SC^{2}ATmd v3: Tutorial

Last updated on 6/13/2012 by Amy Olex

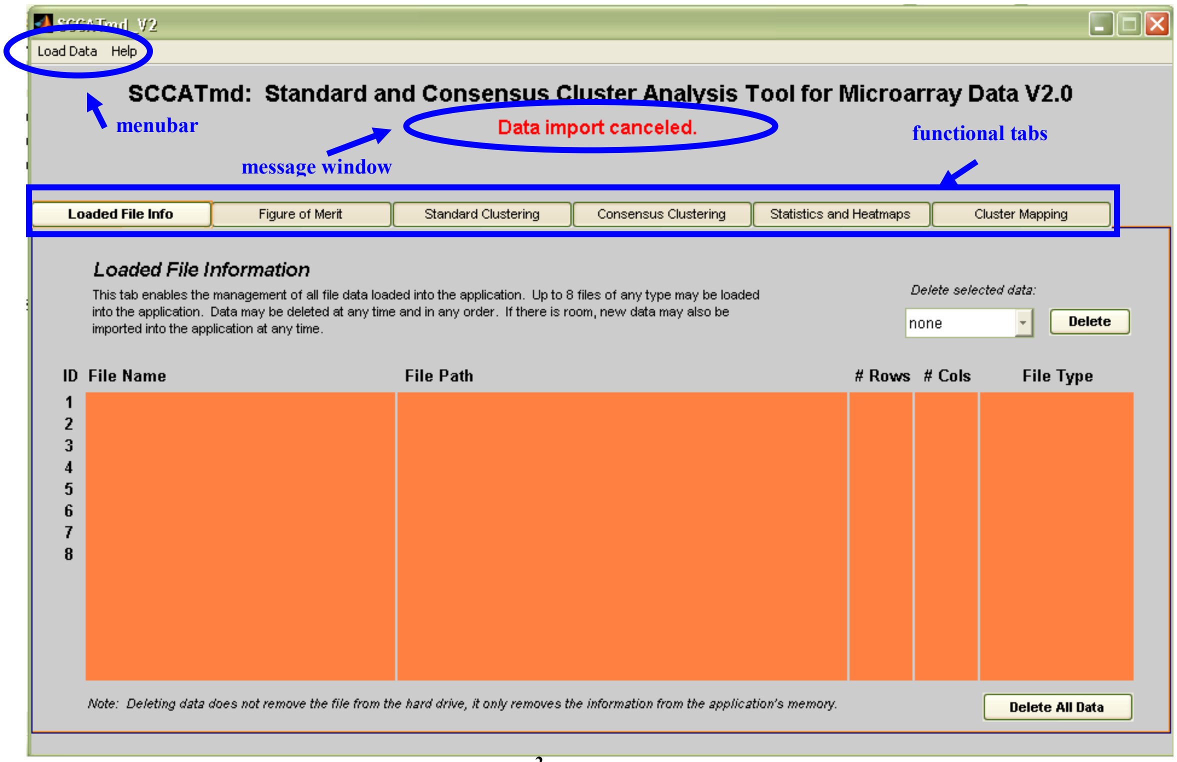
Index

This tutorial is written to walk the user through all the functions of SC^{2}ATmd. The example data files that are used are included with this distribution.

Interface Orientation

top

The SC^{2}ATmd interface in composed of 6 functional tabs, 2 menu bar options, and a message window. These components are identified in Figure 1 below.



SCCATmd: Standard and Consensus Cluster Analysis Tool for Microarray Data V2.0

Data import canceled.

Loaded File Information

This tab enables the management of all file data loaded into the application. Up to 8 files of any type may be loaded into the application. Data may be deleted at any time and in any order. If there is room, new data may also be imported into the application at any time.

ID	File Name	File Path	# Rows	# Cols	File Type
1					
2					
3					
4					
5					
6					
7					
8					

Note: Deleting data does not remove the file from the hard drive, it only removes the information from the application's memory.

Figure 1: SC^{2}ATmd user interface.

The toolbar is located at the very top left of the interface and includes data input functions and help files. Below the toolbar is the message box which initially does not contain anything. The message box will notify the user of the successful completion of a task, warning messages indicating improper input or the cancellation of tasks such as importing a file, and error messages. Down below the message box are 6 tabs; each one provides the user with a different service, and each will be discussed in detail next.

Tab: Loaded File Info

The Loaded File Info tab (shown in Figure 1) is active by default when SCCATmd is started. This tab allows the user to manage all the data that has been loaded into the application for analysis. Up to 8 data files of any type may be loaded into the application at any one time. As each file is loaded into the application's memory, its information (file name, size, format type, etc.) is displayed on the next available line on the Loaded File Info tab. Deletion of one or all files from the application's memory is also allowed. If a data file is deleted, it will only be removed from the applications memory, not the hard drive. This will free up space so that additional files may be imported.

Tab: Figure of Merit

The Figure of Merit tab provides functionality to perform a FOM or cFOM analysis on any dataset imported as a FOM/Clustering file format type (Input File Formats Help page). On-screen directions are provided on the left, and the analysis parameter selection is on the right. A screen-shot of the FOM tab is shown in Figure 2.

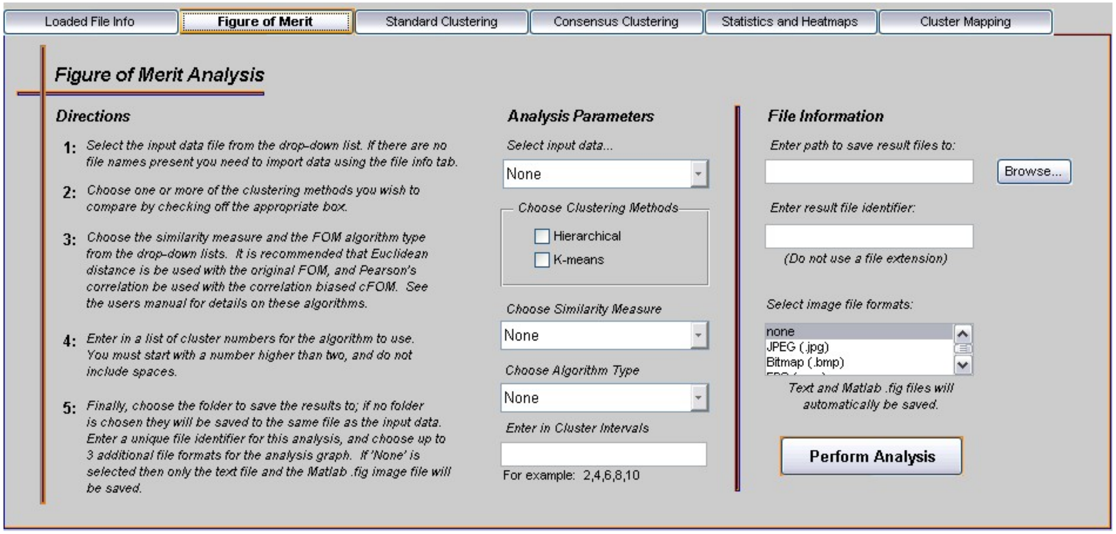


Figure of Merit Analysis

Directions

1: *Select the input data file from the drop-down list. If there are no file names present you need to import data using the file info tab.*

2: *Choose one or more of the clustering methods you wish to compare by checking off the appropriate box.*

3: *Choose the similarity measure and the FOM algorithm type from the drop-down lists. It is recommended that Euclidean distance is be used with the original FOM, and Pearson's correlation be used with the correlation biased cFOM. See the users manual for details on these algorithms.*

4: *Enter in a list of cluster numbers for the algorithm to use. You must start with a number higher than two, and do not include spaces.*

5: *Finally, choose the folder to save the results to; if no folder is chosen they will be saved to the same file as the input data. Enter a unique file identifier for this analysis, and choose up to 3 additional file formats for the analysis graph. If 'None' is selected then only the text file and the Matlab .fig image file will be saved.*

Analysis Parameters

Select input data... None

Choose Clustering Methods

□ Hierarchical

□ K-means

Choose Similarity Measure None

Choose Algorithm Type None

Enter in Cluster Intervals

For example: 2,4,6,8,10

File Information

Enter path to save result files to:

Enter result file identifier:

(Do not use a file extension)

Select image file formats:

none
JPEG (.jpg)
Bitmap (.bmp)

Text and Matlab .fig files will automatically be saved.

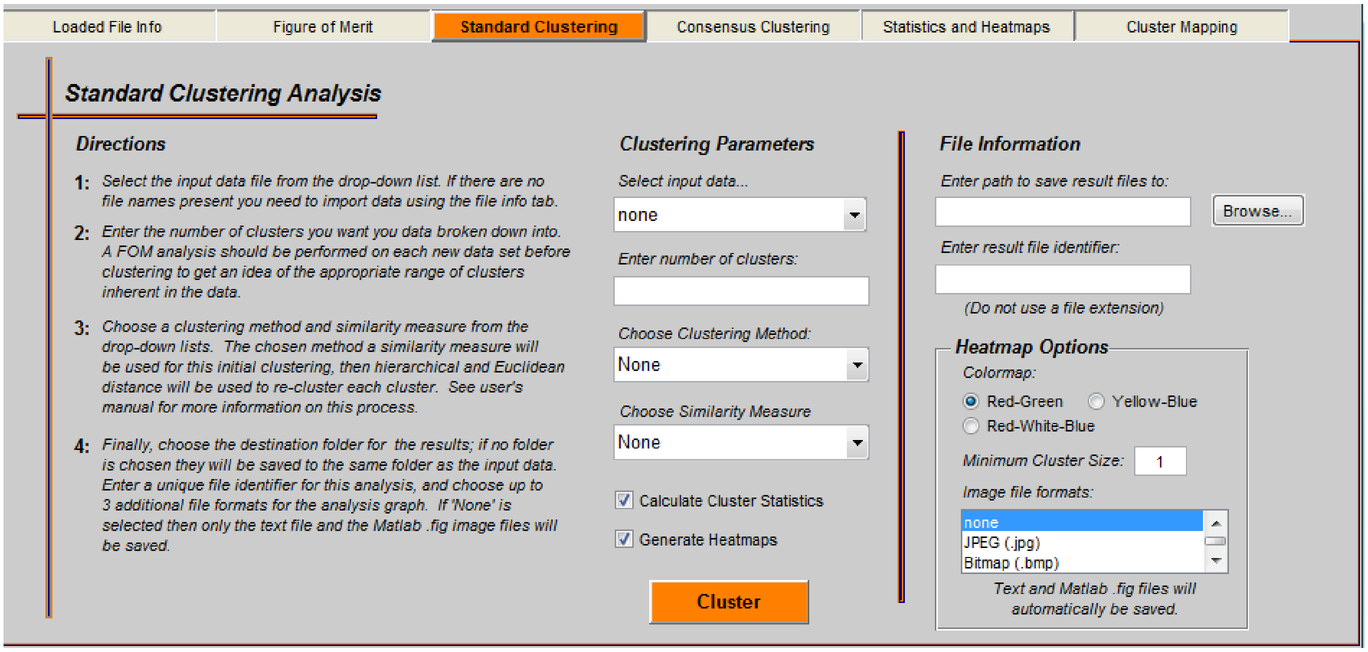
Figure 2: FOM Tab

The first step to performing a FOM or cFOM analysis is to select the Analysis Parameters. The Analysis Parameters that are selected indicate the type of FOM analysis to be performed. The Analysis Parameters include selecting the input data, the clustering methods for comparison, the clustering similarity measure, the figure of merit algorithm type (Original FOM or correlation-biased FOM), and finally the cluster intervals that should be used. This tool implements two versions of the figure of merit, the original Euclidean-biased FOM and a new correlation-biased cFOM; for more information on either of these see (Yeung, Haynor et al. 2001; Olex, John et al. 2007). Unfortunately, the time it takes the original FOM to run is linearly related to the number of genes being clustered while the cFOM is exponential; thus cFOM will take much longer to complete the analysis.

The second step to completing a FOM or cFOM analysis is to specify the output file options. The analysis output may be saved to any location, but if no location is chosen the output will automatically be saved to the same location as the input file. Then, a name for the analysis results is needed, followed by optionally selecting additional image file formats for the analysis graph. See the Walk-through: Figure of merit analysis section for more details on using this tab effectively.

Tab: Standard Clustering *(updated)*

The Standard Clustering tab provides two standard clustering routines, k-means and hierarchical clustering. On-screen directions are provided to the left with parameter selection on the right. A screen-shot of the Standard Clustering tab is shown in Figure 3.



Standard Clustering Analysis

Directions

1: Select the input data file from the drop-down list. If there are no file names present you need to import data using the file info tab.

2: Enter the number of clusters you want you data broken down into. A FOM analysis should be performed on each new data set before clustering to get an idea of the appropriate range of clusters inherent in the data.

3: Choose a clustering method and similarity measure from the drop-down lists. The chosen method a similarity measure will be used for this initial clustering, then hierarchical and Euclidean distance will be used to re-cluster each cluster. See user's manual for more information on this process.

4: Finally, choose the destination folder for the results; if no folder is chosen they will be saved to the same folder as the input data. Enter a unique file identifier for this analysis, and choose up to 3 additional file formats for the analysis graph. If 'None' is selected then only the text file and the Matlab .fig image files will be saved.

Clustering Parameters

Select input data... none

Enter number of clusters:

Choose Clustering Method: None

Choose Similarity Measure: None

☑ Calculate Cluster Statistics

☑ Generate Heatmaps

File Information

Enter path to save result files to:

Enter result file identifier:

(Do not use a file extension)

Heatmap Options

Colormap:

● Red-Green ○ Yellow-Blue
○ Red-White-Blue

Minimum Cluster Size: 1

Image file formats: none, JPEG (.jpg), Bitmap (.bmp)

Text and Matlab .fig files will automatically be saved.

Figure 3: Standard Clustering Tab

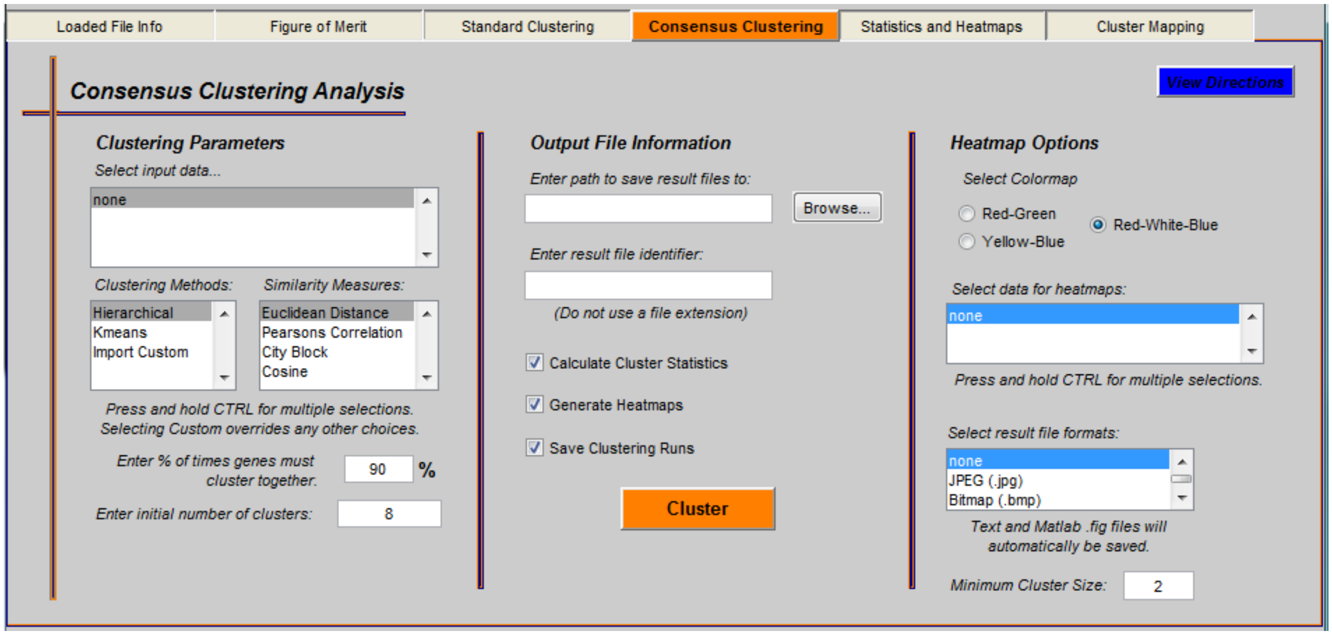
The Clustering Parameters section is used to set all clustering options for both Standard Clustering algorithms. All clustering parameters must be set, as there are currently no default values. The hierarchical clustering algorithm is implemented differently than most other applications. Here a pre-specified number of clusters is required; an explanation of this can be found in Olex *et al.* (Olex and Fetrow 2007). Therefore, the number of clusters to use must be entered for both k-means and hierarchical clustering.

For either clustering algorithm chosen the user may generate a cluster statistics file and/or heatmaps for each cluster generated by checking the boxes above the 'Cluster' button. By default these boxes are checked. If the user chooses to generate heatmaps, then additional options are made available on the right in the 'Heatmap Options' panel. In this panel the user may chose a red-green, yellow-blue or red-white-blue color scheme, multiple image file formats, and a minimum cluster size. The 'minimum cluster size' option allows the user to specify the smallest cluster size that should be considered for heatmap generation. For example, if it is set to 10, then only clusters of size 10 and greater will have a heatmap generated. The user may enter 1 to include all clusters.

Finally, once all clustering options are set the save file information, including a base file name and destination path, must be entered to run the analysis. If no destination path is entered the results will be saved in the same location as the input file. To run the analysis the 'Cluster' button must be pushed.

Tab: Consensus Clustering *(updated)*

The Consensus Clustering tab provides all functions related to performing consensus clustering. On-screen directions can be displayed by pressing the 'View Directions' button in the upper right corner of the tab. A screen-shot of the Consensus Clustering tab is shown in Figure 4.



Consensus Clustering Analysis

Clustering Parameters

Select input data...

none

Clustering Methods: Hierarchical, Kmeans, Import Custom

Similarity Measures: Euclidean Distance, Pearsons Correlation, City Block, Cosine

Press and hold CTRL for multiple selections. Selecting Custom overrides any other choices.

Enter % of times genes must cluster together: 90 %

Enter initial number of clusters: 8

Output File Information

Enter path to save result files to:

Enter result file identifier:

(Do not use a file extension)

☑ Calculate Cluster Statistics

☑ Generate Heatmaps

☑ Save Clustering Runs

Heatmap Options

Select Colormap

○ Red-Green
○ Yellow-Blue
● Red-White-Blue

Select data for heatmaps:

none

Press and hold CTRL for multiple selections.

Select result file formats:

none
JPEG (.jpg)
Bitmap (.bmp)

Text and Matlab .fig files will automatically be saved.

Minimum Cluster Size: 2

Figure 4: Consensus Clustering Tab

The Consensus Clustering tab offers a wide variety of functions and flexibility to the user, and is by far the most complicated in this application. A detailed description with examples of each function can be found in the Walk-through: Consensus Clustering section of this tutorial. A brief description of each function is provided here. There are 3 main sections to this tab: Clustering Parameters, Output File Information and Heatmap Options.

Clustering Parameters: This section is used to set up the type of consensus clustering that is to be performed. First, the user must select one or more input files that should be used to generate the consensus. Each input file must contain the same number of rows and columns, and the entries must be in the same order with matching row labels. Next, the user must select the Clustering Method(s) to use in the analysis. One or both of kmeans and hierarchical clustering may be chosen. If only hierarchical is chosen, the user must have selected either two or more data sets, or two Similarity Measures. Next the similarity measure(s) is chosen, and the user again has the option to choose either one or both of them. The consensus clustering tab has been upgraded to include the additional threshold option ("Enter % of times genes must cluster together"). Here the user types in what percent of the time two genes must be found in the same cluster to be included in a consensus cluster together. Version 2 of SC2ATmd only allowed genes that clustered together 100% of the time to be included in a consensus cluster; however, research showed that was too strict, and allowing for a little fuzziness would accommodate biological variability across replicates. The last section to the Clustering Parameters section changes depending on the Clustering Method chosen. If hierarchical is chosen the user only needs to specify the number of clusters to use in the initial steps of the algorithm. If Kmeans is chosen, the user will also need to specify the number of time the kmeans algorithm should be repeated using a random initialization. Finally, if Import Custom is chosen the user must locate a pre-clustered file from which consensus clusters should be extracted.

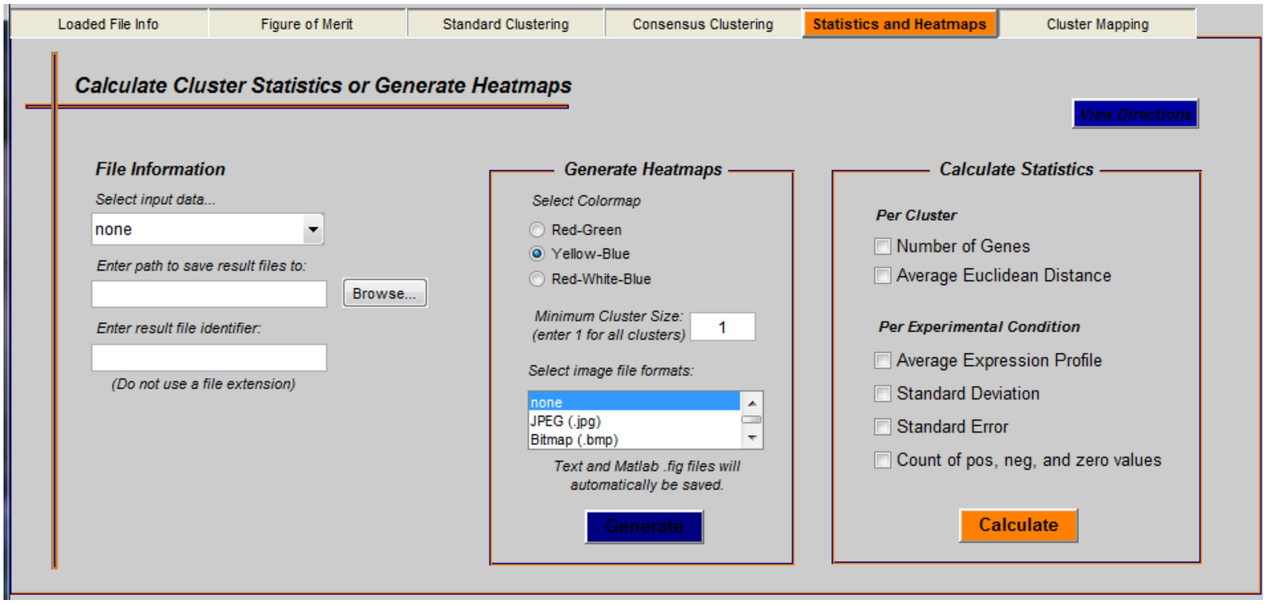
Output File Information: This section is used to specify what additional files should be generated and where they should be saved. A destination path must be specified otherwise an error will occur. The user has the option to generate 3 additional types of files along with the default text file containing the results: a text file containing cluster statistics, heatmap image files, and/or the results of each clustering run that was generated and used to extract consensus clusters. If the user chooses to save the Cluster Runs, this output file may be used as input into the Import Custom function to obtain the same consensus clustering results. This can be used to generate more heatmaps of the same data if they were not generated the first time around.

Heatmap Options: If the user decides to generate heatmaps the 'Heatmap Options' section on the far right will become active. Along with selecting the color scheme, image file types and minimum cluster size (described in the previous section) the user also must choose what input data to use for each heatmap. One or more data files may be selected.

Note: Be careful with using a small minimum cluster size such as 1 or 2. Depending on the size of your data set and the selected parameters, consensus clustering can generate hundreds of consensus clusters.

Tab: Statistics and Heatmaps

This tab provides to distinct functions: the generation of heatmap images from pre-clustered data, and the calculation of cluster statistics for pre-clustered data. On-screen directions can be viewed by pressing the 'View Directions' button at the top right of the tab. A screen-shot of this tab is shown in Figure 5.



File Information

Select input data... none

Enter path to save result files to:

Enter result file identifier:

(Do not use a file extension)

Generate Heatmaps

Select Colormap

○ Red-Green
● Yellow-Blue
○ Red-White-Blue

Minimum Cluster Size: (enter 1 for all clusters) 1

Select image file formats: none (selected), JPEG (.jpg), Bitmap (.bmp)

Text and Matlab .fig files will automatically be saved.

Calculate Statistics

Per Cluster

□ Number of Genes
□ Average Euclidean Distance

Per Experimental Condition

□ Average Expression Profile
□ Standard Deviation
□ Standard Error
□ Count of pos, neg, and zero values

Figure 5: Cluster Statistics and Heatmap Tab

This tab is broken down into 3 sections which are briefly described below.

File Information: For both heatmap and statistics functions an input file must be specified along with a destination path and file name for the resultst to be saved to. If a destination path is not chosen then the results will be saved in the same location as the input file.

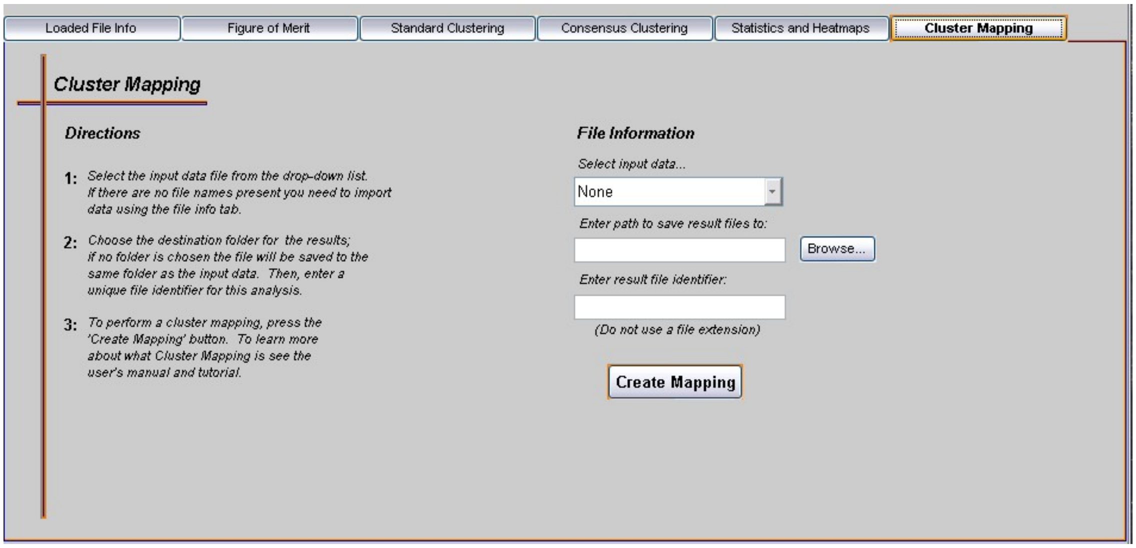
Generate Heatmaps: This section is similar to those on both the Standard and Consensus Clustering tabs. The user selects a pre-clustered data file, and then chooses the heatmap color scheme, minimum cluster size, and any additional image formats each heatmap should be saved as. To generate a heatmap image, each pre-defined cluster is re-clustered using hierarchical clustering with Euclidean distance as the similarity measure to generate the dendrogram and element order.

Please note: The dendrogram DOES NOT reflect the original clustering used to determine the pre-defined clusters. The dendrogram is generated by re-clustering each user-defined cluster using the hierarchical algorithm.

Calculate Statistics: This function is also provided on the Standard and Consensus Clustering tabs, however here the user has the option of selecting which statistics should be calculated. Thus, if the user is not interested in one or more of the statistics these can be left un-checked and will not be included in the output.

Tab: Cluster Mapping top

The Cluster Mapping tab provides a unique function in which one clustering solution is described in terms of another (Olex and Fetrow 2007). To use this function two different clustering solutions of the same data must have been generated (such as using two different clustering algorithms or similarity metrics) and the solutions must be formatted correctly (see the Input File Formats help file). On-screen directions are provided on the left with the File Information section on the right. This analysis is very easy to use as all the user needs to do is specify an input file and an output file name and destination path. A screen-shot of this tab is shown in Figure 6.



Directions

1: *Select the input data file from the drop-down list. If there are no file names present you need to import data using the file info tab.*

2: *Choose the destination folder for the results; if no folder is chosen the file will be saved to the same folder as the input data. Then, enter a unique file identifier for this analysis.*

3: *To perform a cluster mapping, press the 'Create Mapping' button. To learn more about what Cluster Mapping is see the user's manual and tutorial.*

File Information

Select input data... None

Enter path to save result files to:

Enter result file identifier:

(Do not use a file extension)

Figure 6: Cluster Mapping Tab

Walk-through: Figure of merit analysis

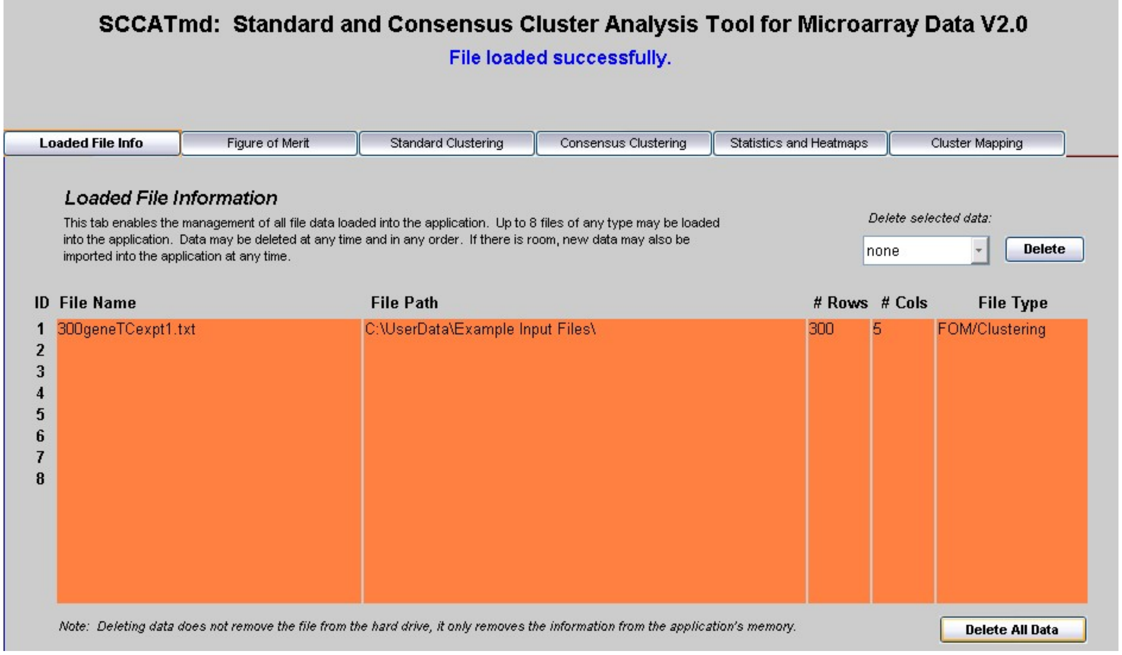
The Figure of Merit analysis is a method that quantitatively compares the performance of several clustering algorithms on one data set. It tells the user which clustering algorithm created the most homogeneous clusters with their data, and suggests an optimal range where the ideal number of groups inherent in the data may lie. For more information on the FOM and it's implementation in this application see Yeung *et al.* and Olex *et al.* (Yeung, Haynor et al. 2001; Olex, John et al. 2007).

The following tutorial will walk the user through performing a FOM analysis on an example microarray time course data set. The example file being used is `300geneTCexpt1.txt` and is located in the tutorial folder provided with this distribution. This data set is composed of 300 randomly selected genes from a microarray time course experiment studying the transcriptional changes during dendritic cell maturation induced by Poly(I:C). Further details of this study can be found in Olex *et al.* (Olex, Hiltbold et al. 2007). Note that this example data set was randomly generated from the data set mentioned in Olex *et al.*; it is not an actual significant data set.

Begin walk-through:

Before any analyses can be performed the proper data file must be loaded into the program. To do this we follow the steps in the Input File Formats help file under the 'FOM/Clustering File Format' section to load the `300geneTCexpt1.txt` file. This section also describes the proper format for all input files. Before loading your own files make sure they are in the right format.

After the file was loaded correctly, the file information should have been updated on the Loaded File Info tab as is shown in Figure 7.



SCCATmd: Standard and Consensus Cluster Analysis Tool for Microarray Data V2.0

File loaded successfully.

Loaded File Information

This tab enables the management of all file data loaded into the application. Up to 8 files of any type may be loaded into the application. Data may be deleted at any time and in any order. If there is room, new data may also be imported into the application at any time.

Delete selected data: none

ID	File Name	File Path	# Rows	# Cols	File Type
1	300geneTCexpt1.txt	C:\UserData\Example Input Files\	300	5	FOM/Clustering
2					
3					
4					
5					
6					
7					
8					

Note: Deleting data does not remove the file from the hard drive, it only removes the information from the application's memory.

Figure 7: Updated file information.

Once the data has been successfully loaded into the system, click on the ‘Figure of Merit’ tab. Follow the steps below to run the FOM analysis. In this example we will run a FOM analysis comparing k-means and hierarchical clustering using Euclidean distance as the similarity metric. If you wish to use Pearson’s correlation coefficient as the similarity metric then it is recommended that a cFOM (correlation-biased FOM) analysis be run (see Olex, John et al. 2007 for more information on why). However, be careful with the cFOM analysis as it takes a lot longer to complete than the original FOM.

1. Under the ‘Analysis Parameters’ section, if the file that was just loaded is not already selected, select it from the ‘Input data’ drop down list.
2. Check both boxes under the ‘Clustering Methods’ section to choose k-means and hierarchical clustering for comparison.
3. Select ‘Euclidean Distance’ from the ‘Similarity Measure’ drop-down box.
4. Select ‘Original FOM’ from the ‘Algorithm Type’ drop-down box.
5. Next we will need to enter a range of cluster numbers for the FOM analysis to iterate over into the ‘Cluster Intervals’ box. What the FOM does is to use each clustering method to divide the data into, say, 2 groups. Then it calculates a score for each algorithm to determine which algorithm generated the most homogeneous 2 clusters. Then the FOM repeats this process using the next number of clusters on the list, say 4, to determine which algorithm generated the most homogenous 4 clusters. This is repeated for each number of clusters we specify in the list. In this example we will set our range of cluster numbers to 2, 6, 10, 14, 18, 22, 26, 30, and 34. The range entered should be evenly spaced; as the algorithm then calculates how many clusters are optimal depends on this. Additionally, it must start with 2 clusters or greater, as it is counter intuitive to generate 1 cluster.

6. Next, the program needs to know where the results should be saved and under what name. Under the 'File Information' section, either enter in a path by hand or use the Browse button to locate the appropriate folder. If no path is entered, the results will be saved in the same folder as the input file. Next, enter in an analysis identifier that is unique to this analysis. A good way is to use the input files name with either FOM or cFOM at the end. We will use this convention in this example, so enter in `300geneTCexpt1_FOM` as the identifier.
7. Finally, both the FOM and cFOM analyses generate a graph that plots the analysis scores. This graph is automatically saved as a Matlab .fig file. If you wish to save it in other formats as well you may select them in the 'Image file formats' box. Hold down the CTRL key to make multiple selections. For this example, we will select JPEG as an additional file format.
8. After everything has been entered, the FOM tab should look like that in Figure 8.
9. Press the 'Perform Analysis' button to initiate the FOM analysis. If there are any errors in your input, an error message will appear in the status window. If this happens simply fix the specified field and press the button again. The status window will show that the analysis is in progress, so just wait until it is finished before proceeding

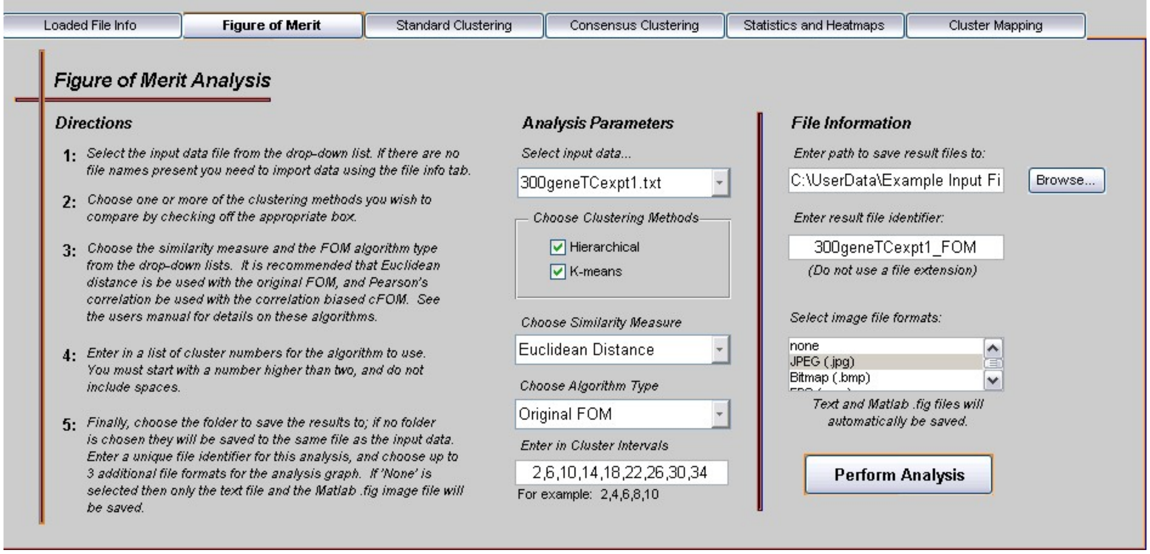


Figure of Merit Analysis

Directions

1: Select the input data file from the drop-down list. If there are no file names present you need to import data using the file info tab.

2: Choose one or more of the clustering methods you wish to compare by checking off the appropriate box.

3: Choose the similarity measure and the FOM algorithm type from the drop-down lists. It is recommended that Euclidean distance is be used with the original FOM, and Pearson's correlation be used with the correlation biased cFOM. See the users manual for details on these algorithms.

4: Enter in a list of cluster numbers for the algorithm to use. You must start with a number higher than two, and do not include spaces.

5: Finally, choose the folder to save the results to; if no folder is chosen they will be saved to the same file as the input data. Enter a unique file identifier for this analysis, and choose up to 3 additional file formats for the analysis graph. If 'None' is selected then only the text file and the Matlab .fig image file will be saved.

Analysis Parameters

Select input data... 300geneTCexpt1.txt

Choose Clustering Methods
☑ Hierarchical
☑ K-means

Choose Similarity Measure: Euclidean Distance

Choose Algorithm Type: Original FOM

Enter in Cluster Intervals: 2,6,10,14,18,22,26,30,34
For example: 2,4,6,8,10

File Information

Enter path to save result files to: C:\UserData\Example Input Fi

Enter result file identifier: 300geneTCexpt1_FOM
(Do not use a file extension)

Select image file formats:
none
JPEG (.jpg)
Bitmap (.bmp)

Text and Matlab .fig files will automatically be saved.

Figure 8: FOM tab with selected analysis options.

Once the analysis is complete a plot of the analysis results will automatically appear on the screen. This plot is shown in Figure 9 where the FOM score (y-axis) is plotted against the range of cluster intervals (x-axis).

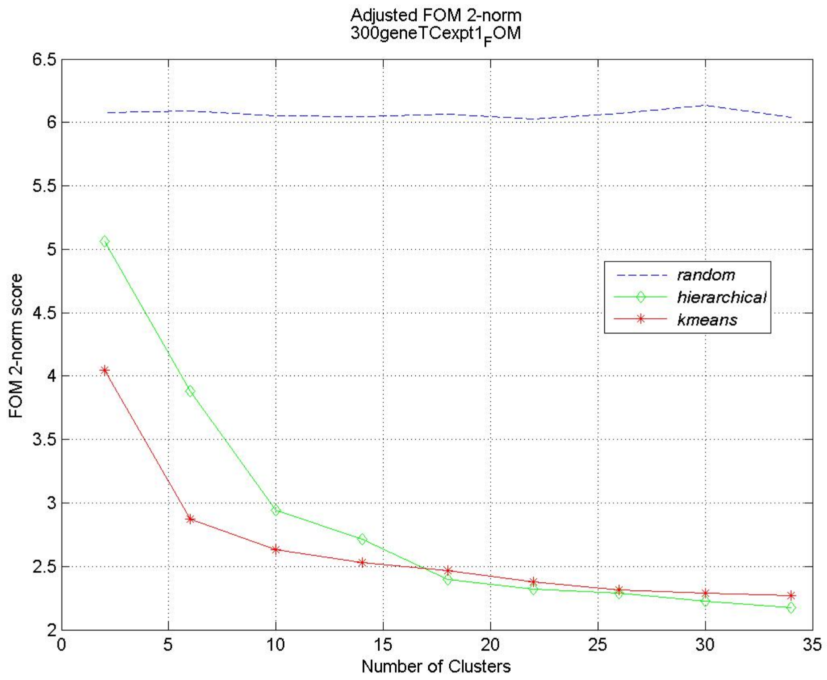
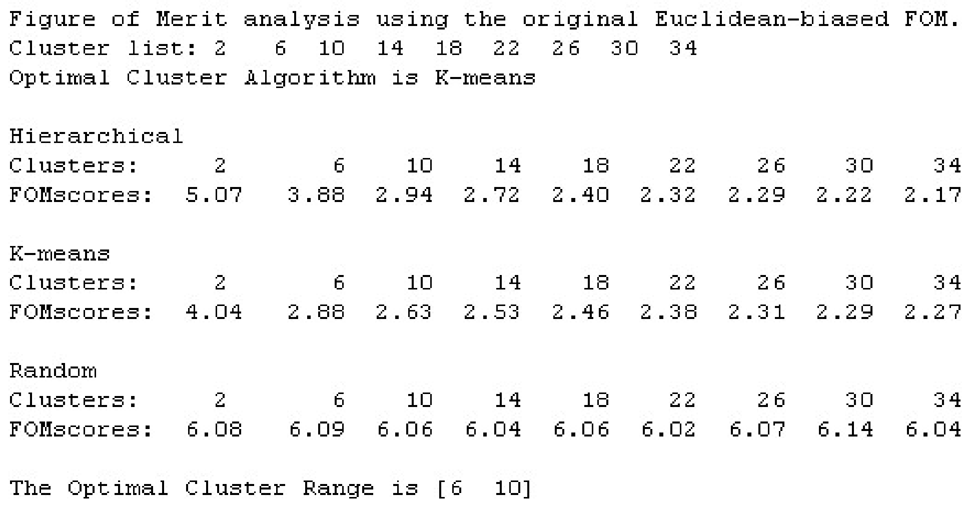


Figure 9: FOM graph output of analysis results.

To interpret this analysis, remember that a lower FOM score indicates higher homogeneity of clusters. Here the k-means algorithm generated higher quality clusters no matter how many clusters were used, thus it is the 'better' clustering algorithm for this data. A message box will appear after the graph has been generated notifying you of the optimal number of clusters that are inherent in this data set. This information can also be found in the results text file that was generated during the analysis. For this analysis the optimal number of clusters is between 6 and 10.

Even though a lower FOM score is better when comparing different algorithms, this cannot be used to determine the ideal number of clusters to use. Inherently the FOM score will decrease as the number of clusters increase (Yeung, Haynor et al. 2001), so we can't just pick the number of clusters that obtains the lowest score. If we did that, then the ideal number of clusters would ultimately equal the number of genes (i.e. every gene is in its own cluster). Therefore, we need to find that point where adding more clusters doesn't drastically change the FOM score. This is the point where there is an 'elbow' in the graph. This application provides a method to calculate this point based on the standard deviations of FOM changes (Olex, Hiltbold et al. 2007). Any number of clusters within this range is acceptable to use, however this is affected by the distance between cluster intervals input by the user. For example, if we would have entered 2, 8, 14, 20, etc. then the optimal range would be 8 to 14 clusters instead of 6 to 10. Thus, it is important to pay attention to the cluster numbers you enter in initially.

This analysis also outputs a text file with the optimal clustering algorithm, ideal cluster range, and all raw FOM scores in it. Using the `300geneTCexpt1.txt` file you should get something like Figure 10:



```
Figure of Merit analysis using the original Euclidean-biased FOM.
Cluster list: 2   6  10  14  18  22  26  30  34
Optimal Cluster Algorithm is K-means

Hierarchical
Clusters:     2      6     10     14     18     22     26     30     34
FOMscores:  5.07   3.88   2.94   2.72   2.40   2.32   2.29   2.22   2.17

K-means
Clusters:     2      6     10     14     18     22     26     30     34
FOMscores:  4.04   2.88   2.63   2.53   2.46   2.38   2.31   2.29   2.27

Random
Clusters:     2      6     10     14     18     22     26     30     34
FOMscores:  6.08   6.09   6.06   6.04   6.06   6.02   6.07   6.14   6.04

The Optimal Cluster Range is [6  10]
```

Figure 10: FOM text output of analysis results.

The FOM scores for k-means may change slightly, but the hierarchical clustering score should be exactly the same. At the top, the version of the FOM is listed; Euclidean-biased is used when Euclidean distance is the similarity measure, and correlation-biased is used when Pearson's correlation coefficient is the similarity measure. Next, the cluster interval list you specified is printed followed by the optimal clustering algorithm. The optimal clustering algorithm is determined to be the one with the lowest average FOM score over all iterations. Then the raw FOM scores for each clustering algorithm used are listed followed by the range for the ideal number of clusters.

This concludes the walk-through of the FOM analysis. The results of the FOM analysis can now be used to actually cluster the data and generate heatmaps. A walk-through for clustering with SC^{2}ATmd is provided next.

Walk-through: Standard Cluster Analysis

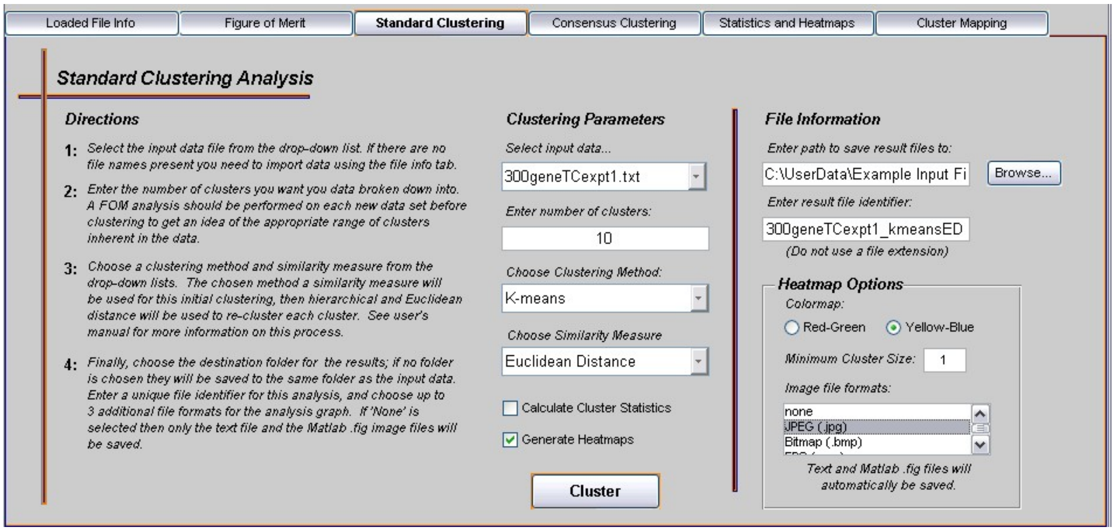
The Standard Clustering tab allows the user to perform standard k-means and hierarchical clustering on their data. The following is a walk-through of performing a standard cluster analysis, and is a continuation of the FOM walk-through above using the `300geneTCexpt1.txt` file.

Begin walk-through:

The FOM analysis previously done on the `300geneTCexpt1.txt` data set suggests that the most appropriate clustering algorithm to use with this data is k-means, and the ideal number of clusters is between 6 and 10 (using Euclidean distance). Any number between 6 and 10 is ok to choose. If you want to narrow the choice down more you can repeat the analysis with smaller cluster intervals between 6 and 10. We will use 10 because 'by eye' this is where the graph starts to look like an 'elbow' in comparison to the look of the graph at 6. Now that we know what our clustering options should be, lets cluster the data. Click on the Standard Clustering tab to start; because clustering and FOM use the same file format it is not necessary to load the file again (unless you skipped the FOM walk-through). Follow the steps below to generate a Standard clustering analysis.

1. Make sure the appropriate data file is selected in the 'Input data' box. If not, then select it.
2. Enter the number of clusters to generate. From the FOM analysis done above we want 10, so enter 10 in the 'number of clusters' box.
3. Select the clustering method from the drop-down box. The FOM analysis indicated that k-means generates clusters with higher homogeneity than hierarchical, so select k-means.

4. Choose the similarity measure to cluster the data with. The similarity measure defines how 2 elements are considered to be similar. For example, Euclidean distance mainly looks at similarity in expression level while Pearson's correlation strictly looks at similar expression patterns or shapes. The FOM analysis is dependent on the similarity measure, so if you did the original FOM analysis choose Euclidean distance, but if you did the correlation-biased FOM analysis choose Pearson's correlation. In this example we had used the original FOM, so choose Euclidean distance.
5. Next you have a choice of generating some additional files besides just the standard text file. The 'Calculate Cluster Statistics' option will generate an additional text file with statistics on each cluster. For this example we don't need this, so uncheck it. Next you have the option of generating heatmap image files for each cluster. For this example we want to see the heatmaps, so leave this box checked.
6. Enter the output file information on the right of the tab. First select a destination folder for the results. Again, if not folder is selected the results will be saved in the same location as the input file. Then, enter in an analysis identifier. Generally it is a good idea to indicate the clustering method and similarity measure used in the analysis. For this example we will use the file name followed by '_kmeansED' which indicates that k-means was used as the method and Euclidean Distance was the similarity measure.
7. Finally, enter in the heatmap image options. You can select a color scheme, the minimum cluster size, and any additional file formats to save the images in. The color scheme can be either one you want. For this example the yellow-blue has been chosen where yellow will indicate up regulation (or positive expression values) and blue will represent down-regulation (or negative expression values). The minimum cluster size tells the program when to stop generating heatmaps. If you only want heatmaps for clusters with 5 or more genes in them, then enter 5. If you want all clusters to be represented as a heatmap then enter 1. Finally, choose any additional image file formats that you want generated. Hold down the CTRL key to make multiple selections. For this example JPEG has been added. *Note: Because we are generating 10 clusters then 10 .fig files and 10 jpeg files will be generated for a total of 20 image files. If you selected additional file types then 10 of each of those will be generated as well.*
8. Once all the fields are filled in press the 'Cluster' button to start the analysis. If anything is missing or wrong an error message will appear in the status window. Again, just fix the problem fields and resubmit the analysis. Once the clustering is started the heatmaps will start appearing in the screen. Wait until all heatmaps are created before you do anything like close them out. Figure 11 shows the Standard Clustering tab with all options set.



Standard Clustering Analysis

Directions

1: Select the input data file from the drop-down list. If there are no file names present you need to import data using the file info tab.

2: Enter the number of clusters you want you data broken down into. A FOM analysis should be performed on each new data set before clustering to get an idea of the appropriate range of clusters inherent in the data.

3: Choose a clustering method and similarity measure from the drop-down lists. The chosen method a similarity measure will be used for this initial clustering, then hierarchical and Euclidean distance will be used to re-cluster each cluster. See user's manual for more information on this process.

4: Finally, choose the destination folder for the results; if no folder is chosen they will be saved to the same folder as the input data. Enter a unique file identifier for this analysis, and choose up to 3 additional file formats for the analysis graph. If 'None' is selected then only the text file and the Matlab .fig image files will be saved.

Clustering Parameters

Select input data... 300geneTCexpt1.txt

Enter number of clusters: 10

Choose Clustering Method: K-means

Choose Similarity Measure Euclidean Distance

□ Calculate Cluster Statistics

☑ Generate Heatmaps

File Information

Enter path to save result files to: C:\UserData\Example Input Fi

Enter result file identifier: 300geneTCexpt1_kmeansED

(Do not use a file extension)

Heatmap Options

Colormap: ○ Red-Green ● Yellow-Blue

Minimum Cluster Size: 1

Image file formats: none, JPEG (.jpg), Bitmap (.bmp)

Text and Matlab .fig files will automatically be saved.

Figure 11: Standard Clustering parameters are set.

Once all the heatmaps have been loaded onto the screen you may start to look at them in more detail. Each one of these images is one of the 10 clusters the data file was broken down into. Whenever SC^{2}ATmd clusters data using k-means or Hierarchical clustering, it re-clusters each cluster using hierarchical clustering and Euclidean distance so the heatmaps of each cluster are organized by expression intensity. Figure 12 is one example of a heatmap generated by SC^{2}ATmd. Note that your clusters will not look exactly the same; k-means is randomly initialized thus a slightly different group of clusters will result each time it is run.

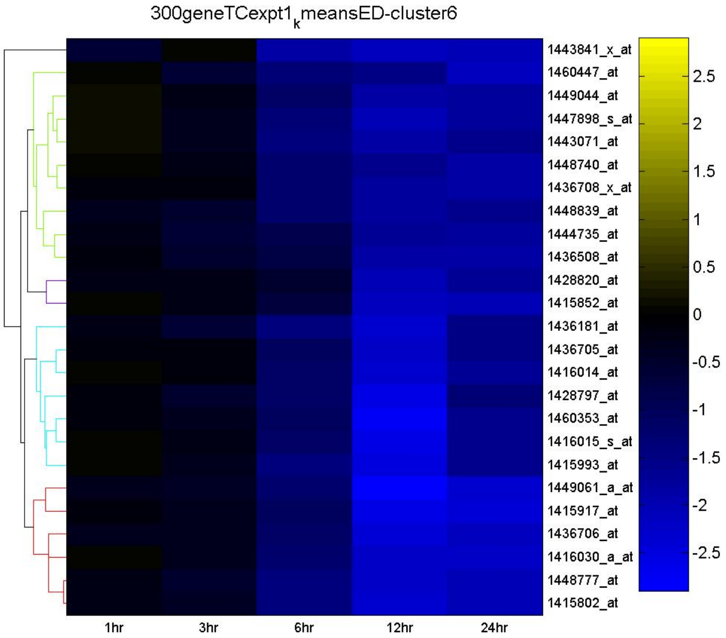


Figure 12: Example heatmap with dendrogram.

Figure 12 is a heatmap where each column represents an experimental condition, and each row is one gene. The file name/figure title that was entered is at the top, gene names/id's are to the right, column labels are at the bottom, the hierarchical dendrogram is to the left, and the color scale is far to the right. This figure happens to be cluster #6, and consists of mostly down regulated or negatively expressed genes. This is time course data, so we can see that all these genes did not have much change in expression until 6 hours after stimulation where they then exhibited a sustained decrease in expression through hour 24. All these genes exhibit a similar pattern and levels of expression so may be related in some way biologically.

Along with heatmaps of every cluster, SC^{2}ATmd outputs all clustering results in a text file which can easily be imported into Excel for further processing. To learn about the organization of this file, see the Output File Formats help file.

The standard clustering analysis is now complete. The Matlab figures may be modified based on the user's preferences and/or their Matlab knowledge. If you selected to output additional image files such as jpeg or PDF, these can be imported and used directly in other documents. Next a walk-through of the Consensus Clustering analysis will be given.

Walk-through: Consensus Cluster Analysis (updated for version 3)

The Consensus Clustering tab allows the user to perform a variety consensus clustering analyses on their data. There are many different ways 'consensus clusters' can be identified. This tab has been designed to be as flexible as possible so that the user may perform any sort of consensus analysis they want. Below there are several walk-through's that explain the basic types of consensus clustering and their purpose. A tab overview is also provided that explains in detail the multiple functions this tab can perform. At any time on-screen directions may be viewed by pressing the 'View Directions' button at the top right of the tab.

Algorithm Overview

In its simplest form consensus clustering takes 2 or more standard clustering solutions (like those you would get from the Standard Clustering Analysis) for the same data set and identifies those sub-groups of elements that were found in the same cluster a specified number of times. Thus, it identifies the most robust and reproducible groups of clustered elements. The algorithm has 2 basic steps: 1) take the input data set(s) and perform a Standard Cluster analysis on each using every combination of the options defined by the user (clustering methods, similarity measures, initial number of clusters, etc.); 2) take these cluster solutions and compare them to identify those sub-groups of elements that were consistently placed in the same cluster a specified number of times.

Tab Overview

Before we begin with the walk-through, the user should become familiar with the tab interface and all its options.

Input data: For consensus clustering the user has the option of using one or more (up to 8) data sets for the extraction of consensus clusters. This option allows the user to identify elements that are clustered together in different or replicate experiments. For example, if a small group of genes are consistently clustered together even when different stimuli are used (different experiments), there is a good chance that these genes a related in some fashion. To select multiple data sets hold down the CTRL key while clicking on those you want to use. Consensus clusters can also be generated from one data set by either selecting multiple clustering methods or similarity measures (described next). Or, if only one data set is used and you don't want to compare clustering methods or similarity measures you may select kmeans as the clustering method and instruct the program to perform multiple repetitions with a random initialization.

Clustering methods: Currently only two clustering methods are available for the consensus analysis, kmeans and hierarchical. If the user is performing a consensus analysis from scratch (i.e. the input data must go through the standard analysis first) then at least one method must be chosen. If the user already has several clustering solutions for which they want to identify consensus clusters from, then the Import Custom method may be chosen. As stated above, multiple clustering methods may be chosen. This enables the user to take one (or more) data set(s), cluster it using both methods, and then see how similar the results are based on consensus clusters returned.

Similarity measures: Four similarity measures are provided for performing the initial standard cluster analysis prior to identifying consensus clusters. Euclidean distance mainly determines the similarity between two elements based on similar levels of expression (for gene expression data) or magnitude of data. City block is similar to Euclidean but measures distance from point A to point B using 'city block' type movements instead of 'as the crow flies' like Euclidean distance does. Pearson's correlation finds similar patterns of expression (for gene expression data) or shape of the data across all conditions. Cosine measures the cosine of the angle between the two vectors of data being compared. One or all of these measures may be used in the identification of consensus clusters.

Consensus Threshold: New to version 3 is the consensus threshold option. Originally, consensus clusters were identified based on a non-optional 100% threshold—that is, two genes were placed in the same consensus cluster if, and only if, they were found in the same cluster for all clustering solutions. Now, the user may specify a lower threshold to identify consensus clusters at various levels of consensus. For example, if 10 clustering solutions were generated and the consensus threshold was set to 90%, then all genes found in the same cluster 9 out of 10 times would be placed in the same consensus cluster. Previously, at the 100% level, genes would have had to clustered together 10 out of 10 times to be included in the same consensus cluster.

Initial number of clusters: Whether kmeans or hierarchical clustering is chosen for the initial standard analysis the number of clusters is needed. Note the initial number of clusters chosen does not determine how many consensus clusters will be generated. The consensus algorithm does a standard analysis first, and then pulls out an undetermined number of sub-clusters from those results as the consensus clusters. The initial number of clusters is used in the standard analysis step of the consensus algorithm, not in the final consensus step. If a FOM or cFOM analysis was done on the data set(s) being used, the initial number of clusters would be that recommended by the FOM or cFOM analysis.

K-means repetitions: If k-means is chosen as the clustering method this field will appear below the 'initial number of clusters' field. K-means is a stochastic clustering method that is randomly initialized for each run (in contrast to the deterministic hierarchical method). Because of this a slightly different clustering solution will be generated each time k-means is run. However, consensus clustering can be used to extract the core sub-groups that always cluster together in every k-mean run by doing a consensus over multiple k-mean solutions of the same data set(s). In order to get an 'averaged' effect it is a good idea to repeat the k-means clustering several times in any consensus analysis that uses this method.

Custom clustering solution: This field will replace the 'Initial number of clusters' and 'k-means repetitions' fields when the 'Import Custom' method is chosen. If you chose to import your own clustering solutions, then this is where you direct the program to the location of the file containing the solutions. Use the Browse button or enter the path manually. If you choose this option you must still have selected at least one data file containing the raw data that was used to generate these solutions. Additionally, any fields that are no longer relevant to this option are deactivated (such as the similarity measure).

Destination path: No matter what analysis is done a location for the output must be selected. For this tab only you must specify a destination for the output. If the field is left blank an error message will occur. This is a result from the ability to use multiple input files from multiple locations. Use the Browse button, or enter the path manually.

Results file name: An analysis identifier, or results file name must also be entered for every analysis. This is what your results will be saved under.

Calculate Cluster Stats: If this box is checked the cluster statistics for each consensus cluster will be calculated and saved. Uncheck this box if you do not want this file generated.

Generate Heatmaps: If you would like heatmaps for each consensus cluster to be generated automatically then check this box. Once the box is checked additional option to the right will become active. These must be filled out if 'Generate Heatmaps' is selected.

Save cluster runs: Checking this option will generate an additional text file that contains each standard clustering solution used in the consensus analysis. The file output by this option can be re-imported back into the consensus algorithm to generate the same consensus clusters as before. If this file is not saved and the same analysis is run again with all the same options the same consensus clusters may not be generated due to random effects caused by k-means.

Heatmap Options: This section contains multiple fields most of which are the same as those found on the Statistics and Heatmap tab. Reference that tab and the walk-through for more details. The one field that differs is the input data for the heatmaps. Because the consensus clustering has the ability to use multiple data files to generate consensus clusters, there is a choice as to which data file is used to generate the heatmaps. This option allows the user to choose one file as a representative, or more than one file. If multiple input files are used then for each consensus cluster the same number of heatmaps will be generated. For example, say there are 3 data sets chosen. Then there will be 3 heatmaps for consensus cluster 1, 3 for cluster 2, etc… Each replicate will use a different set of input data. This comes in handy when a consensus over different experiments is performed and the user wants to look at the differences in expression from data set to data set for an individual cluster.

Walk-through – consensus clustering with one or multiple data file(s)
There are 4 basic types of consensus clustering that can be done using one or more data file(s) other than the Custom Import. When using multiple data files (such as different experiments or replicate experiments) keep in mind that you are not only comparing the methods and measures below, but you are also comparing the different sets of data and pulling out only those sub-groups that show consistency across all data sets. When multiple data sets are used, each data set must contain the same number of row and columns, and the elements must be in the same order. This walk-through will only use one data set in the examples.

1. Multiple k-mean repetitions with 1 similarity measure.
2. 1 clustering method with multiple similarity measures.
3. 2 clustering methods with 1 similarity measure.
4. 2 clustering methods with multiple similarity measures.

Multiple k-mean repetitions with 1 similarity measure – This type of analysis will take the input data set and cluster it multiple times with k-means using a random initialization to get slightly differing solutions each time. These solutions will be compared to identify those elements that were consistently placed in the same cluster every time. The idea here is that these consensus clusters form the core, robust clusters of this data set. No matter how k-means is initialized, these sub-groups are always found together, thus they may be tightly associated with one another in some way.

1 clustering method, multiple similarity measures – This type of analysis would be used to compare the effect different similarity measures have on the same data set. These consensus clusters would indicate that the grouped elements are highly similar in more than just one way (e.g. magnitude and shape of expression profiles instead of just one or the other).

2 clustering methods, 1 similarity measure – This type of analysis can be used to examine the effects different clustering methods have on the same data set. K-means and hierarchical clustering use different algorithms and approaches to clustering data. Thus, elements that form a consensus cluster under these conditions would be impervious to the algorithmic differences of these two clustering methods.

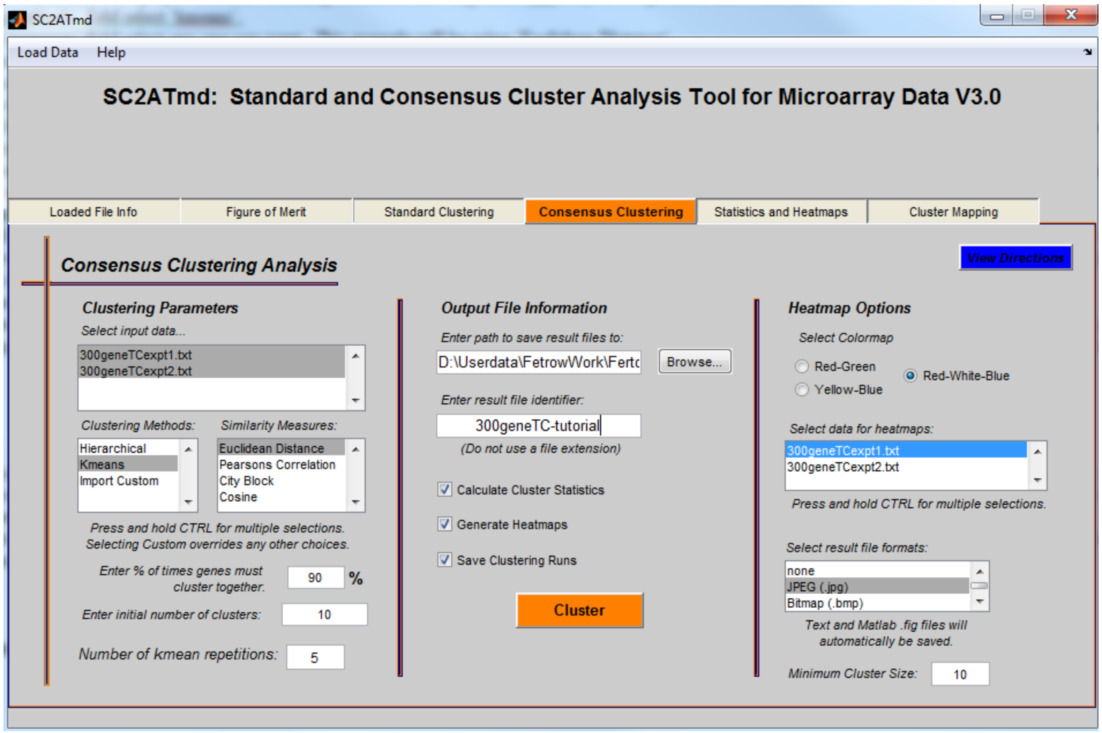
2 clustering methods, multiple similarity measures – This analysis can be done, however it is getting a little too complicated to be able to extract real meaning behind the generated consensus clusters. This analysis is not recommended to anyone who is not very familiar with the algorithmic and mathematical differences of the clustering methods and similarity measures.

Begin walk-through

We will not walk through all 4 analysis types, but will only look at the first one: multiple k-means repetitions with 1 similarity measure using 2 data sets. To perform this analysis, follow the steps below.

1. This example will be using the same file that was used in the Standard Clustering walk-through, `300geneTCexpt1.txt`, as well as its biological replicate, `300geneTCexpt2.txt`. Make sure these files are loaded into the application then click on the Consensus Clustering tab.
2. In the ‘Input data’ field, select the proper data file. If multiple files are showing, make sure only the 2 specified above are selected.
3. In the ‘clustering methods’ field select ‘kmeans’.
4. In the ‘similarity measures’ field select any one you want. This example will be using ‘Euclidean Distance’.
5. Enter the initial number of clusters as 10. This was the FOM analysis results from the previous tutorial.
6. In the ‘consensus threshold’ box, enter 90. This will identify groups of genes that clustered together 90% of the time.

7. You should have noticed that when ‘kmeans’ was selected as the clustering method an extra box appeared at the bottom of that column. Enter the number of time k-means should repeat. We will enter 5 for this example—so each replicate data set will be clustered 5 times resulting in 10 total clustering solutions for consensus. A higher number will generate more robust consensus clusters. You can experiment with this on your own data sets to find a number that works well for your data.
8. In the next column, specify a destination path and a file name for the results that are generated. The file name does not need to contain the input data file name because with consensus clustering the input data file name is automatically appended to the identifier you select.
9. Check the ‘Calculate Statistics’ and ‘Generate Heatmaps’ boxes.
10. In the Heatmap Options section, first choose the color scheme. For this tutorial we will choose the red-white-blue scheme. Select one file in the heatmap data box—both can be selected if you wish, but we will only use one for this example. Finally, enter 10 for the minimum cluster size, and select JPEG for an addition image file type. Consensus clustering generates a lot of clusters so it is a good idea to have a minimum cluster size greater than 2 or 3 so all the singletons that are generated are not displayed as a heatmap.
11. Once all fields have been filled in, your screen should look similar to that in Figure 13.
12. Click the ‘Cluster’ button and wait until all the heatmaps are displayed on the screen. If you are using the file in this tutorial then about 7 to 10 heatmaps should be generated. Since we are using k-means the exact number may vary as each solution is different. If there were any errors, a message will appear in the status window. Fix the errors and submit the analysis again.



SC2ATmd: Standard and Consensus Cluster Analysis Tool for Microarray Data V3.0

Consensus Clustering Analysis

Clustering Parameters

Select input data...
300geneTCexpt1.txt
300geneTCexpt2.txt

Clustering Methods: Hierarchical, Kmeans, Import Custom (selected: Kmeans)

Similarity Measures: Euclidean Distance, Pearsons Correlation, City Block, Cosine (selected: Euclidean Distance)

Press and hold CTRL for multiple selections.
Selecting Custom overrides any other choices.

Enter % of times genes must cluster together: 90 %

Enter initial number of clusters: 10

Number of kmean repetitions: 5

Output File Information

Enter path to save result files to: D:\Userdata\FetrowWork\Fertc

Enter result file identifier: 300geneTC-tutorial
(Do not use a file extension)

☑ Calculate Cluster Statistics
☑ Generate Heatmaps
☑ Save Clustering Runs

Heatmap Options

Select Colormap
○ Red-Green
○ Yellow-Blue
● Red-White-Blue

Select data for heatmaps:
300geneTCexpt1.txt (selected)
300geneTCexpt2.txt
Press and hold CTRL for multiple selections.

Select result file formats:
none
JPEG (.jpg) (selected)
Bitmap (.bmp)
Text and Matlab .fig files will automatically be saved.

Minimum Cluster Size: 10

Figure 13: Consensus clustering with two input files

13. Once the results are generated, you can look at the statistics file that was saved. If you are following this tutorial, it should be named `300geneTC-tutorial-300geneTCexpt1-ClusterStats.txt or 300geneTC-tutorial-300geneTCexpt2-ClusterStats.txt`. Global statistics such as total number of clusters, total number of singleton clusters, average cluster size, etc can be seen. Scrolling down to the bottom, you will notice that only clusters with at least 2 members have stats calculated.
14. The cluster results file (`300geneTC-tutorial-300geneTCexpt1.txt`) contains all consensus clusters including singletons. If you want heatmaps generated for additional clusters in the analysis, this file can be arranged so that it may be used in the Statistics and Heatmaps tab.

15. Finally, the network file (`300geneTC-tutorial_network.txt`) can be used to visualize the consensus clusters as a network. Follow the "Consensus Cluster Network Visualization" tutorial for a walk-through.

Walk-through: Consensus Cluster Network Visualization (new for version 3)

This tutorial is a walk-through of how to visualize the consensus clusters as a network using the output network file (`300geneTC-tutorial_network.txt`) and Cytoscape v2.8.2 (Smoot, Ono et al. 2011). To follow this tutorial, users need to have Cytoscape installed. Cytoscape can be freely downloaded from http://www.cytoscape.org/. Familiarity with Cytoscape is preferable; however, all steps will be covered so that this tutorial can be followed even by novice users. This tutorial will cover the import of the network, import of gene expression data, filtering of network edges to obtain clusters, network layout and formatting based on gene expression data, and the exporting of finalized images.

Begin walk-through

1. Open Cytoscape.
2. Import the network file (`300geneTC-tutorial_network.txt` is used in this tutorial) by going to *File>Import>Network from Table* as shown in Figure 14.

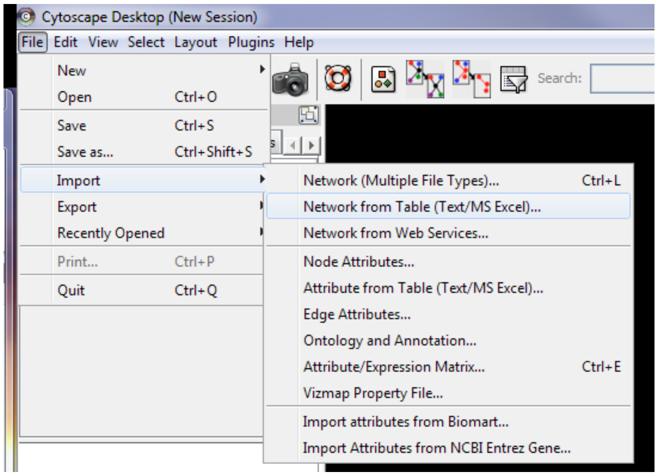
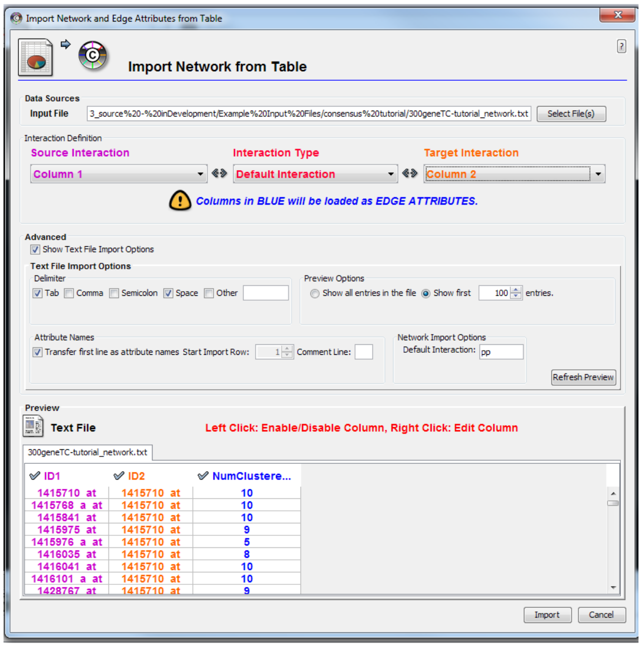

Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
New
Open Ctrl+O
Save Ctrl+S
Save as... Ctrl+Shift+S
Import
Export
Recently Opened
Print... Ctrl+P
Quit Ctrl+Q
Search:
Network (Multiple File Types)... Ctrl+L
Network from Table (Text/MS Excel)...
Network from Web Services...
Node Attributes...
Attribute from Table (Text/MS Excel)...
Edge Attributes...
Ontology and Annotation...
Attribute/Expression Matrix... Ctrl+E
Vizmap Property File...
Import attributes from Biomart...
Import Attributes from NCBI Entrez Gene...

Figure 14

3. Using the "Select File" button, navigate to the network file you want to import. Under the "Advanced" section, check "Show Text File Import Options", then check "Transfer first line as attribute names". Back up at the "Interaction Definition" section, select "Source Interaction" as column 1 and "Target Interaction" as column 2. Finally, to import the edge attribute, just click once on the "NumClusteredTogether" column in the preview section at the bottom—it should turn from gray to blue. Your screen should look like Figure 15 when you hit the "Import" button.



Import Network and Edge Attributes from Table

Import Network from Table

Data Sources

Input File: 3_source%20-%20inDevelopment/Example%20Input%20Files/consensus%20tutorial/300geneTC-tutorial_network.txt [Select File(s)]

Interaction Definition

Source Interaction: Column 1 — Interaction Type: Default Interaction — Target Interaction: Column 2

Columns in BLUE will be loaded as EDGE ATTRIBUTES.

Advanced

☑ Show Text File Import Options

Text File Import Options

Delimiter: ☑ Tab ☐ Comma ☐ Semicolon ☑ Space ☐ Other

Preview Options: ○ Show all entries in the file ● Show first 100 entries.

Attribute Names: ☑ Transfer first line as attribute names Start Import Row: 1 Comment Line:

Network Import Options: Default Interaction: pp

[Refresh Preview]

Preview

Text File — Left Click: Enable/Disable Column, Right Click: Edit Column

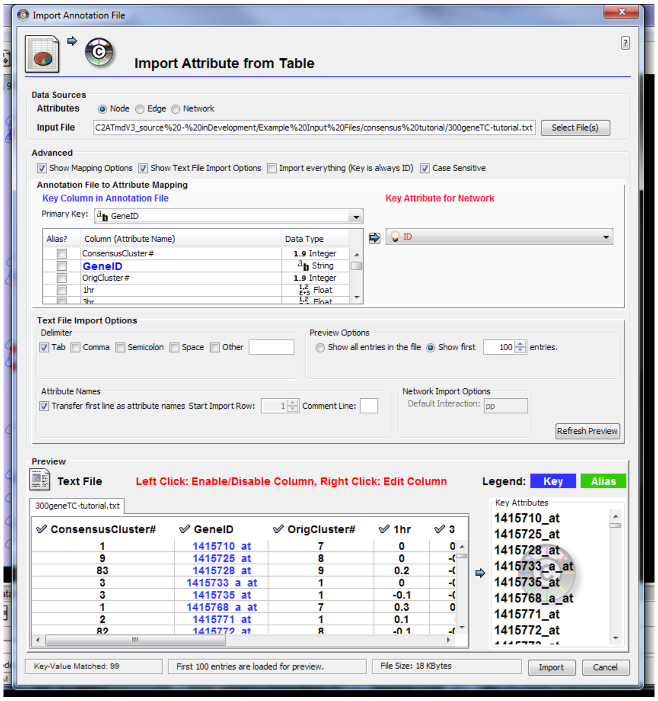
300geneTC-tutorial_network.txt

✓ ID1	✓ ID2	✓ NumClustere...
1415710 at	1415710 at	10
1415768 a at	1415710 at	10
1415841 at	1415710 at	10
1415975 at	1415710 at	9
1415976 a at	1415710 at	5
1416035 at	1415710 at	8
1416041 at	1415710 at	10
1416101 a at	1415710 at	10
1428767 at	1415710 at	9

[Import] [Cancel]

Figure 15

4. Before formatting the network, we need to load in the gene expression data for each node. To do this, go to *File>Import>Attribute from Table.*
5. In the "Input File" box, navigate to the consensus clustering output file. In this example that would be the file named "`300geneTC-tutorial.txt`". Under the "Advanced" section, check "Show Text File Import Options", then check "Transfer first line as attribute names". In addition, under the Advanced options, you need to check "Show Mapping Options", and select the GeneID as the Primary Key. Finally, make sure you don't have any duplicate headings. If you do, you can click on those columns to change the name or remove it from the import. For this example, I chose to remove the last set of time point columns from import (these were from expt2). Your screen should look like Figure 16 before pressing "Import".



Import Annotation File

Import Attribute from Table

Data Sources

Attributes: ● Node ○ Edge ○ Network

Input File: C2ATmdV3_source%20-%20inDevelopment/Example%20Input%20Files/consensus%20tutorial/300geneTC-tutorial.txt [Select File(s)]

Advanced

☑ Show Mapping Options ☑ Show Text File Import Options ☐ Import everything (Key is always ID) ☑ Case Sensitive

Annotation File to Attribute Mapping

Key Column in Annotation File — Key Attribute for Network

Primary Key: GeneID → ID

Alias?	Column (Attribute Name)	Data Type
☐	ConsensusCluster#	1..9 Integer
☐	GeneID	String
☐	OrigCluster#	1..9 Integer
☐	1hr	Float
☐	3hr	Float

Text File Import Options

Delimiter: ☑ Tab ☐ Comma ☐ Semicolon ☐ Space ☐ Other

Preview Options: ○ Show all entries in the file ● Show first 100 entries.

Attribute Names: ☑ Transfer first line as attribute names Start Import Row: 1 Comment Line:

Network Import Options: Default Interaction: pp

[Refresh Preview]

Preview

Text File — Left Click: Enable/Disable Column, Right Click: Edit Column — Legend: Key Alias

300geneTC-tutorial.txt

✓ ConsensusCluster#	✓ GeneID	✓ OrigCluster#	✓ 1hr	✓ 3
1	1415710 at	7	0	0
9	1415725 at	8	0	-(
83	1415728 at	9	0.2	-(
3	1415733 a at	1	0	-(
3	1415735 at	1	-0.1	-(
1	1415768 a at	7	0.3	0
2	1415771 at	1	0.1	
82	1415772 at	8	-0.1	-(

Key Attributes

1415710_at
1415725_at
1415728_at
1415733_a_at
1415735_at
1415768_a_at
1415771_at
1415772_at

Key-Value Matched: 99 — First 100 entries are loaded for preview. — File Size: 18 KBytes

[Import] [Cancel]

Figure 16

6. To view the whole network, you can resize your window, and click on the 1:1 magnifying glass in the tool bar to zoom out. As you can see, there are many edges. This network contains ALL edges regardless of the filtering level chosen.
7. Before filtering, however, I like to keep my original network in case I need to redo something. Therefore, we first need to make a copy of this network. In the menu bar, go to *File>New>Network>Clone Current Network* as shown in Figure 17. Right click on the new network to rename it. I like to name mine with the filtering level.
8. For the clustering, the 90% level was chosen, and there were 10 total clustering solutions (5 for each of the 2 data sets). To calculate the number of time 2 genes had to cluster together just multiple .9*10, or your chosen threshold percentage times the total number of clustering solutions generated. For this data, the cutoff is 9 (90% of 10). Therefore, we only want to keep edges with a "NumClusteredTogether" value of 9 or greater.
9. To do this, go to the "Filters" tab as shown in Figure 17. Under "Filter Definition", search for the option named "edge.NumClusteredTogether" and hit "Add". Then, under the "Advanced" section, double click on the blue bar and enter '9' and the low bound and '10' as the high bound. Check the "Not" box and click the "Apply Filter" button. This selects all edges with a score less than 9—which are the ones we want to delete.
10. Once your screen looks like Figure 18, press the delete button to delete all the red (selected) edges.

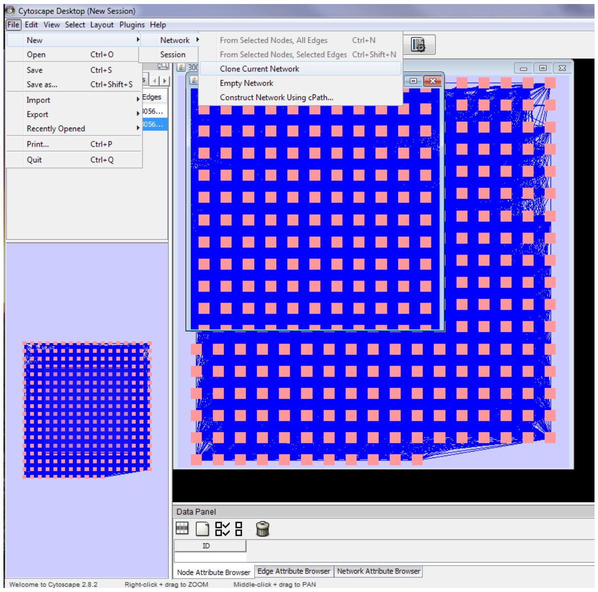
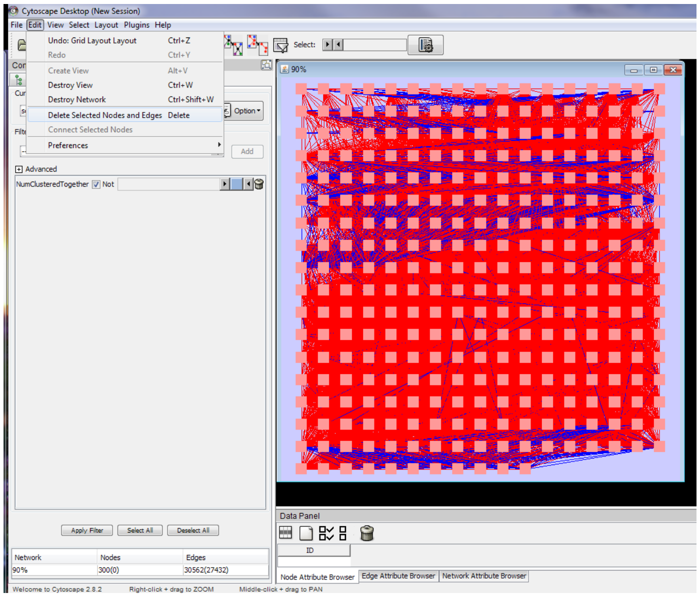

Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
New
Open Ctrl+O
Save Ctrl+S
Save as... Ctrl+Shift+S
Import
Export
Recently Opened
Print... Ctrl+P
Quit Ctrl+Q
Network
Session
From Selected Nodes, All Edges Ctrl+N
From Selected Nodes, Selected Edges Ctrl+Shift+N
Clone Current Network
Empty Network
Construct Network Using cPath...
Edges
Data Panel
ID
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.8.2
Right-click + drag to ZOOM
Middle-click + drag to PAN

Figure 17


Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
Undo: Grid Layout Layout Ctrl+Z
Redo Ctrl+Y
Create View Alt+V
Destroy View Ctrl+W
Destroy Network Ctrl+Shift+W
Delete Selected Nodes and Edges Delete
Connect Selected Nodes
Preferences
Select:
Option
Add
Advanced
NumClusteredTogether ☑ Not
90%
Apply Filter
Select All
Deselect All
Data Panel
ID
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.8.2
Right-click + drag to ZOOM
Middle-click + drag to PAN

Network	Nodes	Edges
90%	300(0)	30562(27432)

Figure 18

11. Now your network looks a lot thinner, but is still a mess to look at. Next, we need to format the layout and node colors to make this network look pretty.
12. To layout the nodes, simply go to the menu bar and select the Layout option. Choose any layout you want. For this tutorial we will use the *Cytoscape Layouts>Force Directed>unweighted* option. Your screen should now look something like Figure 19.

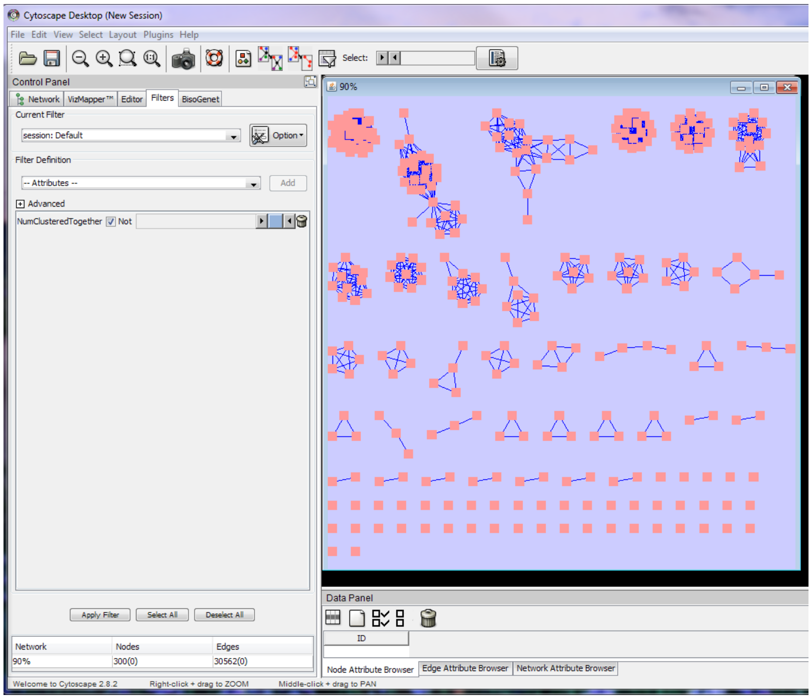


Figure 19

13. To format the node colors, go to the vizMapper Tab. This tab can be used to customize the look of your network to almost anything you can imagine; however, we will just be changing the node color in this example.
14. Under the "Visual Mapping Browser" section, scroll down and double click on the "Node Color" option.
15. Click on "Please select a value!" and choose the 24 hour time point. Under "Mapping Type", choose continuous. Double click on the black and white color bar to customize the colors. You should have a pop up that looks like Figure 20A. Adjust the colors so that positive is red, negative is blue, and 0 is white—it should look like Figure 20B.

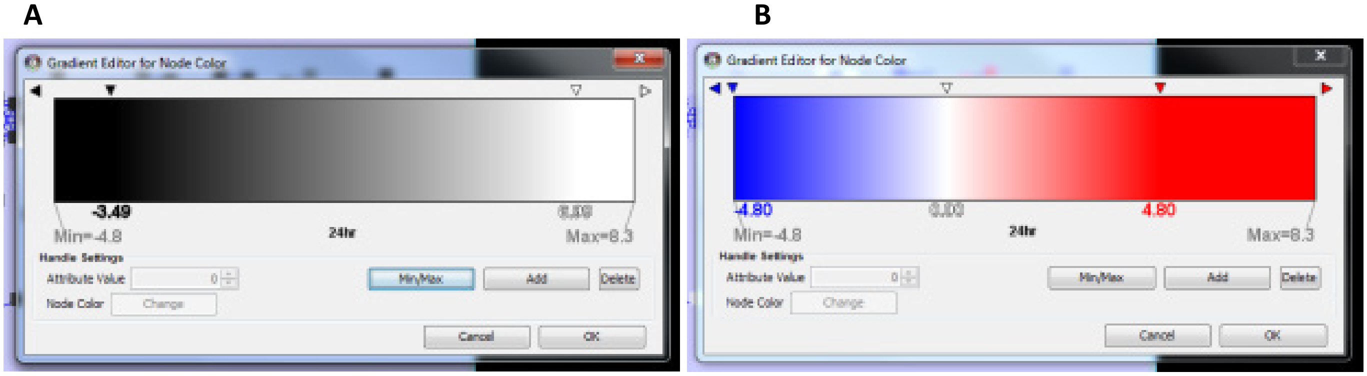
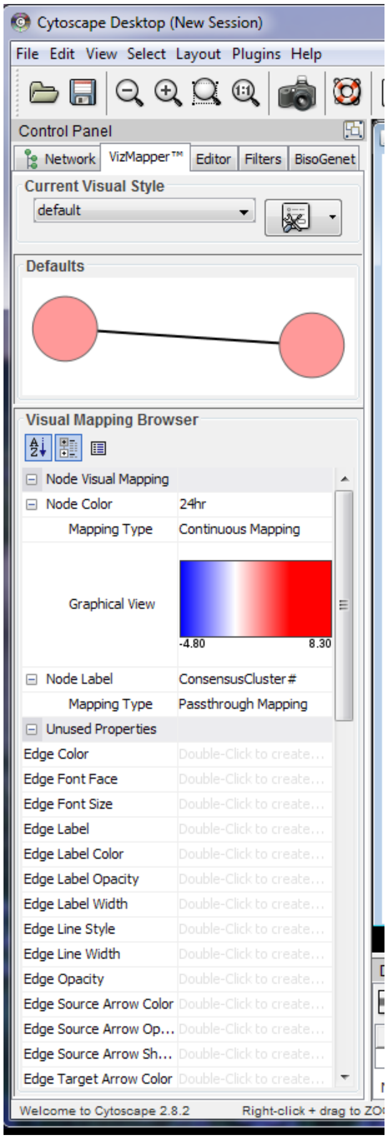


Figure 20

16. Finally, so we know which group of nodes represents which cluster in our heat maps, we need to change the "Node Label" to "ConsensusCluster#".
17. To preview what the printed network will look like, go to the menu bar and select *View>Show Graphic Details*. The node labels are small. To make them readable, and to change other global attributes, double click on the "Defaults" image at the top of the vizMapper tab. I chose to make the font size 30pt, edges color black and the background white. Figure 21 shows what the vizMapper tab should look like after all these changes.
18. The final network should look something like Figure 22.



Current Visual Style: default

Node Visual Mapping	
Node Color	24hr
Mapping Type	Continuous Mapping
Graphical View	-4.80 8.30
Node Label	ConsensusCluster#
Mapping Type	Passthrough Mapping
Unused Properties	
Edge Color	Double-Click to create...
Edge Font Face	Double-Click to create...
Edge Font Size	Double-Click to create...
Edge Label	Double-Click to create...
Edge Label Color	Double-Click to create...
Edge Label Opacity	Double-Click to create...
Edge Label Width	Double-Click to create...
Edge Line Style	Double-Click to create...
Edge Line Width	Double-Click to create...
Edge Opacity	Double-Click to create...
Edge Source Arrow Color	Double-Click to create...
Edge Source Arrow Op...	Double-Click to create...
Edge Source Arrow Sh...	Double-Click to create...
Edge Target Arrow Color	Double-Click to create...

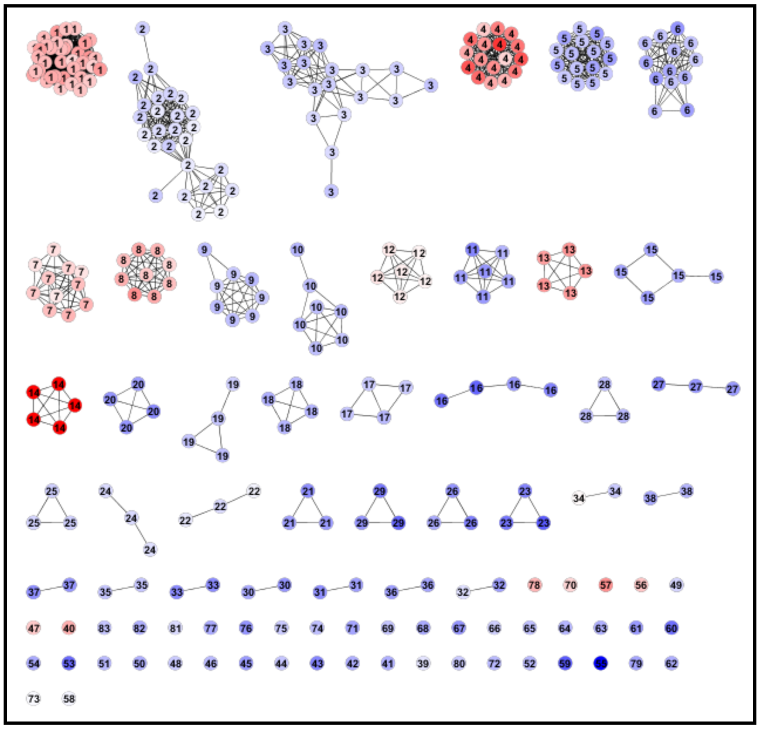


Figure 21 **Figure 22**

19. You can play around with visualization, make a figure for each time point and view them as a series, or create networks for different filtering threshold to see how the genes group together.
20. To export your network, click on the camera button in the tool bar. A variety of formats can be exported.
21. For more information of how to use Cytoscape, please visit www.cytoscape.org.

Walk-through: Heatmap Generation and Cluster Statistics

Under the Statistics and Heatmaps tab are two different functions: Generation of Heatmaps and Calculate Statistics. On-screen directions may be viewed by pressing the 'View Directions' button at the top right of the tab. First we will walk-through the generation of heatmaps followed by the calculation of cluster statistics.

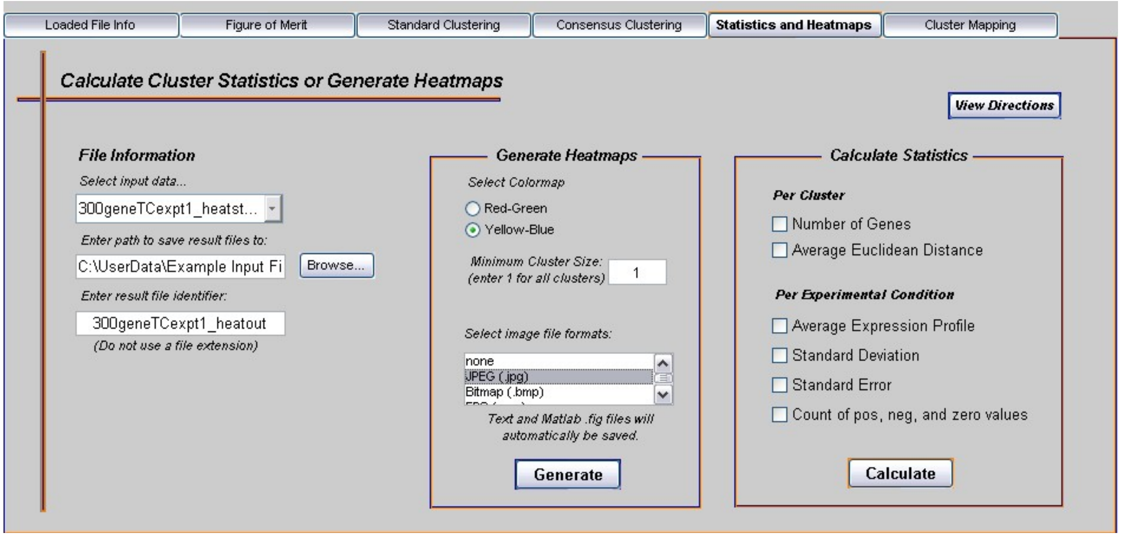
Heatmap Generation

This function is useful when you have a data set that is pre-clustered, but not visualized as a heatmap; or if you hand cluster the data based on functional information and such, and want it to be visualized in heatmap form.

This functionality can also be used to perform hierarchical clustering with Euclidean distance on a data set if the cluster assignment for all genes is set to 1.

To generate heatmaps, your data must first be in the proper format and loaded into the program. See the Input File Formats help file for a description of how to do this. Once the data is loaded, go to the Statistics and Heatmaps tab if you are not already there. Make sure the correct input file is selected, and then choose a destination for the results output and enter a file name that the results should be saved under. If no destination folder is indicated then the results will be saved to the same folder as the input file. Next, under the Generate Heatmaps panel, select the color scheme you would like (red-green or yellow-blue), enter a minimum cluster size, and choose any additional image file formats each heatmap should be saved as. As stated before, the minimum cluster size tells the program when to stop generating heatmaps. For example if you only want clusters with 5 or more genes in them, then you would enter 5 as the minimum cluster size. To include all clusters of any size enter a 1. Once all options have been set press the 'Generate' button at the bottom of the 'Generate Heatmaps' tab. The images will appear on the screen one by one. Wait until all images have been generated before closing any out.

For the example in this tutorial we took the file generated by the clustering algorithm discussed previously (`300genesTCexpt1_kmeansED.txt`), opened it in Excel, and reformatted it so that the columns were in the right order. This new file is named `300genesTCexpt1_heatstats.txt`; it will be used for this heatmap generation example and the generation of cluster statistics in the following section. Once all data is loaded and options set the screen should look like Figure 23:



Calculate Cluster Statistics or Generate Heatmaps

File Information

Select input data...
300geneTCexpt1_heatst...

Enter path to save result files to:
C:\UserData\Example Input Fi

Enter result file identifier:
300geneTCexpt1_heatout
(Do not use a file extension)

Generate Heatmaps

Select Colormap
○ Red-Green
● Yellow-Blue

Minimum Cluster Size:
(enter 1 for all clusters) 1

Select image file formats:
none
JPEG (.jpg)
Bitmap (.bmp)

Text and Matlab .fig files will automatically be saved.

Calculate Statistics

Per Cluster
□ Number of Genes
□ Average Euclidean Distance

Per Experimental Condition
□ Average Expression Profile
□ Standard Deviation
□ Standard Error
□ Count of pos, neg, and zero values

Figure 23: Setting options for heatmap generation.

If you are using the example files, once the Generate Heatmaps button is pressed 10 clusters should appear on the screen. These should be the same 10 clusters generated from the Standard Clustering analysis as we are using the same solution. Cluster heatmaps are automatically saved as .fig files, but additional image formats can be chosen. A text output file is generated that lists the order of genes in each heatmap for each cluster.

If you just want to cluster the data with hierarchical clustering and generate one heatmap that includes ALL genes (instead of one heatmap for each cluster like the Standard analysis does), then just assign all genes to be in cluster 1. This was done for the `300geneTCexpt1_heatstats.txt` file, and the changes were saved under the file name `300geneTCexpt1_1clust.txt`. When this file is run through the Generate Heatmap function, only one heatmap is generated that contains all genes in the data file hierarchically clustered with Euclidean distance. If using the example file, you should get the output shown in Figure 24, and an output text file containing the order of all genes in the heatmap.

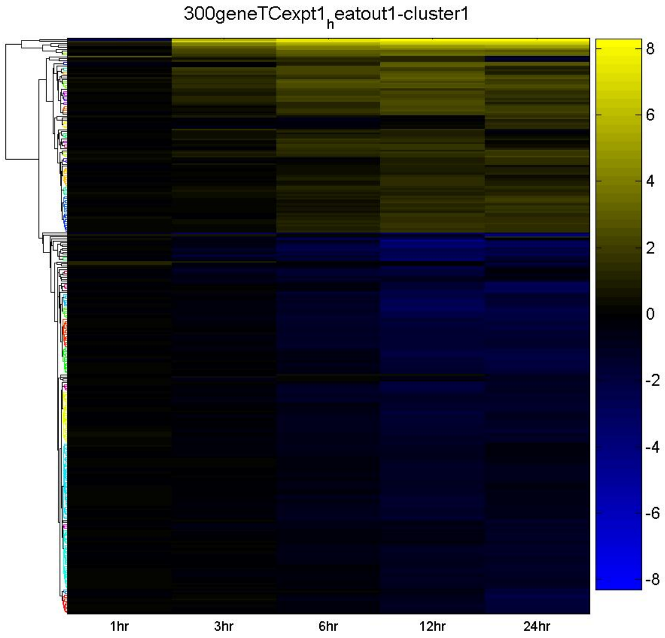


Figure 24: Example of a heatmap generated where all data was assigned to one cluster. This essentially just performs hierarchical clustering using Euclidean distance with the Clustergram function in the MATLAB Bioinformatics Toolbox.

Calculate Cluster Statistics

Calculating cluster statistics is very simple to understand, so not much attention will be paid to it here. Basically the input file must be pre-clustered and in the same format as the heatmap generation input file (the same one can be used). After entering in the input and output file information like with the Generate Heatmap function, just select the boxes beside the information you want calculated for each cluster, and then press the 'Calculate' button. A description of each available output in located in the Output File Formats help file.

Walk-through: Cluster Mapping

Cluster Mapping is an unusual technique that describes one clustering solution in terms of another. One example use of this function is, say you have a large data set with 1,000 genes that has been clustered. You take a subset of these 1,000 genes, say 300, and re-cluster this smaller subset (maybe the large set of genes contains known and unknown, and the small set is just known genes). The question asked is, 'When the smaller data set is re-clustered, how do the clusters change in relation to the clusters generated by the large data set?' In other words you want to know how many large data set clusters make up one small data set cluster, or vice versa.

The input data file for Cluster Mapping has a very specific format that is explained in the Input File Formats help file. The output of cluster mapping is also explained in the Output File Formats help file. Currently, the cluster mapping does not provide specific gene information for each cluster; however, this will be upgraded in future versions so that one may see exactly which genes are located in each cluster.

To run a mapping analysis load the properly formatted data file, then go to the Cluster Mapping tab. Select the input file, choose a destination folder for the results, and enter in a file name for the analysis results to be saved under. Press the 'Create Mapping' button to run the analysis.

References

Olex, A. L. and J. S. Fetrow (2007). "SCCATmd: Implementation and integration of the figure of merit with cluster analysis for gene expression data." manuscript in preparation.

Olex, A. L., E. M. Hiltbold, et al. (2007). "Application of novel filtering and cluster analysis techniques to a dendritic cell maturation time course microarray experiment." manuscript in preparation.

Olex, A. L., D. J. John, et al. (2007). Additional limitations of the clustering validation method figure of merit. 45th ACM Southeast Annual Conference, Winston-Salem, NC.

Smoot, M. E., K. Ono, et al. (2011). "Cytoscape 2.8: new features for data integration and network visualization." Bioinformatics **27**(3): 431-432.

Yeung, K. Y., D. R. Haynor, et al. (2001). "Validating clustering for gene expression data." Bioinformatics **17**(4): 309-318.