

TIGR MeV

MultiExperiment Viewer

Version 3.1
March 8, 2005

Table of Contents

1. General Information.....	5
2. TIGR Software Overview	6
3. Starting TIGR MultiExperiment Viewer	7
4. Loading Expression Data	8
4.1. Loading MeV (.mev) Format Files.....	8
4.2. Loading TIGR Array Viewer (.tav) Files	9
4.3. Loading Tab Delimited, Multiple Sample (.txt) Files (TDMS Format)	10
4.4. Loading Affymetrix Data (.txt, .TXT) Files	11
4.5. Loading GenePix (.gpr) Data Files	12
4.6. Loading Agilent Files.....	13
4.7. Initial View of the Loaded Data, Main Expression Image.....	14
4.8. Result Navigation Tree	15
4.9. The History Node and Log.....	15
5. Adjusting the Data.....	17
5.1. Adjustment / Filter Overview.....	17
5.2. Data Transformations.....	18
5.3. Data Filters (Data Quality and Variance Based Filters)	19
5.4. Data Source Selection.....	22
6. Display Options.....	24
6.1. Sample Annotation.....	24
6.2. Selecting Gene Annotation.....	26
6.3. Color Scheme Selection	26
6.4. Setting Color Scale Limits	27
6.5. Element Appearance	29
7. Viewer Descriptions.....	30
7.1. Overview	30
7.2. Expression Images	30
7.3. Expression Graphs.....	33
7.4. Centroid Graphs	35
7.5. Table Views.....	37

7.6. Common Viewer Activities	39
8. Working with Clusters	41
8.1. Storing Clusters and Using the Cluster Manager	41
9. Utilities Menu	48
Search Utility	48
Import Gene or Sample List	49
Append Sample Annotation	51
Append Gene Annotation	51
10. Creating Output	53
10.1. Saving the Analysis	53
10.2. Saving the Expression Matrix	53
10.3. Saving Viewer Images	54
10.4. Saving Cluster Data	54
11. Modules	55
Description Conventions and General Pointers	55
11.1. HCL: Hierarchical clustering	56
11.2. ST: Support Trees	60
11.3. SOTA: Self Organizing Tree Algorithm	64
11.4. RN: Relevance Networks	69
11.5. KMC: K-Means/K-Medians Clustering	72
11.6. KMS: K-Means / K-Medians Support	74
11.7. CAST: Clustering Affinity Search Technique	76
11.8. QTC: QT CLUST	78
11.9. SOM: Self Organizing Maps	80
11.10. GSH: Gene Shaving	83
11.11. FOM: Figures of Merit	85
11.12. PTM: Template Matching	89
11.13. TTEST: T-tests	92
11.14. SAM: Significance Analysis of Microarrays	100
11.15. ANOVA: Analysis of Variance	104
11.16. TFA: Two-factor ANOVA	107
11.17. SVM: Support Vector Machines	109
11.18. KNNC: K-Nearest-Neighbor Classification	116

11.19. DAM: Discriminant Analysis Module.....	120
11.20. COA: Correspondence Analysis	132
11.21. PCA: Principal Components Analysis.....	134
11.22. TRN: Expression Terrain Maps.....	136
11.23. EASE: Expression Analysis Systematic Explorer	142
12. Scripting	157
<i>Creating a New Script</i>	157
<i>The Script Tree Viewer: Script Construction</i>	158
<i>Adding an Algorithm</i>	159
<i>Script XML Viewer</i>	166
<i>Loading a Script</i>	168
<i>Running a Script</i>	168
13. Working with the Single Array Viewer	171
14. Appendix: File Format Descriptions.....	174
14.1. TAV Files	174
14.2. Tab Delimited, Multiple Sample Files (TDMS files)	175
14.3. GenePix Files.....	177
14.4. MEV Files	177
14.5. Annotation Files.....	180
15. Appendix: Preferences Files	182
16. Appendix: Distance Metrics.....	184
17. Appendix: MeV Script DTD	187
18. License	190
19. Contributors.....	191
20. References.....	193

1. General Information

1.1. Obtaining MeV

www.tigr.org/software/tm4/

Maintainer / Contact Information

TIGR MeV Team – mev@tigr.org

Platform / System Requirements

Java Runtime Environment (JRE) 1.4 or later

Java3D v1.3 or later required for PCA, TRN, DAM and COA functions

1.2. Referencing MeV

Users of this program should cite:

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*. 2003 Feb;34(2):374-8.

http://www.tigr.org/software/tm4/menu/TM4_Biotechniques_2003.pdf

1.3. A note on non-Windows operating systems

The majority of our MeV development and testing was performed on Windows operating systems. Although MeV will run under other operating systems, there may be some incompatibilities or bugs revealed in this manner. Please report any such issues to mev@tigr.org.

MacOSX users can simulate the ‘right-click’ by using <control> click.

2. TIGR Software Overview

TIGR MultiExperiment Viewer is one member of a suite of microarray data management and analysis applications developed at The Institute for Genomic Research (TIGR). Within the suite, known as TM4, there are four programs: MADAM, Spotfinder, MIDAS and MeV. Together, they provide functions for managing microarray experimental conditions and data, converting scanned slide images into numerical data, normalizing the data and finally analyzing that normalized data. These tools are all OSI certified (see section 12) open-source and are freely available through the TIGR website, www.tigr.org/software/tm4.

The Microarray Data Manager (MADAM) is a data management tool used to upload, download, and display a plethora of microarray data to and in a database management system (MySQL). An interface to MySQL, Madam allows scientists and researchers to electronically record, capture, and administrate annotated gene expression and experiment data to be shared with and ultimately used by others within the scientific community.

TIGR Spotfinder is image-processing software created for analysis of the image files generated in microarray expression studies. TIGR Spotfinder uses a fast and reproducible algorithm to identify the spots in the array and provide quantification of expression levels.

TIGR Microarray Data Analysis System (MIDAS) is an application that allows the user to perform normalization and data analysis by applying statistical means and trim the raw experimental data, and create output for MeV.

TIGR MultiExperiment Viewer (MeV) is an application that allows the viewing of processed microarray slide representations and the identification of genes and expression patterns of interest. Slides can be viewed one at a time in detail or in groups for comparison purposes. A variety of normalization algorithms and clustering analyses allow the user flexibility in creating meaningful views of the expression data.

3. Starting TIGR MultiExperiment Viewer

- 3.1. If using Windows, run the TMEV batch file (TMEV.bat) to start the program. This batch file invokes the Java interpreter and stores input parameters. Similarly, if using Linux or Unix, run the tmev.sh file. Macintosh users should double-click on the application file named TMEV_MacOSX_2_2.
- 3.2. A main menu bar will appear, with four menus: *File*, *Display*, *Window* and *References*. Use the *File* menu in the main menu bar to open a new Single or Multiple Array Viewer, load a new preferences file or log in to a database. MeV will continue to run while this menu bar is present. To exit the entire application, select *Quit* from the *File* menu.
- 3.3. Expression data can be viewed from within either a Single or a Multiple Array Viewer. However, the former can open only one set of expression data at a time. The Multiple Array Viewer can display many samples together. The real power of MeV is in the program's analysis modules, found only in the Multiple Array Viewer. That is where the clustering and visualization of data can take place. Therefore, the remainder of this manual will focus on the Multiple Array Viewer. Please see section 13 for details regarding the Single Array Viewer.

4. Loading Expression Data

MeV can interpret files of several types, including the MultiExperiment Viewer format (.mev), the TIGR ArrayViewer format (.tav), the TDMS file format (Tab Delimited, Multiple Sample format), the Affymetrix file format, and GenePix file format (.gpr). See section 14 for details regarding the different file formats.

In addition to being formatted correctly, the input data should already be normalized. Using normalized data as input will result in more statistically valid output. [MIDAS](#), a member of the [TM4 software suite](#), is one program that can do this normalization.

The maximum number of samples that can be loaded into a Multiple Array Viewer at one time depends on the available RAM in the computer running MeV and the number of expression values from the samples.

4.1. Loading MeV (.mev) Format Files

Select *Load File* from the *File* menu to launch the file loading dialog. At the top of this dialog, use the drop-down menu to select the type of expression files to load. On the left-hand side of the dialog is a file browser. Use this browser to locate the files to be loaded.

The default file type to load is the .mev file. This file type is an update of the older .tav file format. Details about this file format can be found in the appendix (14.4). In the section of the loader labeled “TIGR MeV Expression Files (*.mev),” the contents of the folder selected in the file browser will be displayed in the box labeled “available”. Select the .mev files to load and click the “add” button to add them to the list of files to be loaded. Similarly, select the file(s) containing the appropriate annotation information in the section labeled “TIGR MeV Annotation Files (*.ann, *.dat).”

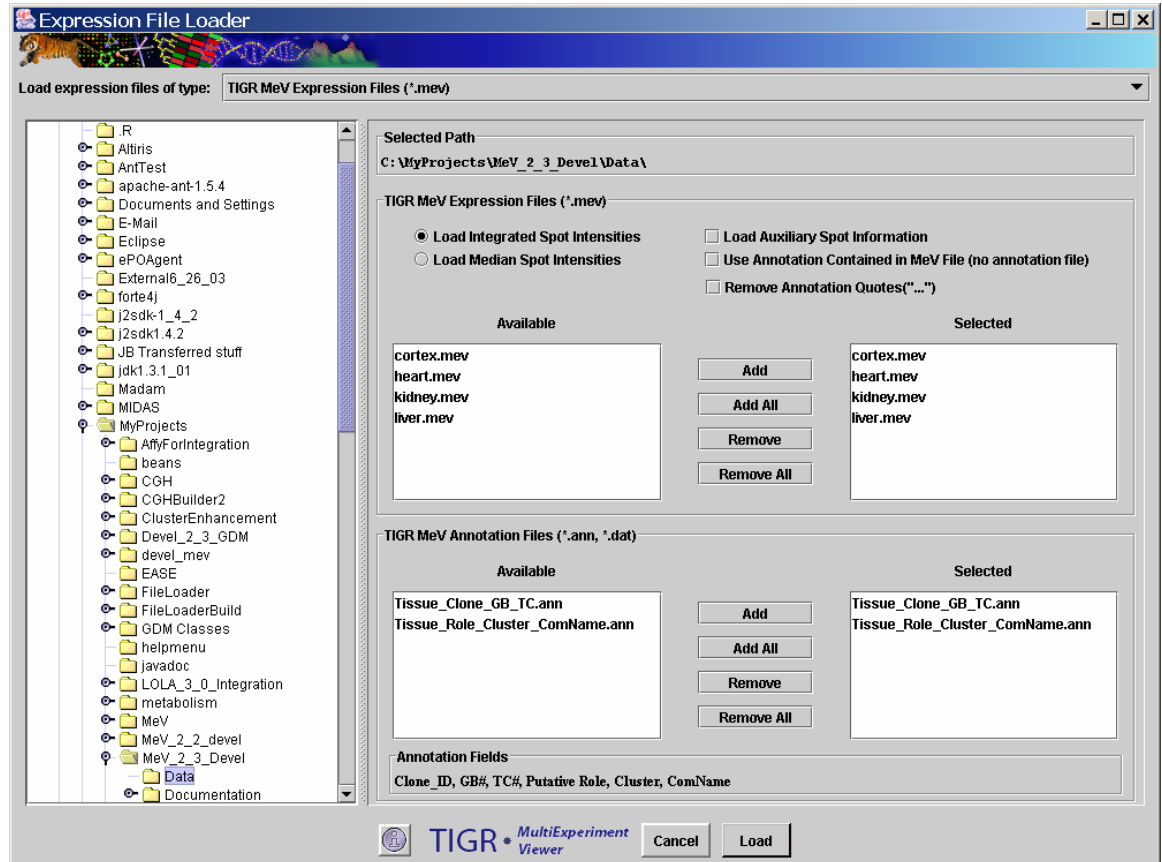
The *Load Integrated Intensities/Load Median Intensities* options specify that MeV should load either integrated intensities (default) or Median intensities. Note that care should be taken to select the intensity measurements, integrated or median, that have been previously normalized in TIGR MIDAS so that the loaded data will have been normalized and possibly trimmed.

The *Load Auxiliary Spot Information* option will load spot background, spot pixel count and other spot specific items. Loading this information will not impact the analysis results but will allow you to view this data when clicking on an expression element. The **default is not to load** this spot information since it consumes significant system memory and can severely limit the number of mev files that can be loaded due to memory constraints.

The *Use Annotation Contained in MeV File* option overrides the use of an annotation file by loading annotation that is contained within the MeV file. This is a specialized case and for ease of annotation updates it is suggested that you adhere to using a separate annotation file even if the mev file contains annotation.

This option is off by default.

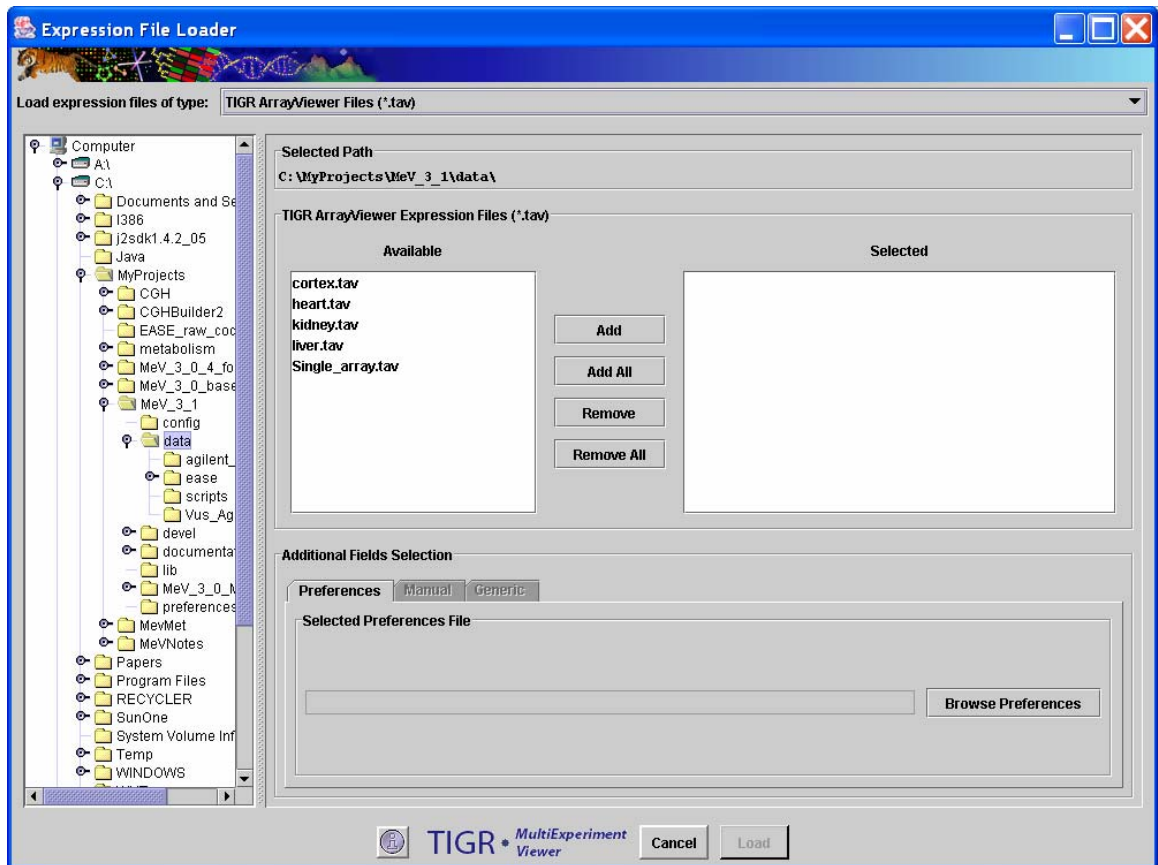
The *Remove Annotation Quotes* option removes quotations from annotation fields where the annotation entries start and end with quotation marks. This option is provided to counteract the behavior of some popular ‘spread sheet’ programs where cells containing text with a delimiter such as a comma are automatically enclosed by quotation marks. If MeV stalls on loading an annotation file, try loading it with this option selected.



4.1.1. The Expression File Loader (MeV Files)

4.2. Loading TIGR Array Viewer (.tav) Files

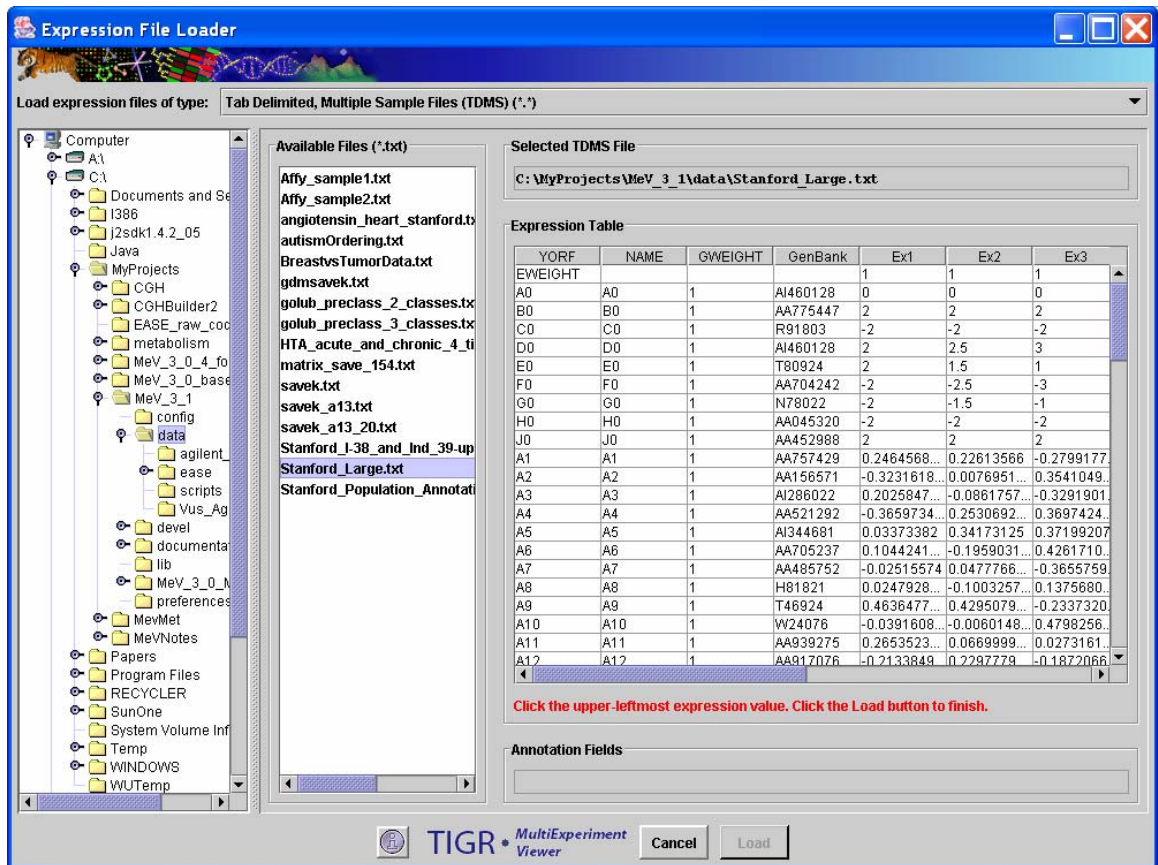
To load .tav-formatted files, use the drop-down menu to select the *TIGR ArrayViewer (*.tav)* option. This loader is very similar to the .mev loader. Use the file browser to select the .tav files to load. Instead of selecting annotation files to load alongside the data files, however, you must select a preferences file. This preferences file contains information that MeV uses to determine what type of .tav file is being loaded. See the Appendix (section 15) for more details on preferences files (15).



4.2.1. The Expression File Loader (.tav Files)

4.3. Loading Tab Delimited, Multiple Sample (.txt) Files (TDMS Format)

Select the *Tab Delimited, Multiple Sample Files (TDMS) (*.*)* option from the drop-down menu to load TDMS format files. Navigate to the folder containing the file and select the desired file in this format. The file will be displayed in a tabular format in the file loader preview table. Click the cell in the table which contains the upper-leftmost **expression value** in the file. The header labels for the annotation fields will be displayed at the bottom of the dialog. Check that the correct fields are listed before clicking *Load*.

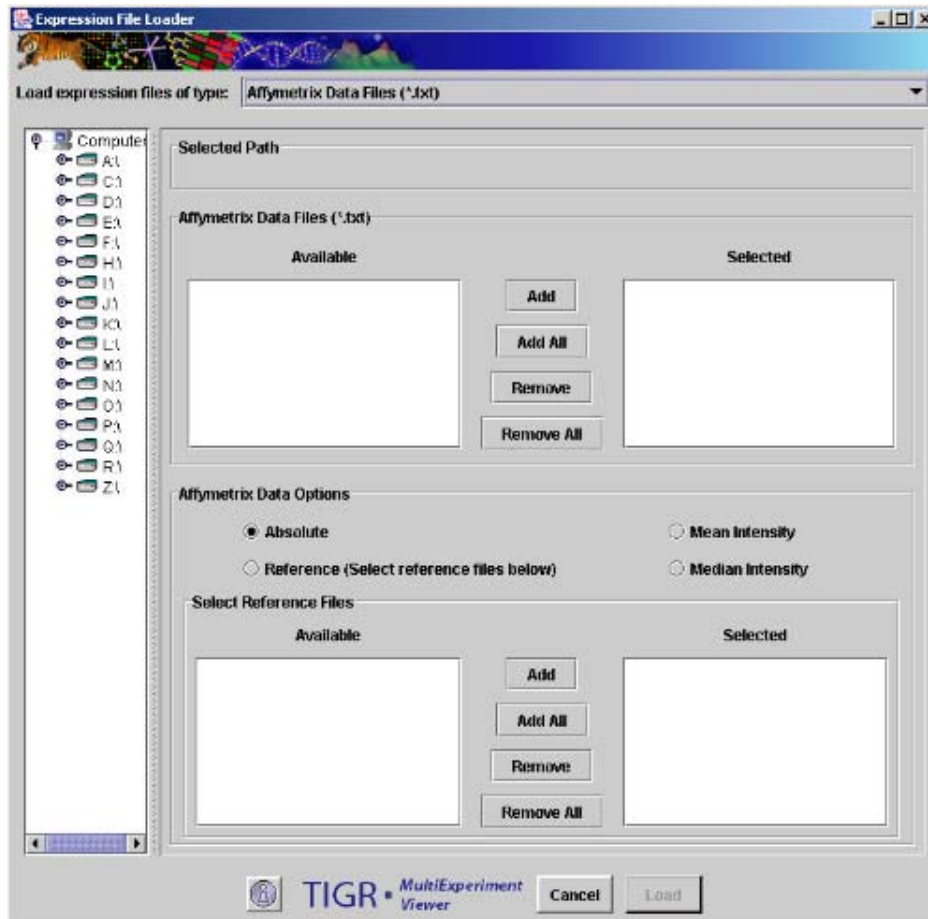


4.3.1. Loading Tab Delimited, Multiple Sample Files (TDMS Format)

4.4. Loading Affymetrix Data (.txt, .TXT) Files

Selecting *Affymetrix Data Files (*.txt)* from the drop-down menu allows the loading of Affymetrix files. Select the Affymetrix files to be loaded as you would when loading .mev files, using the file browser.

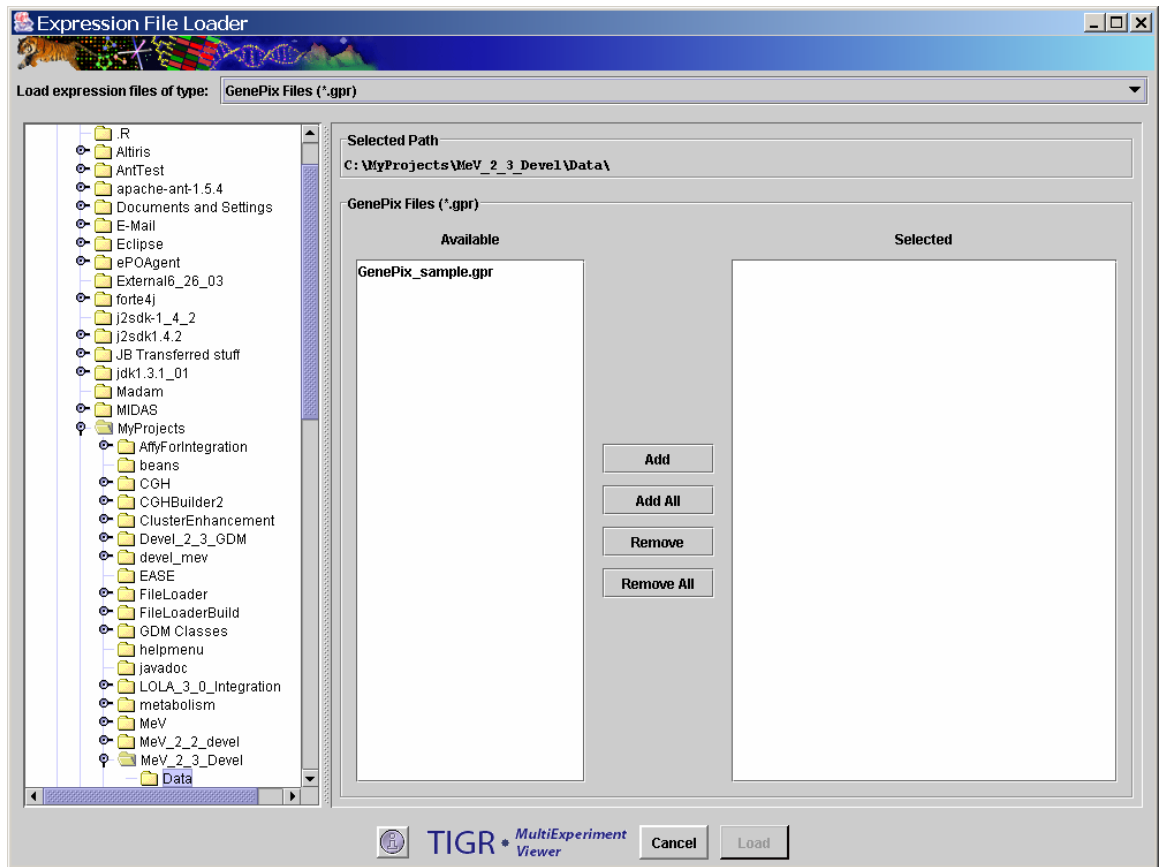
These files contain a single intensity value per spot, instead of the usual two that MeV requires. The values loaded from these files will be use as a Cy5 value, that is, the numerator in the calculation of the ratio of intensities. Therefore there are several options for simulating a second intensity value (the denominator). Select from the radio button options to choose a method. If *Absolute* is selected, the denominator is given a value of 1 for all ratio calculations. If *Mean Intensity* is selected, the average of all intensity values for that gene across all loaded Affymetrix files is used as the denominator for that spot. Similarly, if *Median Intensity* is selected, the median of all intensity values for that spot is used. If *Reference* is selected, a reference Affymetrix file, selected in the file selector at the bottom of the dialog, is used. The intensity value of each record in the Affymetrix file is used as the denominator of the ratio calculation for the corresponding spot in each of the loaded data files.



4.4.1. The Expression File Loader (Affymetrix Files)

4.5. Loading GenePix (.gpr) Data Files

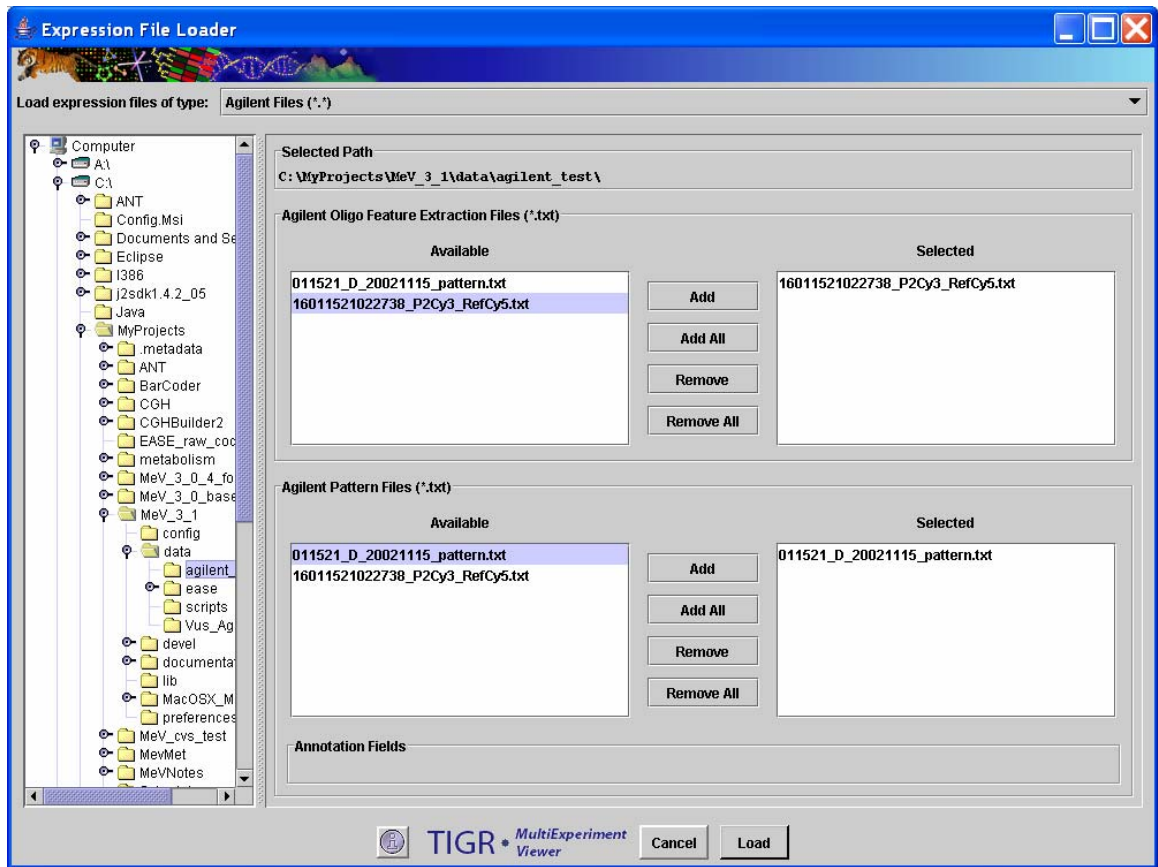
GenePix (.gpr) files can be loaded by selecting the GenePix file loader option from the list of available file formats to load. Use the file system navigation tree on the left to move the directories containing files to load. Files appearing in the *Available* file list can be added to the *Selected* file list using the *Add* or *Add All* buttons.



4.5.1. The Expression File Loader (GenPix Files)

4.6. Loading Agilent Files

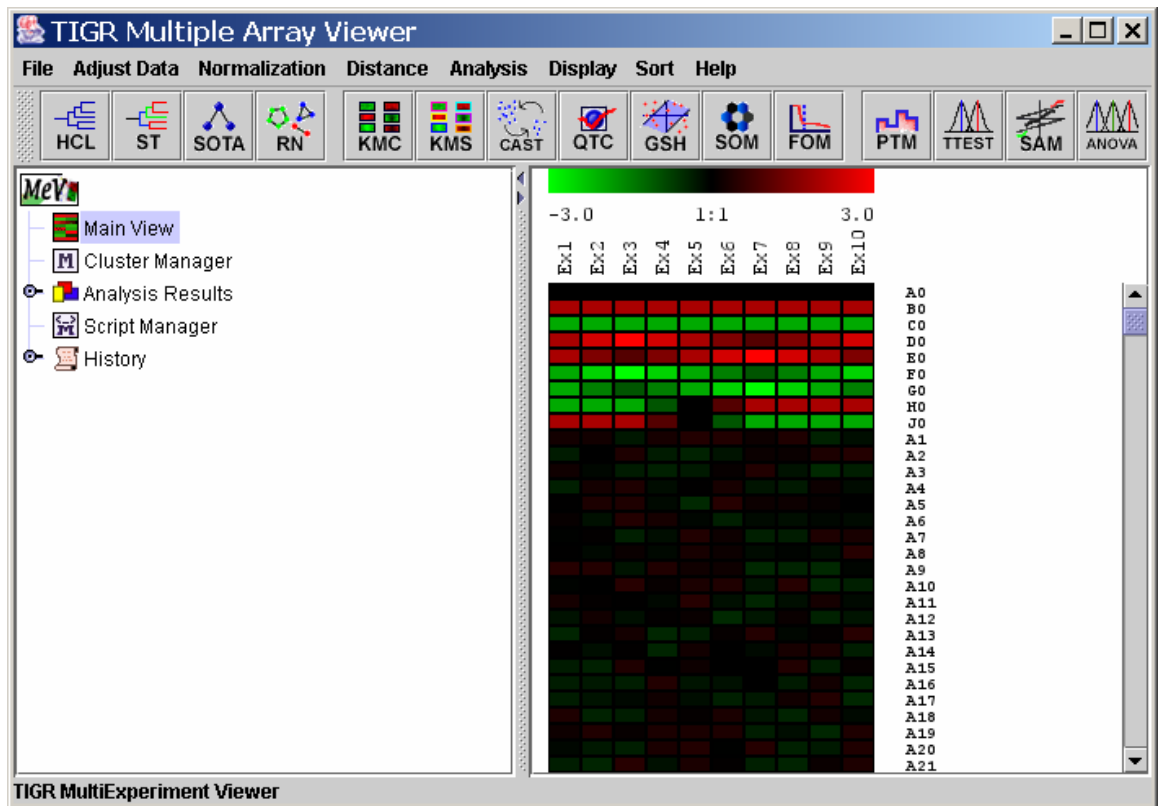
Agilent format text files can be loaded by selecting the Agilent file loader option from the list of available file formats to load. Use the file system navigation tree on the left to move the directories containing files to load. Files appearing in the *Available* file list can be added to the *Selected* file list using the *Add* or *Add All* buttons. The upper file selection area is for the selection of Agilent Oligo Feature Extraction text files and lower file selection area is for the text version of the pattern file that corresponds to your slide.



4.6.1. Agilent File Loader

4.7. Initial View of the Loaded Data, Main Expression Image

For each set of expression values loaded, a column is added to the main display (4.7.1). This display is an *Expression Image* viewer, in which each column represents a single sample and each row represents a gene. The names of the samples are displayed vertically above each column, and any annotation field of interest from the input files can be displayed to the right of each row. MeV expects that each sample loaded will have the same number of elements, in the same order, and that each gene (spot) is aligned with that element in every other sample loaded. For example, using that rule, all input files will have data for gene x in row y . Clicking on a spot displays a dialog with detailed information about that spot. For more detail regarding the *Expression Images* viewer, see section 7.2.



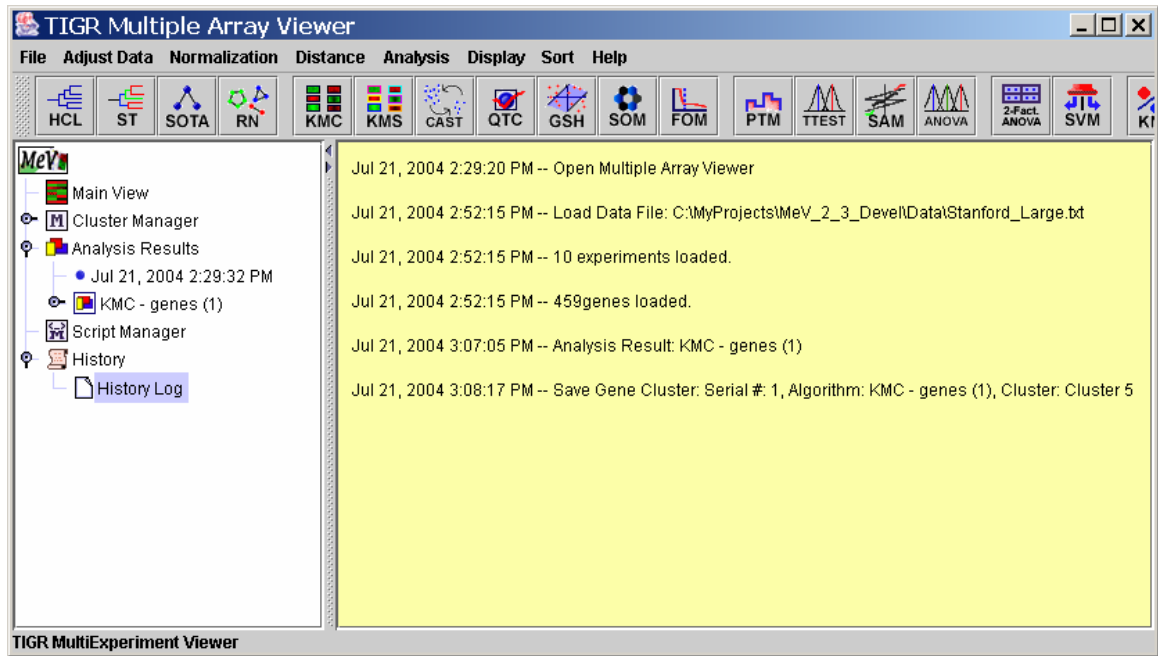
4.7.1. Main View in MeV. (Borders drawn on image.)

4.8. Result Navigation Tree

The left side of the main interface is a navigation tree. At any time, clicking on the *Main View* node will return to the main display. Output from any MeV module will be added as a new subtree under the *Analysis* node. In general, clicking on a node in the tree will navigate to the associated result or data view and that view will be displayed in the right panel. The initial tree will also include a *Cluster Manager* node to manage clusters stored from analysis results. The *Script Manager* manages activities related to analysis script creation, loading, modification, and execution. The Cluster Manager and the Script Manager will be covered in detail in its sections 7.2 and 10 respectively.

4.9. The History Node and Log

The History Node contains a log of most activities. For each log entry a date and time is recorded. Major events such as file loading, analysis loading, script loading, algorithm execution, and cluster storage events are logged to the History Log. If the analysis is stored to a file, then the History is retained and restored so that new events can be logged. The history log can be stored to a text file by right clicking in the viewer and selecting the *Save History to File* menu option.

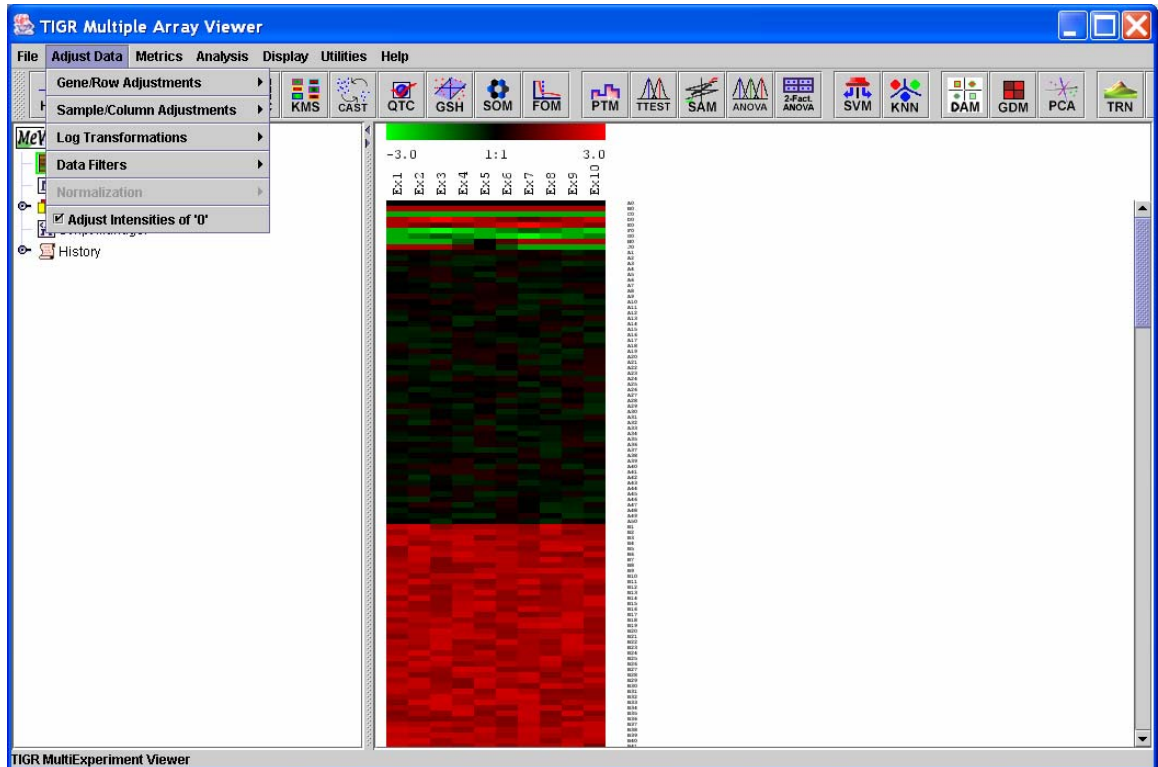


4.9.1. The History Log

5. Adjusting the Data

5.1. Adjustment / Filter Overview

Prior to starting an analysis, certain data adjustments can be done using the items in the *Adjust Data* menu. These include normalization for genes/rows, log transformations, and various filters.



5.1.1. Adjust Data Menu

Adjustments may not necessarily affect the main display or the values displayed when elements are clicked on the matrix displays, but will influence the calculation of the expression matrix, the foundation of all analyses. Adjustments will also be reflected when the entire matrix or individual clusters are saved as text files, although the original data files are not overwritten. Furthermore, with the exception of three options: “Set Lower Cutoffs”, “Set Percentage Cutoffs” and “Adjust Intensities of Zero”, all the changes made to an expression matrix are irreversible for the current MeV session. Different types of adjustments may be applied on top of one another in any sequence, and the same type of adjustment may be applied repeatedly to the matrix, although this may not make sense from the point of view of analysis.

Because of the above features, sometimes it might not be a good idea to apply data transformations halfway through an analysis (except for perhaps the “Set Lower Cutoffs”, “Set Percentage Cutoffs” and “Adjust Intensities of Zero” options), as the post-transformation analyses and displays might not be entirely consistent with the pre-transformation analyses. A good way to use these options might be to apply any required adjustments to the data set, save the entire adjusted matrix as a tab delimited, multiple sample (TDMS) formatted text file (using the “Save Matrix” option under the “File” Menu), and then load this new file in a new

MeV session, during which no further data adjustments will be made. This will ensure consistency throughout the MeV session. Adjustment options are described below:

5.2. Data Transformations

Log2 Transform

This is fairly self-evident, just taking the log2 transform of every element in the matrix. Note that this adjustment should not usually be necessary. When .tav or .mev files are loaded into MeV, the program will automatically compute the log2 ratio of the two intensities and use them in the expression matrix. TDMS files also often contain pre-calculated log2 ratios.

Normalize Genes/Rows

This will transform values using the mean and the standard deviation of the row of the matrix to which the value belongs, using the following formula:

$$\text{Value} = [(\text{Value}) - \text{Mean}(\text{Row})] / [\text{Standard deviation}(\text{Row})]$$

Divide Genes/Rows by RMS

This will divide the value by the root mean square of the current row, where root mean square = square root $[\sum(x_i)^2/(n-1)]$, where x_i is the i^{th} element in the row consisting of n elements.

Divide Genes/Rows by SD

This will divide each value by the standard deviation of the row it belongs to.

Mean Center Genes/Rows

This will replace each value by [value – Mean(row that value belongs to)].

Median Center Genes/Rows

This will replace each value by [value – Median(row that value belongs to)].

Digital Genes/Rows

This will divide up the interval between the minimum and the maximum values in a row into a number of equal-sized “bins”. Each value is now replaced by an integer value of zero or greater, denoting which bin it belongs to (e.g., the minimum value is assigned to bin “zero”, indicating it belongs to the lowest bin;

the maximum value is assigned to the highest bin, and the rest of the values fall in the intermediate bins).

Sample/Column Adjustments

These function in the same way as their corresponding options on genes/rows, except that the current column values, rather than the current row values, are used in the computation.

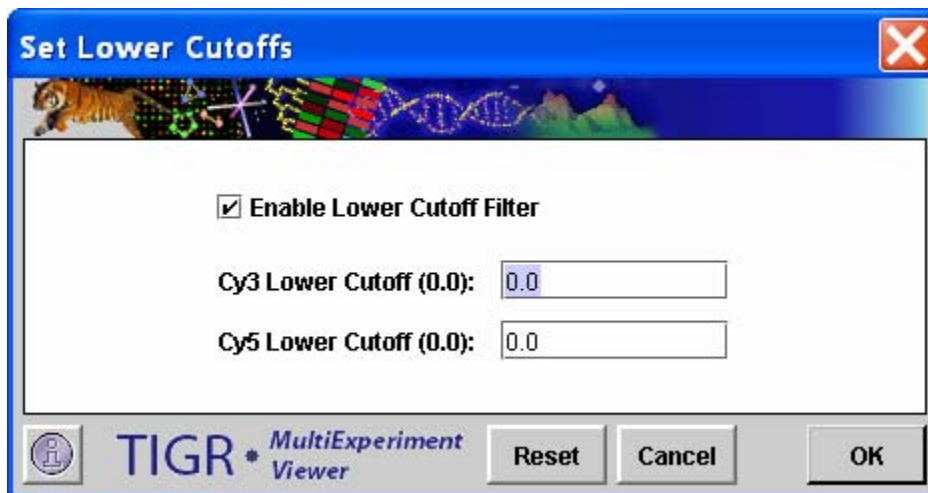
Log10 to Log2

This assumes that the current data are log 10 transformed, and transforms them to log base 2, i.e., it assumes that the input data is in the form $\log_{10}x$, and it outputs \log_2x .

5.3. Data Filters (Data Quality and Variance Based Filters)

Lower Cutoffs

Select *Use Lower Cutoffs* to exclude from analysis any genes for which the expression values (in either the corresponding Cy3 or Cy5 columns) are lower than specified values. Select *Set Lower Cutoffs* to set these values. To enable this option, check the “Use Lower Cutoffs” checkbox just below the “Set Lower Cutoffs” menu option, and uncheck it to disable this option. All subsequent analyses will include only those genes for which all Cy3 and Cy5 values are above the specified thresholds. This option is disabled by default.



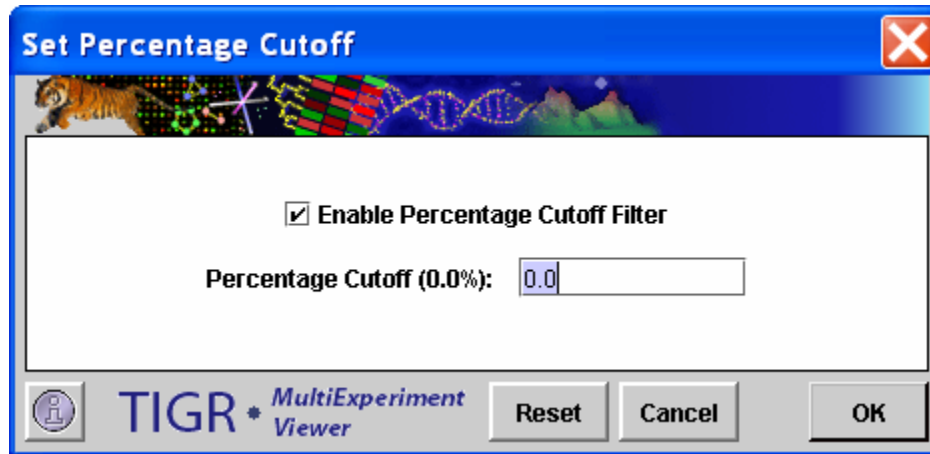
5.3.1. Lower Cutoff Filter Dialog

Percentage Cutoffs

Select *Use Percentage Cutoffs* to ignore the genes for which there are not enough valid (non-zero) expression values across all samples. This will not delete any data, but will only exclude the genes from analysis. This option is sometimes

useful in speeding up module calculation since many zeros will often slow them down.

To determine which genes will be excluded, select *Set Percentage Cutoffs* and enter a percentage value. To enable this option, check the “Use Percentage Cutoffs” checkbox just below the “Set Percentage Cutoffs” menu option, and uncheck it to disable this option. Genes with less than the specified percentage of non-zero values will be ignored. A value of 0.0% indicates that all genes will be used in the analysis. To require that every one of the gene’s expression values must be valid to be included, set the value to 100. This option is disabled by default.



5.3.2. Percentage Cutoff Filter Dialog

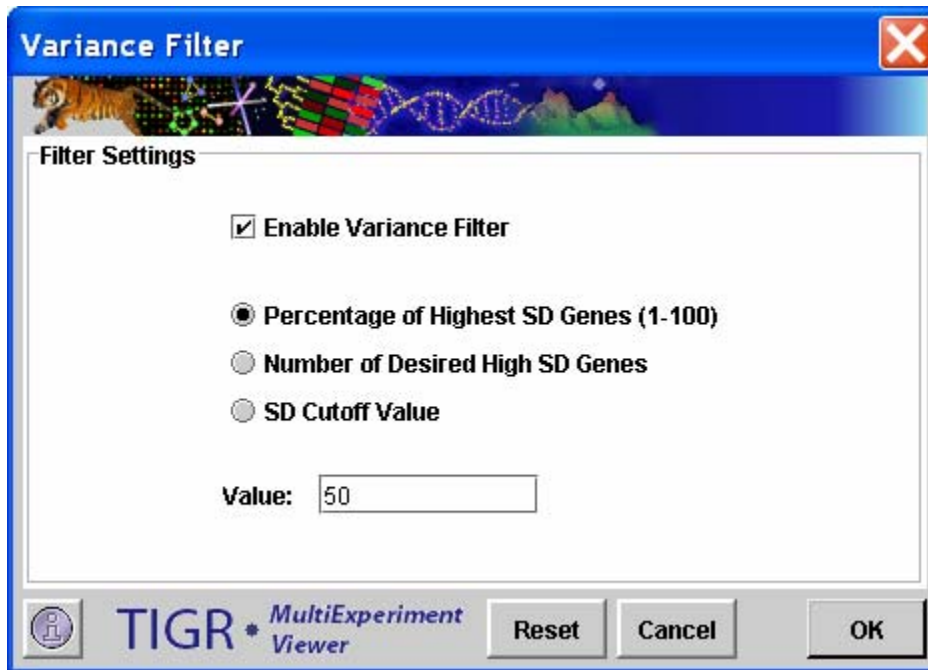
Variance Filter

The variance filter allows the removal of genes with low variation of expression over the loaded samples. This filter is basically used to remove ‘flat genes’ that don’t vary much in expression over the conditions of the experiment. The variance filter has three possible criteria for specifying which genes to keep. The *Enable Variance Filter* check box turns the filter on and off. Be sure to observe the *History Node* log to see the number of genes retained after using the filter. Note that the variance filter is performed after other filters such as *Percent Cutoff Filter* is imposed. This convention insures that the genes that are check for variance also contain some minimum level of ‘good’ (not missing) data.

The *Percentage of Highest SD Genes* option ranks the genes based on standard deviation and then the genes that are kept are some percentage of this ranked list. For and example, if we have 1000 genes and the percentage was set to 20%, then the result would be a final list of the 200 most variable genes.

The *Number of Desired High SD Genes* also ranks the genes based on SD and then the number of genes specified are selected from this SD ordered list such that the highest SD genes are selected.

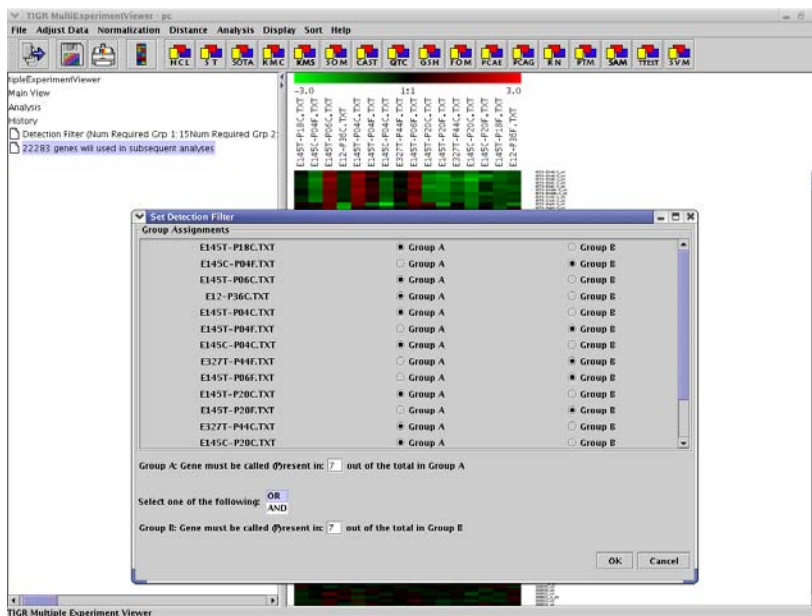
The *SD Cutoff Value* uses an actual SD value such that all genes having an SD greater than this value are selected.



5.3.3. Variance Filter Dialog

Detection Filter (for Affymetrix Data w/ Detection Calls Only)

Select *Use Detection Filter* to ignore genes that are not marked ‘present’ in enough samples. Select *Set Detection Filter* to divide the samples into two groups. Enter the number of times that a gene must be called as ‘present’ in group A. Do the same for group B. Select 'AND' so that each gene must pass both criteria. Select 'OR' so that a gene only must pass one of the criteria in order to be used in further analysis.



5.3.4. Set Detection Filter Dialog

Fold Filter (for Affymetrix Data Only)

Select *Use Fold Filter* to remove genes that do not pass one of three specified criteria based on Fold Change. Fold Change is calculated as: (mean signal in Group A)/(mean signal in Group B). Select *Set Fold Filter* and define the two groups to be filtered. Enter the threshold by which expression of genes in one group should exceed that of the other group. Select the appropriate ‘greater-than’ or ‘less-than’ symbol to define which group should be more highly expressed than the other. Select ‘both’ to keep genes in which either group’s expression level exceeds the other by the defined threshold.

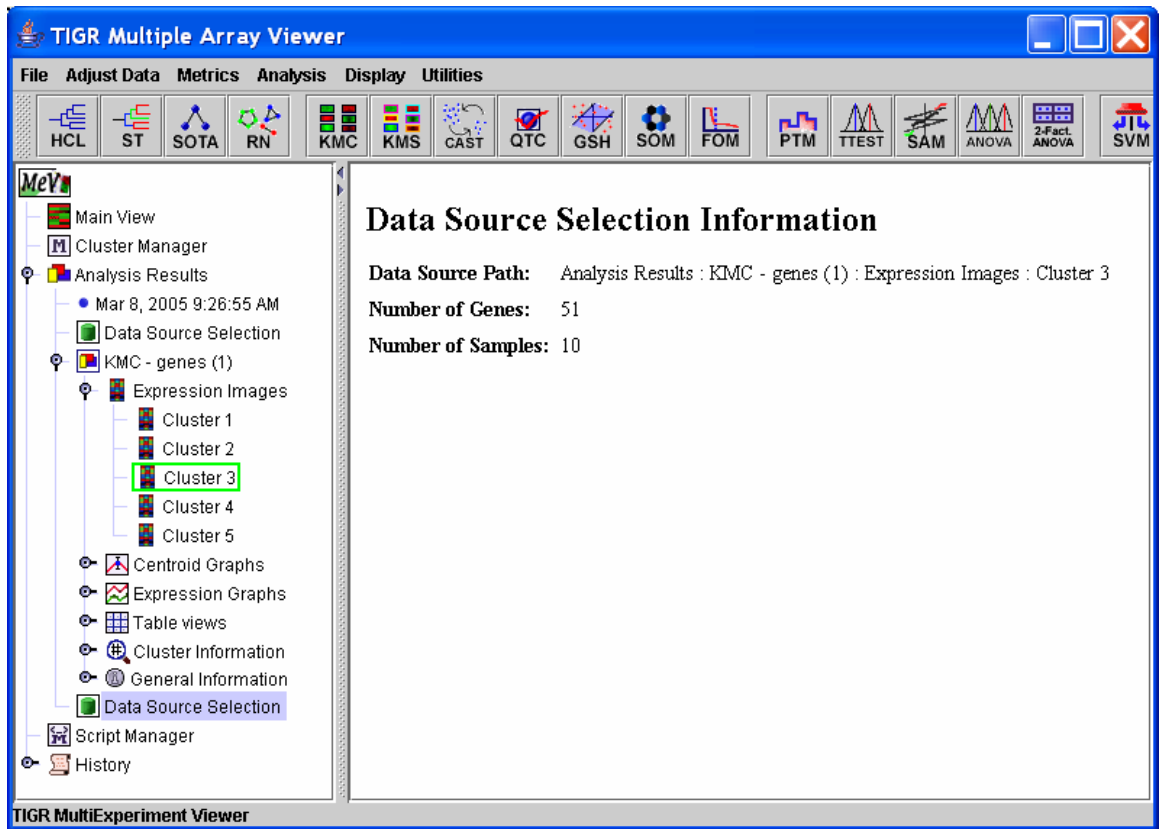
Adjust Intensities of Zero

This option is turned on by default. This means that if either (but not both) of the cy3 or the cy5 intensities for an element is recorded as zero, that intensity value will be reset to 1. In this case, the expression ratios will be calculated as cy5/1 or 1/cy3, depending on which value is zero, and the element is included in subsequent analyses. Sometimes, the user may desire this. However, the user should be aware that the expression ratios for such elements are spurious. You might want to turn this option off, if you want to eliminate all those elements from the analysis that have at least one zero intensity value.

If you deselect this option if either intensity is zero then the log ratio computed for analysis is set to a “Not-A-Value” flag (NaN) and it is not used during any analysis and appears as a gray element in the expression image. Note that with either option any elements with two zero intensities have the computed log ratio is set to the NaN flag since a log ratio cannot be computed.

5.4. Data Source Selection

An important aspect of data analysis or data mining is what might be called analysis branching. This involves the initial selection of a gene or sample set based on a preliminary screen or analysis technique such as a statistical method, and then taking this subset of elements and performing more detailed analysis to find constituent features such as prevalent expression patterns. A right click on a result viewer node in the result navigation tree will present a menu option (when applicable) to set the contained data as the primary data set for subsequent analysis. The selected data source node in the navigation tree will be highlighted by a green rectangle which indicates that this is the primary data source for downstream analysis. As an indication of data source change, a node is placed on the result navigation tree to indicate the source of the data and the number of genes and samples in the selected data set. Subsequent analysis runs will use only this data subset until a new data source is selected.



5.4.1. Data Source Selection Node (source node is marked with the green border)

6. Display Options

The graphical display of data and analysis results is one of MeV's strongest assets. This section of the manual describes options that include selection of gene and sample annotation to display, adjustment of expression image color schemes, and adjustment of expression image element size.

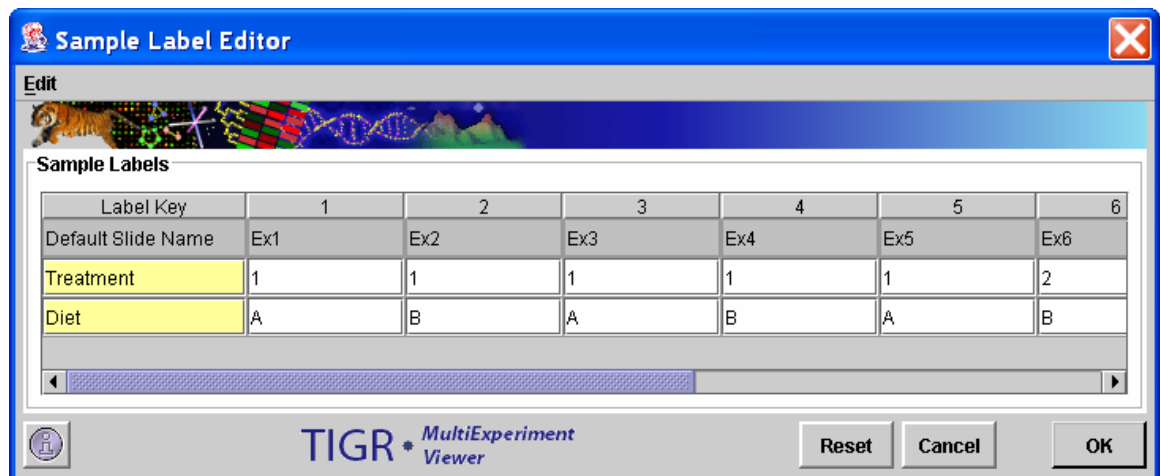
6.1. Sample Annotation

Changing Sample Labels

Samples in MeV can be labeled in various ways to indicate natural groupings based on experimental design. By default, samples are labeled with the file name or in the case of TDMS format files the names are in the header row. Sample labels can be selected from the *Select Sample Label* menu of the *Sample/Column Labels* menu of the *Display* menu. Expression images and expression graphs will label samples by this selected annotation field. TDMS format files can contain additional sample annotation as described in the file format appendix. Two other options for providing additional sample annotation are to use the *append sample annotation* option of the *utilities* menu or to use the Sample Label Editor as described below.

Editing Sample Labels and Sample Reordering

The Sample Label Editor is launched from the *Edit Labels/Reorder Samples* menu option. One key point is that if samples names are added, merged, or if samples are reordered, the changes can be captured by selecting the *Save Matrix* option from the *File* menu. This will save the loaded data in to a TDMS file and will preserve the added sample annotation and sample order, if altered.



6.1.1. Sample Label Editor

The primary function of the Sample Label Editor is to permit the modification of labels (attributes) associated with the loaded samples. Note that the first row in the table contains the default sample name. The Default Name cannot be edited nor can it be deleted. The second major function is to enable the order of loaded

samples to be rearranged. Two menus are available within the editor frame for performing editor actions. The two menus contain the same menu options. One is in the main menu bar and the other is available using a right mouse click with the mouse cursor over a table cell. **Note that the import of additional sample annotation can be done using the *Append Sample Annotation* option from the *utilities* menu.

Sample Reordering

The order of the columns in the table can be altered by clicking on the column header and dragging the column to the desired location. This action alone does not force the loaded data to be reordered. The table order can be imposed on the loaded data if the check box menu item is selected to *Enable Sample Reordering*. This ordering is imposed as soon as the dialog is dismissed.

Note: If analyses have been run on the loaded data, the reordering option is disabled. This is done as a precaution against misrepresentation of prior results that may have relied on the original ordering. Typically, the reordering of samples should be done just after data loading to place associated samples into an order that is reasonable based on study design.

Adding Sample Labels

The *Add New Sample Label* menu item is used to create a new row in the table. Note that the first cell in the new row is light yellow and is used to indicate a *Label Key*. This key will be placed into the *Sample Label* menu in the *Display* menu so that when selected this label type will be displayed for the loaded experiments. Once a Label Key is entered, a double click on the remaining cells will allow editing to add appropriate labels for each sample. Reopening the editor will allow you to alter any label or label key visible in the table except for the default primary label and key (shown in gray). Blank entries are allowed.

Merging Sample Labels

Occasionally it is convenient to merge several attributes or labels to produce a more informative label for each sample. To merge labels start by selecting two or more rows using ctrl-left click and then selecting the *Merge Selected Rows* menu item. A small list will be presented that can be used to order the selected labels before actual merging. Once attribute ordering is done, a new row is inserted in the table to display the merged labels and the merged label key.



6.1.2. Attribute Sorter for Merging Sample Attributes

6.2. Selecting Gene Annotation

The *Gene/Row Label* menu option from the *Display* menu is used to select a gene annotation field from among the loaded annotation types. Expression image viewers and other viewers that display gene annotation will be adjusted to display the selected annotation type.

6.3. Color Scheme Selection

The color scheme selection options pertain to the color gradients that are used to correlate color to an expression value for an element in expression viewers. Color provides a means to easily view patterns of expression. Three preset options can be selected from the *Display* menu's *Color Scheme* menu. The default is a double gradient, green-black-red, and displays under expression (relative to the reference) as green, over expression as red, and spots where there is little differential expression as black. A blue-black-yellow color scheme is available as an alternative scheme. The rainbow scheme is a third selectable option.

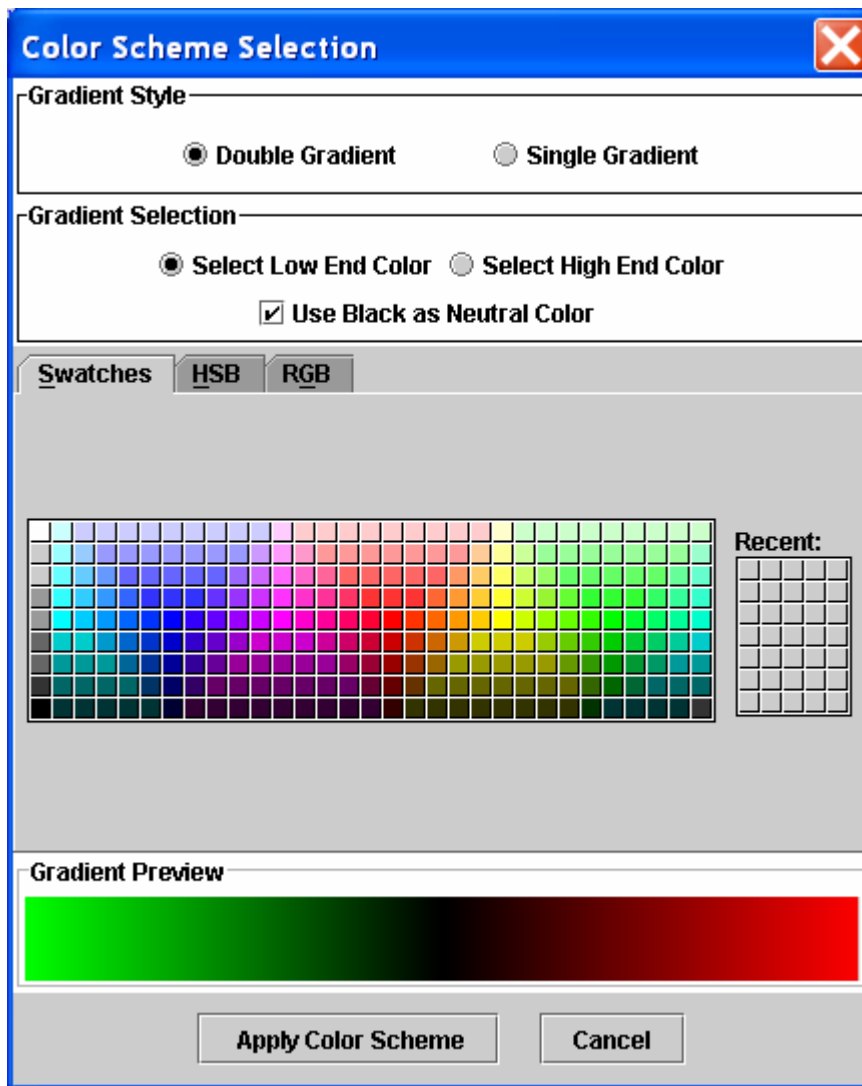
Custom color schemes can be created by selecting the *Custom Color Scheme* option from the *Color Scheme* menu. The form will provide several options:

Gradient Style: Allows the selection of a double gradient or a single gradient. In some cases a single gradient is preferred if values are not compared to a reference.

Gradient Selection: This panel selects the endpoint color that is being selected and allows for the center color on a two gradient scheme to be black or white.

Color Selector: This area presents controls for selecting endpoint colors.

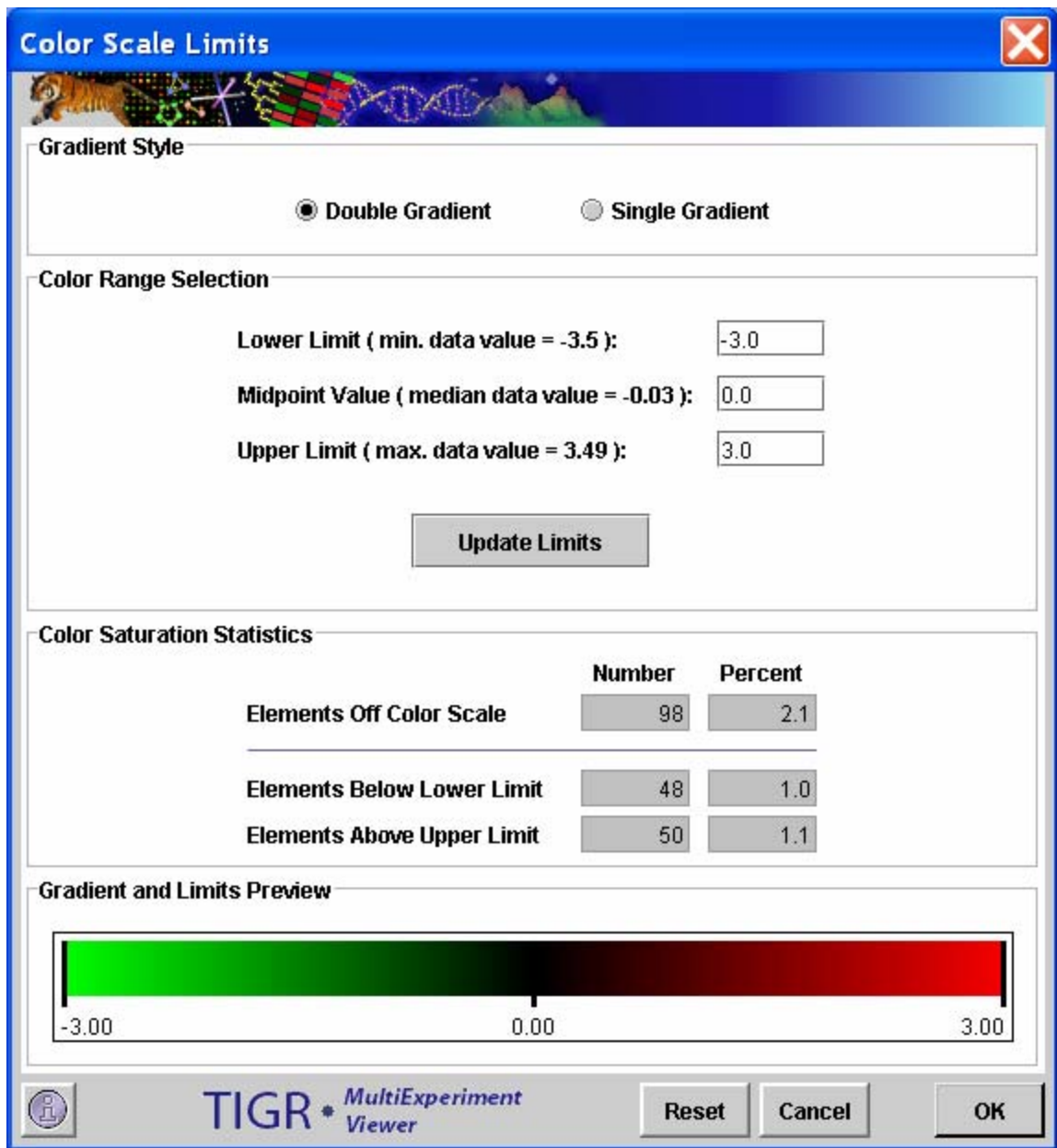
Gradient Preview: This area displays the current settings.



6.3.1. Color Scheme Selection Dialog

6.4. Setting Color Scale Limits

Expression images convey expression levels by converting the numeric expression value ($\log_2(A/B)$ or absolute expression value) as a color that is extracted from the color gradient. By setting numeric limits or endpoints on the displayed gradient (seen in the preview panel above and at the top of expression image viewers), expression values within the limits can be displayed as a color selected from the gradient image. The *Set Color Scale Limits* option from the *Display* menu provides a dialog to set the limits of the color gradient.



6.4.1. Color Scale Limits Dialog

The color scale limits dialog, like the color scheme dialog also provides the choice of using a double or single gradient as appropriate.

The color range panel provides input boxes to set lower and upper limits for the color gradient and a midpoint when using a double gradient. The labels next to the input boxes display the minimum, maximum, and median expression values as a rough guide when setting color scale limits. When the *Update Limits* button is pressed the *Color Saturation Statistics* panel (below) and the gradient preview panel are updated. The limits are also conveyed to MeV and the current viewer is updated to reflect the limits. This permits you to adjust and update the limits while viewing the affect of the new limits on the expression image.

The *Color Saturation Statistics* panel displays the number of spots that are beyond the limits of the color scale. These elements would be represented in the

expression image as the saturated endpoint colors. This panel also reports this information as a percentage of elements that are saturated relative to the total number of elements. These numbers allow you to adjust the limits such that the expression is spread across the gradient with a limited number of saturated (off scale) elements.

The reset button of this dialog will reset the limits in the dialog to the original limits present when the dialog was opened and MeV has its limits rolled back to the original limits.

6.5. Element Appearance

The final two options in the Display menu are options to alter the element size and to draw or omit borders around each element. Element size can be selected from among four preset options or a customized size can be selected by entering an element height and width.

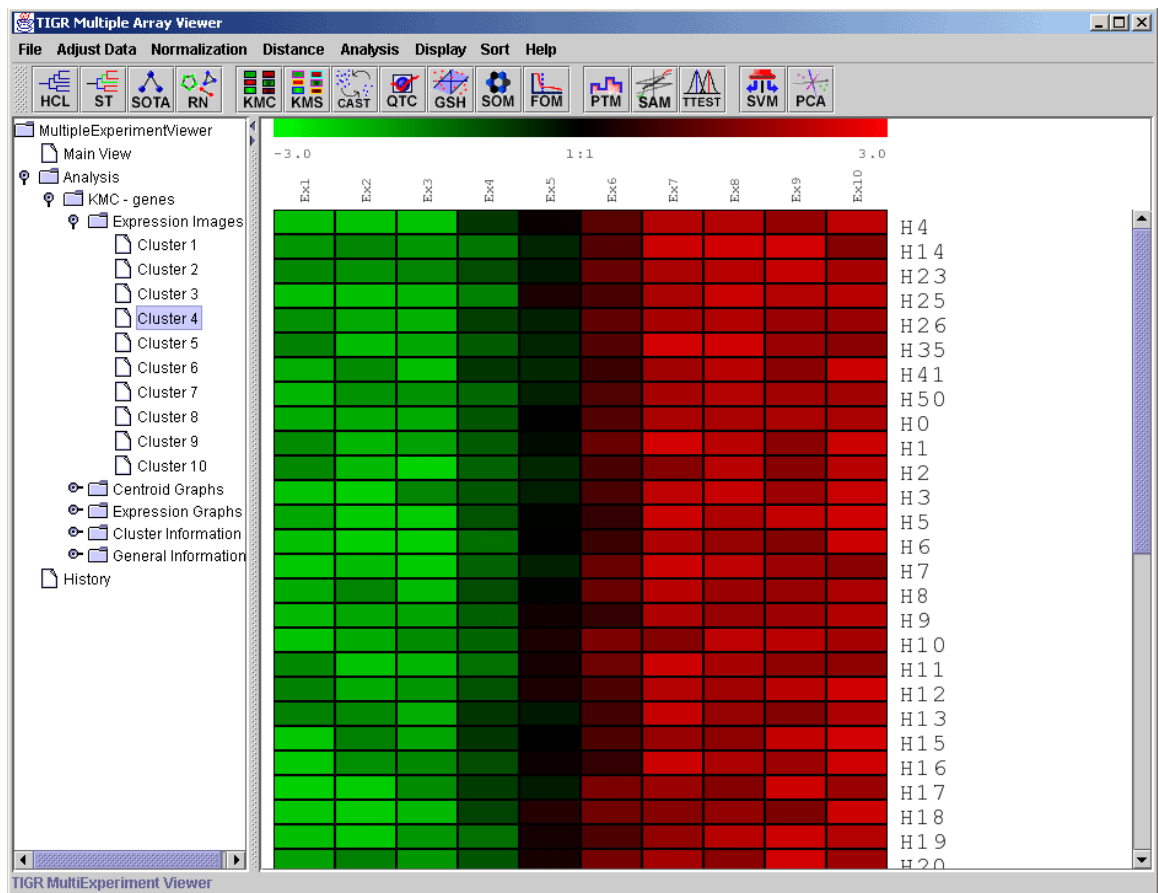
7. Viewer Descriptions

7.1. Overview

Viewers are the graphical displays used by MeV to present the results of the modules' calculations. The viewers will appear as a subtree under the module's Result Tree within the main navigation tree. The viewers listed here are those which are used by more than one of the MeV modules. Custom viewers, used by only one module, are described in that module's section.

7.2. Expression Images

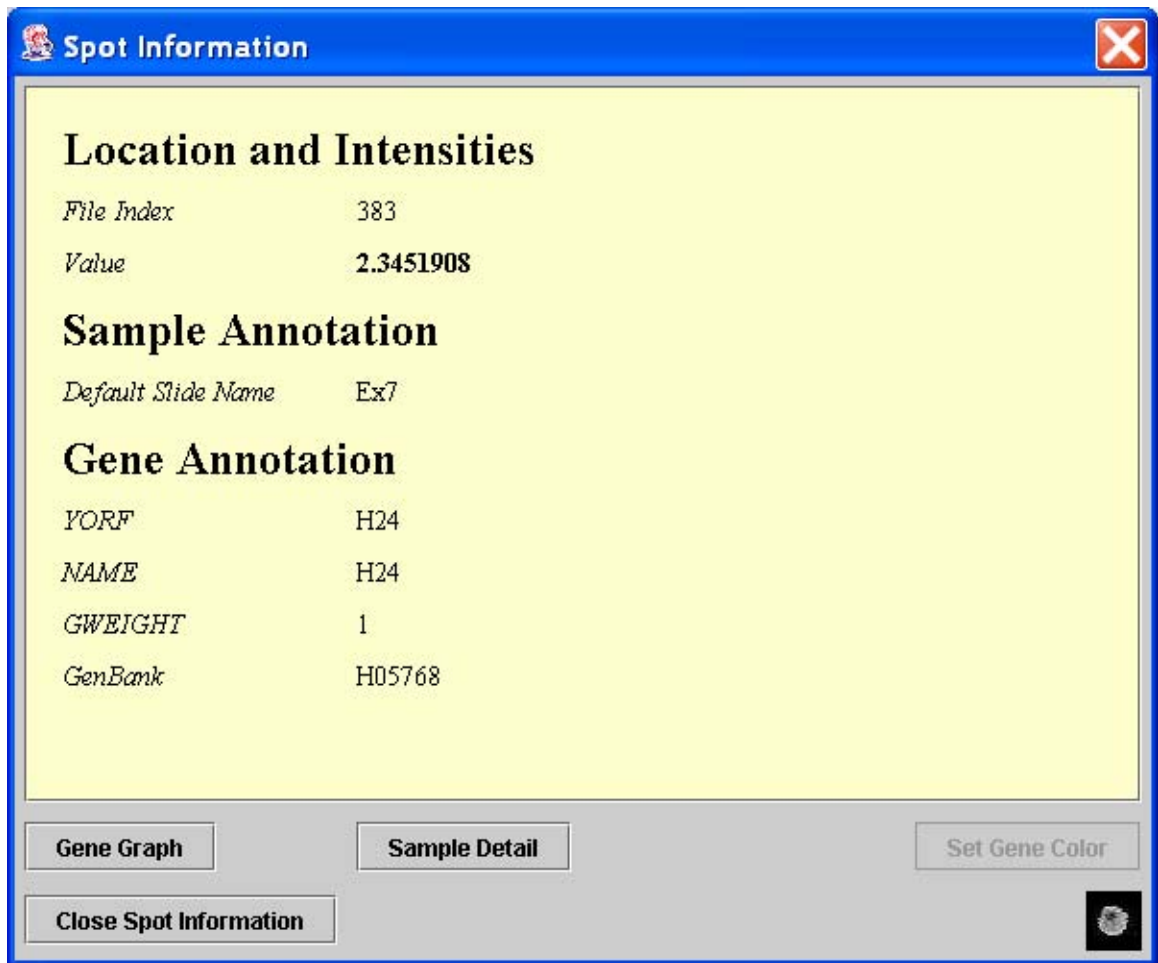
This viewer is used in the main window of the Multiple Experiment Viewer as well as in most of the modules (7.2.1). It consists of colored rectangles, representing genes, in a matrix. Each column represents all the genes from a single experiment, and each row represents the expression of a gene across all experiments. The default color scheme used to represent expression level is red/green (red for overexpression, green for underexpression) and can be adjusted in the *Color Scheme* dialog in the *Display* menu.



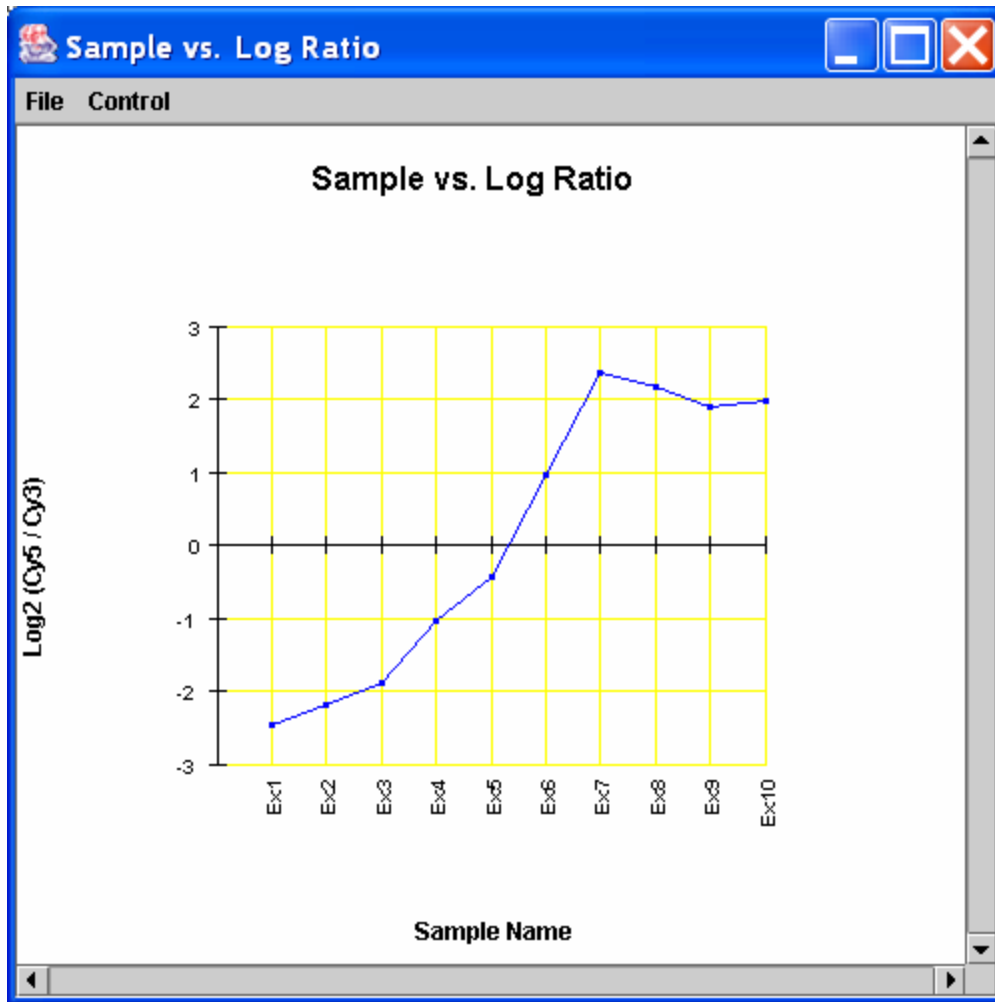
7.2.1. Expression Images Viewer

Clicking on any of the rectangles in this view will open a window reporting information for that spot (7.2.2). This window contains a great deal of information, including expression values and row/column coordinates. Clicking

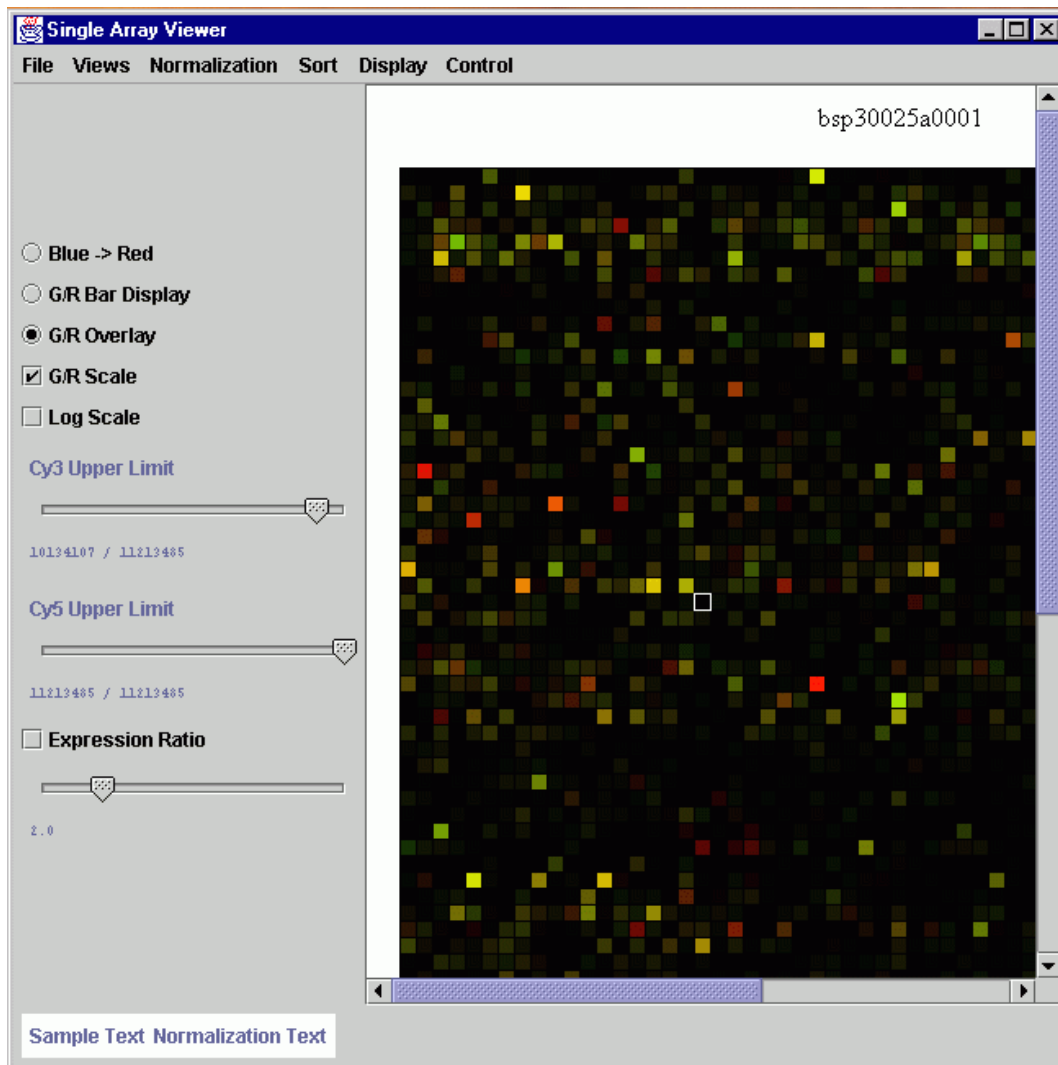
the *Gene Graph* button will open a window containing a graph of this gene's expression level across all samples (7.2.3). To view a Single Array Viewer displaying the entire experiment of which this spot is a part, click the *Sample Detail* button in the *Spot Information* window (7.2.4). The *Set Gene Color* button has not been enabled in this version of MeV.



7.2.2. Spot Information Window



7.2.3. Gene Graph Window



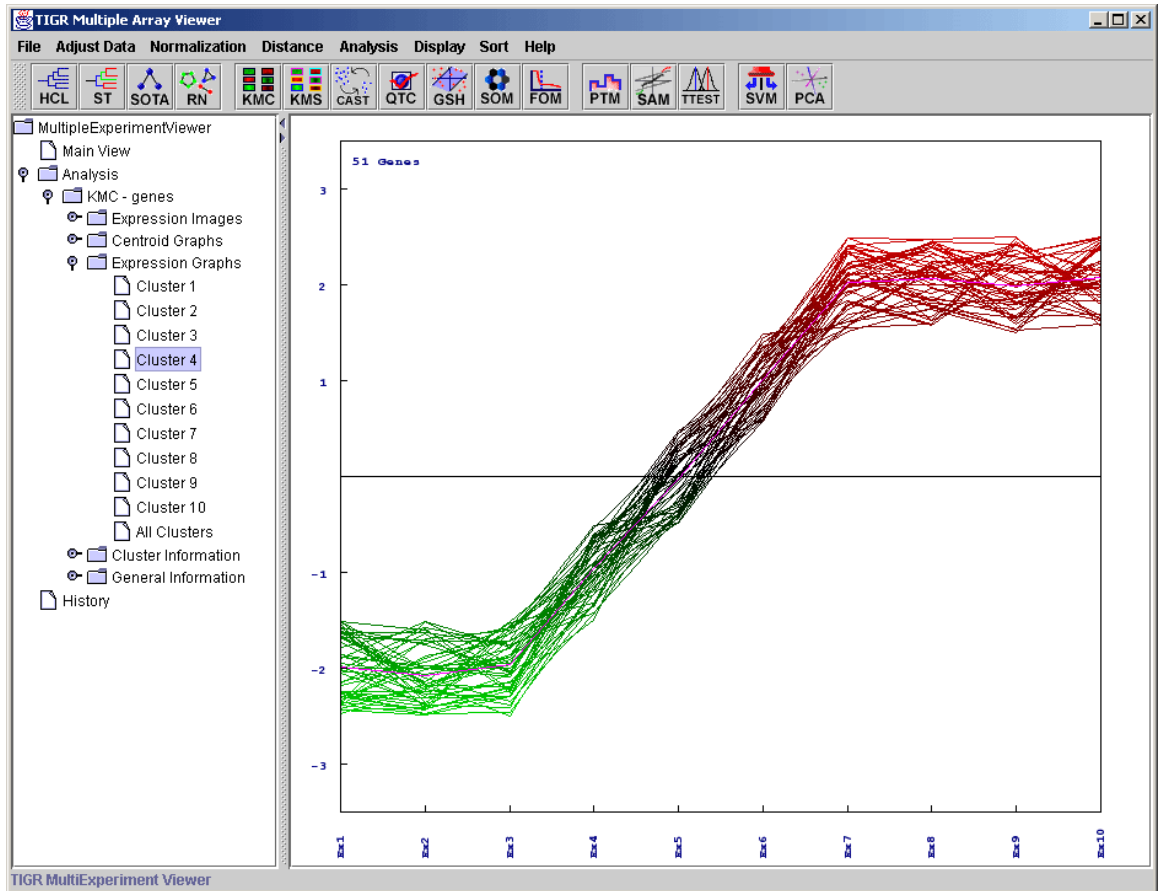
7.2.4. Sample Details displayed in a Single Array Viewer

7.3. Expression Graphs

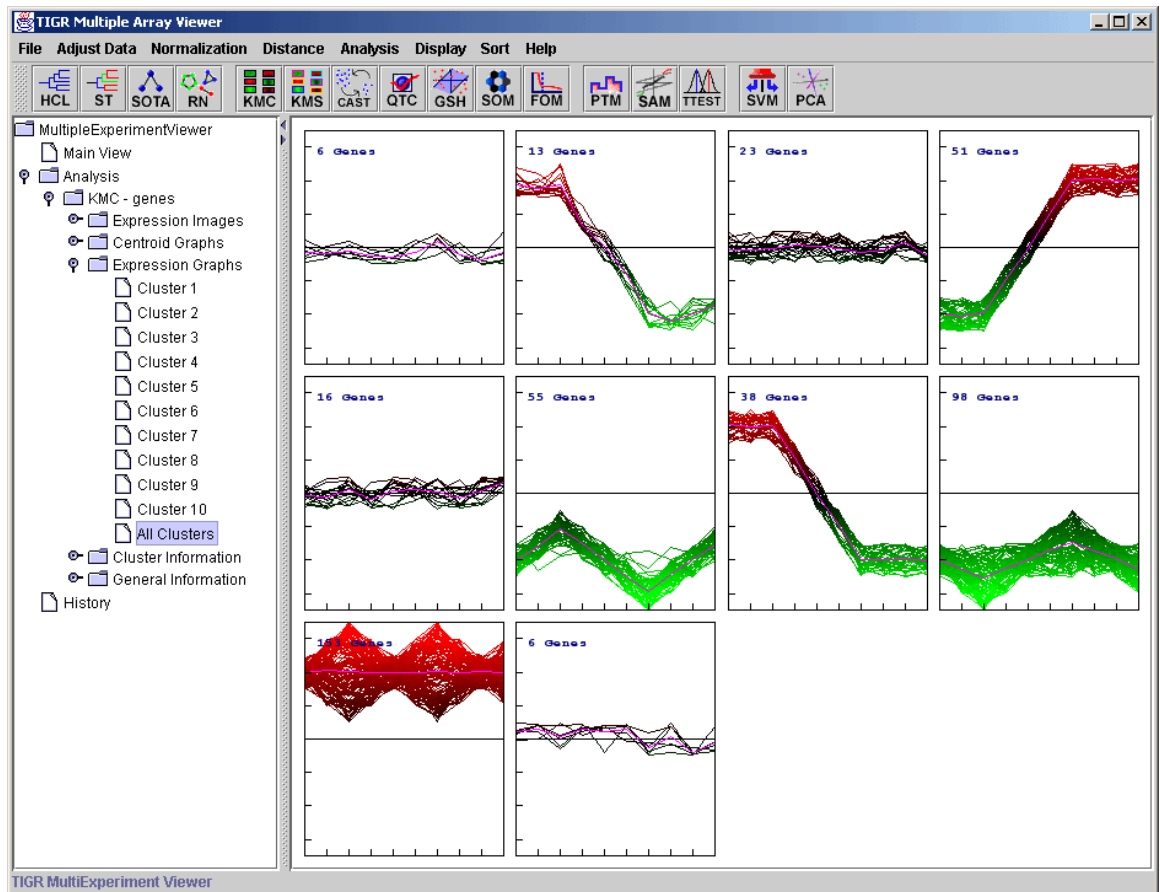
This viewer displays graphs of the expression levels of each gene across the experimental conditions (7.3.1, 7.3.2). By default, the line color is gray or colored according to the cluster membership. The lines can be set to display a gradient coloring option in which the lines are colored according expression level. This option is found in the *Display* menu under *Color Scheme* and selecting the *Use Color Gradient on Graphs* option will toggle this feature on and off.

The mean expression levels of genes in the cluster are shown as a centroid graph overlaid on top of the individual expression graphs.

By default, the range of the Y-axis for an expression graph is the same for each viewer produced from an analysis, regardless of which subset of expression graphs is displayed. The X-axis ($y = 0$) is centered on the Y-axis, and the Y range is set to the distance of the element in the entire data set which is furthest from zero. Use the *Change Y Axis* option in the right-click menu to change the Y range to the maximum distance from zero only for elements within the current expression graph. This option allows the expression graph to expand to increase the resolution of expression changes within a particular cluster.



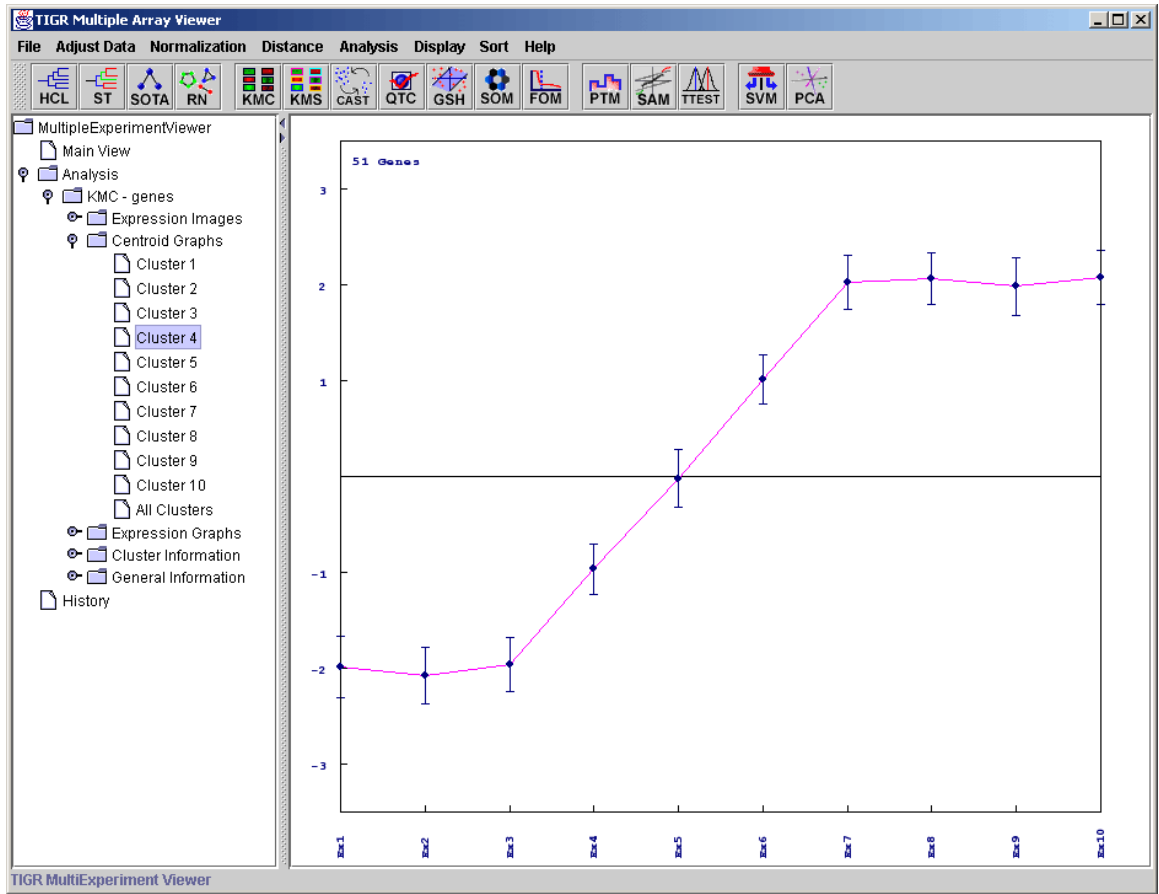
7.3.1. Expression Graph of one cluster (with gradient color selected)



7.3.2. Expression Graphs of all clusters

7.4. Centroid Graphs

This viewer is very similar to the Expression Graph Viewer, except that line graphs for individual genes are omitted, leaving only the centroid graph. Error bars represent the standard deviation of expression within the cluster (7.4.1).



7.4.1. Centroid Graph

7.5. Table Views

These views show element annotation and, where appropriate, auxiliary information such as element-specific statistical information for the elements in a cluster.

7.5.1. Gene cluster table view

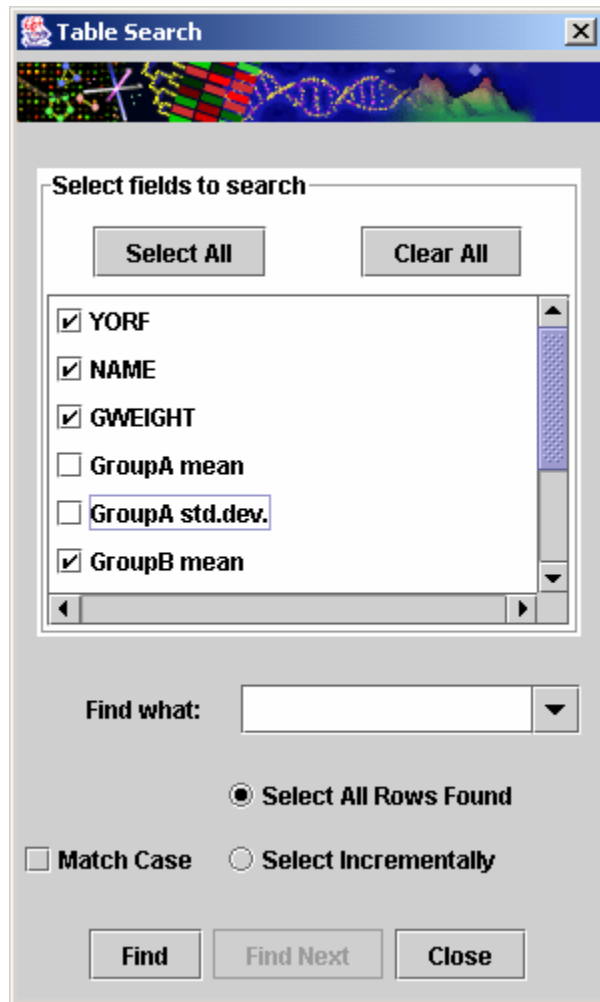
Columns can be dragged horizontally across the table to change their relative ordering. Successive clicks on the header for any column will sort the rows in ascending or descending order of the entries in that column. Sorting of the “Stored Color” column will bring together elements that have been stored with the same cluster color. CTRL-clicking on any column header will sort the table in the original order of elements in that cluster, an option that is also available from the right-click menu (below).

Right-clicking on the table view brings up a menu containing the options available from the right-click menus of other types of viewers, with a few important modifications and additions as follows. In addition to storing the entire cluster and launching the entire cluster as a new MeV session, as in the other types of viewers, users can also select a subset of rows in the table for these operations. Even individual elements can be stored and tracked one at a time. Contiguous row selections can be made by dragging the mouse over these rows, or by selecting the first row of the interval and then SHIFT-clicking on the last row of the interval (in Windows). Non-contiguous rows may be selected by CTRL-clicking on the desired rows.

To delete a cluster stored from this viewer (i.e., either all the rows in the table or a subset of them), select the rows corresponding to the stored cluster and choose the

“Delete ...” option on the right-click menu. Please note that for successful deletion, the stored cluster should have been created within the current algorithm run that the table view belongs to, and the row selection should exactly correspond to the rows in that cluster.

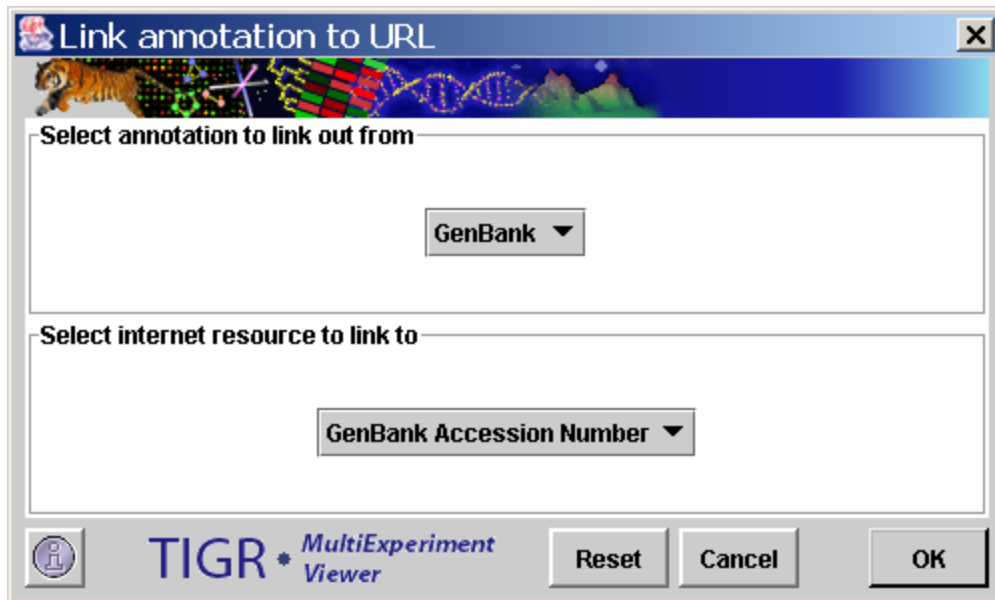
The “Search” function on the right-click menu pulls up a search dialog, as shown below:



7.5.2. Table search dialog

Users may search any combination of columns for specific terms. Other options on this dialog are self-explanatory.

The *Link to URL* option allows users that are connected to the Internet to launch a web browser to open a page related to the selected gene. First select a gene of interest (a row) and then click on the annotation key to use as the gene identifier. MeV will attempt to construct a URL for the selected annotation key by checking a URL configuration file. If it is unable to definitively decide what the selected annotation type is, a dialog will be presented to allow you to select an annotation field name (column header name) and an external resource to open.



7.5.3. Annotation/URL Association Dialog

Within the *config* directory there is a file (**annotation_URLs.txt**) that contains entries to support this feature in mev. This file can be modified with new external resources as needed, and the existing fields can also be modified, with the exception of the UniGene field (see below). The file is arranged such that each row represents a different annotation label:resource pair with the following tab delimited fields (\t indicates a tab).

```
<Annotation Label> \t <URL> \t <Resource Name>
```

The <Annotation Label> indicates the name of the annotation field that is loaded into mev. This is in the header of the annotation file or TDMS file. For example, if in some files you call GenBank GB# and in other files it is referred to as GenBank then two entries in this annotation/url file can be made, one for each key. If an input annotation file had the GenBank numbers identified as gb# then the Annotation/URL Association dialog would launch to allow you to manually indicate the desired external resource.

The <URL> entry is the URL for the resource. Look in the file for examples. It is important to note that the URL has a section labeled 'FIELD1'. This is a placeholder to indicate where the gene identifier should go when constructing the URL. The UniGene URL is the only exception, requiring two variable fields called FIELD1 and FIELD2, which are parsed out of the UniGene identifier. The UniGene Annotation Label is required in the **annotation_URLs.txt** file to correctly parse UniGene Ids; therefore, please do not modify this line.

7.6. Common Viewer Activities

A right click within most result viewers will launch a menu that is specific to the currently displayed viewer. Some of the viewer specific options, such as searching in Table Viewers have been discussed however there are several common viewer options that are shared among the common viewer types.

Store Cluster

The Store Cluster option will save the currently viewed cluster to a cluster manager. This feature will be described in detail in section 8. The main use of this feature is to assign a color to the elements in the cluster so that the location of the marked elements in future results can be assessed.

Launch New Session

The Launch New Session option takes the elements in the current viewer and opens a new multiple array viewer with just those elements represented. This is useful for quickly extracting elements of interest and further characterizing their expression.

Save Cluster

The Save Cluster option saves the currently viewed cluster to a file. The expression values and annotation for the current cluster are saved in a format that can be reloaded as a Tab Delimited, Multiple Sample (TDMS) format file. For several statistical methods this output includes statistics such as F-values or T-values, and p-values depending on the particular statistical algorithm applied.

Save All Clusters

This saves all clusters from a clustering result as described above but where the file name has a cluster index appended to indicate cluster id.

Delete Cluster

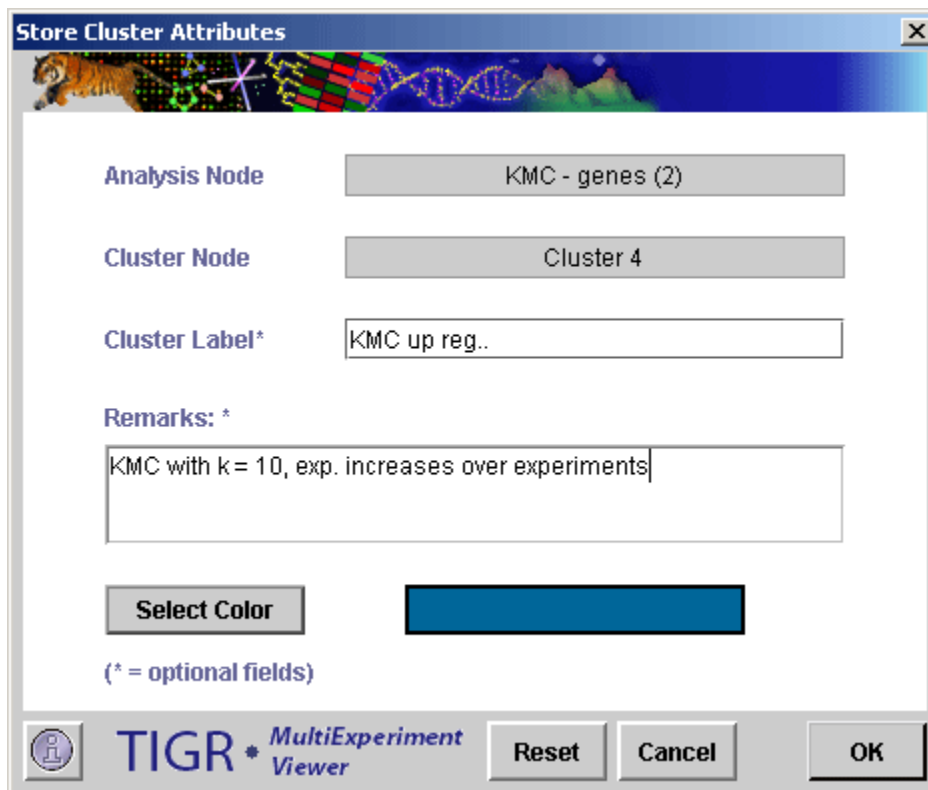
This method is used if the Store Cluster method has been used to store the cluster to the repository. This is a remote method to remove the cluster from the repository. The cluster table viewer in the cluster manager also has a means to remove single or multiple clusters.

8. Working with Clusters

The analysis modules available in MeV subdivide genes or samples into clusters by unsupervised techniques, statistical methods, classification algorithms, or biological relationships. These partitioned sets of elements are then individually displayed in one of the standard cluster viewers.

8.1. Storing Clusters and Using the Cluster Manager

Clusters of interest can be stored to a repository from the basic cluster viewers by opening a menu by right clicking in the viewer and selecting the *store cluster* option. Once a cluster is stored the *Cluster Manager* node on the result navigation tree will contain a list of stored clusters. Gene clusters and sample clusters are maintained in separate spreadsheets which are viewable from the Cluster Manager node. When storing a cluster to the repository an input dialog is presented which allows for three user defined fields to be associated with the cluster. Two optional text fields are used to capture a cluster name and a description of the algorithm or interesting features of the cluster. The third user input is a color used to identify genes or experiments which are members of the clusters. These colors can be tracked while performing analyses so that clustering consensus can be established.



Store Cluster Attributes

Analysis Node KMC - genes (2)

Cluster Node Cluster 4

Cluster Label* KMC up reg..

Remarks: *
KMC with k = 10, exp. increases over experiments

Select Color [Blue Color Swatch]

(* = optional fields)

TIGR MultiExperiment Viewer [Reset] [Cancel] [OK]

8.1.1. Cluster Attributes Dialog

The cluster tables contain the following columns:

-The *Serial Number* is a unique number which is sequentially assigned to easily identify a particular cluster.

-The *Source* field describes whether the cluster source was an algorithm, a cluster operation, or some other means of selecting a group of elements.

-The *Algorithm Node* field identifies the algorithm used, if the source was an algorithm, and includes the navigation tree result index (in parentheses).

-The *Cluster Node* field identifies the specific cluster node under the Algorithm Node from which the cluster was stored.

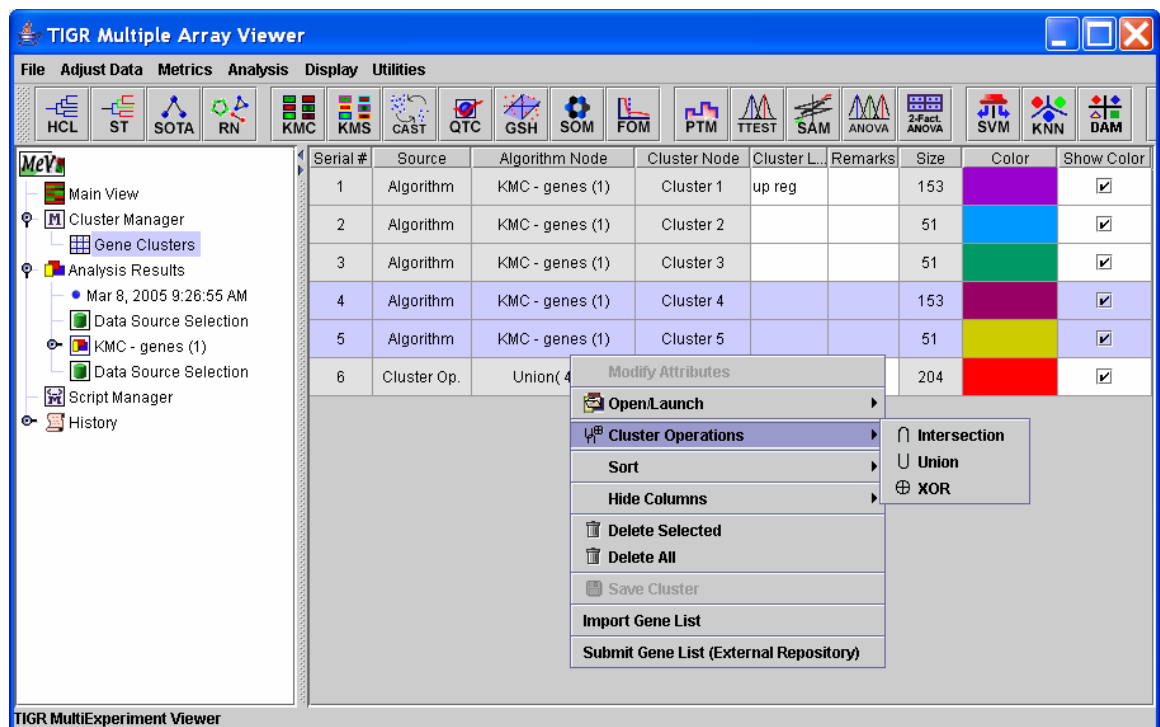
-The *Cluster Label* is an optional user defined name for the cluster.

-The *Remarks* field can be used to contain details about the process used to create the cluster or specific features of interest in the cluster.

-The *Size* field shows the number of elements in the cluster.

-The *Color* displays the user defined color for the cluster.

-The *Show Color* check box allows you to show or repress the displayed color. This option can be useful when visualizing cluster intersections in viewers. Selecting only one cluster color to view can simplify interpretation.



8.1.2. Cluster Table with option menu open

The spreadsheet allows single or multiple row selection (by holding down the control key when left clicking the mouse). A right click with one or more rows selected will display a menu that contains several options detailed below.

-The *Modify Attributes* option allows the user to modify cluster label, remarks or the cluster color by displaying the input form with the current settings displayed.

-The *Open/Launch* menu has two options, *open cluster viewer* will pull up the source cluster viewer. The second option is *launch mev session* which opens a new multiple array viewer containing only the data from the selected cluster or the union of the members of several clusters if several clusters are selected.

-The *Cluster Operations* menu allows for three possible operations to be performed if two or more clusters are selected. The *Union* operation combines the members of the selected clusters and stores the resulting cluster on the list. Elements represented in more than one cluster of the input clusters are only represented once in the output cluster. The *Intersection* operation takes the elements from two or more clusters and produces a cluster containing all elements which are common to all clusters. The *XOR* (exclusive OR) operation produces a cluster containing elements that are members of one cluster or another but not members of more than one cluster.

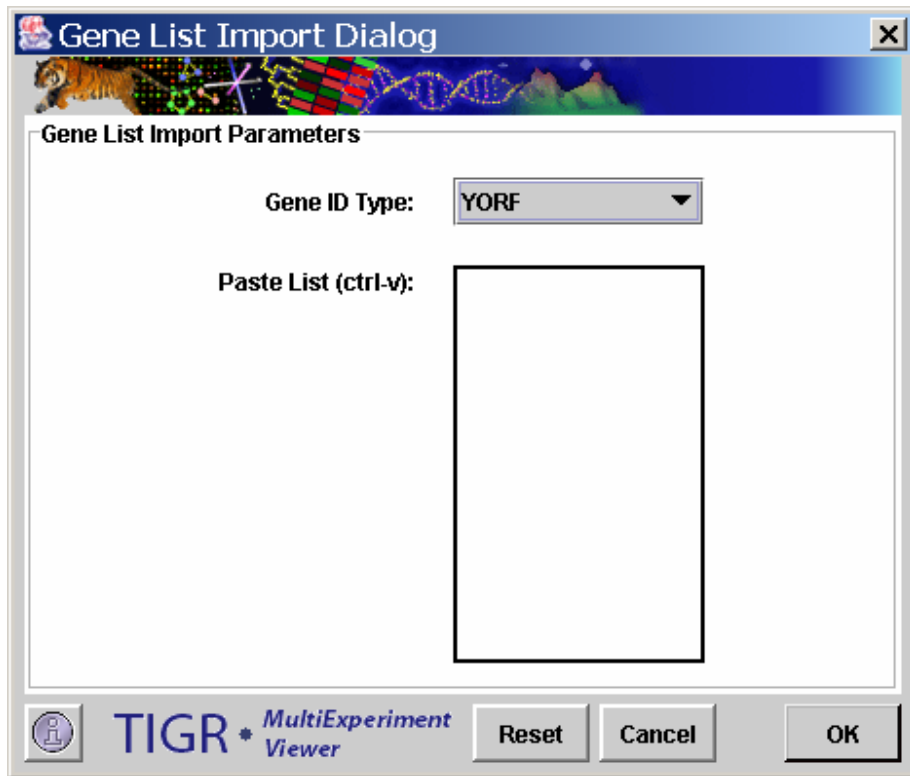
-Options also exist to delete selected clusters or all clusters in the list as well as to save a selected cluster to a specified file.

Deleting clusters can be performed by selecting a single or multiple clusters in the cluster table or by selecting delete public cluster option from the menu in the viewer which contains the cluster.

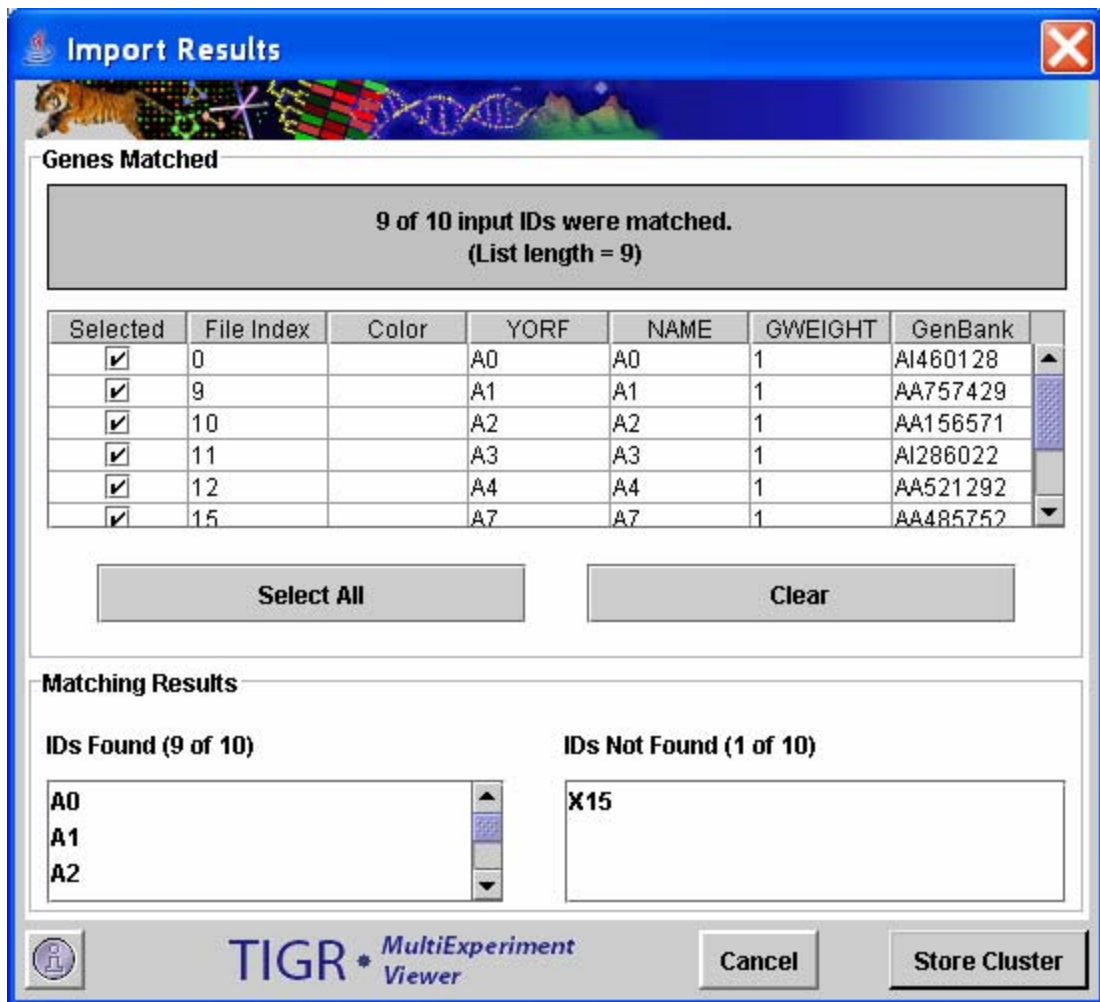
You can also *Save Cluster Data* to a tab-delimited text file. Selecting this option from the right-click menu will cause a file chooser to appear. Select a file name and a place to save row/column data, log ratio expression values, and (optionally) Cy3 and Cy5 values for each gene in the cluster. Selecting *Save All Clusters* will allow you to save the genes in all clusters in a similar way. This option is available from the cluster table as well as in the viewer.

One additional option is the option to delete all gene clusters or sample clusters. These global operations which effect all colored clusters is selected from the *Utilities* menu in the multiple array viewer by selecting *Delete All Gene Clusters* or *Delete All Sample Clusters* or can be done from the cluster tables.

The *Import Gene/Experiment List* allows one create a cluster based on supplied identifiers. Identifiers belonging to the cluster are pasted into the text area. The drop down list indicates the type of annotation being loaded. An intermediate dialog (*Import Result Dialog*) will appear to display the results of the import and to allow you to select a subset of the identified elements before saving the elements as a cluster. After review of identified elements a *cluster attributes dialog* will be presented so that a cluster name, description, and color can be defined for the new cluster.

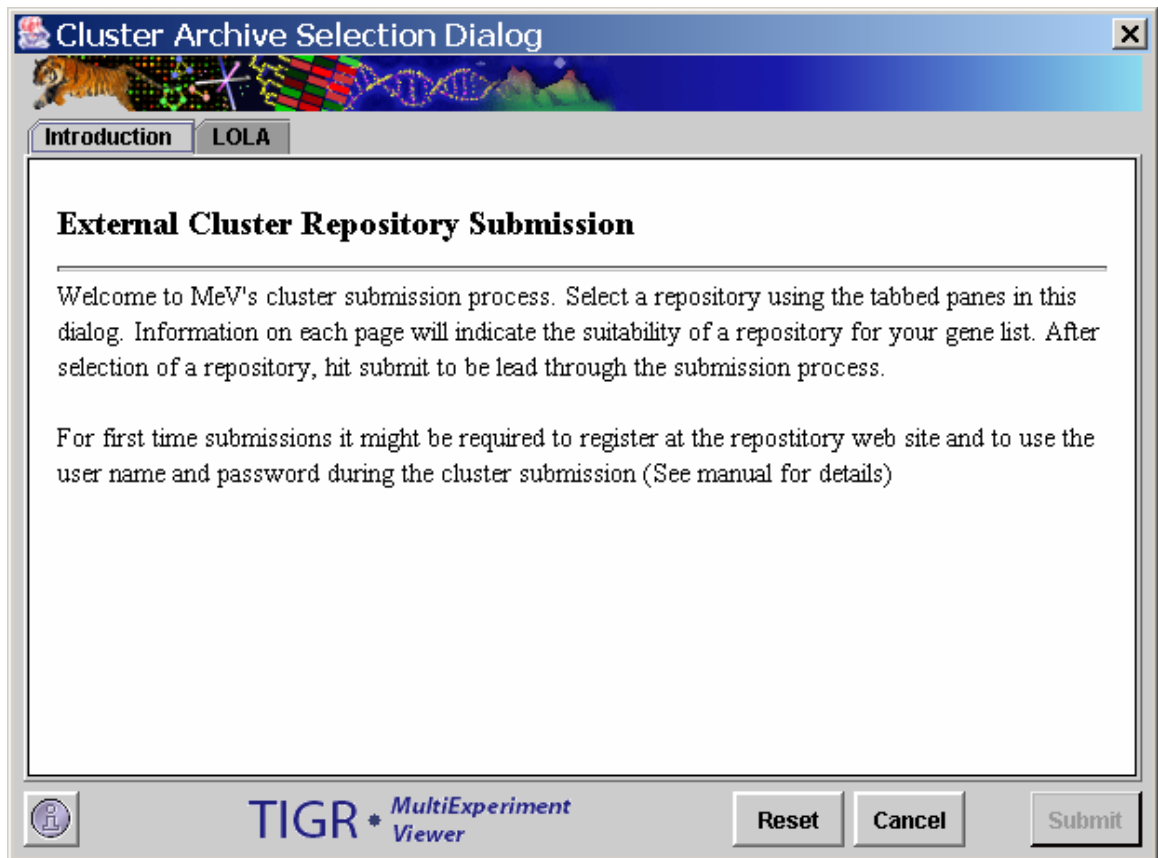


8.1.3. List Import Dialog



8.1.4. Import Result Dialog

The *Submit Gene List (External Repository)* option permits gene lists to be submitted to external database repositories. The data, including the gene identifier types, possibly organism under study, and other factors can determine if the current data is suitable for submission to a repository. To submit a cluster, first select the cluster to submit from the cluster table by left clicking on the appropriate row. Modify the cluster's name and description attributes if required since they may be important for cluster identification within the external repository. Select the *Submit Gene List (External Repository)* menu option to start the submission process. The initial dialog provides a set of panels indexed by repository name. At the time of the version 3.0 release, only one repository was available for submission but more can be added in future releases. On each repository page a general description of the repository is given as well as some guidelines and requirements that should be met prior to cluster submission. Please adhere to the requirements of the specific repository. Once the repository is selected the user will be guided through the submission process unique to the selected repository. Once a repository is selected hit the submit button to be lead through the submission process for the selected repository.



8.1.5. Repository Selection Dialog

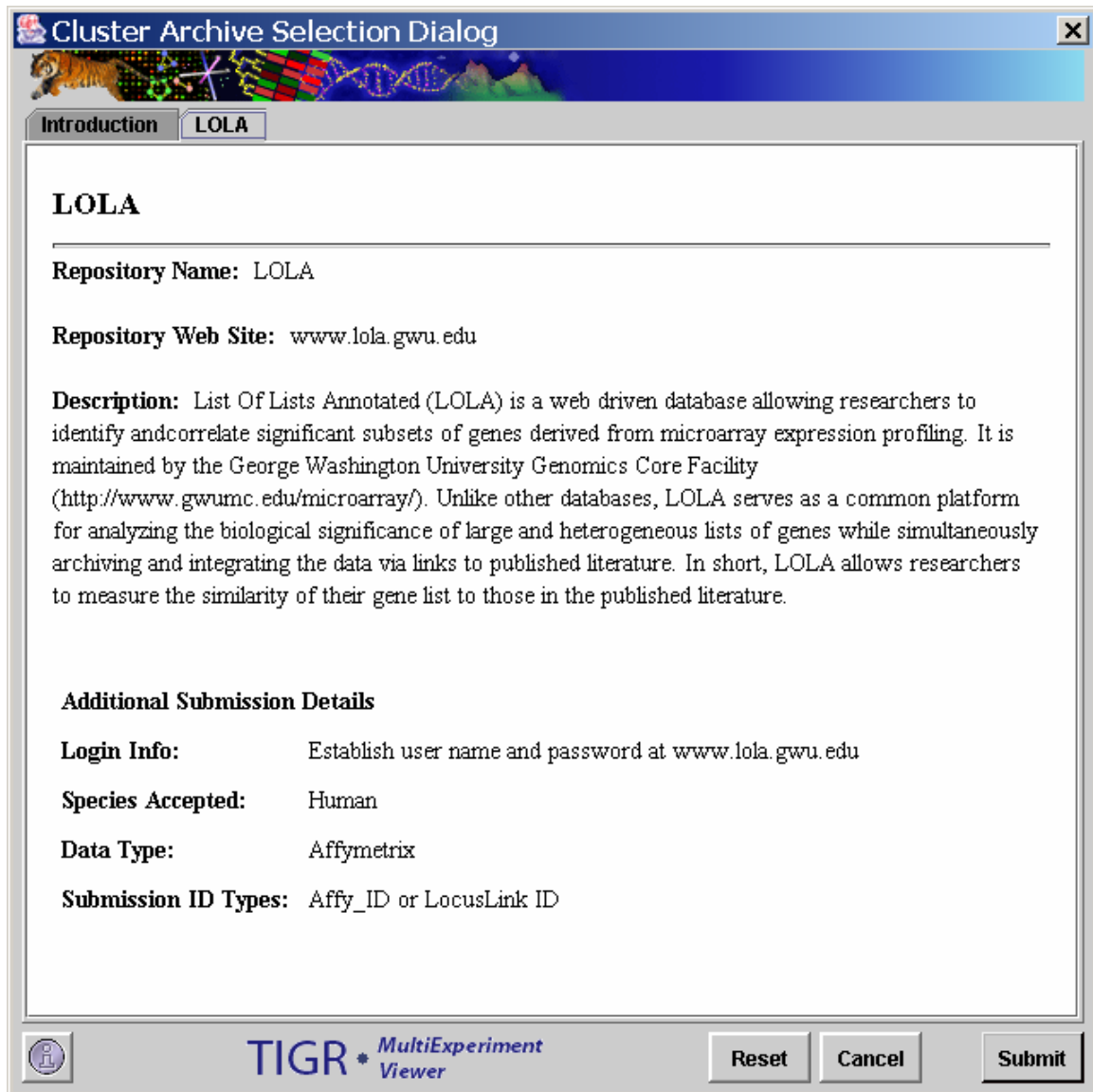
Note that some repositories may require user accounts with passwords to be established before the submission can be enabled. If a password is required, a dialog will be provided to enter a user name and password for the repository. Note that one can enter login information in the repository configuration file to have MeV remember the user name and password. To enable this one can edit the `archive_submission_config.xml` file in the config directory. Open this file in a text editor and move to the repository entry for the specific repository to add user information. Alter the `user` tag from:

```
<user/>
```

to:

```
<user user_name="your_name"  
password="password" email="your_email"/>.
```

A sample of this tag is in the xml file as a guide. Note that some repositories might not require an email or might in that case the email attribute can be removed from the tag.



8.1.6. Sample Repository Information Page

9. Utilities Menu

Search Utility

The Search feature permits the user to search the data for genes or samples for a search term given search criteria. Once the search is complete, the elements are returned in a table. Navigation shortcuts provide a means to open cluster viewers that contain the elements found in the search.

The search initialization dialog allows the option of finding genes or samples. The search criteria include a search term, a selection to make the search case sensitive, and a selection to permit the search term to be an exact match or simply a contiguous portion of a larger annotation term.

The screenshot shows a dialog box titled "Global Search" with a close button in the top right corner. Below the title bar is a decorative banner featuring a tiger, a DNA double helix, and a chromosome. The dialog is organized into three main sections:

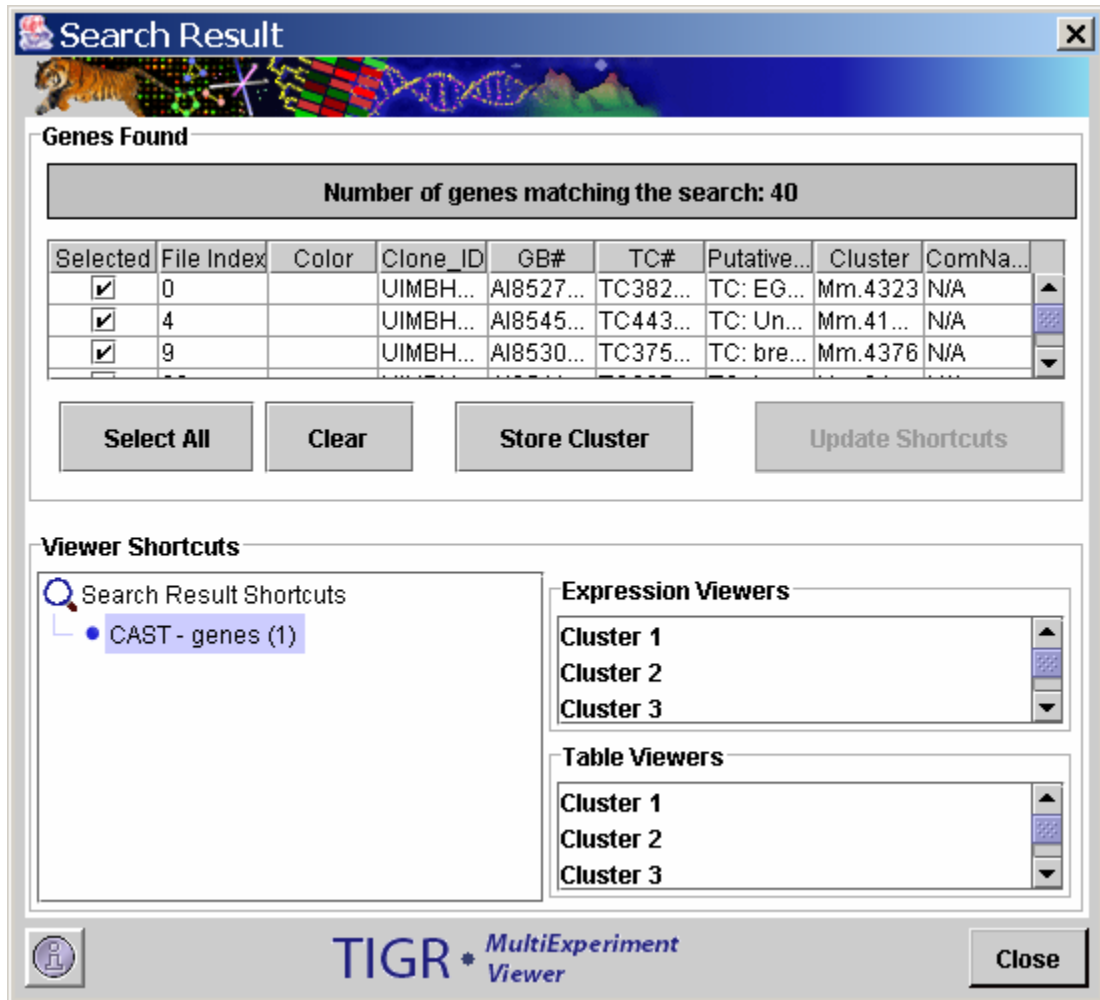
- Search Mode:** Contains two radio buttons. "Gene Search" is selected, and "Sample Search" is unselected.
- Search Term:** Includes a text input field labeled "Search Term:" and two checkboxes: "Case Sensitive" and "Exact Match", both of which are currently unselected.
- Searchable Fields:** Features two buttons, "Select All Fields" and "Reset Selection", positioned above a list of checkboxes. The checked fields are:
 - Clone_ID
 - GB#
 - TC#
 - Putative Role
 - Cluster
 - ComName

At the bottom of the dialog, there is a logo for "TIGR MultiExperiment Viewer" and three buttons: "Reset", "Cancel", and "OK".

9.1.1. Search Initialization Dialog

The search result is presented in a new window that is split into an upper section that contains the list of genes or samples identified as matching the search criteria and a lower section providing shortcut links to cluster viewers that contain the identified samples or genes.

The upper panel displays the list of genes or samples retrieved by the search. The list of genes or samples found could include some entries that not of interest. Elements in the list can be deselected using the checkboxes. Clicking on the *Update Shortcuts* button will produce a new search result window with just the previously selected entries and the associated viewer shortcuts. This allows one to prune unwanted elements out of the search result. The *Store Cluster* button will store the selected items as a cluster and assign a user selected color. Note that this is one method to help identify the elements found during the search within the cluster viewers.

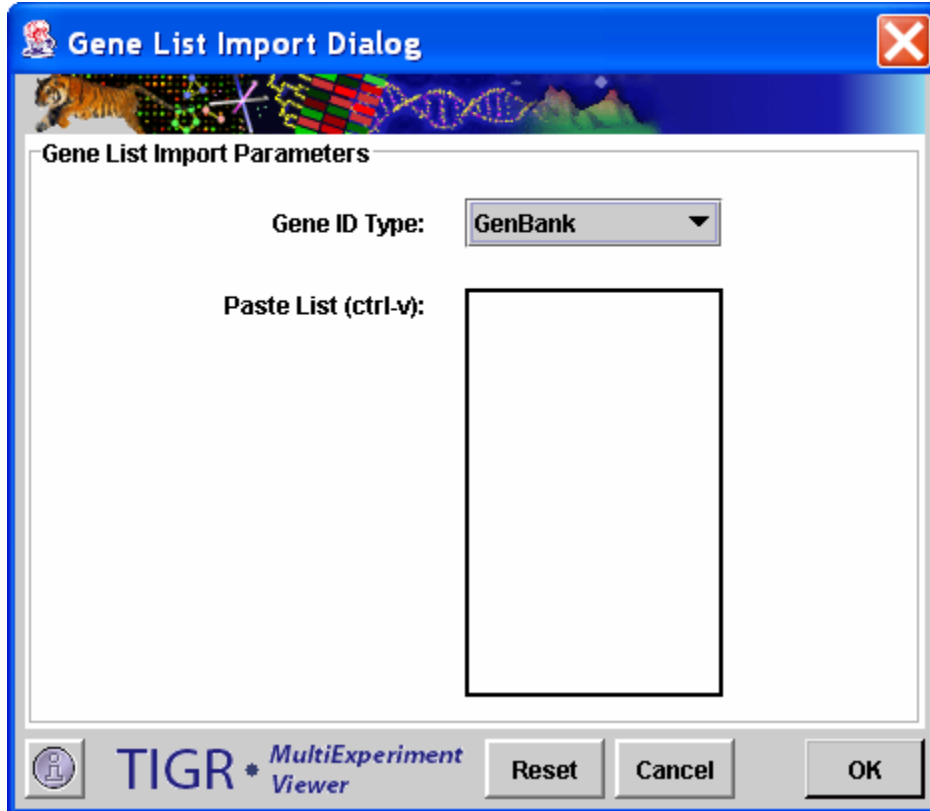


9.1.2. Search Result Window

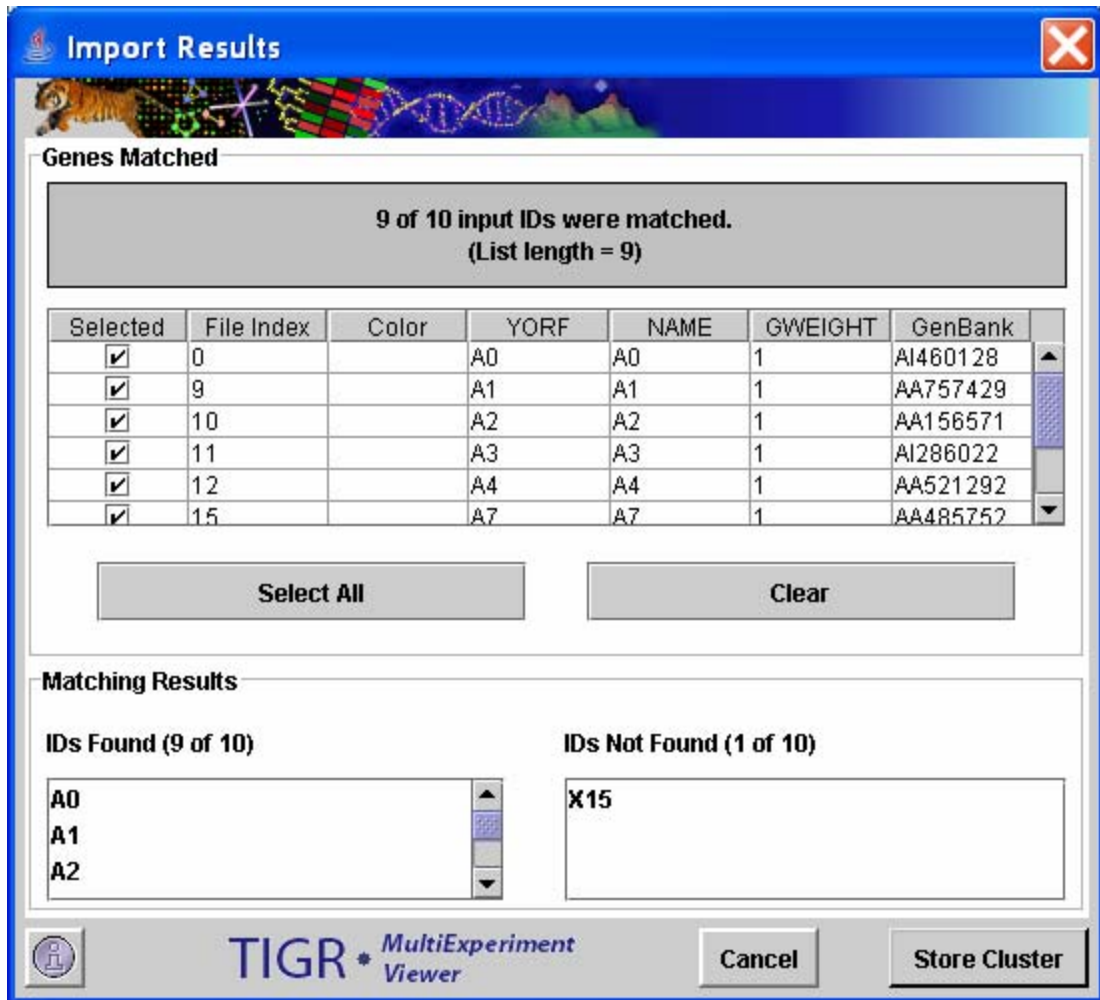
Import Gene or Sample List

The *Import Gene/Sample List* allows one to create a cluster based on supplied identifiers. Identifiers belonging to the cluster are pasted into the text area. The drop down list indicates the type of annotation being loaded. After searching for matches, the List Import Result dialog will be displayed. This intermediate dialog

will report the number of input identifiers that were matched, the number of matches found (note that this might be greater than the number of identifiers in the case where genes are duplicated on the array. This dialog also displays a table that contains matching elements. The rows in the table can be selected to remove unwanted entries before hitting the *Store Cluster* button to store the items to a cluster. The bottom section of the dialog also reports which indices were found and which were not found in the loaded data set.



9.1.3. List Import Dialog



9.1.4. List Import Result Dialog

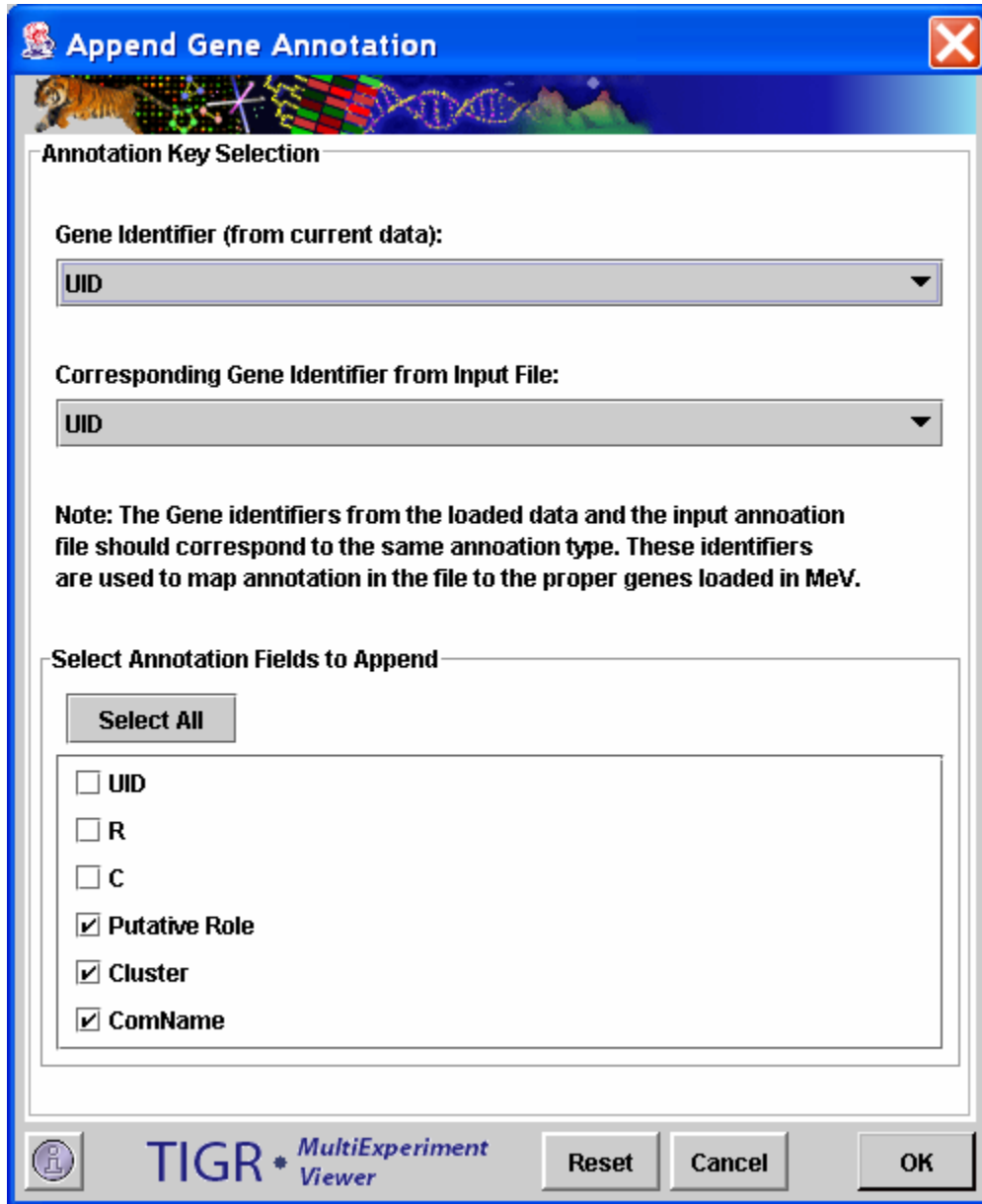
Append Sample Annotation

This feature allows the import of additional sample annotations. Often these would be more descriptive sample names that distinguish the samples based on a study factor, condition, or some measured variable. The sample annotation file should be a tab-delimited text file containing one header row for annotation labels (field names). The file may contain multiple columns of annotation with each column containing a header entry that indicates the nature of the annotation. The annotation for each sample is organized in rows corresponding to the order of the loaded samples. If annotation is missing for a sample the entry in that sample row may be left blank. Please see the manual appendix on file formats for more information and a small example.

Append Gene Annotation

The Append Gene Annotation feature is used to append additional gene annotation from an MeV style annotation (.ann, or .dat) file. The annotation file format is described in detail in the File Format Appendix. The main parameter selection dialog permits the selection of two key fields to be used to map the annotation from the input file to the proper gene already loaded in MeV. One key

should be specified for the currently loaded data. This key should be a gene identifier of some sort and should correspond to the annotation field selected as the file's primary key. The values of these annotation keys is used to map or correlate gene annotation in the file to the loaded genes. The lowest section of the parameter dialog specifies the annotation fields from the file to import. Import status will be reported and an entry that logs the import will be added to the history node's history log.



9.1.5. Gene Annotation Import Dialog

10. Creating Output

While creating output files is not the main purpose of MeV, several options exist for saving files from the Multiple Array Viewer for later use. For details regarding file output with the Single Array Viewer, please see section 13.10.

10.1. Saving the Analysis

Saving the state of an ongoing analysis to file can be accomplished by selecting *Save Analysis As* in the Multiple Array Viewer's File menu. The analysis file will contain the loaded data, the current analysis results, and any clusters that are stored in the cluster manager. If the analysis has been saved previously then the *Save Analysis* option will be enabled. If this option is taken then the analysis will be saved to the analysis file that was opened or the last saved analysis file. To reopen an analysis, select the *Open Analysis* menu option in the File menu. This will restore the data and all algorithm results.

One word of caution on analysis saving, the results are saved in a format that captures the detailed structure of the result. Some of the objects stored in the file are developed outside of TIGR including some basic Java classes. The version of these Java classes is critical to reconstructing the result and errors could occur if an analysis was created using one Java version and then opened using another Java version. For this reason MeV reports Java version mismatches when appropriate. Due to this Java version match requirement, sharing of analysis files between collaborators can lead to problems if each is using a different Java version. Even with this drawback the ability for individuals to store an analysis is beneficial.

Note that for sharing analysis techniques and results with collaborators, one option is to use MeV's scripting ability to create an algorithm execution script so that others analyzing the same data can share techniques and view interesting results produced by specific analysis routines. (See Scripting for details. Chapter 10).

10.2. Saving the Expression Matrix

The expression matrix can be saved as a tab delimited text file by selecting the *Save Matrix* item from the *File* menu. Enter a name for the file in the save file dialog that is displayed. The matrix saved reflects any data adjustments that are currently imposed on the data set such as percentage cutoffs or low intensity cutoffs.

UniqueID	Name	Ex2	Ex3	Ex4	Ex5	Ex6	Ex7	Ex8	Ex9	Ex10
A0	A0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
B0	B0	1.9999992	1.9999992	1.9999992	1.9999992	1.9999992	1.9999992	1.9999992	1.9999992	1.9999992
C0	C0	-2.0000039	-2.0000039	-2.0000039	-2.0000039	-2.0000039	-2.0000039	-2.0000039	-2.0000039	-2.0000039
D0	D0	2.4999998	2.9999998	2.9999998	2.4999998	1.9999992	1.9999992	1.4999998	0.99999934	1.4999998
E0	E0	1.4999998	0.99999934	1.4999998	1.9999992	2.4999998	2.9999998	2.9999998	2.4999998	2.4999998
F0	F0	-2.5000008	-3.000142	-2.5000008	-2.0000039	-2.0000039	-1.5000007	-1.0000256	-1.0000256	-1.5000007
G0	G0	-1.5000007	-1.0000256	-1.5000007	-2.0000039	-2.0000039	-2.5000008	-3.000142	-2.5000008	-2.5000008
H0	H0	-2.0000039	-2.0000039	-1.0000256	0.0	0.99999934	1.9999992	1.9999992	1.9999992	1.9
J0	J0	1.9999992	1.9999992	0.99999934	0.0	-1.0000256	-2.0000039	-2.0000039	-2.0000039	-2.0
A1	A1	0.22613108	-0.27992555	0.27894148	0.41660944	0.40449798	0.16748707	0.37853232	0.37853232	0.37853232
A2	A2	0.0076903533	0.35410163	-0.324028	-0.40811864	-0.22405647	0.15208569	0.13762924	0.13762924	0.13762924
A3	A3	-0.086178996	-0.32919875	-0.38246173	-0.297692	0.09950139	0.40959656	-0.21474145	-0.21474145	-0.21474145
A4	A4	0.25306734	0.3697412	-0.1458932	0.0320121	0.28635117	0.28635117	-0.30051067	-0.30051067	-0.30051067
A5	A5	0.34172982	0.37199104	-0.11216137	-0.48228675	0.46387237	0.16378413	0.16980185	0.16980185	0.16980185
A6	A6	-0.1959054	0.42616928	0.27522734	-0.078372434	-0.37152755	-0.12526849	-0.16877374	-0.16877374	-0.16877374
A7	A7	0.047770597	-0.36558595	-0.1554148	0.40188536	0.16643775	-0.36906567	-0.33950883	-0.33950883	-0.33950883
A8	A8	-0.10032875	-0.13755952	-0.10321837	0.28750706	0.15527576	-0.13934253	-0.04870717	-0.04870717	-0.04870717
A9	A9	0.4295071	-0.23373696	0.38981664	0.38981664	0.20473514	0.18789765	-0.420528	-0.420528	-0.420528
A10	A10	-0.0060100724	0.479824	0.095255636	0.33857003	0.38170505	-0.23830795	0.46950975	0.46950975	0.46950975
A11	A11	0.066994876	0.027313566	-0.097083755	0.46378434	-0.18517362	-0.44181275	-0.1755565	-0.1755565	-0.1755565
A12	A12	0.2297775	-0.18720949	0.05287704	0.11959473	-0.48246494	-0.2592242	0.13789062	0.13789062	0.13789062
A13	A13	0.05785373	0.2539055	-0.4502324	-0.3468802	0.0924789	0.4314772	-0.0851767	-0.0851767	-0.0851767
A14	A14	-0.17265245	0.040681183	-0.43022922	0.2302701	0.0044699945	-0.11539963	0.28847426	0.28847426	0.28847426
A15	A15	-0.384472	0.38051647	-0.015448719	0.16840056	1.0999395E-4	0.002087819	0.40919814	0.40919814	0.40919814
A16	A16	-0.2920516	-0.34303975	0.4230365	-0.2422754	-0.18991432	0.015518956	-0.31080058	-0.31080058	-0.31080058
A17	A17	-0.23262587	-0.0951242	0.20357737	-0.1628484	-0.36485115	-0.29194447	0.2217666	0.2217666	0.2217666

10.2.1. Expression matrix saved as text file.

10.3. Saving Viewer Images

To save a tiff file of the currently displayed image in the main view, select *Save Image* from the *File* menu. To print the image, select *Print Image* instead.

10.4. Saving Cluster Data

You can also *Save Cluster Data* to a tab-delimited text file. Selecting this option from the right-click menu will cause a file chooser to appear. Select a file name and a place to save row/column data, log ratio expression values, and (optionally) Cy3 and Cy5 values for each gene in the cluster (or each sample if the cluster is produced via clustering samples). Selecting *Save All Clusters* will allow you to save the elements in all clusters in a similar way.

11. Modules

Description Conventions and General Pointers

Each of these modules can be launched from the *Analysis* menu or from a button on the Multiple Array Viewer toolbar. Common to all clustering modules the selected algorithm can be performed to cluster genes or samples. In the title of the analysis module descriptions which follow, the acronym used to label the module's button is followed by the name of that module. Below this is a line containing the reference(s) used in implementing the algorithm.

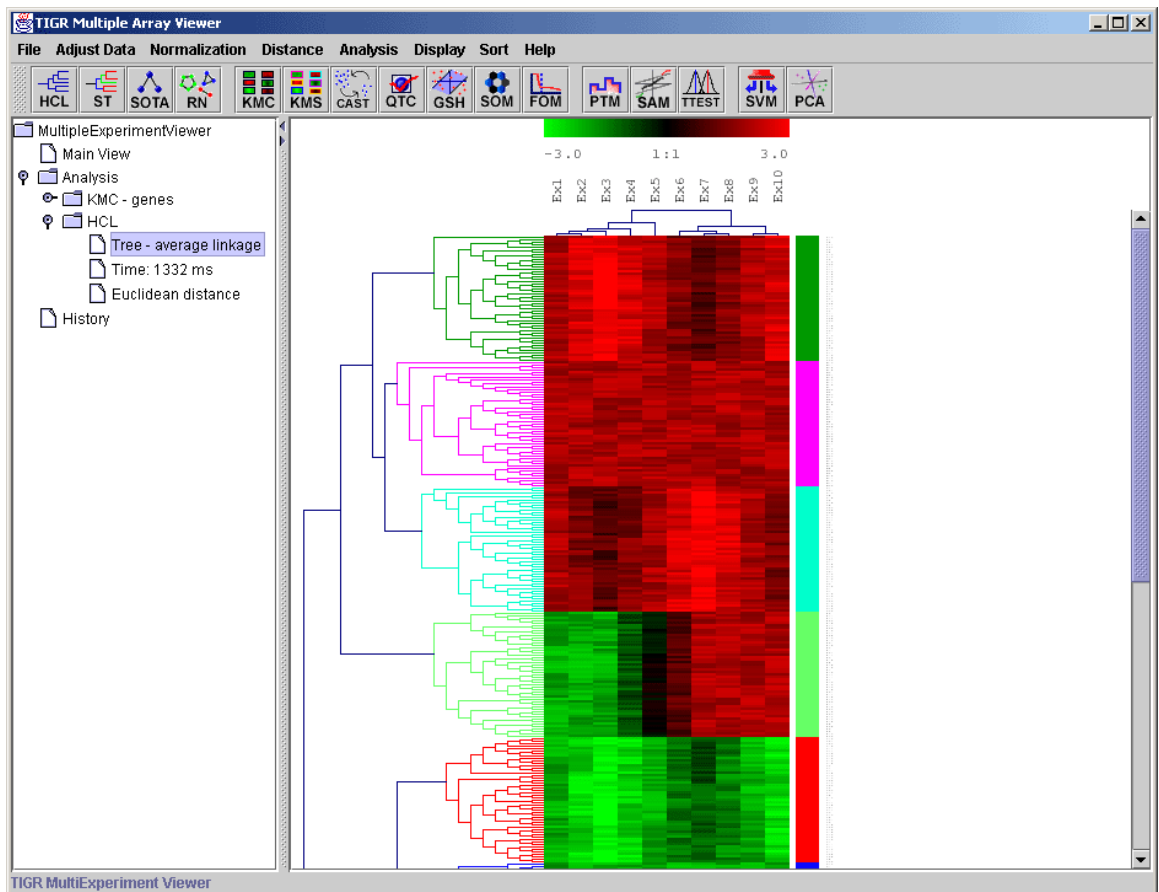
Each module, when run, will create a subtree labeled with its acronym and a label indicating whether the result was created by clustering genes or samples. This subtree will be placed under the *Analysis* tab in the result navigation tree. The tabs within this subtree contain the results of the module's calculations. These tabs vary greatly depending on the module which creates them, but the *General Information* tab always contains a summary of the parameters used in the analysis.

Each algorithm run will present a dialog or form to use to input parameters specific to the algorithm being performed. An information button on each dialog (lower left corner) can be used to retrieve a reference page describing the required parameters.

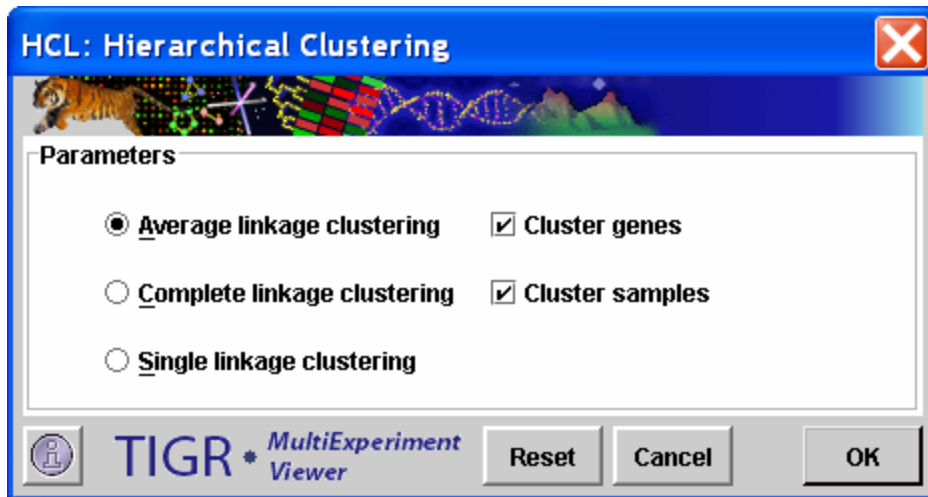
11.1. HCL: Hierarchical clustering

(Eisen *et al.* 1998)

Selecting this analysis will display a dialog that allows three different linkages and options to cluster genes, samples or both. Once the computations are complete, select the *Tree* node under *Analysis* to view the hierarchical tree. The display is similar to the main display, but similar genes and experiments are connected by a series of ‘branches’. Labels are displayed on the right side. Clicking under a branch intersection (node) will select that node and the subtree below that node. Once selected, right clicking in the same area will display a popup menu that allows the user to set the highlighted area as a cluster, name the cluster, save the cluster and set several tree options. Clusters set and named in this display can propagate to other displays. Saving a cluster will display a dialog where a tab delimited text file containing the data for the highlighted cluster can be named. The algorithm also produces a *Node Height Graph* which displays the number of terminal nodes in the tree given a particular inter-node distance threshold.



11.1.1. Hierarchical tree with clusters selected.



11.1.2. HCL Initialization Dialog

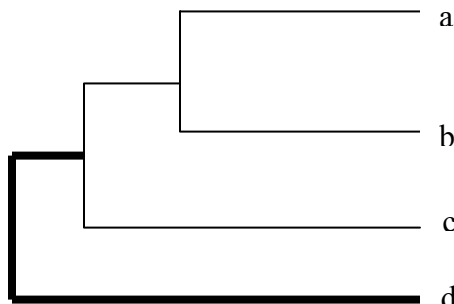
Parameters:

Linkage Method

This parameter is used to indicate the convention used for determining cluster-to-cluster distances when constructing the hierarchical tree.

Single Linkage: The distances are measured between each member of one cluster each member of the other cluster. The minimum of these distances is considered the cluster-to-cluster distance.

Average Linkage: The average distance of each member of one cluster to each member of the other cluster is used as a measure of cluster-to-cluster distance. Note that this option in MeV actually is determined by a weighted average of distances of cluster members. Example: Consider the distance from node 'd' to cluster (a,b,c)...



Unweighted Average Linkage:

$$d_{d,(a,b,c)} = \frac{(d_{d,a} + d_{d,b} + d_{d,c})}{3} = \frac{d_{d,a}}{3} + \frac{d_{d,b}}{3} + \frac{d_{d,c}}{3}$$

Weighted Average Linkage:

$$d_{d,(a,b,c)} = \frac{[(d_{d,a} + d_{d,b})/2 + d_{d,c}]}{2} = \frac{d_{d,a}}{4} + \frac{d_{d,b}}{4} + \frac{d_{d,c}}{2}$$

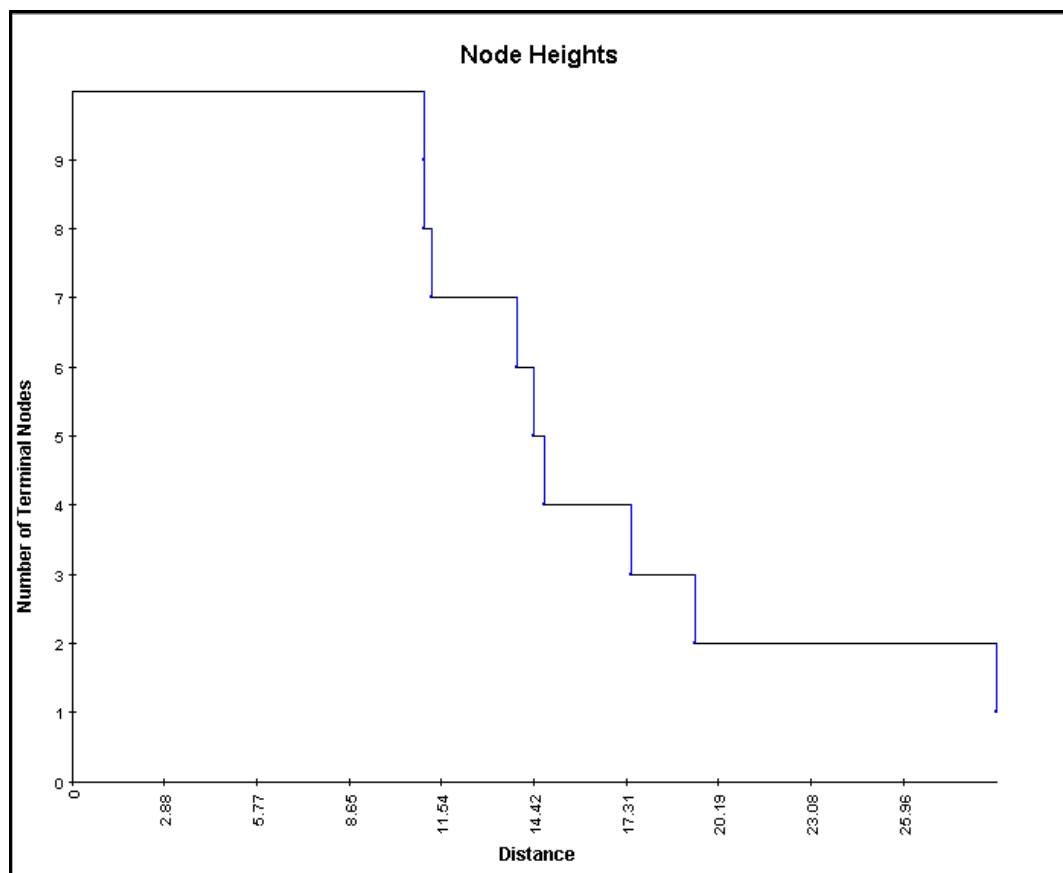
Nodes on are weighted unequally where nodes deeper in the sub-tree contribute less to the overall computed distance.

Complete Linkage: The distances are measured between each member of one cluster each member of the other cluster. The maximum of these distances is considered the cluster-to-cluster distance.

Cluster Genes / Cluster Samples Options

These checkboxes are used to indicate whether to cluster genes, samples, or both.

Default Distance Metric: Euclidean

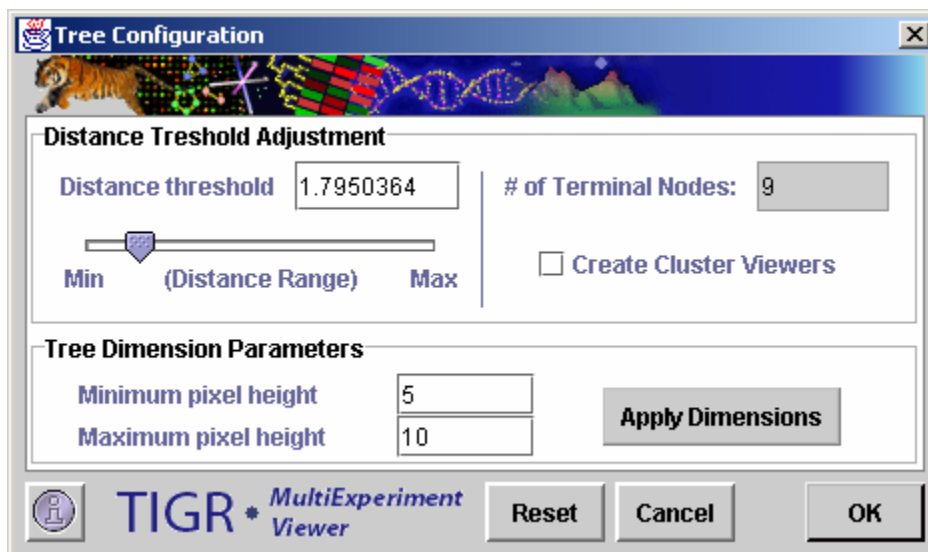


11.1.3. Node Height Graph

Adjusting the Tree Configuration and Viewing Clusters

A right click in the Tree Viewer will produce a menu which includes an option to

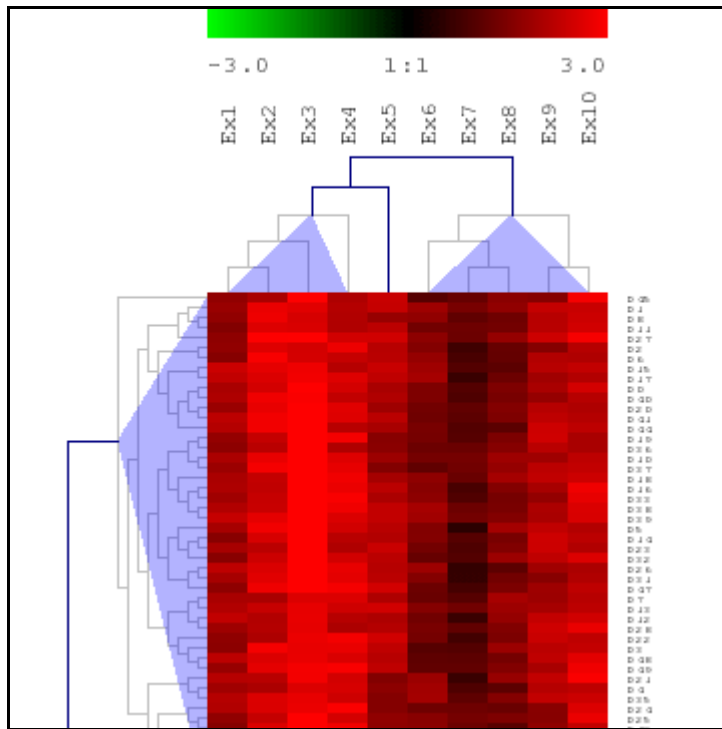
alter the displayed tree, *Tree Properties*. This option permits the user to change the tree's appearance and to reduce the complexity of the tree by imposing a distance threshold. Elements on nodes which have distances below this threshold can be considered as one entity (or cluster). Consequently the lower level detail of the tree is ignored. As the value is adjusted the corresponding HCL tree will have nodes below this threshold appear as light gray in color and a translucent 'wedge' from that node to all enclosed elements will be drawn on the tree. This representation of the tree will persist unless the dialog is dismissed by hitting cancel. The distance threshold can be entered into a text field or can be adjusted with a slider over the maximum range of inter-node distances. The number of terminal nodes (clusters) using the current distance threshold is displayed in the upper right quadrant of the dialog.



11.1.4. HCL Tree Configuration Dialog

The *Create Cluster Viewers* option allows you to create viewers based on the distance threshold. This option collects groups of elements falling below terminal nodes in the tree using the current distance threshold. The clusters of elements are represented as nodes in the result navigation tree under the HCL result node. The results are added once the HCL Tree Configuration dialog has been dismissed.

The minimum and maximum pixel distance imposes limits on the minimum and maximum displayed inter-node distance. This alters the appearance of the tree. The *Apply Dimensions* button causes the entered tree dimensions to be applied to the HCL tree. This allows one to fine tune the tree's appearance without dismissing the dialog.

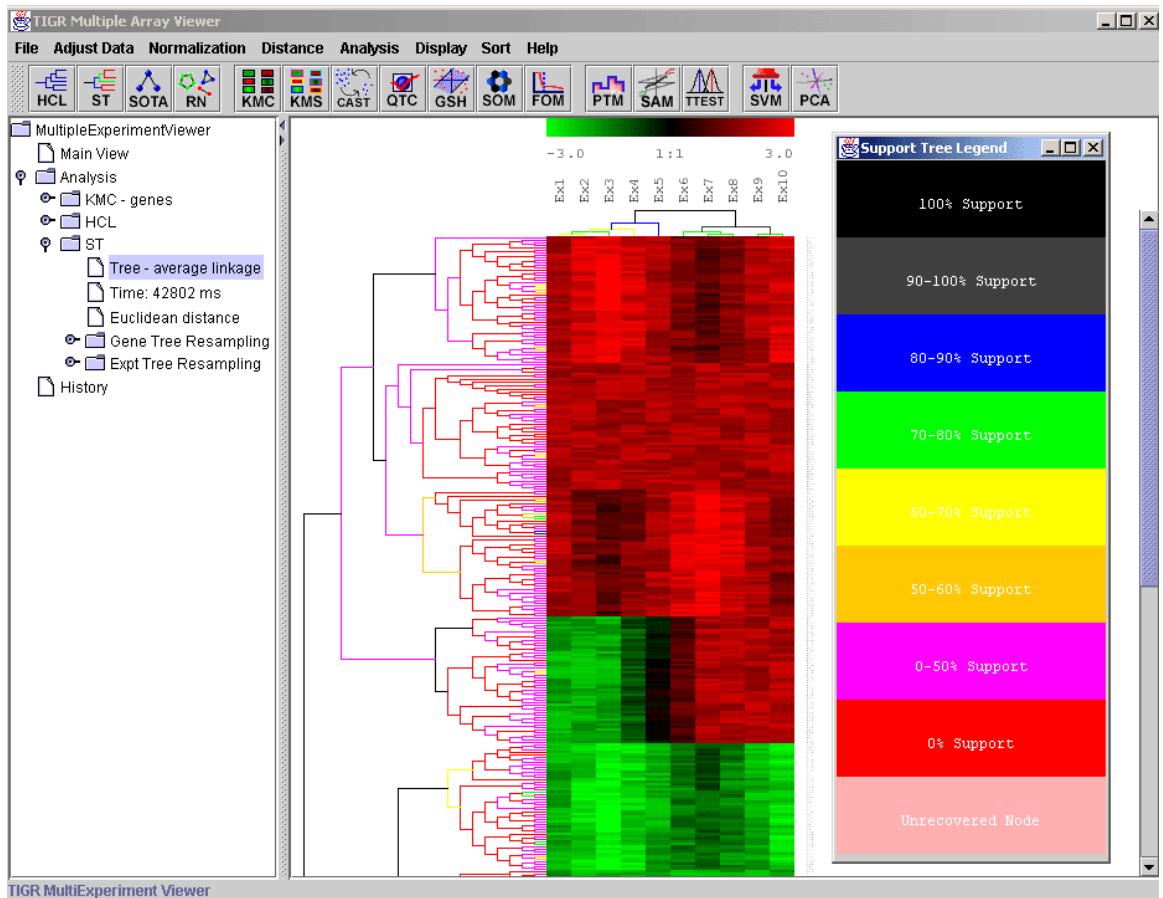


11.1.5. HCL Tree with distance thresholds applied.

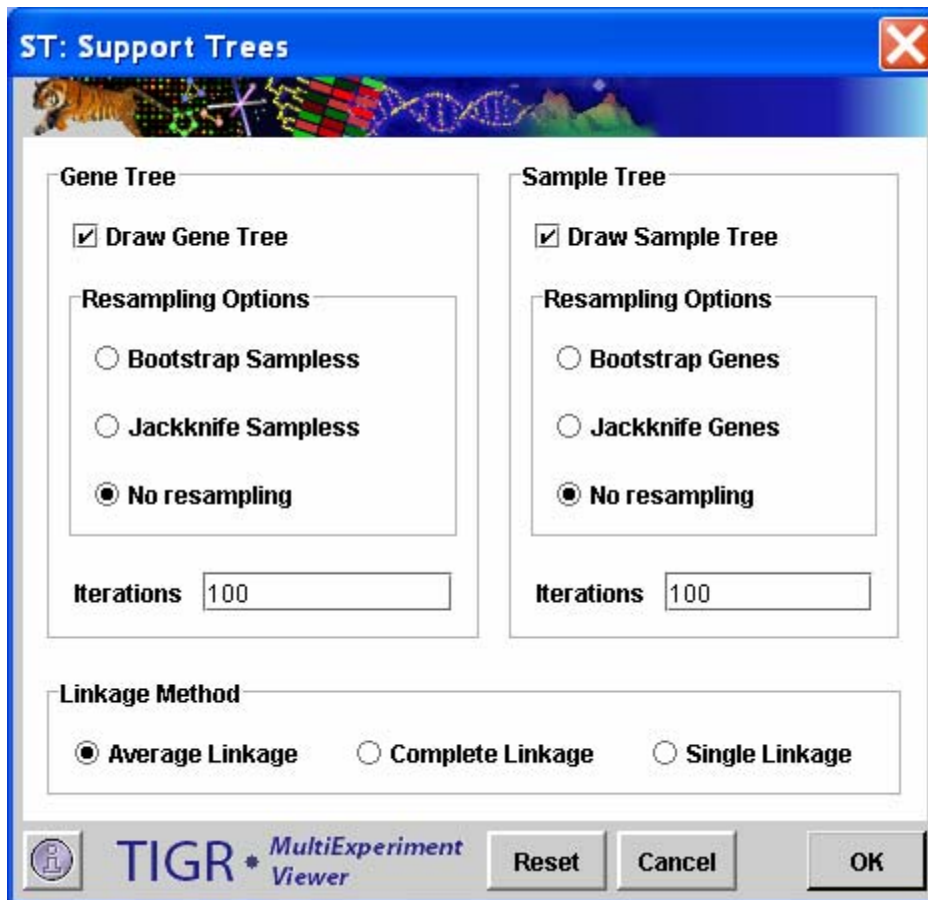
11.2. ST: Support Trees

This option shows the hierarchical trees obtained using the previous module, but it also shows the statistical support for the nodes of the trees, based on resampling the data. The user can select two resampling methods: bootstrapping (resampling with replacement), and jackknifing (resampling by leaving out one observation in this implementation). Resampling can be conducted on genes and / or experiments for a user-specified number of iterations. The branches of the resulting tree are color-coded to denote the percentage of times a given node was supported over the resampling trials. The legend for the color code corresponding to a given level of support can be found under the Help menu.

The two most useful options for support trees are likely to be bootstrapping genes to build experiment trees, and bootstrapping experiments to build gene trees.



11.2.1. Support trees for Hierarchical Clustering with Support Tree Legend displayed on right.



11.2.2. Support Trees Initialization Dialog

Parameters:

The Support Tree algorithm permits the resampling options to be set separately for the gene tree and the sample tree.

Tree Construction Options

The Draw Gene Tree and Draw Sample Tree options allow you to select to construct a gene tree, an experiment tree, or both.

Resampling Options

You can elect to resample either genes or samples or neither using either a bootstrapping or a jackknifing method.

Bootstrapping

The matrix is reconstructed such that each expression vector has the original number of values but the values are a random selection (with replacement) of the original values. Values in the original expression vector may occur more than once since the selection uses replacement.

Jackknifing

Jackknifing takes each expression vector and randomly selects to omit an element. This method produces expression vectors that have one fewer element and this is often done to minimize the effect of single outlier values.

Iterations

This indicates how many times the expression matrix should be reconstructed and clustered.

Linkage Method

This parameter is used to indicate the convention used for determining cluster-to-cluster distances when constructing the hierarchical tree.

Single Linkage: The distances are measured between each member of one cluster each member of the other cluster. The minimum of these distances is considered the cluster-to-cluster distance.

Average Linkage: The average distance of each member of one cluster to each member of the other cluster is used as a measure of cluster-to-cluster distance.

Complete Linkage: The distances are measured between each member of one cluster each member of the other cluster. The maximum of these distances is considered the cluster-to-cluster distance.

Support Tree Legend

A legend to relate support tree output colors to % support is displayed by selecting the support tree legend menu item in the main help menu.

Default Distance Metric: Euclidean

11.3. SOTA: Self Organizing Tree Algorithm

(Dopazo *et al.* 1997, Herrero *et al.* 2001)

The initialization form shown below (11.3.1) is divided into four main areas. The SOTA algorithm constructs a binary tree (dendrogram) in which the terminal nodes are the resulting clusters.

SOTA: Self Organizing Tree Algorithm

Sample Selection

Cluster Genes Cluster Samples

Growth Termination Criteria

Max. Cycles: 10 Max. Cell Diversity: 0.01
Max. epochs/cycle: 1000 Min. Epoch Error Improvement: 0.0001

Run Maximum Number of Cycles (unrestricted growth)

Centroid Migration and Neighborhood Parameters

Winning Cell Migration Weight: 0.01
Parent Cell Migration Weight: 0.005 Neighborhood Level: 5
Sister Cell Migration Weight: 0.001

Cell Division Criteria

Use Cell Diversity (mean dist(gene,centroid))
 Use Cell Variability (max(dist(g(i), g(j)))) p Value: 0.05

Hierarchical Clustering

Construct Hierarchical Trees

TIGR * MultiExperiment Viewer [Reset] [Cancel] [OK]

11.3.1. SOTA Initialization Dialog Box.

Parameters and Basic Terminology:

SOTA Terminology and Concepts

Topology

The topology of the resulting tree is a binary tree structure where each terminal node represents a cluster.

Node

A structure which contains a Centroid Vector and a number of associated expression profiles (members).

Cell
A Node which is the terminal Node in a branch of the tree (a.k.a. leaf node). The members of the cell are considered members of an expression cluster.

Centroid Vector
A vector that is representative of the membership of a node.

Members
Expression Elements associated with a Node.

Node
A structure which contains a Centroid Vector and a number of associated expression profiles (members).

Cell
A Node which is the terminal Node in a branch of the tree (a.k.a. leaf node). The members of the cell are considered members of an expression cluster.

Growth Termination Criteria Parameters

Max Cycles
This integer value represents the maximum iterations allowed. The resulting number of clusters produced by SOTA is (Max Cycles +1) unless other criteria are satisfied prior the indicated maximum number of cycles.

Max epochs/cycle
This integer value indicates the maximum number of training epochs allowed per cycle.

Max. Cell Diversity
This value represents a maximum variability allowed within a cluster. All resulting clusters will fall below this level of 'diversity' (mean gene to cluster centroid distance) if diversity is used as the cell division criteria. (Unless Max cycles are reached at which time some clusters may still exceed this parameter)

Min Epoch Error Improvement
This value is used as a threshold for signaling the start of a new cycle and a cell division. The tree diversity is monitored during a training epoch and when the diversity fails to improve by more than this value then training has been considered to have stabilized and a new cycle begins.

Run Maximum Number of Cycles (unrestricted growth)
The algorithm will run until Max Cycles or until all of the input set are fully partitioned such that each cluster has one gene or several identical gene vectors.

Centroid Migration and Neighborhood Parameters

Migration Weights
These values are used to scale the movement of cluster centroids (characteristic gene expression patterns) toward a gene vector which has been associated with a neighborhood. When a gene is associated with a cluster the centroid adapts to become more like the newly associated gene vector. The parent and sister cell

migration weights should be smaller than the weight for the winning cell (Cell to which the gene vector is associated.).

Neighborhood Level

This value determines which cells are candidates to accept new expression elements. When elements are considered for redistribution to new node during a cell division candidate cells are determined by moving up the tree toward the root this number of levels. From that node, all cells (terminal nodes) within this subtree are targets for possibly accepting expression vectors. (Each vector moves into the cell to which it is most similar).

Cell Division Criteria Parameters

Use Cell Diversity

Cell diversity is the mean distance between the cell's members (expression profiles) to the cell's centroid vector. When considering which cell to divide, the cell with the greatest diversity is split. (providing it's diversity exceeds Max Cell Diversity (see above))

Use Cell Variability

Cell variability is the maximum element-to-element distance within a cell. The cell having the largest internal gene-to-gene distance is selected as the next cell to divide. In this case the stopping criteria is changed so that growth continues until the most variable cell falls below a variability criteria generated using the provided pValue (see below)

pValue

This value is used when using variability as the cell division criteria. A distribution of all element to element distances is generated by resampling the data set with each expression vector having randomized ordering of vector elements. The resulting distribution represents random gene to gene distances. The pValue supplied is applied to this resampled distribution to generate a variability cutoff. Clusters falling below this variability cutoff have a probability of having members that are paired by chance at or below the supplied pValue.

Hierarchical Clustering

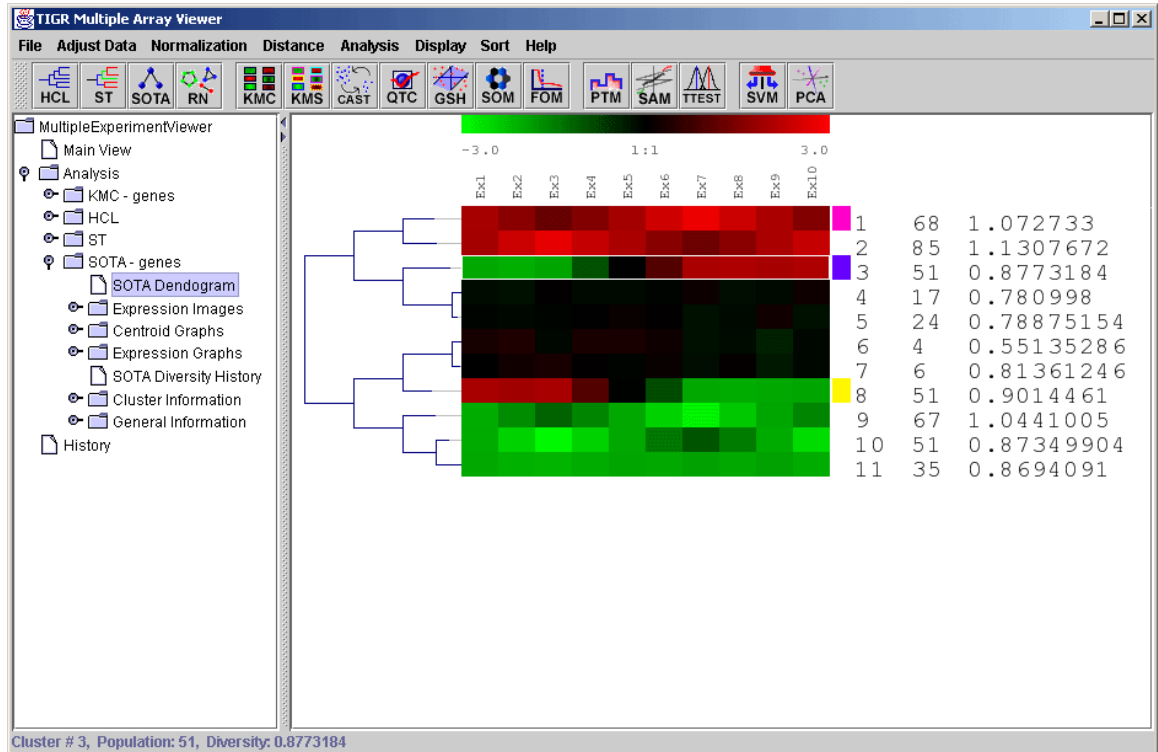
This check box selects whether to perform hierarchical clustering on the elements in each cluster created.

Default Distance Metric: Euclidean

The result views created by the SOTA algorithm include the basic viewers with the addition of two SOTA specific viewers and enhancements to expression image viewers to include more cluster information.

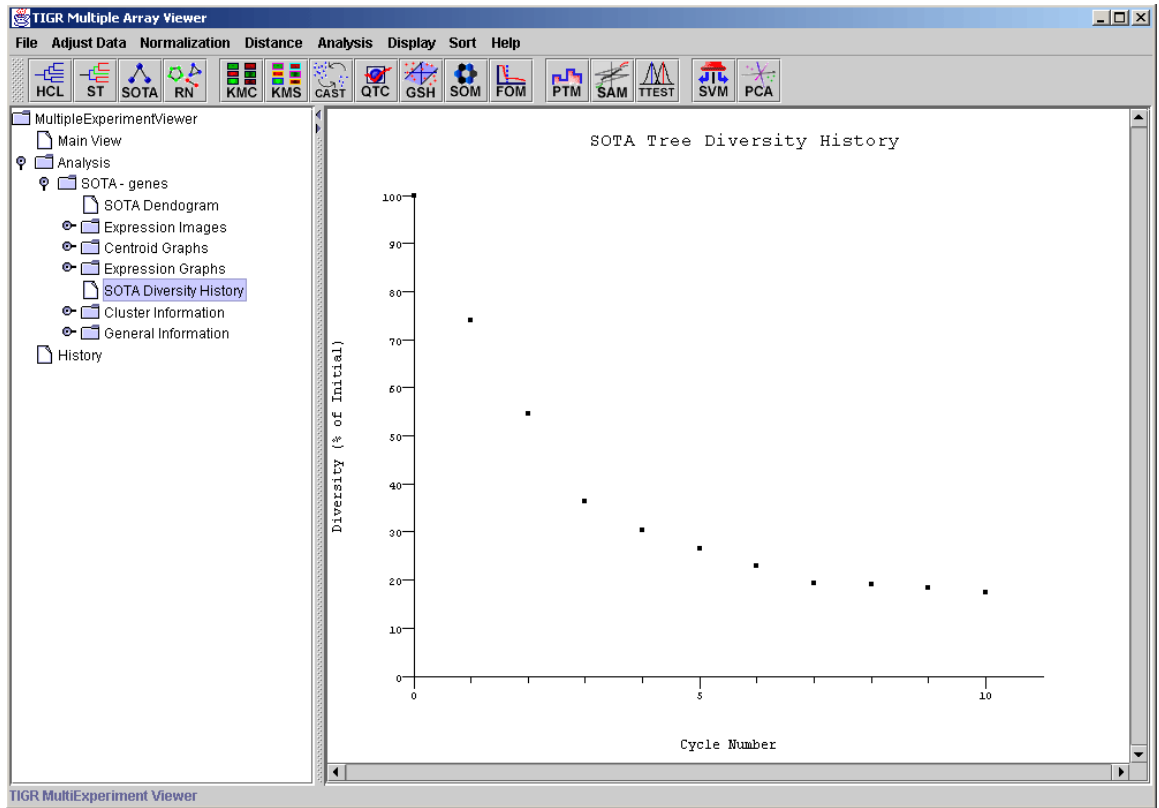
One of the SOTA specific viewers is the SOTA dendrogram (below) which displays the generated tree with the expression image of each resulting cluster's centroid gene. The text to the right of the centroid expression image includes a

cluster id number, the cluster population (number of genes in the cluster), and the cluster diversity (mean gene to centroid distance). Clusters can be colored and saved from this viewer and a left click over a cluster centroid jumps to the expression image for that cluster.



11.3.2. SOTA dendrogram

The SOTA diversity viewer shows the change in the summation of gene to associated centroid distance vectors for all genes in the tree. This is a measure of overall tree diversity. This can reveal how much diversity improvement is achieved with each cycle (new cluster addition).

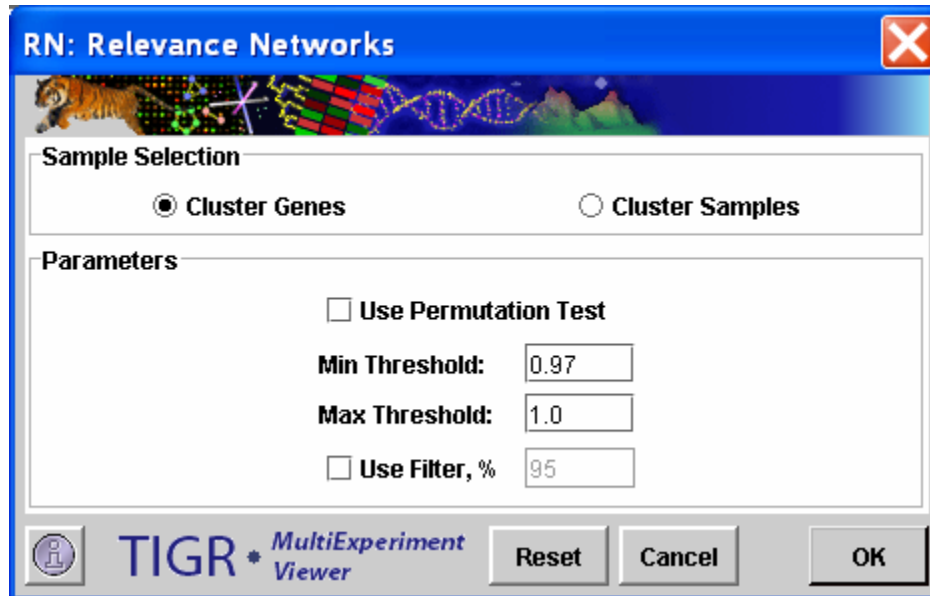


11.3.3. SOTA diversity viewer

11.4. RN: Relevance Networks

(Butte *et al.* 2000)

A relevance network is a group of genes whose expression profiles are highly predictive of one another. Each pair of genes related by a correlation coefficient larger than a minimum threshold and smaller than a maximum threshold (assigned in the initialization dialog box) is connected by a line. Groups of genes connected to one another are referred to as networks.



11.4.1. RN Initialization Dialog

Parameters:

Sample Selection

The sample selection option indicates whether to cluster genes or samples.

Use Permutation Test

This check box is used to indicate that the minimum threshold R^2 value should be selected based on a distribution constructed from element to element R^2 values derived following permutation of the expression vectors.

Min Threshold

This value ranging from 0 to 1.0 indicates the smallest R^2 possible between two elements to permit a link between the elements in a subnet.

Max Threshold

This value ranging from 0 to 1.0 indicates the greatest R^2 possible between two elements to permit a link between the elements in a subnet.

Use Filter

This option allows the user to filter out elements with little dynamic change thus removing flat or uninteresting elements. A measure of entropy is used to rank the elements. The percentage value entered (1 to 100) indicates what percentage of the elements to retain for the construction of the network. A value of 25 will retain the 25% of elements having the greatest entropy.

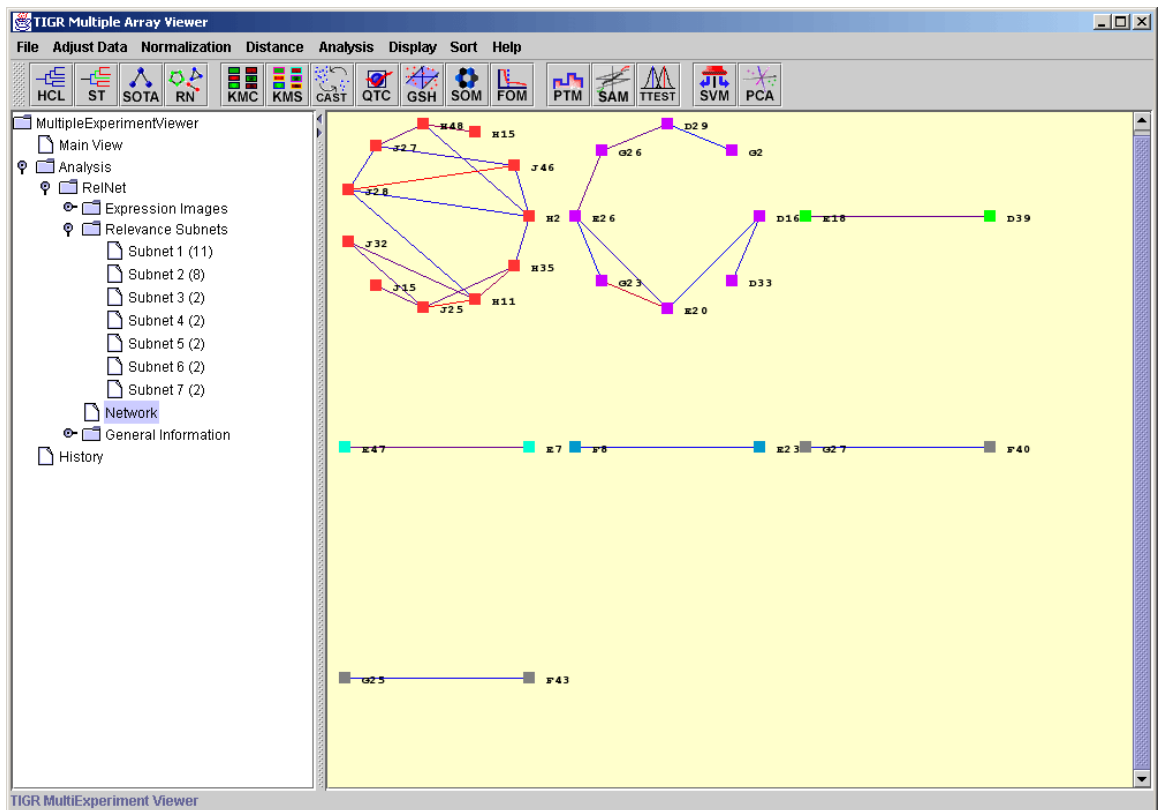
Distance Metric: Pearson squared

Other Acceptable Metrics: None. Changing the metric in the *Distance* menu

will not affect the calculations done by this module.

The module calculates the correlation coefficient between genes by comparing the expression pattern of each gene to that of every other gene. The ability of each gene to predict the expression of each other gene is measured as a correlation coefficient. Genes are represented as nodes in a network and edges are drawn between them if their correlation coefficient falls between the minimum and maximum thresholds specified in the initialization dialog.

The experiment subtree created by this module contains information regarding the networks predicted. Under the *Network* tab is a graph of all of the subnets generated (fig. 10.5.1). A subnet is a group of genes in which each gene is connected to at least one other gene. The *Relevance Subnets* tab contains network diagrams for each of the individual subnets, and the *Expression Images* folder contains expression views for the genes in each of them.



11.4.2. Network View.

Several options can be launched from the network viewer to enhance the view or to better characterize the results by using a right click context menu. Note that links colored in red represent elements that are positively correlated while links colored in blue represent elements that are negatively correlated. Options from the menu to enhance the view include the ability to zoom in and out on the subnets, alter the color of the background, alter the element shape, and alter the thickness of the links for better visibility.

Other options reveal the nature of the subnets. You can select to alter the threshold to make it more stringent. Using this option under the *links-threshold*

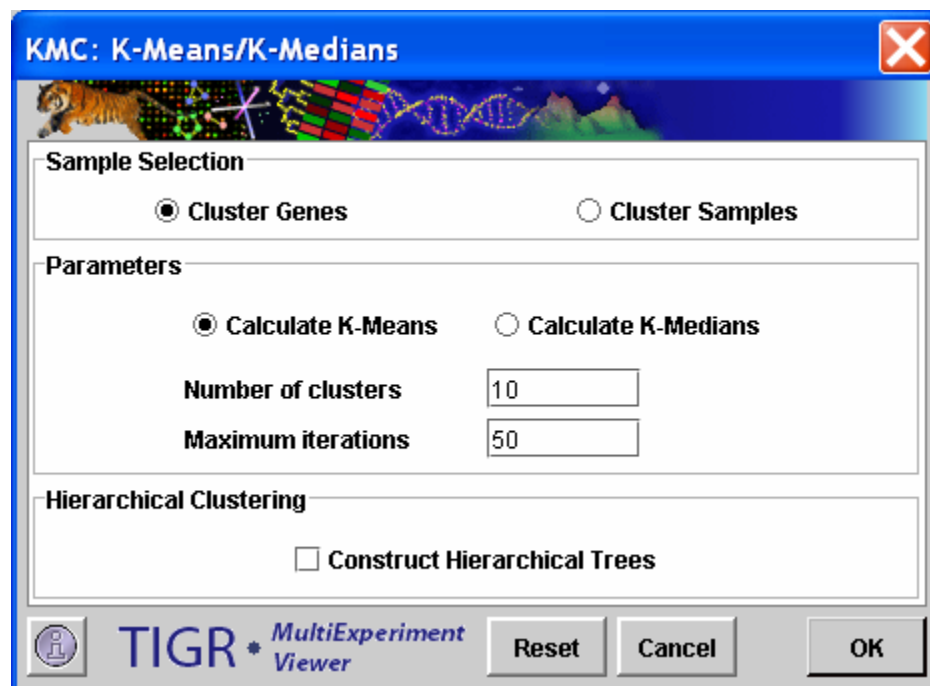
the viewer is instructed to only show links for elements correlated with R^2 greater than the new threshold. Selection of elements in the network can be done by using the *select* option in the right click menu. Two options are offered for selecting elements, one by providing a gene identifier label (*Element ID* option), and the second option by specifying all elements with a specified minimum number of links (*Feature Degree* option). As in other viewers selected elements may be assigned a color (*Set Selection* option).

11.5. KMC: K-Means/K-Medians Clustering

(Soukas *et al.* 2000)

Selecting this analysis will display a dialog that allows the user to specify whether to use means or medians, as well as the number of clusters and iterations to run. Once the computations are complete, select the KMC node under Analysis to view the results. There are several sub-nodes beneath KMC, further divided by the clusters created based on the KMC input parameters. Hierarchical trees shows trees constructed for each cluster, if the option to draw hierarchical trees for clusters is selected. Expression images are similar to the main display. Cluster Information is a summary of each cluster based on size and % composition. Centroid graphs show the centroids for each cluster and experiment, individually or all at once. Expression graphs are similar to centroid graphs, but with each gene's expression levels displayed alongside the centroids. Right clicking within an expression image displays a popup menu that allows the user to propagate the cluster to other displays (Set Public Cluster), save and delete the cluster.

This method of clustering is useful when the user has an a priori hypothesis about the number of clusters that the genes should subdivide into.



11.5.1. KMC Initialization Dialog

Parameters:

Sample Selection

The sample selection option indicates whether to cluster genes or samples.

Means/Medians option

The Means or Medians option indicates whether each cluster's centroid vector should be calculated a mean or a median of the member expression patterns.

Number of Clusters

This positive integer value indicates the number of clusters to be created. Note that FOM can be used to estimate an appropriate value.

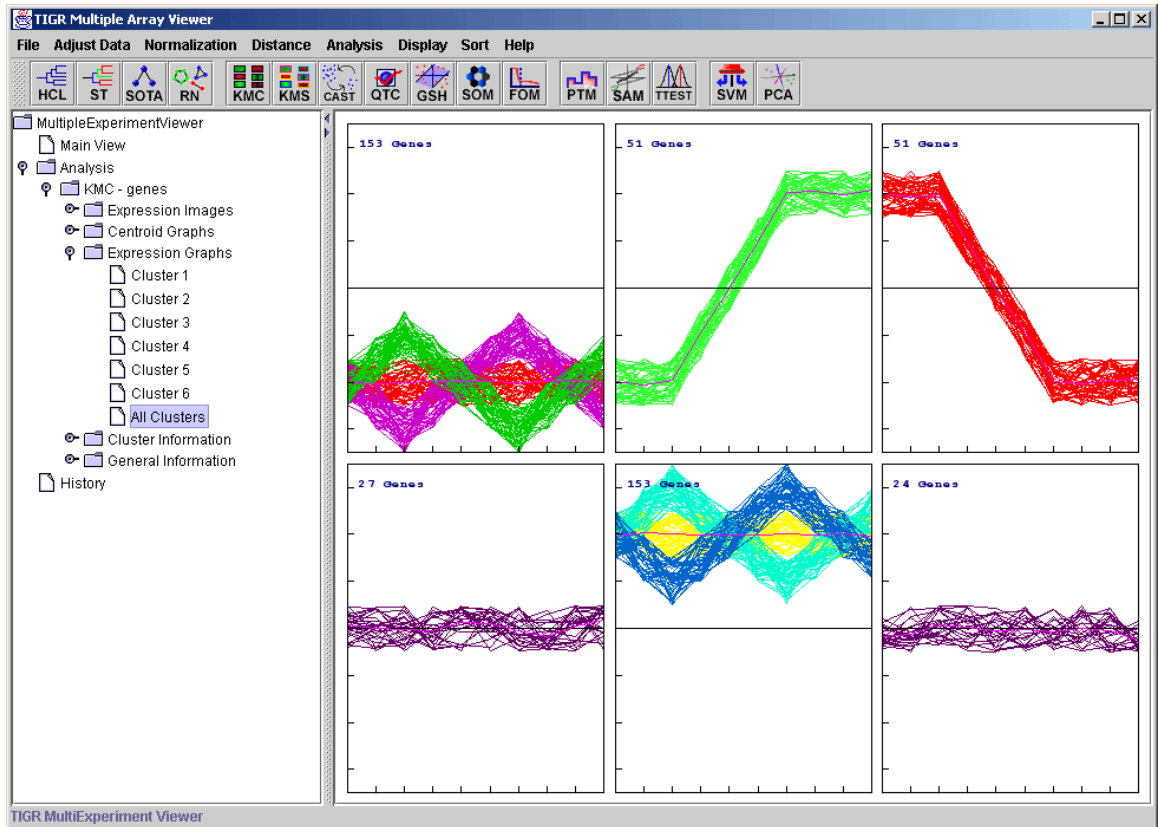
Number of Iterations

This positive integer value is the maximum number of times that all the elements in the data set will be tested for cluster fit. On each iteration each element is associated with the cluster with the closest mean (or median). Note that the algorithm will terminate when either no elements require migration (reassignment) to new clusters or when the maximum number of iterations has been reached.

Hierarchical Clustering

This check box selects whether to perform hierarchical clustering on the elements in each cluster created.

Default Distance Metric: Euclidean



11.5.2. K-Means / K-Medians Clustering: Expression Graphs.

11.6. KMS: K-Means / K-Medians Support

This module allows the user to run the K-Means or K-Medians algorithms multiple times using the same parameters in each run. Owing to the random initialization of K-Means and K-Medians, the clusters produced may vary substantially between runs, depending on the data set and the input parameters. The KMS module allows the user to generate clusters of genes that frequently group together in the same clusters (“consensus clusters”) across multiple runs. The output consists of consensus clusters in which all the member genes clustered together in at least $x\%$ of the K-Means / Medians runs, where x is the threshold percentage input by the user (see screenshot below).

Parameters:

Sample Selection

The sample selection option indicates whether to cluster genes or samples.

Means/Medians option

The Means or Medians option indicates whether each cluster's centroid vector should be calculated a mean or a median of the member expression patterns.

Number of k-means/k-medians runs

This integer value indicates how many times KMC should be run.

Threshold % of occurrence in same cluster

This parameter indicates the minimum percentage of times that two elements should cluster together in order consider the two elements in a cluster. For instance, if 10 KMC runs were run, and the percentage was 80% then a pair of expression elements found together at least 8 times would be considered to pass a criteria to be included in a cluster.

Number of Clusters (K)

This positive integer value indicates the number of clusters to be created during each KMC run. Note that for K-Means support the final number may turn out to be slightly smaller or larger than this entered value depending on the nature of the input data and the appropriate selection of K (number of clusters to create). Note that FOM can be used to estimate an appropriate value for K.

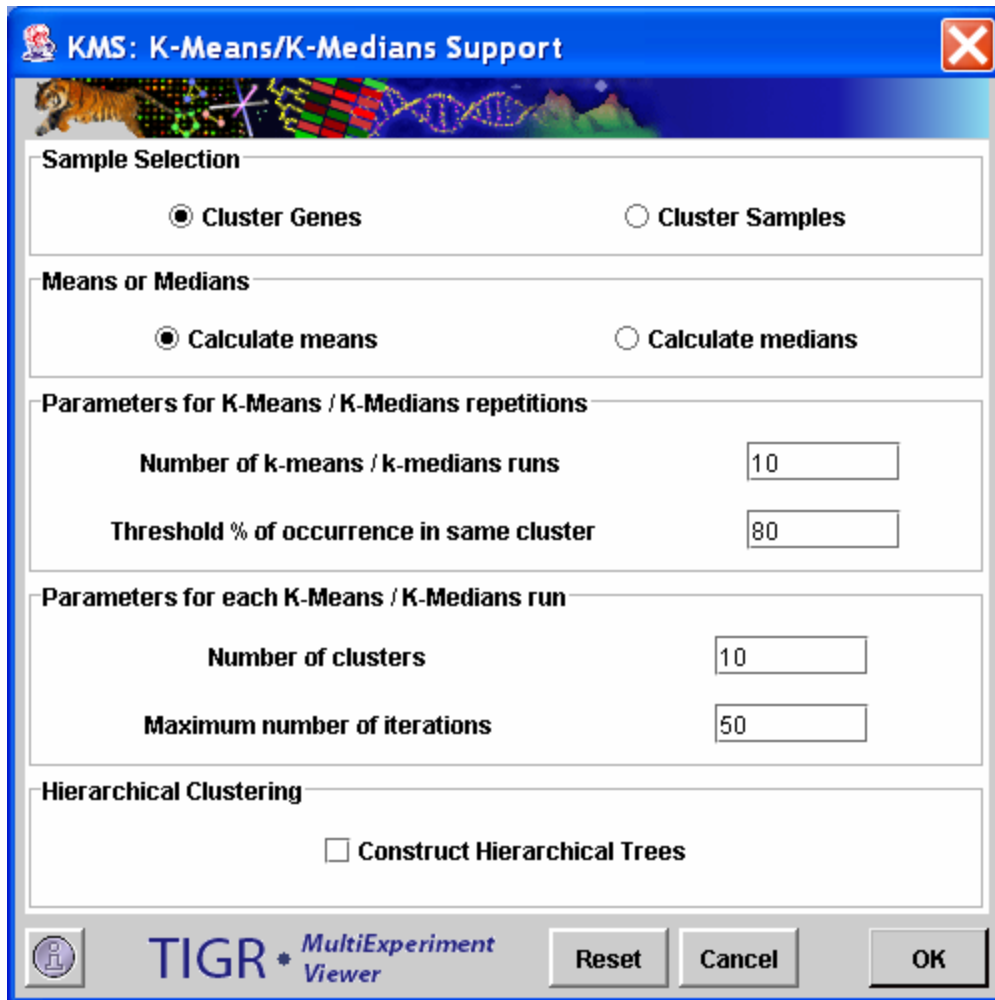
Number of Iterations

This positive integer value is the maximum number of times that all the elements in the data set will be tested for cluster fit. On each iteration each element is associated with the cluster with the closest mean (or median). Note that a KMC run will terminate when either no elements require migration (reassignment) to new clusters or when the maximum number of iterations has been reached.

Hierarchical Clustering

This check box selects whether to perform hierarchical clustering on the elements in each cluster created.

Default Distance Metric: Euclidean



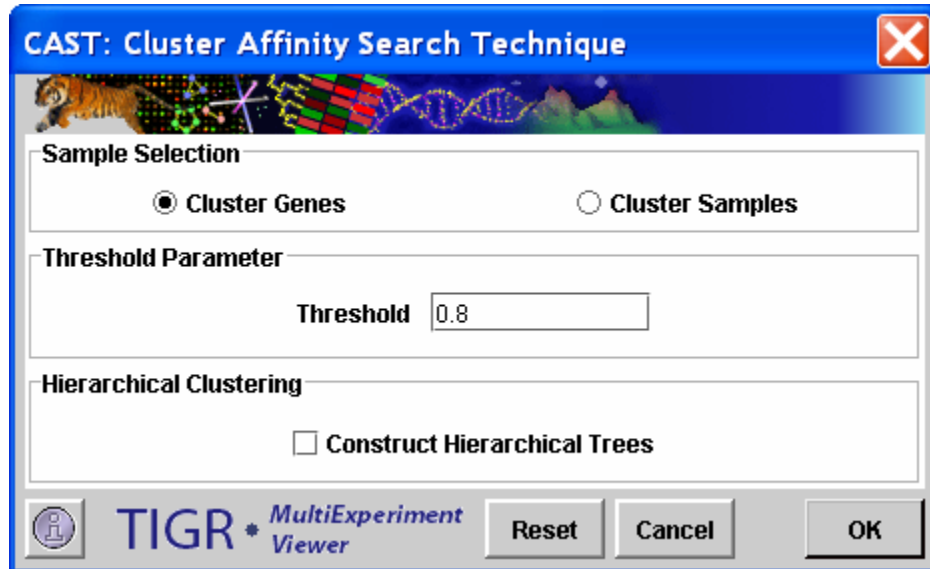
11.6.1. K-Means / K-Medians Support: Initialization Dialog Box.

The number of consensus clusters generated may be more than the input number of clusters per run. This is because some genes may cluster together frequently, yet they may form a subset of different clusters in different runs. Hence, a set of genes that appeared as a single cluster in any given run may be split up into two or more consensus clusters over several runs. Some genes may remain unassigned because they did not cluster with any other genes in enough runs to exceed the threshold percentage.

11.7. CAST: Clustering Affinity Search Technique

(Ben-Dor *et al.* 1999)

The user is prompted for a threshold affinity value between 0 and 1 (which may be thought of as the reciprocal of the distance metric between two genes, scaled between 0 and 1), that has to be exceeded by all genes within a cluster. The algorithm works by both adding and removing genes from a cluster, each time adjusting the affinities of the genes to the current cluster, and continuing this process until no further changes can be made to the current cluster.



11.7.1. CAST Initialization Dialog

Parameters:

Sample Selection

The sample selection option indicates whether to cluster genes or samples.

Threshold

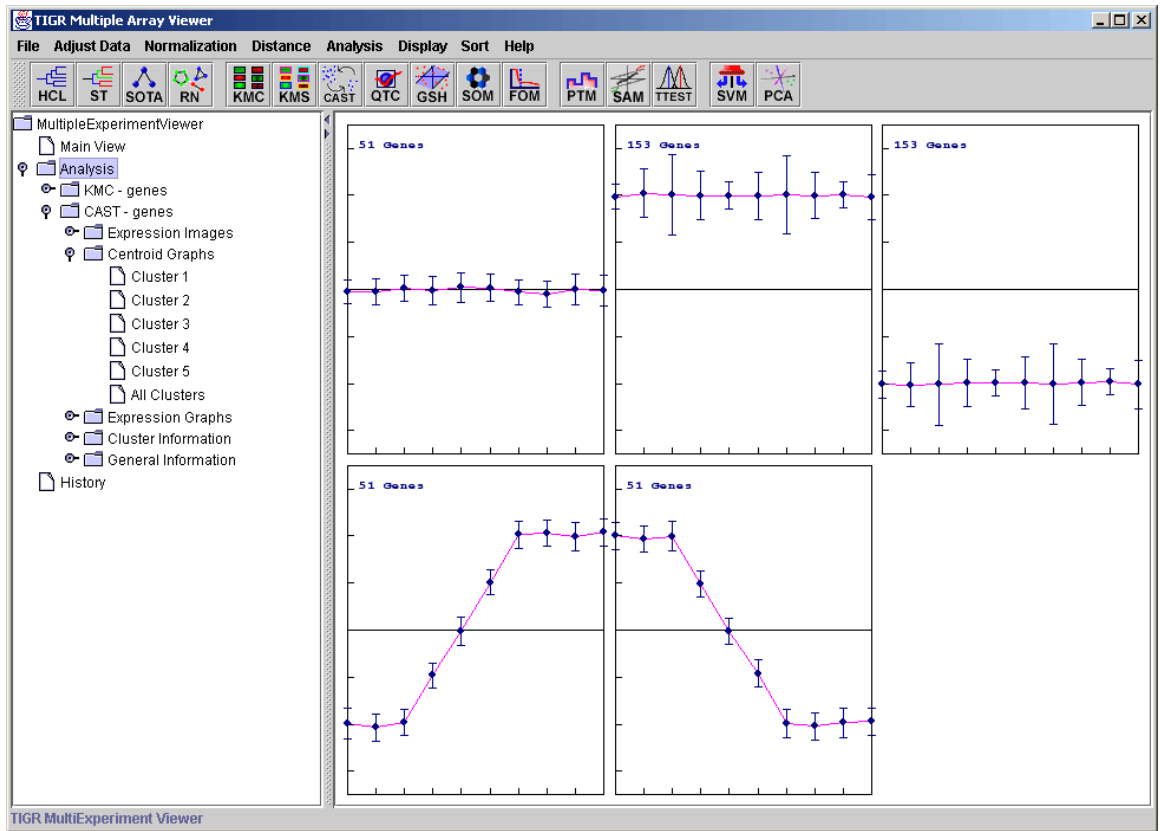
The threshold parameter is a value ranging from 0.0 to 1.0 which is used as a cluster affinity threshold. Each expression element will have an affinity for the current cluster being created based on its relationship to the elements currently in the cluster. If that affinity is greater than the supplied threshold the gene is permitted to be a member of the cluster. Note that thresholds near 1.0 are more stringent and tend to produce many clusters with rather low variability. Conversely, using a lower threshold will produce fewer, more variable clusters. A balance by trial-and-error should be found between these extremes.

Note that in the algorithm expression elements are repeatedly tested for their affinity to the cluster being formed. In that way elements can be added or subtracted based on the current cluster membership.

Hierarchical Clustering

This check box selects whether to perform hierarchical clustering on the elements in each cluster created.

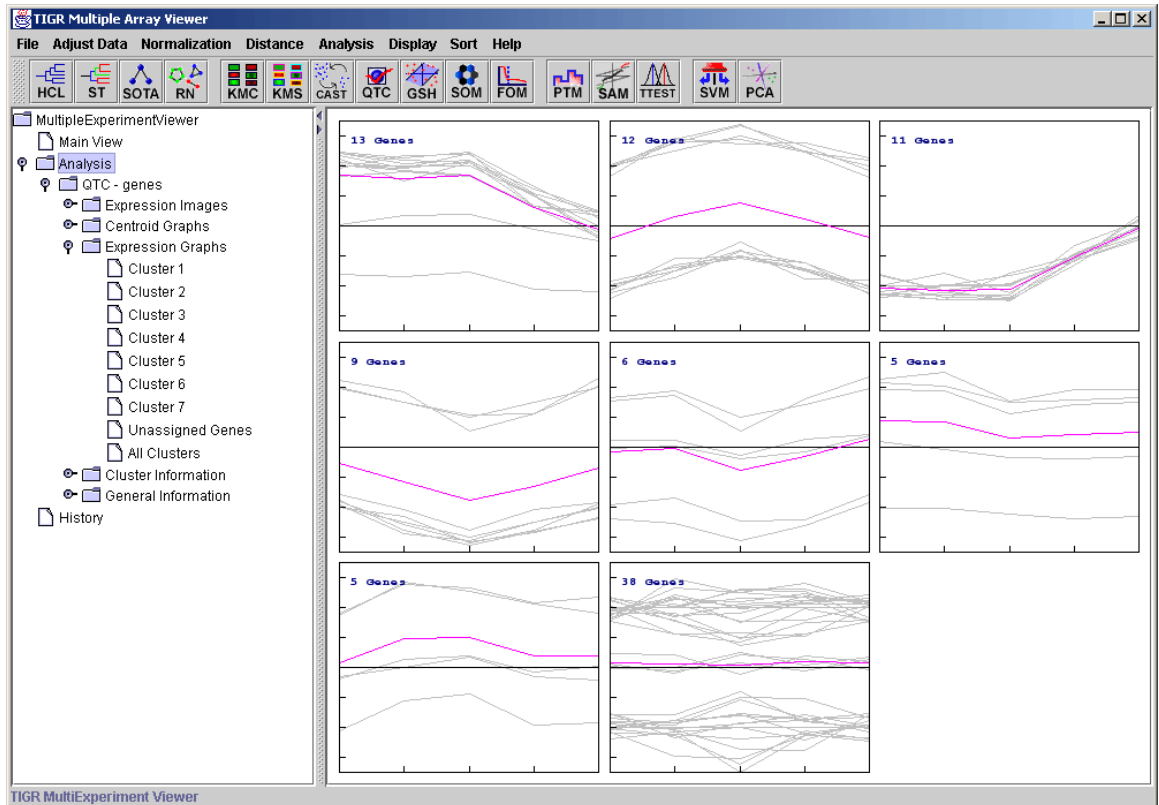
Default Distance Metric: Euclidean



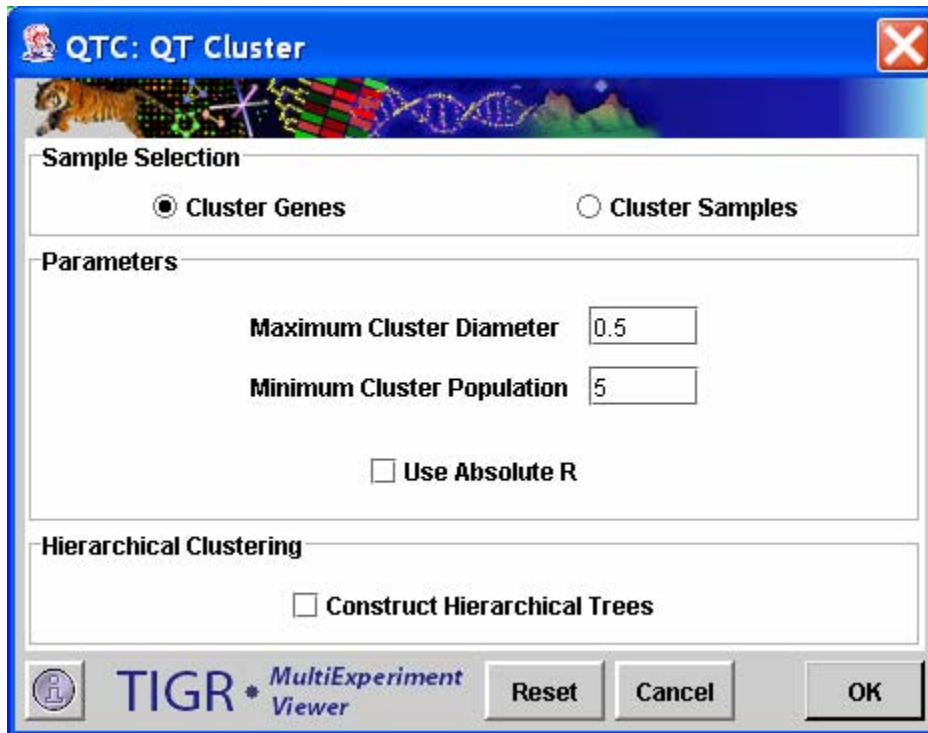
11.7.2. CAST: Centroid View.

11.8. QTC: QT CLUST (modified from Heyer *et al.* 1999)

The dialog will prompt the user for the cluster diameter and the minimum cluster size. The **cluster diameter** is the largest distance allowed between two genes in a cluster, expressed as a fraction between 0 and 1. A diameter of 1 corresponds to the largest possible distance between two genes. For Pearson Correlation, Pearson Uncentered, Pearson Correlation Squared, Kendall's Tau and Cosine Correlation, the maximum possible distance is 1, and therefore the user-input diameter is the actual maximum allowed distance between two genes in a cluster. For the other distance metrics, which do not have a fixed upper bound, the maximum distance between two genes in the current dataset is set to 1, and the diameter is calculated accordingly. To reduce bias resulting from outliers, the distances used for computing clusters are jackknifed, i.e., each experiment is left out in turn while computing the distance between two genes, and the maximum of the distances is taken. The **minimum cluster size** specifies the stopping criterion for the algorithm as it searches through the data set finding smaller and smaller clusters with each iteration. Checking the **Use Absolute** checkbox will include genes with similar as well as opposing trends in a cluster (e.g., if the distance metric selected is Pearson Correlation, both positively and negatively correlated genes will be considered for inclusion in a cluster). If the **Use Absolute** box is unchecked, only genes of similar trends will be considered for inclusion in the same cluster. As with the previous two methods, it is possible to construct hierarchical trees from the clusters. The last displayed group of genes consists of genes that remain unassigned to any cluster. The subnodes on the left panel are similar to the ones previously described for k-means and SOM.



11.8.1. QT CLUST: Expression Graphs.



11.8.2. QTC Initialization Dialog

Parameters:

Sample Selection

The sample selection option indicates whether to cluster genes or samples.

Maximum Cluster Diameter

Cluster Diameter is related to the overall variability between the member elements within a cluster. Maximum Cluster Diameter is a constraint on that variance such that all formed clusters must have a diameter (variability) smaller than the entered maximum. Increasing the maximum diameter tends to make larger and more variable clusters and decreasing the maximum diameter tends to produce smaller less variable clusters.

Minimum Cluster Population

The minimum number of elements required to be present in order to form a cluster. For instance a Minimum Cluster Population of 10 insures that all formed clusters will have at least 10 members.

Use Absolute R

Using this option will group expression patterns that are positively and negatively correlated.

Hierarchical Clustering

This check box selects whether to perform hierarchical clustering on the elements in each cluster created.

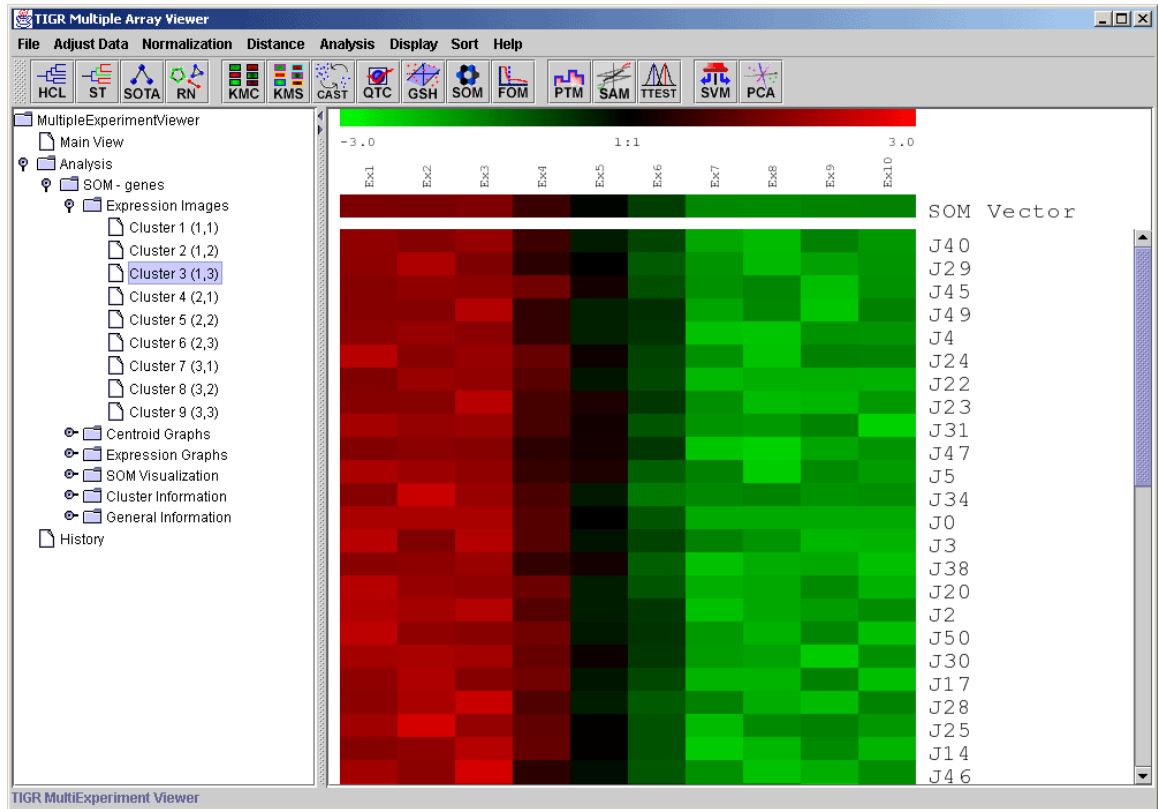
Default Distance Metric: Pearson

This module works very poorly with the average dot-product and covariance distance metrics.

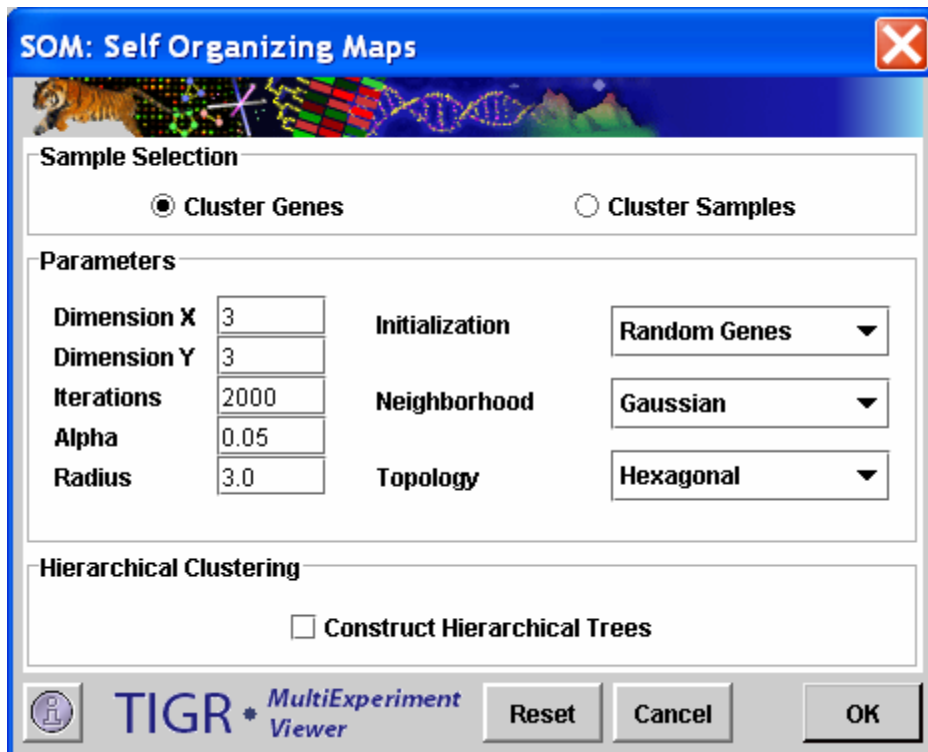
11.9. SOM: Self Organizing Maps

(Tamayo *et al.* 1999, Kohonen, T., 1982)

Selecting this analysis will display a dialog that allows the user to set up the size, topology and behavior of the SOM. Once the computations are complete, select the SOM node under *Analysis* to view the SOM results. The subnodes under this node are very similar in form and function to those found beneath the KMC node. K-Means / K-Medians Support: Initialization Dialog Box.



11.9.1. SOM: Expression Image.



11.9.2. SOM Initialization Dialog

Basic Terminology

Node

An SOM structure to which expression elements are associated to form clusters. Each node contains an SOM Vector.

SOM Vector

A vector of size n which represents its node's location in the n dimensional expression space. Distances from this vector to expression vectors in the input data are used to determine to which node an expression vector should be associated.

Training/Adaptation

The process of repositioning the SOM Nodes by altering their SOM Vectors. The adaptation process is a result of an expression element being associated with a node. The new position is determined by the distance between the expression element and the SOM Vector, the Alpha value, and the neighborhood convention (see below).

Topology

A two dimensional topology used to define how node-to-node distances are calculated.

Note that a cluster is a collection of expression elements associated with a Node.

Parameters:

Sample Selection

The sample selection option indicates whether to cluster genes or samples.

Dimension X

This positive integer value determines the X dimension of the resulting topology.

Dimension Y

This positive integer value determines the Y dimension of the resulting topology. Note that Dimension X times Dimension Y gives the number of clusters that will be produced.

Iterations

This positive integer value indicates the total number of times that the data set will be presented to the network (or Map, Graph). Each expression element will be presented this number of times to train the Nodes.

Alpha

This value is used to scale the alteration of SOM vectors when a new expression vector is associated with a node.

Radius

When using the *bubble neighborhood* parameter this float value is used to define the extent of the neighborhood. If an SOM vector is within this distance from the winning node (the cluster to which an element has been assigned) then that Node (and SOM vector) is considered to be in the neighborhood and it's SOM vector is adapted.

Initialization

Random Genes or Random Samples: Indicates that the initial SOM vectors will be selected at random as actual elements in the data.

Random Vector: Indicates that the initial SOM vectors will be constructed as random vectors generated to reflect the magnitude of the data set. These initial vectors are not actual expression vectors in the data set.

Neighborhood

The neighborhood options indicate the conventions (formulas) used to update (adapt) an SOM vector once an expression vector has been added into a Node's neighborhood.

Bubble: This option uses the provided radius (see above) to determine which surrounding SOM nodes are in the neighborhood and therefore are candidates for adaptation. When this option is selected the Alpha parameter for scaling the adaptation is used directly as provided from the user.

Gaussian: This option forces all SOM vectors in the network to be adapted regardless of proximity to the winning node. In this case the Alpha parameter is scaled based on the distance between the SOM vector to be adapted and the winning node's SOM vector.

Topology

Indicates whether the topology should be rectangular or hexagonal. If rectangular topology is selected the node-to-node distance is determined as Euclidean distance within the two dimensional x-y grid. If hexagonal distance is used an appropriate formula is used to determine the distance given the coordinates of the two nodes.

Hierarchical Clustering

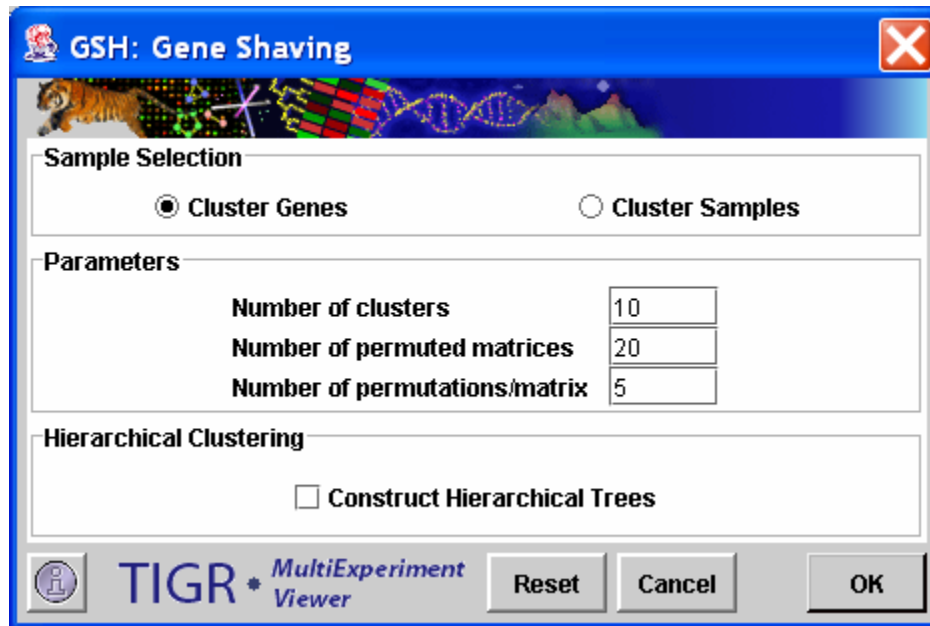
This check box selects whether to perform hierarchical clustering on the elements in each cluster created.

Default Distance Metric: Euclidian

11.10. GSH: Gene Shaving

(Hastie *et al.* 2000)

The clusters that are created by this method differ from the results of other clustering algorithms in several ways. Clusters are constructed such that they show a large variation across the set of samples and small variation between the expression levels of the individual genes. Each cluster is independent of the others and they may overlap other clusters; each gene may belong to several clusters or none at all. One particularly interesting feature of this algorithm is that it will associate genes whose expression levels change by a similar magnitude across experiments, but in the opposite direction. For example, a gene with a given expression pattern across a series of experiments will be clustered with other genes whose expression pattern is the exact opposite.



11.10.1. Gene Shaving initialization dialog.

Parameters:

Sample Selection

The sample selection option indicates whether to cluster genes or samples.

Number of Clusters

This integer value indicates the number of clusters to produce. Note that in the GSH algorithm the clusters do not necessarily represent disjoint sets. Some elements may be represented in more than one cluster while other elements may not be represented at all.

Number of Permuted Matrices

This integer value indicates the number of permuted matrices used to generate an average R^2 (measure of cluster variance) used to generate the gap statistic.

Number of Permutations/Matrix

This integer value represents the number of alterations to each permuted matrix produced to generate the gap statistic.

Hierarchical Clustering

This check box selects whether to perform hierarchical clustering on the elements in each cluster created.

Default Distance Metric: Euclidean

The first principle component of the expression matrix is calculated, and those genes whose variance is in the bottom 10% are removed. These two steps are repeated using the remaining genes until only one gene remains. This results in a series of nested clusters. One cluster is chosen from this series using the gap statistic (see below for details). The expression matrix is then orthogonalized, another series of nested clusters generated and one cluster chosen from it. The process is repeated until the number of chosen clusters reaches the number specified in the “Number of clusters” parameter.

The method used to select one cluster out of a nested series is maximization of the Gap Statistic. Randomized clusters are created from the existing expression matrix. The ratio of expression variance of a given gene between experiments versus the variance of each gene about the cluster average is calculated. The cluster whose ratio is furthest from the average ratio of the randomized matrices is chosen.

This module is computationally intensive, so it may be several minutes before results are displayed. The experiment subtree created by the module contains expression images, centroid graphs and expression graphs of each of the clusters predicted and the genes not assigned to clusters. It also contains a *Cluster Information* tab which reports the sizes of each cluster.

11.11. FOM: Figures of Merit

(Yeung *et al.* 2001)

The Figure of Merit is, in concept, a measure of fit of the expression patterns for the clusters produced by a particular algorithm. MeV's FOM implementation provides FOM results for running the KMC and CAST clustering algorithms. Each algorithm is initialized by selecting either the K-Means/K-Medians tab or the CAST tab.

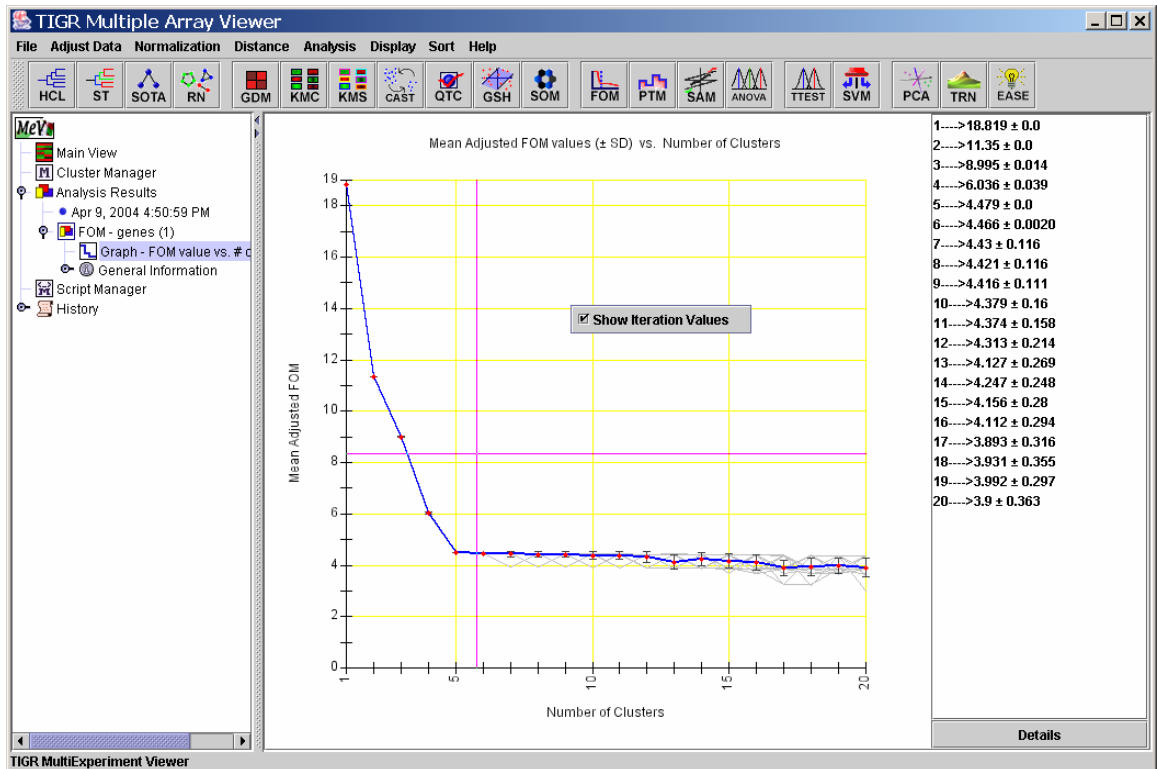
Currently, FOM is available for the CAST, K-means and K-medians algorithms. A figure of merit is an estimate of the predictive power of a clustering algorithm. It is computed by removing each sample in turn from the data set, clustering genes based on the remaining data, and calculating the fit of the withheld sample to the clustering pattern obtained from the other samples. The lower the adjusted FOM value is, the higher the predictive power of the algorithm. The "Maximum number of clusters" input field under the K-Means / K-Medians tab in the initialization box is used to determine how many times FOM values should be calculated for the k-means/k-medians algorithm. Each time, the number of clusters computed is increased by one, starting with one cluster in the first iteration. The "Interval" input field under the CAST tab allows the user to specify the increase in threshold affinity in successive iterations of CAST. If the "Take Average" box is checked, in case there is more than one clustering outcome for a given number of clusters, the average FOM for that number of clusters will be used to draw the FOM-vs.-number of clusters curve. If the "Take Average" box is unchecked, each FOM value will be represented in case of such a tie, and curves will be drawn through each value.

In the figure below, the value of the adjusted FOM for a K-means run decreases steeply until the number of clusters reaches 4, after which it levels out. This suggests that, for this data set, K-means performs optimally for 4 clusters and that any additional clusters produced will not add to the predictive value of the algorithm. FOM is useful in determining the best input parameters for a clustering algorithm.

FOM Input Parameters

FOM Iteration Selection

This field specifies a number of FOM iterations to run during the analysis. When the K-Means mode is selected and the number of FOM iterations is greater than one, the mean FOM values are reported with the standard deviation on the output graph. This is useful in the case of K-means where the initialization step involves an initial random partitioning. Each K-means run is potentially unique although each should be similar. When using this option the result can be based on several runs. A right click in the graph will provide the option to show or hide the individual lines representing FOM iterations.



11.11.1. FOM vs. No. of Clusters graph for KMC algorithm. (1-20 clusters, 20 iterations)

FOM: Figure of Merit

Sample Selection

Gene Cluster FOM Sample Cluster FOM

FOM Iteration Selection

Number of FOM Iterations:

K-Means / K-Medians **CAST**

Calculate means Calculate medians

Maximum number of clusters (enter an integer > 0):

Maximum number of iterations (enter an integer > 0):

K-Means / K-Medians will be run using a starting K (number of clusters) = 1, with K being incremented by 1 in each subsequent iteration, up to the maximum number of clusters specified above

TIGR * MultiExperiment Viewer

11.11.2. FOM Initialization Dialog (CAST)

CAST Parameters

Threshold Interval

For FOM an interval is used to perform a series of CAST runs in which the Affinity Threshold is incremented from 0.0 by the interval indicated. The default of 0.1 is often a good value since it provides 11 CAST results from 0.0 to 1.0 incremented by 0.1.

The threshold parameter is a value ranging from 0.0 to 1.0 which is used as a cluster affinity threshold. Each expression element will have an affinity for the current cluster being created based on its relationship to the elements currently in the cluster. If that affinity is greater than the supplied threshold the gene is permitted to be a member of the cluster.

FOM: Figure of Merit

Sample Selection

Gene Cluster FOM Experiment Cluster FOM

FOM Iteration Selection

Number of FOM Iterations:

K-Means / K-Medians **CAST**

Calculate means Calculate medians

Maximum number of clusters (enter an integer > 0):

Maximum number of iterations (enter an integer > 0):

K-Means / K-Medians will be run using a starting K (number of clusters) = 1, with K being incremented by 1 in each subsequent iteration, up to the maximum number of clusters specified above

TIGR MultiExperiment Viewer

Reset Cancel OK

11.11.3. FOM Initialization Dialog (K-means/K-medians)

KMC Parameters

Means/Medians Option

The Means or Medians option indicates whether each cluster's centroid vector should be calculated as a mean or as a median of the member expression patterns.

Maximum Number of Clusters

This positive integer value indicates the maximum number of clusters to be created. For instance, if the entered value is 10 then KMC is run 10 times to produce 1,2,3...,10 clusters. An FOM value is returned for each run.

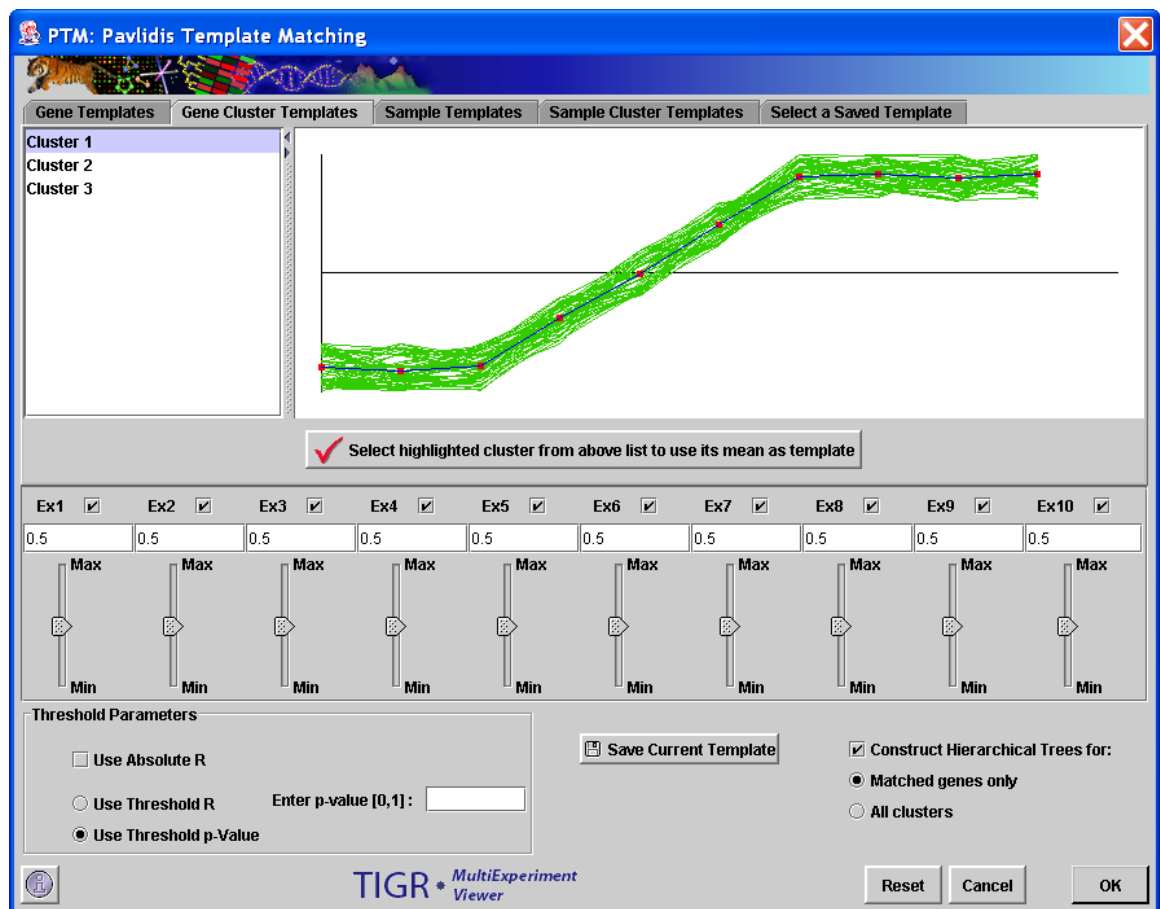
Maximum Number of Iterations

This positive integer value is the maximum number of times that all the elements in the data set will be tested for cluster fit within a single KMC run. On each iteration each element is associated with the cluster with the closest mean (or median). Note that a KMC run will terminate when either no elements require migration (reassignment) to new clusters or when the maximum number of iterations has been reached.

11.12. PTM: Template Matching

(Pavlidis and Noble 2001)

The user can specify a template expression profile for a gene (a series of relative expression ratios between 0 and 1), and the data set will be searched for matches to the template, based on the Pearson Correlation between the template and the genes in the data set. The template profile can be specified in one of several ways: 1) by selecting one of the genes in the data set as a template from the list on the upper left, and then clicking the "Select highlighted gene" button; 2) doing the same thing with one of the cluster means, assuming that clusters have already been set by some other method; 3) entering values between 0 and 1 in the text input fields above the slider bars corresponding to each experiment; or, 4) Adjusting the slider bars to the desired values. Matches can be made by considering either the signed or the unsigned values of correlation coefficient (using the checkbox labeled "Match to Absolute R?"), and the threshold criterion for matching can be either the magnitude of the correlation coefficient, or the significance (p-value) of the correlation coefficient.



11.12.1. Template Matching (PTM) initialization dialog.

Parameters:

Template Selection Tabs

The five tabbed panels at the top of the dialog select to view candidate templates from various sources.

The *Gene Template tab* provides a list of genes in the data set and their expression profile graphs. Selection of elements in the list display the expression pattern for that gene.

The *Gene Cluster Template tab* provides a list and view of templates which are the mean values of stored (colored) clusters.

The *Sample Template*, and *Sample Cluster Template tabs* provide the same functionality but use experiment templates.

The *Saved Template tab* provides an interface for loading gene and experiment templates. Templates loaded from files will populate a list from which a template may be selected.

A button in each of these areas is used to select the displayed template for matching.

Threshold Parameters

Use Absolute R

Using this option will select expression patterns that are either positively or negatively correlated with the template.

Use Threshold R

This option will indicate that the threshold value for determining a match is the R value between the expression vector and the template.

Use Threshold p-Value

This option will indicate that the threshold value for determining a match is the p value on R between the expression vector and the template.

Threshold Input Value

This is either a supplied value for R or a p-value ranging from 0.0 to 1.0. R values closer to 1.0 are more stringent. p-values closer to 0 are more stringent.

Save Template

This button launches a file browser to allow the user to save the current template to a tab delimited text file. These files can be loaded from the Saved Template tab interface.

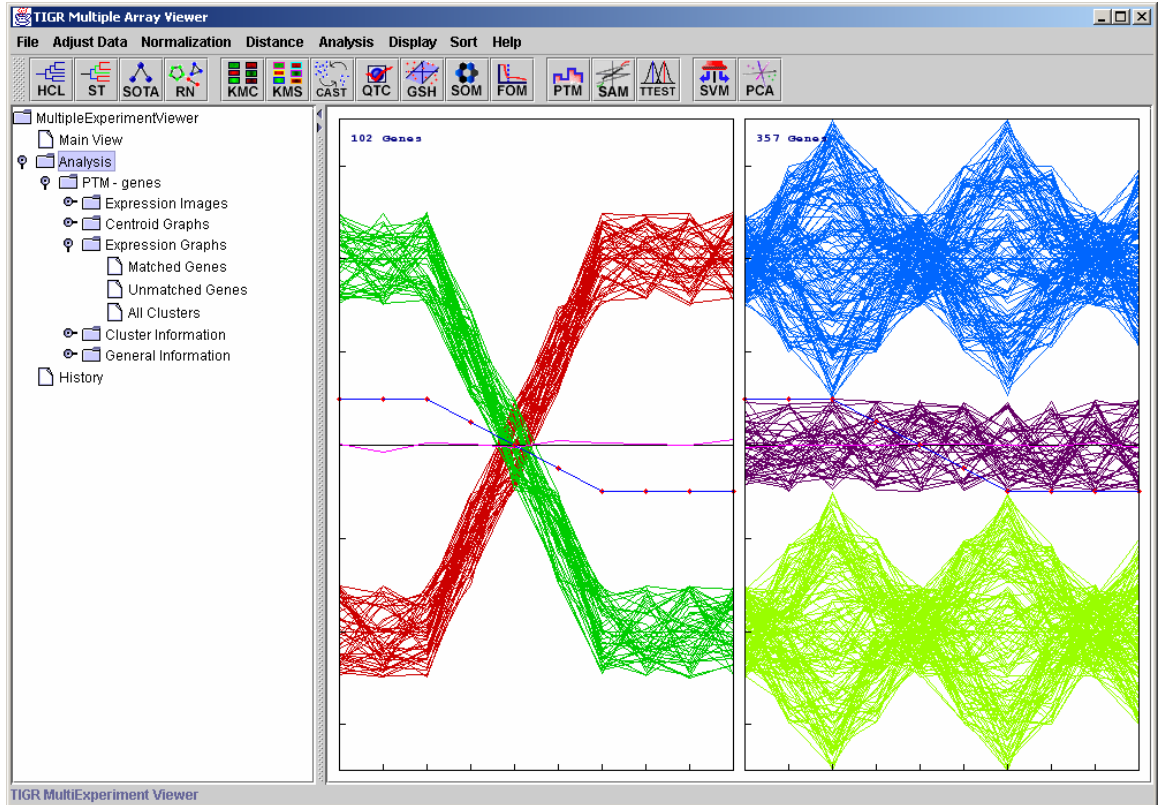
Hierarchical Clustering

This check box selects whether to perform hierarchical clustering on the elements in the resulting matched and unmatched element sets.

Default Distance Metric: Pearson (fixed, will not correspond to the distance menu)

Template matching is particularly useful when the researcher is searching for a specific expression pattern. Applying this method with the input parameters in the previous figure gives the following output, where the first panel on the right

corresponds to genes that matched the template, and the second panel to genes that did not match:



11.12.2. PTM results: Expression Graphs.

11.13. TTEST: T-tests

(Dudoit *et al.* 2000, Pan 2002, Welch 1947, Zar 1999, Korn, *et al.* 2001, 2004)

Three t-test designs are implemented: one-sample, paired and between-subjects. In the one-sample design, the user specifies a mean. Each gene whose mean log₂ expression ratio over all included samples is significantly different from the user-specified mean is assigned to one cluster, while those genes whose means are not significantly different from the user-specified mean are assigned to another cluster. To exclude a sample from the analysis, uncheck the box next to that sample's name in the left pane of the one-sample screen.

In the between-subjects design, samples can be assigned to one of two groups, and genes that have significantly different mean log₂ expression ratios between the two groups are assigned to one cluster, while the genes that are not significantly different between the two groups are assigned to another cluster. The user may choose to exclude some samples from the analysis, which can be done by selecting the "neither group" option for those samples in the initialization dialog (see screenshot below). For the between-subjects t-test, we can use the Welch t-test for small samples with unequal variances in the two groups (Welch 1947), or assume equal variances.

In the paired design, samples are not only assigned to two groups, but there is also a one-to-one pairing between a member of group A and a corresponding member of group B (e.g., gene expression measurements on a group of subjects, where measurements are taken before (Group A) and after (Group B) drug treatment on each subject).

T-values are calculated for each gene, and p-values are computed either from the theoretical t-distribution, or from permutations of the data for each gene. Whether a gene's mean expression level is significantly different between the two groups is determined either by directly comparing the gene's p-value with the user-specified critical p-value or alpha, or by adjusting the p-values using a correction for multiple testing (see screenshot below).

TTEST: T-test

One-class | **Between subjects** | Paired

Ex1
 Ex2
 Ex3
 Ex4
 Ex5
 Ex6
 Ex7
 Ex8
 Ex9
 Ex10

Enter the mean value to be tested against:

Save settings | Load settings | Reset

Variance assumption (for between subjects t-test only)

Welch approximation (unequal group variances)
 Assume equal group variances

P-Value Parameters

p-values based on t-distribution
 p-values based on permutation:

- Randomly group samples times
- Use all permutations

Overall alpha (critical p-value):

p-value / false discovery corrections

just alpha (no correction)
 standard Bonferroni correction
 adjusted Bonferroni correction

Step-down Westfall and Young methods (for permutations only): minP maxT

False discovery control (permutations only)

With confidence of [1 - alpha]:

EITHER, The number of false significant genes should not exceed


 OR, The proportion of false significant genes should not exceed

 Fast approximation (but possibly conservative)
 Complete computation (possibly slow)

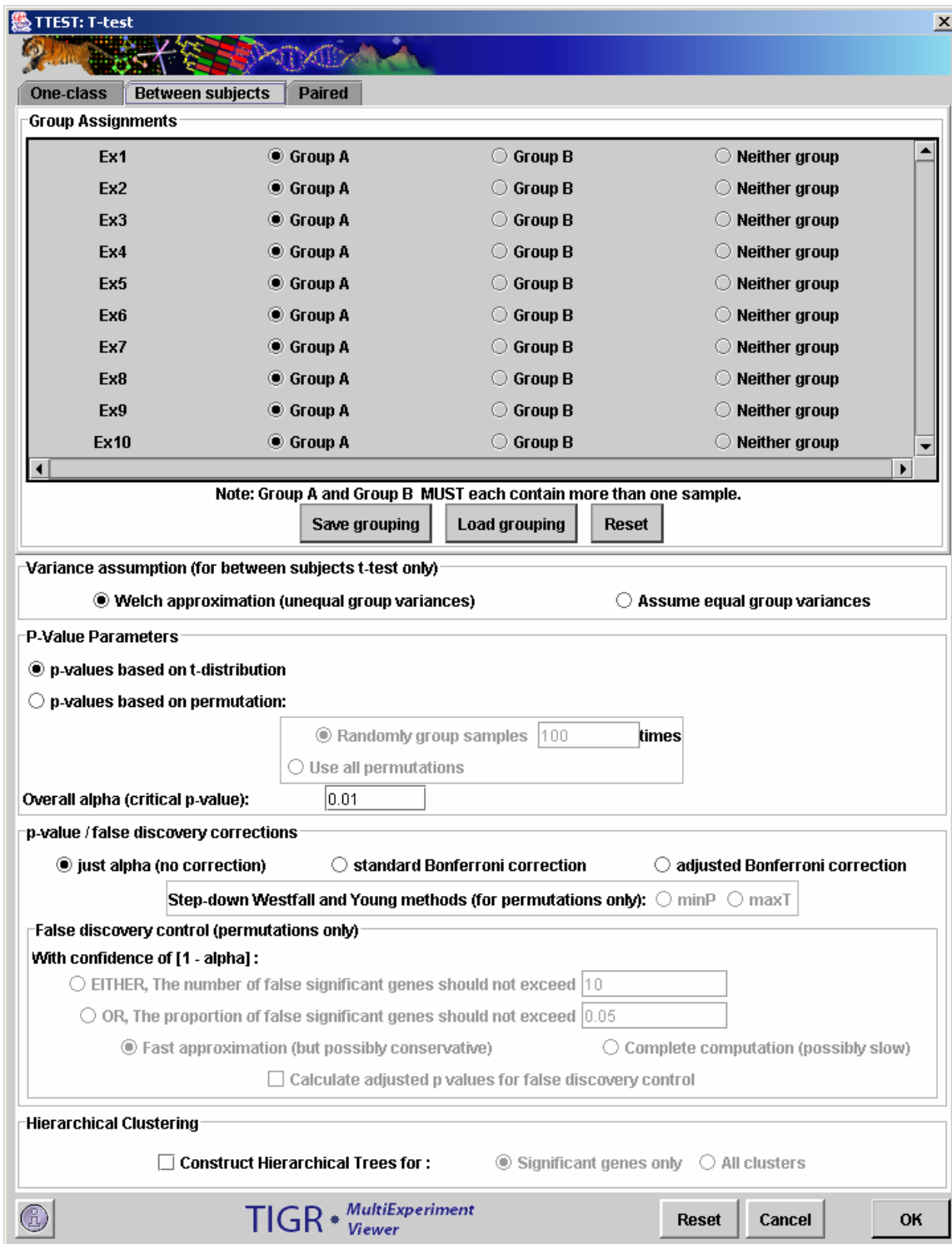
Calculate adjusted p values for false discovery control

Hierarchical Clustering

Construct Hierarchical Trees for :
 Significant genes only
 All clusters


TIGR * MultiExperiment Viewer

(a)



(b)

11.13.1. a and b. TTEST Initialization Dialog Box.

Parameters:

One-class / Between Subjects / Paired Panel

In the one-class design, samples can be included or excluded by checking or unchecking the checkboxes next to each sample name. The user can also specify the test mean. In the between-subjects panel, the buttons permit each sample to be placed into group A, group B, or neither group.

If an experiment is placed in neither group it will be ignored for the purposes the analysis. Note that groups A and B must each have at least two members following the assignment. The paired panel allows the specification of pairs of experiments.

Save Grouping / Setting

The save grouping / setting button allows you to save the grouping or setting to file. This is particularly useful when there are many experiments.

Load Grouping / Setting

This button allows you to select and load a saved grouping or setting.

Reset

The reset button returns all of the settings to the original settings.

P-Value Parameters

This set of controls are used to indicate the method by which p-values are determined for each gene and allows the input of the critical p-value. p-values can be computed either from the theoretical t-distribution, or from permutations of the data for each gene between the two groups.

p-values based on t-distribution

Using this option a gene's p-value is taken directly from the theoretical t-distribution based on the gene's calculated t-value.

p-values based on permutation

Using this option, a gene's p-value is determined by forming a distribution based on permutations of the data for that gene. For the one class t-test, in each round of permutation, some of the values in the expression vector are picked at random to be replaced by the following quantity: $(\text{original value}) - 2 \times (\text{original value} - \text{hypothesized mean})$. Thus, the randomized vectors have some of their elements randomly “flipped” about the hypothesized mean. For the between subjects t-test, the permutations allow each value in the expression vector in group A or group B to be randomly placed into either group (the size of each group is conserved). t-values are constructed following each permutation to construct a distribution which is used to generate p values for each gene based on its t-value. If permutations are used, two buttons allow you to select to permute the values a number of times indicated, or to permute the values a number of times equal to the maximum number of permutations possible.

Critical p-value

This text field allows you to enter the alpha or critical p-value.

P-Value / false discovery Corrections

The p-values for each p can be adjusted to correct for the large number of observations (genes) and the increased possibility of considering a gene without a real significant change to be considered significant.

Alternatively, a false discovery threshold can be set such that the number

or proportion of false positives in the significant gene list does not exceed a specified level with a certain confidence.

Just Alpha (no correction)

Using this option the alpha is not altered.

Standard Bonferroni Correction

In the standard Bonferroni correction, the user-specified alpha is divided by the number of genes to give the critical p-value. This is much more stringent than using an uncorrected alpha.

Adjusted Bonferroni Correction

In the adjusted Bonferroni correction, the t-values for all the genes are ranked in descending order. For the gene with the highest t-value, the critical p-value becomes (α / n) , where n is the total number of genes; for the gene with the second-highest t-value, the critical p-value will be $(\alpha / n-1)$, and so on. The stringency of this correction falls somewhere between no correction and the Standard Bonferroni.

Step-down Westfall-Young MaxT correction (Dudoit et al. 2003):

In this method, the genes are ranked in descending order of their absolute t-values, and the adjusted p-values are computed by an algorithm described in Dudoit et al. 2003.

False discovery control:

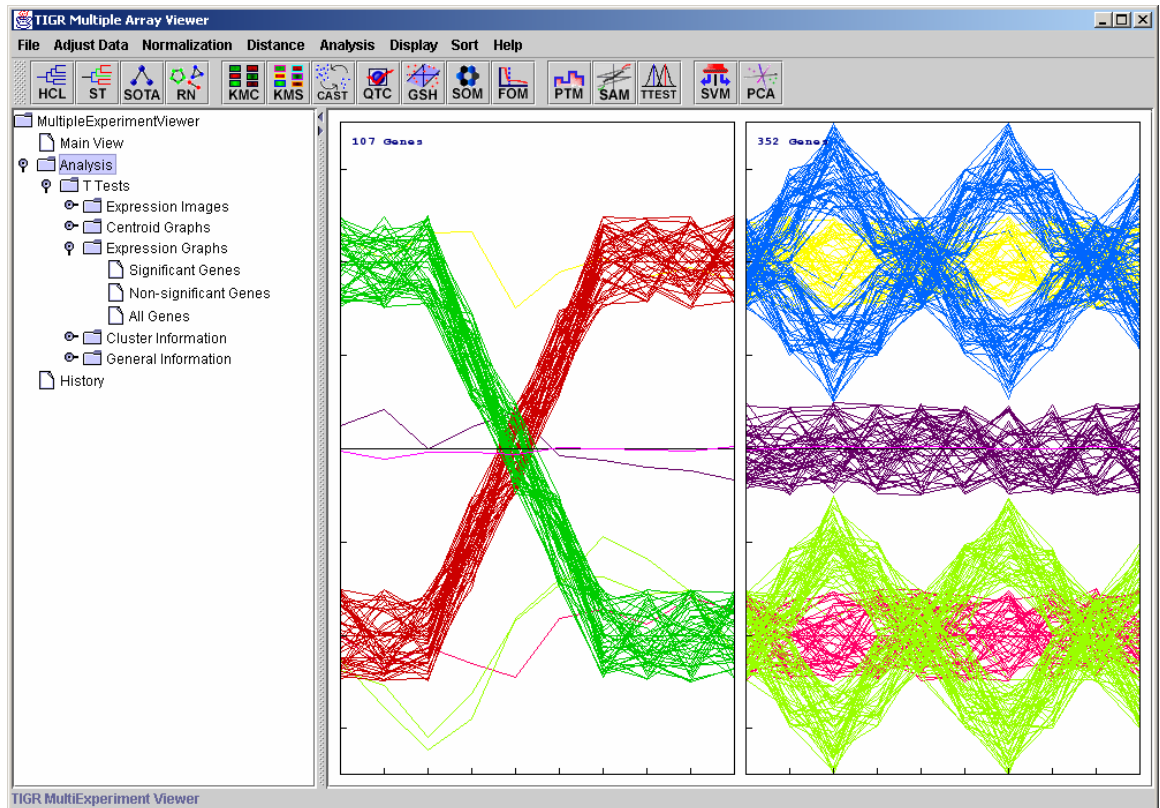
The algorithms are described in Korn *et al.* 2001, 2004.

Hierarchical Clustering

This check box selects whether to perform hierarchical clustering on the elements in each cluster created.

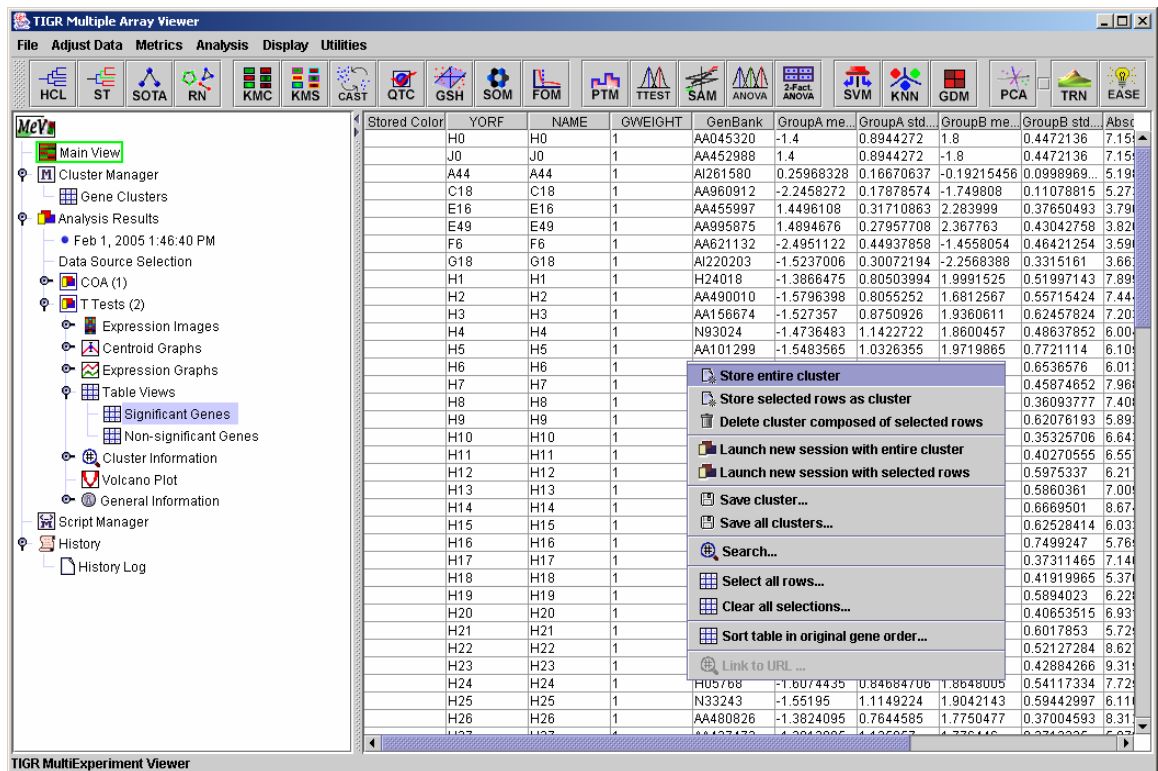
P-value corrections reduce the probability that a non-significant gene will be erroneously picked as significant. This can be a serious issue when many tests are done (which is usually the case in microarray analyses, as there are as many tests as there are genes in the analysis). The standard Bonferroni correction is very stringent and may exclude many genes that are really significant, whereas the adjusted Bonferroni correction is less conservative, and more likely to include significant genes while still controlling the error rate. The step-down Westfall-Young MaxT correction is also less conservative than the standard Bonferroni correction, and still provides statistical power. False discovery control is a useful option as p-value corrections can be too stringent for microarray analysis.

Sample output from this module is shown below:



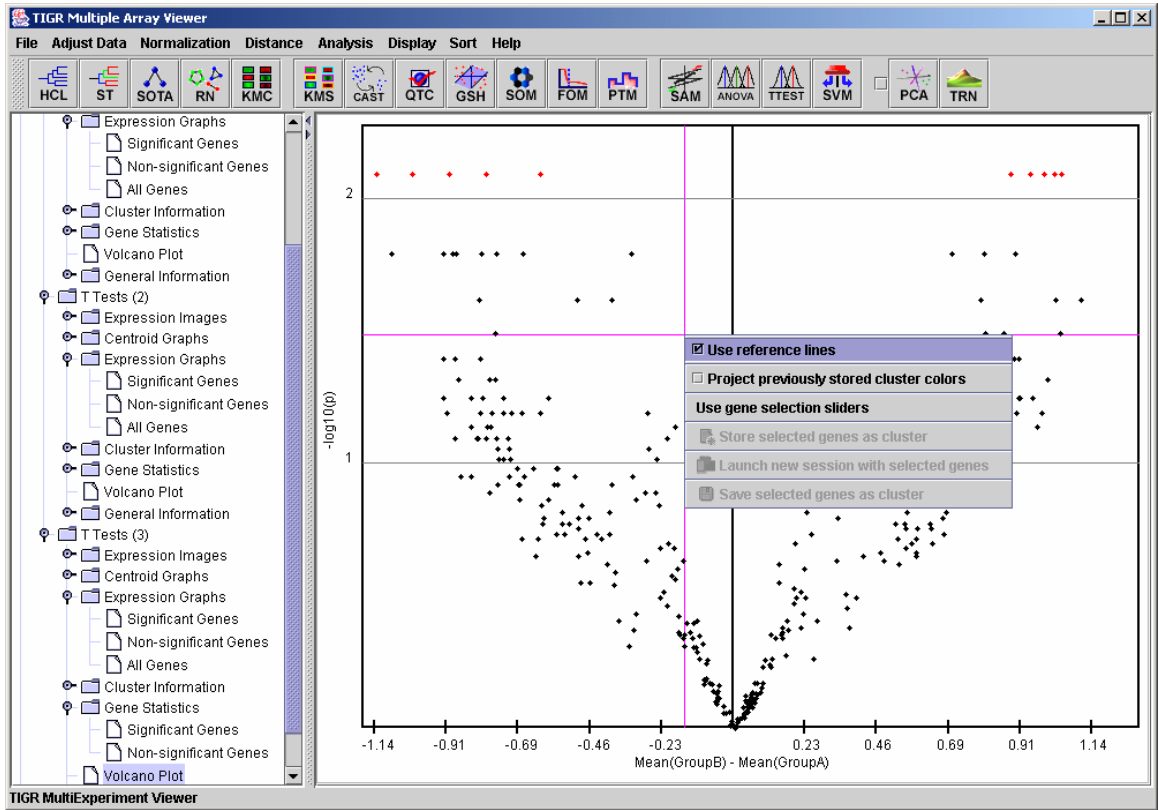
11.13.2. TTEST Results: Expression Graphs (significant genes are in the graph on the left).

The TTEST module also outputs table viewers with gene-specific statistics as shown below. These tables can be saved as tab-delimited text files by right-clicking on them. All the columns in the tables can be sorted in ascending or descending order. Successive clicks on a column header re-order the rows in ascending or descending order of the values in the selected column. Holding down the CTRL key while clicking anywhere on the header will restore the original ordering.



11.13.3. TTEST Gene Statistics Table Viewer

Another non-standard TTEST viewer is the volcano plot viewer. This plot shows the difference between the means of groups A and B for each gene plotted against the negative \log_{10} p-value associated with its t-value. A volcano plot gives an intuitive visual sense of the nature of the relationship between the mean differences between groups and the statistical significance of those differences, for the data set as a whole. Right-clicking on this plot brings up options for toggling the reference lines on and off, selecting genes from the plot using slider bars, storing these selected genes as clusters, and projecting cluster colors from previous analyses on to the volcano plot.



11.13.4. TTEST Volcano plot

11.14. SAM: Significance Analysis of Microarrays

(Tusher et al. 2001 implemented as in Chu et al. 2002)

SAM can be used to pick out significant genes based on differential expression between sets of samples. It is useful when there is an *a-priori* hypothesis that some genes will have significantly different mean expression levels between different sets of samples. For example, one could look at differential gene expression between tissue types, or differential response to exposure to a perturbation between groups of test subjects. A valuable feature of SAM is that it gives estimates of the False Discovery Rate (FDR), which is the proportion of genes likely to have been identified by chance as being significant. Furthermore, SAM is a very interactive algorithm. It allows users to eyeball the distribution of the test statistic, and then set thresholds for significance (through the tuning parameter delta) after looking at the distribution. The ability to dynamically alter the input parameters based on immediate visual feedback, even before completing the analysis, should make the data-mining process more sensitive.

Currently, SAM is implemented for the following designs:

1) **Two-class unpaired**, where samples fall in one of two groups, and the subjects are different between the two groups (analogous to a between subjects t-test). The initialization dialog box is similar to the t-test dialog (Fig. 11.14.1).

The user inputs the group memberships of the samples in the top panel. In the two-class design, genes will be considered to be “positive significant” if their mean expression in group B is significantly higher than in group A. They will be considered “negative significant” if the mean of group A significantly exceeds that of group B.

2) **Two-class paired**, in which samples are not only assigned to two groups, but there is also a one-to-one pairing between a member of group A and a corresponding member of group B (e.g., gene expression measurements on a group of subjects, where measurements are taken before (Group A) and after (Group B) drug treatment on each subject).

3) **Multi-class**, where the user specifies the number of groups (>2) that samples fall into. Genes will be considered significant if they are significantly different in expression across some combination of the groups.

4) **Censored survival**, where each sample is associated with a time and a state (censored or dead). Censored samples are those for which the subject was alive at the time the data were collected, and no further data are available for those subjects.

5) **One-class**, in which the user specifies a value against which the mean expression of each gene is tested. A gene is considered significant if its mean log₂ expression ratio over all included samples is significantly different from the user-specified mean. To exclude a sample from the analysis, uncheck the box next to that sample’s name in the left pane of the one-class screen.

The data for each gene are permuted, and a test statistic d is computed for both the original and the permuted data for each gene. In the two-class unpaired design, d is analogous to the t-statistic in a t-test, in that it captures the difference among mean expression levels of experimental conditions, scaled by a measure of variance in the data. Missing values in the input data matrix are imputed by one of two methods: 1) **Row average**: replacing missing expression measurements with the mean expression of a row (gene) across all columns (experiments), OR 2) **K-nearest neighbors**: where the “K” most similar genes (using Euclidean distance) to the gene with a missing value are used to impute the missing value.

SAM Initialization

Two-class unpaired | Two-class paired | Multi-class | Censored survival | One-Class

Group Assignments

Experiment	Group A	Group B	Neither group
Ex1	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ex2	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ex3	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ex4	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ex5	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ex6	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ex7	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ex8	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ex9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>

Note: Group A and Group B MUST each contain more than one experiment.

Save grouping | Load grouping | Reset

Number of permutations

Enter number of permutations:

S0 and Q Value parameters

Select S0 using | OR Enter s0 percentile (0-100)

Calculate q-values? No (quick) | Yes (slow!)

Imputation Engine

K-nearest neighbors imputer | Number of neighbors:

Row average imputer

Save Imputed Matrix

Hierarchical Clustering

Construct Hierarchical Trees

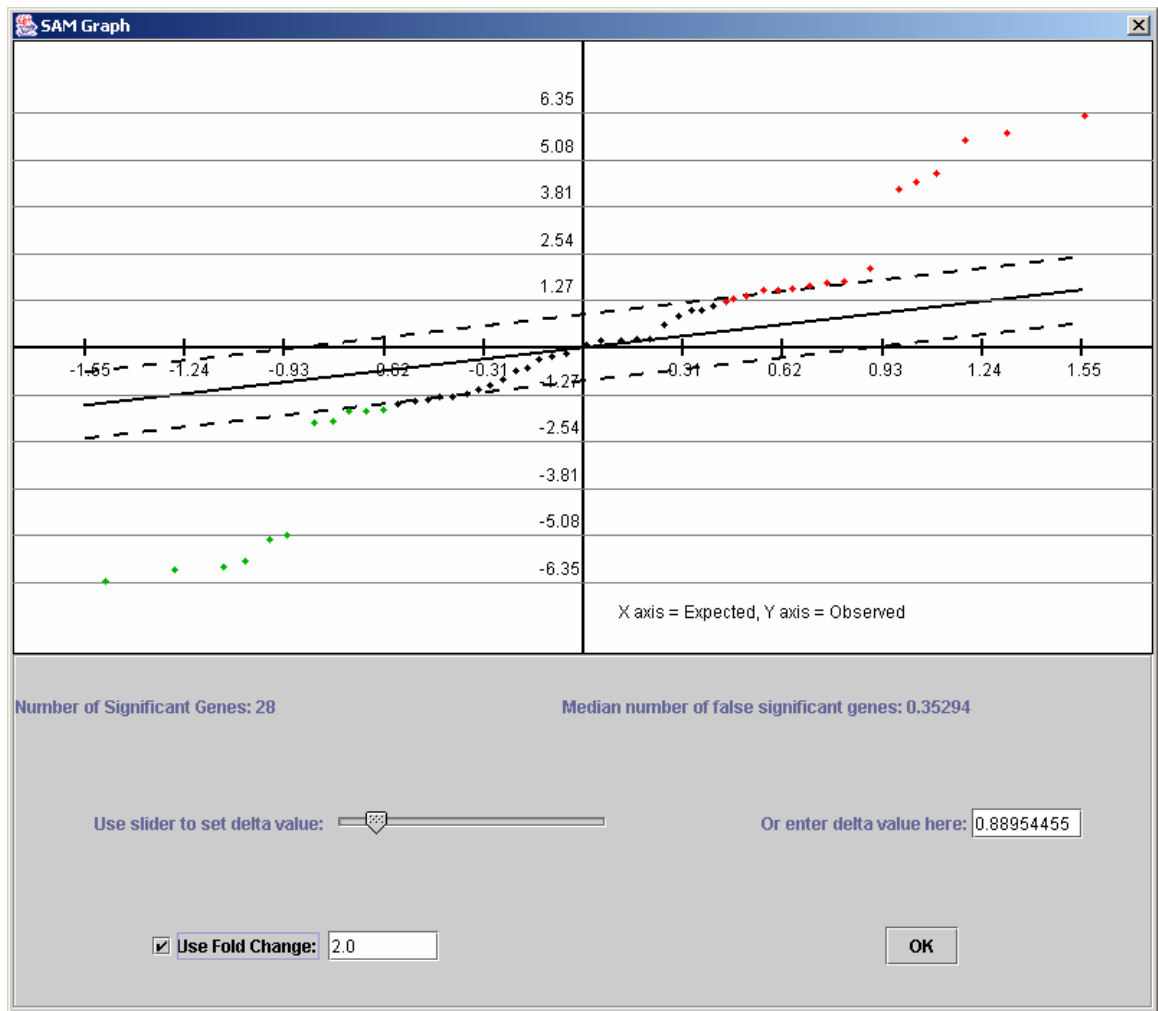
TIGR MultiExperiment Viewer | Reset | Cancel | OK

11.14.1. SAM Initialization Dialog

SAM generates an interactive plot (Fig. 11.14.2) of the observed vs. expected (based on the permuted data) d -values. The user can change the value of the tuning parameter delta using either the slider bar or the text input field below the plot. *Delta* is a vertical distance (in graph units) from the solid line of slope 1 (i.e., where observed = expected). The two dotted lines represent the region within +/- delta units from the “observed = expected” line. The genes whose plot values are represented by black dots are considered non-significant, those colored red are positive significant, and the green ones are negative significant. The user can also choose to apply a **fold change** criterion for the two-class paired and unpaired designs. In this case, in addition to satisfying the delta criterion, a gene will also have to satisfy the following condition to be considered significant:

For a given fold change F ,

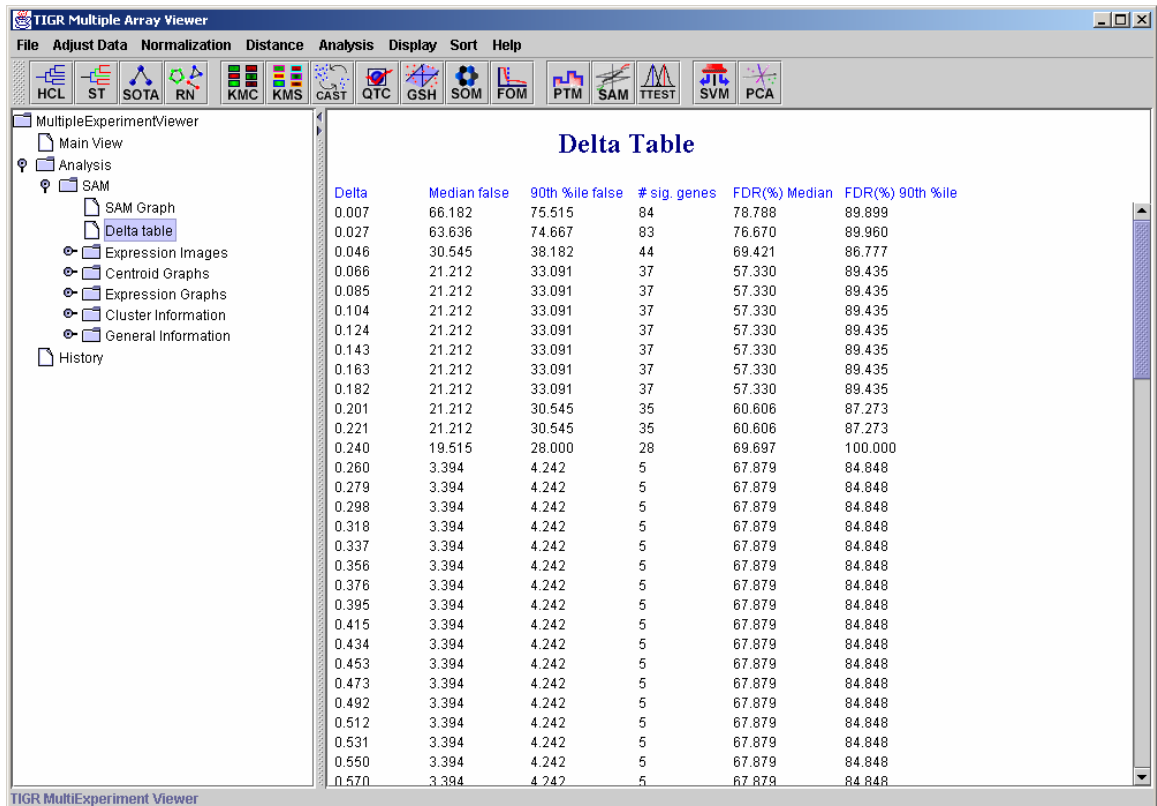
[Mean (unlogged group B) / Mean (unlogged group A values)] $\geq F$ (for positive significant genes), or $\leq 1/F$ (for negative significant genes), where F is the fold change.]



11.14.2. SAM Output

If SAM has been used at least once during a run of MeV, the input parameters and SAM graph of the last run can be called up by default, thus bypassing the need to run SAM again for that set of parameters.

In addition to the standard viewers and information tabs, SAM also outputs a SAM graph viewer, as well as a Delta table viewer (Fig 11.14.3), which contains output information for a range of SAM values. This information can be saved as a tab-delimited text file by right-clicking on the table. The clusters saved from the other viewers will store gene-specific SAM statistics in addition to the annotation and expression measurements stored in clusters from most other modules.



11.14.3. SAM Delta Table Viewer

11.15. ANOVA: Analysis of Variance

(Zar 1999, pp 178-182)

ANOVA is an extension of the t-test to more than two experimental conditions. It picks out genes that have significant differences in means across three or more groups of samples. Currently, only one-way or single-factor analysis of variance is implemented. The user is initially required to enter the number of groups, following which a sample grouping panel similar to the t-test panel, with the appropriate number of groups, is created. Samples can be assigned to any group or excluded from the analysis. F-statistics are calculated for each gene, and a gene is considered significant if p-value associated with its F-statistic is smaller than the user-specified alpha or critical p-value. Currently, p-values are computed only from the F-distribution.

One-way ANOVA Initialization

Number of groups

Group Assignments

Ex1	<input checked="" type="radio"/> Group 1	<input type="radio"/> Group 2	<input type="radio"/> Group 3	<input type="radio"/> Not in groups
Ex2	<input checked="" type="radio"/> Group 1	<input type="radio"/> Group 2	<input type="radio"/> Group 3	<input type="radio"/> Not in groups
Ex3	<input checked="" type="radio"/> Group 1	<input type="radio"/> Group 2	<input type="radio"/> Group 3	<input type="radio"/> Not in groups
Ex4	<input checked="" type="radio"/> Group 1	<input type="radio"/> Group 2	<input type="radio"/> Group 3	<input type="radio"/> Not in groups
Ex5	<input checked="" type="radio"/> Group 1	<input type="radio"/> Group 2	<input type="radio"/> Group 3	<input type="radio"/> Not in groups
Ex6	<input checked="" type="radio"/> Group 1	<input type="radio"/> Group 2	<input type="radio"/> Group 3	<input type="radio"/> Not in groups
Ex7	<input checked="" type="radio"/> Group 1	<input type="radio"/> Group 2	<input type="radio"/> Group 3	<input type="radio"/> Not in groups
Ex8	<input checked="" type="radio"/> Group 1	<input type="radio"/> Group 2	<input type="radio"/> Group 3	<input type="radio"/> Not in groups
Ex9	<input checked="" type="radio"/> Group 1	<input type="radio"/> Group 2	<input type="radio"/> Group 3	<input type="radio"/> Not in groups
Ex10	<input checked="" type="radio"/> Group 1	<input type="radio"/> Group 2	<input type="radio"/> Group 3	<input type="radio"/> Not in groups

Note: Each group MUST each contain more than one experiment.

P-value parameters

Enter alpha (critical p-value):

Hierarchical Clustering

Construct Hierarchical Trees

TIGR * MultiExperiment Viewer

11.15.1. One-way ANOVA initialization dialog box.

Parameters

Group Assignments

Group Selection Controls

This set of buttons permits each experiment to be placed into any group, or no group. If a sample is placed in no group it will be ignored for the purposes the analysis.

Note that each group must each have at least two members following the assignment.

Save Grouping

The save grouping button allows you to save the grouping to file. This is particularly useful when there are many experiments.

Load Grouping

This button allows you to select and load a saved grouping.

Reset

The reset button returns all of the group selection control buttons to Group 1, the initial state.

P-Value Parameters

This is used to input the critical p-value. P-values are computed from the theoretical F-distribution.

Hierarchical Clustering

This check box selects whether to perform hierarchical clustering on the elements in each cluster created

In addition to the standard viewers, this module outputs gene-specific statistics under the “F-Ratio Information” tab, as shown below. These tables can be saved as tab-delimited text files by right-clicking on them. All the columns in the tables can be sorted in ascending or descending order. Clicking on a column header re-orders the rows in ascending order of the values in the selected column. Holding down the SHIFT key while clicking on the column header will re-order the rows in the descending order of that column. Holding down the CTRL key while clicking anywhere on the header will restore the original ordering.

TIGR Multiple Array Viewer

File Adjust Data Normalization Distance Analysis Display Sort Help

HCL ST SOTA RN KMC KMS CAST QTC GSH SOM FOM PTM SAM ANOVA TTEST SVM PCA

MultipleExperimentViewer

- Main View
- Analysis
 - One-way ANOVA
 - Expression Images
 - Centroid Graphs
 - Expression Graphs
 - Cluster Information
 - F-Ratio information
 - Significant Genes
 - Non-significant Genes
 - General Information
- History

Unique ID	Spot Name	Spot Weight	Group1 me...	Group1 std...	Group2 me...	Group2 std...	Group3 me...	Group3 std...	
A0	A0	1	0.0	0.0	0.0	0.0	0.0	0.0	N
B0	B0	1	2.0	0.0	2.0	0.0	2.0	0.0	N
C0	C0	1	-2.0	0.0	-2.0	0.0	-2.0	0.0	N
D0	D0	1	2.5	0.5	2.0	0.5	1.75	0.6454972	1
E0	E0	1	1.5	0.5	2.0	0.5	2.25	0.6454972	1
F0	F0	1	-2.5	0.5	-2.0	0.5	-1.75	0.6454972	1
G0	G0	1	-1.5	0.5	-2.0	0.5	-2.25	0.6454972	1
A1	A1	1	-2.0	0.0	0.0	1.0	2.0	0.0	1
A2	A2	1	2.0	0.0	0.0	1.0	-2.0	0.0	6
A3	A3	1	0.06422493	0.29820946	0.36668697	0.0762278...	-0.0035144...	0.34363487	0
A4	A4	1	0.0128793...	0.3386631	-0.3187289	0.0921475...	0.26822692	0.14683537	0
A5	A5	1	-0.07092709	0.26621515	-0.19354802	0.25730923	-0.16372713	0.3987915	0
A6	A6	1	0.0856127...	0.39541218	0.05749334	0.21724758	-0.12817432	0.20692855	0
A7	A7	1	0.24915238	0.18717049	-0.0435245...	0.4768001	0.1118623	0.07317513	0
A8	A8	1	0.11156401	0.31109852	-0.0582168	0.32384735	-0.1247288...	0.0339455...	0
A9	A9	1	-0.11431832	0.2206299	0.13763689	0.27976358	-0.0116957...	0.39679977	6
A10	A10	1	0.0206784...	0.1190003	0.11319309	0.1987352	0.0248037...	0.29638064	1
A11	A11	1	0.21980788	0.39314786	0.26082134	0.1120318...	-0.3627567	0.1986139	0
A12	A12	1	0.14488335	0.2905416	0.27184868	0.15444125	-0.14973351	0.42260122	0
A13	A13	1	0.1198895...	0.1275276	0.0605114...	0.35201153	-0.11698475	0.29195407	1
A14	A14	1	-0.0569378...	0.24864803	-0.10332832	0.33003187	-0.19109274	0.23106217	0
A15	A15	1	-0.0394280...	0.35220957	-0.23487449	0.2881723	0.22049832	0.27518356	0
A16	A16	1	-0.0298313...	0.12368584	-0.06515752	0.33570597	0.03818571	0.32704136	1
A17	A17	1	-0.11072978	0.42636153	0.0510278...	0.1019578...	0.03710229	0.3177015	0
A18	A18	1	-0.32198283	0.0266310...	-0.0030479...	0.3699321	-0.12320718	0.2973118	1
A19	A19	1	-0.22161625	0.12136618	-0.1080355	0.28815243	-0.0273934...	0.41843417	3
A20	A20	1	-0.12653272	0.4331264	0.2179629	0.1242599...	-0.13588768	0.27482247	1
A21	A21	1	0.2264994	0.1551286	0.33907738	0.01729145	-0.14871097	0.3682818	0
A22	A22	1	-0.24507831	0.14308676	0.23679535	0.22913224	0.0385911...	0.43945697	2
A23	A23	1	-0.09871387	0.49816275	0.06311054	0.25755474	-0.0733670...	0.383684	1
A24	A24	1	0.0156969...	0.22506227	-0.23113425	0.27522767	0.2801697	0.3184198	0
A25	A25	1	-0.26904163	0.09083017	-0.31129918	0.11583544	0.0285251...	0.43259102	0
A26	A26	1	-0.0241408...	0.47091904	0.34565696	0.14735182	0.11396349	0.35808244	2
A27	A27	1	0.0281856...	0.12657094	0.0229531...	0.09858432	0.07923959	0.4439097	0
A28	A28	1	-0.23826828	0.14737579	0.0205084...	0.09012849	0.08178065	0.27997315	0

TIGR MultiExperiment Viewer

11.15.2. One-way ANOVA: F-Ratio Information viewer

11.16. TFA: Two-factor ANOVA

(Keppel and Zedeck 1989, pp 183-196, 536-541; Manly 1997, pp 125-131; Zar 1999, pp 248-250.)

Two-factor ANOVA can be used to find genes that vary significantly across levels of two independent variables (factors), as well as their interaction. The first initialization dialog prompts for the names and number of levels of the factors, following which an initialization dialog very similar to that for t-tests and one-way ANOVA is displayed. The only difference is that the top panel of this dialog contains two sub-panels instead of one for group assignments:

Two-factor ANOVA Initialization

Group Assignments

Factor A assignments

Ex1 Group 1 Group 2 Not in groups

Ex2 Group 1 Group 2 Not in groups

Ex3 Group 1 Group 2 Not in groups

Ex4 Group 1 Group 2 Not in groups

Ex5 Group 1 Group 2 Not in groups

Ex6 Group 1 Group 2 Not in groups

Ex7 Group 1 Group 2 Not in groups

Ex8 Group 1 Group 2 Not in groups

Ex9 Group 1 Group 2 Not in groups

Ex10 Group 1 Group 2 Not in groups

Factor B assignments

Ex1 Group 1 Group 2 Not in groups

Ex2 Group 1 Group 2 Not in groups

Ex3 Group 1 Group 2 Not in groups

Ex4 Group 1 Group 2 Not in groups

Ex5 Group 1 Group 2 Not in groups

Ex6 Group 1 Group 2 Not in groups

Ex7 Group 1 Group 2 Not in groups

Ex8 Group 1 Group 2 Not in groups

Ex9 Group 1 Group 2 Not in groups

Ex10 Group 1 Group 2 Not in groups

Save settings Load settings Reset

P-Value Parameters

p-values based on t-distribution

p-values based on permutation: Enter number of permutations 1000

Enter critical p-value 0.01

Alpha Corrections

just alpha (no correction) standard Bonferroni correction adjusted Bonferroni correction

Step-down Westfall and Young methods (for permutations only): minP maxT

Hierarchical Clustering

Construct Hierarchical Trees

TIGR * MultiExperiment Viewer

Reset Cancel OK

11.16.1. TFA initialization dialog.

The p-value parameters and alpha corrections (not yet implemented) are similar in function to the corresponding features in t-tests and one-way ANOVA. The cluster views in the output are similar to those of most other modules. Table viewers display the annotation, F-values and p-values of genes. Two or three p-values are generated for each gene: one each for the effects of the two factors, and an interaction p-value if relevant (see below). A significant gene cluster is

generated for each significant effect. F-values and p-values are saved when clusters are saved as text files from the right-click menu on any viewer.

A few points should be kept in mind while running this analysis:

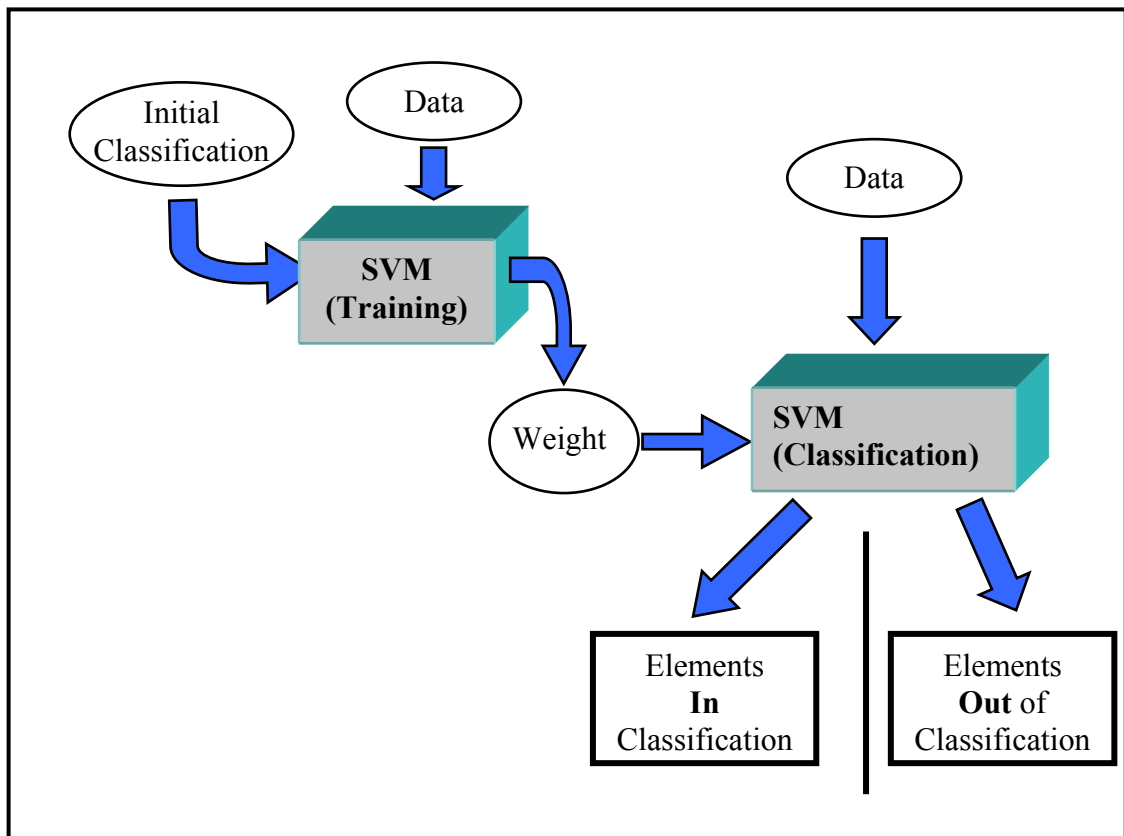
- Optimally, there should be equal numbers of samples in each cell (i.e, for each factorA - factorB combination). If samples sizes in cells are unbalanced, F-tests are biased, the degree of bias depending on the amount of imbalance. In such a case, the F-tests might be evaluated at a more stringent critical p-value than the one originally intended.
- Unbalanced designs, as described above, can occur in two ways: (1) by initially specifying the factor assignments in such a way that they are unbalanced, or (2) due to missing values for a gene, so that even if the original assignments are balanced, some cells have missing values for the gene.
- F-tests using the F-distribution (as opposed to using permutations) are quite fast.
- However, missing values in the expression matrix will greatly slow down permutation tests. The reason for this is, if a gene has missing values, it has to be permuted individually. In the permutations, values are randomly reassigned to cells, making sure that the missing values remain in their original cell. As each gene has to be permuted one a case-by-case basis, the total number of permutations will be (number of genes with missing values)*(number of permutations).
- On the other hand, for those genes that have complete data, the columns of the expression matrix are permuted, and all of the permuted F-values for those genes are computed at one go in a given permutation. Thus, the computation time for permutation tests is orders of magnitude less for a complete matrix than for one with significant numbers of missing values.
- Thus, the ideal data set for this kind of analysis would be one with balanced factor assignments, and no missing values (if you want to do permutation tests).
- Designs with just one sample in each factor A-B combination (cell) are also handled; however, in this case, only the A and B factor main effects are tested. Interaction is not tested in this case.
- Unbalanced designs where one or more cells have only one sample, or no samples, are not tested for any effects.

11.17. SVM: Support Vector Machines

(Brown *et al.*, 2000)

Although SVMs have been used in various fields of study, the use of SVMs for gene expression analysis was described in detail by Brown *et al.* SVM is a supervised learning classification technique. The algorithm uses supplied information about existing relationships between members of a subset of the elements to be classified. The supplied information, an initial presumed relationship between a set of elements, coupled with the expression pattern data leads to a binary classification of each element. Each element is considered either in or out of the initial presumptive classification.

The algorithm proceeds through two main phases. The first of these phases, **training**, uses the presumptive classification (supplied knowledge) and the expression data as inputs to produce a set of weights which will be used during the next phase. The second phase, **classification**, uses the weights created during training and the expression data to assign a discriminant or score to each element. Based on this score each element is placed into or out of the class. (Fig. 11.17.1)



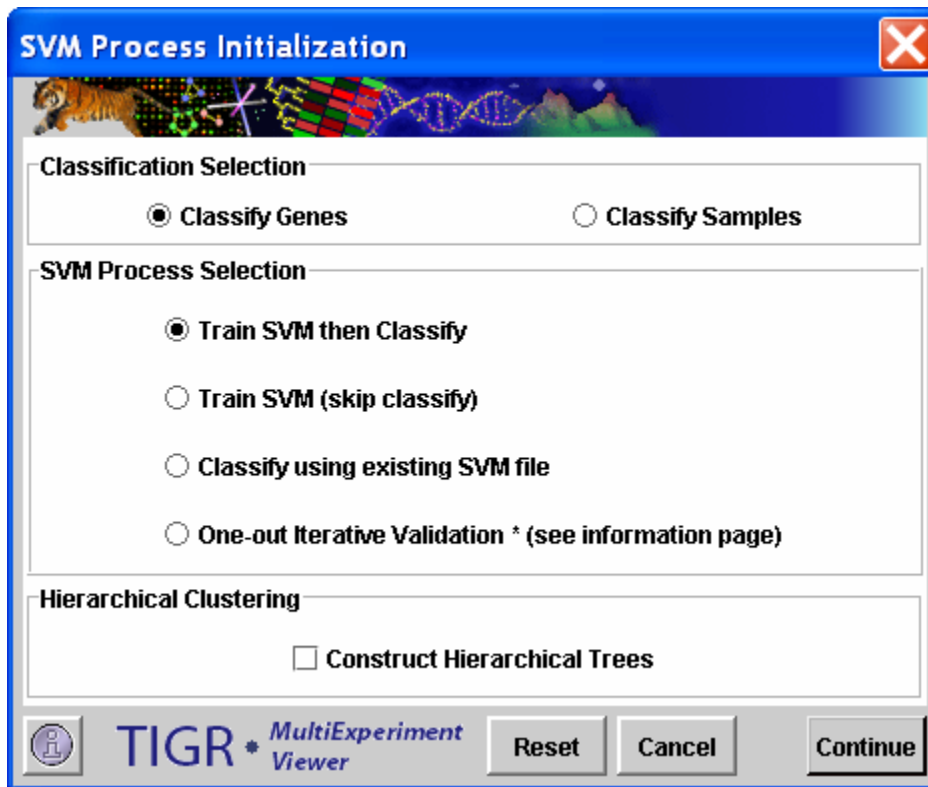
11.17.1. SVM Process Overview

SVM Dialog Overview

The initial dialog (Fig. 11.17.2) is used to define the basic SVM mode. One can select to classify genes or experiments and can select to perform one or both phases of the algorithm. The Train and Classify option allows one to run both

phases of the algorithm. Starting with a presumptive classification and expression data the result is a final classification of each element. The Train only option produces a list of weights which can be stored as an 'SVM' file along with training parameters so that they can be applied to data to classify at a later time. The Classify only option prompts the user for an SVM file of weights and parameters and results in final classification. The user also has an option to produce hierarchical trees on the two resulting sets of elements.

The second dialog (Fig. 11.17.3) is used during either the Train and Classify mode or the Train Only mode. The upper portion is used to indicate whether the initial presumptive classification will be defined using the SVM Classification Editor or supplied as an SVC file.



11.17.2. SVM Process Selection Dialog

Process Initialization Parameter Information

Sample Selection

The sample selection option indicates whether to cluster genes or experiments.

Process Selection

The SVM algorithm works by performing two main processes, training and classification. One can elect to perform training only, classification only, or both phases of the SVM classification technique.

The *Training Only* option results in a set of numerical weights which can be stored as an SVM file and used for classification at a later time.

The *Classification Only* option takes a file input of weights generated from training and results in a binary classification of the elements.

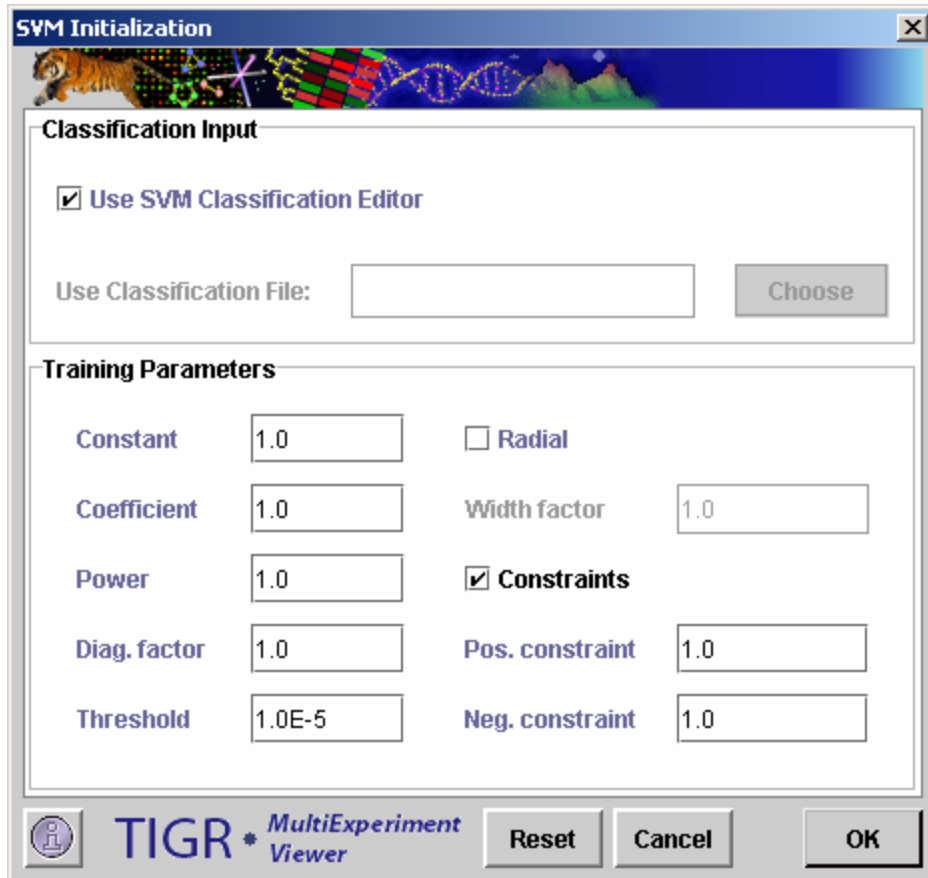
The *Training and Classification* option provides the ability to use the input set as a training set to produce weights which are immediately

applied to perform the classification.

The *One-out Iterative Validation* iteratively performs an SVM training and classification run. On each iteration one element is moved to the neutral classification and therefore will not impact the SVM training nor the classification of elements. The final classification will not be biased by an initial classification of the element.

Hierarchical Clustering

This check box selects whether to perform hierarchical clustering on the elements in each cluster created.



11.17.3. SVM Training Parameter Initialization Dialog

Classification Input

The SVM training process requires the supplied expression data and an additional initial presumptive classification which indicates which elements are initially presumed to have a relationship. Two options are provided for selecting members of the initial classification.

Use SVM Classification Editor

This option causes an editor application to be launched in order to allow a flexible tool for finding and marking elements to be positive members of the initial classification. This classification can be saved as an SVC file for later recovery of these initial settings.

Use Classification File

This allows the loading of an initial classification from an existing SVC file.

Kernel Matrix Construction

One can select to construct a polynomial or a radial kernel matrix.

Polynomial Kernel Function Parameters

The polynomial option is the default and three parameters are used to define the kernel construction.

Constant

An additive constant. (c)

Coefficient

A multiplicative constant. (w)

Power

A power factor. (p)

Polynomial Kernel Function

$$K(i,j) = [w*(Dist(i,j)+c)]^p$$

Radial Basis Function Parameters

The Radial Basis checkbox is used to select to use this type of Kernel generating function.

Width Factor

Radial width factor (w, see in below formula).

Radial Basis Kernel Function

$$K(x,y) = e^{-(\|x - y\|^2)/(2w^2)}$$

Training Parameters

Diagonal Factor

Constant added to the main diagonal of the kernel matrix. Adding this factor to the main diagonal of the kernel is required to force the matrix to be 'positive definite'. The definition of a positive definite matrix is best reviewed in books devoted to linear algebra but this state is achieved by selecting a constant of sufficient magnitude.

This positive definite state of the kernel matrix is required for the SVM algorithm to yield meaningful results. Testing values starting at 1.0 and increasing may be required to find an appropriate value. If the value is too low all elements will be partitioned in the negative class. For a range of values for this factor a stable set of elements may be classified as positive. At very high values there is a tendency to force all positive examples to be in the positive class regardless of their similarity of expression.

Threshold

This value is used as a stopping criteria for the weight optimization phase of training. Optimizing the weights produced during training is an iterative process which converges on an optimal set of weights to separate the positive and negative examples. This threshold dictates how stable the weights must be before the optimization process is terminated. Selection of a threshold that is very low could cause the optimization process to take an extremely long time and yet yield similar results to those where a higher threshold value was used which terminated the process earlier.

Constraints

This check box selects to apply limits to weights produced during training.

Positive Constraint

The upper limit to produced weights.

Negative Constraint

The lower limit to produced weights.

Distance Metric: Dot Product using normalized expression vectors so that the norm of each vector is 1. This metric is fixed for this algorithm and will not correspond to the distance menu.

The SVC file format is a tab delimited text file with the following columns for each element,

- 1.) Index - a sequential integer index.
- 2.) Classification - an integer value indicating class membership.
(1 = in initial classification, 0 = neutral, -1 = out of initial classification)
- 3.) Optional annotation columns.

The SVM Classification Editor (Fig. 11.17.4) allows one to use searches on supplied annotation as well as SVC files to assign membership to the initial presumptive classification. The editor allows the user to sort the list based on classification or annotation fields. The constructed initial classification can be stored in SVC format and later reloaded to allow alterations to produce what could be several initial classifications for a given study. The SVC files, once created, can be used to supply the initial classification thereby skipping the editor step. If the editor is used a button or menu selection launches the algorithm based on the current classification selection.

The screenshot shows the SVM Classification Editor window with a menu bar (File, Edit, Tools) and a toolbar containing icons for undo, redo, new, search, and run. Below the toolbar is a table with the following data:

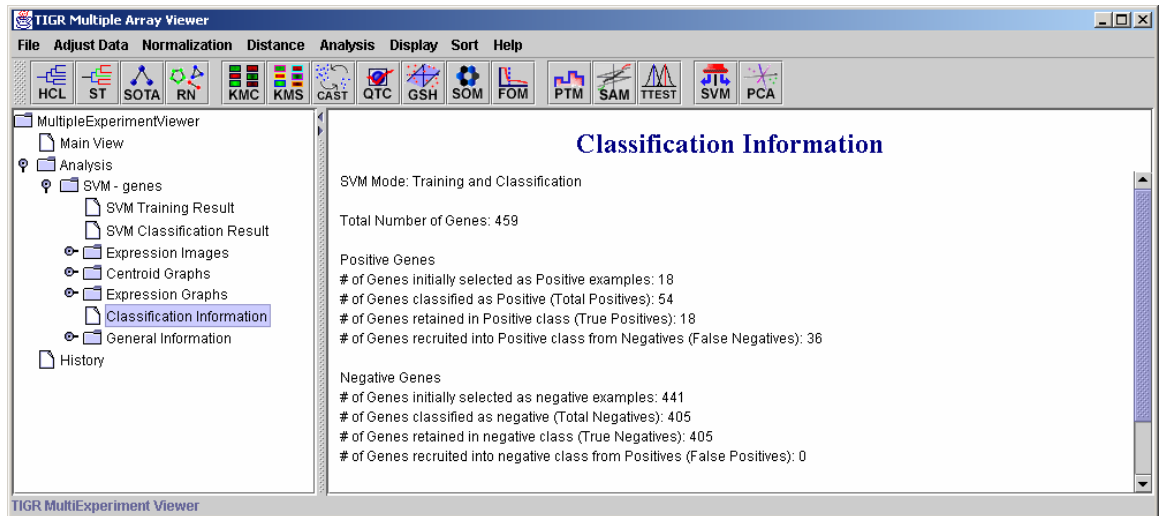
Index	In Class	Out of Class	Neutral	Unique ID	Spot Name	Spot Weight
0	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	A0	A0	1
1	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	B0	B0	1
2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	C0	C0	1
3	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	D0	D0	1
4	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	E0	E0	1
5	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	F0	F0	1
6	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	G0	G0	1
7	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	H0	H0	1
8	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	J0	J0	1
9	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	A1	A1	1
10	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	A2	A2	1
11	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	A3	A3	1
12	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	A4	A4	1
13	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	A5	A5	1
14	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	A6	A6	1
15	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	A7	A7	1
16	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	A8	A8	1
17	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	A9	A9	1

11.17.4. SVM Classification Editor

The second dialog also defines parameters used for creating the kernel matrix. The following is an overview of the training parameters.

SVM Output

The final result of an SVM run depends upon the process run. Training results in a set of weights that can be viewed along with the parameters for kernel construction. Note that from this viewer the training results can be saved as an SVM file. Classification results in a viewer that indicates each element's discriminant value and a final classification. The SVM Classification Information Viewer describes how many elements were initially selected as positive examples and how many elements were later recruited into the positive and negative classifications as well as other overview statistics.



11.17.5. Classification Information Viewer

Expression image viewers reveal which elements have been recruited into each of the final classification partitions by coloring the annotation red. Other result viewers are essentially the same as those describe in the K-Means clustering section.

11.18. KNNC: K-Nearest-Neighbor Classification (Theilhaber et al. 2002)

KNN Classification is a supervised classification scheme. A subset of the entire data set (called the training set), for which the user specifies class assignments, is used as input to classify the remaining members of the data set. The user specifies the number of expected classes, and the training set should contain examples of each class. If the **Classify** option is chosen from the initial dialog, an input dialog box is displayed for parameter input.

The screenshot shows a dialog box titled "KNN Classification" with a blue header and a close button (X) in the top right corner. The dialog is divided into several sections:

- Classify genes or samples:** Two radio buttons are present. "Classify genes" is selected (indicated by a filled circle), and "Classify samples" is unselected (indicated by an empty circle).
- Variance filter:** A checkbox labeled "Use variance filter (if unchecked, use all genes)" is unselected. Below it, a text input field contains "1000" with the label "Use only the following number of highest-variance genes:".
- Correlation filter:** A checkbox labeled "Use correlation filter" is unselected. Below it, two text input fields are present: the first contains "0.01" with the label "Cutoff p-value for correlation:", and the second contains "1000" with the label "Number of permutations for correlation test:".
- KNN classification parameters:** Two text input fields are present. The first is labeled "Number of classes" and contains "5". The second is labeled "Number of neighbors" and contains "3".
- Create / import training set:** Two radio buttons are present. "Create new training set from data" is selected (indicated by a filled circle), and "Use previously created training set from file" is unselected (indicated by an empty circle).
- Hierarchical Clustering:** A checkbox labeled "Construct Hierarchical Trees" is unselected.

At the bottom of the dialog, there is a footer area containing the TIGR MultiExperiment Viewer logo on the left, and three buttons: "Reset", "Cancel", and "Next >" on the right.

11.18.1. KNN Classification Initialization Dialog

KNN Classification Parameter Information

Classify genes or samples

This is self-explanatory. Although the following description refers to genes, the same steps will apply to experiments if “Classify samples” is chosen.

Variance filter

This is the first of two noise-reduction filters that can optionally be applied before classification. The variance filter keeps only those genes in the entire data set (including the classifier set) that have the highest variance across all samples. The number of genes to be retained is specified by the user.

Correlation filter

The correlation filter is used to filter out those genes of the set to be classified, that are not significantly correlated with at least one member of the training set. The significance of correlation is determined by the p-value, which is calculated by a permutation test in which each gene is permuted a user-specified number of times.

KNN classification parameters

This is where the user specifies the expected number of classes (which is also the number of classes present in the training set).

The number of neighbors is the number of genes from the training set that are chosen as neighbors to a given gene. Euclidean distance is used to determine the neighborhood. Let’s say we want to classify a gene *g*. Gene *g* is assigned to the class that is most frequently represented among its *k* nearest neighbors from the training set (where *k* is specified by the user). In case of a tie, gene *g* remains unassigned.

Create / import training set

If the user chooses to import a previously created training set (for instructions on saving a training set, see below), on hitting the “Next” button a file chooser is displayed from which the training file can be chosen. If an appropriate file is chosen, the KNN classification editor shown below in Fig xx is displayed with the class assignments from the file. If the option to create a new training set from data is chosen, on hitting the “Next” button the classification editor is directly displayed with all genes set to neutral.

Hierarchical Clustering

This checkbox selects whether to perform hierarchical clustering on the elements in each cluster created.

Default Distance Metric: Euclidean (for finding nearest neighbors), Pearson (for Correlation test). Fixed, will not correspond to distance menu.

KNN Classification Editor

Edit Tools

Index	Class 1	Class 2	Class 3	Class 4	Class 5	Neutral
0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
3	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
11	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
12	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
13	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
14	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
15	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
16	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
17	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
18	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
19	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
20	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
21	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
22	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
23	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
24	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
25	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
26	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
27	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
28	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
29	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
30	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
31	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
32	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
33	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
34	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
35	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
36	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
37	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
38	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Save classification to file
 Do not save classification to file

Next >

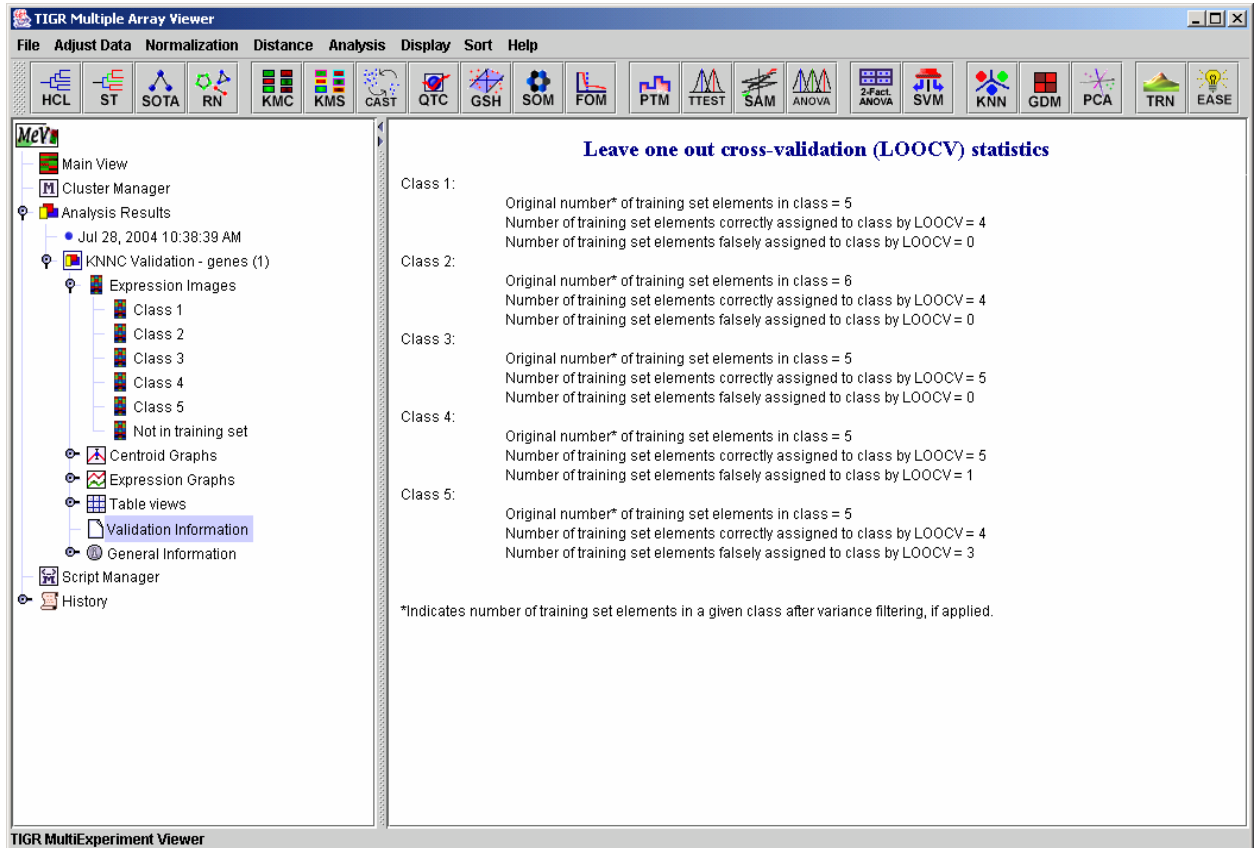
11.18.2. KNN Classification editor

Genes can be assigned to any of the specified classes by checking the box in the appropriate column for that gene. Genes designated as “Neutral” will be classified in subsequent steps, whereas those assigned to classes in this step will be treated as the training set. The “Edit” and “Tools” menus at the top of the classification editor allow searching, sorting and selection of data. The classification scheme

can be saved in a text file if needed, to be loaded in a future KNNC run (as explained above).

On hitting the “Next” button in the classification editor, the calculations proceed, and the output has the usual viewers as most other MeV runs. The main difference is that each type of viewer has several sub-viewers for Used Classifiers, Unused Classifiers (those that are possibly weeded out by variance filtering), Classified, Used Classifiers + Classified, Unclassified and All. This is done so that users can visually compare training vectors to trained vectors.

The **Validate** option in KNNC provides dialogs very much like the above, and is used to perform leave-one-out cross validation on the training set. The output viewers are also similar to the ones obtained in classification, with an additional viewer that provides cross-validation statistics.



10.18.3 KNN Classification cross-validation statistics viewer

11.19. DAM: Discriminant Analysis Module

(Danh V. Nguyen and David M. Rocke, 2002)

DAM is a method for classification of genes/experiments into more than two groups or classes. DAM incorporates a gene dimensional reduction method, Multivariate Partial Least Squares (MPLS), and two classification methods, Polychotomous Discriminant Analysis (PDA) and Quadratic Discriminant Analysis (QDA). Either PDA or QDA can be performed after MPLS depending on the user's selection. A three-dimensional plot of the most significant gene components is generated and displayed in the DAM results viewer. The results viewer also contains Expressions Images, Expression Graphs, Centroid Graphs and Cluster Information for experiments that are used as classifiers and for experiments that are to be classified. DAM can be launched from the Multiple Array Viewer toolbar by selecting the DAM button or by using the DAM menu option in the analysis menu.

DAM Initialization

Classification Selection

Classify Genes Classify Experiments

Data Screening

Enable Data Screening Step (ANOVA)

Alpha Value

Classification Algorithm Selection

PDA QDA

DAM Classification Parameters

Number of Classes

Number Of Components

Validation Selection

Enable Validation

A0 A1 A2

TIGR * MultiExperiment Viewer

11.19.1. DAM Initialization Dialog

Parameter Selection

The *Classification Selection* panel provides two options for DAM: classify genes or classify experiments.

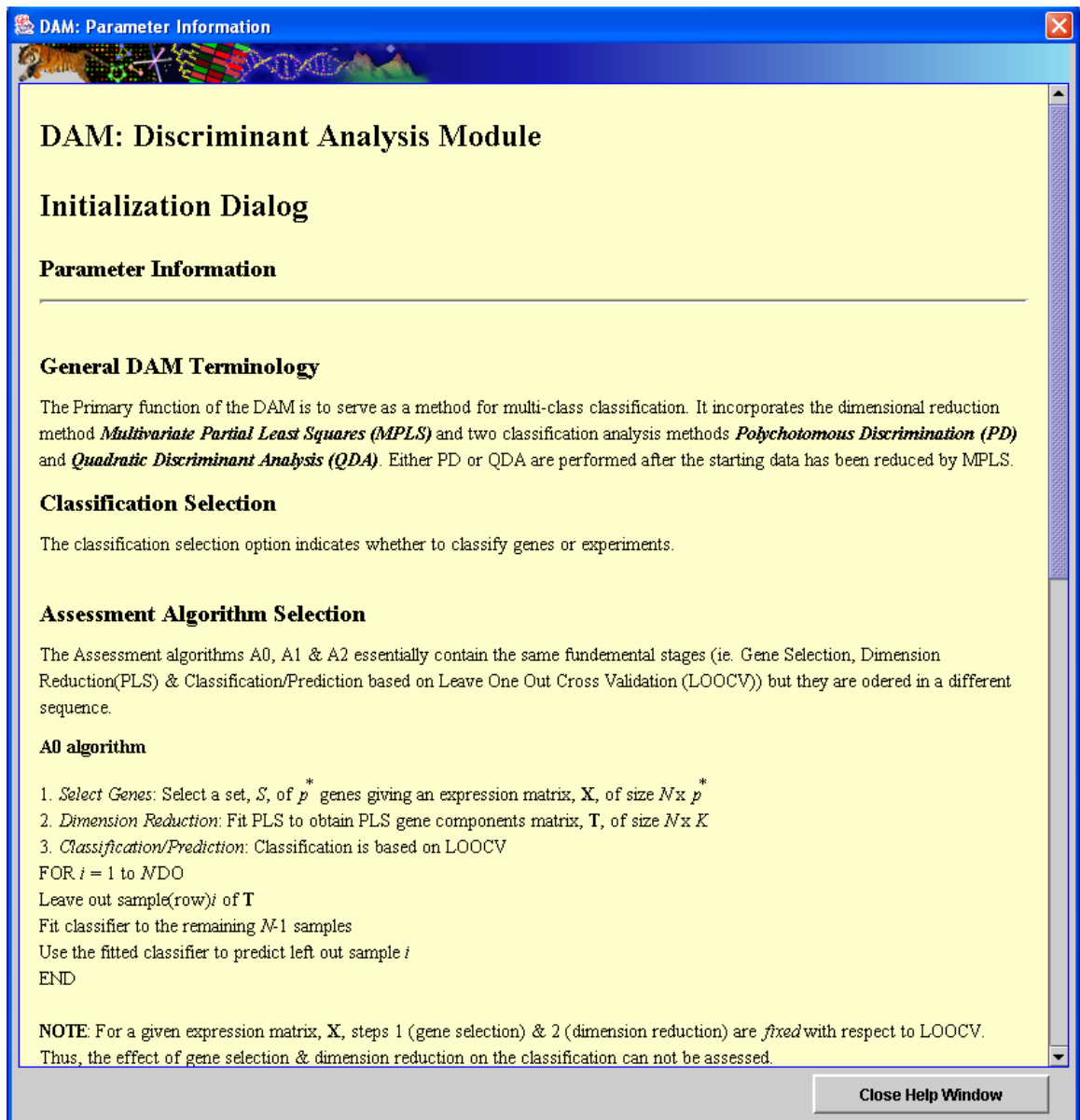
The *Data Screening* panel allows the users to select to apply a filter to select genes (or samples if classifying genes) from the loaded set that should be near optimal for partitioning the elements to classify based on Analysis of Variance (ANOVA) on permutations of the known (training) class members. The list of selected elements and number of selected elements will be reported after classification and the alpha value can be adjusted to apply a more or less stringent criterion and will impact the number of elements selected to enter the classification algorithm.

The *Classification Algorithm Selection* panel is for selecting either between PDA and QDA as the primary classification algorithm.

The *DAM Classification Parameters* panel has fields to enter the number of expected classes that should be found. The data will be partitioned into this many classes. The *number of components* indicates the number of representative expression vectors that should be generated from the data using MPLS. These components can roughly be described as components that represent major features of the data or correlate to variance found in the data. These components, once determined, are actually used for partitioning the data rather than the actual input expression vectors. Usually about 3 components is adequate to describe or cover most of the variance in the data. This step in the algorithm is described as the dimensional reduction step.

The *Validation Algorithm Selection* panel allows the user to select if validation should be performed and which assessment algorithm in A0, A1 and A2 are to be used in validation. The algorithms are described in the cited DAM reference and briefly in the information help page that is launched when from the information button in the lower left corner.

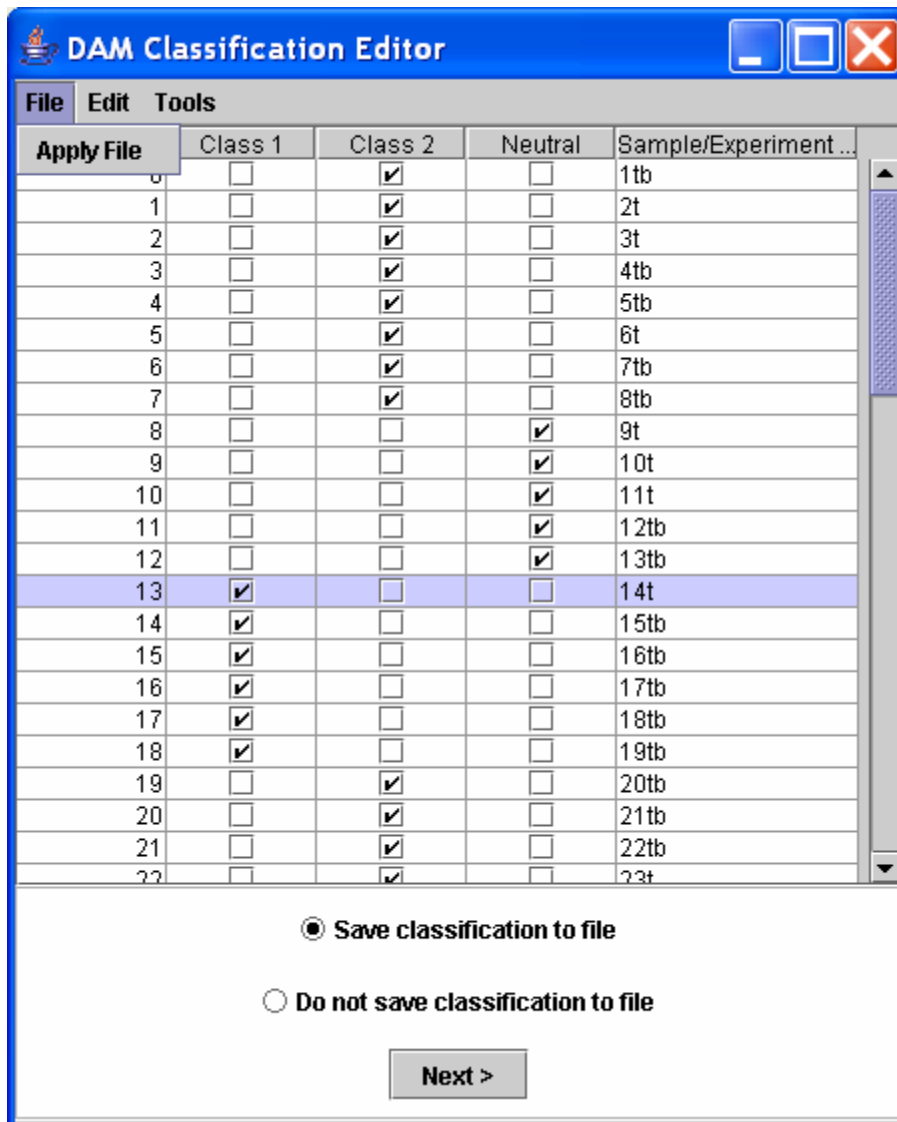
When the Information button at the bottom-left corner of the DAM Initialization Dialog is depressed, an Information Dialog screen is popped up. This Dialog contains brief description of the terminologies in DAM, and references page describing the methods and parameters used in DAM.



11.19.2. DAM Parameters Information Page

The DAM Classification Editor

When the *Next* button is depressed from the initialization dialog, a DAM Classification Editor screen shall pop up. This screen allows the user to identify which samples are known examples of a class and to which class they should be assigned. Samples that are left as *neutral* are assumed to be of unknown classification and will be partitioned based on the known members of the class using the selected algorithm and parameters. Note that the default is to save these selections to a file when exiting the editor. If the settings have been set previously and saved you can choose to load or apply those settings from the editor's File menu. This will apply the saved settings but you will still have the ability to alter class memberships before proceeding.

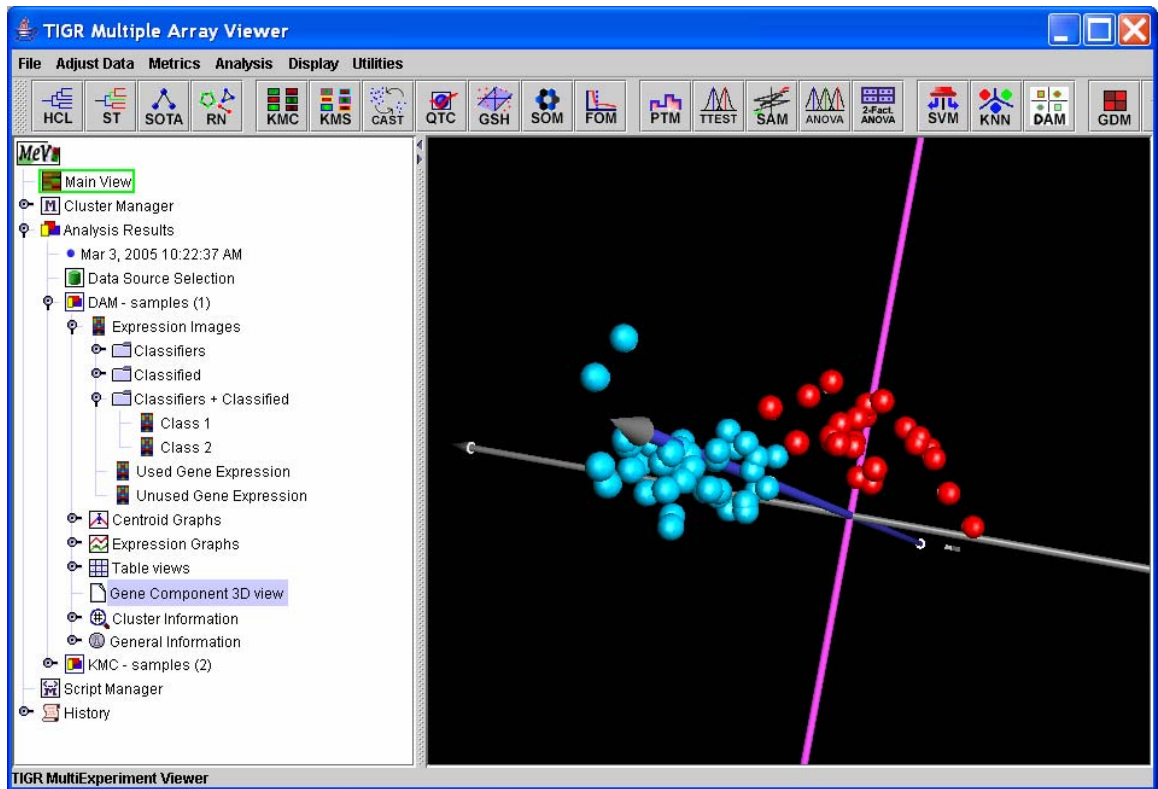


11.19.3. DAM Classification Editor

DAM Result Viewers

When the DAM module has run to completion, a sub-tree labeled “DAM” will be created and placed under the Analysis tab in the navigation tree. The tabs within the DAM sub-tree contain results of the module’s calculations. The results include Expression Images, Centroid Graphs, Expression Graphs, Gene Component 3D view, Cluster Information and General Information. Of these viewers, all have been described previously except the Gene Component 3D view.

The Gene Component 3D view is a three dimensional view representing the 3 most significant gene components obtained from Dimensional Reduction. The display can be rotated and shifted by left dragging or right dragging respectively. Right clicking on the 3D view **node** will display a popup menu that allows the user to change the 3D view’s display options and create a selection area to define a cluster. The points are projections of the elements being classified into 3D space using the first three expression components generated during MPLS determination of representative components.



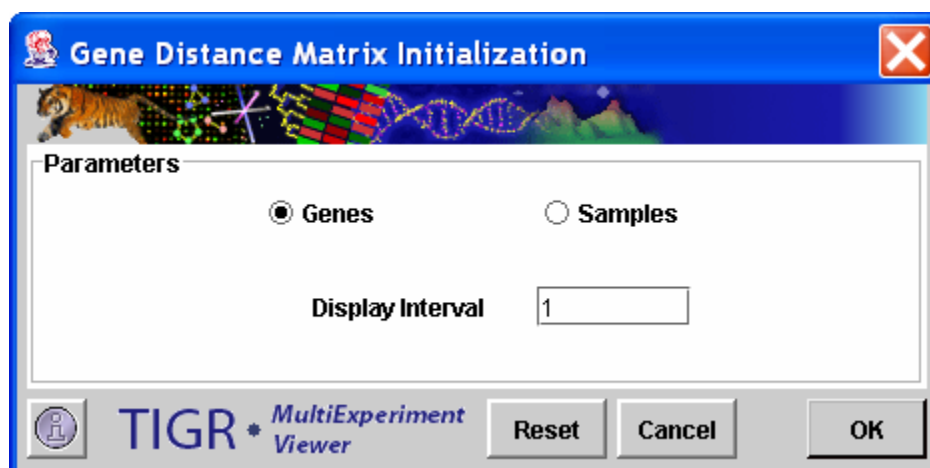
11.19.4. 3D Component Projection View

GDM: Gene Distance Matrix

Most of the clustering methods found in MeV form clusters by algorithms that group genes based on similarity of expression pattern. The distance, inverse of similarity, between two genes is calculated using a distance metric (see ‘Distance’ menu and manual section 13, the appendix on metrics). The GDM gives an intuitive and comprehensive view of the distance (or similarity) between any two genes loaded into MeV by creating a colored matrix representing all gene-to-gene distances. The GDM module is useful for taking a distance survey as well as discovering which genes are similar in expression pattern to a particular gene of interest. Like most of the MeV modules, the GDM module can also be used with experiments as input.

GDM Initialization

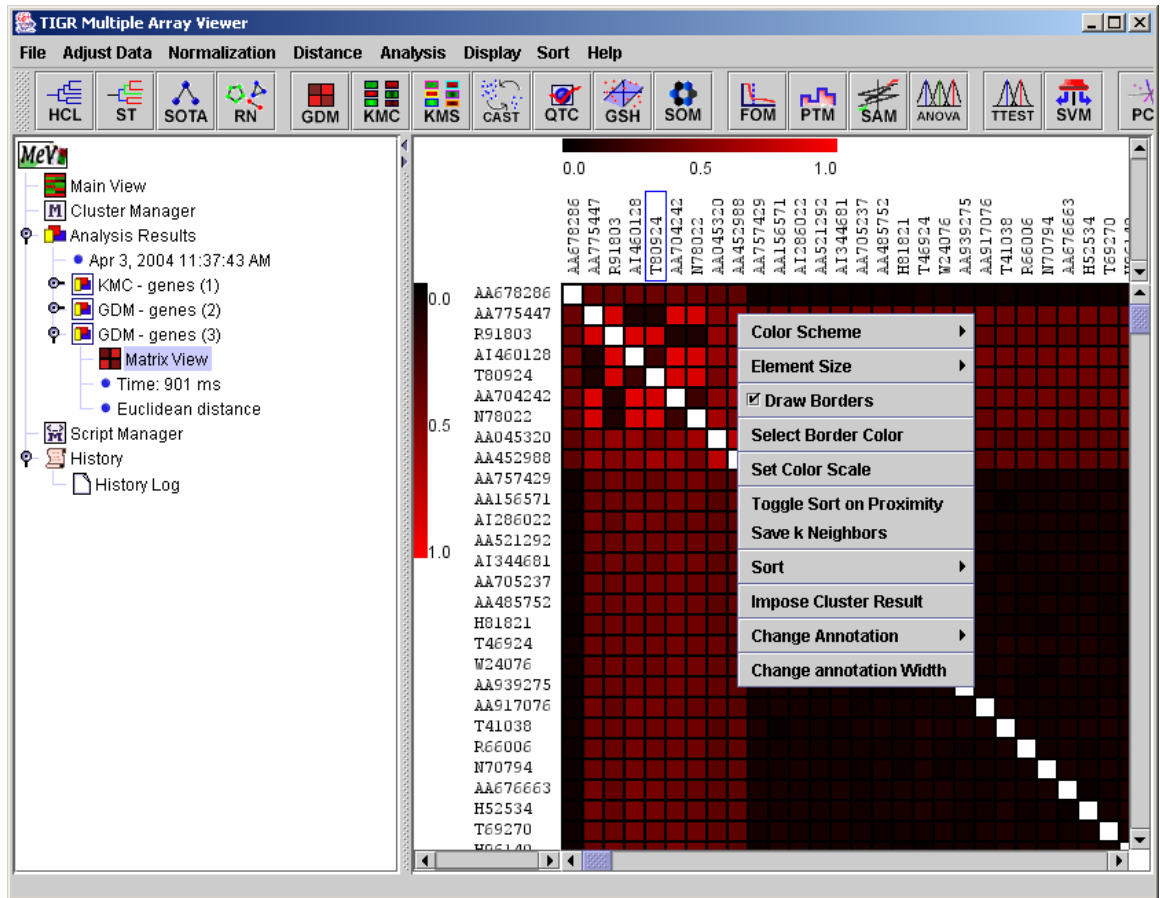
The GDM module can be started by using the GDM button or by selecting the GDM menu item from the analysis menu. When creating a gene matrix it is possible to display a subset of the full data set. The creation of an nxn matrix is expensive from a computer memory standpoint and by using the “Display Interval” option it is possible to make a smaller matrix and conserve memory.



11.19.5. GDM Initialization Dialog

Matrix Viewer Basics

The matrix viewer has annotation headers that can be used to identify the gene associated with a column or row. Each square element within the matrix is rendered as a color that represents the distance between the two genes associated with the element. The main diagonal is simply rendered as white for identification.



11.19.6. GDM showing borders and popup menu.

Menu Options

Like most viewers in MeV, the GDM has a right click menu that provides options for extracting information from the viewer and manipulating the appearance of the viewer. In the sections below each of the menu options are described.

Color Scheme

Two default color schemes are available as well as the option to select a custom color scheme.

Element Size

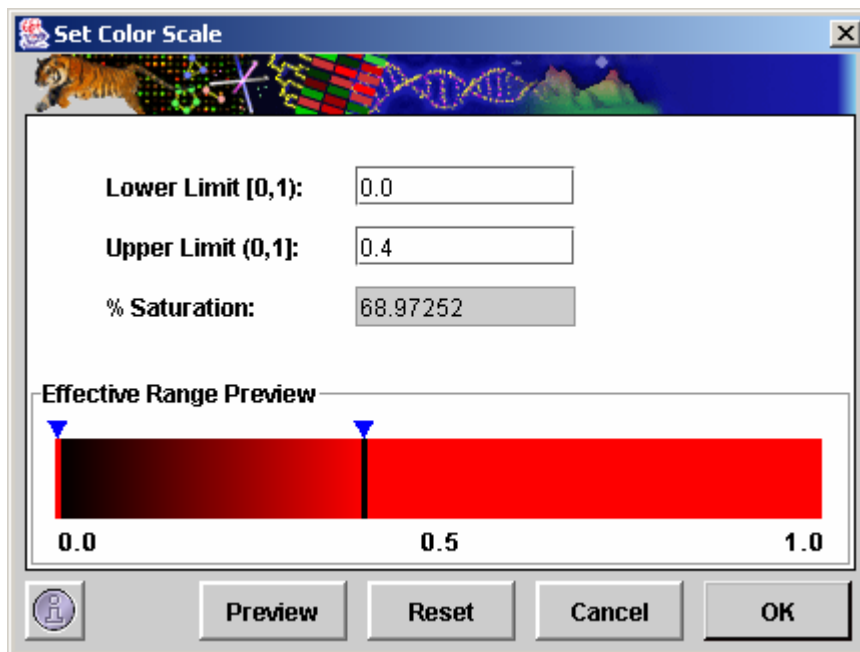
Five preset element sizes are offered as well as the option to select a custom size. Changing the size is a good way to get either a detailed look at specific genes or to take a broad survey of the matrix.

Draw Borders

The Draw Borders options places borders between elements and can help with visual alignment when viewing the distance of one gene to several other genes on a single row of the matrix. The color of the border can be selected from the menu to contrast the colors used to represent distance.

Set Color Scale

The GDM menu provides the option of selecting the limits of the displayed color scale. By altering the lower and upper limits of the scale, gene distances map to a different location on the color scale. The figure of the Color Scale Dialog illustrates this. The lower limit is at 0.0 while the upper limit is set to 0.4. Any element having a distance greater than 0.4 will appear *saturated* in color at the bright red end of the gradient. The effective range is 0.0 to 0.4 and this will accentuate distances in this low-end range. By imposing different limits it is possible to get better resolution (color differentiation) within a given range. When setting the upper or lower limit to values other than 0.0 and 1.0 respectively, there will always be some gene pairs that have distances that fall off of the upper or lower effective range. The percentage of elements in the matrix that are saturated is displayed as a guide to the percentage of elements that are off of the effective range. When altering color scale it is often useful to hit the *preview* button to view the effects on the actual matrix. *Reset* will return the values to the values that were in effect when the dialog was launched. *Cancel* will return the limits to the original values in effect when the dialog was launched and will dismiss the dialog. Note that the current limits are always displayed in the labels on the header color gradient.



11.19.7. Color Scale Dialog

Toggle Sort on Proximity

In the main matrix figure you may note that a gene in the column header is selected as denoted by the rectangle around the gene identifier. Moving the mouse cursor over the header enacts item selection. If you click on the label in the header then that element is moved to the left (or top if in row header) and the remaining elements are ordered by proximity to the selected element. Typically you should see the first row and column appear as a gradient since the neighbors are ordered by proximity. The sort menu options can be used to impose other

orderings. The Toggle Sort on Proximity menu item turns this capability on or off.

Save k Neighbors

The Save k Neighbors menu option is used in conjunction with proximity sort. Once sorted by proximity the Save k Neighbors options saves any number of the selected gene's (or sample's) nearest neighbors as displayed in the viewer.

Sort

The Sort menu provides methods to sort the genes or samples according to the order specified in the input file (default) or by a selected annotation key. This provides a useful method to be used with proximity sort. First one can order by annotation which enables easy selection of the gene of interest. Once found, clicking on the gene's label will shift it to the corner position with its nearest neighbors in order of proximity.

Impose Cluster Result

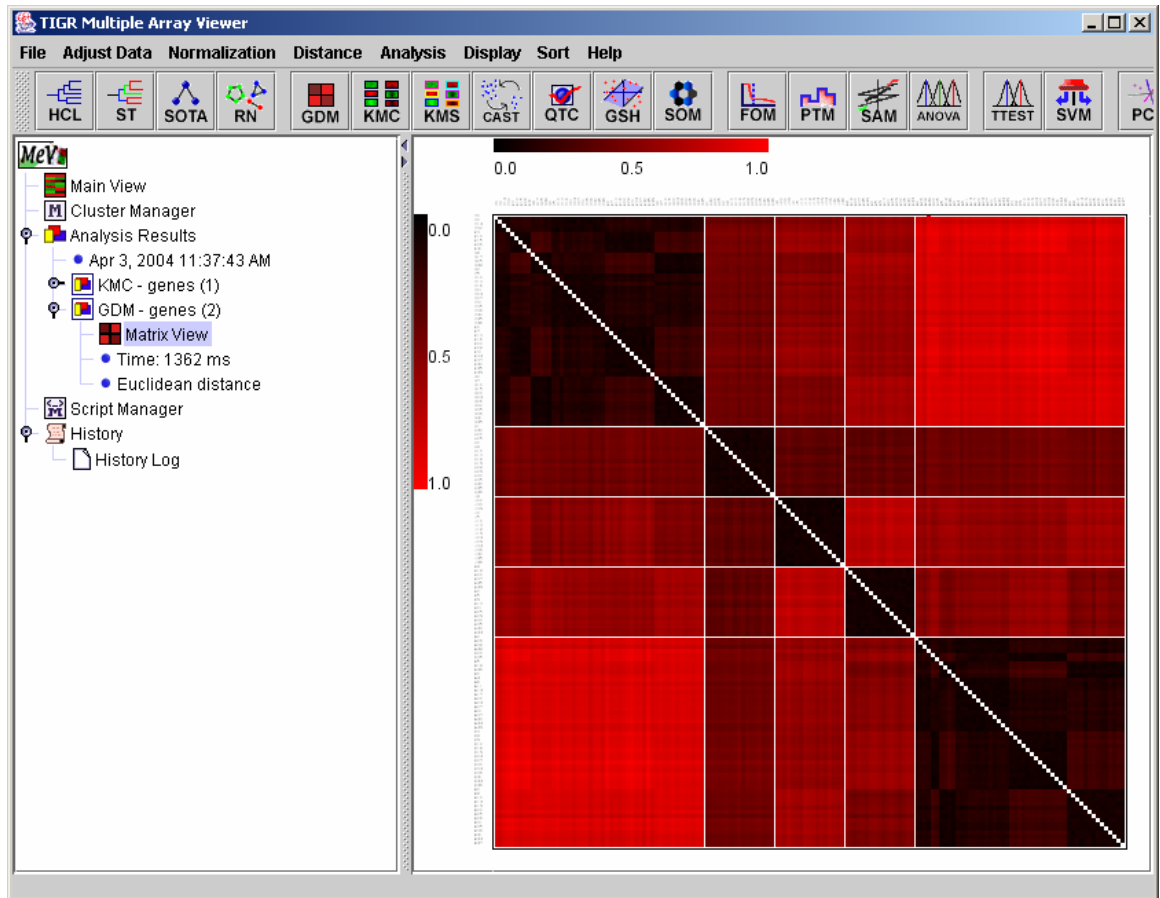
The Impose Cluster Result option is among the most useful features of the GDM. When selected the current MeV session is surveyed for appropriate clustering results to apply to the GDM. Application of the clustering result reorders the genes such that the rows and columns are grouped by cluster membership. Cluster boundaries are represented by a white border. Notice the figure displaying a distance matrix with a K-means clustering result imposed with five clusters. The elements within a cluster are similar as evidenced by the very dark squares on the main diagonal. Note that the element size was reduced in order to view the entire matrix and that every third gene was displayed.

Change Annotation

The change annotation feature allows the selection of an annotation type to be displayed in the headers.

Change Annotation Width

This option allows the expansion or contraction of the header to permit viewing of the header without excessive scrolling or if contracted more of the matrix will be visible.



11.19.8. GDM matrix with a KMC clustering result imposed ($k = 5$)

Additional GDM Features

If one clicks on a spot an information record is displayed describing several attributes of the element. The annotation identifying each element as well as the raw and scaled (0-1) distances and parameters related to taking the distance such as distance metric are reported. From this report page a graph can be displayed using the *Expression Graph* button. The graph overlays the expression graphs of the two genes.

Gene Distance Spot Information

Annotation

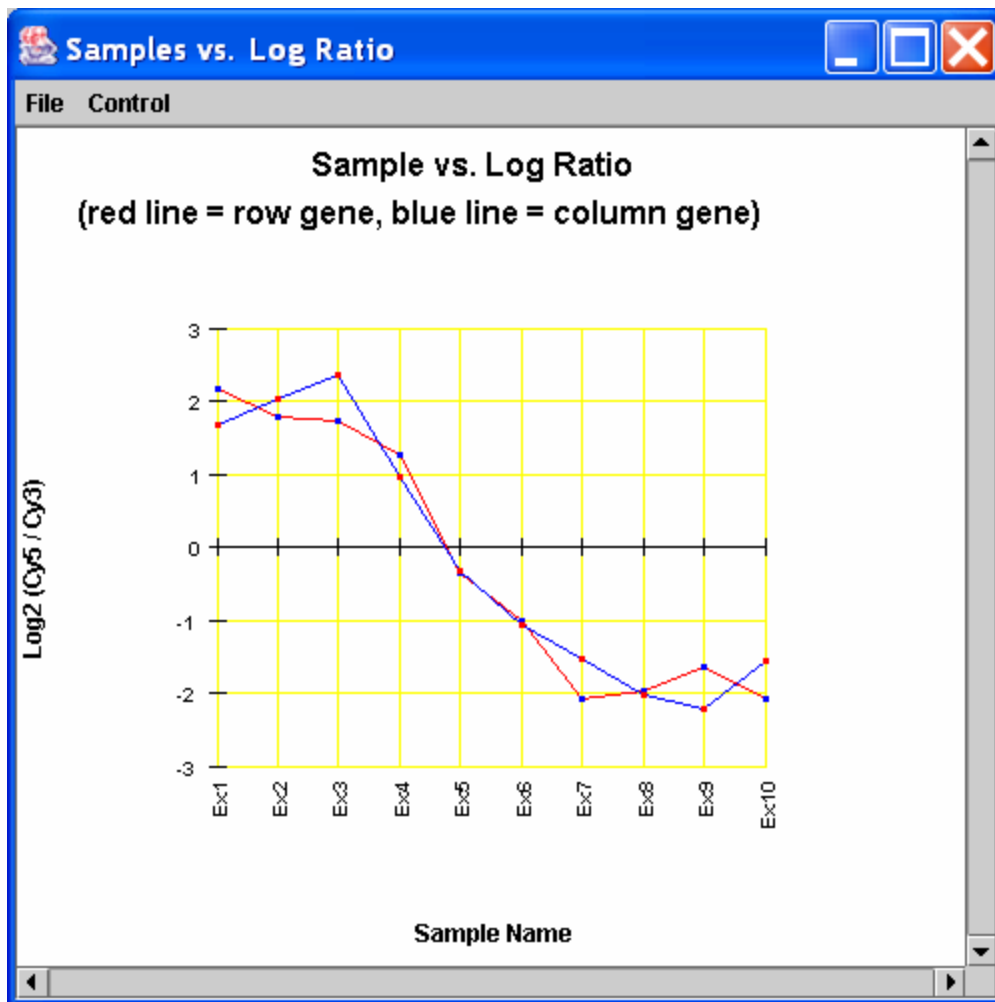
	Column Gene	Row Gene
Row	129	132
Column	1	1
YORF	H29	H37
NAME	H29	H37
GWEIGHT	1	1
GenBank	AA488715	AA448711

Distance Information

GDM Matrix Row	132
GDM Matrix Column	129
Scaled Gene Distance	0.06876421
Actual Gene Distance	1.0257703
Distance Metric	Euclidean distance
Vector Size	10
Missing Values	0, 0
Distance Based on	10

Expression Graph Close

11.19.9. GDM Element Information Record

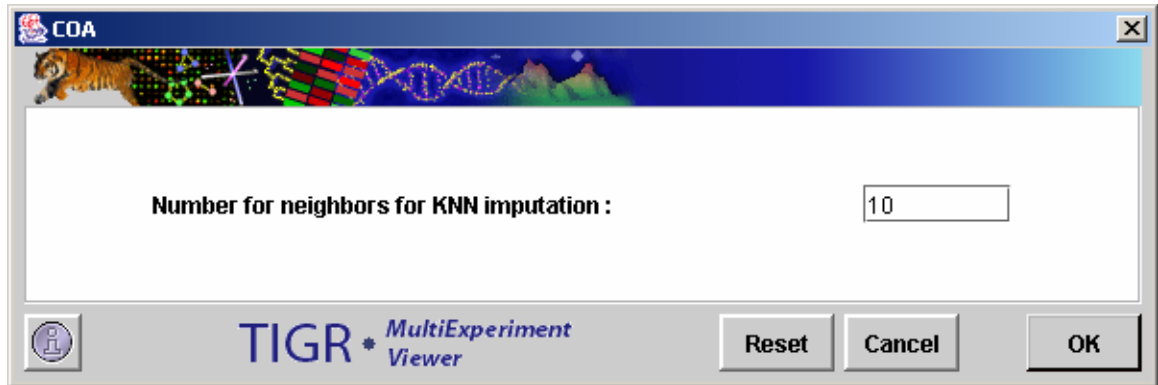


11.19.10. GDM Element Information Expression Graph

11.20. COA: Correspondence Analysis (Fellenberg *et al.* 2001, Culhane *et al.* 2002)

Correspondence analysis is an explorative method to study associations between variables. Like principal components, it displays a low-dimensional projection of the data. However, in this case, both genes and samples can be projected onto the same space, revealing associations between them.

Correspondence analysis requires an expression matrix with no missing values. Therefore, any missing values have to be imputed first. We use the k-nearest neighbors algorithm to impute missing values. The only user input in the initialization dialog is the desired number of neighbors for imputation.

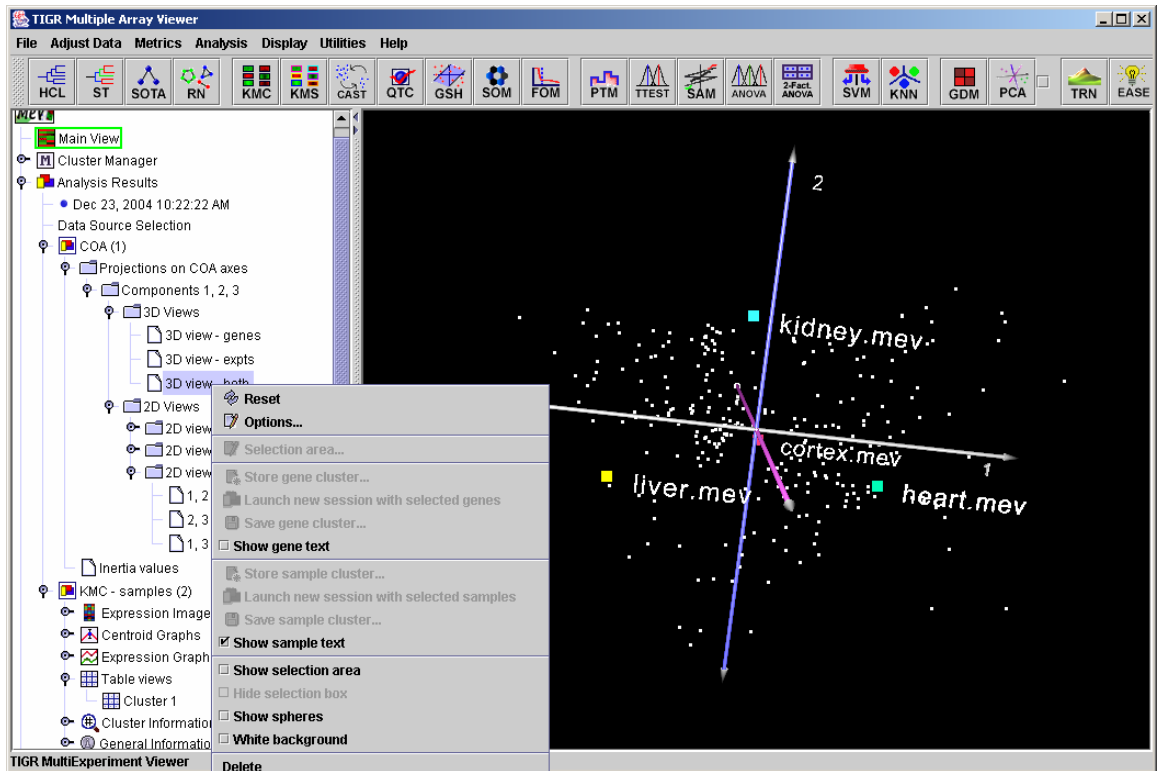


11.20.1. COA initialization Dialog

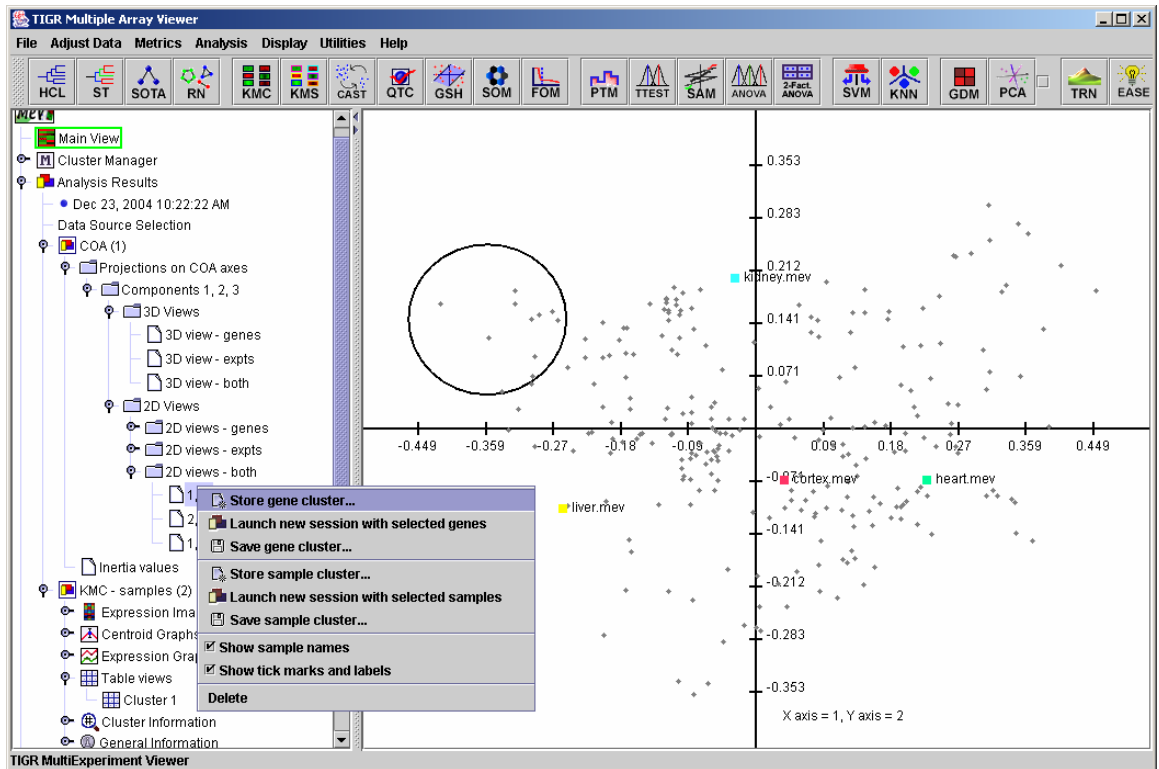
The displays in this module are very similar to the PCA displays, except that 2D and 3D plots are shown for genes and experiments separately as well as together on the same plot. Menus for creating new plots, selecting data points and customizing displays are available from the corresponding nodes on the navigation tree just as in PCA.

Genes that lie close to one another on the plot tend to have similar profiles, regardless of their absolute value. The same is true for samples. If some genes and samples lie close to one another on the plot, then these genes are likely to have a high expression in the nearby samples relative to other samples that are far away on the plot. On the other hand, if a set of genes are on the opposite side of the plot from a set of samples relative to the origin, then the expression of that set of genes is likely to be depressed in those samples relative to samples that might be positioned close to those genes. The farther the points are from the origin, the stronger the association between genes and samples.

Correspondence analysis works by decomposing a matrix of chi-squared values derived from the rows and columns of the expression matrix. The first two or three axes are the most informative in showing associations among genes and experiments. The amount of information explained by a given axis is quantified by its inertia, which may be thought of as the proportion of the total chi-squared value of the matrix explained by that axis. The inertia values are provided under the corresponding node under the main COA analysis node.



11.20.2. COA: 3D display



11.20.3. COA: 2D display

11.21. PCA: Principal Components Analysis

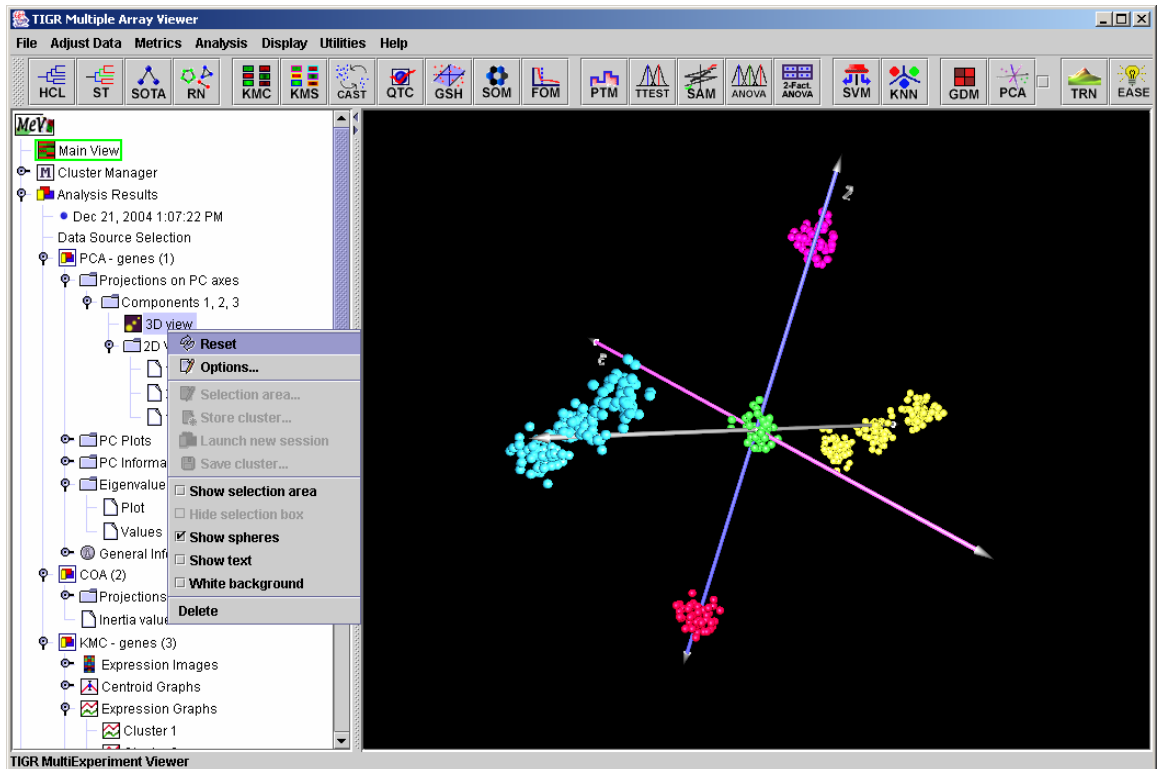
(Raychaudhuri *et al.* 2000)

PCA is used to attribute the overall variability in the data to a reduced set of variables termed principal components. To each principal component a certain fraction of the overall variability of the data is attributed such that each successive component determined accounts for less of the variability than the previous one. This ranks the components in order of decreasing determination of data variability. The first three principal components are used to map each element into a three dimensional viewer.

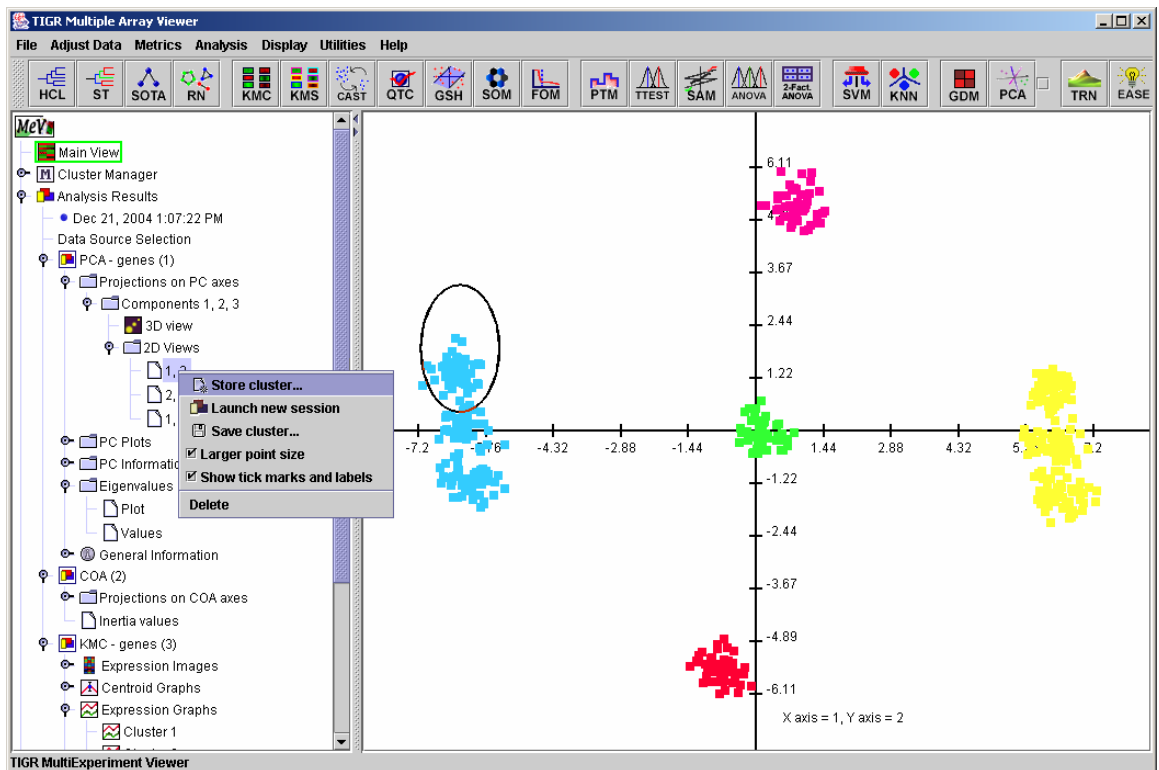
Once the calculations are complete, select the PCA node under Analysis to view the PCA results. Under the node called “Projections on PC Axes” are the default plotting of components 1, 2 and 3. Right-clicking on this node will allow other components to be chosen for plotting. These new plots will show up as new nodes under this node.

3D view is one of the primary PCA displays, and is a three dimensional view. The display can be rotated and shifted by left dragging or right dragging respectively. Right clicking on the 3D view **node** will display a popup menu that allows the user to change the 3D view’s display options and create a selection area (essentially a cube) to define a cluster. The 2D views will display plots of any two components at a time. Dragging the mouse over the 2D view will create a selection area, which can be used to define a cluster. Cluster options and other features are available by right-clicking on the 2D view node on the navigation tree I the left pane.

PC plots, PC information and Eigenvalues detail the calculations behind the construction of the display. Often some meaning such as overall expression level, expression trends, or some other aspect of the data set can be found to correlate to the principal components. Using the PC plots, and noting where clusters of elements showing various trends labeled in other algorithms fall in the 3D viewer can help to assign some tentative meaning to each component. Note that interpretation of the components is not exact and is somewhat subjective.



11.21.1. PCA: 3D View.



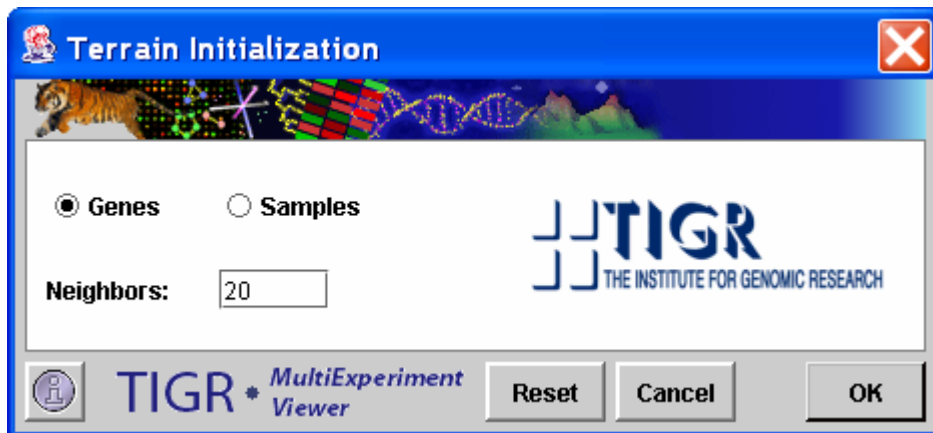
11.21.2. PCA: 2D view

11.22. TRN: Expression Terrain Maps

(Kim *et al.* 2001)

Terrain maps provide a three dimensional overview of the major clusters inherent in the data. Terrain maps can represent gene or experiment groupings depending on the mode selected in the input dialog. The input elements are first mapped into a two dimensional grid in which the placement of each element is influenced by a user selected number of nearest neighbors. Once the two dimensional layout is finished the third dimension is determined by the density of points over discrete areas of the 2d grid. This value is projected as a surface in the third dimension. The higher peaks indicate large numbers of very similar elements while the lower peaks are composed of fewer elements that tend to be not so similar in expression.

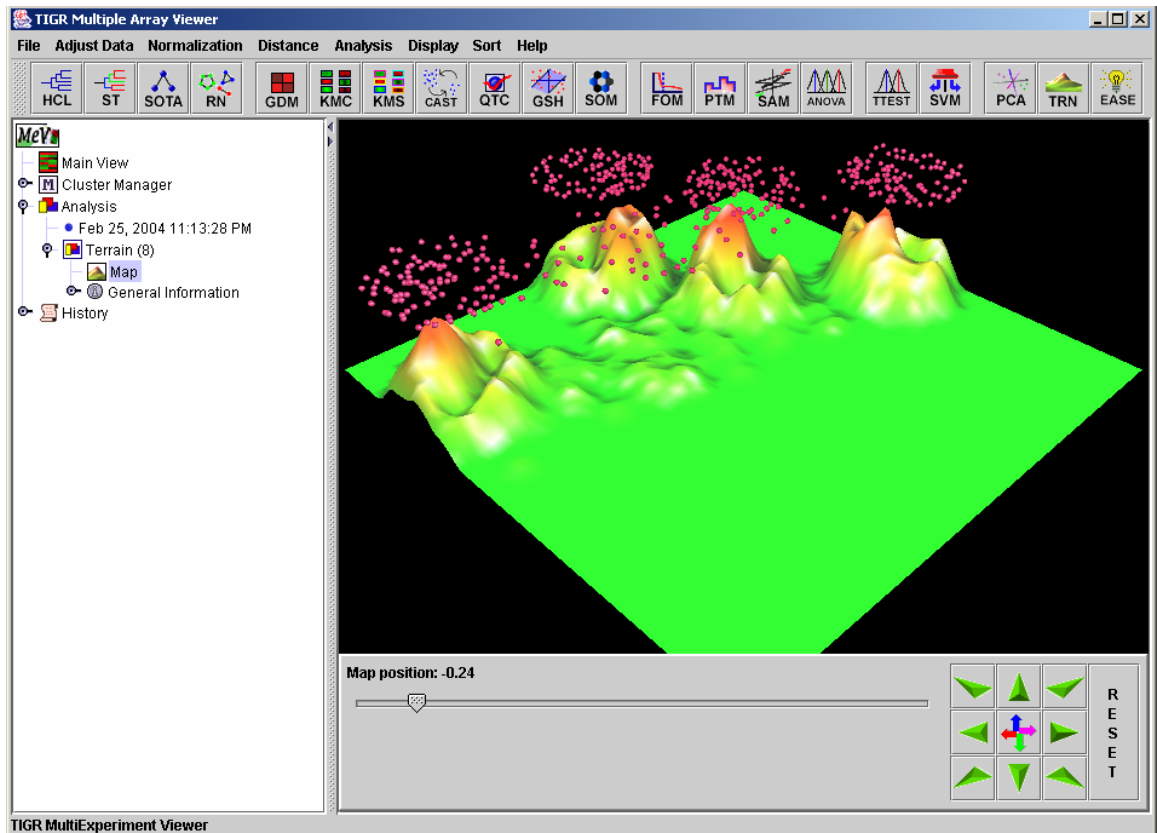
It is very important to note that like many algorithms TRN uses a default distance metric that greatly affects the outcome of the terrain created. The default metric for TRN is Pearson Squared, which groups elements based on correlation and tends to place strongly correlated and strongly anti-correlated elements in similar groups. Explicit selection of Euclidean distance can sometimes reveal more obvious groupings. See section 13, the distance metric appendix, for more details on available metrics in MeV.



11.22.1. Terrain Map Initialization Dialog.

Data Element Identification

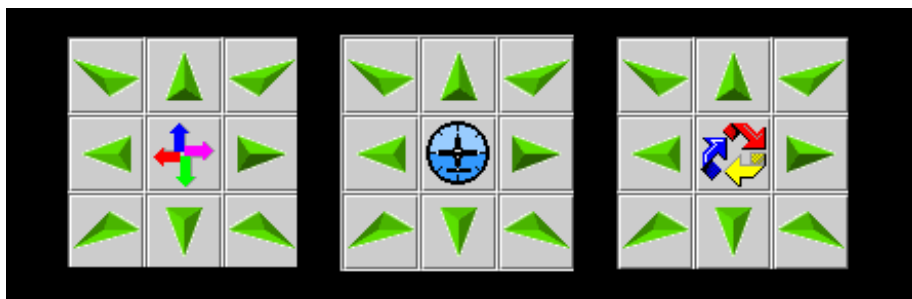
Data element identification can be achieved by four major methods. The right click menu allows one to turn on labels. The utility of the labels depends on how close the elements are to one another. It is also possible to move the mouse cursor over an element and reveal the annotation label currently selected in the main *Display* menu (*Label* submenu). A third option is to click on an element, which will open an element information table listing annotation for the selected element. The fourth option is to select one or more elements in an area and either write them to a file or open a new mev session with the selected elements. Details about element selection are provided below.



11.22.2. Terrain Map with Navigation Controls.

Terrain Navigation

The main visible controls below the terrain view are related to navigation through the terrain. The navigation panel consists of nine buttons that will alter the point of view (pov) relative to the terrain. There are three major navigation modes for moving through the terrain viewer. Each mode is selected by clicking on the center square of the navigation control.



11.22.3. Terrain Map Navigational Controls, Various Navigation Modes

Linear Axis Mode

This mode selection is indicated by the four straight arrows icon in the center of the navigation control panel. The arrow buttons move the point of view relative to the terrain in predictable straight line movements. The up and down arrows move the pov higher or lower relative to the terrain. The upper left and right arrows move the pov directly toward the terrain view while the lower right and left

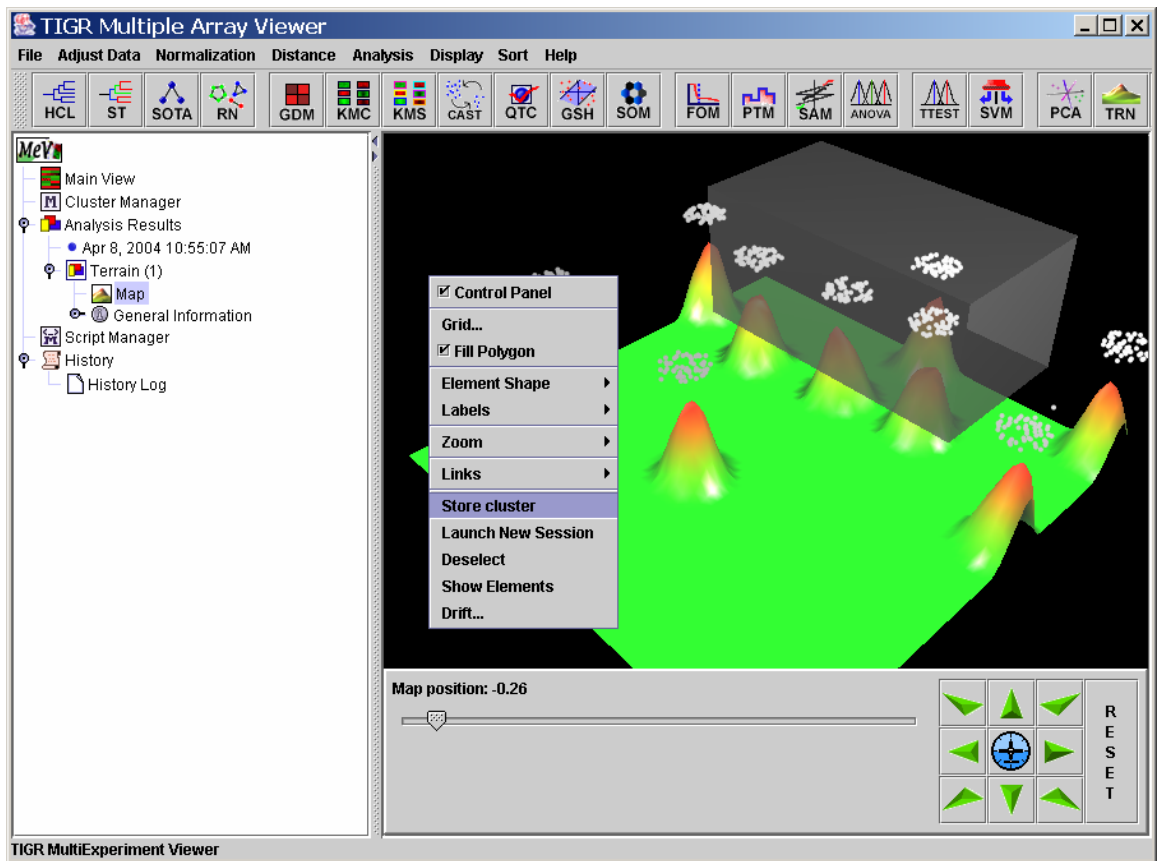
arrows move the pov directly away from the terrain. The left and right arrows move the pov left and right relative to the terrain.

Attitude Mode

This mode controls the pitch angle and roll of the pov relative to the terrain. The corner arrow buttons make the terrain roll about the center point of the viewer. The left and right force a banking motion and the up and down arrows control the pitch or angle of view.

Rotational Mode

This mode forces rotation of the terrain about a central point in the terrain. The up and down buttons tilt the terrain toward and away from the pov. The left and right button forces the terrain to rotate about the central point while staying constant relative to the horizon. The corner buttons permit a tight rolling motion of the terrain relative to the pov.



11.22.4. Terrain Map with Menu and Selection Area Visible (Euclidean Distance).

Altitude Slider

This moves the data points up and down relative to the terrain. Sometimes it is preferable to have the points higher than the terrain to view the distribution.

The Terrain Viewer Menu

The terrain viewer has a right click menu providing several options for altering aspects of data presentation and for extracting information such as gene lists. Each option is describe below.

Control Panel

This option determines whether the navigation control panel is visible or not.

Grid...

The Grid menu item permits the selection of the terrain's resolution or smoothness. Reducing the grid value can sometimes improve viewer response by saving computational expense on the computer. Lowering the grid value will render the terrain as a series of planes.

Fill Polygon

This option alters whether the terrain's surface is render or if simply the outline of adjacent planes. Like the Grid option deselecting this can produce faster rendering of the viewer during navigation.

Element Shape

Three optional element shapes are supported, point, cube, and sphere.

Labels

Labels can be displayed or hidden. If displayed there is a billboard option to help render the labels such that they can be seen from the current point of view. Note that if identification of a cluster of elements is desired the best way to verify is to select the elements and store them as a cluster or launch a new viewer containing the elements. Both options are described below.

Links

Elements passing at threshold criteria can display a link between the points. The links menu option allows the user to display or hide displayed links between elements passing the threshold criteria. Within the links menu the current threshold can also be set in order to visualize the strength of the associations within element groupings. The thickness of the links is also adjustable from 1 to 10 on a relative scale.

Drift (Auto-navigation Control)

Flying through the terrain to a selected point on the terrain or a data element is performed by holding down the *Ctrl* key and *left clicking* on the desired destination. The point of view will automatically navigate to the destination. If the selected destination is a point on the terrain such as a peak or a plateau the

point of view will orient such that the view is orthogonal to the plane on the terrain that contains the destination point. This can be useful to quickly get to a top down view when clicking on flat areas between peaks. If the selected destination is a data point, the final point of view on the element and in the plane containing the data elements. The route will move toward the element and attempt to orient in a position to more easily read the label associated with the element. Some further navigation may be required to avoid elements that may obstruct the element of interest.

The *Drift* menu option permits entry of a relative distance parameter that can range from 0 up to but not including 1.0. This represents the final distance between the point of view and the destination when drifting. Small values will cause the drift auto-navigation mode end position to be close to the destination and is often preferred when visiting a specific data element.

Cluster Selection and Related Operations

The following four options are only enabled if a group of elements has been selected in the selection area. Points are selected by holding down shift and dragging the mouse in the viewer. Once selected the following options are enabled:

Store Cluster

The store cluster option will store the elements in the selection box in the Cluster Manager. The clusters in the repository are viewable in the cluster table and if incomplete selections are made in TRN, several clusters can be made and then joined (Union operation) in the cluster table to capture all of the desired elements.

Launch New Session

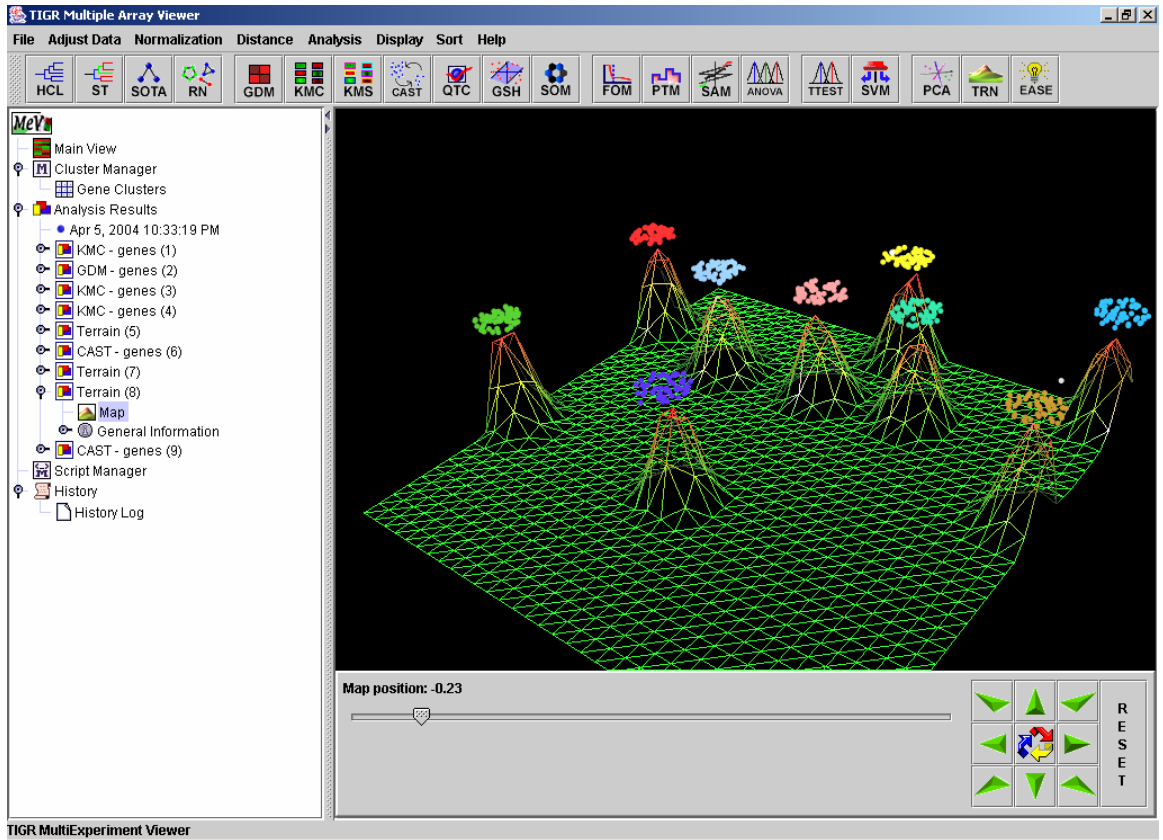
This feature launches a new Multiple Experiment Viewer containing the selected elements. This is perhaps the fastest and most convenient means of viewing the selected elements. An alternative to this is the Show Selection option described below.

Deselect

This option dismisses the selection area.

Show Selection

This option shows the selected elements within a small window. Note that the window supports the selection of multiple elements using shift-click or ctrl-click and the list can be copied and pasted into other applications by ctrl-c.



11.22.5. Terrain Viewer Without Surface Rendering

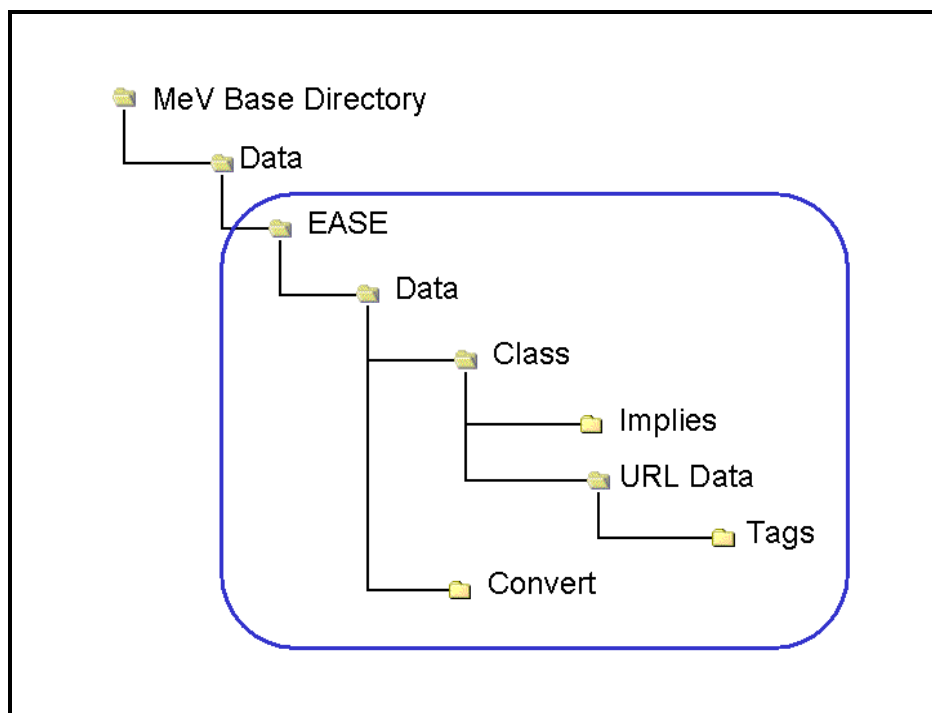
11.23. EASE: Expression Analysis Systematic Explorer (Hosack et al. 2003)

The implementation of EASE within the MeV framework provides a method to give the researcher an initial biological interpretation of gene clusters based on the indices provided in the input data set and information linking those indices to biological “themes”. These themes are generally GO terms, KEGG pathways, or any other descriptive term related to biological role or biochemical pathway information.

The result of the analysis is a group of biological themes which are represented in the cluster. A statistic reports the probability that the prevalence of a particular theme within the cluster is due to chance alone given the prevalence of that theme in the population of genes under study (all “genes” loaded into MeV).

This implementation is based on the over-representation analysis feature of the EASE application available at <http://david.niaid.nih.gov/david/ease.htm>. Two classes have been utilized from the EASEOpenSource package with modifications to enable some of the options described below. A full description to the theory behind EASE and test studies can be found in the EASE reference, Hosack, et al., 2003.

Behind the MeV version of EASE there is a file structure that contains files required for annotation conversion and linking indices to biological themes. The file structure mimics much of the file structure behind the stand-alone version of EASE.



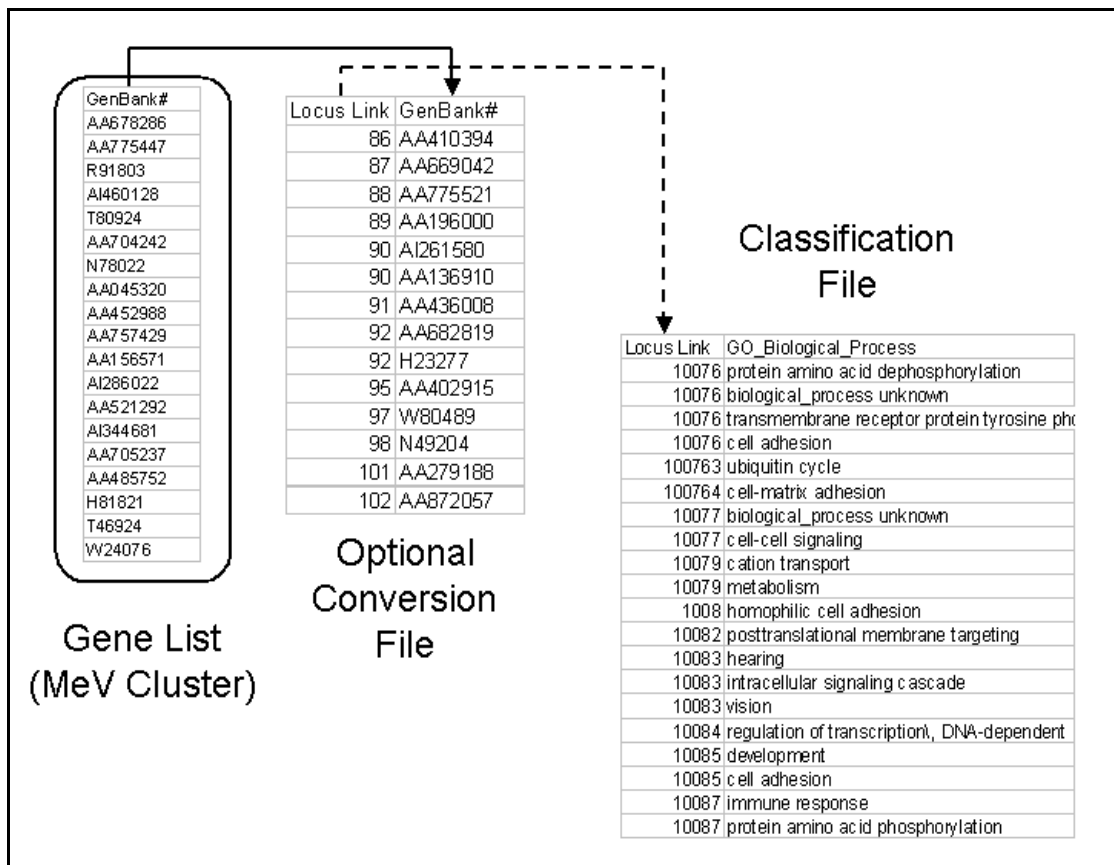
11.23.1. EASE Directory Structure

The following table lists the directories used in the EASE implementation within MeV. The primary directories, *Convert*, *Class*, *Implies*, *URL Data*, and *Tags* each contain files used by EASE. The minimal requirement is a file in the *Class* directory to map gene indices to themes. The optional annotation conversion files are located in the *Convert* directory and serve to convert annotation types. This is only required if the annotation within MeV used to identify genes differs from that contained in the file in the *Class* directory linking annotation to biological theme. The *Implies* directory contains optional files describing the hierarchy of biological theme terms where one theme description might imply another. The *URL* related files are also optional but having these allow the resulting tables to link to web sites which describe the findings such as GO terms or KEGG pathways. See the table below for a summary of the directory structure.

Directory	Description
EASE	Root of the EASE file structure.
Data	Encompasses EASE data files.
Convert	Contains files linking indices (e.g. GenBank# → Locus Link ID) These files are <i>optional</i> and are not needed if the annotation indices for each gene are the same as the keys in the Classification File (below).
Class	Contains files linking indices to biological themes (e.g. locus link id → GO Biological Process) (MINIMAL REQUIREMENT)
Implies	Contains <i>optional</i> files relating themes to other themes (e.g. hydrogen ion transport “Implies” cation transport)
URL Data	Contains <i>optional</i> files describing a url indicating the tag (accession placement). The contents of these files act as a models when constructing links to resources using annotation accession numbers.
Tags	Contains <i>optional</i> files linking biological theme or pathway to accession numbers

11.23.2. EASE Directory Descriptions

The files behind EASE are often used to link one annotation key to another annotation value and for that purpose most files are arranged in rows containing key/value pairs separated by a tab delimiter. The figure below (9.18.3) shows a scenario which demonstrates how a primary index in MeV can be mapped through a secondary index to a biological theme, in this case a GO term.



11.23.3. Example of files linking annotation indices to biological themes

EASE Input Parameters

File Updates and Configuration Options

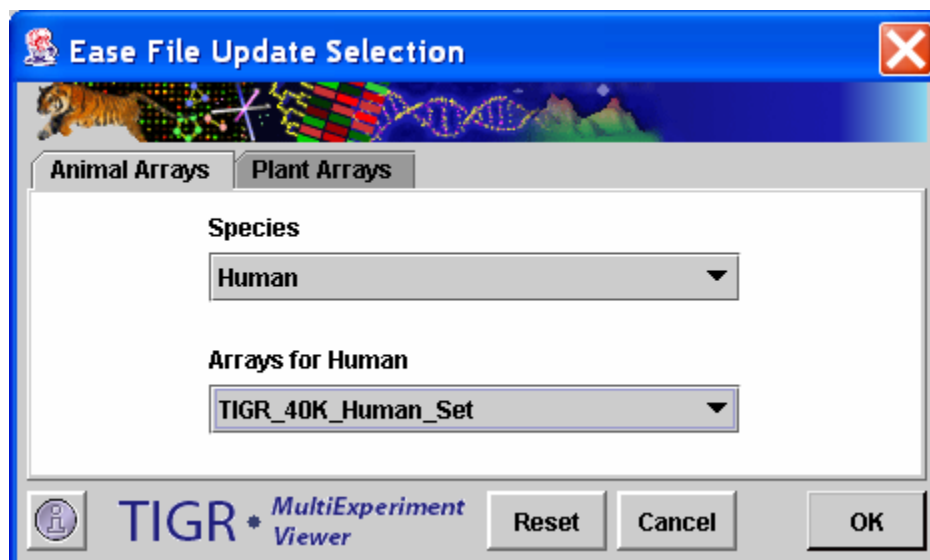
Selecting the EASE File System

This button enables the selection of a local directory to be used as the source for annotation files for EASE analysis. Multiple file systems can be present to support a variety of array types. Selection of a file system directs EASE related file choosers to that area but file selections may be made outside of selected base file system if appropriate. Note that the selected directory should be the directory that contains the EASE "Data" directory. In MeV's default data directory this would be the 'ease' directory.

Updating/Downloading the EASE File System

This feature downloads EASE annotation file systems for a selected species and array. A selection dialog will allow species selection from a variety of plant and animal species. A list of many commercially available arrays for the selected species is also presented for selection. After species and array selection, a dialog will be presented to select a directory as the destination directory for the EASE file system. Zip files will then be downloaded and automatically extracted into the destination directory. The new base directory will be labeled with "ease_" and

the selected array name. This new file system can be selected as the default system.



11.23.4. EASE File Update Array Selection Dialog

Analysis Mode Selection

The EASE implementation in MeV provides two major modes of operation, *Cluster Analysis* and *Slide Annotation Survey* modes.

Cluster Analysis

This mode performs annotation analysis on a selected subset (sample list or cluster) of the full data set loaded in MeV. The output is a list of biological 'themes' represented in the cluster and a statistic reporting the probability that a particular theme is over represented in the cluster relative to it's representation in the entire data set. The resulting table will initially be sorted by this statistic.

Slide Annotation Survey

The survey mode simply produces a list of biological themes that are represented in the slide. The initial ordering of the output table is based on the prevalence of a theme in the data set (hit count). This mode can be used to cluster genes based on biological themes. The clusters can then be stored and marked (colored) for tracking during cluster analysis.

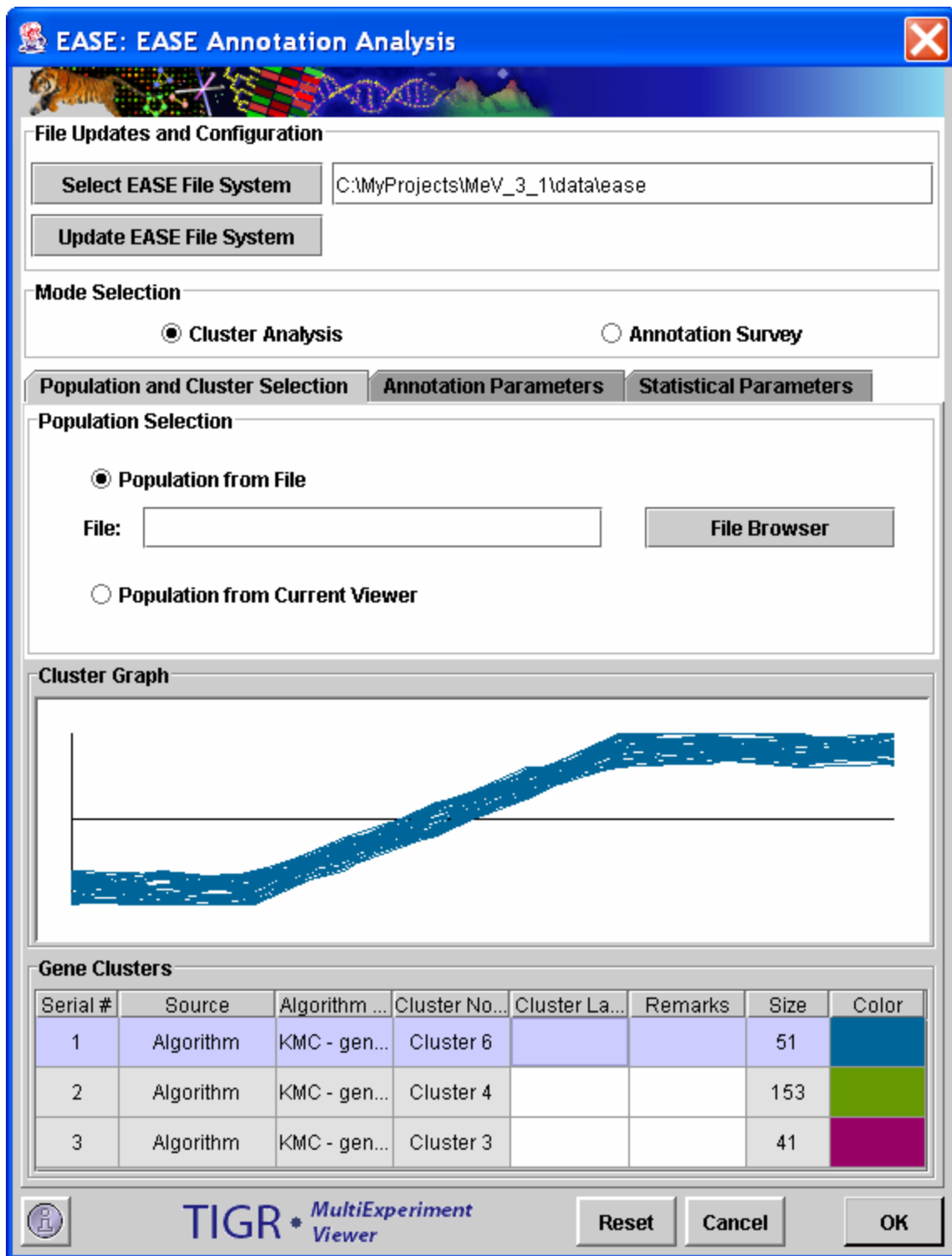
Population and Cluster Selection Page

This panel provides options to specify a gene population list and a gene cluster list. The top portion is devoted to selection of a population list and has two options for list selection. The default selection is to use a population file which is simply a file containing all gene indices from which the cluster was selected.

Often this includes all slide annotation or a large subset of the slide with bad data and control spots removed.

The file format is a simple list of indices with one entry per line. The other option for population definition is to simply use all of the genes loaded into MeV. It is often necessary to use the file to define the gene population because often the current viewer may not contain all genes considered to be part of the population. This is the case when the viewer was launched as a new viewer on a data subset or if the viewer was initially loaded with a previously saved cluster.

The Population and Cluster Selection Panel also displays gene clusters currently stored in MeV's cluster repository. If no clusters have been saved then a blank browser page will be displayed on this page and the Cluster Analysis mode option will be disabled. Selecting a row in the cluster table will display the cluster in the expression graph area of the browser. EASE cluster analysis will operate on the selected cluster.



11.23.5. Population and Cluster Selection Page

Annotation Parameters Page

MeV Annotation Key

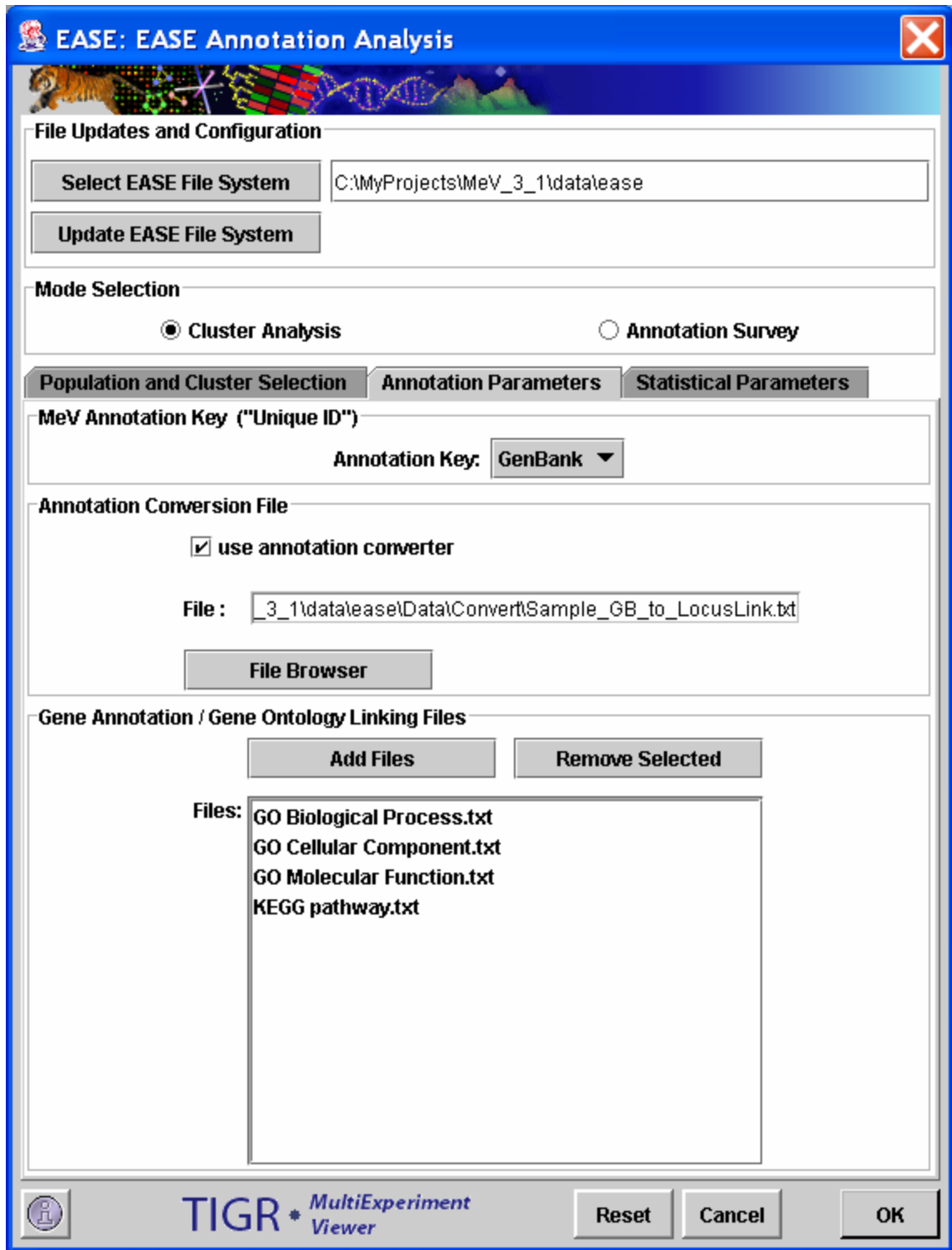
This area contains a drop down list which contains a list of available annotation types which can be used identify genes. Generally it's best to use an index or accession 'uniquely' identifying the spotted material.

Annotation Conversion File

This optional file provides the mapping from your annotation key (above) to the index used to map to biological themes (GO terms, KEGG pathways, etc.). If your annotation key type is the one used in the linking file (below) then this conversion (mapping) is not needed. These files if needed are typically stored in the *Convert* directory.

Gene Annotation / Gene Ontology Linking Files

This section allows one to specify one or more annotation files. These files contain gene indices paired with biological themes such as GO terms. These files typically reside in the *Class* directory.



11.23.6. Annotation Parameters Page

Statistical Parameters Page

Several sections on this page are used to specify the reported statistic, optional multiplicity corrections, and optional result trimming parameters.

Reported Statistic

Fisher Exact Probability

The Fisher's Exact Probability reports the probability that a biological theme is over-represented in the cluster of interest relative to the representation of that theme in the total gene population. For example, suppose that one has a gene list of 50 genes from a population of 10,000 genes. Now suppose that 10 of the 50 genes were related to pathway "A" but only 13 genes in the total population were associated with pathway "A". This scenario would yield a low probability that the observed number of hits (occurrences of pathway "A") within the small sample could be due to chance alone. This statistic is based on the hypergeometric distribution and has benefits over chi-square in that it is appropriate for finite populations. The reference cited for EASE describes this statistic at length.

EASE Score

The EASE Score reported is essentially a jackknifed Fisher's Exact Probability that is arrived at by calculation of the Fisher's Exact where one occurrence (list hit for a term) has been removed.

Multiplicity Corrections

Bonferroni Correction:

This correction simply multiplies the statistic by the number of results generated. This is the most stringent correction of the three options.

Bonferroni Step Down Correction:

This modified Bonferroni correction ranks the results by the statistic in ascending order. Each value is multiplied by (n-rank) where n is the number of results. In the case of a tie, where two results have the same probability the rank is kept constant until the next element occurs having a higher probability value. The rank is then adjusted for the number of tied elements where rank was constant.

Sidak Method:

This correction uses the following formula where v' is the corrected value and k is the rank of the result in terms of original statistic value. In this case ties in rank are handled as described in the step down Bonferroni correction described above.

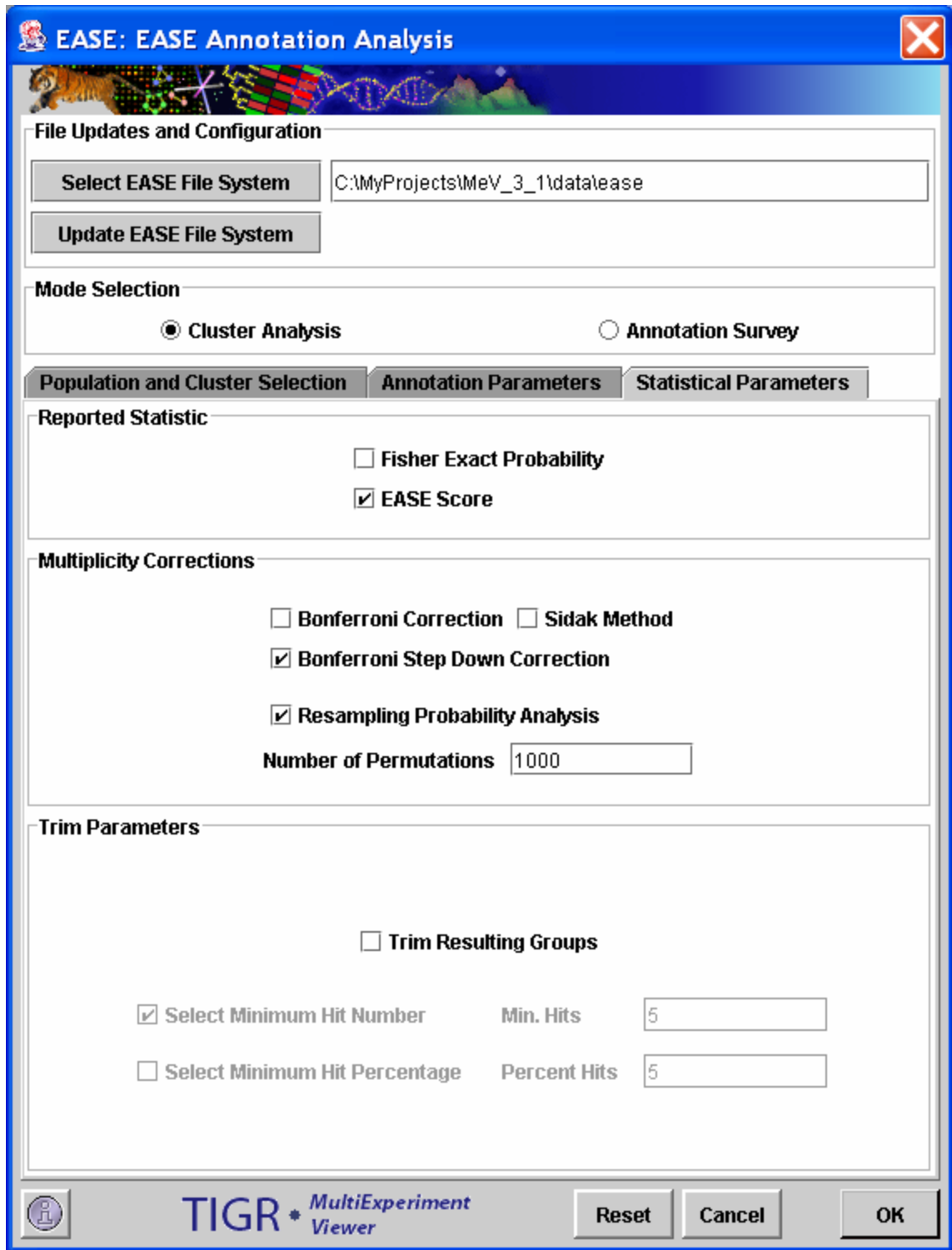
$$v' = 1 - (1 - v)^k \text{ (Sidak method formula)}$$

Resampling Probability Analysis:

The resampling option performs a number of resampling iterations where random gene lists of the initial input size are selected from the population without replacement and run through the analysis. The result for each biological theme is a probability which indicates the probability of the original significance level (EASE score or Fisher Exact) occurring by chance alone.

Trim Parameters

The trim parameters can be applied to filter the analysis results based on either the *number* of hits within the cluster or on the *percentage* of genes in the cluster that are represented by an annotation term. Sometimes a term can be found significant but does not represent a large segment of the cluster of interest. These options can be applied to be certain that a minimum number of genes in the cluster fall under that particular annotation class. This feature should be used with caution so that biological themes represented by very few genes are not excluded.



11.23.7. Statistical Parameters Page

Results of EASE Analysis

The primary result is reported in a table in which entries are ordered based on the reported statistic. The table can be sorted on any column. A right click in the table will launch a menu allowing several operations:

Store Selection as Cluster: Stores the genes associated with a biological theme as a cluster that will be stored in the cluster manager.

Open Viewer: Opens one of three possible viewers containing the genes within the biological theme. These viewers are also accessible from a node in the result tree which follows the table node in result navigation tree.

The screenshot shows the 'TIGR Multiple Array Viewer' window. The interface includes a menu bar (File, Adjust Data, Normalization, Distance, Analysis, Display, Sort, Help) and a toolbar with various analysis tools like HCL, ST, SOTA, RN, GDM, KMC, KMS, CAST, QTC, GSH, SOM, FOM, PTM, SAM, ANOVA, TTEST, SVM, PCA, TRN, and EASE. Below the toolbar is a table with the following columns: Index, File, Acc., Term, List Hits, List Size, Pop. Hits, Pop. Size, and Fisher's Ex... The table contains 26 rows of data, primarily related to GO Biological Processes and KEGG pathways, such as 'Oxidative phosphorylation - Homo sapiens' and 'ATP synthesis - Homo sapiens'.

Index	File	Acc.	Term	List Hits	List Size	Pop. Hits	Pop. Size	Fisher's Ex...
1	KEGG pathway	hsa00190	Oxidative phosphorylation - Homo sapiens	18	20	18	120	1.748E-19
2	GO Molecular Function	GO:0015078	hydrogen ion transporter activity	18	38	18	323	2.378E-19
3	KEGG pathway	hsa00193	ATP synthesis - Homo sapiens	17	20	17	120	6.003E-18
4	GO Biological Process	GO:0006818	hydrogen transport	18	38	20	325	3.549E-17
5	GO Biological Process	GO:0015992	proton transport	18	38	20	325	3.549E-17
6	GO Molecular Function	GO:0008324	cation transporter activity	22	38	32	323	1.239E-16
7	GO Molecular Function	GO:0015077	monovalent inorganic cation transporter activity	18	38	21	323	2.608E-16
8	GO Molecular Function	GO:0015075	ion transporter activity	22	38	34	323	9.45E-16
9	GO Biological Process	GO:0006812	cation transport	22	38	36	325	5.126E-15
10	GO Biological Process	GO:0006811	ion transport	22	38	37	325	1.198E-14
11	KEGG pathway	hsa01120	Energy Metabolism - Homo sapiens	18	20	25	120	7.342E-14
12	GO Biological Process	GO:0009206	purine ribonucleoside triphosphate biosynthesis	12	38	12	325	1.147E-12
13	GO Biological Process	GO:0009108	coenzyme biosynthesis	12	38	12	325	1.147E-12
14	GO Biological Process	GO:0009142	nucleoside triphosphate biosynthesis	12	38	12	325	1.147E-12
15	GO Biological Process	GO:0009201	ribonucleoside triphosphate biosynthesis	12	38	12	325	1.147E-12
16	GO Biological Process	GO:0009145	purine nucleoside triphosphate biosynthesis	12	38	12	325	1.147E-12
17	GO Biological Process	GO:0006753	nucleoside phosphate metabolism	12	38	12	325	1.147E-12
18	GO Biological Process	GO:0006754	ATP biosynthesis	12	38	12	325	1.147E-12
19	GO Biological Process	GO:0015672	monovalent inorganic cation transport	18	38	30	325	8.436E-12
20	GO Biological Process	GO:0006752	group transfer coenzyme metabolism	12	38	13	325	1.377E-11
21	GO Biological Process	GO:0009205	purine ribonucleoside triphosphate metabolism	12	38	14	325	8.896E-11
22	GO Biological Process	GO:0009141	nucleoside triphosphate metabolism	12	38	14	325	8.896E-11
23	GO Biological Process	GO:0009144	purine nucleoside triphosphate metabolism	12	38	14	325	8.896E-11
24	GO Biological Process	GO:0009199	ribonucleoside triphosphate metabolism	12	38	14	325	8.896E-11
25	GO Biological Process	GO:0006732	coenzyme metabolism	12	38	14	325	8.896E-11
26	GO Biological Process	GO:0046034	ATP metabolism	12	38	14	325	8.896E-11

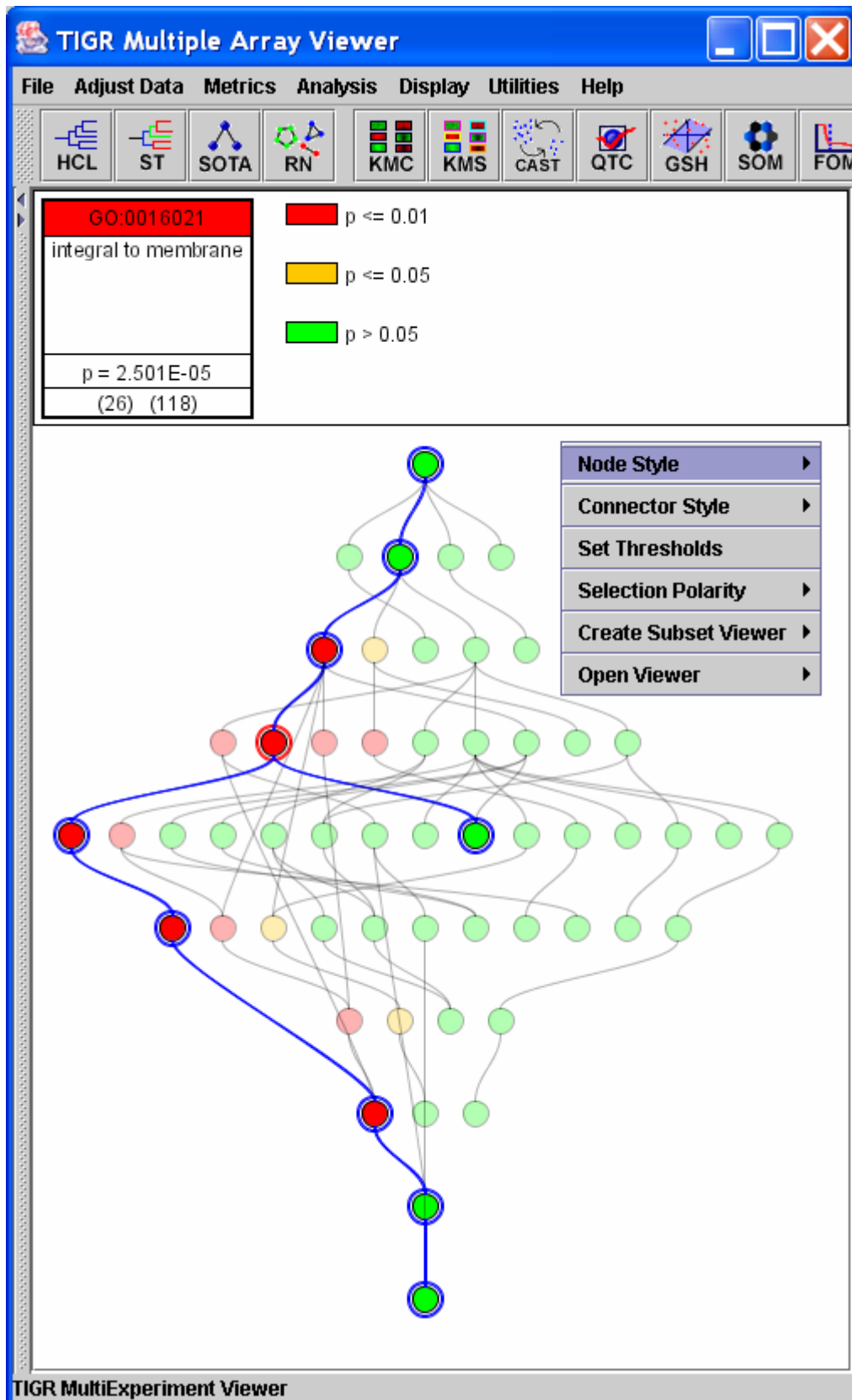
11.23.8. EASE Result Table

Save EASE Table: Stores the result table to a tab delimited file.

Open Web Page : Opens the default web browser using the URL associated with the theme file (e.g. KEGG pathway.txt) using the available accession or index. If the accession numbers are not available (in the Tags directory) or no URL file has been entered in the URL Data directory this feature will be disabled.

GO Hierarchy Viewer

The GO Hierarchy Viewer is a hierarchical representation of GO terms resulting from an EASE cluster analysis. Each of the GO terms found in the cluster under analysis is represented by a node in the hierarchy and as one descends down a path in the hierarchy the terms represented by the nodes in the path tend to become narrower in scope in terms of identifying a particular biological theme. The color of each node represents the theme's p-value relative to user-defined thresholds.



11.23.9. EASE GO Hierarchy Viewer with Selected Path (Non-verbose nodes)

The header represents the currently selected node in verbose rendering or the hierarchy's root in the case where no node is selected. The header also provides the key to relate node color to the two user defined thresholds.

The viewer allows several options to customize the view and to extract information. A right click menu provides the following options.

Node Style

Node style dictates the rendering of the go tree nodes. *Minimal* rendering represents each node as a circle with the color representing the p-value's relation to the defined thresholds. The *Verbose* rendering provides more information about the identity of the GO term including, GO id, GO term name, the p-value, and the number of genes in the cluster and in the population that are related to this GO term. Note the example of verbose rendering in the header of the Tree Viewer figure above.

Connector Style

Two connector styles are available, curved and strait. Under some circumstances, the strait rendering provides a more easily traced path.

Set Thresholds

Two user-defined p-value thresholds can be set using this option. The initial default levels are 0.05 and 0.01 for the upper and lower thresholds respectively. Representation of p-value significance by discrete colors provides a quick means to focus on significant results. Alterations to the thresholds are immediately represented in the header and the tree's node colors.

Selection Polarity

The selection polarity option provides the option to modify tree path selection. The default is *Bipolar Selection* in which the nodes above (to the root node) and below the selected node are all selected. The *Select Ancestors* forces path selection to extend from the selected node up to the root node of the tree. The *Select Successors* option forces path selection to extend from the selected node down the tree.

Create Subset Viewer

Extracting subsets of nodes from the main hierarchy is perhaps the most important feature of this viewer. After selection of a tree path, it is possible to use this option to build a new GO Hierarchy Viewer from the nodes in the selected path. The extracted trees can be rendered in a new window or docked in MeV's main viewer panel. Note that if the option is taken to render the new tree docked in MeV's viewer panel then a node is placed in the result tree so that the subtree can be saved in the analysis.

Open Viewer

The open viewer option provides a shortcut method to jump from a node of interest to a viewer containing all of the cluster's genes that are related to the selected theme node.

12. Scripting

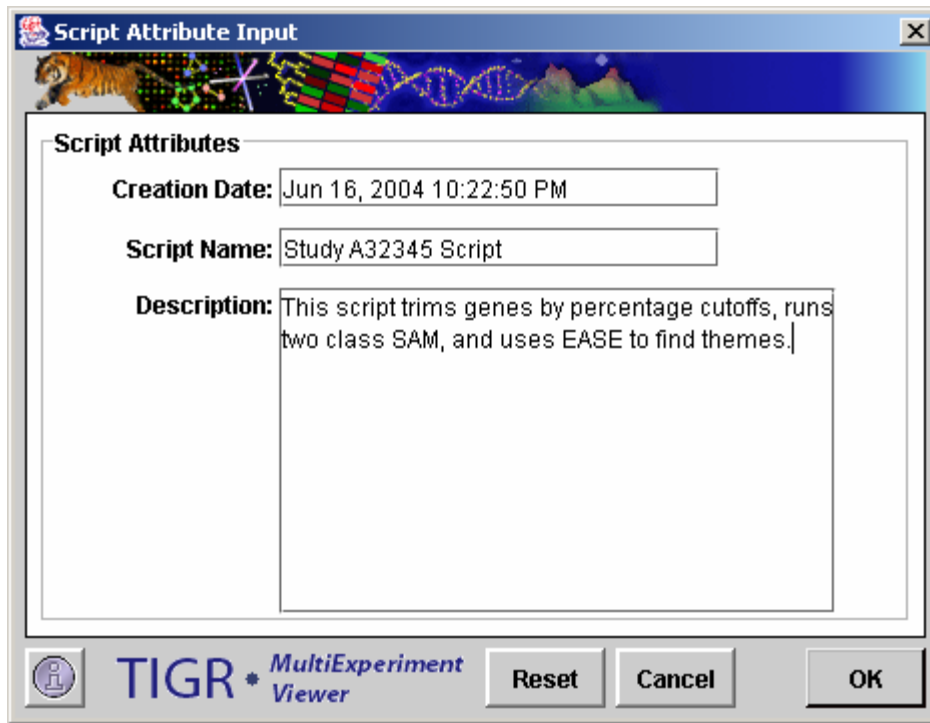
The scripting capabilities within MeV permit the execution of multiple algorithms to be performed without user oversight or intervention once processing begins. The execution steps are dictated by a user-defined script that describes the parameters to use for the selected algorithms. Scripting in MeV allows one to document the algorithms run and the selected parameters during data analysis. The script document can be shared with collaborators so that analysis steps can be replicated on the common data set. Scripting is also useful when running several long analysis steps that would normally require monitoring in MeV's interactive mode. Each algorithm and the parameters are pre-selected in the script so the next algorithm kicks off as soon as the previous run finishes. Despite the advantages of scripting, there may be times when careful evaluation of a result before deciding on the next algorithms is needed. In this setting scripting might be used as a first pass analysis and the multiple results of the script run can lead to the selection of different algorithms or new parameter selections.

The Script

The MeV script is an XML based text document containing information about which algorithms to run, the order of the algorithms, and the source data for each algorithm. Script creation is accomplished through a graphical representation of the script to eliminate the need for the user to understand the complex structure of the script. The Document Type Definition (DTD) can be found in Section 14 Appendix (MeV's Script DTD) for those interested in the details of script structure.

Creating a New Script

Creating a script is a simple process that can be initiated by selecting *New Script* from the *File* menu in the Multiple Array Viewer. Data must be loaded before this menu option can be enabled since many algorithms require data-specific information (e.g. group assignments for TTEST or SAM depend on the number and order of the loaded experiments). Once the New Script menu option has been selected, an initial dialog form will come up to allow one to enter a script name and description.

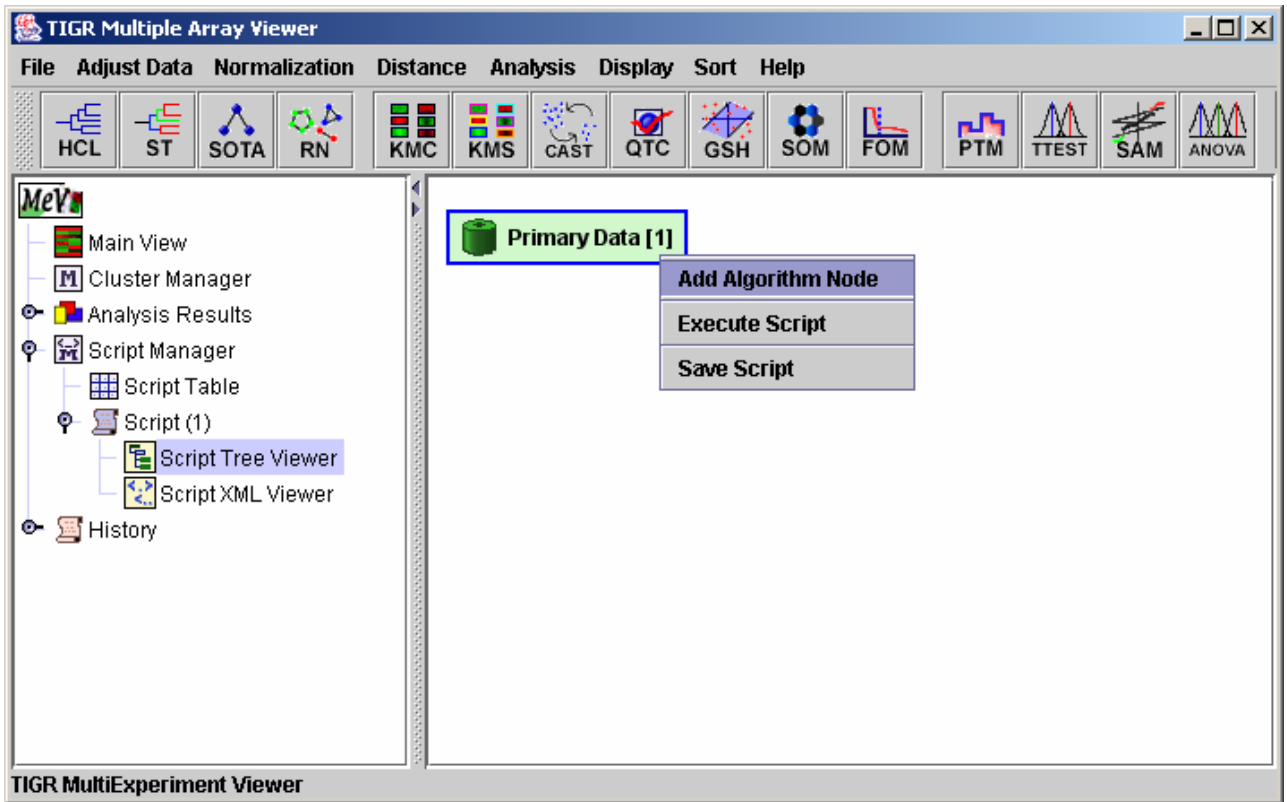


12.1.1. New Script Attribute Dialog

The script name, description, and the creation date will be stored in the script as comments. Once the initialization dialog is dismissed the script manager node will become populated with a script table and two viewers associated with the new script. The viewer that opens automatically is called the *Script Tree Viewer*. This viewer is a graphical representation of the script and it is from this viewer that the user constructs the script. The other script viewer is the *Script XML viewer*, which displays the actual text of the script during script creation.

The Script Tree Viewer: Script Construction

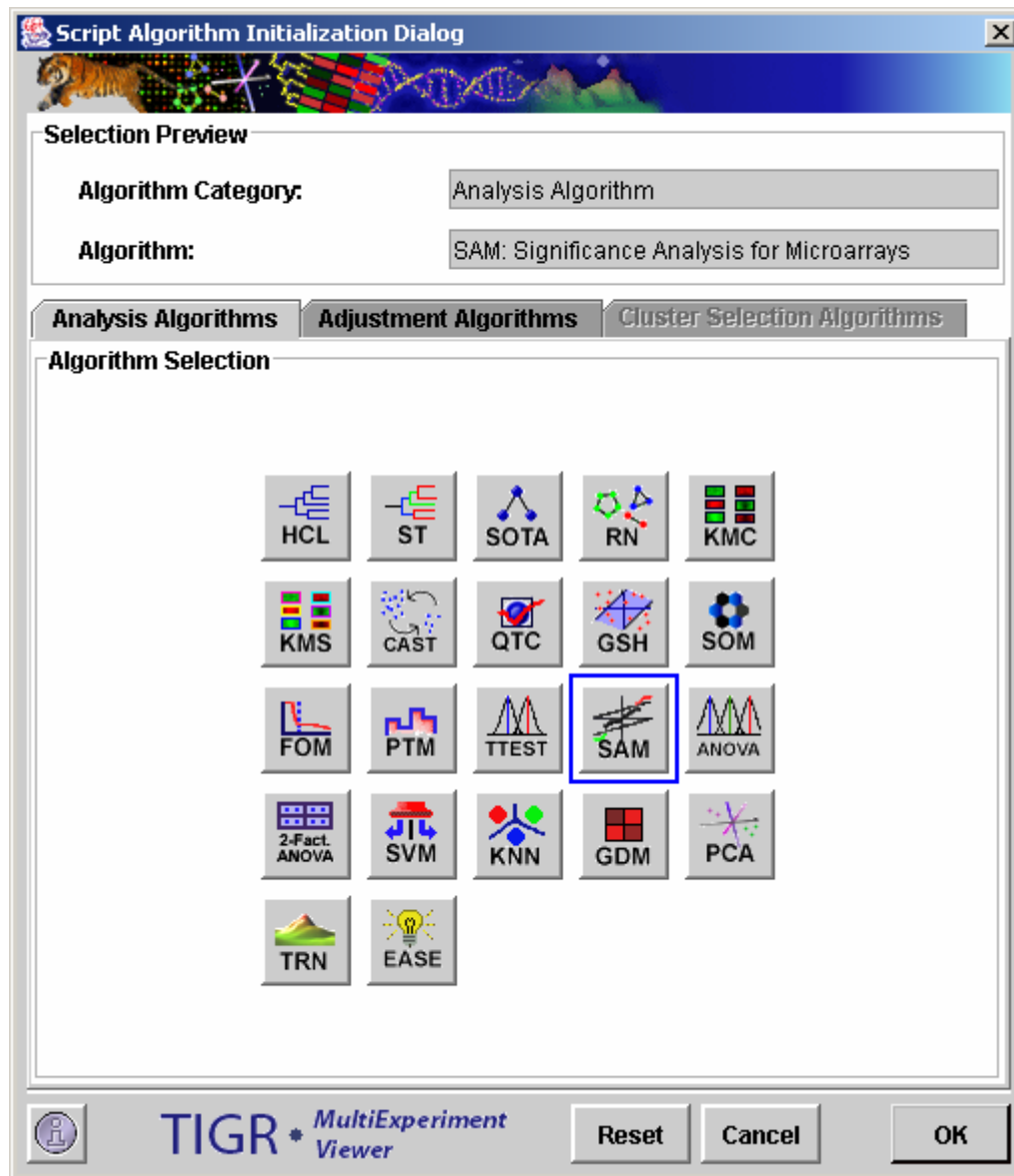
The Script Tree Viewer is the main viewer used to construct the script. The viewer's graphical nature permits the user to focus on script creation without undue consideration of complex script syntax. The Script Tree Viewer represents the script as a set of connected nodes. Each node is either a *data node* or an *algorithm node*. Data nodes, shown as light green, are sources for data for attached algorithms. Any number of algorithms can be attached to a data node. Algorithm nodes, shown as light yellow, represent processes that transform the data or act on the data to produce results.



12.2. Script Tree Viewer with Initial Primary Data Node and Pop-up Menu

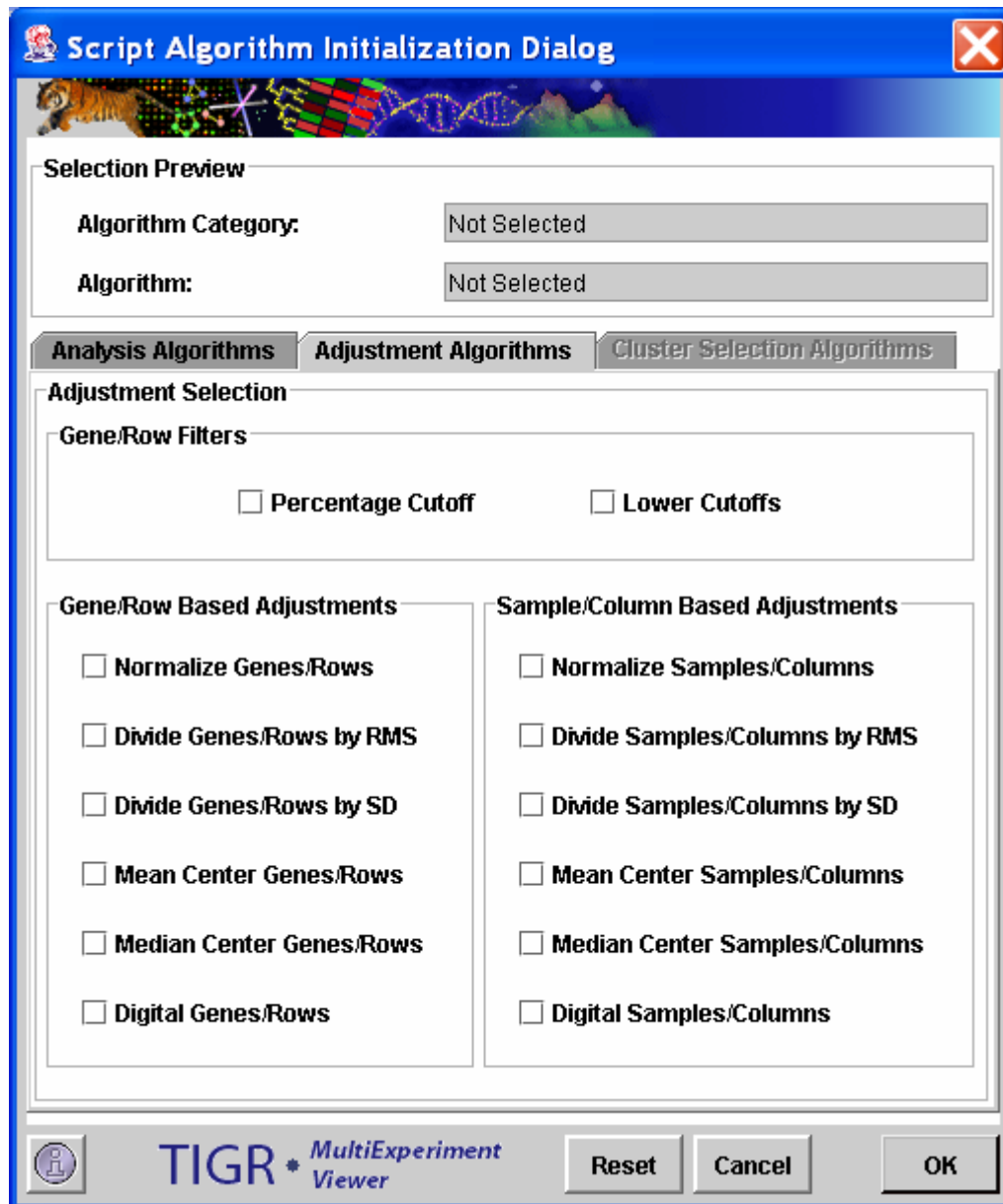
Adding an Algorithm

Select a data node to use as source data by left clicking a data node. Selected nodes will have a blue highlighted border when selected. A right click will reveal a menu containing an *Add Algorithm Node* menu option. Add Algorithm Node will present a dialog used to select the algorithm and parameters to append to the data node.



12.3. Script Algorithm Selection Dialog, Analysis Algorithm Panel

The algorithms fall into three main categories represented on three tabbed panels in the Algorithm Selection Dialog. Analysis Algorithms include gene and experiment clustering algorithms, classification algorithms, statistical algorithms, and data visualizations. The analysis algorithms are all described in the *Modules* section of the manual (section 9). The Adjustment algorithms are those algorithms found in the *Adjustment* menu of the Multiple Array Viewer interface and include the Affymetrix™ based filters if Affymetrix™ data is currently loaded. The Adjustment algorithms either filter the data based on some criteria or are used to perform a mathematical transformation of the data.



12.4. Algorithm Selection Dialog, Adjustment Algorithm Panel

The cluster selection algorithms are specific to scripting in MeV. Automatic Cluster Selection allows the user to provide criteria for evaluating cluster results where clusters have no intrinsic identity such as “significant genes”. One scenario is the result from K-Means Clustering (KMC) where $K=10$. In this case, 10 clusters will be produced and the cluster selection algorithms could be used to extract clusters based supplied criteria.

More on Cluster Selection Algorithms

Two main options are available for cluster selection. *Diversity Ranking Cluster Selection* computes cluster diversity for each of the input clusters and then ranks the clusters from least variable to most variable. Clusters are selected that satisfy a minimum size (population) but are as least variable as possible. In Diversity

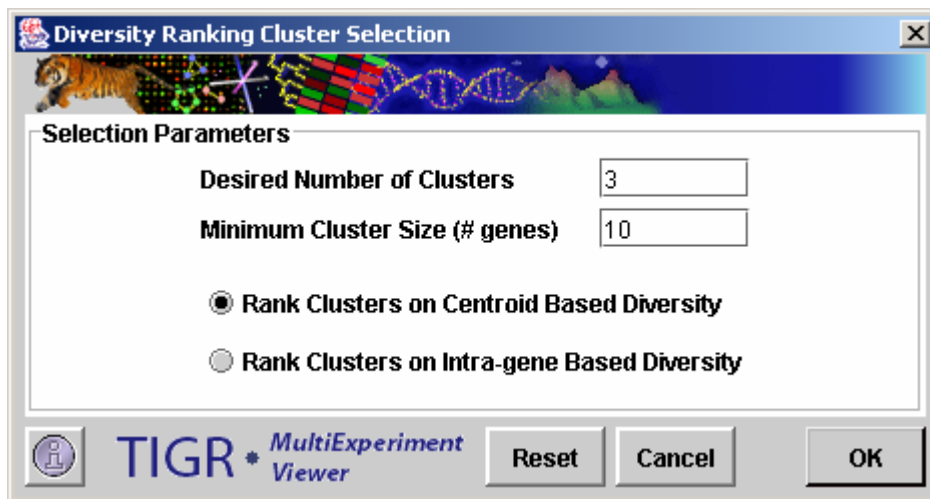
Ranking Cluster Selection two possibilities exist for determination of cluster diversity:

(1) **Centroid Based Diversity** (mean gene to centroid distance)

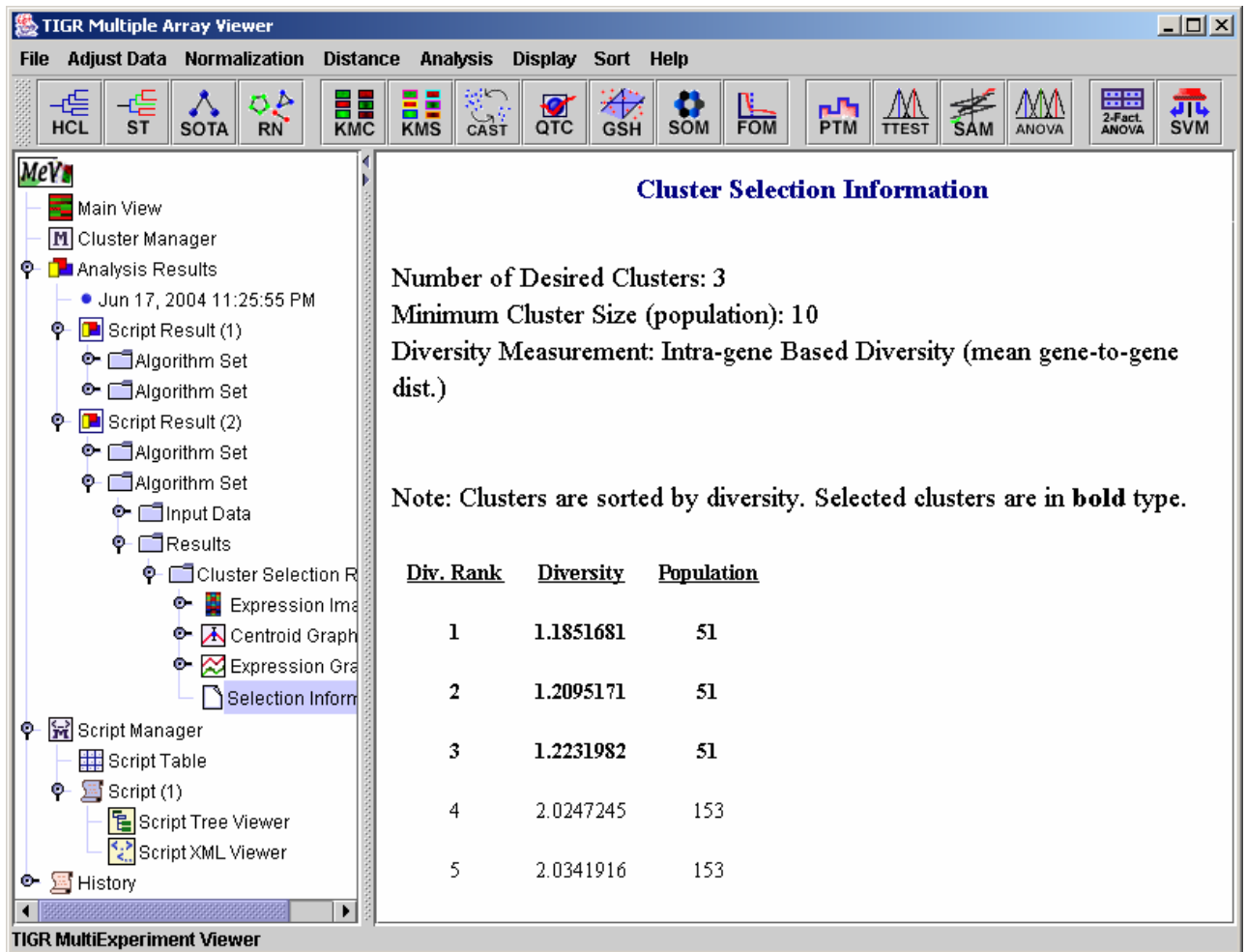
$$D = \left[\sum_{i=1}^n \text{dist}(\bar{g}_i, \bar{c}) \right] / n ; c \text{ is the cluster centroid, } g_i \text{ is the } i\text{th expression vector of } n \text{ vectors.}$$

(2) **Intra-gene Based Diversity** (mean of all gene-to-gene distances in the cluster)

$$D = \left[\sum_{i=1}^n \sum_{j=1}^i \text{dist}(\bar{g}_i, \bar{g}_j) \right] / [1/2(n^2 - n)]; i \neq j$$



12.5. Diversity Ranking Cluster Selection Dialog.



12.6. Cluster Selection Information Viewer.

The output is a list ranking the clusters by diversity with cluster population listed. Clusters that pass the size criteria and are least diverse are selected and are indicated in the list by bold type. The output nodes from cluster selection on the Script Tree can be used as input data to new algorithms.

The other option for cluster selection is *Centroid Entropy/Variance Ranking Cluster Selection*. This method places either a variance or an entropy value on the cluster's **centroid** (mean expression pattern). While the previous method selects tightly constructed clusters, this method focuses on finding clusters having variable centroids. The selected clusters are clusters that have a lot of variability on average over the expression measurements. The clusters are ranked on decreasing centroid entropy or variance, and clusters are selected with the highest centroid variability. The selected clusters must also pass a minimum cluster population. A Cluster Selection Information Viewer is created to describe the selection process. This viewer is similar to the viewer pictured for the Diversity Ranking Cluster Selection algorithm. Two options are available to describe centroid behavior, variance and entropy:

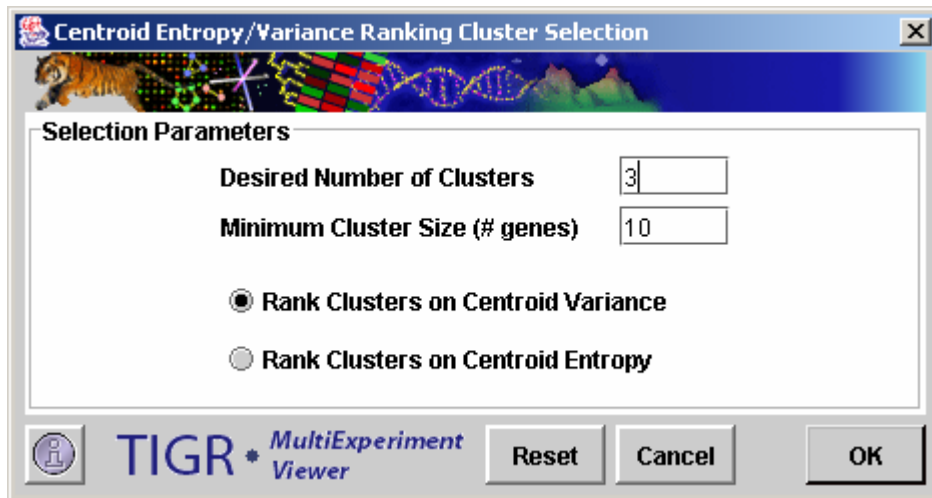
- (1) Centroid Variance

$V_c = \sum_{i=1}^m (c_i - \bar{x})^2$; V_c is centroid variance where \bar{x} is the centroid mean, c_i is the i th centroid value of m values.

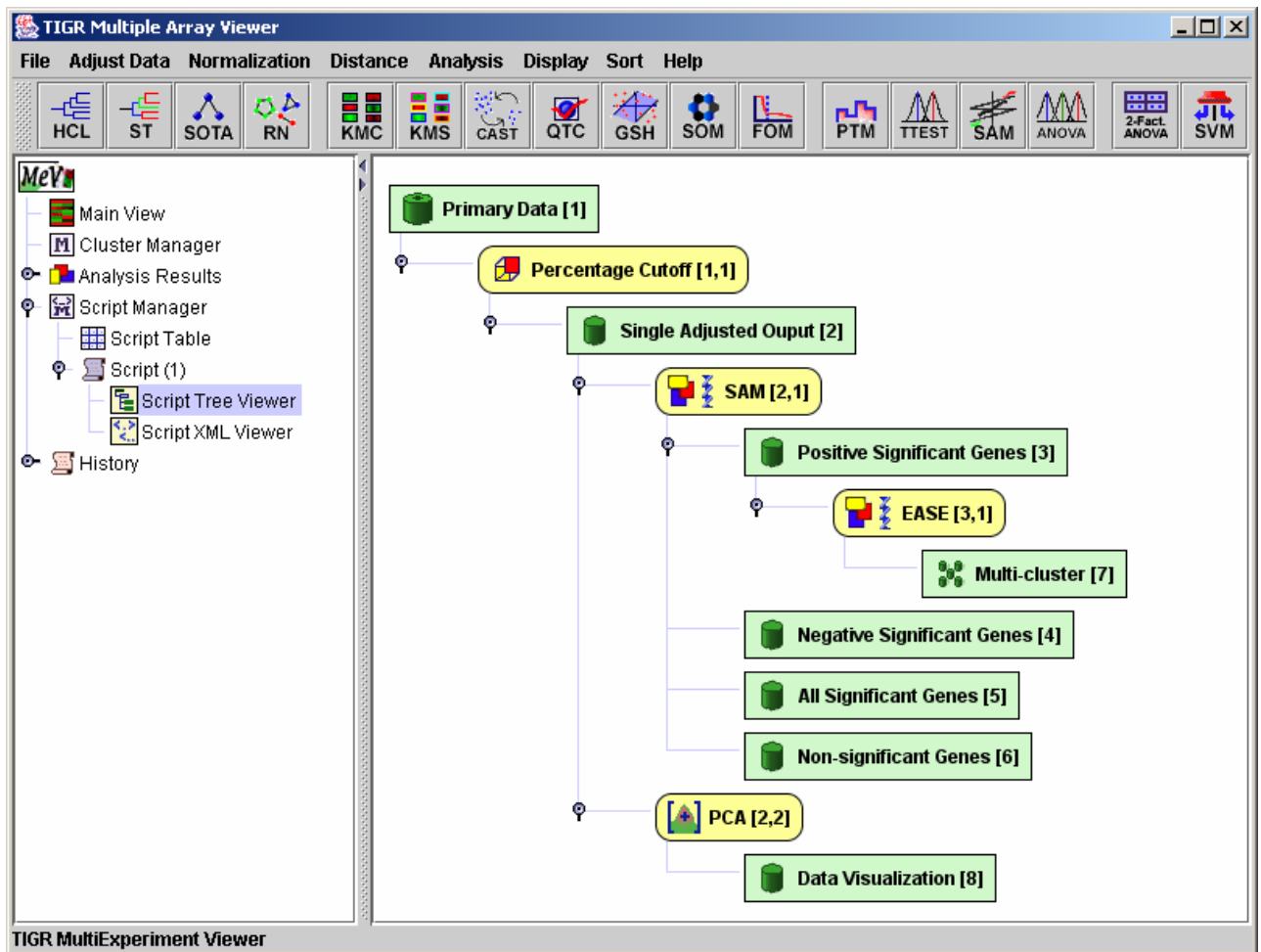
(2) Centroid Entropy

Entropy in this case describes the dispersion of expression values within the expression limits of the centroid. Centroid values are binned within 10 bins evenly dividing the expression value range for the centroid. $P(x)$ is the fraction of centroid points falling in bin x .

$$H = -\sum_{x=1}^{10} p(x) \log_2(p(x))$$



12.7. Centroid Entropy/Variance Ranking Cluster Selection Dialog.



12.8. Script Tree Viewer with Constructed Script

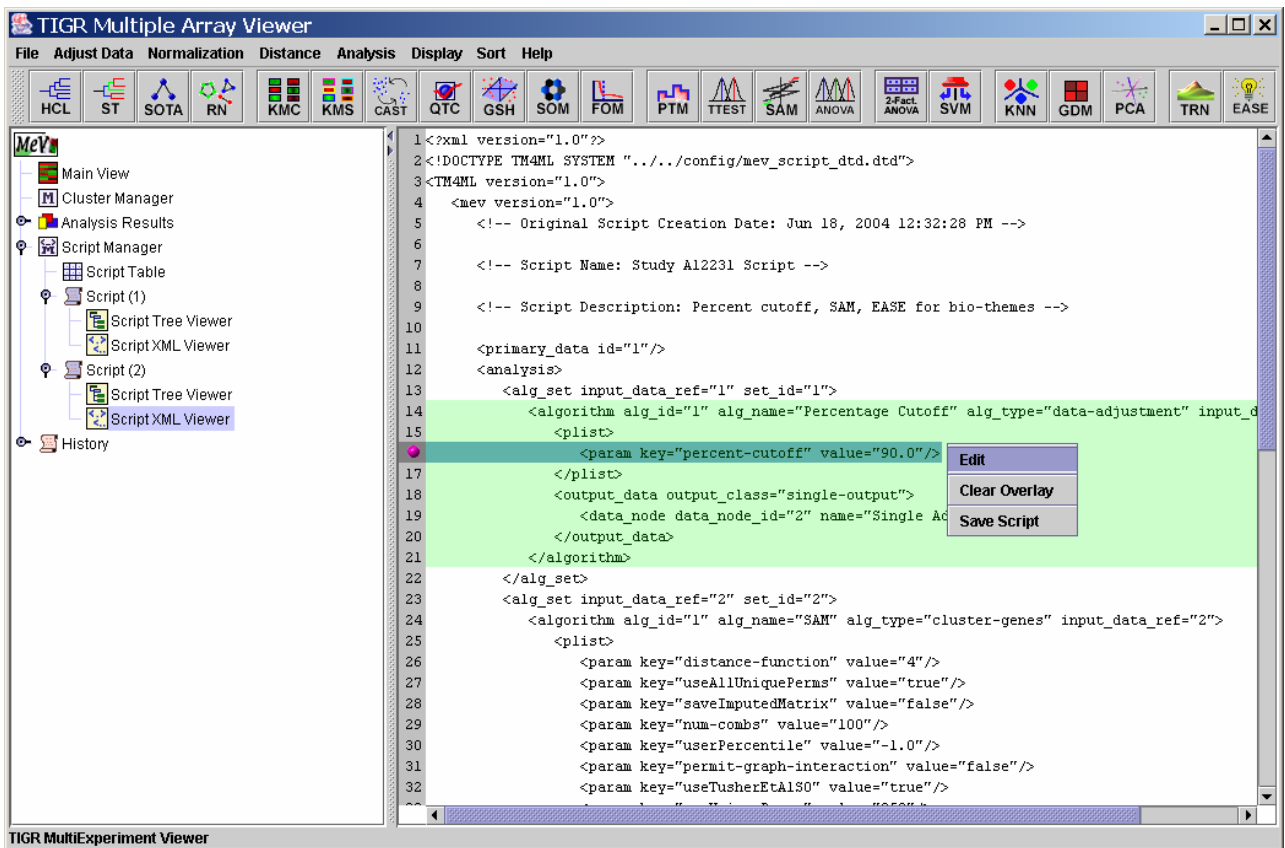
Script Tree Viewer Options

Right click menus displayed from the script tree viewers vary depending on whether the selected node is a data node or an algorithm node. The menu displayed from a data node provides the *Add Algorithm Node*, *Execute Script*, and *Save Script* options. The menu displayed from an algorithm node provides the *Delete Algorithm*, *View XML Section*, *Execute Script*, and *Save Script* options.

The Delete Algorithm option deletes the algorithm, the associated output data nodes and any downstream algorithms that rely on the output of the deleted algorithm. The View XML Section option is a shortcut to the *Script XML Viewer*. When selected this option will open the XML viewer and will highlight the algorithm section associated with the node selected in the Script Tree Viewer. **One note on saving a script, scripts should be saved to the *script* directory inside MeV's *Data* directory. This location of the script ensures that when the script is loaded the files supporting script validation are located and used to validate script integrity.**

Script XML Viewer

The Script XML Viewer is a text rendering of the script, as it would appear when saved to an output file. The main purpose of the viewer is to get a view of the script during script creation and to review parameter value selections for particular algorithms. When the XML viewer is opened via the Script Tree Viewer the selected algorithm in the XML viewer is highlighted in light green as shown in the Script XML Viewer figure. Script lines are selected by clicking on the row number displayed on the left side of the viewer. If the selected line corresponds to a parameter key-value pair, then the *Edit* menu option will be enabled so that the value can be altered.

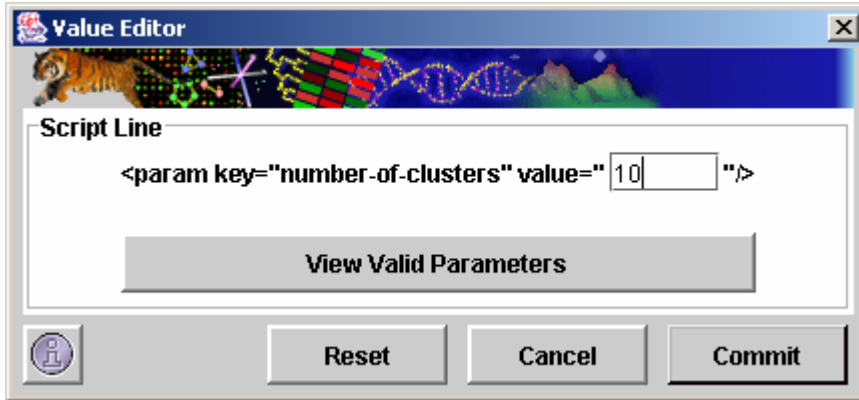


12.9. Script XML Viewer, highlighted algorithm and script line selected.

Editing Parameter Values within the XML Viewer

Limited editing capabilities are available in the Script XML Viewer. If the selected script line is a parameter line then hitting the *Edit* menu option will display an input dialog to permit altering the value of the parameter. Keep in mind that caution should be used when changing script parameters so that downstream algorithms that are dependent on results are still valid. The parameter input dialog only permits alteration of the parameter's value. A button on the dialog, labeled *View Valid Parameters* will produce a table of possible parameters for the algorithm being modified. The list contains parameter names, parameter types,

optional constraints (value limits), and whether the parameter is required in all cases or if it is dependent on other parameter selections.



12.10. Script Parameter Value Editor

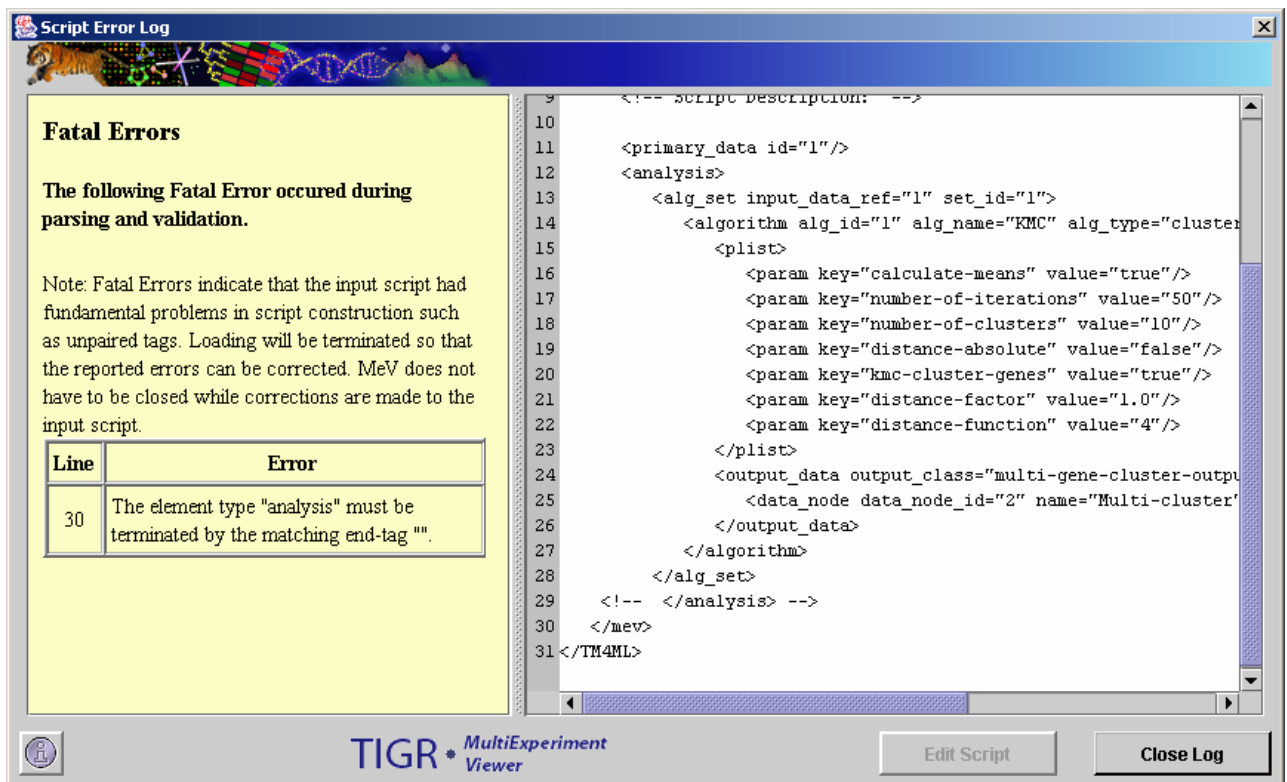
Key	Value Type	Min	Max	Required
number-of-clusters	int	1		Always
number-of-iterations	int	1		Always
calculate-means	boolean			Always
kmc-cluster-genes	boolean			Always
distance-factor	float			Always
distance-absolute	boolean			Always
distance-function	int	0	9	Always
hierarchical-tree	boolean			Dependant
method-linkage	int	0	2	Dependant
calculate-genes	boolean			Dependant
calculate-experiments	boolean			Dependant

12.11. Script Algorithm Parameter Table

Loading a Script

Loading of saved scripts is done by selecting *Load Script* from the File menu of the Multiple Array Viewer. A file selection dialog will automatically launch to prompt for the selection of a script file. Scripts should be stored in the script directory of MeV's data directory to ensure proper validation.

During script loading several levels of script validation occur. *Fatal Errors* (usually malformed XML), *Validation Errors* (script that does not match the Document Type Description (DTD)), *Parser Warnings*, and algorithm *Parameter Errors* (missing required parameters, parameter type mismatch, or parameter out-of-bounds errors) are caught during the validation. If multiple validation errors exist, all will be reported. All validation errors are reported in a Script Error Log dialog. The Error Log initially lists the errors and indicates a line number for each error. The *Edit Script* button launches an XML viewer that can be modified and saved to address the errors. Once the dialog is closed, the script should be re-loaded using the File menu to begin a fresh script loading and validation.

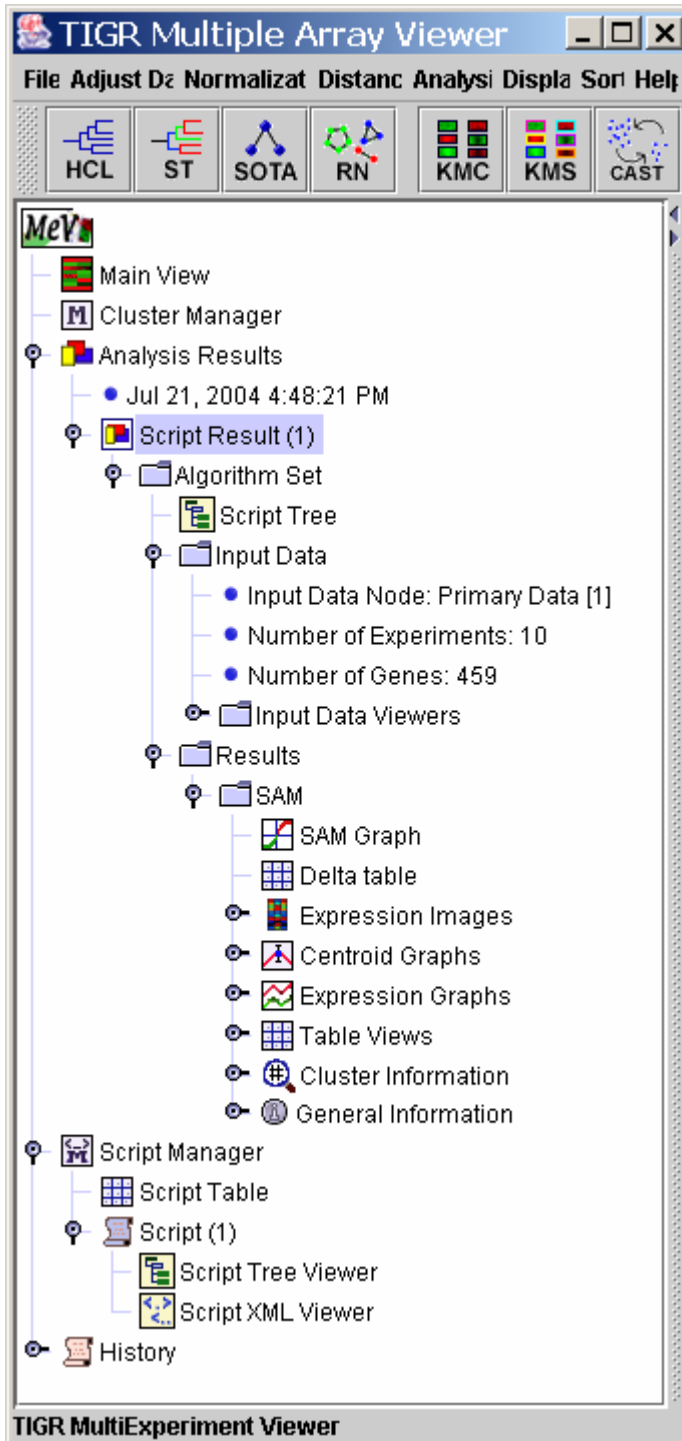


12.12. Script Error Log (with XML editor opened)

Running a Script

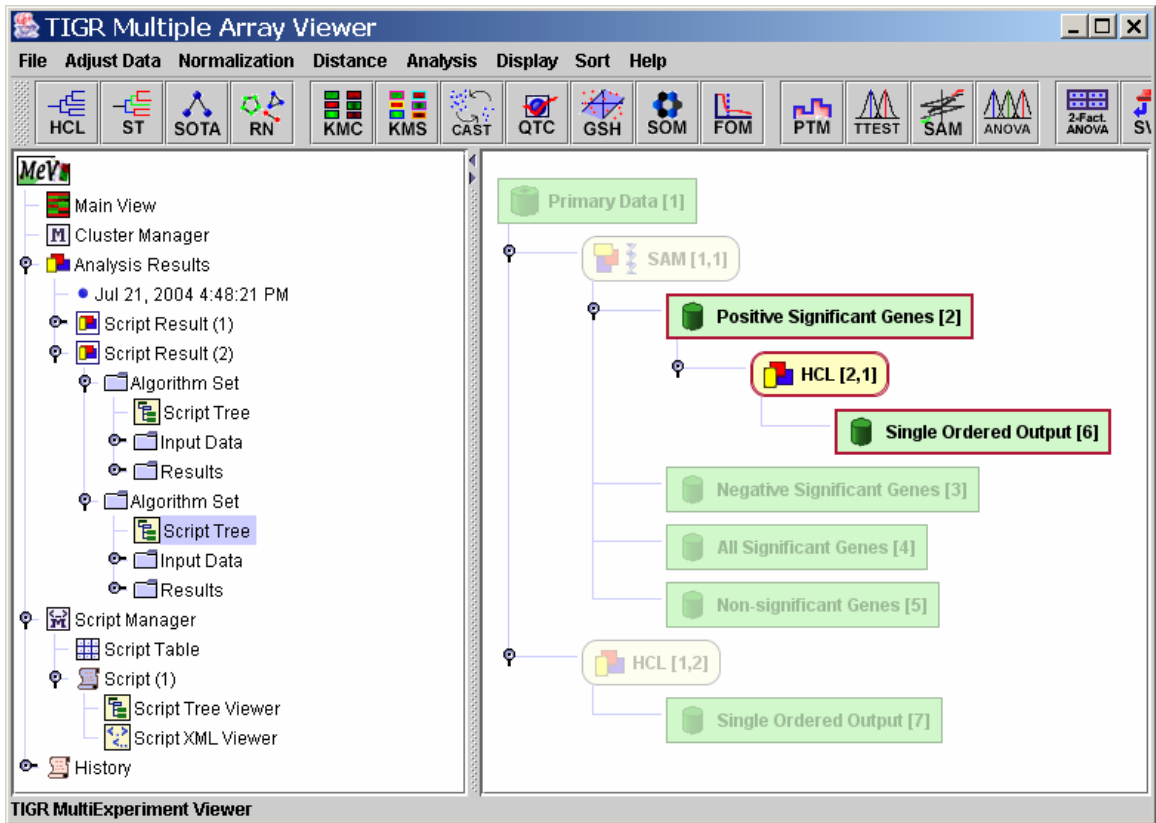
Script execution can be initiated from the Script Table viewer, the Script Tree Viewer, or the Script XML Viewer by using the right click menu option, *Execute Script*. The script is logically split up into units called *Algorithm Sets* that represent one or more algorithms sharing a common input data source. The output results are grouped into algorithm sets and since each algorithm set has a unique input data set that can be a subset of the loaded data, an input data node is used to display the input data for the algorithm set. The figure displaying script analysis

output clearly shows the expected output from a script containing one algorithm (SAM). The *Input Data* node shows the number of experiments and the number of genes as well as three cluster viewers. The *Script Tree* node in the output for an algorithm set helps to orient the researcher as to which part of the script falls within the enclosing algorithm set. The algorithm set, input data, attached algorithms and the result data nodes, are highlighted while other script nodes are semi transparent.



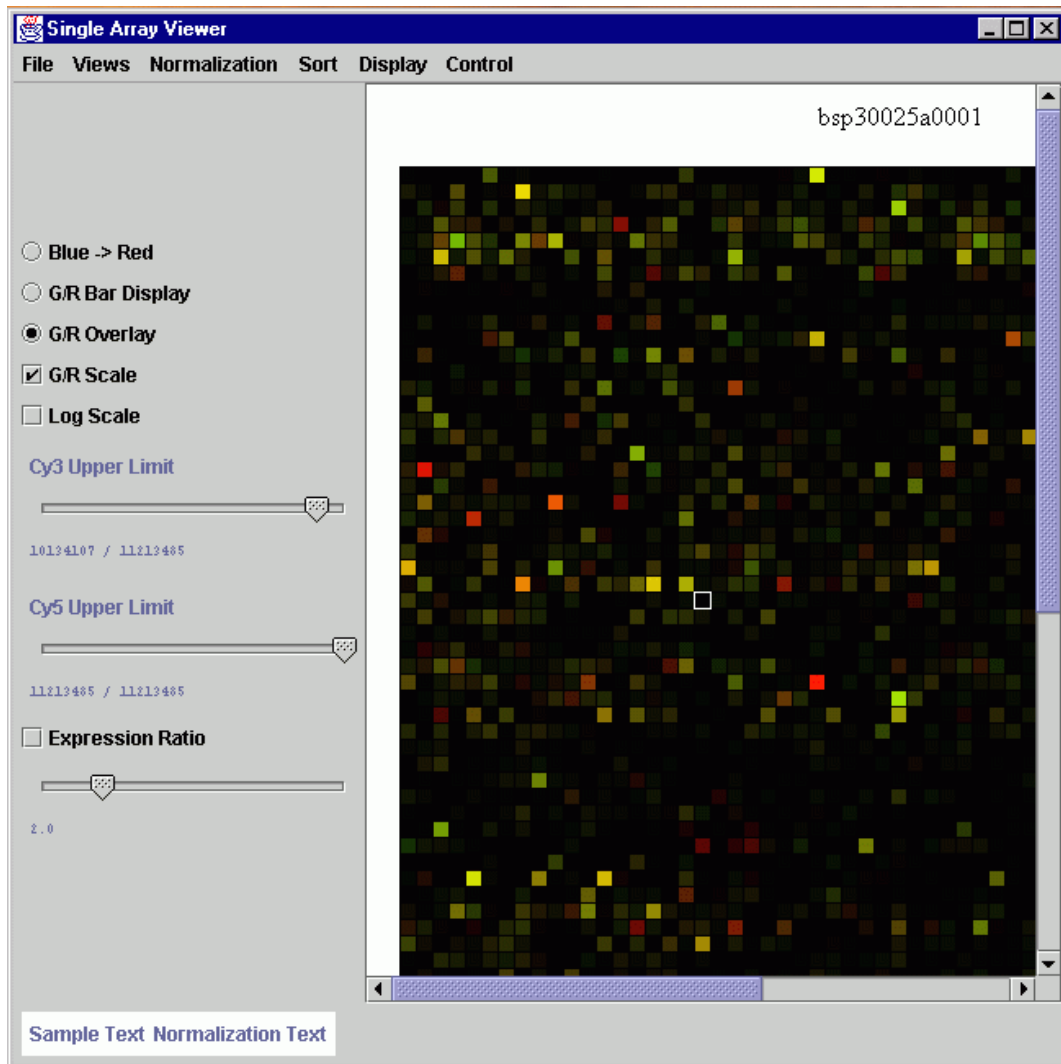
12.13. Script Output Nodes on the Result Tree (Single SAM run)

If an algorithm fails to produce a data node that is a source data node for another algorithm set then that algorithm set (using the null input) is aborted and an empty node with a text label indicating empty source data is displayed.



13. Working with the Single Array Viewer

- 13.1. The Single Array Viewer displays one slide at a time. Open a Single Array Viewer by choosing *New Single Array Viewer* from the *File* menu in the MeV main toolbar. Once a Single Array Viewer window is open, use its *File* menu (different from the main menu bar's *File* menu) to load a slide. *Open Experiment From File* loads a flat file (in .tav format) and *Open Experiment From DB* loads array data from a relational database using several stored procedures. If the user selects the former option, an open file dialog will be displayed, prompting the user to select a flat file to load. The latter option displays a list of experiment names and then analysis ids from the database. By selecting an experiment name (which represents a labeled slide) and analysis id (which represents a particular image analysis session) a unique dataset is specified.

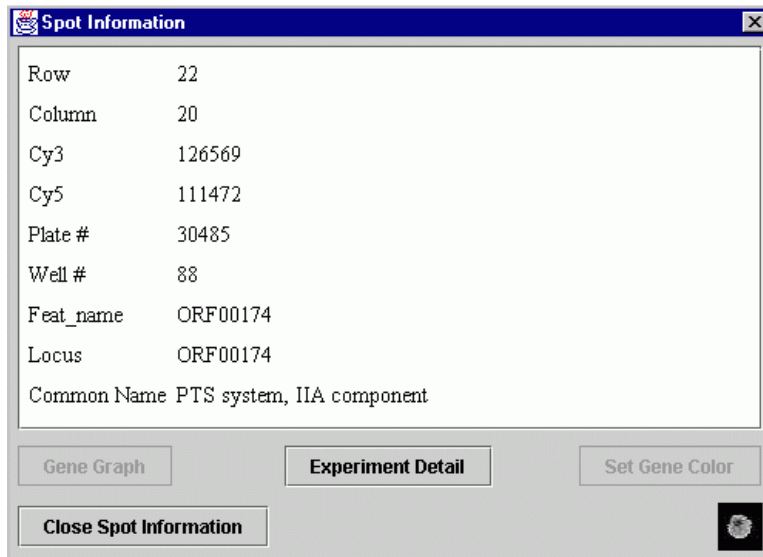


13.1.1. Single Array Viewer

- 13.2. Once a slide has been loaded, a representation of the slide will be displayed in the window. Each colored rectangular bar (an element) corresponds to a spot on the array, and is in the same position. The display can be changed using the same menus as in the Multiple Array Viewer. One display unique to the Single Array Viewer is the false-color display. It shows the two channels in separate areas,

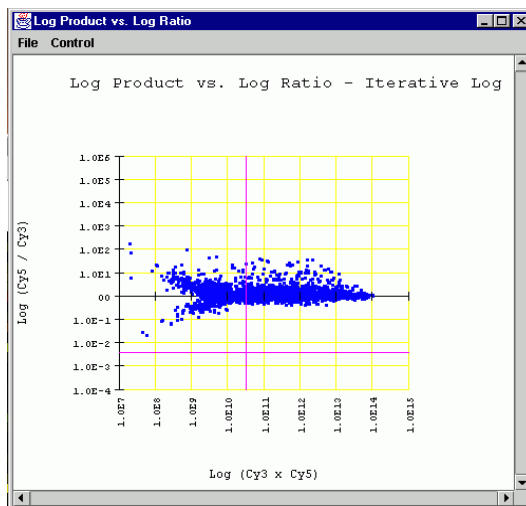
where elements are colored based on a scale where low intensities are dark and blue, while high intensities are bright and red.

- 13.3. Clicking on a spot will display a dialog that shows detailed information about the target spot. This information includes the row and column of the spot, intensity values and the extra fields specified in the preference file. Other elements of this dialog include a compressed version of the actual spot image (where available) and a graph showing the expression levels of the gene across multiple experiments.



13.3.1. Spot information

- 13.4. Several graphs can be created using the *View Graph* item in the *Views* menu. The choices include scatter plots of the two intensities, ratio histograms and a log ratio vs. log product graph. These graphs are displayed in a separate window, and mention the dataset and normalization scheme that spawned them.



13.4.1. Graph view from Single Array Viewer.

- 13.5. A sub array can be created to view just those genes that are still displayed after applying the expression ratio. Select the *View Sub Array* item from the *Views*

menu to create a new Single Array Viewer window with the selected elements. These elements will be rearranged to eliminate gaps in the display.

- 13.6. The View Region item in the *Views* menu displays a dialog for inputting the coordinates of a metablock. A new Single Array Viewer is created, using the targeted metablock as the new dataset. In this way, the user can focus on a particular defined area of the slide.
- 13.7. The elements of the array can be sorted by location (row and column), ratio or any of the additional fields that were specified in the preferences file. These sort options are available in the *Sort* menu. *Sort by location* is the default sorting method.
- 13.8. Differentially expressed genes can be identified by checking the *Expression Ratio* checkbox on the panel on the left side of the window. The slider below the checkbox controls the expression ratio used to determine differential expression. When the checkbox is checked, only those genes which have one intensity value greater than the other by a factor greater than or equal to the expression ratio will be displayed. Other genes will be blacked out. For example, a ratio of 2.0 will exclude genes where the two intensities do not differ by a minimum factor of two.
- 13.9. The array representation can be saved as an image file or sent to a printer. Select *Save Image* from the *File* menu and choose a name and graphic format in the dialog that appears. To print the image, select *Print Image* from the *File* menu and set up the printer dialog.
- 13.10. To write a flat file as output, select the *Generate Report* item from the *File* menu. A save file dialog will be displayed, prompting for a name for the report file. This report will contain the data for each spot that is currently visible in a tab delimited format similar to the .tav format. The first lines of the report contain the name of the original input file and the normalization method used.

Report for SlideFile: bsp30025a001
Normalization: Iterative Log Thresholds: 0.25/0.75

Row	Column	MetaRow	MetaColumn	SubRow	SubColumn	Cy3	Cy5	Cy5/Cy3	Plate #	Well #	Feat_name	Locus	Common Name
9	1	2	1	2	1	133186	31977	0.24009280254681423	30486	2	ORF00238	ORF00238	conseru
13	1	2	1	6	1	86283	295625	3.426225328280194	30490	2	ORF02650	ORF02650	phospho
33	1	5	1	5	1	18732	71427	3.8131005765534915	30489	5	ORF00171	ORF00171	pyruvate
19	2	3	1	5	2	817686	8912752	10.899968936731215	30489	27	ORF01391	ORF01391	DNA pro
30	2	5	1	2	2	50196	1818544	36.228862857598216	30486	29	ORF02553_K0	ORF02553_K0	blank
51	2	8	1	2	2	137002	1512402	11.039269499715333	30486	32	ORF02341	ORF02341	conseru
59	2	9	1	3	2	382220	2104471	5.505915441368845	30487	33	ORF00527	ORF00527	conseru
78	2	12	1	1	2	108151	28170	0.2604691588612218	30485	36	ORF02404_K0	ORF02404_K0	blank
2	3	1	1	2	3	5000	105925	21.185	30486	49	ORF02654	ORF02654	vanZ protein, putative
18	3	3	1	4	3	30371	219357	7.222580751374666	30488	51	ORF01030	ORF01030	blank
23	3	4	1	2	3	419257	1583533	3.776998356616586	30486	52	ORF01653	ORF01653	hypothet
34	3	5	1	6	3	302230	3137163	10.380051616318697	30490	53	ORF02305	ORF02305	membran
48	3	7	1	6	3	25636	84253	3.2865111561866125	30490	55	ORF00579	ORF00579	5-methyl
15	4	3	1	1	4	18084	61818	3.418380089183809	30485	75	ORF00896	ORF00896	blank
59	4	9	1	3	4	34997	107820	3.08803550018573	30487	81	ORF01589	ORF01589	hypothet
78	4	12	1	1	4	64202	267589	4.167923117659886	30485	84	ORF00870	ORF00870	hypothet
82	4	12	1	5	4	267145	1376204	5.151524453012409	30489	84	ORF00931_K0	ORF00931_K0	blank
20	5	3	1	6	5	14240	116249	8.163553370780516	30490	99	ORF01914	ORF01914	neuramin
22	5	4	1	1	5	1674688	7643260	4.563990426873543	30485	100	ORF02196_K0	ORF02196_K0	blank
25	5	4	1	4	5	33697	181702	5.3922207623824075	30488	100	ORF01757	ORF01757	pyridin
27	5	4	1	6	5	8252	51364	6.224430041105187	30490	100	ORF00120	ORF00120	hypothet
64	5	7	1	5	5	57601	1005437	17.455200430548082	30489	103	ORF00026_K0	ORF00026_K0	blank
47	5	10	1	1	5	200872	913185	4.546103986618344	30485	106	ORF02343_K0	ORF02343_K0	blank
72	5	11	1	2	5	78964	25220	0.31938604933893927	30486	107	ORF02274	ORF02274	primase
8	7	2	1	1	7	77627	458243	5.9031393716103935	30485	146	ORF00525	ORF00525	Blp2 pr
14	7	2	1	7	7	30505	463043	15.179249303392886	30491	146	ORF00508	ORF00508	blank
51	7	8	1	2	7	147921	456389	3.0853563726583784	30486	152	ORF01180	ORF01180	glutami
71	7	11	1	1	7	217842	1859668	8.536774359398096	30485	155	ORF02346_K0	ORF02346_K0	blank
32	8	5	1	4	8	386199	1566320	4.055732925253561	30488	173	ORF01737	ORF01737	conseru
58	8	9	1	2	8	14278	52618	3.6852500350189104	30486	177	ORF02505	ORF02505	L-Fucul
59	8	9	1	3	8	379532	1424193	3.752497813096129	30487	177	ORF00506_K0	ORF00506_K0	blank
72	8	11	1	2	8	16334	134138	8.212195420595078	30486	179	ORF00066_K0	ORF00066_K0	blank
6	9	1	1	6	9	49538	154453	3.1178691105817755	30490	193	ORF02476	ORF02476	conseru

13.10.1. Flat file output from Single Array Viewer

14. Appendix: File Format Descriptions

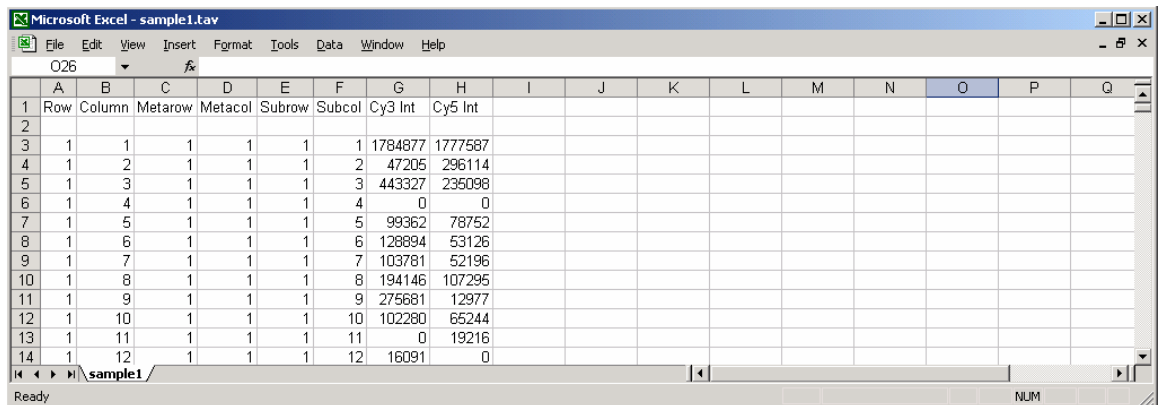
14.1. TAV Files

The original TAV (TIGR ArrayViewer) file type was an eight-column, tab-delimited text format developed at TIGR for the purposes of storing the intensity values of the spots on a single slide. It is written out by the program TIGR Spotfinder and contains one row for each spot. The first six columns of the file contain positional data for the spots and are followed by two columns of intensity data.

These eight columns are required by MeV for display and analysis of experimental data. Optional columns can contain flags, annotation, Genbank numbers, etc. It is the variability of .tav files caused by these optional columns that make Preferences files necessary (see section 15). Optional columns can be used to sort the spots in the *Main View*, by choosing the appropriate column from the *Sort* menu.

A flag is simply a letter code corresponding to a description of the spot:

- A – 0 non-saturated pixels in the spot
- B – 0-50 non-saturated pixels in the spot
- C – 50 or more non-saturated pixels in the spot
- X – spot is rejected, due to spot shape and intensity relative to background
- Y – background is higher than spot intensity
- Z – spot not detected by Spotfinder.



1	Row	Column	Metarow	Metacol	Subrow	Subcol	Cy3 Int	Cy5 Int									
2																	
3	1	1	1	1	1	1	1784877	1777587									
4	1	2	1	1	1	2	47205	296114									
5	1	3	1	1	1	3	443327	235098									
6	1	4	1	1	1	4	0	0									
7	1	5	1	1	1	5	99362	78752									
8	1	6	1	1	1	6	128894	53126									
9	1	7	1	1	1	7	103781	52196									
10	1	8	1	1	1	8	194146	107295									
11	1	9	1	1	1	9	275681	12977									
12	1	10	1	1	1	10	102280	65244									
13	1	11	1	1	1	11	0	19216									
14	1	12	1	1	1	12	16091	0									

14.1.1. A TAV file containing only the required fields

1	Row	Column	Metarow	Metacol	Subrow	Subcol	Cy3 Int	Cy5 Int	Flag 1	Flag 2	Ratio	Plate#	Well#	clone_id	amplified	GB#	TC#	Com_name
2																		
3	1	1	1	1	1	1	1784877	1777587	B	B	3903	574	73	49570	0	M86720	null	null
4	1	2	1	1	1	2	47205	296114	C	C	0.925	491	265	4035	2	AA126115	THC1082463	Chloride condu
5	1	3	1	1	1	3	443327	235098	C	C	0.9532	494	85	5124	1	AA598884	THC1058838	NADH-ubiquinc
6	1	4	1	1	1	4	0	0	Y	X	0.8092	497	277	6261	1	R11499	null	null
7	1	5	1	1	1	5	99362	78752	C	C	0.9086	501	73	7741	1	R16600	null	null
8	1	6	1	1	1	6	128894	53126	C	C	0.7329	504	265	8916	1	R06746	THC1119429	unnamed prote
9	1	7	1	1	1	7	103781	52196	C	C	0.8043	507	85	10026	1	AA009791	null	null
10	1	8	1	1	1	8	194146	107295	C	C	0.9263	510	277	11226	1	AA412691	THC1118802	CCAAT-binding
11	1	9	1	1	1	9	275681	12977	C	C	0.8931	514	73	12708	1	T89094	THC1067707	RGP4; regulat
12	1	10	1	1	1	10	102280	65244	C	C	1.0188	517	265	13908	1	AA457232	THC1134142	unnamed prote
13	1	11	1	1	1	11	0	19216	X	C	0.8443	520	85	15018	1	N49263	null	null
14	1	12	1	1	1	12	16091	0	C	X	1.2875	523	277	16218	2	AA443940	null	null

14.1.2. A TAV file with several extra fields

14.2. Tab Delimited, Multiple Sample Files (TDMS files)

TDMS files encapsulate the expression data from multiple samples into a single tab delimited file. Each sample represented in the file will have a single dedicated column that contains the expression data for that sample. Each row, below the header rows, represents information relating to a particular spot on the slide.

The following sections describe the format of the file in detail. The image following the description contains color coded sections that relate to each of the distinct areas of the file.

Header Rows (yellow, light blue, and cyan, top 4 rows in the example)

TDMS files must contain one or more header rows. These header rows must be at the top of the file. The first header row is used primarily to contain the default sample name for each sample contained in the file. The first header row also contains descriptive gene annotation field names (yellow) over the columns dedicated to gene annotation. Additional header rows may be present and each additional header row contains additional sample annotation.

Gene Annotation Columns (green (annotation), yellow (annotation field names))

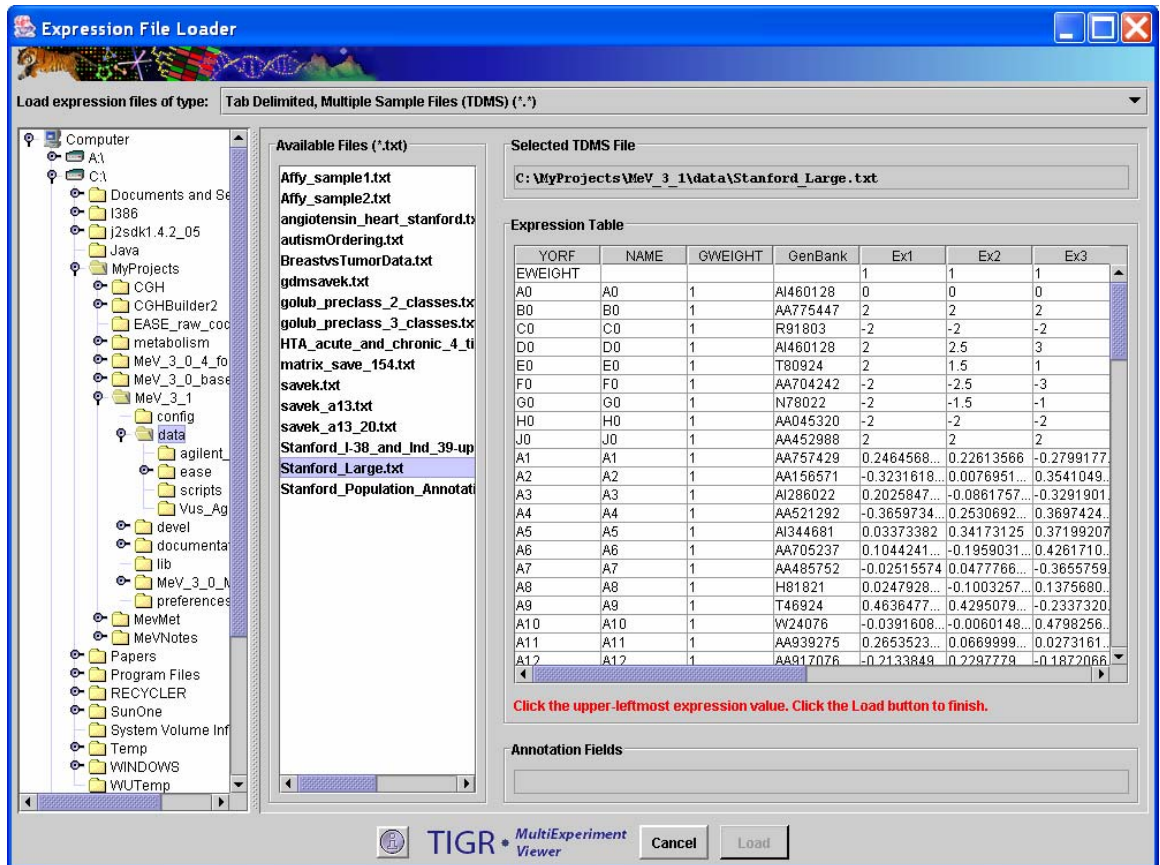
The TDMS format permits any number of annotation columns, which must occupy the left most columns of data in the file. Each of these annotation columns contains annotation corresponding to the spot represented by that row of the file. Each of the annotation columns has a label in the top header row to indicate the annotation type.

Expression Data (uncolored section with numeric data)

The Expression data is arranged such that there is one column for each sample represented in the file. The position of the expression data column for a particular sample is beneath the header's sample label for that sample. See the TDMS file loader figure below as an example file displayed within the preview window.

NAME	GenBank	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
	Treatment	placebo	placebo	placebo	placebo	placebo
	Patient ID	101	102	103	104	105
	Gender	F	F	M	M	M
A1	AA757429	0.246457	0.226136	-0.27992	0.278945	0.416614
A2	AA156571	-0.32316	0.007695	0.354105	-0.32402	-0.40812
A3	AI286022	0.202585	-0.08618	-0.32919	-0.38246	-0.29769
A4	AA521292	-0.36597	0.253069	0.369742	-0.14589	0.032014
A5	AI344681	0.033734	0.341731	0.371992	-0.11216	-0.48229
A6	AA705237	0.104424	-0.1959	0.426171	0.275234	-0.07836
A7	AA485752	-0.02516	0.047777	-0.36558	-0.15541	0.401885
A8	H81821	0.024793	-0.10033	0.137568	-0.10322	0.287513
A9	T46924	0.463648	0.429508	-0.23373	0.389819	0.204742
A10	W24076	-0.03916	-0.00601	0.479826	0.095263	0.338572
A11	AA939275	0.265352	0.067	0.027316	-0.09708	0.463785
A12	AA917076	-0.21338	0.229778	-0.18721	0.052883	0.119597
A13	T41038	-0.43006	0.057859	0.253913	-0.45023	-0.34688

14.2.1. Tab Delimited, Multiple Sample File (TDMS) Format Image



14.2.2. Loading Tab Delimited, Multiple Sample Files (TDMS Format)

14.3. GenePix Files

Microarray data in the GenePix file format can be easily converted to tab file format using TIGR's GenePix Converter application. This program is freely available from the TIGR website as part of the MADAM package). The latest version of the converter can convert multiple files in batch mode. See the help files for the GenePix Converter for more details. Files converted when the "Keep all GenePix data" box is checked will be extremely large and may cause MeV to run slowly. Leaving this box unchecked will cause only the data required by MeV to be written to the output tab file.

14.4. MEV Files

A MultiExperimentViewer or .mev file is a tab-delimited text file that contains coordinate and expression data for a single microarray experiment. A single header row is required to precede the expression data in order to identify the columns below. With the exception of optional comment lines, each remaining row of the file stores data for a particular spot/feature on the array.

MeV and other TM4 software tools will consider comment lines non-computational. A comment line must start with the pound symbol '#', and can be included anywhere in the file. If the pound symbol is the first character on a line, the entire line (up to the newline character '\n') will be ignored by the software tool.

The mev files created at TIGR will typically contain at least one comment at the top of the file with the following information. This information is optional to non-TIGR users. The format and fields contained within these comments are subject to change.

version	Version number based on revisions of expression data
format_version	The version of the .mev file format document
date	Date of file creation or update
analyst	Owner or the person responsible for creating the file
analysis_id	<i>id</i> from the <i>analysis</i> table that corresponds to this set of expression values
slide_type	<i>slide_type</i> from the <i>slide_type</i> table that this array is based on
input_row_count	Number of rows of expression (eg. non-header) data in input files
output_row_count	Number of rows of expression (eg. non-header) data in this file
created_by	Software tool used to create the file
description	Common name or other details about the experiment

An example of the leading comments:

```
# version: V1.0
# format_version: V4.0
# date: 10/06/2004
# analyst: aisaeed
```

```

# analysis_id: 10579
# slide_type: IASCAG1
# input_row_count: 32448
# output_row_count: 32448
# created_by: TIGR Spotfinder 2.2.3
# TIFF files processed: gpc30025a_532_nm.tif, gpc30025a_635_nm.tif
# description: Tumor type comparison
# This is the 4th experiment in a series of 20 to identify tissue-specific genes.

```

The header row consists of the field names for each subsequent row in this file (with the exception of comment lines). A minimum of seven columns must be present, and these must use a set of specifically named headers. Any number of additional columns may be included. The seven required column headers are:

UID	Unique identifier for this spot
IA	Intensity value in channel A
IB	Intensity value in channel B
R	Row (slide row)
C	Column (slide column)
MR	Meta-row (block row)
MC	Meta-column (block column)

As of version 4.0 of this file format the IA and IB columns can be substituted with MedA and MedB. The new requirement is that at least one integrated intensity (IA, IB, etc.) **or** one median (MedA, MedB, etc.) value be reported for each channel in the microarray. For example, a two channel microarray .mev file would require either IA and IB **or** MedA and MedB.

MedA	Median intensity in channel A
MedB	Median intensity in channel B

The mev files created at TIGR may use one of the following formats for the header row, depending on the origin of the mev file. The non-required columns (i.e. anything after the 7th column) may be rearranged and their names are subject to change at this time.

1) Database created mev file:

```

UID \t IA \t IB \t R \t C \t MR \t MC \t SR \t SC \t FlagA \t FlagB \t SAA \t SAB \t
SFA \t SFB \t QCS \t QCA \t QCB \t BkgA \t BkgB

```

UID	Unique identifier for this spot
IA	Intensity value in channel A
IB	Intensity value in channel B
R	Row (slide row)
C	Column (slide column)
MR	Meta-row (block row)
MC	Meta-column (block column)
SR	Sub-row
SC	Sub-column

FlagA	<i>TIGR Spotfinder</i> flag value in channel A
FlagB	<i>TIGR Spotfinder</i> flag value in channel B
SAA	Actual spot area (in pixels) in channel A
SAB	Actual spot area (in pixels) in channel B
SFA	Saturation factor in channel A
SFB	Saturation factor in channel B
QC	Cumulative quality control score
QCA	Quality control score in channel A
QCB	Quality control score in channel B
BkgA	Background value in channel A
BkgB	Background value in channel B

2) Spotfinder created mev file:

```
UID \t IA \t IB \t R \t C \t MR \t MC \t SR \t SC \t FlagA \t FlagB \t SAA \t SAB \t
SFA \t SFB \t QCS \t QCA \t QCB \t BkgA \t BkgB \t SDA \t SDB \t SDBkgA \t
SDBkgB \t MedA \t MedB \t AID
```

UID	Unique identifier for this spot
IA	Intensity value in channel A
IB	Intensity value in channel B
R	Row (slide row)
C	Column (slide column)
MR	Meta-row (block row)
MC	Meta-column (block column)
SR	Sub-row
SC	Sub-column
FlagA	<i>TIGR Spotfinder</i> flag value in channel A
FlagB	<i>TIGR Spotfinder</i> flag value in channel B
SAA	Actual spot area (in pixels) in channel A
SAB	Actual spot area (in pixels) in channel B
SFA	Saturation factor in channel A
SFB	Saturation factor in channel B
QC	Cumulative quality control score
QCA	Quality control score in channel A
QCB	Quality control score in channel B
BkgA	Background value in channel A
BkgB	Background value in channel B
SDA	Standard deviation for spot pixels in channel A
SDB	Standard deviation for spot pixels in channel B
SDBkgA	Standard deviation of the background value in channel A
SDBkgB	Standard deviation of the background value in channel B
MedA	Median intensity value in channel A
MedB	Median intensity value in channel B
AID	Alternative ID

The first seven fields (UID, IA, IB, R, C, MR and MC) are required as specified above.

This flexible format allows users to track slide-specific data of interest, such as background, spot size and alternate intensities without requiring them of all users or adopting a limited ‘vocabulary’ of field names. This header row serves to identify the required and additional data columns. UID must be the left-most column in the mev file. Other columns do not need to be present in a fixed order.

For mev files generated at TIGR, the UIDs may be of the form: *database_name:spot_id* (eg. cage:20238). For any given microarray database, the *id* field in the *spot* table will be unique. The combination of database and *spot_id* will therefore uniquely identify any spot on any array created at TIGR. It is important to note that this is not enough information to distinguish between spots in the same location on two slides of the same *slide_type*, as this would typically require an *analysis_id*. Since annotation data is based on *slide_type*, it is not necessary to make this distinction, as all slides of a given type will use the same annotation file.

The AID column will usually contain an incremental sequence of numbers starting at 1. These can be used to return the file to the original sorted order and can function as a unique row identifier if necessary.

Applications that generate files of expression data (commonly in tav format) by retrieving records from the database access the *spot* table. *TIGR Spotfinder*, *Midas* and *Madam* are all capable of generating UIDs of the form described above in addition to the typical coordinate and intensity data.

mev files are required to end with the extension ‘.mev’. At this time there are no further naming conventions for mev files.

14.5. Annotation Files

An annotation file is a tab-delimited text file containing annotation data for a specific *slide_type*. mev files can be associated with an annotation file only if both types of files are based on the same *slide_type*. The keys to this association are the unique ids in both files. Rows of mev and annotation files can be associated with each other if the unique ids are identical. A single header row is required to precede the annotation data in order to identify the columns below. Each remaining row of the file stores annotation data for a particular spot/feature on the array.

Annotation files may contain any number of non-computational comment lines. These lines, starting with ‘#’, will be treated identically to comment lines in mev files, and should precede the header row.

Annotation files created at TIGR will use UIDs that match the format used in the mev files, most likely *database_name:spot_id*. The structure of each annotation file is detailed below.

The header row consists of headers that identify each column of data. Each subsequent row of the file stores data for a particular spot/feature on the array. The annotation files created at TIGR will typically contain at least one comment at the top of the file with the following information:

version	Version number based on revisions of annotation data
format_version	The version of the .mev file format document
date	Date of file creation or update
analyst	Owner or the person responsible for creating the file
created_by	Software tool used to create the document
gi_version	Version of the Gene Indices (or db?) that produced this annotation data
slide_type	<i>type</i> from the <i>slide_type</i> table that this array is based on
output_row_count	Number of rows of annotation (eg. non-header) data
description	Common name or other details about the experiment

An example of the leading comments:

```
# version: V3.0
# format_version: V4.0
# date: 04/20/2004
# analyst: jwhite
# created_by: Database script
# gi_version: 3.0
# slide_type: IASCAG1
# output_row_count: 32448
# description: Standard annotation file
```

The header row consists of the field names for each subsequent row in this file. Only the UID field is required. It must be the first field present and it must be named 'UID'. Any number of additional fields may be included. Annotation files created at TIGR will always contain the following columns:

UID	unique identifier for this line of annotation
R	row (slide row)
C	column (slide column)

The remaining fields may vary, and a standard set has yet to be determined. Such a list will be published on a future date. R and C have been included to allow for manual alignment of the mev and corresponding annotation files in the event that the mev files were not generated in a traditional manner (ie. using *Madam*, etc.).

Some varieties of annotation files follow. The format may vary depending on the purpose of the file:

```
UID \t R \t C \t FeatN \t GBNum \t TCNum \t ComN \t ...
UID \t R \t C \t GeneN \t Rxn \t PathwayN \t ...
UID \t R \t C \t FeatN \t End5 \t End3 \t ChrNum \t ...
```

Of course, it would be possible to combine the fields of these files, or add fields that have not been mentioned here. The goal is to keep the annotation flexible and the processing seamless.

There are not any naming conventions for annotation files at this time. If such a standard is introduced in the future, it will be detailed here.

15. Appendix: Preferences Files

Preferences files store information about a data input file's format. The number of file format variations, including the Stanford file format and various flavors of TAV file format, make it necessary to provide MeV with an instruction set for reading those files. Preferences files are human-readable, tab-delimited text and contain the information MeV needs to understand the data layout in a microarray data file. Three sample preference files are included with the MeV installation, serving as templates for TAV files (section 14.1), Stanford files (section 14.1.2) and cluster files. Use a text editor such as Notepad to customize one of these files for a particular file type. The names of Preferences files should always end with the word "Preferences" and have no period (.).

Most lines in the preferences file are preceded by a double-slash (//) indicating that the following text is part of a comment and will be ignored by MeV. These comments contain descriptions of the parameters listed below them. Lines containing a parameter have no double-slash and consist of a parameter description followed by a tab and the parameter value. Only the parameter value should be altered when customizing a preferences file. It is also extremely important that the label and value are separated by **one** tab character. Below is a list of Preferences file parameters.

Input Preference

Sets the method with which data is entered into TIGR MultiExperimentViewer

'Database' – Connects to database and loads slides from there.

'File' – Loads slides from flat files as default, but can also connect to database.

'Only File' – Loads slides from flat files only. No database connection.

Database Server Name

Sets the path to the database. This is not necessary if *Input Preference* is set to 'Only File'.

For sybase, use the following string:

jdbc:sybase:Tds:<yourhost>:<yourport>

For oracle, use the following string:

jdbc:oracle:thin:<yourhost>:<yourport>

Database Names

Sets the list of databases to choose arrays from. All stored procedures must be accessible from all of the available databases on this list. The database names should be separated by a colon, ':'. This is not necessary if *Input Preference* is set to 'Only File'.

Element Info

Sets the number of row and column pairs and the number of intensities per element, separated by a colon ':'. For example, 3:2 would indicate that each element has three pairs of row and column values, such as (row, column meta_row, meta_column, sub_row, sub_column), and two intensities, such as (Cy3, Cy5). The input file must have the row and column pairs listed at the

beginning of the line, followed by the intensities and then by any additional fields such as common name or Genbank number (see below for details). This parameter is mandatory.

Headers

Sets the number of non-element row and column headers in the input file, separated by a colon ':'. For example, 2:4 would indicate the first two rows and first four columns of every input file are considered headers.

Unique ID

Sets the number of the column that represents the unique identifiers for each element. This column should not contain duplicate values, and every element should have a value in this column.

Spot Name

Set the number of the column that represents the name of each element. These are usually descriptive and human decipherable strings. These values do not need to be unique.

Additional Fields

Indicates the names of additional data fields (after row, column, cy3, cy5) to be stored and displayed. The field names should be separated by a colon ':'. If you have additional fields you want to have MeV process, each line of your input from the flat file or row returned from the database must have a number of additional columns equal to the number of additional fields. Each field name must be unique.

Algorithm Factory

The name of the Algorithm Factory implementation class

16. Appendix: Distance Metrics

MeV provides eleven distance metrics from the *distance* menu on the menu bar. While Euclidean Distance and Pearson Correlation are by far the most utilized metrics this appendix summarizes all available metrics. Note that in the following equations u and v are expression vectors of size m .

Euclidean Distance

Euclidean distance is perhaps the most familiar distance metric since it reflects the distance between two objects in space. The definition of Euclidean distance extends to as many dimensions as present in the expression vectors to be compared. Distances can range from 0 to positive infinity.

$$d(u, v) = \sqrt{\sum_{i=1}^m (u_i - v_i)^2}$$

Manhattan Distance

Manhattan distance (or ‘City-Block’ distance) describes the distance as the sum of the differences of each element pair or dimension. In two dimensions it is like ‘going up one and over three blocks’ to get to a destination in a city where one can not traverse a block diagonally. Distances can range from 0 to positive infinity.

$$d(u, v) = \sum_{i=1}^m (u_i - v_i)^2$$

Pearson Correlation

The Pearson Correlation and other related metrics are very commonly used to evaluate trends of expression over a set of conditions. This metric allows one to group trends or patterns irrespective of their overall level of expression. Two genes having different levels of expression but having ‘parallel’ expression patterns would be considered closely related. Values can vary from -1 to 1. Correlations near 1 indicate a strong positive correlation between the two vectors. Meaning that when one increases the other increases. Values closer to -1 indicate a negative correlation, when one vector has a relative increase in expression, the other vector has decreasing expression.

$$\rho(u, v) = \frac{\text{cov}(u, v)}{\sigma_u \sigma_v}$$

where

$$\text{cov}(u, v) = \frac{\sum_{i=1}^m (u_i - \bar{u})(v_i - \bar{v})}{(m-1)}$$

and

$$\sigma_w = \frac{\sqrt{\sum_{i=1}^m (w_i - \bar{w})^2}}{(m-1)}$$

Pearson Uncentered

This is a variation of the standard Pearson correlation in which the σ_w values computed are the standard deviation from zero rather than the mean intensity for that vector. The difference between this metric and the Pearson correlation is partially dependent on how different the mean expression is from zero.

Pearson Squared

This variant of the Pearson correlation performs Pearson uncentered and then squares the result. Using this metric will cause patterns that are strongly positively and strongly negatively correlated to possibly cluster together. Values for this metric range from 0 to 1.0.

Cosine Correlation

This metric produces values that range from -1 to 1 with values toward 1 indicating a strong positive relationship and values toward -1 indicating a strong negative relationship.

$$corr_{\cosine}(u, v) = \frac{\sum_{i=1}^m (u_i v_i)}{\sqrt{\sum_{i=1}^m u_i^2 \sum_{i=1}^m v_i^2}}$$

Covariance

Covariance can produce values which are unbounded. Values are not scaled by factors representing the variance within u and v. Covariance values for vectors with a strong positive relationship should be large and positive. If the two vectors have a strong negative relationship then the covariance will be negative and large. If the two vectors have little relationship then the terms of the summation above tend to vary between positive and negative and the covariance tends toward zero.

$$cov(u, v) = \frac{\sum_{i=1}^m (u_i - \bar{u})(v_i - \bar{v})}{(m-1)}$$

Average Dot Product

Average Dot Product can produce values which are unbounded. This metric has been used to compute similarity between expression vectors which have been normalized such that all elements range in value from 0 to 1 and each vector has a norm of 1.

$$u \bullet v = \frac{\sum_{i=1}^m (u_i v_i)}{m}$$

Spearman's Rank Correlation

Spearman's Rank Correlation, as the name implies, ranks the expression within each vector based on increasing expression level. Each vector in this manner is transformed to reflect the ordering of expression level. If two elements have exactly the same expression then both elements get assigned to the same level falling 0.5 levels above the next lower level. For example an expression vector containing five values (0.3, 1.2, 2.2, 1.1, 1.2) would have a ranking of (1.0, 2.5, 3.0, 2.0, 2.5). These ranking vectors are then used to compute the distance via:

$$corr_{Spearman}(u, v) = 1 - 6d / n(n^2 - 1)$$

where

$$d = \sum_{i=1}^m (x_i - y_i)^2 ; \text{ x and y are the ranking vectors corresponding to u and v.}$$

The spearman rank correlation makes no assumptions about the distribution of the data and the magnitude of expression becomes unimportant as ranking of expression level is used to determine the correlation.

Kendall's Tau

Kendall's Tau is a measure of correlation based on the tendency of the two vectors to vary in the same direction from one element to the next. In the case of gene expression vectors, for each observation of expression, the expression is compared to the previous measurement to determine if it is a relative increase or decrease in expression. If both expression vectors change in the same direction, both increasing or both decreasing, then the metric is incremented. If the expression vectors, for that element, change in expression in opposite directions, then the metric is decremented. The measure is finally scaled by the number of observations. This metric ignores the magnitude of the expression levels but purely looks at inflections in the expression patterns.

17. Appendix: MeV Script DTD

17.1. DTD

```
<!--<?xml encoding="UTF-8" ?> -->

<!ELEMENT TM4ML (midas?, dbi_controller?, mev?)>
<!ATTLIST TM4ML version CDATA #REQUIRED>

<!-- midas and dbiController place holders -->
<!ELEMENT midas EMPTY>
<!ELEMENT dbi_controller EMPTY>

<!ELEMENT mev (primary_data, analysis)>
<!ATTLIST mev version CDATA #REQUIRED>

<!ELEMENT analysis (alg_set+)>

<!ELEMENT alg_set (algorithm*)>
<!ATTLIST alg_set set_id CDATA #REQUIRED
                input_data_ref CDATA #REQUIRED>

<!ELEMENT algorithm (plist, mlist?, output_data)>
<!ATTLIST algorithm alg_id CDATA #REQUIRED
                    input_data_ref CDATA #REQUIRED
                    alg_name CDATA #REQUIRED
                    alg_type ( cluster | cluster-genes | cluster-
experiments | data-visualization | data-adjustment | cluster-selection
| data-normalization ) #REQUIRED>

<!ELEMENT plist (param*)>

<!ELEMENT param EMPTY>
<!ATTLIST param key CDATA #REQUIRED
                value CDATA #REQUIRED>

<!ELEMENT mlist (matrix*)>
<!ELEMENT matrix (element+)>
<!ATTLIST matrix name CDATA #REQUIRED
                type ( int-array | string-array | FloatMatrix )
#REQUIRED
                row_dim CDATA #REQUIRED
                col_dim CDATA #REQUIRED>

<!ELEMENT element EMPTY>
<!ATTLIST element row CDATA #REQUIRED
                col CDATA #REQUIRED
                value CDATA #REQUIRED>

<!ELEMENT output_data (data_node+)>
<!ATTLIST output_data output_class
                    ( single-output | multi-cluster-output | multi-gene-
cluster-output
                    | multi-expteriment-cluster-output | partition-output)
#REQUIRED>

<!-- single-output indicates that the result is one set
      (usually the result of normalization, filtering, or transform.

      multi-cluster-output is produced by many clustering algorithms and
```

represents multiple clusters in which each cluster contains vectors that are similar. There is no clear ordering of results. Generally to act on this output a selection algorithm should be used to select a cluster.

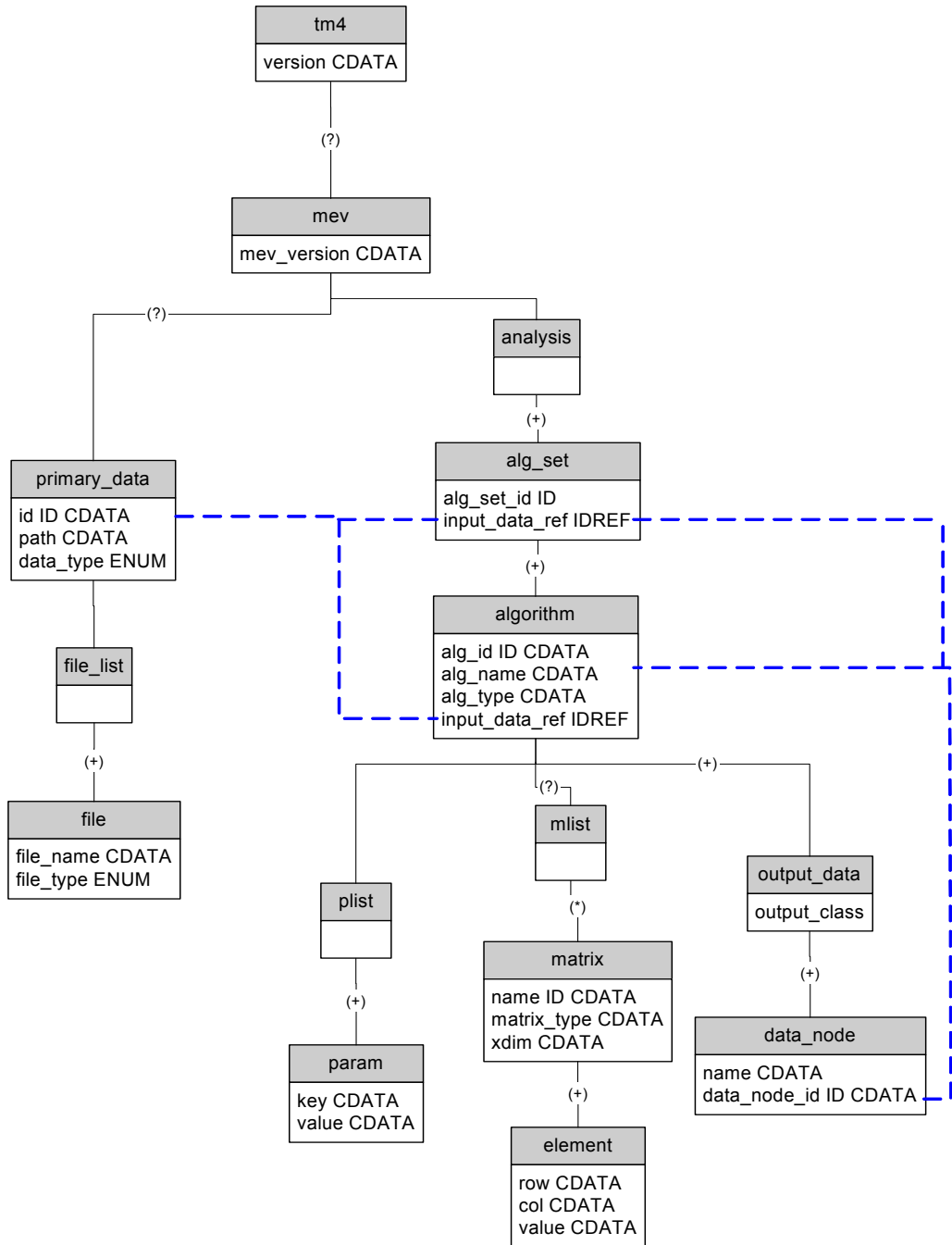
partition-output is a multi cluster output where the clusters are ordered and cluster members have a particular shared quality. e.g. Significant genes by a statistical algorithm, elements partitioned by classification algorithms. -->

```
<!ELEMENT data_node EMPTY>
<!ATTLIST data_node data_node_id CDATA #REQUIRED
                name CDATA #REQUIRED>

<!ELEMENT primary_data (file_list?)>
<!ATTLIST primary_data id CDATA #REQUIRED
                data_type ( mev | tav | stanford | gpr |
affy_abs |
                                affy_ref | affy_mean) #IMPLIED>
<!-- want an enumeration of data types (mev|tav|stanford|affy|gpr) -->
<!ELEMENT file_list (file+)>
<!ELEMENT file EMPTY>
<!ATTLIST file file_path CDATA #REQUIRED
                file_type ( data | annot | preference ) #REQUIRED>
```

17.2. DTD UML Schema

UML SCHEMA REPRESENTATION OF MEV SCRIPTING XML
 The dashed lines represent possible references from data id's to input_data IDREF.
 (?) = 0 or 1 element, (+) = 1 or more elements, (*) = 0 or more elements
 (unmarked links represent exactly one element)



18. License

Copyright © 1999-2005, The Institute for Genomic Research (TIGR).
All rights reserved.

This software is OSI Certified Open Source Software.
OSI Certified is a certification mark of the Open Source Initiative.

Please view the license (Artistic_License.pdf) in the root MeV directory.

19. Contributors

<u>Contributors</u>	<u>Affiliation</u>	<u>Contributions</u>
Alexander I. Saeed, Nirmal Bhagabati, John Braisted, John Quackenbush	The Institute for Genomic Research	System design, core and module development and implementation, project coordination, documentation, optimization, and usability assurance
Alexander Sturn, Zlatko Trajanoski	Institute of Biomedical Engineering, Graz University of Technology	Initial system architecture and viewer design, initial module development (HCL, KMC, SOM, PCA, SVM)
Mark Snuffin, Aleksey Rezantsev, Dennis Popov, Alex Ryltsov, Edward Kostukovich, Igor Borisovsky	DataNaut, Inc. (http://www.datanaut.com)	System architecture, module architecture, parallel processing system, primary module development (RN, TRN)
Stu Golub, Zaigang Liu, Jane Ruan, Minhas Siddiqui	Syntek Systems Corporation, Inc. (http://www.syntek.com)	Primary module development (CAST, FOM, GSH, GDM)
Wei Liang	The Institute for Genomic Research	Genomic Research Normalization algorithm development, TM4 website maintenance
Jerry Li, Vasily Sharov, Joe White, Mathangi Thiagarajan, Tracey Currier, Eleanor Howe	The Institute for Genomic Research	Additional development, software testing, documentation, support
Patrick Cahan, Tim McCaffrey	The George Washington University	Affymetrix data loader and filters, updated Affymetrix data loader, gene submission capability for the LOLA cluster repository.
Todd Peterson	National Center for Genome Resources	GeneX-Lite to MeV connectivity
Luke Somers	Fox Chase Cancer Center	QTC algorithm optimization and bug fix
Jim Johnson, Ernest Retzel	Center for Computational Genomics and Bioinformatics, University of Minnesota	Java WebStart configuration, scripts and consultation to support using ANT technologies for MeV development.
Douglas Hosack, Glynn Dennis Jr., Brad T. Sherman, H., Clifford Lane,	NIH/NIAID LTB Lab	EASEOpenSource package and consultation for development of EASE in

Richard A. Lempicki, Wei Gao		MeV.
Eric Albert	Independent Development	Implementation of BrowserLauncher.java, utility to launch default web browser from within MeV
Sally Gaddis	University of Texas, MD Anderson Cancer Center	Report and basic testing of Java3D support for mev 3D viewers on the Mac platform.
Vu T. Chu	University of Washington, Microbiology	Agilent data loading design and implementation.
Stephen C. Harris	FDA, National Center for Toxicological Research	T-test bug identification, and suggestion for improvement of adjusted Bonferroni correction.

20. References

- Ben-Dor, A., R. Shamir, and Z. Yakhini 1999. Clustering gene expression patterns. *Journal of Computational Biology* 6:281-297.
- Brown, M.P., W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, Jr., and D. Haussler 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences USA* 97: 262-267.
- Butte, A.J., P. Tamayo, D. Slonim, T.R. Golub, I.S. Kohane 2000. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences USA* 97:12182-12186.
- Chu, G., B. Narasimhan, R. Tibshirani and V. Tusher 2002. SAM "Significance Analysis of Microarrays" Users Guide and Technical Document. <http://www-stat.stanford.edu/~tibs/SAM/>
- Culhane A.C. et al. 2002. Between-group analysis of microarray data. *Bioinformatics* 18:1600-1608.
- Dopazo J., J. M. Carazo 1997. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol.* 44:226-233.
- Dudoit S., J.P. Shaffer and J.C. Boldrick 2003. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science* 18: 71-103
- Dudoit, S., Y.H. Yang, M.J. Callow, and T. Speed 2000. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 2000, Statistics Dept., Univ. of California, Berkeley.
- Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA* 95:14863-14868.
- Fellenberg K. et al. 2001. Correspondence analysis applied to microarray data. *Proceedings of the National Academy of Sciences USA* 98(19)10781-10786.
- Hastie, T., R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, P. Brown 2000. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* 1:RESEARCH0003.
- Herrero, J., A. Valencia, and J. Dopazo 2001. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* 17(2):126-136
- Heyer, L.J., S. Kruglyak, and S. Yooseph 1999. Exploring expression data: identification and analysis of co expressed genes. *Genome Research* 9:1106-1115.

- Hosack, D.A., G. Dennis Jr., B.T. Sherman, H.C. Lane, R.A. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biol.*, 4:R70-R70.8, 2003.
- Keppel, G., and S. Zedeck. 1989. *Data Analysis for Research Designs*. W. H. Freeman and Co., NY.
- Kim, S.K., J. Lund, M. Kiraly, K. Duke, M. Jiang, J.M. Stuart, A. Eizinger, B.N. Wylie, and G.S. Davidson 2001. A Gene Expression Map for *Caenorhabditis elegans* *Science* 293: 2087-2092
- Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43:59-69
- Korn, E.L., J.F. Troendle, L.M. McShane, R. Simon (2001). Controlling the number of false discoveries: application to high-dimensional genomic data. Technical report 003, Biometric Research Branch, National Cancer Institute.
<http://linus.nci.nih.gov/~brb/TechReport.htm>
- Korn, E.L., J.F. Troendle, L.M. McShane, R. Simon (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* 124: 379-398.
- Manly, B.F.J. 1997. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. 2nd ed. Chapman and Hall / CRC , FL.
- Nguyen, D.V., D.M. Rocke. Multi-class Cancer Classification via Partial Least Squares with Gene Expression Profiles. *Bioinformatics*, 18(9):1216-1226, 2002
- Pan, W. 2002. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 18: 546-554.
- Pavlidis, P., and W.S. Noble 2001. Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biology* 2:research0042.1-0042.15
- Raychaudhuri, S., J. M. Stuart, & R. B. Altman 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing 2000*, Honolulu, Hawaii, 452-463.
Available at http://smi-web.stanford.edu/pubs/SMI_Abstracts/SMI-1999-0804.html
- Soukas, A., P. Cohen, N.D. Socci, and J.M. Friedman 2000. Leptin-specific patterns of gene expression in white adipose tissue. *Genes and Development* 14:963-980.
- Tamayo, P., D. Slonim, J. Masirov, Q. Zhu, S. Kitarewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences USA* 96:2907-2912.

Theilhaber, J., T. Connolly, S. Roman-Roman, S. Bushnell, A. Jackson, K. Call, T. Garcia, R. Baron. 2002. Finding Genes in the C2Cl2 Osteogenic Pathway by K-Nearest-Neighbor Classification of Expression Data. *Genome Research* 12:165-176.

Tusher, V.G., R. Tibshirani and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA* 98: 5116-5121.

Welch B.L. 1947. The generalization of 'students' problem when several different population variances are involved. *Biometrika* 34: 28-35.

Yeung, K.Y., D.R. Haynor, and W.L. Ruzzo 2001. Validating clustering for gene expression data. *Bioinformatics* 17:309-318.

Zar, J.H. 1999. *Biostatistical Analysis*. 4th ed. Prentice Hall, NJ.