Cluster software and Java TreeView

To download the software:
http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm
http://bonsai.hgc.jp/~mdehoon/software/cluster/manual/TreeView.html

Cluster 3.0 Manual:
http://bonsai.hgc.jp/~mdehoon/software/cluster/manual/

Introduction
Cluster 3.0 Software takes gene expression matrices as input and looks for similarities in gene expression patterns across a set of samples, putting together the genes with most similar patterns and building a similarity tree to show relationships across the gene set. Cluster is available as a GUI that is simple to use, and which can do quite a lot of sophisticated calculations quite rapidly for a small geneset (on the order of hundreds or thousands of genes). The results can be displayed in programs like Java TreeView, in the form of "Heat Maps". Both Cluster and Java Treeview are freeware and are very easy to install and use. Cluster can also be run quite simply on the command line to handle larger datasets, but these take a lot more time. Larger datasets are also difficult to view meaningfully.

The correlations can be conducted using various types of statistics that you can choose from a menu, although the choice of statistical program is not simple and requires some dedicated study. You can also choose to analyze and display gene patterns after log2 transformation, and "centering", which simply shows the relative up-down pattern for a particular gene in shades of red and green (or other colors if you choose). Cluster will carry out the log2 transformation and centering for you if you choose the options.

A typical use of Cluster is to create a simple heat-map of differentially expressed genes for simple viewing of similar gene groups in a dataset. For this purpose, most people use "default" settings: Pearson correlation with Average Linkage. Cluster offers many other options which can yield slightly different results (mostly in terms of the tree that describes similarity between, rather than within, groups of genes). For more detail please see the cluster 3.0 manual.

1. Cluster requires an input "expression matrix" text file. This file will be of the type:

ID	Expt1	Expt2	Expt3	Ctrl1	Ctr2	Ctrl3
Gene1	v1*	v2	v3	v4	v5	v6
Gene2	v7	v8	v9	v10	v11	v12
Gene3	v13	v14	v15	v16	v17	v18
.	.	.	.	.	.	.
.	.	.	.	.	.	.
GeneN .	.	.	.	.	.	vN

*v=Expression values (for RNAseq will be cpm or fpkm) in each sample.

More samples, e.g individual experimental and control samples including especially different types of conditions (e.g. timepoints etc) means that the correlations will be much more robust and trustworthy. However, it is possible to generate a heat map for small numbers of samples that can still be useful as a way to group genes for display.

1 **Using Cluster**

For this exercise I have created an expression matrix file for you to use: "Knox_SeriesMatrix-DEGs_noprobesets.txt", on the server. The data are from a paper in which duplicate or triplicate RNA samples from mouse maternal placenta (MP) or Fetal Placenta (FP) were examined in microarrays, over a period of development between embryonic day 8.5 (E8.5) and birth (postnatal day 0, or P0). (Knox and Baker, 2008; Genome Research 18: 695-705.

I analyzed expression data between E12.5 and E17.5 from FP, and identified all differentially expressed genes (DEGs). Then I extracted just the DEGs from the expression matrix in GEO to create the file. Below, I have outlined how I created the input file from any whole-genome expression matrix (arrays or RNAseq), using a Galaxy-based method that you can follow if you are interested.

Note that this same matrix file, with a few formatting modifications, is the file you can use for the Cytoscape Network Correlation exercise.

Cluster is easy to use. Once you open the program a menu at the top of the page will give you direct access to the help files and the (very hefty) cluster manual. There you can find out more for any of the options you can use in this program.

A. Open the GUI and in the file menu, select "**Open Data**", which will allow you to select the matrix file from your computer. The file should be in tab-delimited text format, which you can save from excel or similar formats.

B. **Filter Data.** When the file is loaded, you will see several buttons that can be selected starting with "Filter Data". You do not have to use this filter but can, e.g. cut off genes that *do not* have expression values of at least 1 cpm for at least 1 of the samples (this is commonly done for many gene expression analysis programs because genes with very low expression contribute the most noise). You can also filter out genes with very high or very low standard deviations (SD) or select out the genes that show max and min values that are below some cutoff value (would be one way to get rid of genes that are not significantly differentially expressed). Or you can do this kind of filtering before you load the expression matrix file (which is what most people do).

   Each time you want to execute a new filter you should select "Apply" then go on to the next filter. Cluster gives you the operations in their preferred order of execution so proceed from top to bottom.
   Once you have selected and applied filters, you will need to reload the dataset if you want to go back and change them. Your original dataset remains unmodified, but the version stored in Cluster has changed.

C. **Adjust data:** If your values are not already log transformed you can do it here (*note that the dataset I provided is already converted to log2 values, it is not a good idea to calculate the log again).*

Here you can also choose to "**Center genes**". This will allow the display to show, for each gene, a pattern of up- or down-regulation across the samples, based on its mean or medium expression value. As described in detail in the Cluster manual **Median** is the most robust form of centering and should generally be used.

D. **If your dataset is not already normalized you can do it here.** The dataset I gave you is already normalized. Genes are also already normalized when they come off of the Cuffdiff or HTSEQ count programs (cpm or fpkm).

Note that Cluster also allows for "arrays" (experiments) to be clustered, e.g. if you wanted to analyze a bunch of individual experiments and see how similar they are to each other. So there are choices for this kind of analysis too.

E. For this exercise, which is gene-centric, we will center the genes. First leave log transform and normalize unchecked, but check "Center genes" and "Median".

Now select the tab" Heirarchical".

a. Under Genes select "Cluster" and Correlation "centered". Then in buttons below, select "Average linkage". This should produce several files.
b. The .cdt file is the one to be opened in Treeview. In the Java TreeView menu, select "File" and "Open". TreeView will display your clustering results as heatmap.
c. Once in TreeView you can select a particular clade of genes (group with similar expression) and get the gene names as a list by selecting 'Export->save list" in the menu. You can also save the image, or subimages (selected clades), change colors, and many other things.
d. Now repeat the process, but first, change the name of the Job by adding "log2" at the end (if you don't do this you will overwrite the first file). Go back to "adjust data" and select "log transform data", then apply. Go to "hierarchical" again and select Cluster/correlation(centered), then average linkage. This shows the effect of log2 transformation. This display is not very helpful for this particular dataset, because the gene expression differences across timepoints are not very high. The log2 transform gives you better statistics and smooths out noise, but it also dampens the signal.
e. Explore the different display functions in Treeview and Download the two cluster images as .png

Appendix: Preparing matrix files for Cluster analysis

Cluster works best when you do not input too many genes at a time; its very hard to use any tool to visualize all >20K human genes (or all ~14K Drosophila genes for that matter. So it's a good idea to go into these programs with a carefully selected dataset.

For this exercise, we will first select a superset of all genes that are significantly differentially expressed in either the fetal placenta (FP) or the maternal placenta (MP) E12 and E17 datasets. To make this easy we will **select the top 500 genes from each dataset** (ranked by p-value).

1. From the excel files you prepared above, copy the names of **probesets (column 1 of the GEO2R output)** of the top 500 genes from each set into a single text editor .txt file (a good text editor for this purpose that works in both PC and MAC is Text Wrangler). Save the file as "probesets.txt" making sure the file is saved with Unix-compatible "line breaks" and "encoding" (the default for Text Wrangler).
2. Open your terminal on your computer.
   #First you need to get to the same directory that your genes.txt file is located in. On my computer, I have stored the file on the desktop. So I need to "change directory" (cd) to the desktop. At the prompt, type, and then hit return:

   cd desktop

   # now we are going to employ a simple unix command set that will generate a list of unique genes from this set (the set contains many redundant gene entries).

   cat genes.txt | sort | uniq >genes_unique.txt

   # this set of "piped" commands (the symbol | is the "pipe") has called up your genes file, sorted the gene names, and then identified a unique list of gene names from that redundant list. The > command tells your computer to generate an output file called probesets_unique.txt.
   # take a look at your new file; it should have genes sorted alphabetically and each gene name should be listed only once.

   How many unique genes are present in this unique list?
3. Now, you want to use this unique list of genes to extract expression profiles for the DEGs across all tissues in the Knox/Baker expression experiments.
   a. First you need to download the "expression matrix" file of all genes and all tissues from GEO. Every gene expression experiment should have such a file available for download. Go back to the "Accession display" page you started GEO2R from. If you scroll down to the "Download family" section, you will see a link to the "Series Matrix file". To use this file for our purposes I had to do a lot of manipulations, like extracting the headers and getting gene names to correspond to the Affymetrix IDs. I also calculated averages over the expression values for each gene and each set of duplicates for each sample

(each tissue and timepoint). Rather than take you through all that I have uploaded a processed matrix file called "Knox_series_matrix.txt" to the wiki. **Download this file, then upload into your Galaxy instance.**

b. **Next, you want to use the "Join subtract and group/join two datasets" function** to extract the rows from the whole gene set that correspond to your unique set of genes. Galaxy needs both files to be in the same format, so easiest thing to do is to add a column to the probeset list and make it become a "tabular" file. Open the probesets_unique.txt file in excel and add a column with ... in every row. Save as text.
   Join the two files, using Probesets_unique.txt as file 1 and the gene expression matrix as file 2. **You will be joining on probeset names** because these are unique identifiers (unlike the gene names) so you need to figure out which column to join the two files on. Select "yes" for keeping lines and … to fill empty cells.

c. **Clean up the file for Cluster and Cytoscape use.** Rename the resulting Galaxy file as 'SeriesMatrix_Knox_DEGS.txt". Open as xls
   i. You will need to copy the sample names from the first row of the original "matrix" file to label your new rows according to sample type and developmental timepoint.
   ii. You also now can trim the first two rows from this file (Leftover from Galaxy). Sort the file on gene name and eliminate probesets that do not have an associated name.

To actually use these data in cluster and Cytoscape, you want to switch over to the file I have made with that same name, which is on the wiki. Affymetrix arrays had the irritating property of having multiple probesets for each gene name, which cluster and cytoscape do not like. I have trimmed the file to select the best probesets for each gene. You can ask me how I did this, but its not fun and most of the time in future your will be working with RNAseq data where you will not have this issue.