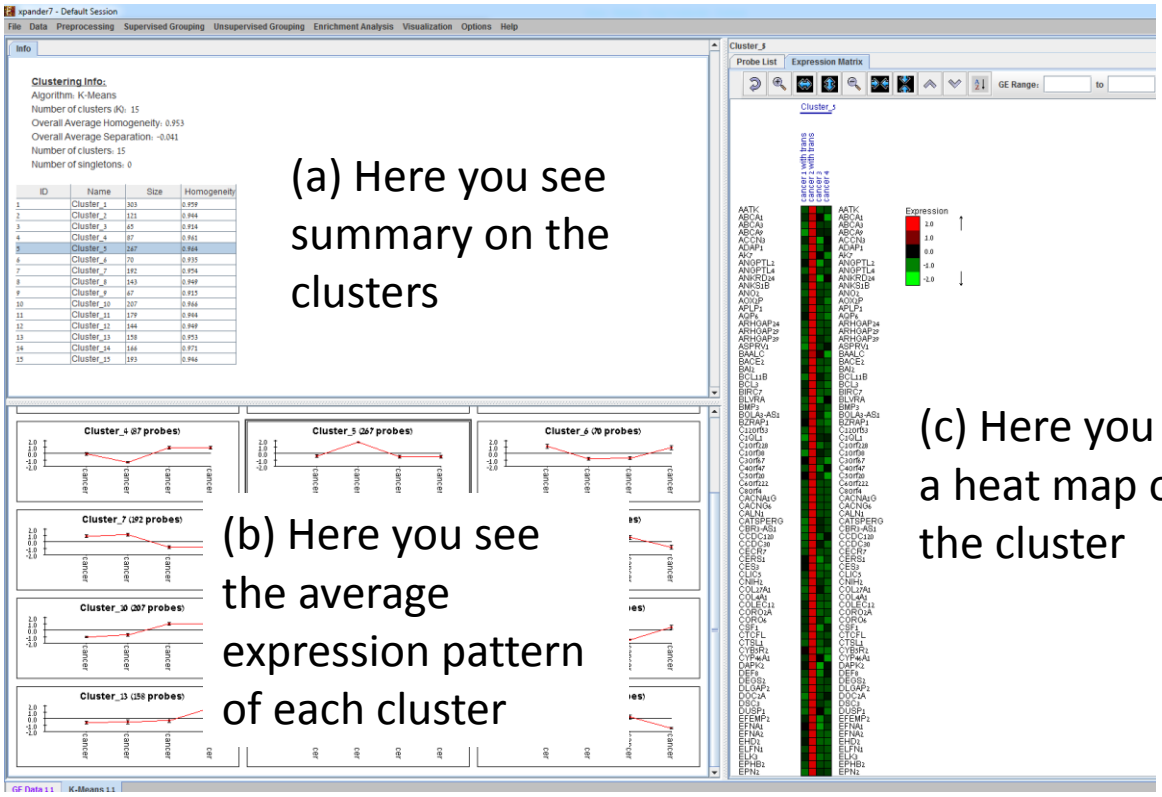


Now try to cluster the data with the K-means algorithm:



Try to cluster the data into 15 clusters

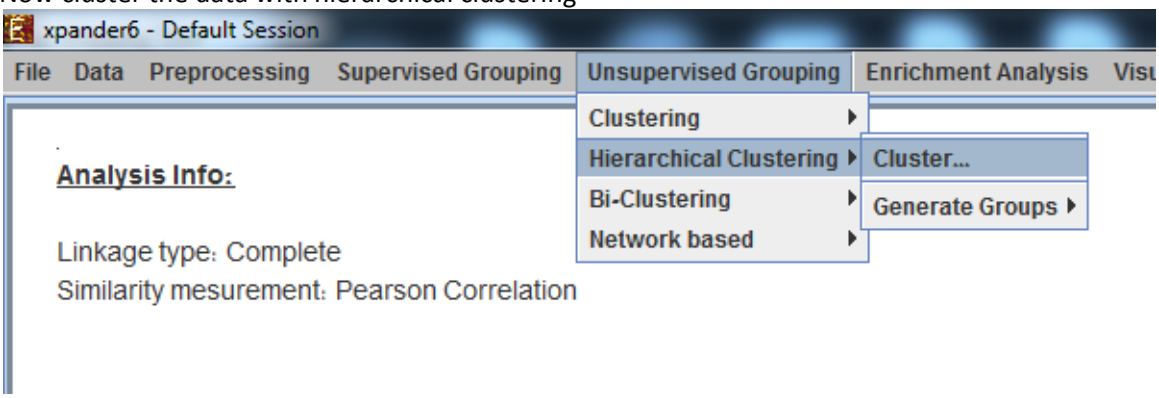
You will get the following screen with panes (a) (b) and (c):



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.

Look at the different patterns you got.

Now cluster the data with hierarchical clustering



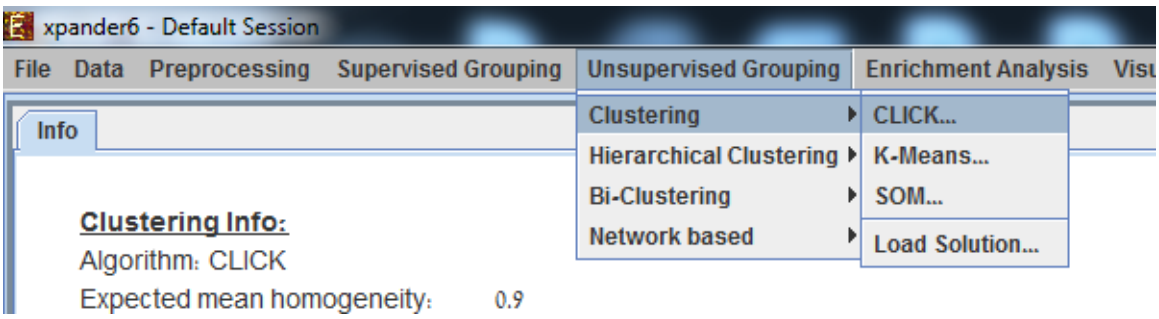
Choose Complete for linkage.

Check the Probes and uncheck the Conditions (we will cluster only on the genes).

[The linkage function specifies how the distance between two clusters will be calculated. In complete-linkage clustering, the link between two clusters contains all element pairs, and the distance between clusters equals the distance between those two elements (one in each cluster) that are farthest away from each other.]

Look on the right pane (c) on the different patterns. 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 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.

How many clusters did you get?

Now try to cluster the data with CLICK, with a higher homogeneity value: 0.9.

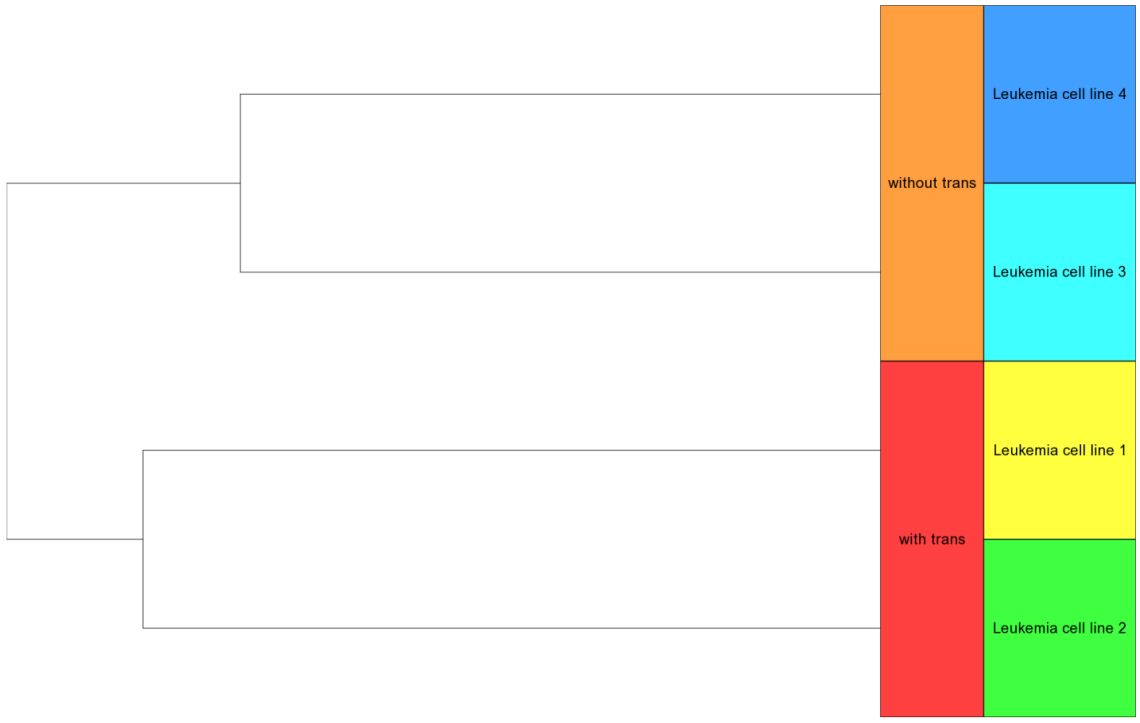
How many clusters did you get now? Do they look better (more homogenous)?

Can you find clusters in which the expression pattern is similar in the two cell lines with the translocations, and similar in the 2 cell lines without the translocation?

If yes – which ones?

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.

Hierarchical clustering on the samples looks like this:



This clustering was done on all genes, using Pearson correlation as distance measure.

We can ask the following question: which genes behave the same in each set, and are differentially expressed between the 2 sets. In order to do this, DESeq was run on these samples: the 2 samples with the translocation were compared to the 2 samples without the translocation.

After DESeq we get the average fold change, a p-value and an adjusted p-value for multiple comparisons.

In order to find such genes, that are differentially expressed in the cells with the translocation versus the without translocation, the data-set was filtered in the following way:

1. Genes with average log2ratio (with translocation) versus (without translocation) above 1 or below -1 (means fold change is above 2)
2. Adjusted p-value<0.05
3. Maximal count for that gene was above 10

This leaves us with 288 genes. You will find the normalized gene counts of these genes in the file:

norm_counts_288_genes.txt

What dominant patterns would you expect to see here?

(Hint: remember how we filtered the list: average fold change above 1 or below -1 and adj p-vale<0.05)

Let's look at the heat map of the expression pattern of these genes:

Choose File -> Load data -> Expression Data -> Tabular Data File:

Choose the file norm_counts_288_genes.txt

Take log of the counts:

Pre-processing -> Log Data

Now standardize the data:

Preprocessing => Standardization => Mean 0 and Variance 1

Cluster the data with CLICK :

Unsupervised Grouping => Clustering => Click

Choose default homogeneity.

Look at the patterns of the clusters you got, and make sure you understand the difference between the current filtering and the previous one.
Before, we didn't manage to get such clean groups. This example demonstrates the importance of the filtering of the genes before the clustering. It is essential to filter the genes according to the question we are asking.
---------- **THE END** ----------