Profiler Tutorial

Dvir Netanely, June 2006

Eytan Domany's group, WIS

1. **Load data**
    a. **From a Matlab DGS file** - Use the "Load data…" button on the upper left corner (or use Import->'Load data from Matlab file'). For this tutorial you may load the file 'ALL_D2ESC_data_with_absents' which is a Matlab file in a DGS format (containing 3 variables: Data, Genes and Samples).
    b. **From a text file** – Tab delimited file containing one row header (sample names) and one column header (probe-set names).
2. **Remove 'All-Absent' probe-sets** – After constructing a tab-delimited text file containing detection calls (A/P), use Data->'Remove all absent genes based on A/P file'. This file should not contain any headers (SLOW).
3. **Process samples** – Remove samples or change their label using the Tools->'process sample labels menu'
    a. **Remove unwanted samples:** Remove the 3 ESC samples
    b. **Change sample labels:** Sample label prefix is used to group samples in various analyses. Unite the fatInduction and FatControl samples to one group names 'Fat'. Do the same for the various Bone samples.
    c. **Reorder the samples if needed.** Move the MSC samples to the left. Samples can also be reordered by applying hierarchical clustering on the samples by using Tools->'cluster samples' (this clustering operation will use the parameters specified in the hierarchical clustering panel on the screen).
4. **Examine data properties using common graphs**

a. Use Tools->**'Chip value distribution'** to display histogram of dataset values.
b. Use Tools->'Gene **variability plot'** to examine a histogram of gene variability.
c. Use Tools->'Gene level **sample comparison'** to compare two sample groups on the array level.
d. Use Export->'Display cluster gene **expression matrix'** to display the current expression matrix next to a color bar.
e. You may **sort the probesets** by their expression on a given samples by specifying column number and pressing the 'Sort by col.' Button on the left. Enter more than column number to sort by the averaged expression on these columns.

5. Data preprocessing

a. Use Data->'Set lower threshold' to set a lower threshold of 32.
b. Apply log2 using Data->'Apply log2 transformation'.

6. Filter out genes in order to reduce dataset dimensionality

a. Already removed all absent genes?
b. **Keep only top 12000 variable probe-sets** by clicking on the 'Top var. filter' button on the left after specifying the number of probe-sets in the text-box under it.
c. **Keep only probe-sets whose variance is above a certain threshold** – First examine the variability plot mentioned in section 4b above. The use Data-'Remove low variability genes' to specify variance threshold. The given default value represents the variance corresponding to the 1st percentile.
d. **Use ANOVA to keep only genes whose inter-group variance is higher than their inner group variance**. Define FDR alpha in the textbox (default is 5%) and press the ANOVA button on the left.

e. Use other **supervised tests** to keep only genes that significantly change between sample groups (t-test, rank-sum, fold change). Tools->'Conduct supervised tests'.

f. Keep only probe-sets specified by an **external probe-sets list**. Choose the Import->'Import probe-set list as current cluster'. This option assumed that the probe-set list in the text file is a subset of the loaded dataset probesets.

g. Manually select probe-sets subset by **dragging a box with the mouse over the expression matrix**. To ensure that all subsequent operation will be conducted on the zoomed in probe-sets use Data->'Make current cluster prime data'. To go back to the previous expression matrix click the 'Back' button on the left. To show all probe-sets in the full expression matrix stored in the program's memory, choose View-'Show all genes'.

7. **Normalize the rows if relative expression patterns are what you want.**
8. **Cluster the data**
    a. **Hierarchical clustering –** Simple, implements many distance and linkage functions, limited to ~8000 probesets.
    b. **SPC clustering**
    c. **Correlation clustering –** Detects clusters including probe-sets whose expression is inverse. Applicable to up to 1500~ probe-sets.
    d. **Profiling –** Applicable to more than 10,000 ps.
9. **Load annotations to memory –** On the 'Gene annotation analysis' on the lower right corner of the screen, choose the microarray version your dataset is using from the upper list-box. This will load all annotations into memory.
10. **Look for genes on the clustered dataset –** Enter a search string in the textbox on the left and press 'Gene lookup' or 'Title lookup' to search for the specified string in the gene symbol or gene title respectively. If a

match is found, the results will specify all the clusters in which the probe-sets appear.

11. **For each cluster you can now ….**
    a. **Display annotation enrichment analysis on screen-** On the lower list-box in the 'Gene annotation analysis' choose an annotation class (for example – 'Biological function') to display a GO enrichment analysis results for the currently displayed probe-sets in the onscreen expression matrix. Marking the 'Gene based' check box will make sure the analysis is not using more than one probe-set per gene. In the opened chart, pressing a bar will open a figure specifying all probe-sets that belong to the selected GO class.
    b. **Display cluster probe-set lists on the screen** (Export->'Display cluster PS, GS, Title).
    c. **Save the current cluster's annotated probe-set list to a tab-delimited text file – Probeset ID, Gene symbol and Gene title** (Export->'Save annotated cluster probesets to a text file')
    d. **Save the current cluster's data to a Matlab DGS file** (Export->'Save cluster data as Matlab file')
    e. **Export the current cluster full information to an HTML file (including data, probe-set annotations and links to external databases)** (Export -> 'Export current 'official' cluster to HTML).
12. **Export ALL clusters to HTML** by using Export->'Export all clusters to HTML'
13. Or even better – **FILTER the clusters first, based on their profile using** Tools->'filter clusters'.
    a. Clusters produced by all clustering methods can be filtered. After every clustering operation, profiles are automatically calculated for each cluster using the parameters specified in the 'Profile clustering section'. To recalculate only the post-clustering profile envelope

choose Tools->'recalculate profile envelope for last clustering' before opening the 'filter clusters' panel.

b. In the **'filter clusters' panel** you should define cluster size range, profile patterns (such as monotonically increasing...) and press the GO button. The filtered clusters appear in the list box. You can manually select subsets (default: all filtered clusters are selected).
c. You can then do 2 things with the selected filtered clusters:
    i. Use Export->'Export filtered clusters to HTML' to **export to HTML** only the filtered clusters.
    ii. Use View->'View filtered clusters' followed by Data->'Make current cluster prime data' to **keep as the main dataset only probe-sets that compose the selected filtered clusters**.

Good luck!

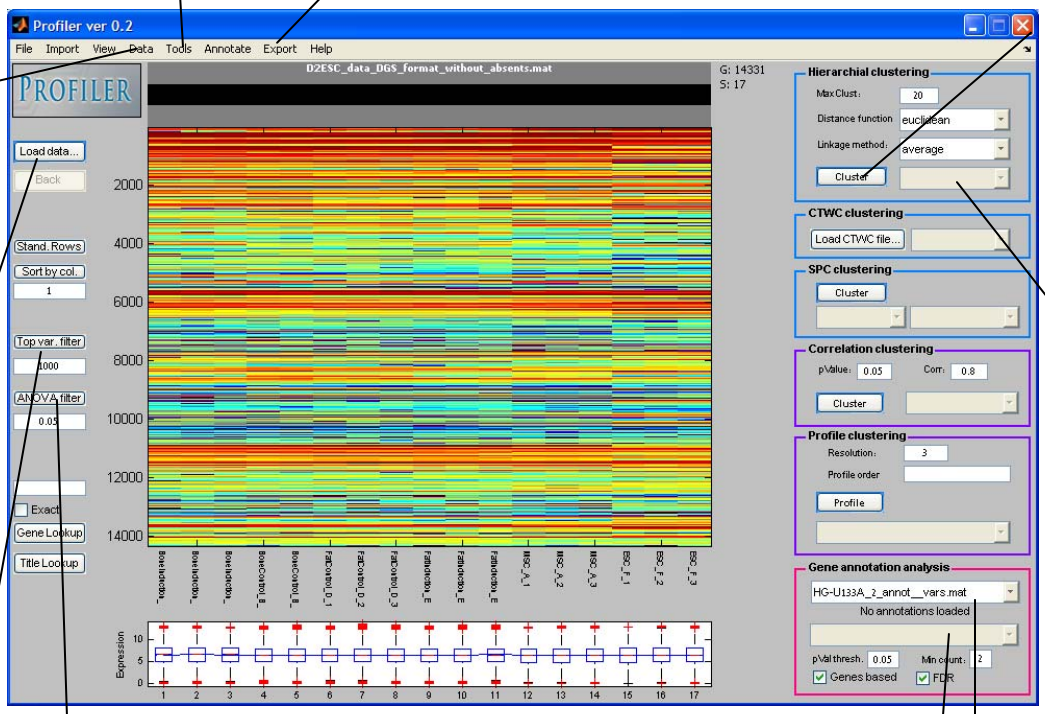
Quick start scheme

(8)
Filter clusters
Open Cluster-Filter (from the Tools menu, or by clicking Control+F) to filter detected clusters by profile or by size

(9)
Export clusters to HTML
By using the Export menu, or simply by clicking Control+H to export ALL clusters. Click Control+E to export only filtered clusters

(3)
Preprocess data
Remove all absent genes
Set thresholds
Log the data
Normalize Rows

(5)
Cluster the genes using any of the 5 methods



(6)
Choose a cluster to display it

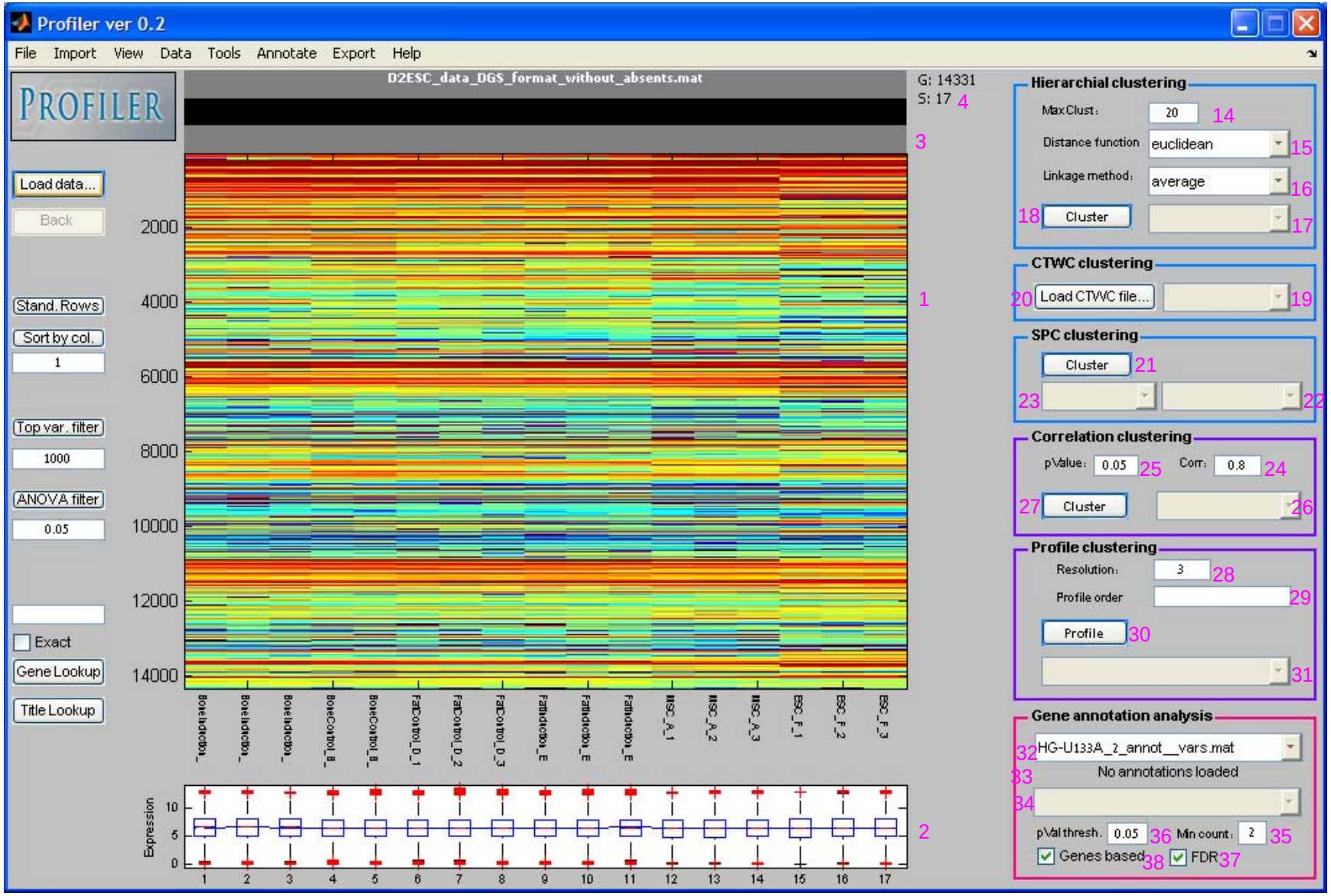
(1)
Load data (mat file)
Use the data menu to import from other formats

(4a)
Reduce number of genes (primitive filtering)
Keep only top X variable genes

(4b)
Reduce number of genes (supervised tests)
Keep only gene that vary between sample groups more than they vary within groups (ANOVA, samples groups are defined by prefix before underscore)
You can also conduct other supervised the test from the Tools->Conduct supervised

(7)
Check GO enrichment for displayed cluster

(2)
Choose chip type


Profiler ver 0.2
File Import View Data Tools Annotate Export Help
PROFILER
D2ESC_data_DGS_format_without_absents.mat
G: 14331
S: 17
Load data...
Back
Stand. Rows
Sort by col.
1
Top var. filter
1000
ANOVA filter
0.05
Exact
Gene Lookup
Title Lookup
Hierarchial clustering
Max Clust:
20
Distance function
euclidean
Linkage method:
average
Cluster
CTWC clustering
Load CTWC file...
SPC clustering
Cluster
Correlation clustering
pValue:
0.05
Corr:
0.8
Cluster
Profile clustering
Resolution:
3
Profile order
Profile
Gene annotation analysis
HG-U133A_2_annot__vars.mat
No annotations loaded
pVal thresh.
0.05
Min count:
2
Genes based
FDR
Expression

1. Main expression matrix:
    a. Rows represent probesets, columns represent samples.
    b. Color spans from 1st percentile to the 95th percentile of the displayed data.
    c. If less than 40 genes are displayed, probesets IDs displayed on the right.
    d. If annotation table was loaded, Gene symbols are also displayed.
2. Gene profile box plot.
    a. Displays the averaged expression of the cluster.
    b. Red line in the center represent sample expression median.
    c. Red crosses represent outliers.
    d. For more information, check out Matlab's box plot help.
3. Information bar
    a. First row – displays the file name from which the dataset was loaded
    b. Second row – Information about currently displayed probesets.
    c. Third row – Parameters used for clustering operation that generated the displayed cluster.

    The following figure shows the information bar after selecting cluster 4 (of size 120) of Hierarchical clustering.

4. Displayed dataset dimensions. G stands for probesets, S for samples.
5. Loads data from a Matlab file in DGS format.
6. Goes back to the previous cluster displayed.
7. Applies rows standardization (both centering and normalization).
8. Reorders the genes based on the column indices specified in the textbox below. If more than one column index is specified, genes will be sorted based on the average of specified columns.
9. Filters out low variability genes by keeping only the top X variable genes.
10. ANOVA filter
    a. Calculates one way ANOVA (Analysis Of Variance) on the data, followed by FDR correction, and keeps only the genes that passed the test.
    b. List-box determines FDR Q.
    c. ANOVA groups are determined based on the sample names. Samples are considered to be in the same group if they have the same prefix before the underscore (_). You can use the Tools->process sample labels command to change sample labels.
11. Check the checkbox to specify exact string lookup.
12. Looks for probesets by Gene symbol
    a. If the 'exact' check-box unchecked, approximated matches are also returned.
    b. If matches are found, expression of those genes will be displayed in addition to specification to the cluster in which each gene is found.
    c. Gene lookup should be applied only after loading data, loading annotations and portioning the data using any clustering method.
13. Same as above, searches the Gene Title list.

Hierarchical clustering

14. Specifies the maximal number of clusters to be found.
15. Specifies distance function to be used.
16. Specifies linkage method to be used.
17. Cluster selector
    a. Select cluster to display on screen
    b. Cluster size is shown in parenthesis.
    c. First cluster includes all probesets.
18. Initiates hierarchical clustering operation on loaded data.

Comments:

- Advantage - Simple and well established
- Advantage – you can use many types of distance and linkage methods.
- Disadvantage – you need to specify number of clusters beforehand.
- Can be applied on large datasets.

CTWC clustering

19. Cluster selector
    a. Select cluster to display on screen
    b. Cluster size is shown in parenthesis.
    c. First cluster includes all probesets.
20. Import a CTWC output file
    a. CTWC output file is named “matlab_workspace.mat” and it located inside the results folder of the CTWC working folder.
    b. By importing a CTWC output file you are both loading a new dataset, and its associated clustering structure.

SPC clustering

21. Initiates SPC clustering operation on loaded data.
22. Cluster selector
    a. Select cluster to display on screen
    b. Cluster size is shown in parenthesis.
    c. First cluster includes all probesets.
23. SPC temperature selector
    a. Each temperature level is associated with a set of clusters.
    b. On low temperature levels a small number of clusters is yielded (low resolution portioning of the data), whereas on high temperature levels, a large number of clusters is yielded.
    c. Default temperature is the highest temperature.

Comments:

- Advantage – Detects delicate signals.
- Advantage – returns a dendrogram spanning many temperatures.
- Advantage – No need to specify maximal cluster number.
- Can be applied on Medium-Large datasets.

Correlation clustering

24. Specifies correlation coefficient value (R) to be used as a lower threshold in correlation clustering (a probeset will be added to a cluster only if its correlation with this cluster’s founder is above this value).
25. Specifies p-value to be used as an upper value in correlation clustering (a probeset will be added to a cluster only if the p-value of the correlation with this cluster’s founder is below this value).
26. Cluster selector
    a. Select cluster to display on screen

b. Cluster size is shown in parenthesis.
c. First cluster includes all probesets.

27. Initiates correlation clustering operation on loaded data.

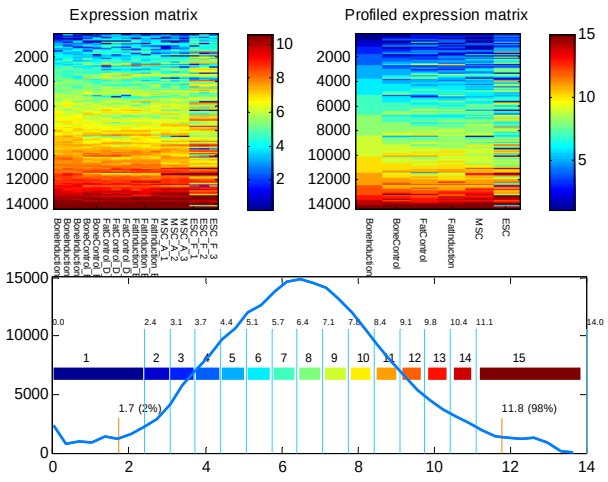
Comments:

- Advantage – Detects delicate signals.
- Advantage – Can detect inversely correlated genes.
- Can be applied on small datasets.

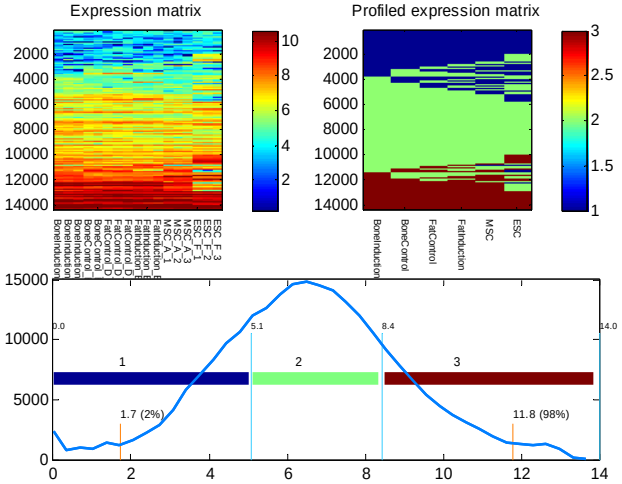
Profile clustering

28. Specifies profile resolution
    a. Profile resolution defines how many levels of expression are used in mapping expression values.
    b. Low values (2-3) will generate small number of profile clusters, whereas high values (>4) will generate many profile clusters, dissecting the information in a more delicate manner.
    c. Default is 3.
    d. Profile resolution is equivalent to fold change.

Resolution = 15



Resolution = 3



29. Determines sample group priority by which genes are ordered after the profile clustering is done. Leave blank to sort the profiles from left to right.
30. Initiates a profile clustering operation.
31. Cluster selector
    a. Select cluster to display on screen
    b. Cluster size is shown in parenthesis.
    c. Cluster profile is shown in brackets.
    d. First cluster includes all probesets.

Comments:

- Advantage – Intuitive and biologists friendly.
- Advantage – Extremely fast.

- Advantage - Can be applied on large datasets.
- Not useful for datasets with many sample groups.

Note that the Resolution parameter is also used when calculating post-clustering profiles for Clusters that were generated by any other clustering method.

Cluster annotation enrichment testing

32. Selecting an Affymetrix chip version will load the annotation table. Make sure to select a chip version that matches chip version of the loaded expression data.
33. Specifies last annotation table loaded.
34. Select an annotation class to display the enrichment significance results page for the displayed cluster.
35. Minimal number of probesets (or Genes If Genebased checkbox is checked) to be included in the cluster.
36. Maximal hyper-geometric function p-value threshold.
37. If checked, FDR (with Q=0.05) is applied on the detected annotations found (reduces number of yielded annotations).
38. If checked, calculation is conducted on the gene level, rather than on the probeset level (This method is more correct due to the fact that there are more than one probeset for each gene; usually reduces number of yielded annotations).

Comment:

- Click Ctrl+Z to annotate current cluster.
- When exporting to HTML, annotation enrichment significance is calculated using the parameters chosen in this frame.