SC^{2}ATmd: Tutorial

Last updated on 12/4/07 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.

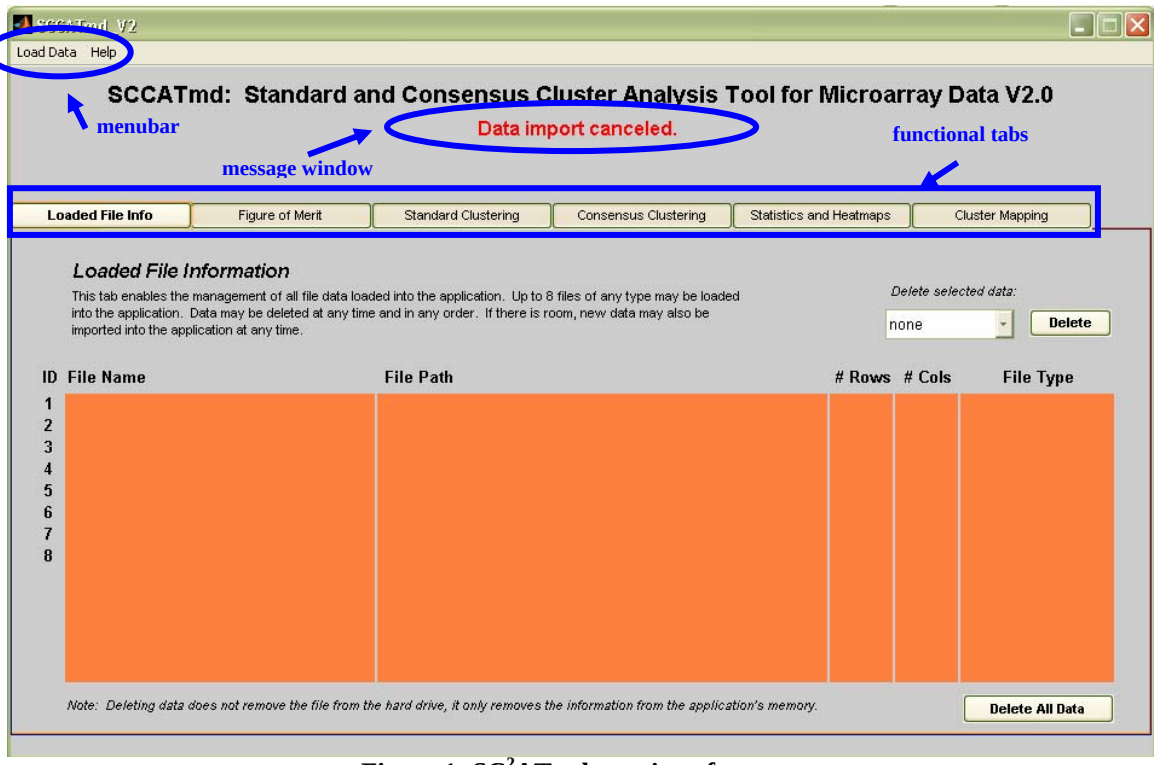

Figure 1: SC^{2}ATmd user interface.

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.

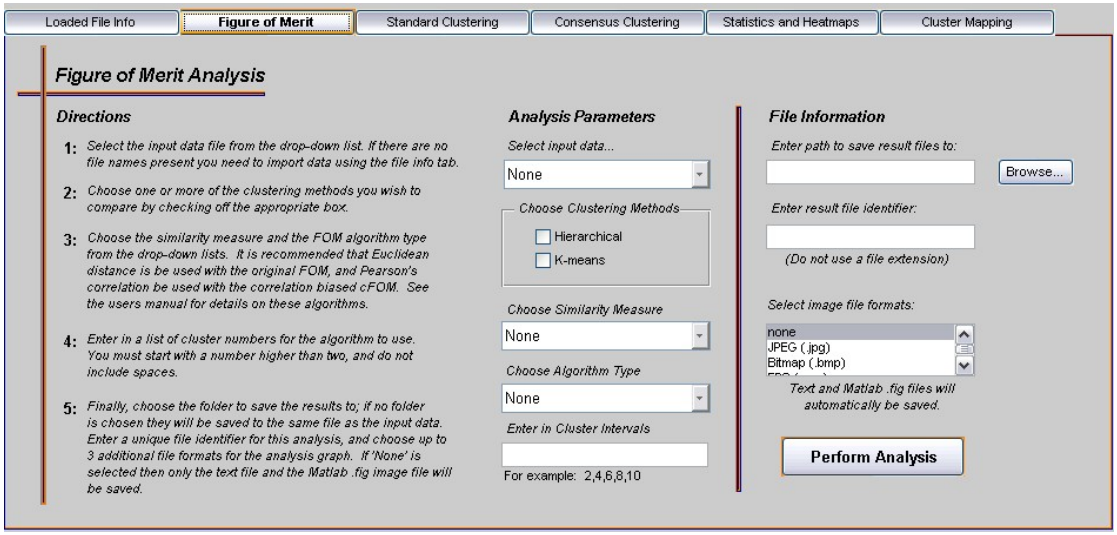

Figure 2: FOM Tab

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 top

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.

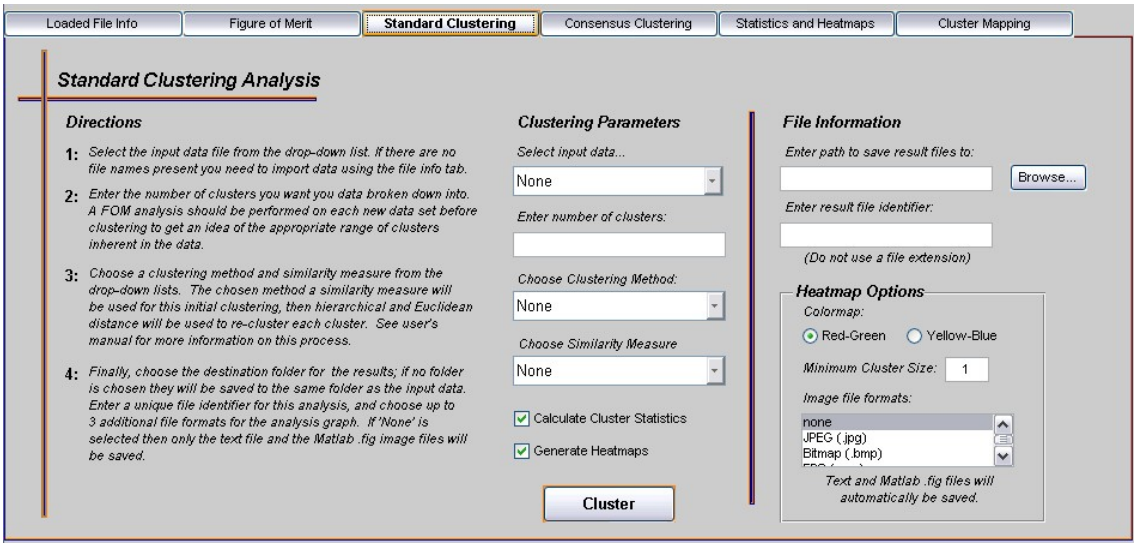

Figure 3: Standard Clustering Tab

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 or yellow-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

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.

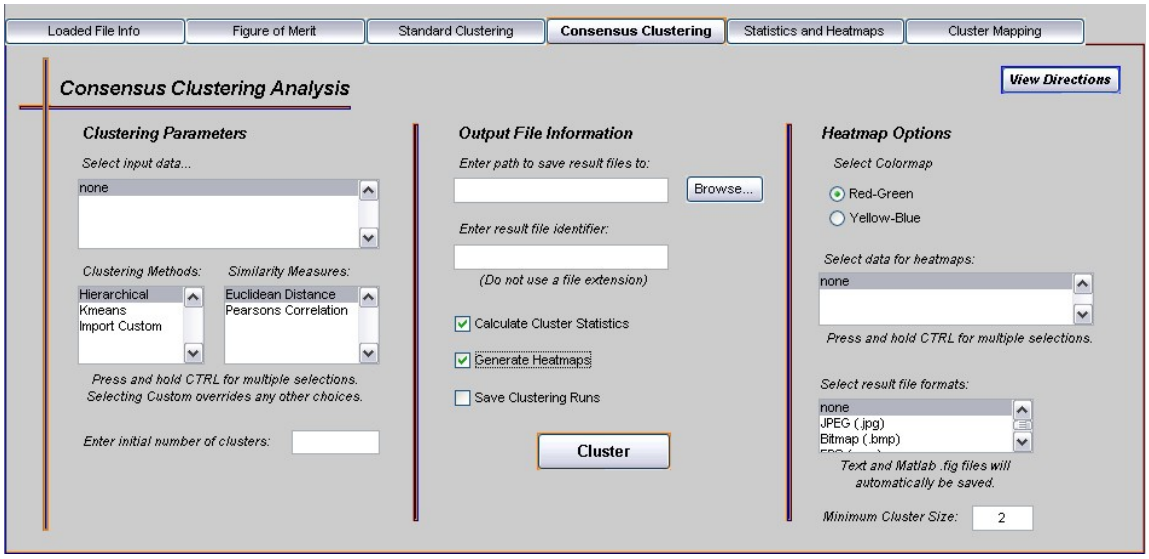

Figure 4: Consensus Clustering Tab

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 either selected 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 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 chose 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.

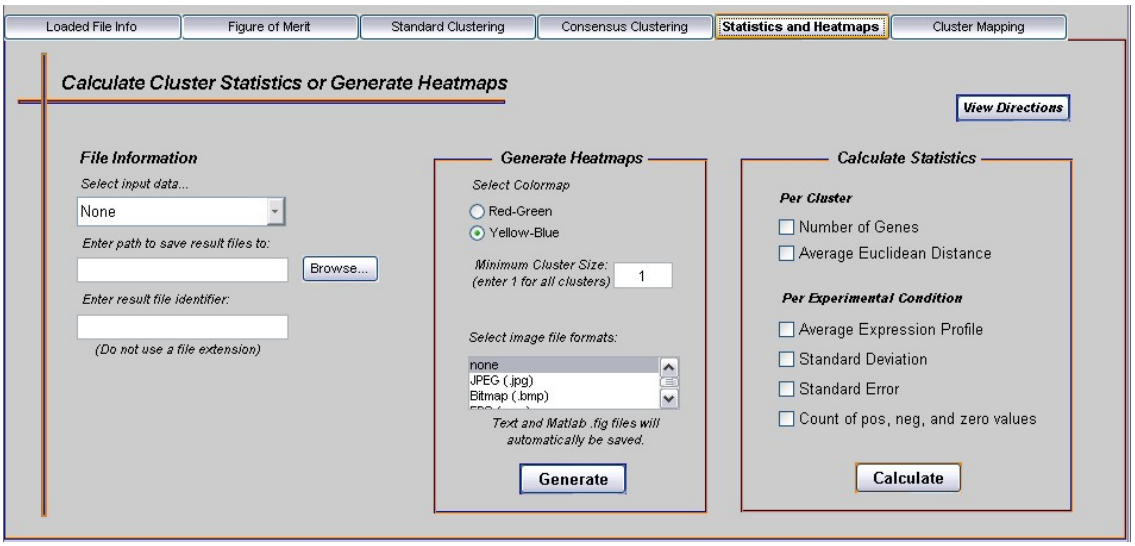

Figure 5: Cluster Statistics and Heatmap Tab

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.

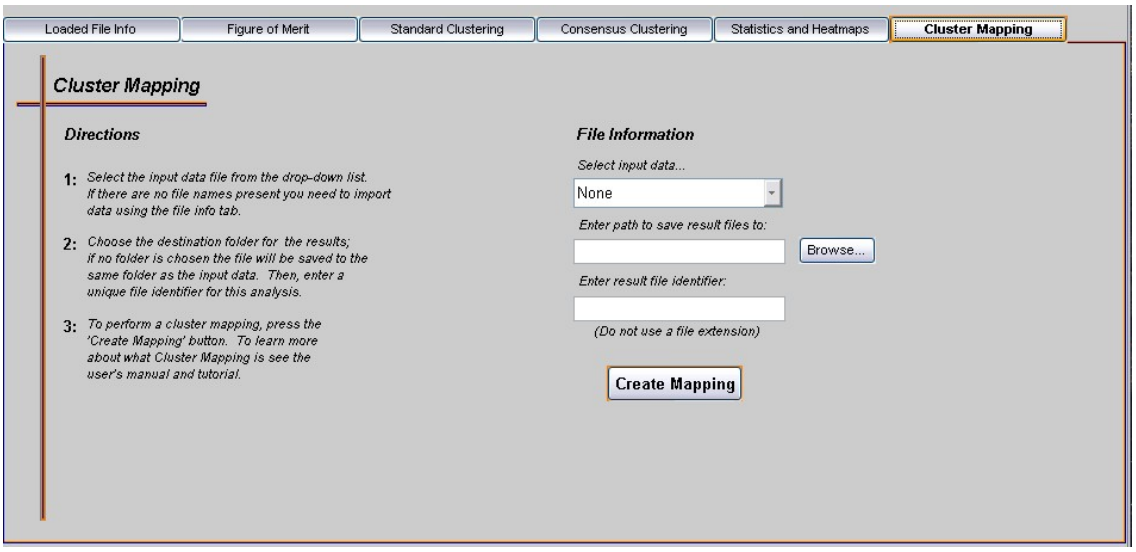

Figure 6: Cluster Mapping Tab

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.

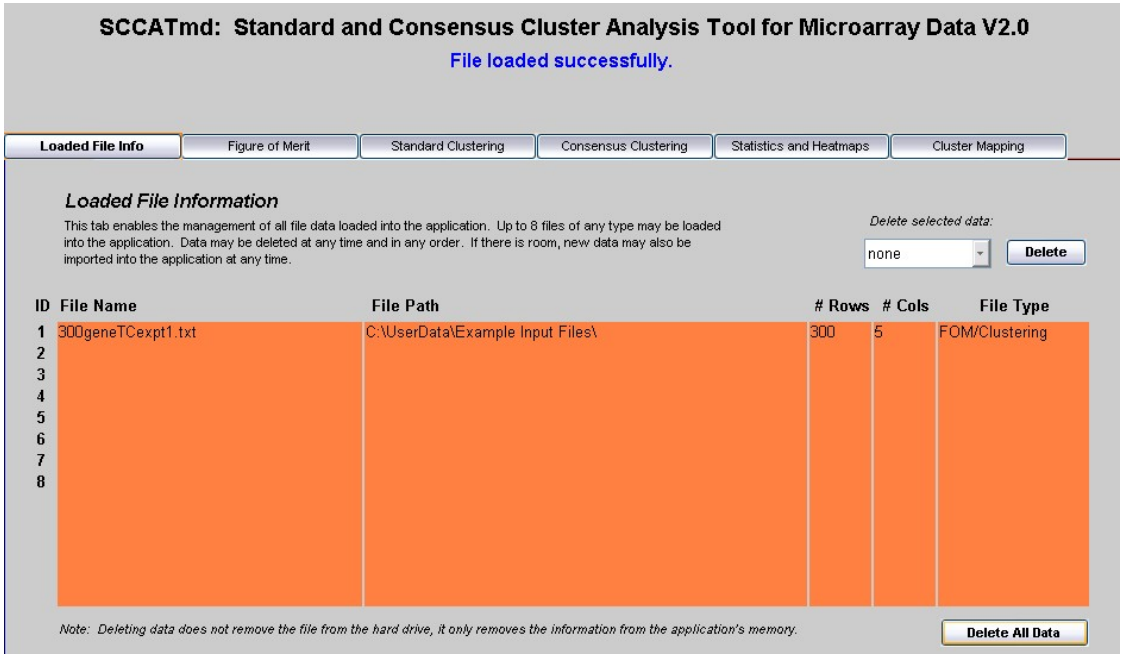

Figure 7: Updated file information.

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

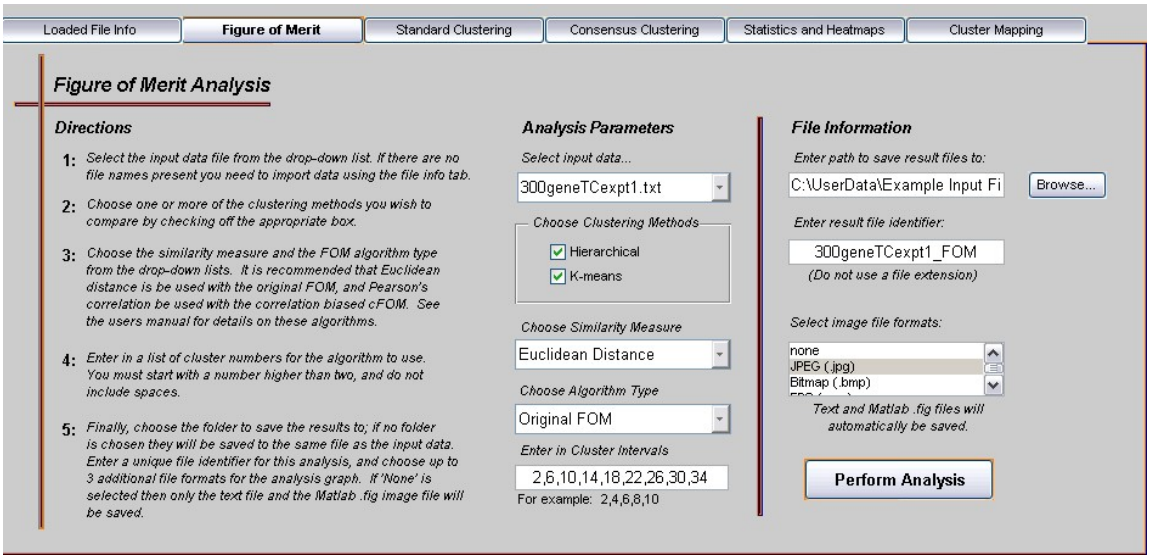

Figure 8: FOM tab with selected analysis options.

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).

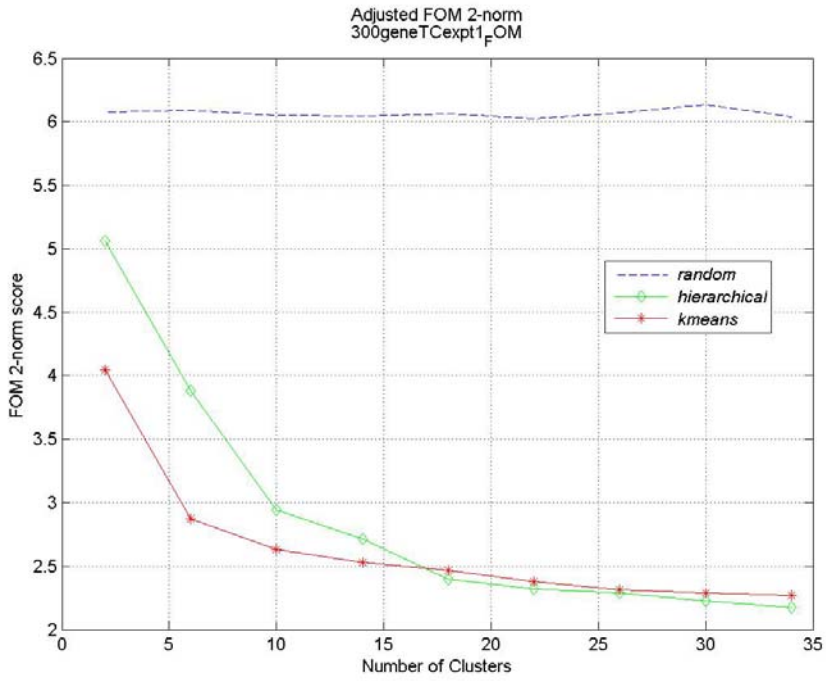

Figure 9: FOM graph output of analysis results.

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.

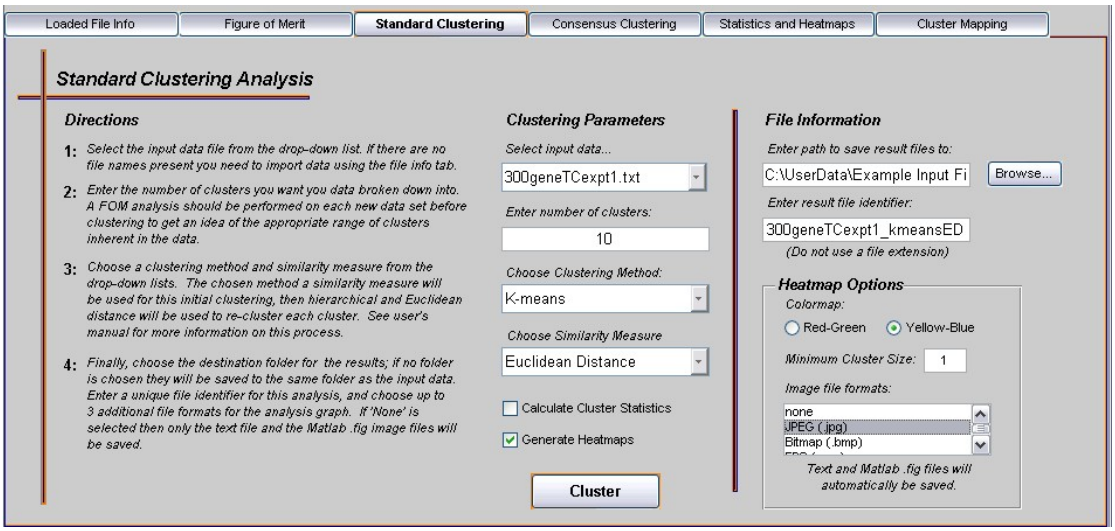

Figure 11: Standard Clustering parameters are set.

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.

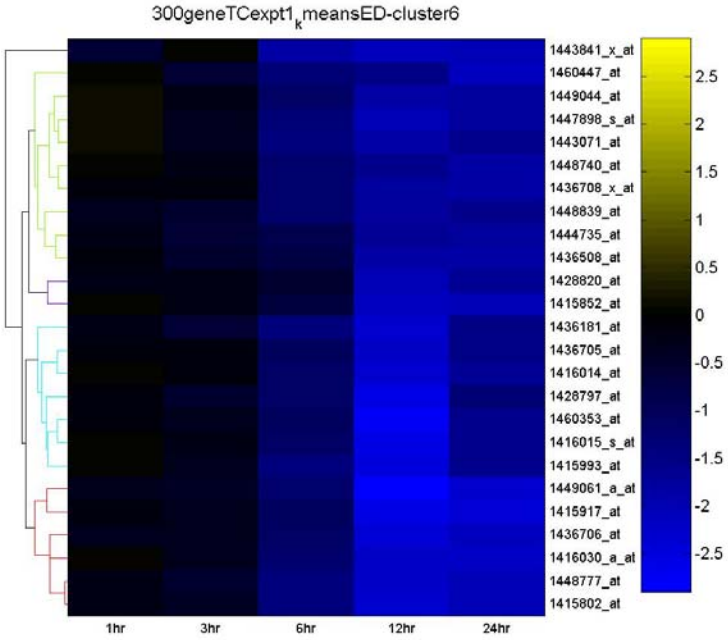

Figure 12: Example heatmap with dendrogram.

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

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 in all solutions. 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 the 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 in all solutions.

Tab Overview
Before we begin with the walk-through's 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: Two 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, and Pearson's correlation finds similar patterns of expression (for gene expression data) or shape of the data across all conditions. One or both of these measures may be used in the identification of consensus clusters. One may choose to use both if a consensus across two different similarity measures is wanted. Otherwise, only one is necessary.

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 2 similarity measures.
3. 2 clustering methods with 1 similarity measure.
4. 2 clustering methods with 2 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, 2 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, 2 similarity measures – This analysis can be done, however it is getting a little to 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. 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`. Make sure this file is 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 one is selected.
3. In the ‘clustering methods’ field select ‘kmeans’.
4. In the ‘similarity measures’ field select either one. 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. You should have noticed that when ‘kmeans’ was selected as the clustering method and 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. A higher number will generate more robust consensus clusters, but going too high could generate none at all. You can experiment with this on your own data sets to find a number that works well for your data.
7. 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.
8. Check the ‘Calculate Statistics’ and ‘Generate Heatmaps’ boxes.
9. In the Heatmap Options section choose the color scheme, make sure the one file is highlighted in the heatmap data box, 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.
10. Once all fields have been filled in your screen should look similar to that in Figure 13.
11. 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 9 or 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.

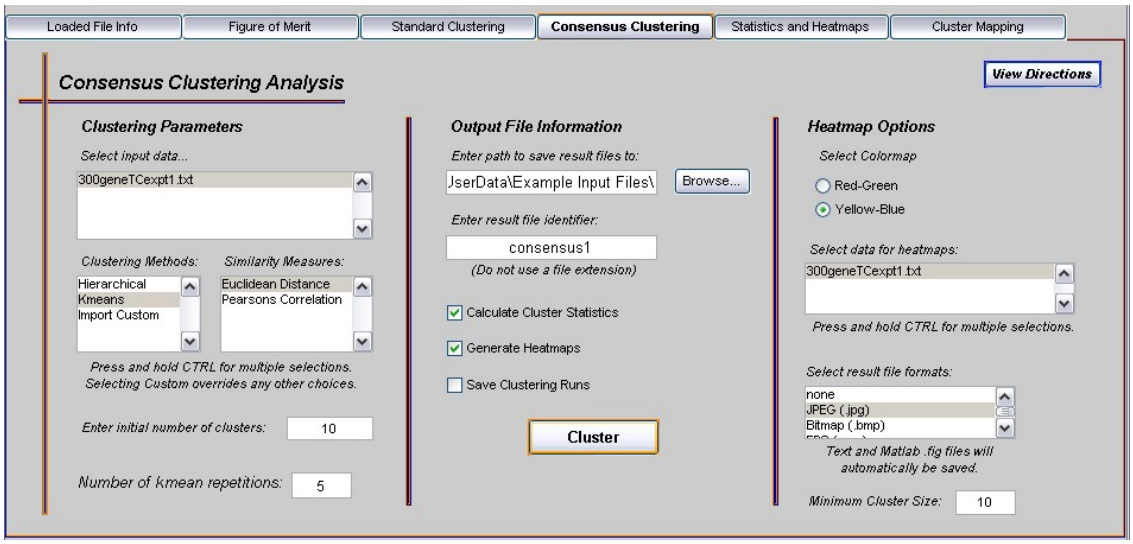

Figure 13: Consensus clustering with one input file

12. 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 `consensus1-300geneTCexpt1-ClusterStats.txt`. Global statistics such as total number of clusters, total number of singleton clusters, average cluster size, etc can be see. Scrolling down to the bottom you will notice that only clusters with at least 2 members have stats calculated.

13. The cluster results file (`consensus1-300geneTCexpt1.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.

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 simply 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 14:

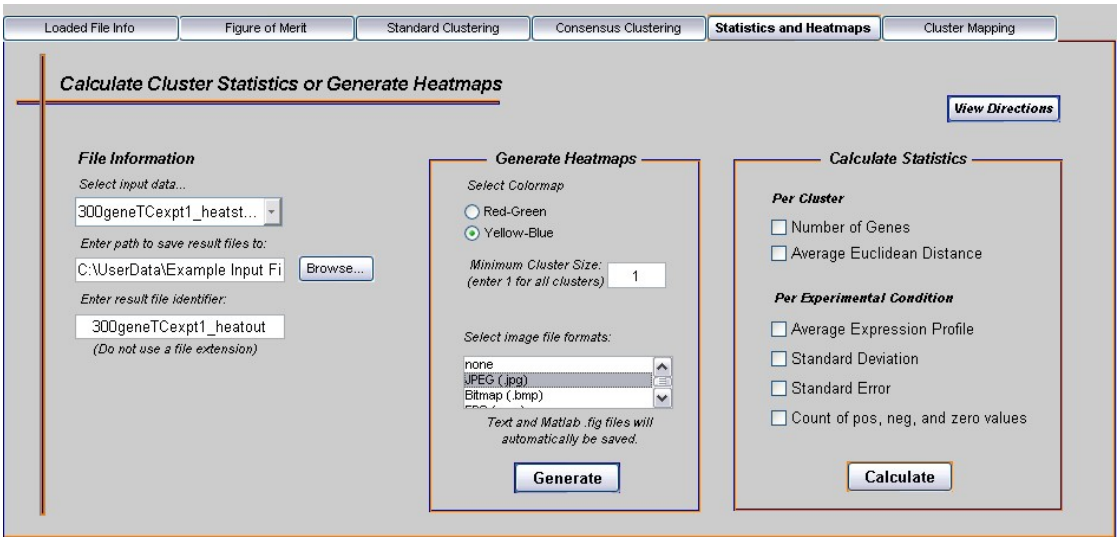

Figure 14: Setting options for heatmap generation.

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 15, and an output text file containing the order of all genes in the heatmap.

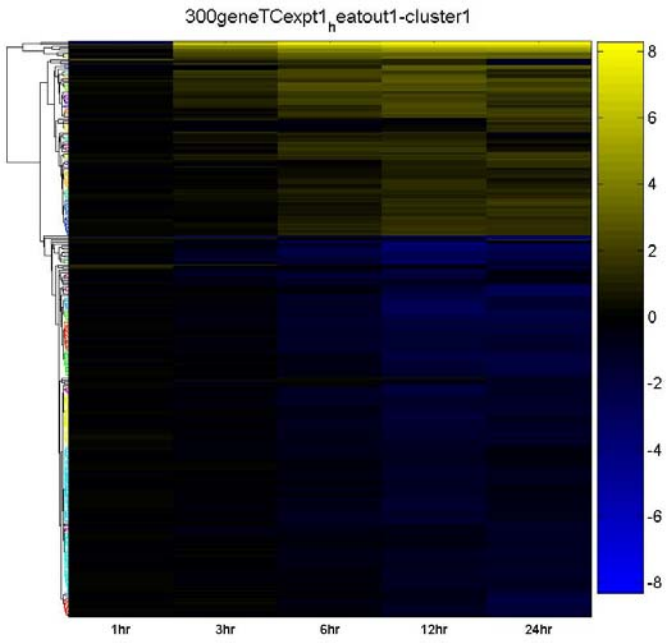

Figure 15: Example of a heatmap generated where all data was assigned to one cluster. This essentially just performs hierarchical clustering using Euclidean distance with the Clustergram function in the MATLAB Bioinformatics Toolbox.

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.

Yeung, K. Y., D. R. Haynor, et al. (2001). "Validating clustering for gene expression data." Bioinformatics **17**(4): 309-18.

Input File Formats

Last updated on 12/4/2007 by Amy Olex

Index of main import file types

A tab-delimited textfile is recommended for all input files as this is the default for Matlab's Import Wizard, however other standard delimiters may be used (such as CSV) if the default is changed during the file import process (see below).

Important Notes:

- ***There can be no missing or invalid data in any of the input files. If there is missing data, please remove these elements before performing any analyses.***
- ***All row labels MUST be unique. If there are duplicate row labels the application will not process your data correctly.***
- ***All row labels and column headers must have text elements in them (i.e. letters, punctuation, etc.), they cannot be all numeric as the data will not be imported properly (this will be fixed in future versions).***
- ***An incorrectly formatted file may be imported under any of the file types. It is up to the user to ensure the files are formatted properly. If an incorrect file format is loaded into the system and used for analysis, the analysis results will not be correct. In future versions of this application the file format will be checked prior to importation.***

FOM/Clustering File Format

top

The FOM/Clustering file format type is to be used for input into the Figure of Merit, Standard Clustering and Consensus Clustering analyses tabs.

The FOM/Clustering file format is illustrated below (Figure 1) where rows are data set features (e.g. genes, proteins, or any other element with measurements) and columns are the data set conditions (e.g. timepoints, cell types, etc.). The very first row always contains the column headers, and the first column always contains the feature/row labels (e.g. geneID's, protein name, etc.); all row labels and column headers must contain some type of text (i.e. cannot be all numbers). There can be any number of rows or columns in the data set as long as your computer has the memory to handle processing it.

Note: All labels must uniquely identify each row/column.

GeneID	1hr	3hr	6hr	12hr	24hr	
1444203_at		5.2	5.7	7	6.7	5.7
1450297_at		5.5	5.9	6.7	7.1	3
1418930_at		4.6	7.7	8	8.2	5.7
1429563_x_at		3.6	4.8	5.3	5.4	5.1
1436576_at		2.6	7.1	9.2	9.5	5.9
1439114_at		2.8	4.9	6.5	7.4	7.1
1442130_at		3	3.4	5.1	5.3	5.3
1449497_at		3.1	3.7	4.3	5.7	3.5
1450783_at		2.2	7.4	8	8.1	7.9
1452639_at		3.2	5	6.4	6.9	5.4
1419530_at		3.5	3.9	4.2	5.6	3.5
1422305_at		3.7	3.6	7.1	7.3	1.7
1423579_a_at		4.2	1.3	3.9	5	5.7
1434350_at		1.8	1.5	2	3	2
1437054_x_at		4	1	5.8	7.2	8.2
1440815_x_at		2.5	2.1	2.3	2.1	1.3
1444588_at		1.7	3.6	5.2	6	4.1
1445431_at		2.5	4.1	3.8	4	3.9
1447914_x_at		1.2	1.8	2.2	2.3	2.8
1448436_a_at		1.2	3.2	3.1	3.2	2.3
1449028_at		2.6	2.6	2.4	3.5	4.6
1450213_at		3.9	3.4	5	5.6	4.7
1450291_s_at		1.3	4.1	8.1	10.3	10.4
1451609_at		1.8	2.7	2.4	2.8	-0.3
1454043_a_at		1.1	1.4	4.2	6.1	5.9
1457404_at		4	1.9	3	2.1	-0.6
1457764_at		3.8	2.9	2.4	3.2	3.5
1459398_at		2.6	3.3	4.5	3.3	1.5

Figure 1: FOMAnalysis and Clustering input file format

Follow these steps to import data under the FOM/Clustering file format type:

1. Click on the 'Load Data' menu option and select 'FOM/Clustering'. See Figure 2.

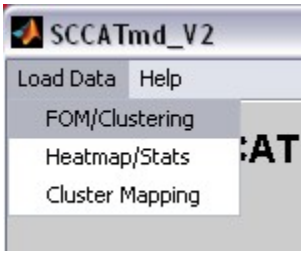

Figure 2: Loading FOM and clustering input files.

2. Use the file browser to find your file, then click 'Open'. See Figure 3.

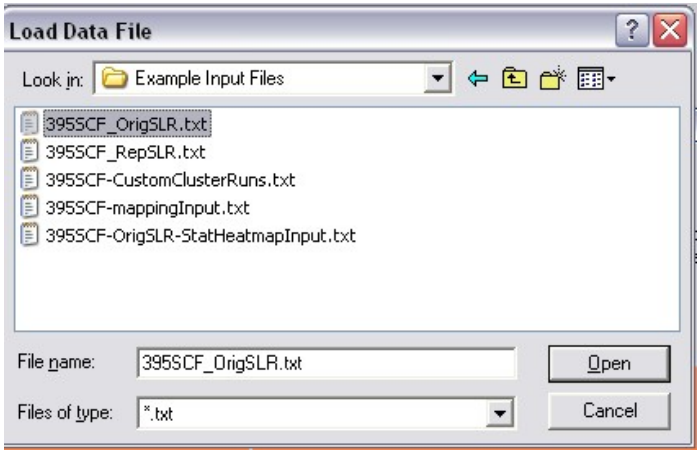

Figure 3: Browse to data file.

3. If the text file was not tab-delimited, choose the appropriate delimiter in the 'Select Column Separator' panel. Then check the preview window to make sure Matlab is reading your file correctly, and click 'Next'. See Figure 4

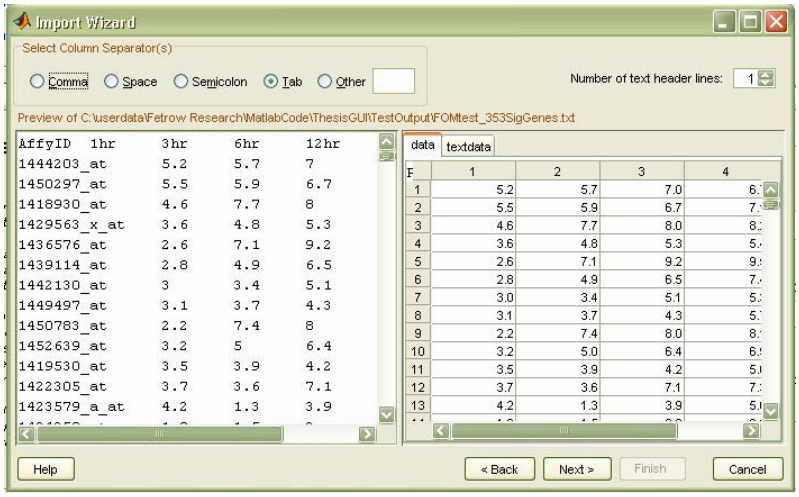

Figure 4: Matlab Import Wizard

4. If the file was loaded correctly there should be two variables listed in the window; data and textdata. If so, just click 'Finish'. If this is not the case make sure your source file is in the correct format, remove any unessesary white space and try to reload it. See Figure 5.

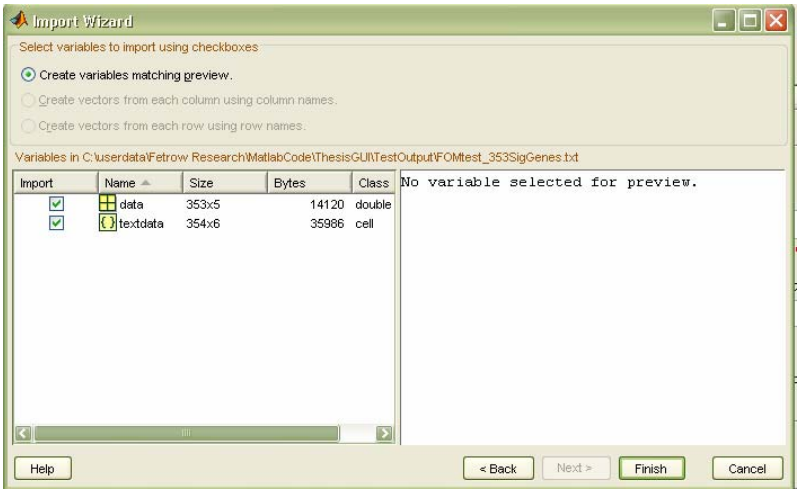

Figure 5: Matlab Import Wizard

5. After the file has been loaded into the GUI you should see its information appear on the File Info tab. See Figure 6.

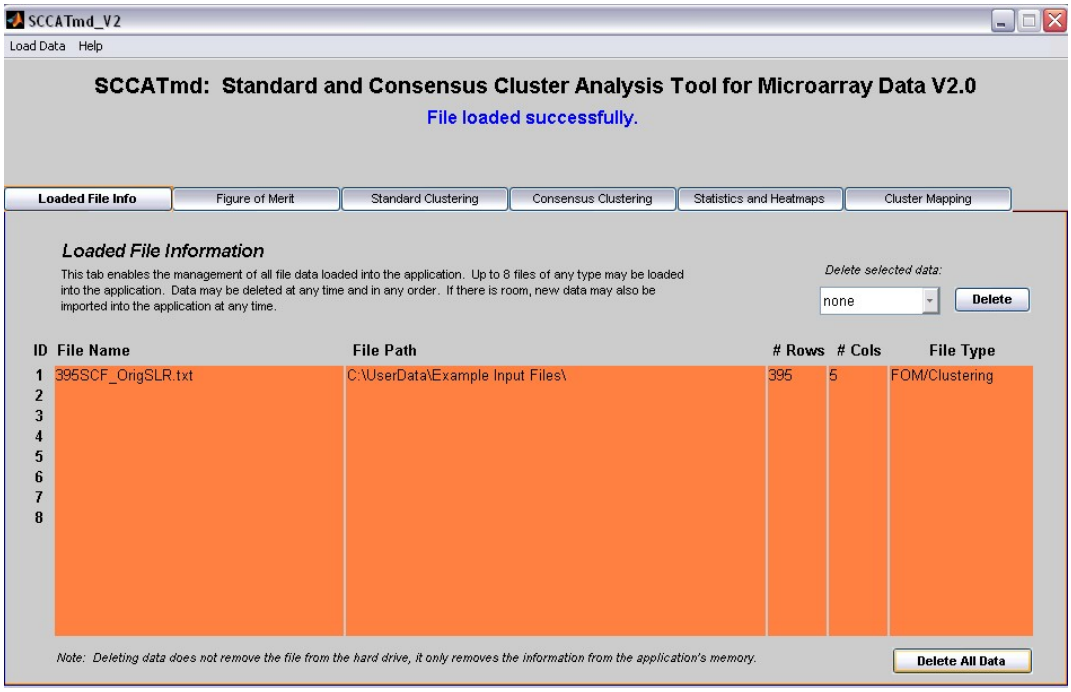

Figure 6: Updated file information on FileInfo Tab.

Heatmap/Stats File Format top

The Heatmap/Stats file format is used for the Heatmap Generation and Cluster Statistics tab.

The Heatmap/Stats file format is illustrated below (Figure 7). The first column contains the row labels where the rows are features (e.g. genes, proteins, etc.), the second column contains the cluster assignment for each row, and the rest of the columns are the experimental data for each condition (e.g. timepoints, cell types, etc.) with the first row containing the column headers.

Note: All labels must uniquely identify each row/column.

GeneID	cluster#	1hr	3hr	6hr	12hr	24hr
1436058_at	1	1.5	6.9	7.4	7.5	7.4
1424339_at	1	1.7	6.4	7.3	7.6	7.6
1450484_a_at	1	0.7	7.2	7.9	8.4	8.1
1450783_at	1	2.2	7.4	8	8.1	7.9
1421009_at	1	2.2	7.9	8.4	8.6	8.5
1418930_at	1	4.6	7.7	8	8.2	5.7
1436576_at	1	2.6	7.1	9.2	9.5	5.9
1449317_at	2	0.9	1.7	2.3	3	2.3
1450971_at	2	0.9	1.9	2.4	2.8	2.2
1448063_at	2	1	1.6	2.4	2.8	2
1449773_s_at	2	1.1	2	2.1	2.5	1.9
1447914_x_at	2	1.2	1.8	2.2	2.3	2.8
1442015_at	2	1.7	1.7	2.1	2.2	2.6
1434350_at	2	1.8	1.5	2	3	2
1432795_at	2	1.7	1.7	1.6	2.5	1.7
1459219_at	2	1.4	1.8	2.3	2.9	1.3
1449078_at	2	1.7	1.7	2.9	3	2.2
1425079_at	2	2	2.2	2.9	2.7	1.9
1444782_at	2	1.1	1.3	1.6	2	1.9
1449449_at	2	0.6	1.3	1.3	2.7	2.2
1437658_a_at	3	0.7	-1.1	-1.6	-2.4	-2.4
1454703_x_at	3	0.8	-0.8	-1.5	-2.8	-1.9
1457528_at	3	0.5	-0.4	-1.1	-2	-1.1
1438838_at	3	0.8	-0.3	-0.6	-4.4	-0.3
1460731_at	3	0.5	0.1	-0.1	-2.6	-0.6
1453683_a_at	3	0.8	-1.2	0.3	-3.3	-4
1437917_at	3	1.3	0.6	0.2	-0.3	-2
1453431_at	3	1.1	0.5	1.3	-0.5	-2.6

Figure 7:Cluster statistics and heatmap generation input file format

Follow these steps to import data for cluster statistics and heatmap generation:

1. Click on the 'Load Data' menu option and select 'Heatmap/Stats'. See Figure 8.

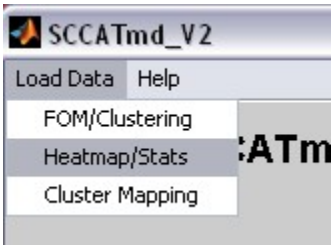

Figure 8: Loading Heatmap and Stats input file.

2. Follow steps 2-4 of the FOM/Clustering directions listed above.
3. After the file has been loaded into the GUI the file information should be updated on the File Info tab. See Figure 9.

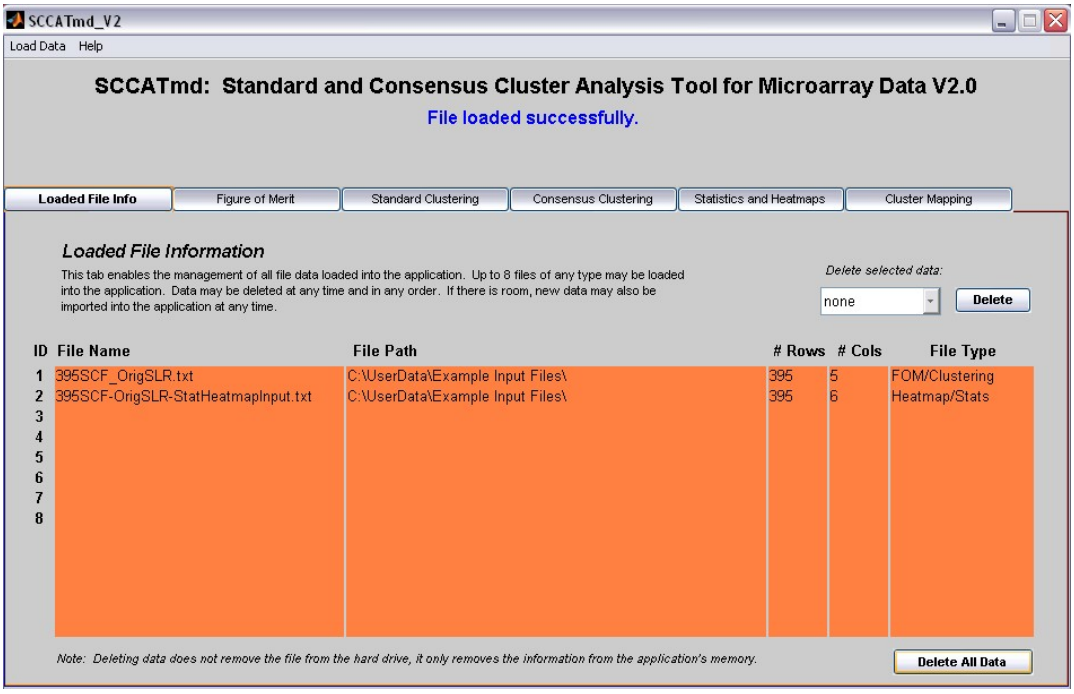

Figure 9: Updated file information

Cluster Mapping File Format top

The Cluster Mapping file format is used for the Cluster Mapping tab only.

The Cluster Mapping file format is illustrated below (Figure 10) where the rows are features (e.g. genes, proteins, etc.), the first column contains the first clustering solution with cluster assignments for each gene, and the second column contains the second clustering solution with another set of cluster assignments for the same genes. The file must be sorted in ascending order by the first clustering solution, and then sorted in ascending order by the second. In other words, the first solution in column 1 is sorted all the way; then for each sorted cluster of the first solution, the corresponding cluster assignments in the second solution are sorted in ascending order. This ordering can easily be done in Excel with the Sort function under Edit -> Sort.

Note: All labels must uniquely identify each row/column.

GeneID	Solution1	Solution2
1437218_at	1	2
1437658_a_at	1	2
1437917_at	1	2
1438527_at	1	2
1420579_s_at	2	1
1423175_s_at	2	1
1441302_at	2	1
1442157_at	2	1
1442830_at	2	1
1447989_at	2	1
1434350_at	2	4
1440815_x_at	2	4
1447914_x_at	2	4
1451609_at	2	4
1419530_at	3	1
1423579_a_at	3	1
1445431_at	3	1
1448436_a_at	3	1
1459973_x_at	3	1
1460605_at	3	1
1422408_at	3	3
1439310_at	3	3
1445840_at	3	3
1434152_at	3	3
1444203_at	4	3
1450297_at	4	3
1418930_at	4	3
1429563_x_at	4	3

Figure 1: Cluster mapping input file format

Follow these step to import data for the cluster mapping function in the HeatmapGeneration tab:

1. Click on the 'Load Data' menu option and select 'Cluster Mapping'. See Figure 11.

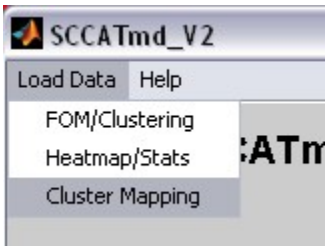

Figure 11: Loading Cluster Mapping input file.

2. Follow steps 2-4 of the FOM/Clustering directions listed above in the Figure of Merit and Cluster analysis section.
3. After the file has been loaded into the GUI the file info should be updated in the File Info tab. See Figure 12.

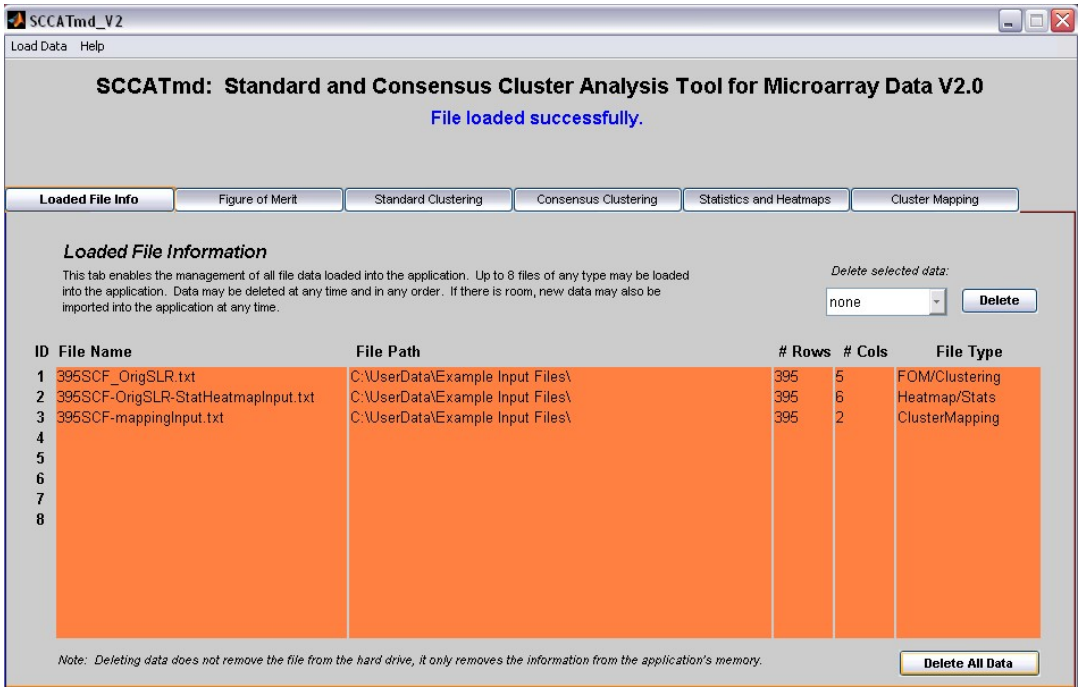

Figure 12: Updated file information

Other File Types

top

Consensus Clustering Custom Import file:

The custom import file under the Consensus Cluster tab is formatted like the Cluster Mapping file format, and is used to import custom clustering solutions for the extraction of consensus clusters. As with all the other import files, the first column contains row labels and the first row contains column headers. The data portion of the matrix is similar to the Cluster Mapping format except it does not have to be ordered and there can be any number of clustering solutions. All clustering solutions must contain the same number of total clusters, must have the same number of rows, and must not have any missing data. See Figure 13.

GeneID	1Run1	1Run2	1Run3	1Run4	1Run5	2Run1	2Run2	2Run3	2Run4	2Run5
1415802_at	8	7	8	3	8	2	4	3	3	3
1415829_at	4	3	8	3	3	2	4	8	3	1
1415917_at	6	7	6	3	2	2	4	8	3	3
1415922_s_at	1	5	1	7	4	8	8	4	2	4
1415945_at	4	3	8	3	3	2	4	8	3	3
1416014_at	8	7	8	3	8	2	4	8	3	1
1416015_s_at	8	7	8	3	8	2	4	8	3	1
1416016_at	5	4	4	5	6	6	5	4	8	4
1416123_at	1	5	5	7	4	8	8	4	2	4
1416150_a_at	4	3	8	3	3	2	6	3	3	3
1416151_at	4	3	8	3	3	2	6	8	3	3
1416152_a_at	4	3	8	3	3	2	6	3	3	3
1416221_at	1	8	2	1	4	8	7	4	7	4
1416283_at	8	7	8	3	8	2	4	8	3	1
1416333_at	8	7	8	3	8	2	4	8	3	3
1416380_at	5	1	4	5	6	6	5	4	8	4
1416653_at	1	8	5	4	4	8	8	4	2	4
1416684_at	6	7	8	3	8	7	4	2	6	5
1416685_s_at	4	3	8	3	3	7	4	2	6	5
1417057_a_at	8	7	8	3	8	7	4	2	6	5
1417172_at	5	4	4	5	6	6	5	4	8	4
1417185_at	2	1	4	5	5	3	5	5	8	6

Figure 13: Consensus Clustering Custom Import file format

Output File Formats

Last Updated on 12/4/2007 by Amy Olex

Index

This section describes the format and content of all the output files for each analysis function. Output files come in 3 types: text files, image files, and Matlab .fig files.

Important Notes:

- ***All output will be saved in the same location as the input file unless the user specifies another destination.***

Figure of Merit Analysis top

By default the figure of merit analysis produces a text file (.txt) and an image file (.fig). The image may be saved in additional image file formats according to user preference. A description of each follows.

Text file: *myfom.txt*

A summary of the FOM analysis is output in a text file that is similar to Figure 1.

```
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:  9.02    5.61  5.31  5.14  4.81  4.59  4.49  4.42  4.30

K-means
Clusters:     2       6    10    14    18    22    26    30    34
FOMscores:  7.12    5.44  4.95  4.68  4.57  4.47  4.36  4.30  4.24

Random
Clusters:     2       6    10    14    18    22    26    30    34
FOMscores: 10.65   10.64 10.58 10.69 10.67 10.57 10.68 10.56 10.66

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

Figure 1: FOM text output file.

The first line identifies the figure of merit version that was used. If Euclidean distance was chosen as the similarity measure, then the original Euclidean-biased version of the FOM will be used, else if Pearson's correlation coefficient is used then the correlation-

biased FOM will be used. The second line reiterates the list of cluster numbers the user entered. The third line indicates the clustering algorithm that performed the best on this data; this is the recommended clustering algorithm. The next three sections list the sequence of FOM scores for each clustering algorithm chosen. This is followed by the identification of the optimal number of clusters to use with the input data set.

Image files: *myfom.fig, .jpg, .tiff, .bmp, .eps, .ai and .pdf*

For each iteration of the FOM algorithm the scores are plotted on a graph that is automatically saved as a .fig file; the other image file formats may also be generated if the user chooses. In Figure 2 the x-axis lists the number of clusters used in each iteration of the FOM algorithm, and the y-axis is the FOM score. Each line on the plot indicates the series of FOM scores calculated for a clustering algorithm.
The .fig file can be opened and manipulated in Matlab; otherwise the other image files can just be imported into documents as-is.

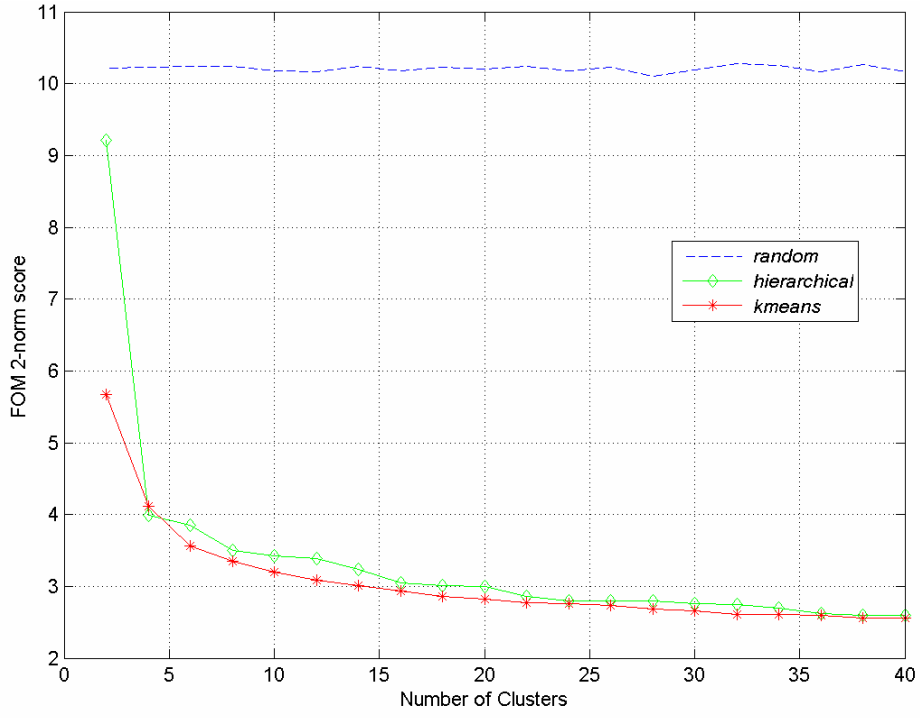

Figure 2: FOM graph output.

What is a FOM analysis?
The FOM analysis is a quantitative analysis that compares the performance of different clustering algorithms on a set of data. The clustering method that gets the lowest FOM score creates the most homogeneous clusters, and is therefore the best method. The FOM analysis reveals which clustering algorithm is best suited for each data set, and it gives a range for how many clusters are optimal. For more details on what the FOM is see (Yeung, Haynor et al. 2001).

Determining the Optimal Clustering Method

The lower the FOM score, the better; therefore the line closest to the bottom of the graph is the optimal clustering method. This program calculates the average FOM score for each method, and the method with the lower average is chosen as optimal; this information can be found in the text output file discussed previously.

Choosing the Optimal Number of Clusters
The number of clusters is chosen based on where the 'elbow' in the graph is. This 'elbow' indicates that increasing the number of clusters is not improving the overall cluster homogeneity, so the FOM score is not improving much and is flattening out. The optimal number of clusters is indicated in the text summary file; any number of clusters that fall in the enclosed range are acceptable.

Customizable Graph Features (Matlab users)
Matlab's .fig file gives the user the capability to customize the look of the FOM plot. Matlab tutorials can be found on the web; below is a short, non-comprehensive list of customizable graph features.

- Font and font size
- Grid lines on/off
- Axis labels
- Title and Legend
- Axis markers
- Line color, shape and size
- Data marker color, shape and size
- Background color
- Graph dimensions

Standard and Consensus Clustering Analysis top

This section describes the output generated by the standard clustering and the consensus clustering analyses. For each of these analyses one text output file is generated automatically. The user has the option of also creating a text file containing the cluster statistics, and image files representing each cluster as a heatmap.

The standard clustering algorithm performs the selected clustering method on the loaded data set once, and generates a text file with all clustering results. If the ‘Generate Heatmap’ option is chosen, one heatmap with dendrogram will be generated for each cluster and saved as a Matlab .fig file; other image file formats may also be selected. If hierarchical clustering is selected a global dendrogram is output in addition to the heatmaps as a .fig and .jpeg file; this dendrogram relates each cluster to the others so that the entire hierarchical tree can be reconstructed if desired.

The consensus clustering algorithm performs the selected type of consensus clustering (see Tutorial), and generates a text file with all clustering results. If the ‘Generate

Heatmap' option is chosen, one heatmap for each cluster of appropriate size (see Tutorial) will be generated and saved as a Matlab .fig file; other image file formats may also be chosen. The consensuses clustering also lets the user save the multiple clustering runs that were used to extract the consensus clusters as a text file.

Text file: *myclusters.txt*

Figure 3 is an example of the standard and consensus clustering output file if heatmaps are also generated. If heatmaps are not generated, then the second column, 'clusterOrder' will not be included.

cluster#	clusterOrder	AffyID		1hr	3hr	6hr	12hr	24hr
1	1	1455581_x_at	0.8	2.3	3.1	3.2	2.4	
1	2	1436172_at	0.9	2.4	3.1	3.2	2.5	
1	3	1446090_at	0.9	2.3	3.4	3.9	2.6	
1	4	1448436_a_at	1.2	3.2	3.1	3.2	2.3	
1	5	1432548_at	1.3	3.1	3.7	3.4	2.6	
1	6	1446457_at	1.1	3	3.7	3.3	2.4	
1	7	1450446_a_at	0.7	3	3.4	3.3	1.9	
1	8	1458512_at	1.4	2.8	3	4.2	1.9	
2	1	1459659_at	3.1	1.8	0.3	0.4	-0.3	
2	2	1460312_at	3.1	1.5	0.5	0.4	-0.1	
2	3	1443392_at	2.3	2	0.1	1	-0.2	
2	4	1447301_at	2.1	3.1	0.8	0.2	-0.9	
2	5	1458202_at	2.2	2.4	1	0.1	-0.1	
2	6	1459147_at	2.1	3.6	0	0.2	0	
2	7	1457235_at	2.2	2.2	-1.6	0.6	0.8	
2	8	1445471_at	3.5	0.4	0.8	0.9	0.7	
2	9	1456720_at	4.4	1	1.2	0.6	-0.1	
2	10	1423175_s_at	3.7	3.8	0.3	1.9	2.9	
2	11	1432808_at	3	3.7	1.3	1	1.9	
2	12	1432904_at	3.4	2.9	0.3	0.6	1.5	
2	13	1443789_x_at	2.3	2.5	0.7	0.6	2.2	
2	14	1459044_at	2.2	2.8	0.6	0.5	1.8	
2	15	1450823_at	2.4	2.6	0.4	1.4	1.7	
3	1	1437754_at	2	1.5	2.7	1.8	2.2	
3	2	1458737_at	1.6	2	2.9	1.7	2.5	
3	3	1443694_at	0.8	1.1	2.6	1	2.7	
3	4	1446990_at	0.8	0.5	3	1.2	3.2	

Figure 3: Standard and Consensus Clustering text output file.

Text file description:

- Column 1: This column contains the cluster assignments for each gene in the file.
- Column 2: This column contains the order of genes in each cluster on the heatmaps.
- Column 3: The unique gene labels provided by the user.
- Columns 4-n: The imported data for each gene that was used to generate the clusters.

Heatmap Image files: *myclusters-clusterX.fig*
The standard and consensus clustering algorithms provide the option to generate one hierarchically clustered heatmap for each cluster. The image files may be used as-is, or the Matlab .fig file may be customized by advanced Matlab users. An example heatmap and dendrogram are shown in Figure 4.

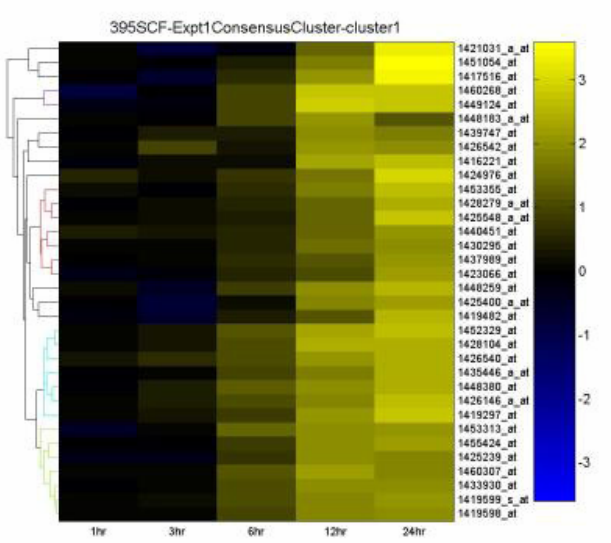

Figure 4: Clustering heatmap output.

Each column of the heatmap represents an experimental condition, and each row is one gene. Column labels are located at the bottom, row labels on the right, and the file name is at the top. The dendrogram to the left relates each gene to the others that are within the same cluster where the height of each branch indicates how similar two genes or groups of genes are (shorter branches indicate a stronger similarity). The color index is located on the right; either a red-green or yellow-blue color map can be selected where red and yellow indicate an increase from the control, green and blue represents a decrease from the control, and black is no change.

Global Dendrogram files: *myclusters-gden.fig and myclusters-gden.jpg*

This unique implementation of the traditional hierarchical clustering algorithm results in the complete hierarchical tree being divided into the pre-selected number of clusters. To relate the individual clusters, the 'top' of the complete dendrogram is saved as a .fig and jpeg file, so that the user is able to relate each reported cluster and reconstruct the entire tree. An example dendrogram is shown in Figure 5 where each numbered leaf represents a cluster of genes for which a heatmap was created as above.

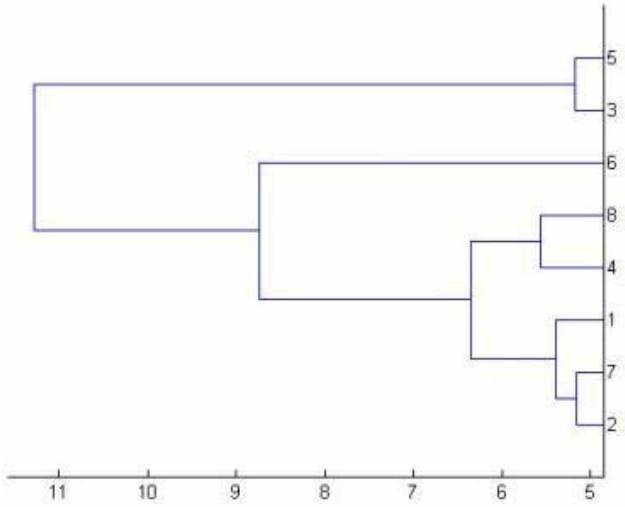

Figure 5: Hierarchical clustering global dendrogram output.

Cluster Mapping Output top

The cluster mapping function outputs one tab-delimited text file that describes one clustering solution in terms of another. An example output file is shown in Figure 6

```
Solution1   Solution2:NumGenes
1      2:21
2      1:6     4:153
3      1:135   3:4
4      3:34
```

Figure 6: Cluster Mapping text output file.

The first column lists each cluster in the first solution. The rest of the columns list how many genes in each clustering solution1 are located in the corresponding clustering solution2. The file can be read as follows: the solution1 cluster #1 is composed of 21 genes from the solution2 cluster #2; the solution1 cluster #2 is composed of 6 genes from the solution2 cluster #1 and 153 genes from the solution2 cluster #4; the solution1 cluster 3 is composed of 135 genes from the solution2 cluster #1 and 4 genes from the solution2 cluster #3; and the solution1 cluster #4 is composed of 34 genes in solution2 cluster #3.

Generate Heatmap Output top

The heatmap generation function outputs *n* .fig files, one for each of the *n* clusters that are specified in the input file; other image file formats may also be chosen as well as the minimum cluster size (see Tutorial). These files are the same as those output by the

standard clustering function, only the input clusters are pre-defined. Also, a text output file, similar to that shown in Figure 3 is output.

Cluster Statistics Output top

The ClusterStats function outputs one text file containing information about each cluster. An example file is shown in Figure 7.

```
General Cluster Information
Total number of clusters: 8
Total number of singleton clusters: 0
Average cluster size (excluding singletons): 49
Largest cluster size: 92 genes
Smallest cluster size: 14 genes


------------------------------
Cluster 1:
Average Euclidean Distance Score:    1.64
Number of genes in cluster:   92
Average Expression Profile:    0.08  0.53  1.35  2.18  1.94
Standard Deviations:     0.46  0.58  0.58  0.44  0.81
Standard Errors:   0.05  0.06  0.06  0.05  0.08
Count of positive expressions:      31    75    91    92    89
Count of negative expressions:      38    15    1     0     2
Count of zero expression:     23    2     0     0     1

------------------------------
Cluster 2:
Average Euclidean Distance Score:    1.64
Number of genes in cluster:   76
Average Expression Profile:    0.14  1.45  2.51  3.13  2.93
Standard Deviations:     0.38  0.61  0.48  0.49  0.81
Standard Errors:   0.04  0.07  0.06  0.06  0.09
Count of positive expressions:      36    76    76    76    76
Count of negative expressions:      21    0     0     0     0
Count of zero expression:     19    0     0     0     0

------------------------------
```

Figure 7: Cluster Statistics text output file.

The very first section is a global summary of the clusters contained in this analysis.

A description of each per-cluster element in the output file is below:

- *Average Euclidean Distance Score:* This score reflects the overall homogeneity of each cluster with respect to Euclidean distance (ED). ED looks for clusters that have highly similar levels of expression; thus, a lower ED score indicates higher homogeneity with respect to similar expression levels. Note that if the clusters were generated using Pearson's correlation coefficient then the ED score will most likely be high since correlation clusters are not homogeneous with respect to Euclidean distance.
- *Number of genes in cluster:* The size of the cluster.
- *Average Expression Profile:* The expression values in each condition/column were averaged to obtain an average profile for each cluster.

- *Standard Deviations:* The standard deviation for each condition/column in a cluster is calculated. This is a measure of the variance at each time point. Lower StdDev indicate more homogeneity for a given condition.
- *Standard Errors:* The StdDev for each condition was divided by the total number of genes in the cluster to obtain the standard error.
- *Count of positive transcripts:* A count of how many transcripts were up-regulated for each condition (expression > 0).
- *Count of negative transcripts:* A count of how many transcripts were down-regulated for each condition (expression < 0).
- *Count of zero transcripts:* A count of how many transcripts showed no change for each condition (expression = 0).

References top

Yeung, K. Y., D. R. Haynor, et al. (2001). "Validating clustering for gene expression data." Bioinformatics **17**(4): 309-18.