Input File Formats for SC^{2}ATmd v3

Last updated on 6/13/2012 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.

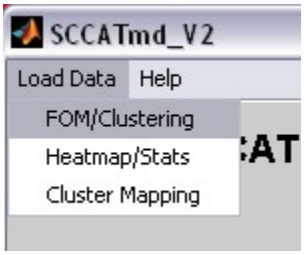

Figure 2: Loading FOM and clustering input files.

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

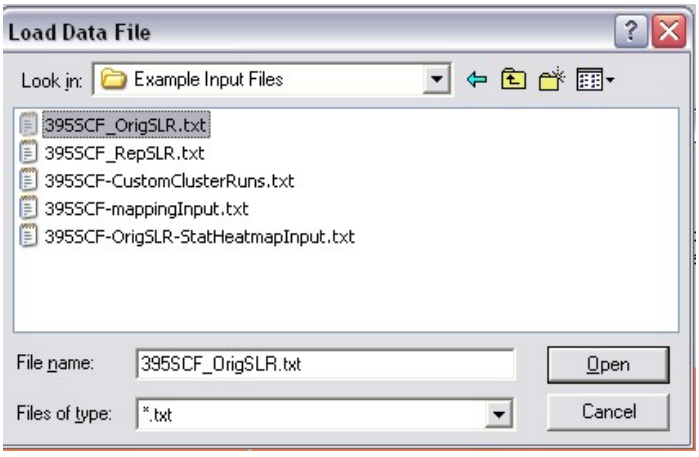

Figure 3: Browse to data file.

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

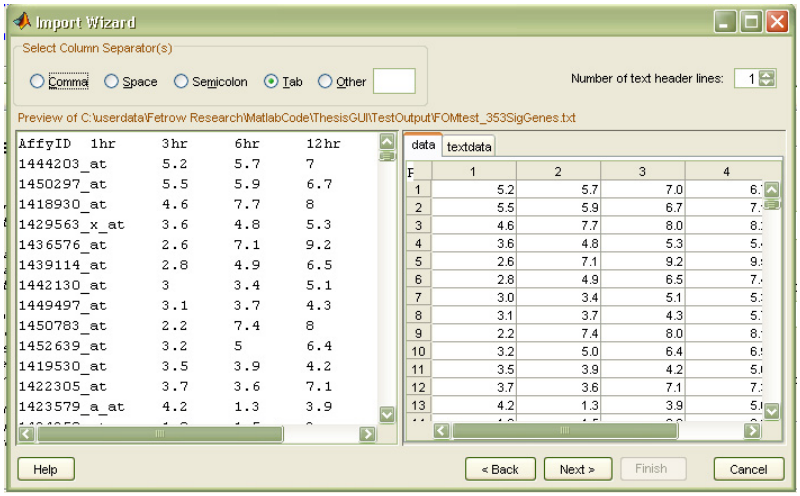

Figure 4: Matlab Import Wizard

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 unnecessary white space and try to reload it. See Figure 5.

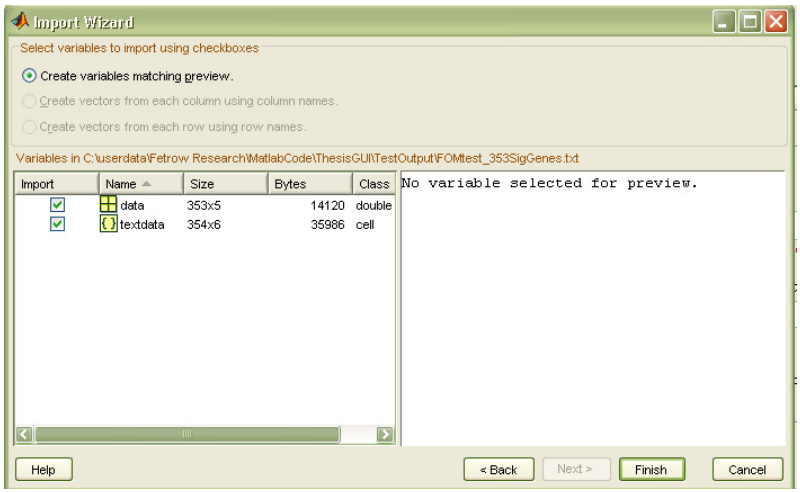

Figure 5: Matlab Import Wizard

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

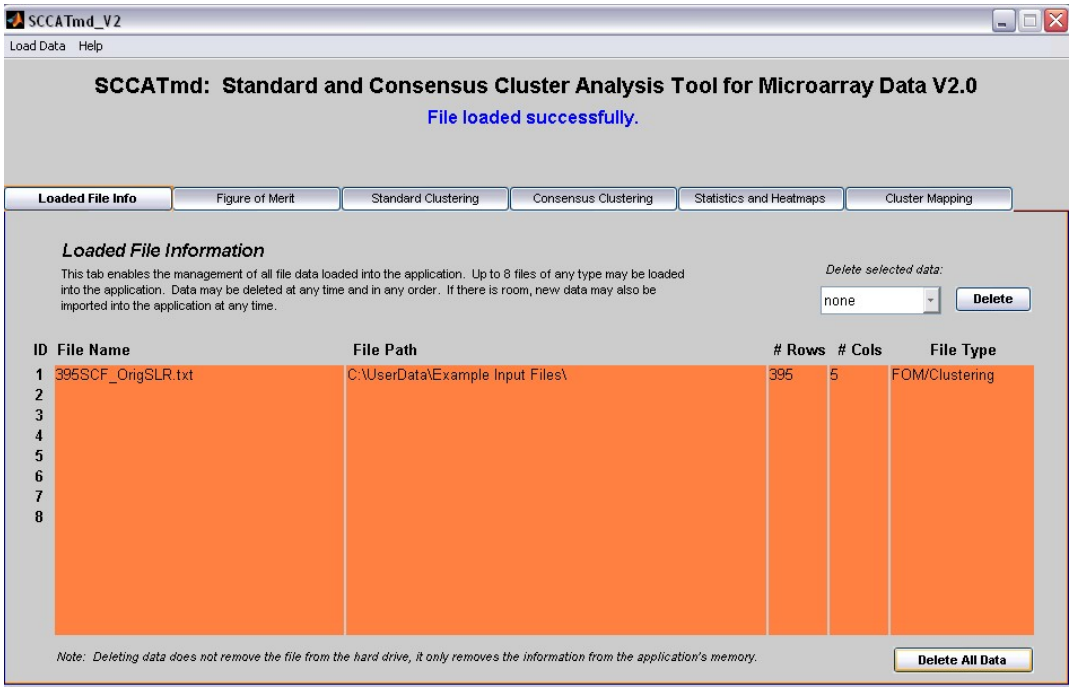

Figure 6: Updated file information on FileInfo Tab.

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. time points, cell types, etc.) with the first row containing the column headers.

Tip: If you have data that you want to visualize as a heat map without breaking it down into different clusters, simply format the file as below, but assign all genes to cluster 1. SC^{2}ATmd will then generate only 1 heat map with all of your genes in it using HAC with Euclidean distance to organize the heat map.

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 heat map 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.

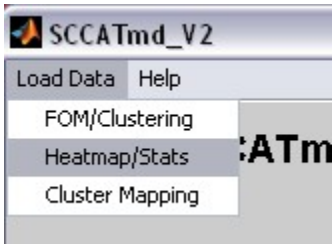

Figure 8: Loading Heatmap and Stats input file.

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.

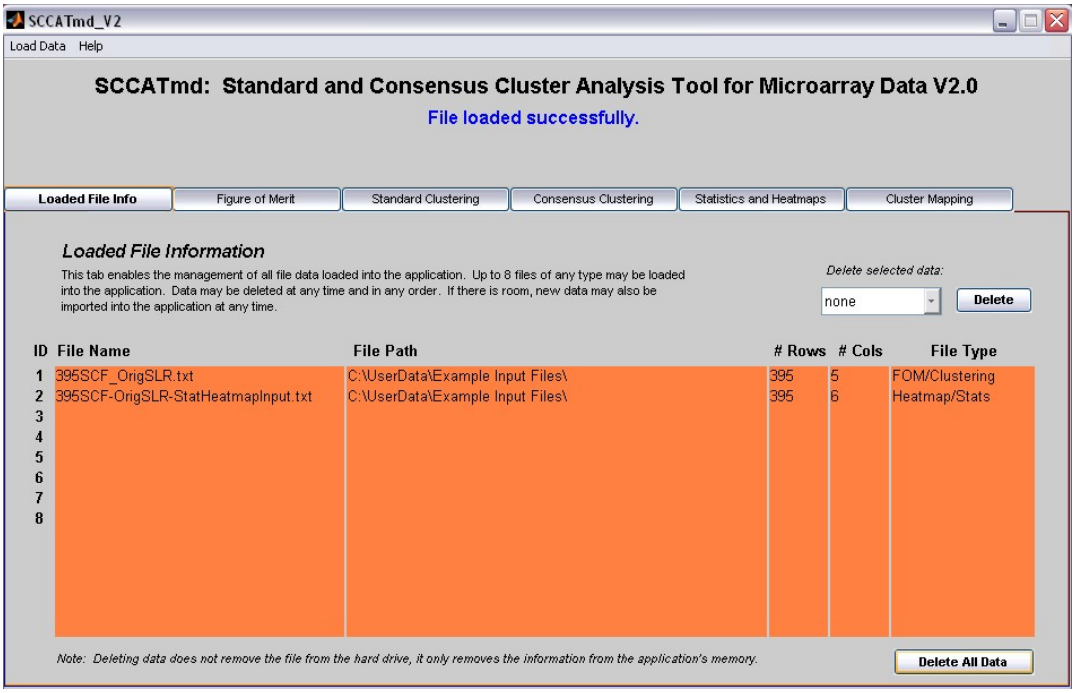

Figure 9: Updated file information

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 steps 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.

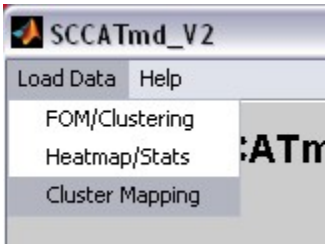

Figure 11: Loading Cluster Mapping input file.

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.

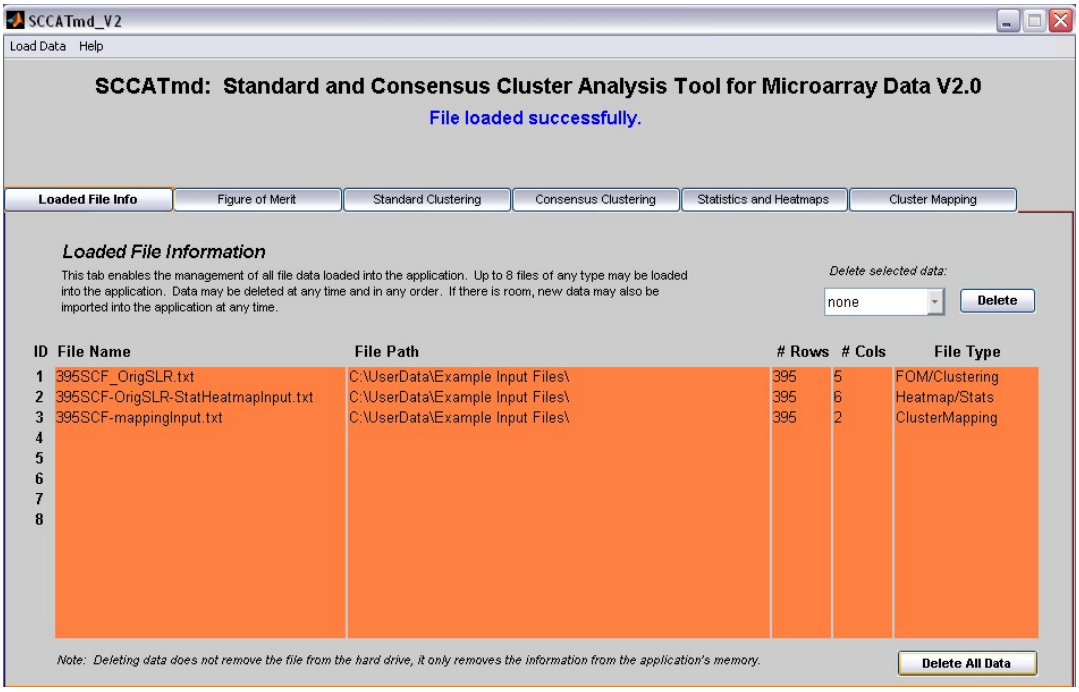

Figure 12: Updated file information

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