MeV

MultiExperiment Viewer

Version 4.6
July 2, 2010

Table of Contents

1. General Information

1.1. Obtaining MeV

http://mev.tm4.org

Quick Start Guide

There is a Quick Start guide on the website for beginner users. This may be better to look at first since it contains the basics on how to start using MeV. The full manual can be referenced for more detailed information.

Maintainer / Contact Information

Please look for answers or ask a question in the forums, at
http://www.tm4.org/forum/

Platform / System Requirements

Windows and Linux: Java Runtime Environment (JRE) 1.6 or later
Mac OSX: Java Runtime Environment (JRE) 1.5 or later
Java3D v1.3 or later required for PCA, TRN, DAM and COA functions

Installation Path

Users are strongly discouraged **not** to install MeV under a directory which has space in the location name. E.g. of such directories on a Windows machine would be 'C:\Documents and Settings', 'C:\Program Files' etc. Some modules will not work if MeV is installed in any such directory.

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.

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

1.4. For more help

A link to the online copy of this manual can be found in Help → MeV Manual in the main MultipleExperiment Viewer toolbar. For more help beyond the scope of this manual, or to submit a bug report, see the MeV website, at http://mev.tm4.org.

2. TM4 Software Overview

MultiExperiment Viewer is one member of a suite of microarray data management and analysis applications originally 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 TM4 website, **http://www.tm4.org**.

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.

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.

MultiExperiment Viewer (MeV) is an application that allows the user to view processed microarray slide representations and identifies 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 MultiExperiment Viewer

3.1. If using Windows, run the TMEV batch file (TMEV.bat) to start the program. Similarly, if using Linux or Unix, run the tmev.sh file. Macintosh users should double-click on the application file named MeV. Be careful not close the window to the command prompt window or else the whole program will close.

A main menu bar will appear, with five menus: *File*, *Display*, *Window*, *About* and *Help*. A Multiple Array Viewer should also open. To open a new Single or Multiple Array Viewer, load a new preferences file, or log in to a database, use the *File* menu in the main menu bar,. MeV will continue to run while this menu bar is present. To exit the entire application, select *Quit* from the *File* main menu.

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 2 for details regarding the Single Array Viewer.

3.2. Another way to start MeV is by launching it via Java WebStart. In this way, MeV can be launched from a web page by clicking a link. A pre-selected data set and annotation can be automatically loaded without user intervention. Setting up this sort of system is the responsibility of the web site's administrator, and so is beyond the scope of this manual. More information and technical details about this feature and MeV's command-line options are available at http://www.tm4.org/mev_webstart.html.

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 4 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 Data* 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 by choosing *Select File Loader → TIGR Files → MeV Files* Use this browse button in the *File* panel 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 (4.4). In the section of the loader labeled "MeV Expression Files (*.mev)," the contents of the folder selected in the file selection dialog box 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 "MeV Annotation Files (*.ann, *.dat)."

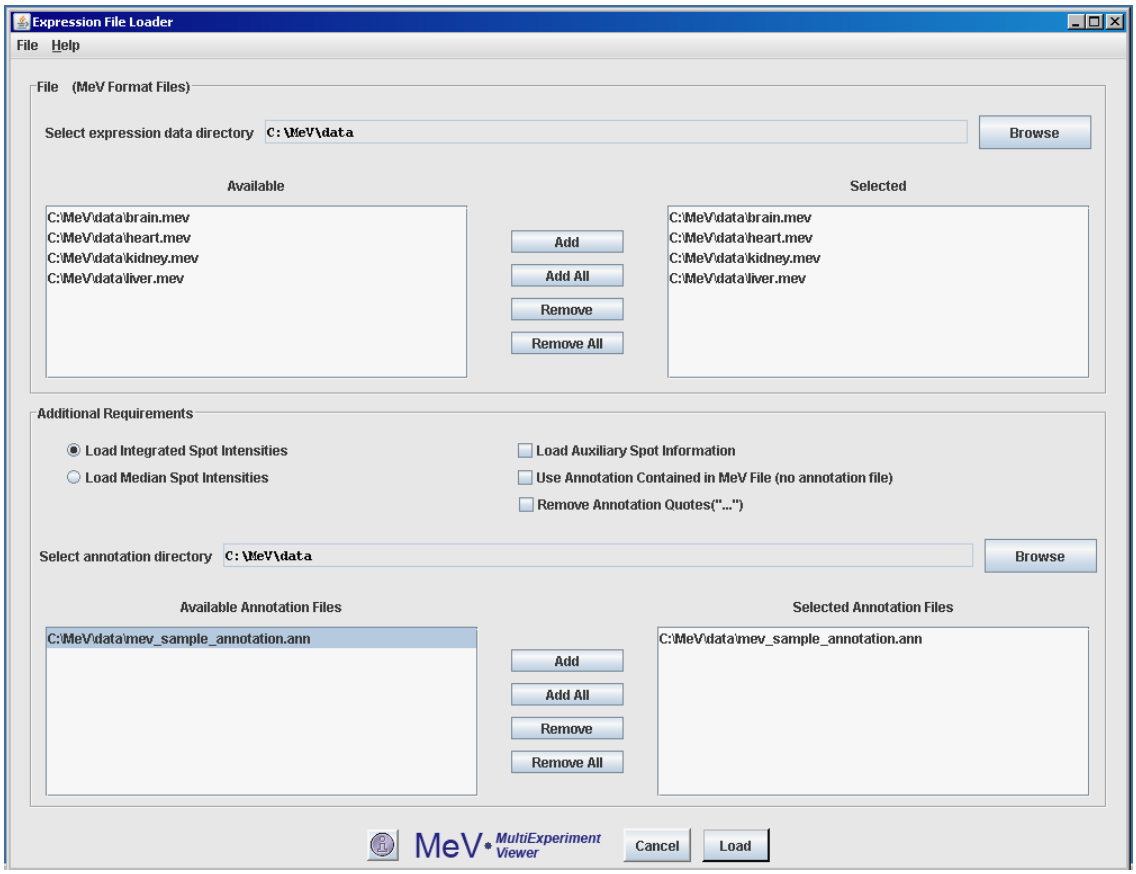
The *Load Integrated Intensities/Load Median Intensities* options specify that MeV should load either integrated intensities (default) or Median intensities. Note that care should be taken to select the intensity measurements, integrated or median, that has been previously normalized in MIDAS so that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

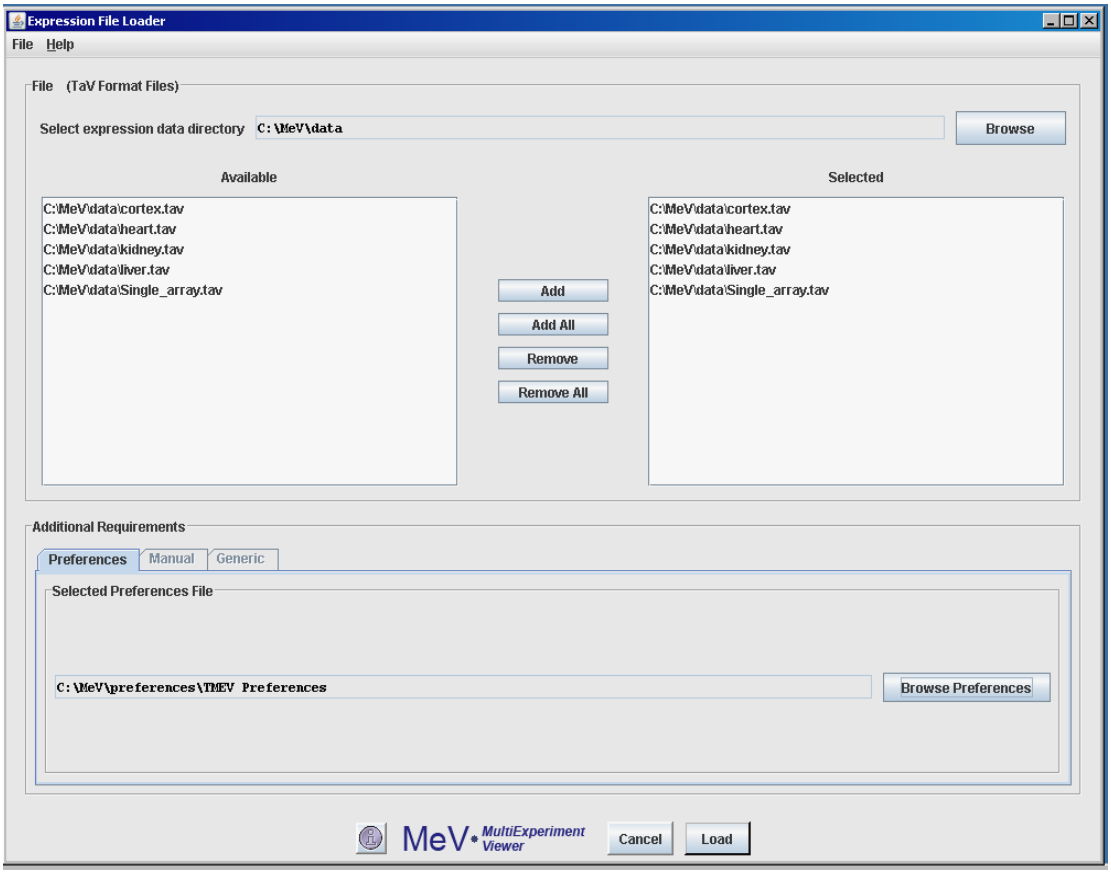
The *Remove Annotation Quotes* option removes quotations from annotation fields where the annotation entries start and end with quotation marks. This option is provided to counteract the behavior of some popular 'spread sheet' programs where cells containing text with a delimiter such as a comma are 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

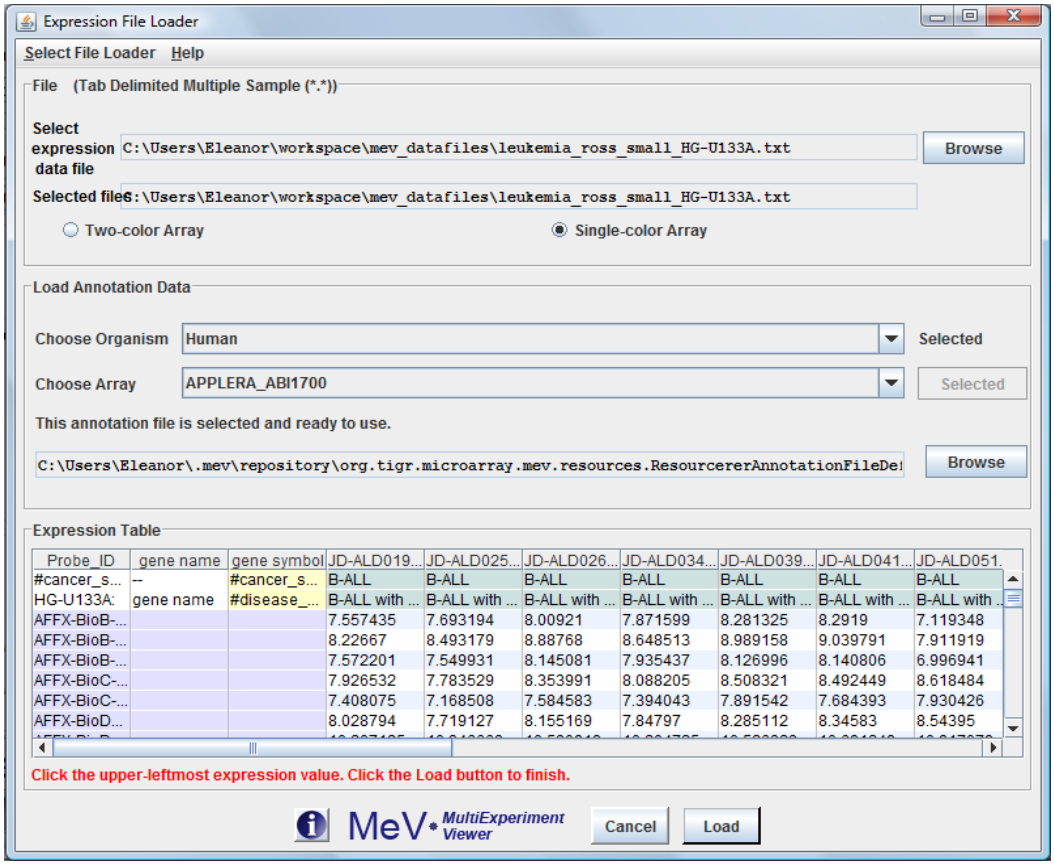
To load .tav-formatted files, use the drop-down menu to select the *TIGR Array Viewer (*.tav)* option. This loader is very similar to the .mev loader. Use the browse button in the *File* panel to select the .tav files to load. Instead of selecting annotation files to load alongside the data files, however, 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 4.11) for more details on preferences files (4.11).



4.2.1. The Expression File Loader (.tav Files)

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

The Expression File Loader should be the screen that automatically comes up once you select *Load Data* from the *File* menu. If not, click *Select File Loader* → *Tab Delimited, Multiple Sample Files (TDMS) (*.*)* option from the drop-down menu to load TDMS format files. Use the browse button in the *File* panel to select the desired file. The file will be displayed in a tabular format in the file loader preview table. Select either the *Two-color Array OR Other* or the *Single-color Array* radio button to indicate the data type of the expression data. Please read Using the Annotation Feature, for information on the same. Click the cell in the table which contains the upper-leftmost **expression value** in the file. That is, click the upper-leftmost cell that contains measuered data of the genes. MeV will color-code the cells of the display table to indicate which cells it will load as expression data (white and gray stripes), gene/row annotation (purple), sample annotation (blue) or sample annotation labels (yellow). Gene annotation labels will be displayed in the column headers. Data in cells with white backgrounds will not be loaded. MeV allows you to have as many columns as you want of labels and annotation, but you must tell it where the labels end and the expression data begins. Check that the correct fields are listed before clicking *Load*.



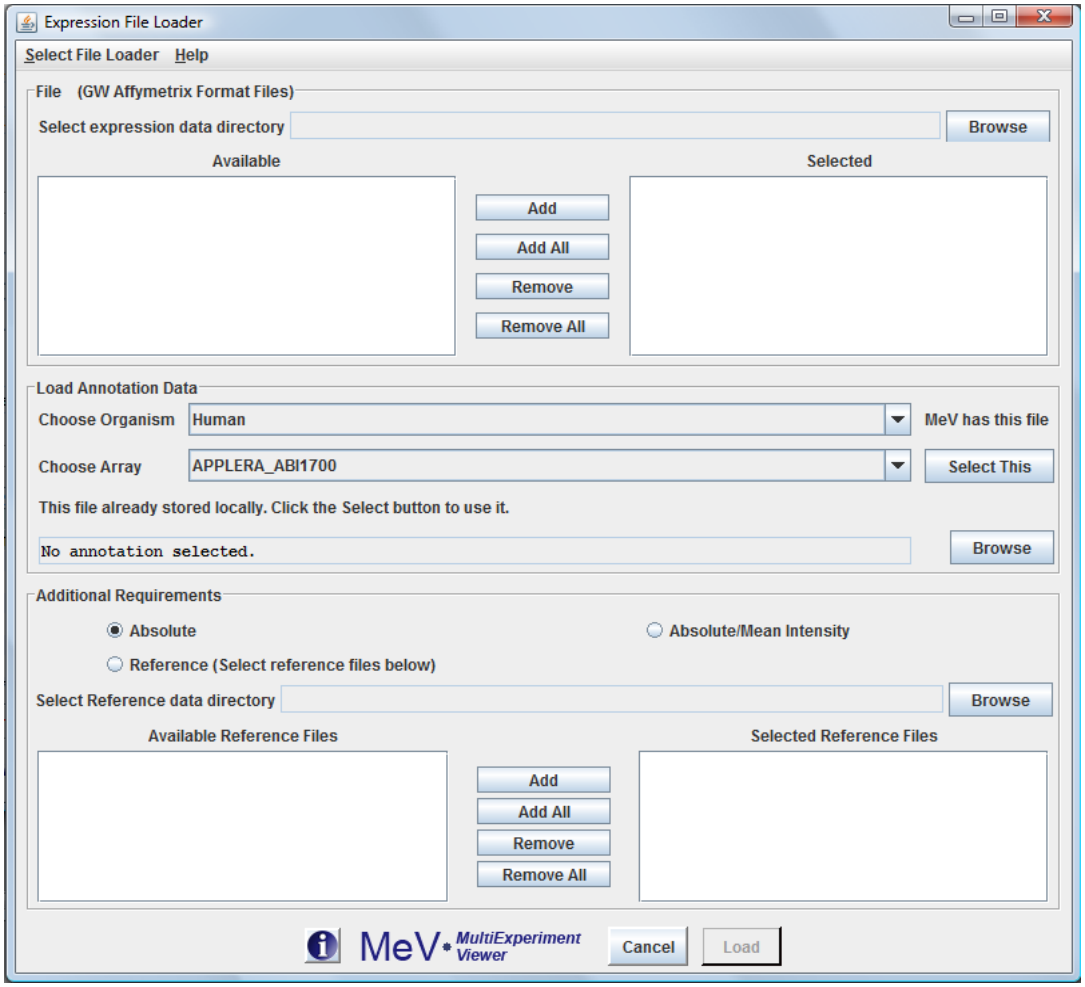
4.3.1. Loading Tab Delimited, Multiple Sample Files (TDMS Format)

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

Selecting *Affymetrix Data Files (*.txt)* from the drop-down menu allows the loading of Affymetrix files. We currently support five different Affymetrix file format namely: GW Affymetrix, dChip/DFCI_Core, Affymetrix GCOS, Bioconductor (using MAS5) and RMA Files. Select the directory containing files to be loaded using the browse button in the *File* panel. The contents of the folder selected in the file selection dialog box will be displayed in the box labeled "Available". Select the files to load and click the "Add/ Add All" button to add them to the list of files to be loaded. Please read Using the Annotation Feature section, for information on the same.

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 used 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. Automatically Setting Color Scale Limits

When data are loaded in MeV, it automatically generates an expression image. 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. For Affymetrix data and two channels data, each has different data range in the data set. Now MeV has a new function to automatically sets the color scale limits according to different data sets.

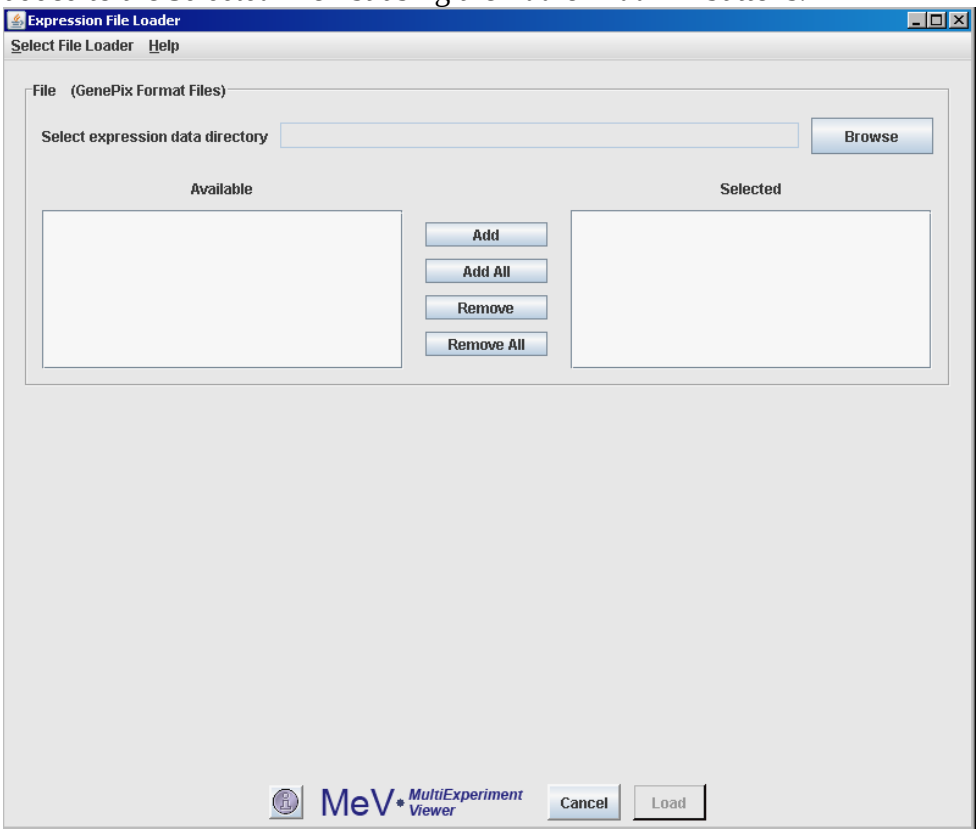
For Affymetrix data the low end is set to 0 and midpoint is set as median and high end is set to the value so that 80% of the data fall below this value.

For two color array data the low end is set to -3 and midpoint is set to 0 and high end is set to 3.

Users can change color scale limits after loading as instructed in the following section.

4.6. Loading GenePix (.gpr) Data Files

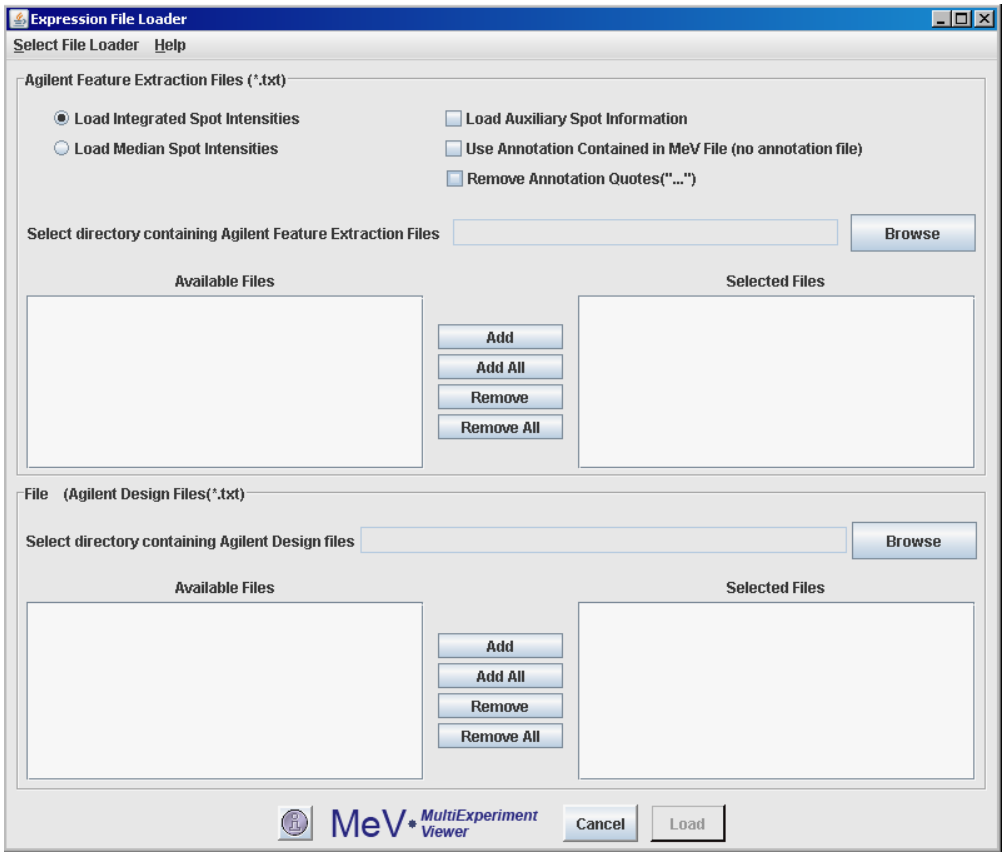
GenePix (.gpr) files can be loaded by selecting the GenePix file loader option from the list of available file formats to load. Use the *Browse* button to select the directory containing the .gpr files. Files appearing in the *Available* file list can be added to the *Selected* file list using the *Add* or *Add All* buttons.



4.6.1. The Expression File Loader (GenePix Files)

4.7. 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 *Browse* button to select the directory containing the files. 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.



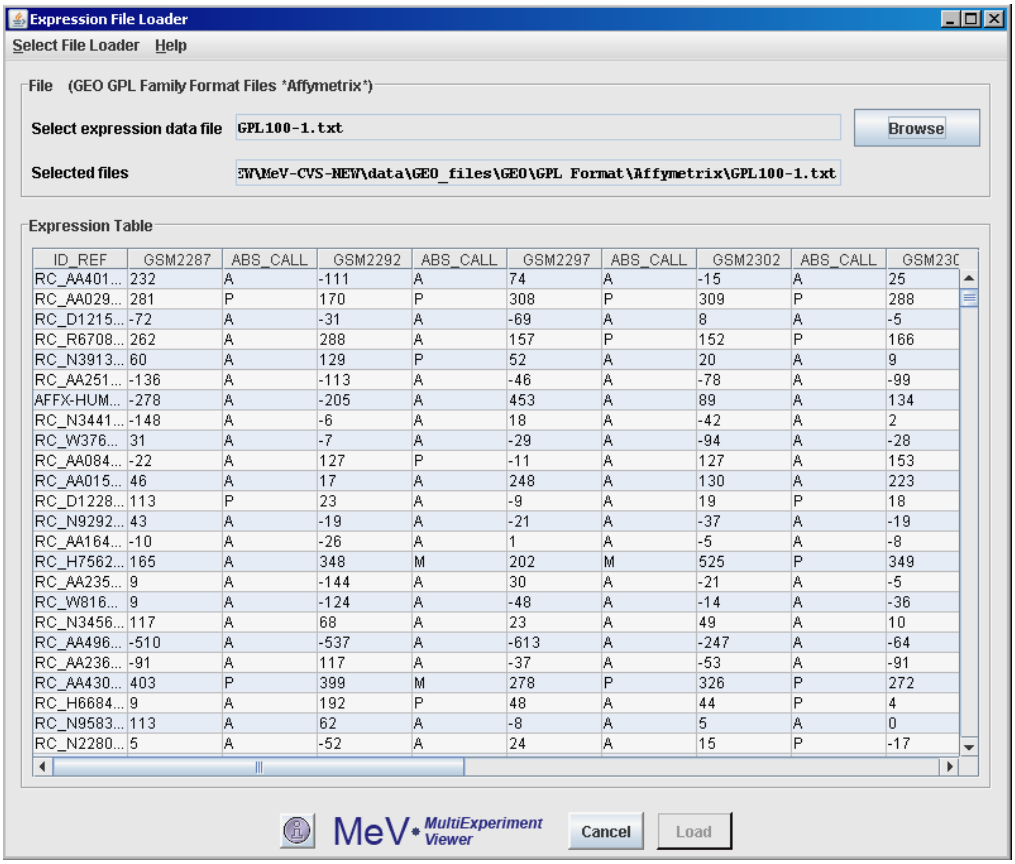
4.7.1. Agilent File Loader

4.8. Loading GEO Simple Omnibus Format in Text (SOFT) Affymatrix format File

After Expression File Loader dialog is launched, SOFT Affymetrix format file can be loaded by selecting the GEO SOFT Affymetrix file loader option from the list of available file formats to load.

Use the *Browse* button in the *File* panel to locate the file to be loaded. The selected file will be displayed in a tabular format in the file loader preview table. We currently allow loading GPL and GSM format files using this loader. The platform information contained in the GPL files is automatically loaded. There is no provision to load the platform information separately. Click the cell in the table which contains the upper-leftmost **expression value** in the file, then click the

Load button to load the file.

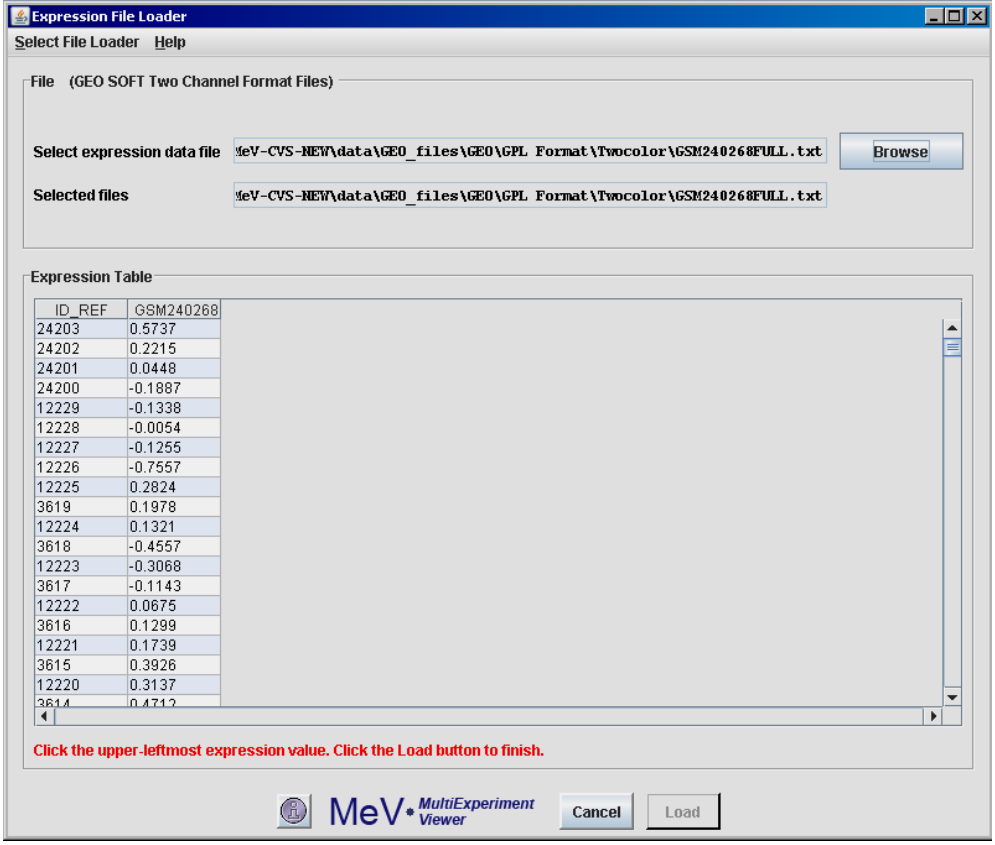


4.8.1. GEO SOFT File Loader

4.9. Loading GEO Simple Omnibus Format in Text (SOFT) two channel format File

SOFT two channel format file can be loaded by selecting the GEO SOFT two channel file loader option from the list of available file formats to load.

Use the *Browse* button in the *File* panel to locate the file to be loaded. We currently allow loading GPL and GSM format files using this loader. The platform information contained in the GPL files is automatically loaded. There is no provision to load the platform information separately. The selected file will be displayed in a tabular format in the file loader preview table. Then click the cell in the table which contains the upper-leftmost **expression value** in the file, then click the Load button to load file.

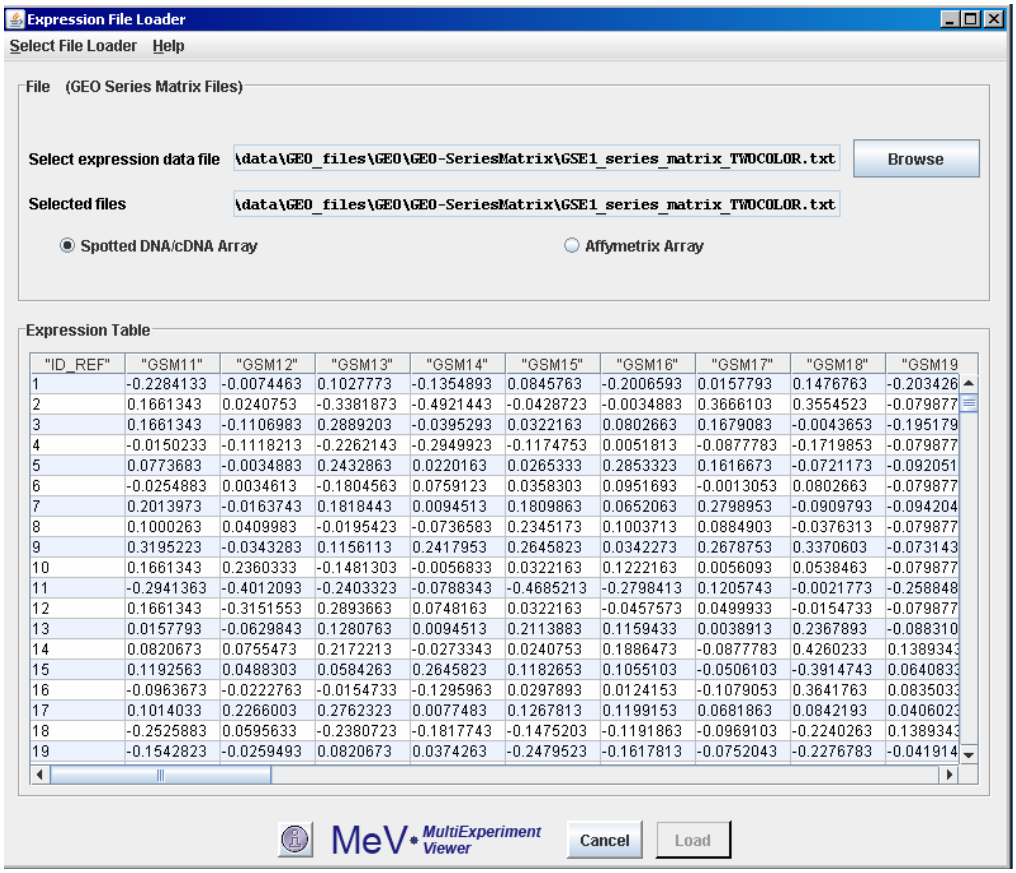


4.9.1. GEO SOFT two-channel file loader

4.10. Loading GEO Series Matrix Files

GEO Series Matrix files can be loaded using the *Browse* button in the *File* panel. Select *"Spotted DNA/cDNA Array" or "Affymetrix Array"* to indicate the data-type of the loaded expression data.

Series_matrix files are summary text files that include a tab-delimited value-matrix table generated from the 'VALUE' column of each Sample record, headed by Sample and Series metadata. These files include SOFT attribute labels. Data generated from multiple Platforms are contained in separate files. It is recommended to view Series_matrix files in a spreadsheet application like Excel. CAUTION: value data are extracted directly from the original records with no consideration as to whether the values are directly comparable.

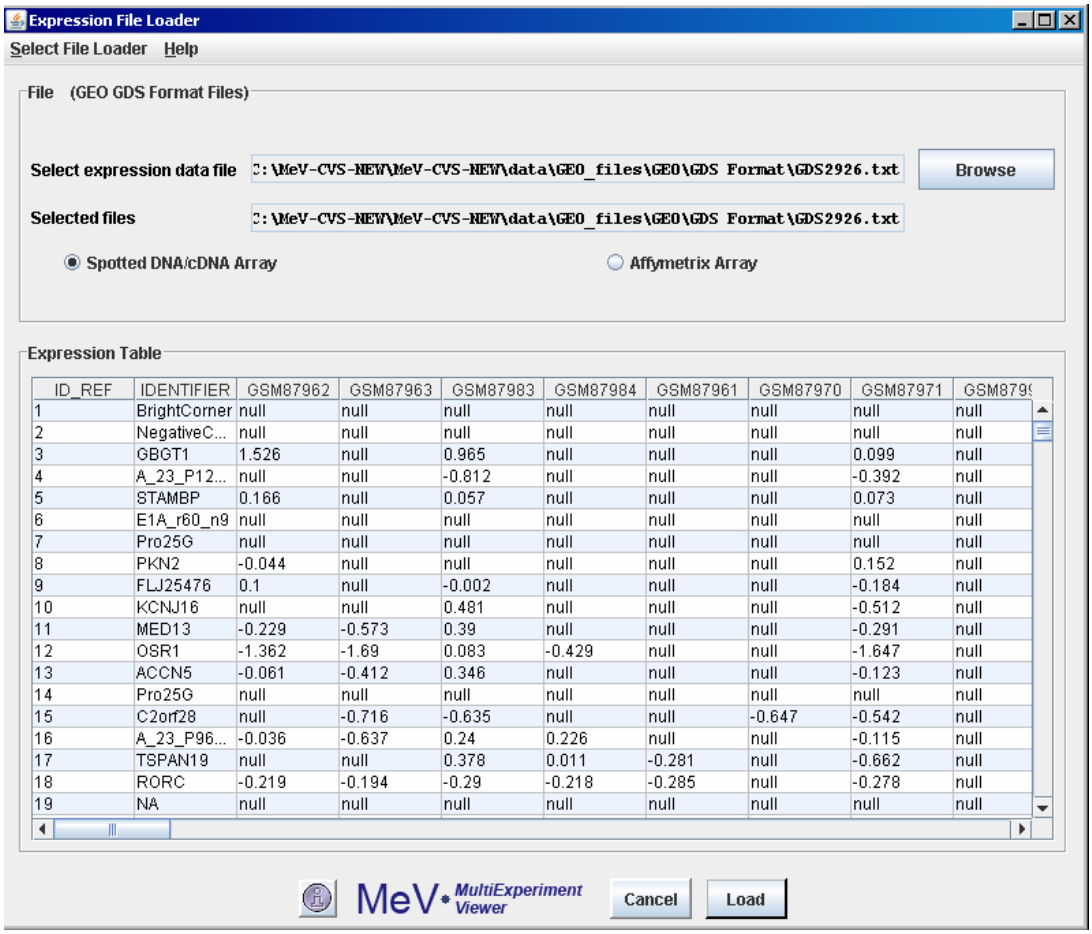


4.10.1. GEO Series Matrix File Loader

4.11. Loading GEO GDS Format Files

GEO GDS format file can be loaded using the *Browse* button in the *File* panel. Select *"Spotted DNA/cDNA Array" or "Affymetrix Array"* to indicate the data- type of the loaded expression data.

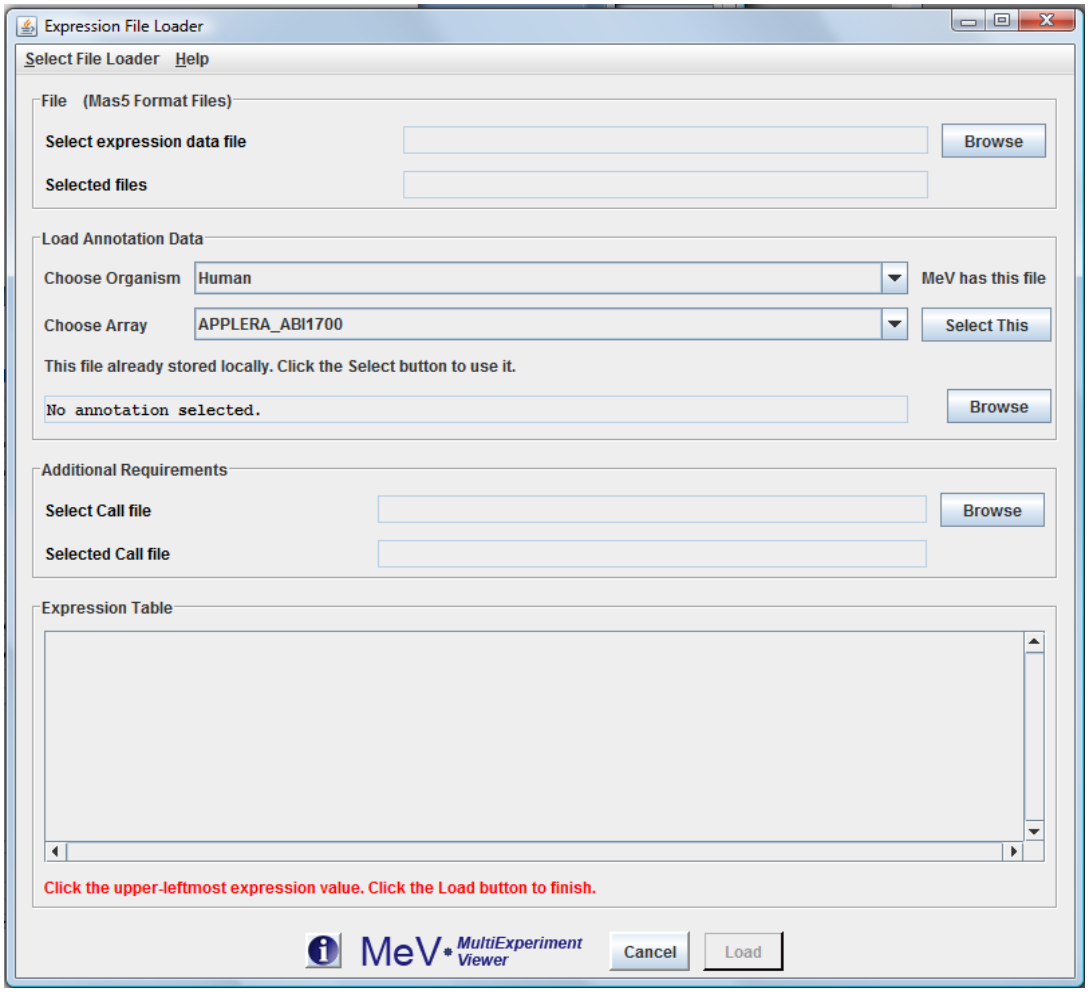
GEO Datasets (GDS) are curated sets of comparable GEO Sample (GSM) data. GDS data tables contain VALUE measurements extracted from original Sample records.



4.11.1. GEO GDS Format File Loader

4.12. Loading Bioconductor(MAS5) Format Files

Bioconductor (MAS5) format file can be loaded by selecting the Bioconductor (MAS5) file loader option from the list of available file formats to load. Use the browse button to locate the files to be loaded. The file will be displayed in a tabular format in the file loader preview table. The call format file generated by Bioconductor can be selected using the browse button in the “Additional Requirements” panel. Please read Using the Annotation Feature section, for information on the same. Click the cell in the table which contains the upper-leftmost **expression value** in the file. Click button Load to load file.



4.12.1. Bioconductor (MAS5) Format file loader

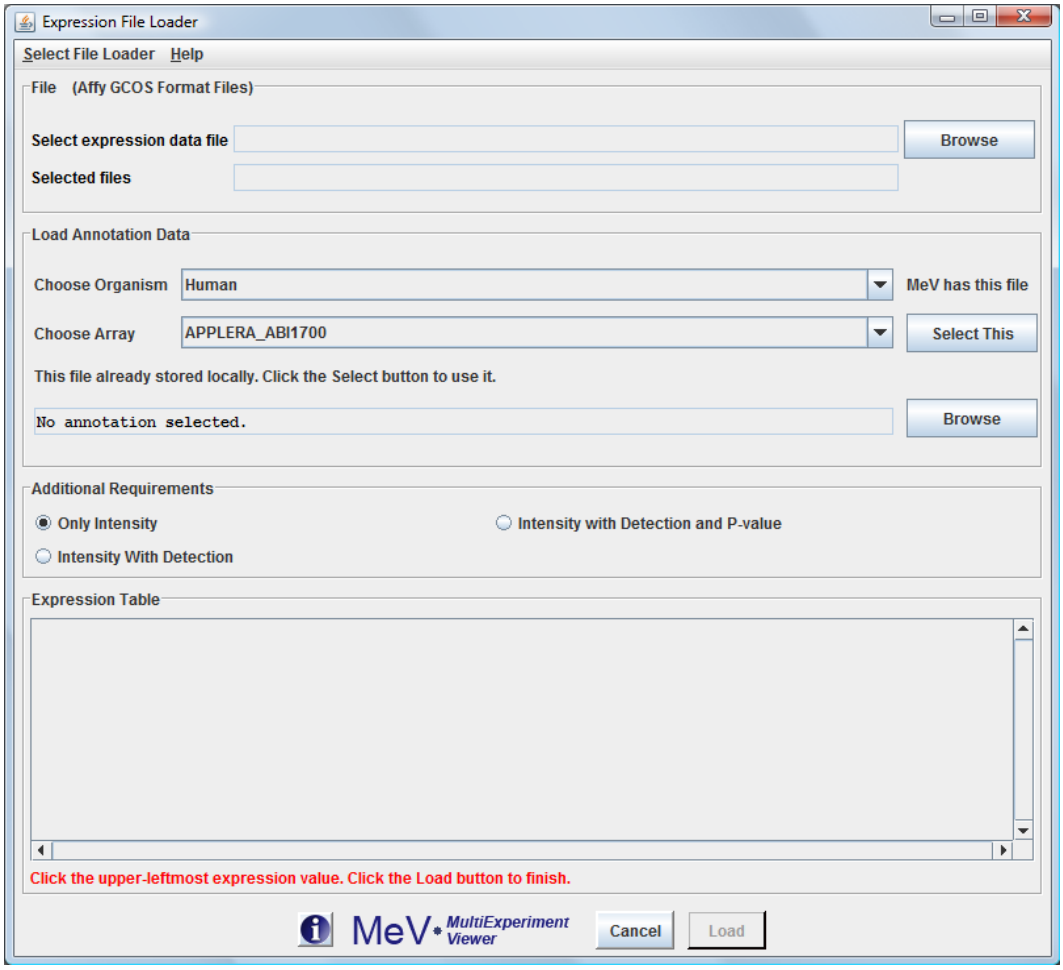
4.13. Loading Affymatrix GCOS Format Files

Affymetrix GCOS (Gene chip operating software) output (pivot data) files can be loaded by selecting the Affymetrix GCOS file loader option from the list of available file formats to load.

This file loader can actually load three similar file formats by choosing data option radio buttons.

Only Intensity-----only containing signal intensity for every experiment
Intensity with Detection -----containing detection for every experiment
Intensity with Detection and p-value-----containing detection and P-value for every experiment

Use the *Browse* button to locate the files to be loaded. The file will be displayed in a tabular format in the file loader preview table. Please read Using the Annotation Feature section, for information on the same. Choose the data options by clicking proper radio button. Then click the cell in the table which contains the upper-leftmost **expression value** in the file. Then click the Load button to load the file.

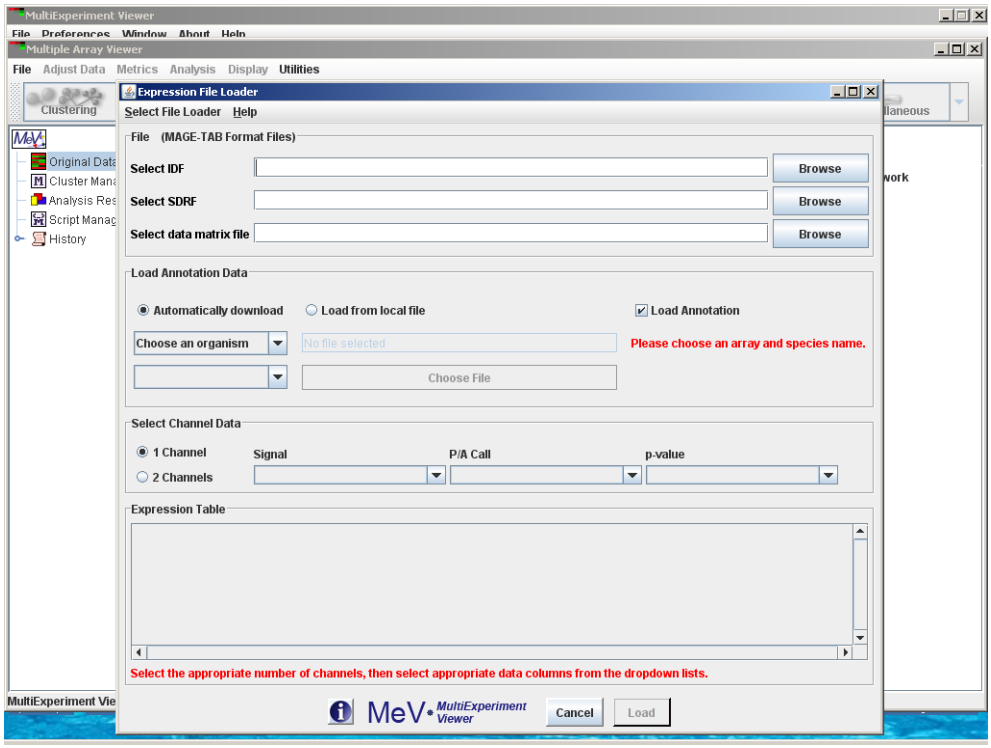


4.13.1. Affymatrix GCOS Format file loader

4.14 Loading MAGE-TAB formatted data files.

The MeV MAGE-TAB file loader allows a user to select and load expression data from MAGE-TAB (version 1.0) formatted files. The MAGE-TAB specification can be found at: http://www.mged.org/mage-tab/spec1.0.html . This format includes files that describe an experimental series, the samples used in the experimental series, optionally the array design of the microarray chips employed in the study, and a data matrix file containing gene expression data values.

The MeV MAGE-TAB file loader can be accessed by selecting 'File > Load Data' from the file menu, and then 'Select Loader' from the loader menu. The panel itself is different from other MeV file loaders in that it must load at least 3 files (defined in the MAGE-TAB spec.). There are 4 regions on the panel: 1) the files section, 2) the annotation section, 3) the channel data section, and 4) the data section.



In the files section, the individual data entry fields are initially blank. Once the user browses to and identifies the location of one of the required files (IDF, SDRF or data matrix files), MeV will check for the existence of the others and report back with a file name or a message indicating that one of the file types was not found. The user can then browse to find each file individually. At this point MeV will only search for the file currently being selected.

The annotation section allows the user to either select gene annotation from a list of arrays that exist on the current system, or can be downloaded from the internet. (Network access is required for the latter.)

The channels section will be populated once a user selects a data matrix file. In fact the user should not set any values until selecting a data matrix because they will be changed once the file is selected. After the file is selected and the panel is updated, the user can select appropriate values if they have not already been selected.

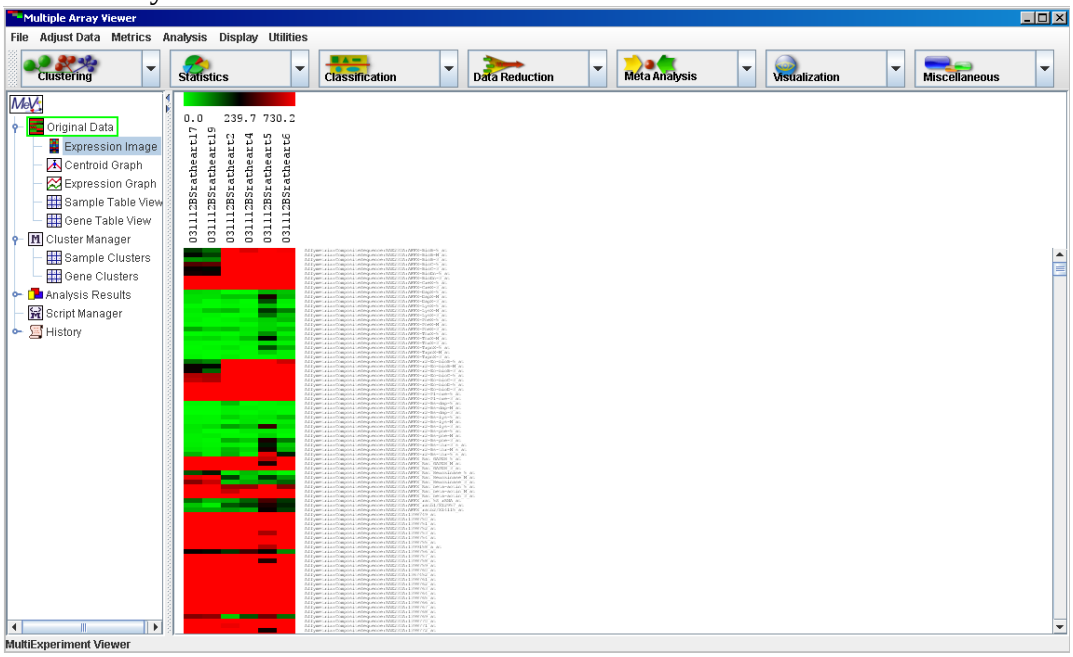
The 1 channel / 2 channel radio buttons indicate whether the data is derived from 1- or 2- channel hybridizations. Depending upon which radio button is selected, MeV will display either Signal, Detection and p-Value combo boxes, or Intensity1, Intensity2 and Log Ratio/Ratio combo boxes. The Log Ratio/Ratio combobox may be used for either Log(ratio) or ratio data. Each drop-down list includes all of the data type items from the second row of the MAGE-TAB data matrix file plus one additional entry, “none”, to allow the user to select no value for the combobox. The user must know and understand the data well enough to make appropriate selections for these lists. The data section merely displays the data from the selected data matrix file. No interaction with this section is

necessary. The 'Load' button is inactive until the user provides a data matrix file. The Load button will then be activated.

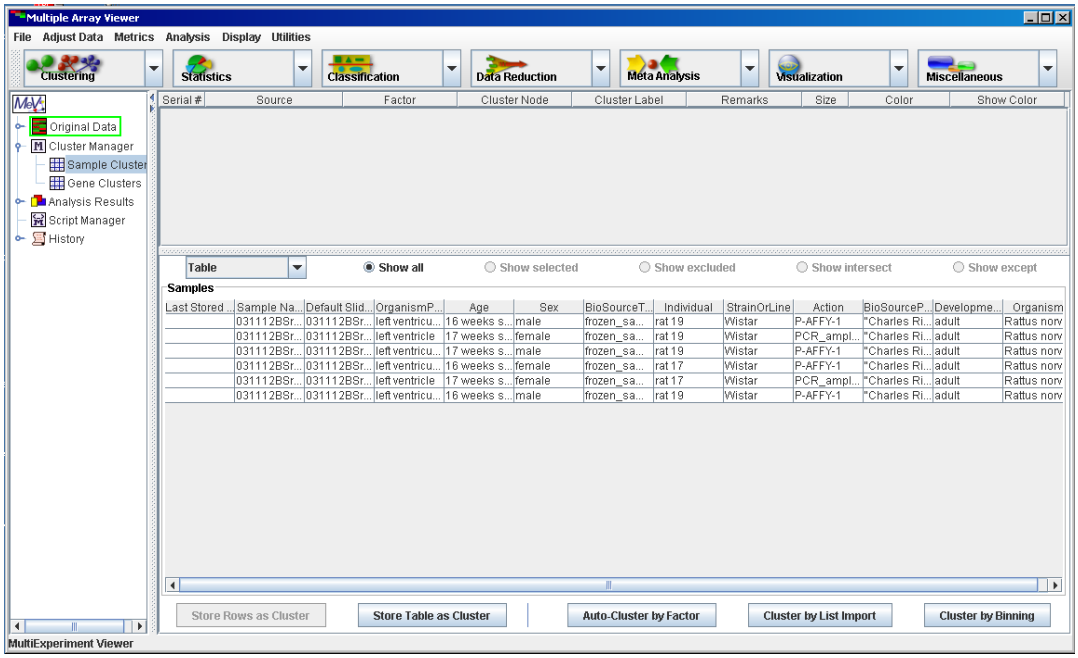
Once a data matrix file has been selected and displayed, the user can select 'Load' to load and parse the data matrix file. MeV will display the data matrix as a heat map with row labels derived from the first column of the data matrix file. Once the data have been presented as a heat map, all other MeV functionalities will be available for use.

The MAGE-TAB format defines categorical items that MeV can use to group data. This data is found in the ExpermentalFactor tags of the MAGE-TAB IDF and the FactorValue columns in the SDRF. MeV loads this data into a ClusterAnnotation viewer which will allow the user to view and edit this information. The user can there specify which fields should be used to group the data. This is useful for analyses such as t-test, ANOVA, etc.

Where can you view all the data in the SDRF file?

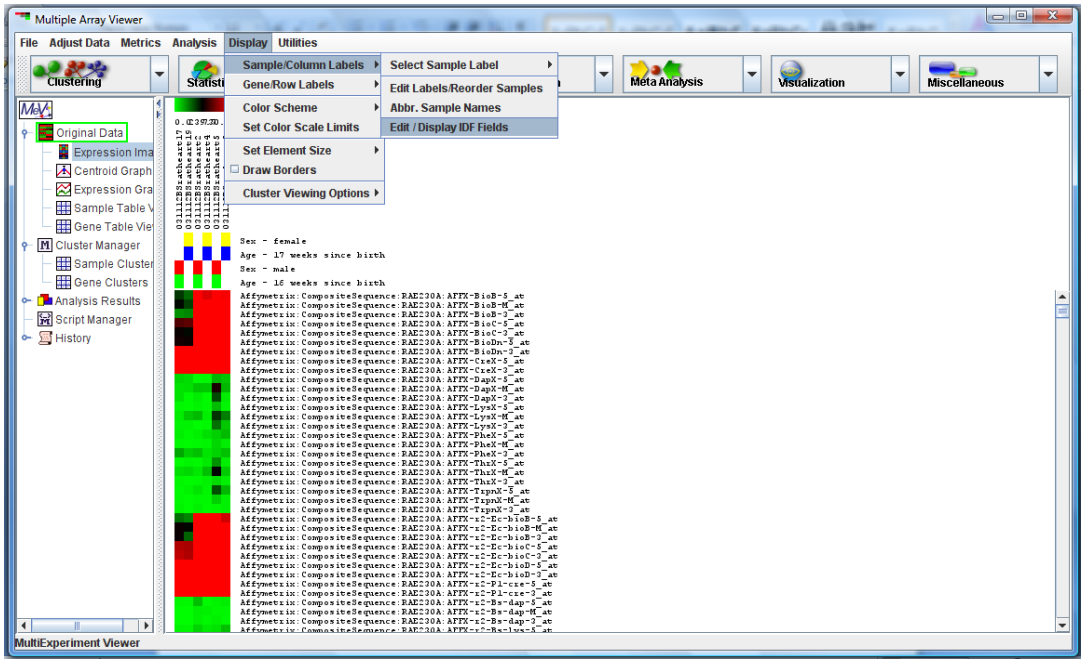


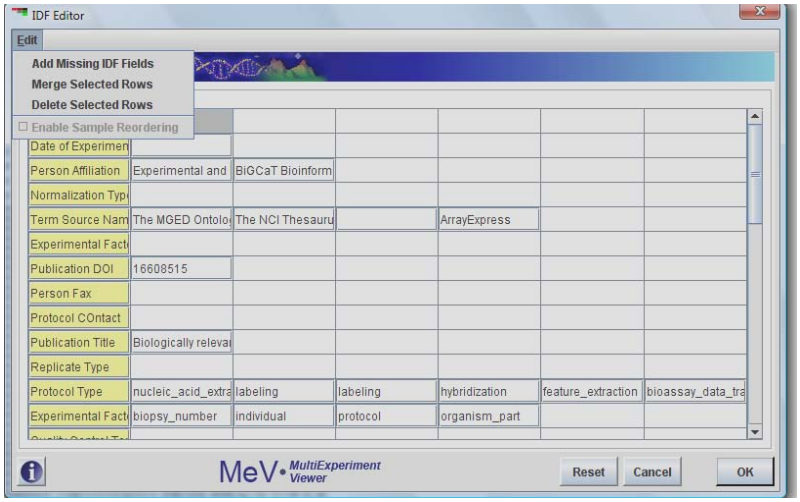
The Cluster Manager ->Sample Clusters is where you will be able to view the data that you loaded from the SDRF file.



The contents of the SDRF file (Characteristics and FactorsValues) are displayed in a tabular view in Cluster Manager.

Where can I view and edit the contents of IDF file?





The IDF Editor lets you view and edit your IDF files. The first column (IDF fields) is not editable. If you want to add any missing fields to your data, click on the "Add Missing IDF Fields" option. This will automatically insert the missing rows. Changes to the IDF or SDRF data are retained in memory until you exit MeV. None of this data is saved in the current version of MeV.

The current version of MeV will not save data in MAGE-TAB format; however, this feature is planned for a future release of MeV.

Example files can be found in the following links:
MAGE-TAB files supplied along with MeV (in the "data/magetab" folder), e.g.
E-ATMX-12.idf.txt
E-ATMX-12.sdrf.txt
E-ATMX-12-processed-data-1343527784.txt

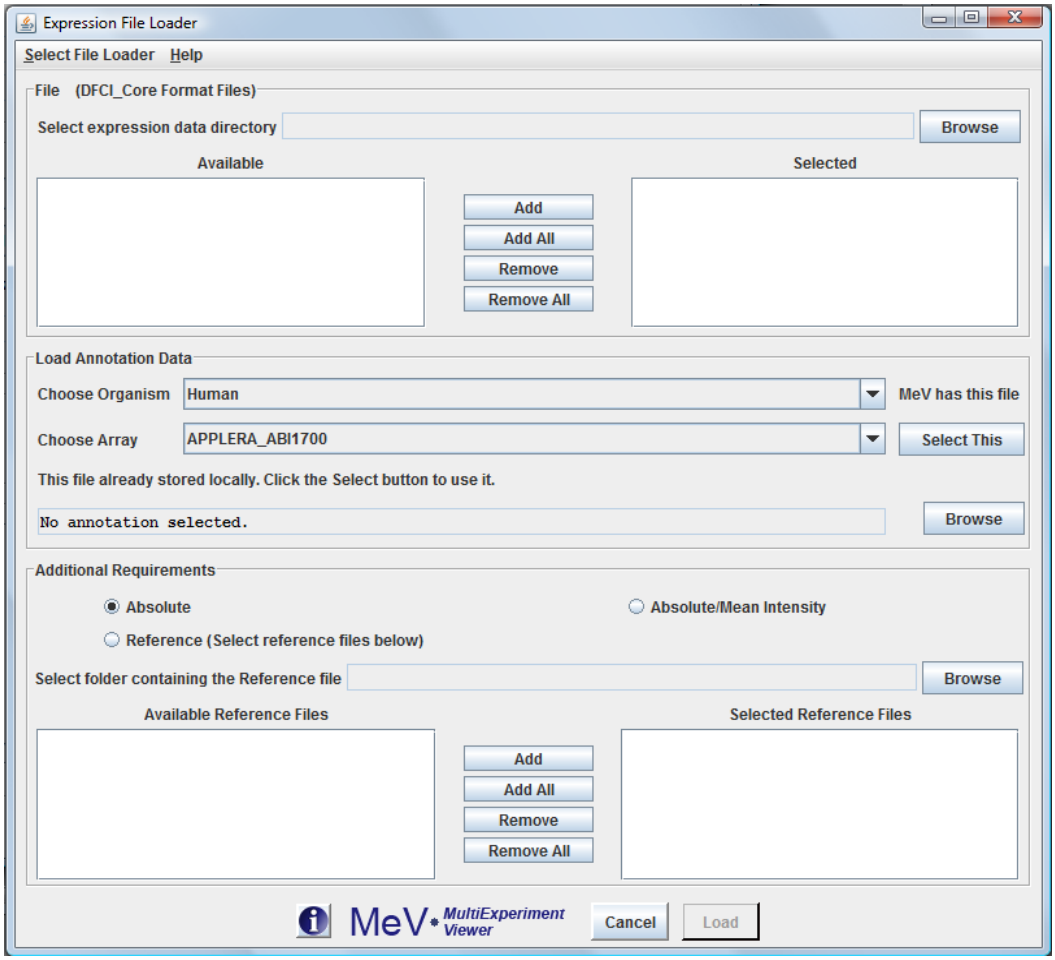
NOTE: Although we have tested a variety of MAGE-TAB files with MeV, there are some MAGE-TAB files that may not load into MeV. This may be due to flexibility in the MAGE-TAB v.1.0 specification. Also note that not all the files available on the ArrayExpress web site that are in MAGE-TAB format can be loaded into MeV because the parser we currently use supports the MAGE-TAB v.1.0 specification. The ArrayExpress web site includes files that conform to the as yet unpublished MAGE-TAB v.1.1 format.

4.14. Loading dChip Output Files

Selecting *dChip/DFCI_core Format Files (*.*)* from the drop-down menu allows the loading of dChip or DFCI_core output files. Select the directory containing data files using the *Browse* button. Files appearing in the *Available* file list can be added to the *Selected* file list using the *Add* or *Add All* buttons. Please read Using the Annotation Feature section, for information on the same.

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 used 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. 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.



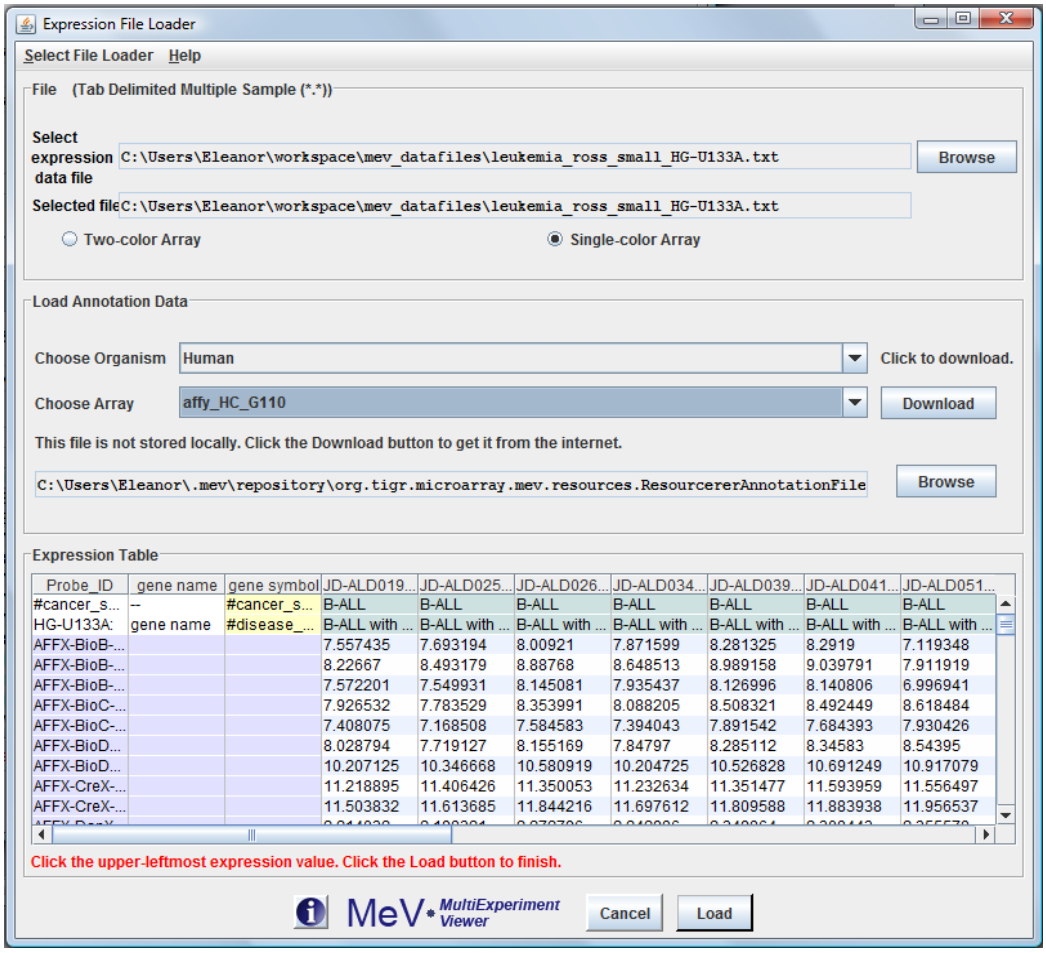
4.14.1. dChip File Loader

4.15. Using the Annotation Feature

One of the features of MeV release 4.1 is enhanced support for gene annotations. MeV provides users an alternative to fetch additional gene annotation, from the web based resource providing annotation based on The Gene Indices (TGI) for commonly available microarray resources, including widely used clone sets and Affymetrix GeneChip Arrays namely; RESOURCERER. Please refer to the developer's manual for further information on the Annotation model. This feature is currently provided only for Affymetrix data.

Steps for using this feature

1. Load the data file/files of your choice. If you are using the Tab delimited file loader (TDMS), check the "*Affymetrix Array*" radio button. The TDMS loader considers the array platform to be "*Spotted DNA/cDNA Array OR other*" by default.



Step 1.

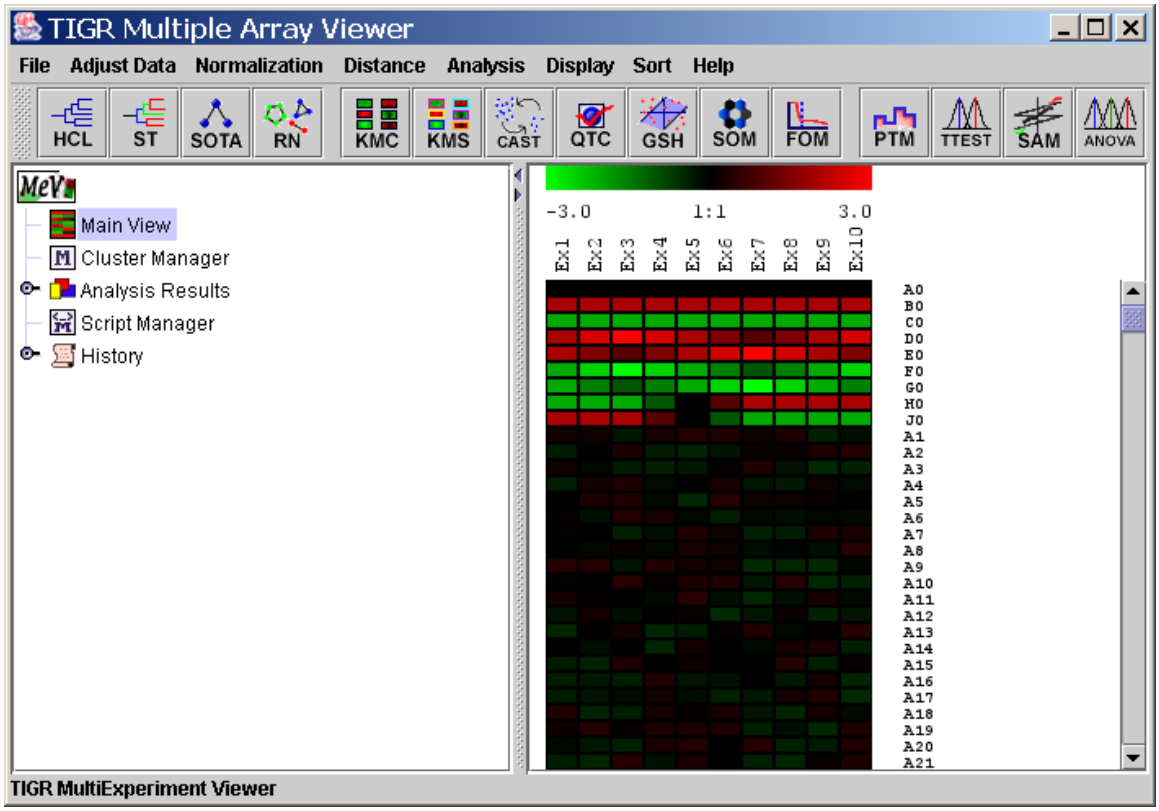
2. Select the organism your expression data comes from using the Choose Organism drop-down menu. The list of arrays in the Choose Array drop-down menu will be updated with the available array types. Choose the one that your data was collected from. The button to the right of the Choose Array drop-down menu will update, and will ready "Download" if the file has not already been downloaded. Click this button and MeV will attempt to download the file from the internet. If the file has been previously downloaded, this button will read "Select" instead of Download, indicating that internet access is not required to use this annotation. Click "Select" to choose this annotation type.

If internet access is not available and you have an annotation file of the appropriate format, you can still select an annotation file from your filesystem. Click the "Browse" button, located beneath the "Download" button, and select the file. Annotation file formats that are accepted include the Resourcerer file format and the Affymetrix .csv file format.

Follow any further instructions present in the file loader panel before clicking the *Load* button. In the example shown here, you will click on the upper leftmost expression value in the *Expression Table* panel before clicking the *Load* button.

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

For each set of expression values loaded, a column is added to the main display (4.16.1). 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.



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

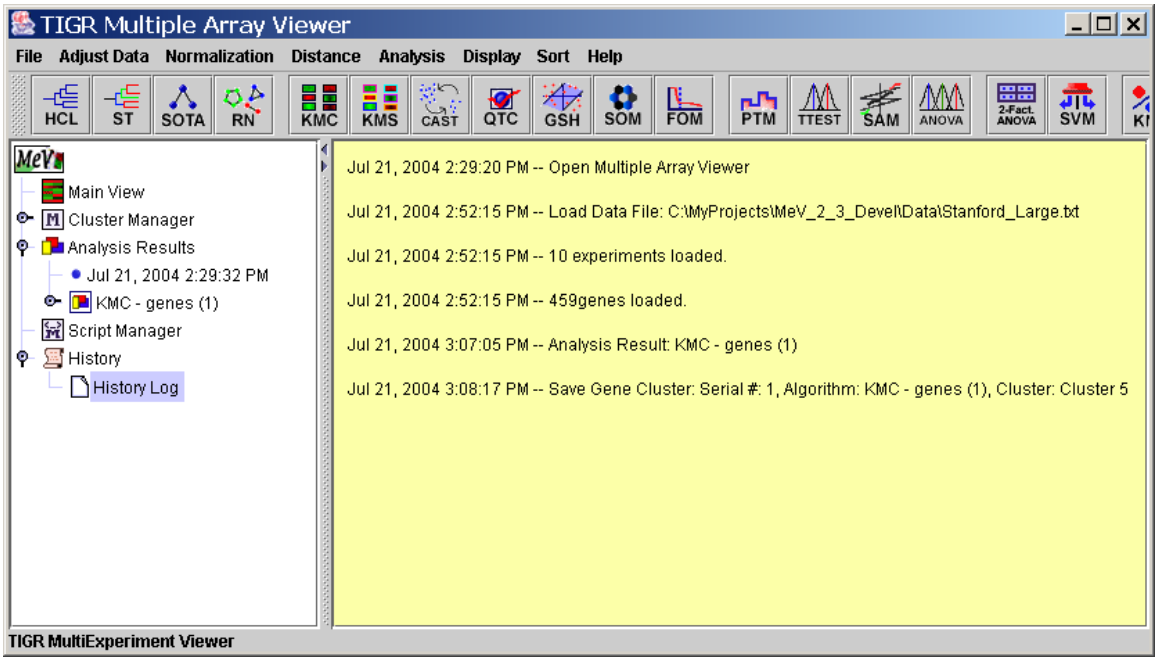
4.17. Result Navigation Tree

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 is sections 7.2 and 10 respectively.

4.18. 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.

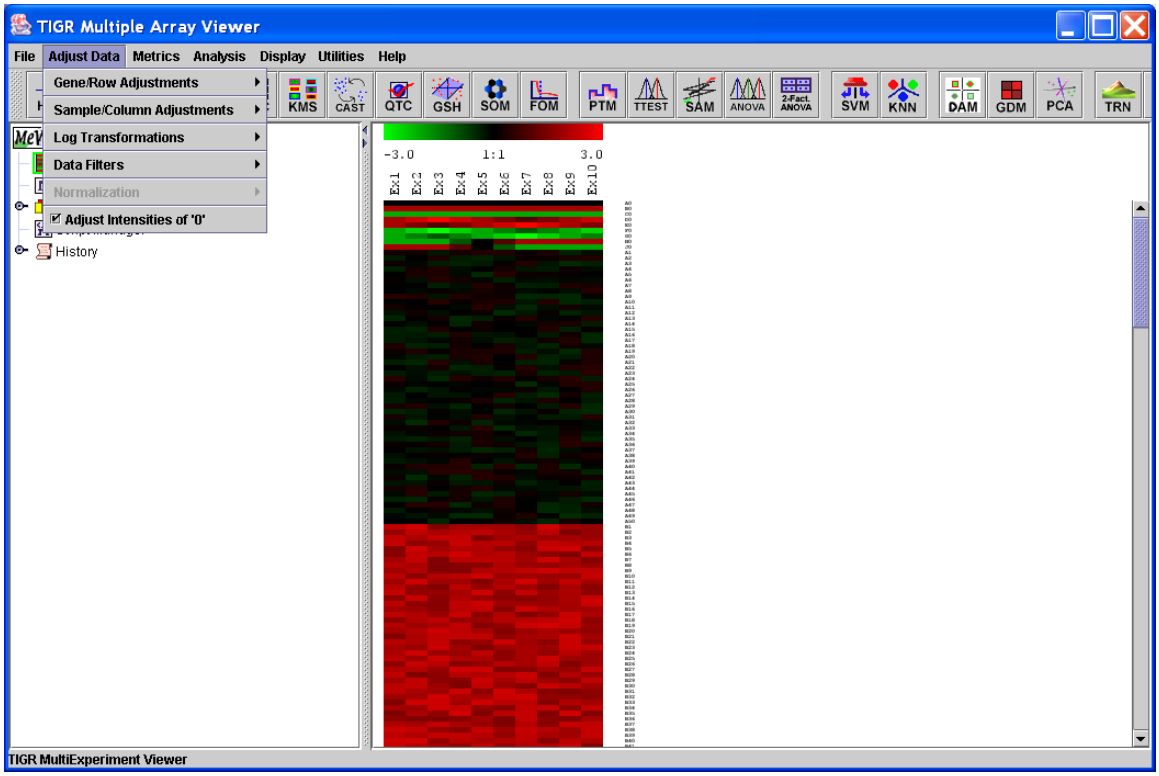


4.18.1. The History Log

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. However, adjusts will influence the calculation of the expression matrix, the foundation of all analyses. They will also be reflected when the entire matrix or individual clusters are saved as text files (*.txt), 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, as the post-transformation analyses and displays might not be entirely consistent with the pre-transformation analyses.. (The exception for this would the “Set Lower Cutoffs,” “Set Percentage Cutoffs” and “Adjust Intensities of Zero” options which will be discussed in section 5.4.) 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,

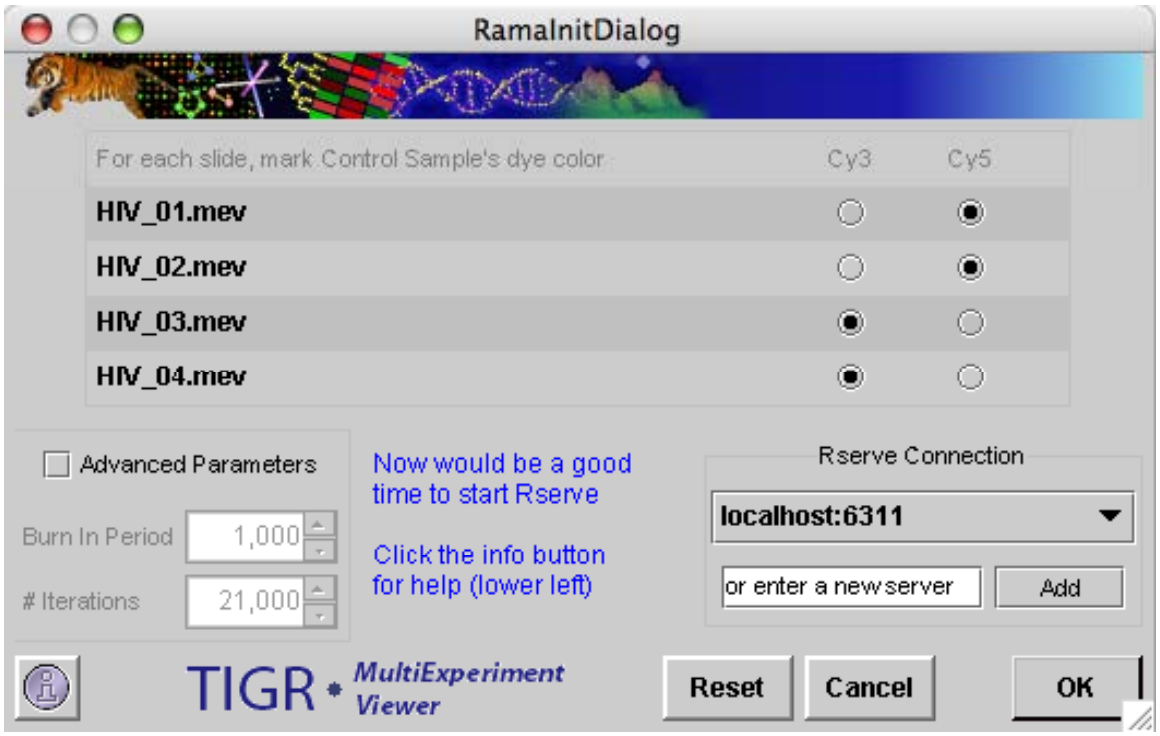
multiple sample (TDMS) formatted text file by choosing *File → Save Matrix*, and then load this new file in a new MeV session. During this time no further data adjustments will be made. This will ensure consistency throughout the MeV session. Adjustment options are described below.

5.2. Replicate Analysis

RAMA: Robust Analysis of MicroArrays

Robust estimation of cDNA microarray intensities with replicates. The package uses a Bayesian hierarchical model for the robust estimation. Outliers are modeled explicitly using a t-distribution, and the model also addresses classical issues such as design effects, normalization, transformation, and nonconstant variance.

The initialization dialog shown below allows the user to denote the dye labeling scheme used in the experiment.



RAMA Initialization Dialog

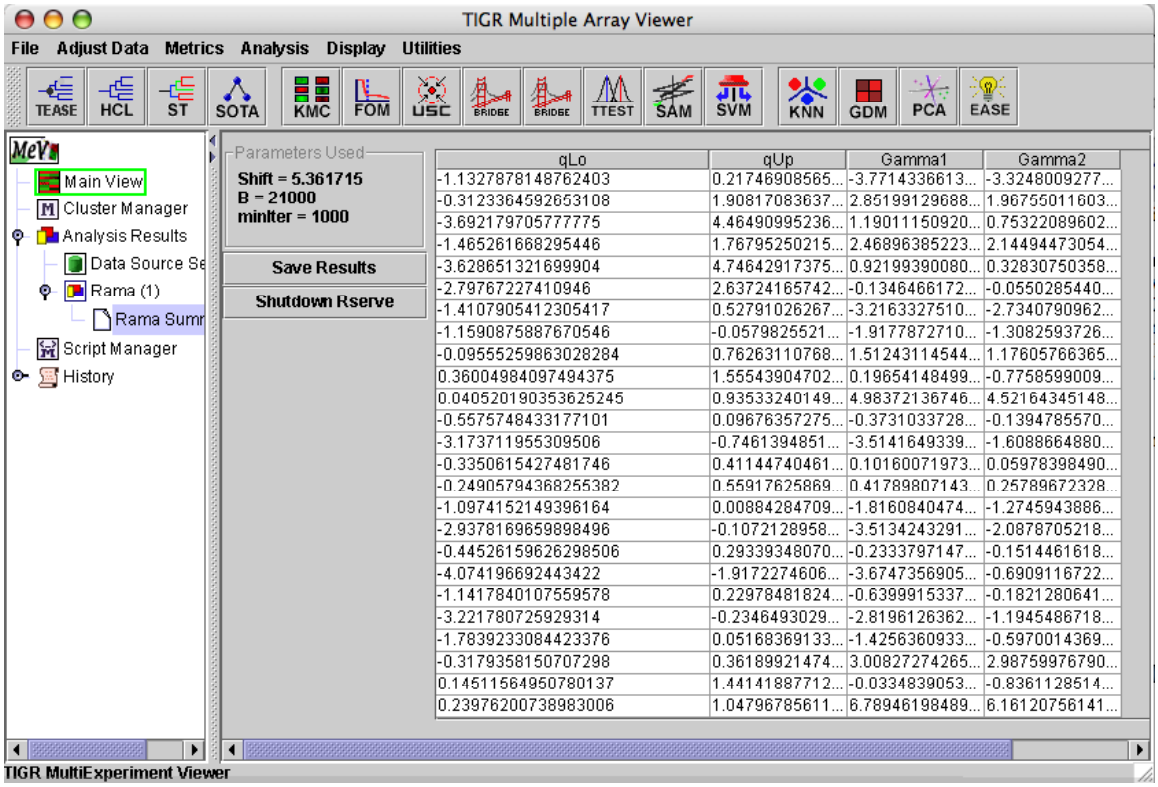
Rserve Connection
RAMA is a package written in the R programming language and requires a connection to a computer running RServe to function. See Section 8 for details on installing R and RServe.

By default, Bridge will look on the local machine for an Rserve server. However, since Rserve is a TCP/IP server, theoretically it could be running anywhere. The user need only enter an IP address and port number separated by a : in the Text Field “Enter a new location”. By clicking “Add”, the new location will populate

the pull down menu. It will be saved to the user's config file and be available for later use.

Rama Results

Sample output from this module is shown below (11.26.2). Rama estimates intensities based on the data input. The new set of intensities replaces the loaded dataset and is available for further analysis. A table of the results is also made available.



RAMA Results: Table of results

5.3. Data Transformations

Log2 Transform

This is fairly self-evident, because it just takes 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:

Value = [(Value) – Mean(Row)]/[Standard deviation(Row)]

Hoever, most data is already normalized when loaded.

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 Adjustmenst

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 log10x, and it outputs log2x.

Log2 to Log10

This assumes that the current data are log 2 transformed, and transforms them to log base 10, i.e., it assumes that the input data is in the form log2x, and it outputs log10x.

Unlog2 transformation

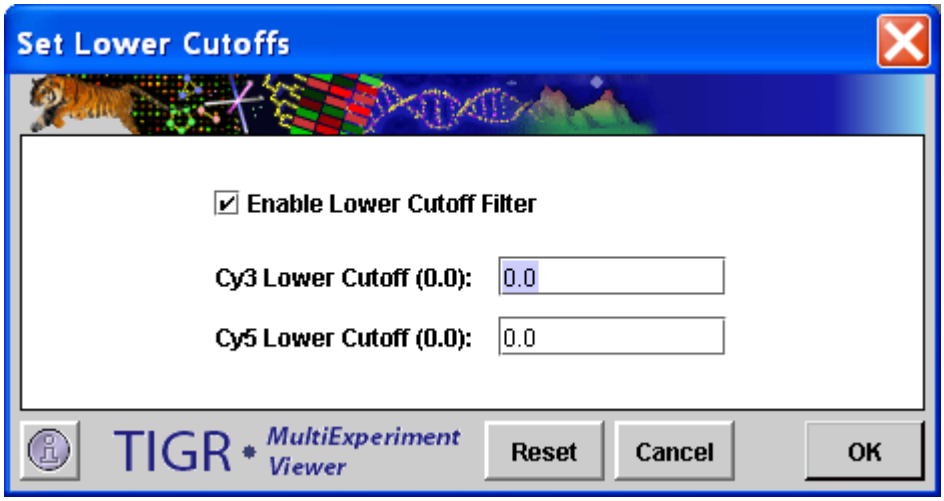
This assumes that the current data are log2 transformed, and removes the log2 transformation.

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

Lower Cutoffs

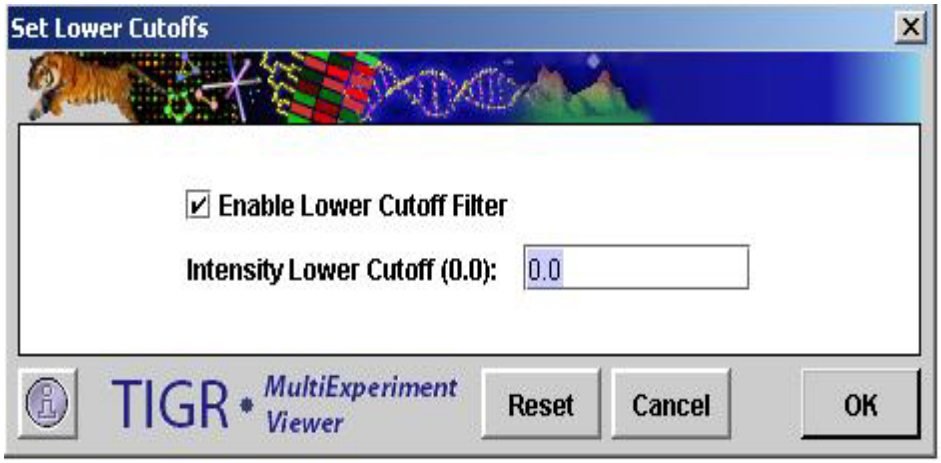
Select *Use Lower Cutoffs* to exclude from analysis any genes for which the expression values are lower than specified values. There are two options in this menu. One is for two color arrays and one is for single-color arrays.

For two-color arrays: select *Adjust Data* → *Data Filters* → *Low Intensity Cutoff Filter* → *two color microarray* to set either the corresponding Cy3 or Cy5 columns. To enable this option, check the "Enable Lower Cutoff Filter" 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.



5.4.1. Lower Cutoff Filter Dialog for two-color arrays

For single-color arrays: select *Low Intensity Cutoff Filter* to set intensity lower cutoff. To enable this option, check the "Enable Lower Cutoff Filter" 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 intensity is above the specified threshold. This option is disabled by default.

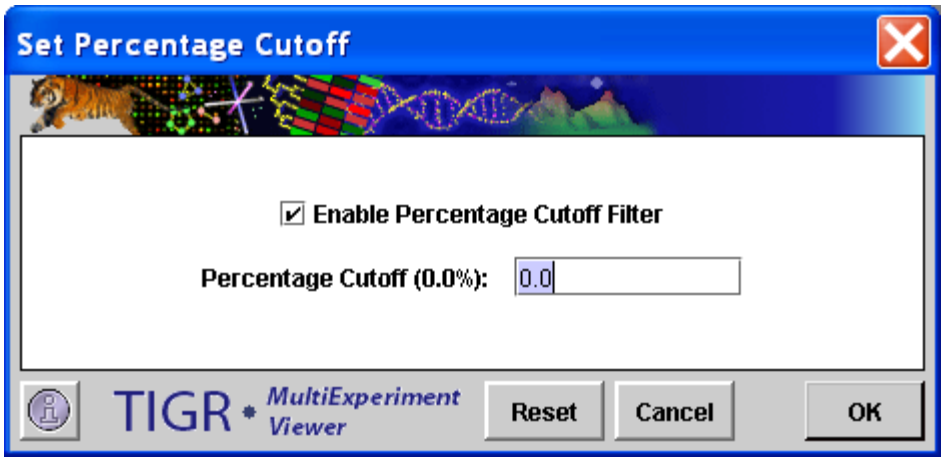


5.4.2. Lower Cutoff Filter Dialog for One Color Microarray

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 *Adjust Data* → *Data Filters* → *Percentage Cutoff Filter* and enter a percentage value. To enable this option, check the "Enable Percentage Cutoff Filter" 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.



5.4.3. Percentage Cutoff Filter Dialog

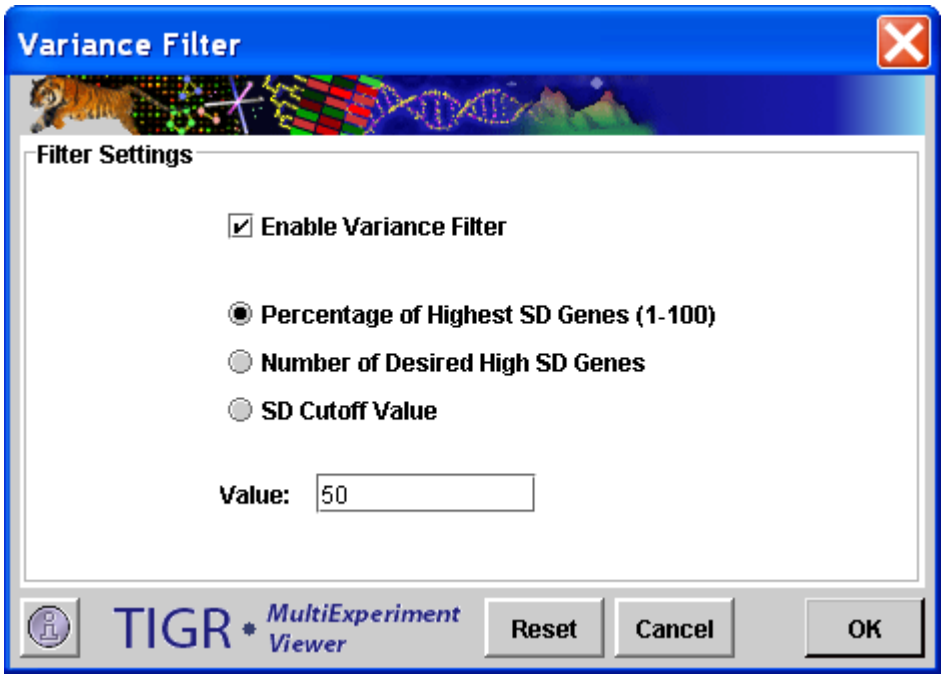
Variance Filter

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. *Select Adjust Data → Data Filters → Variance Filter*. 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 checked 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.

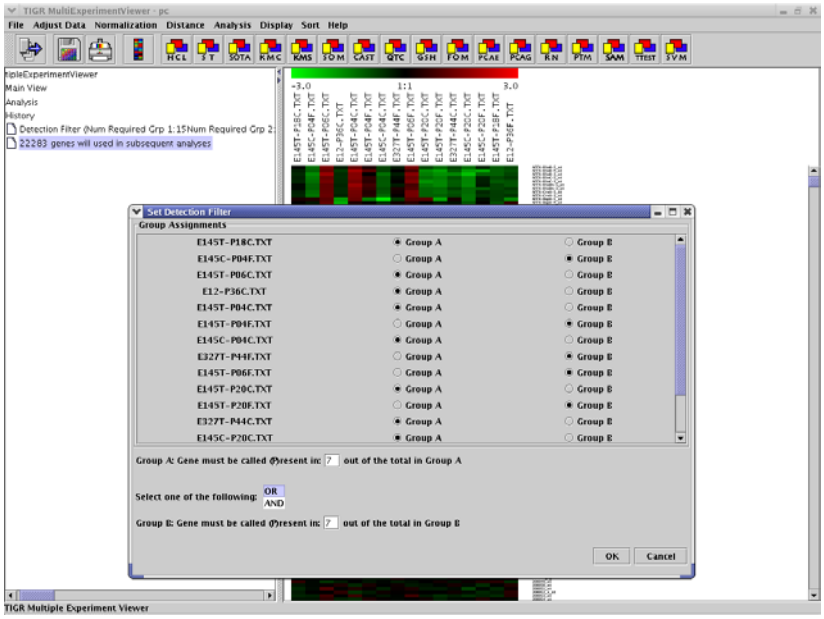


5.4.4. Variance Filter Dialog

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.



5.4.5. 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.

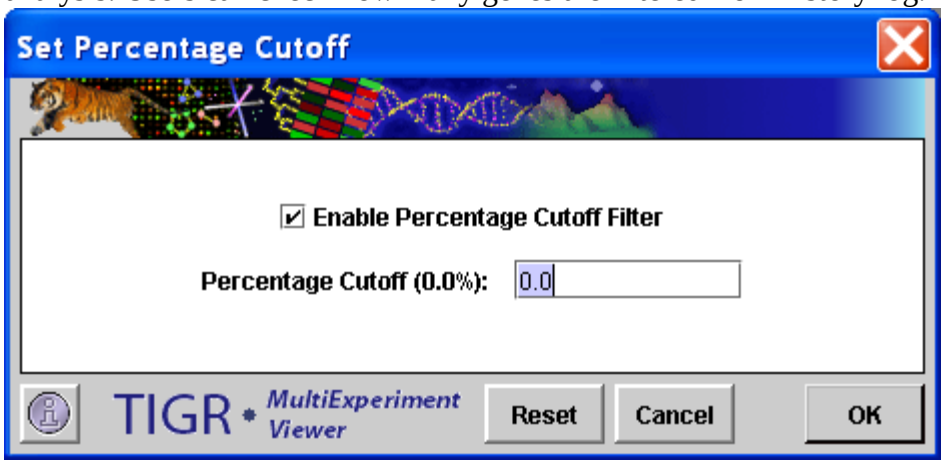
Adjust Intensities of Zero

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.

Bioconductor detection call noise filter

Select Bioconductor detection call noise filter to filter the genes for which the absent call percentage across all samples is above the level users define in the dialog. This will not delete any data, but will only exclude the genes from analysis. Users can check how many genes are filtered from history log.

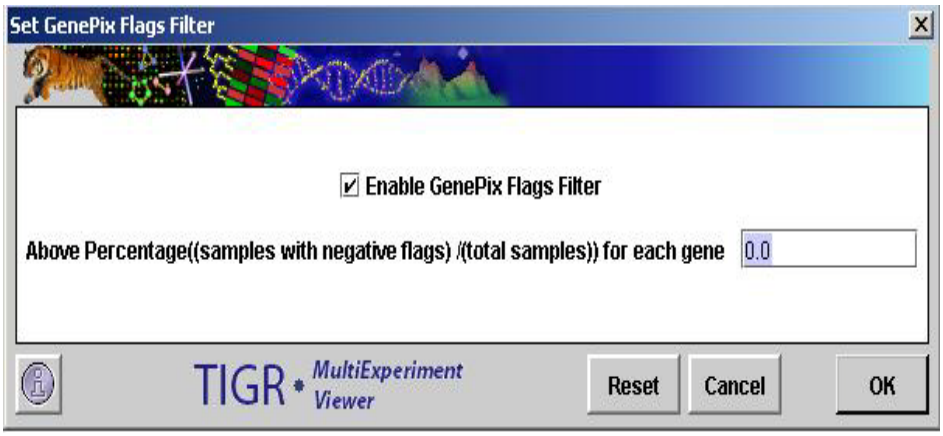


5.4.6. Set Detection Filter Dialog

GenePix Flags Filter

The GenePix Flags Filter allows the removal of genes for which the samples with negative Flags percentage across all samples is above the level users define in the dialog. 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 *Adjust Data* → *Data Filters* → *GenePix Flags Filter* and enter a percentage value. To enable this option, check the "Enable GenePix Flags Filter" checkbox just below the "Set GenePix Flags Filter" menu option, and uncheck it to disable this option. 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.

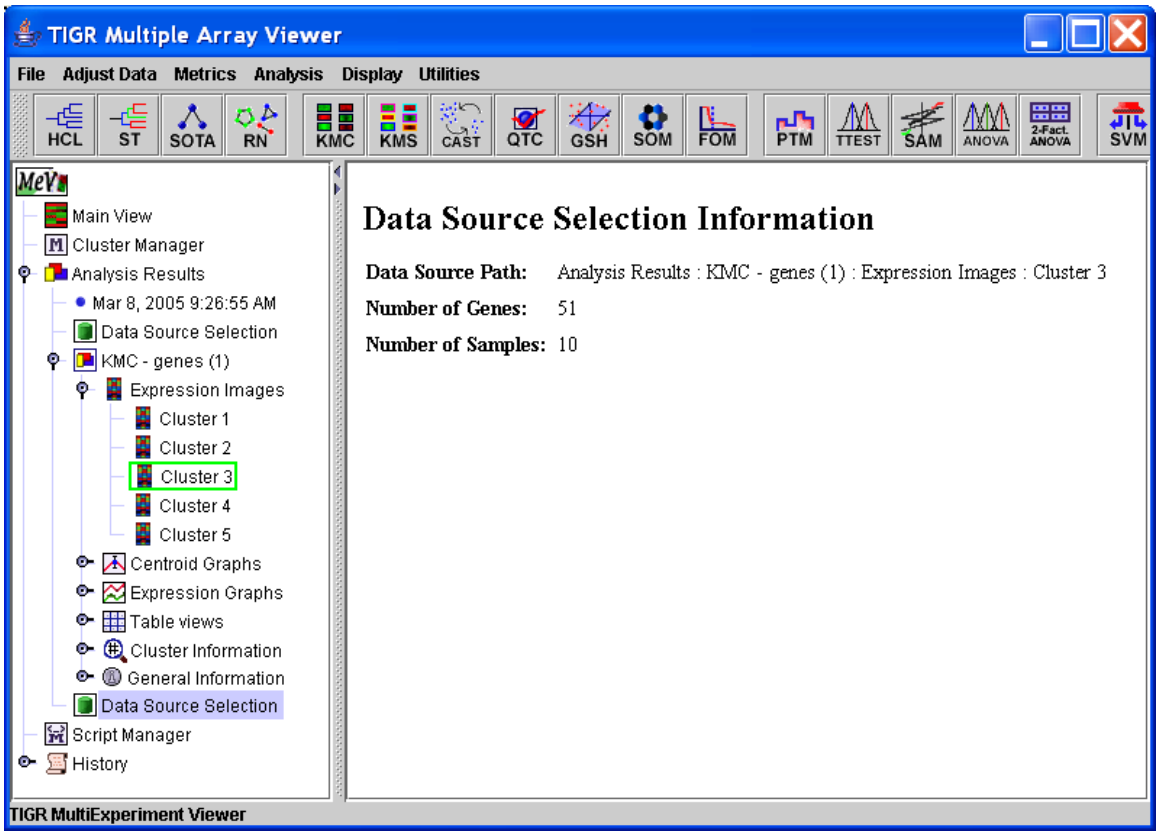


5.4.7. GenePix Flags Filter Dialog

5.5. 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.



5.5.1. 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.

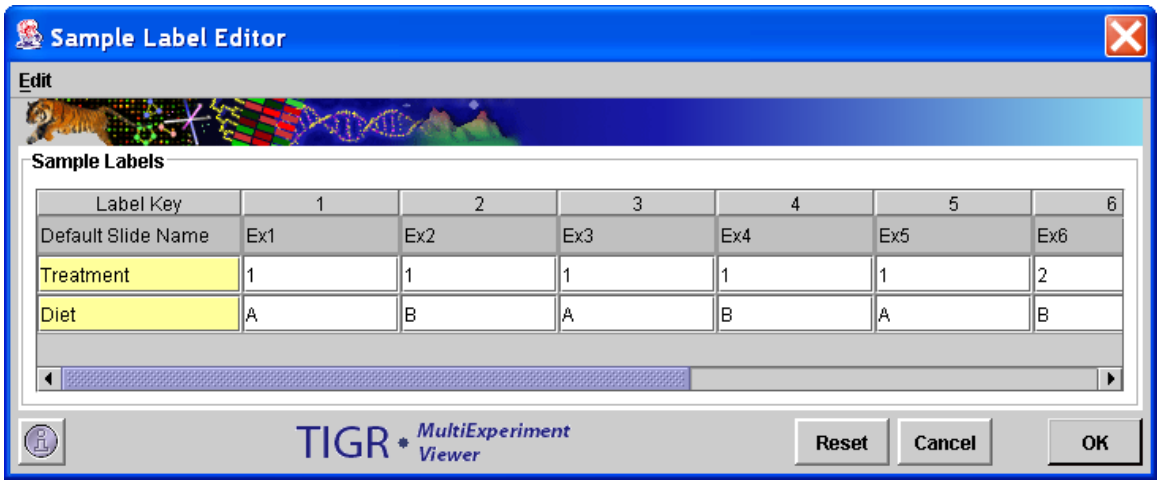
6.1. Sample Annotation

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 select by choosing *Display* → *Sample/Column Labels* → *Select Sample Label*. Expression images and expression graphs will label the 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 select U*tilities* → *Append Sample Annotation* 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 *File* → *Save Matrix* from the main 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.

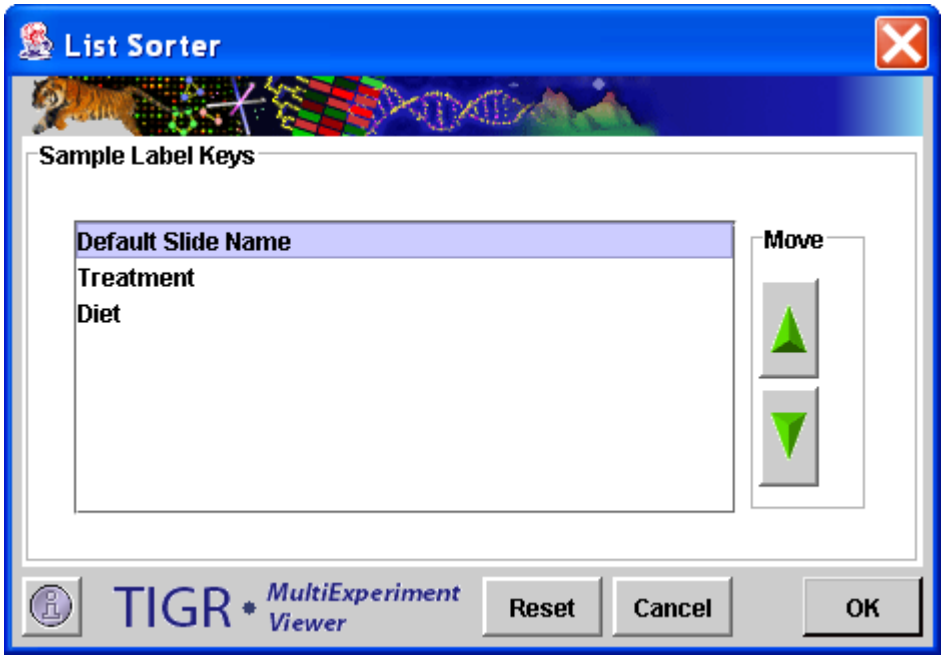
Note: If analyses have been run on the loaded data, the reordering option is disabled. This is done as a precaution against misrepresentation of prior results that may have relied on the original ordering. Typically, the reordering of samples should be done just after data loading to place associated 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 by choosing *Display* → *Color Scheme*. 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. MeV also provides "Custom color schemes" and "Accessible Color Scheme."

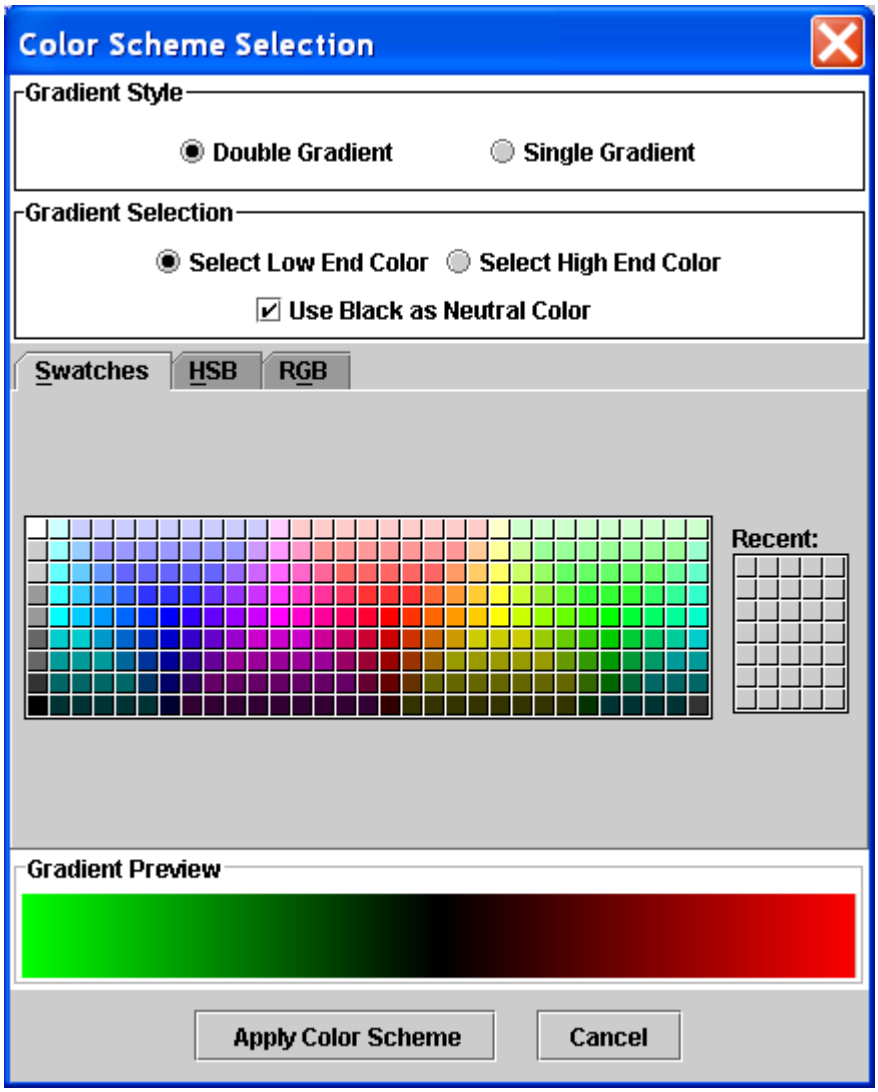
Custom color schemes can be created by selecting *Color Scheme* → *Custom Color Scheme*. The form will provide several options:

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.

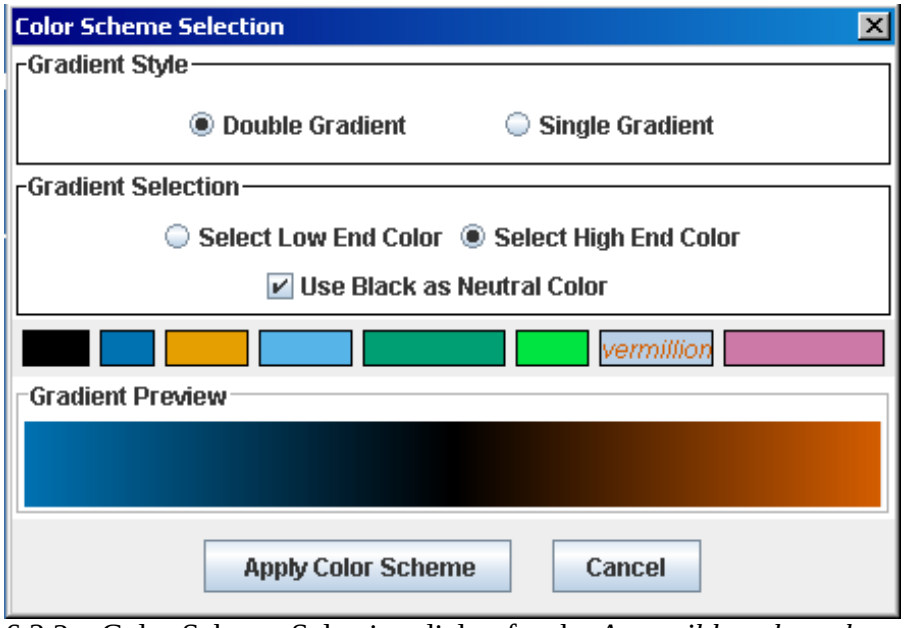
Gradient Preview: This area displays the current settings.



6.3.1. Color Scheme Selection Dialog

The Accessible Color Scheme was developed to make the images that MeV generates accessible for people with the color blindness. The color palette contains eight colors adapted from the web based resource: http://jfly.iam.u-tokyo.ac.jp/color/index.html

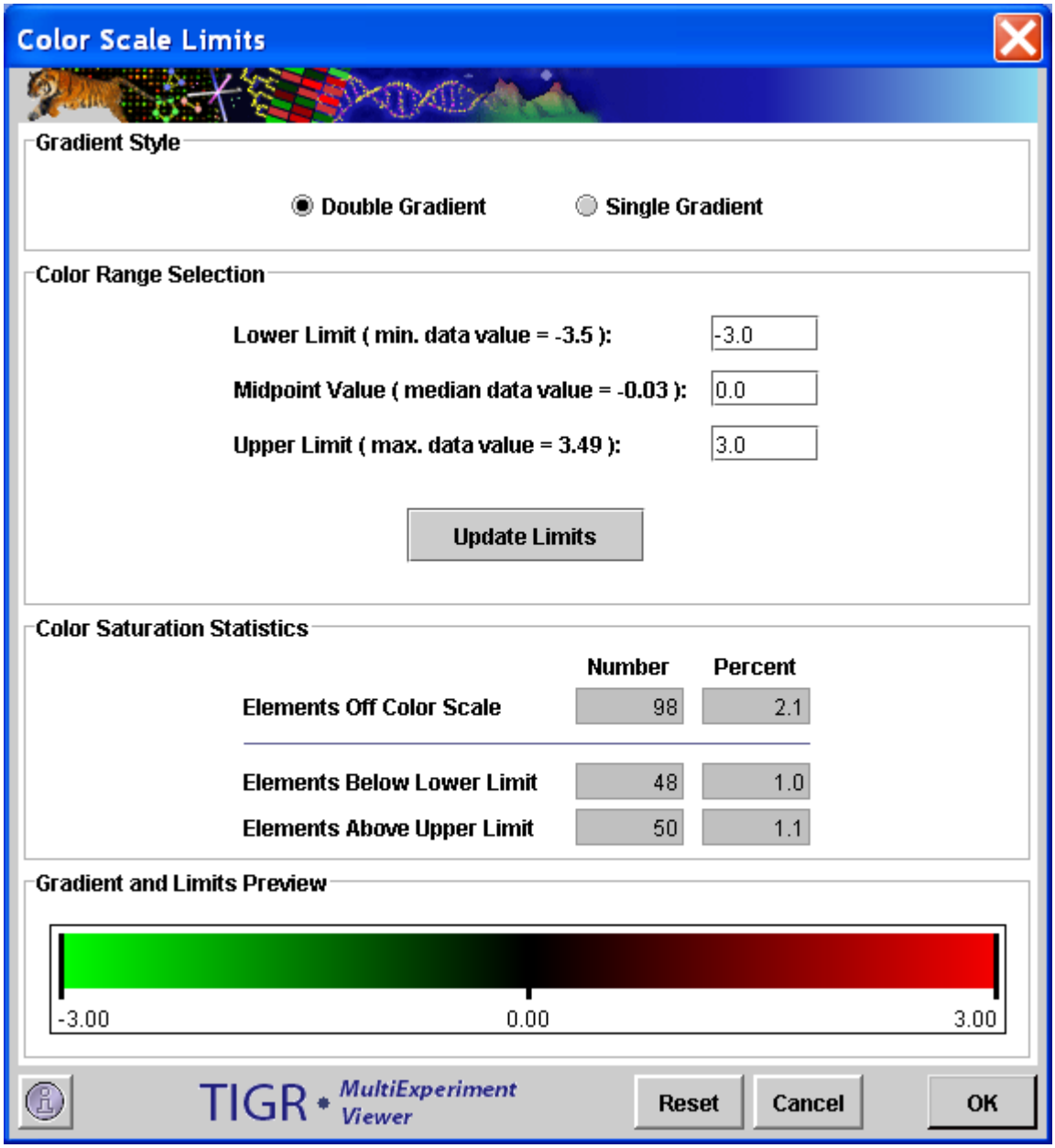
The Color Scheme Selection dialog (shown below), which appears when you select the *Accessible Color Scheme* option from the *Color Scheme* menu; provides choices similar to the *Custom Color Scheme* option (Gradient Style, Gradient Selection, Color Selector and Gradient Preview).



6.3.2. Color Scheme Selection dialog for the *Accessible color scheme*

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 it's 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.

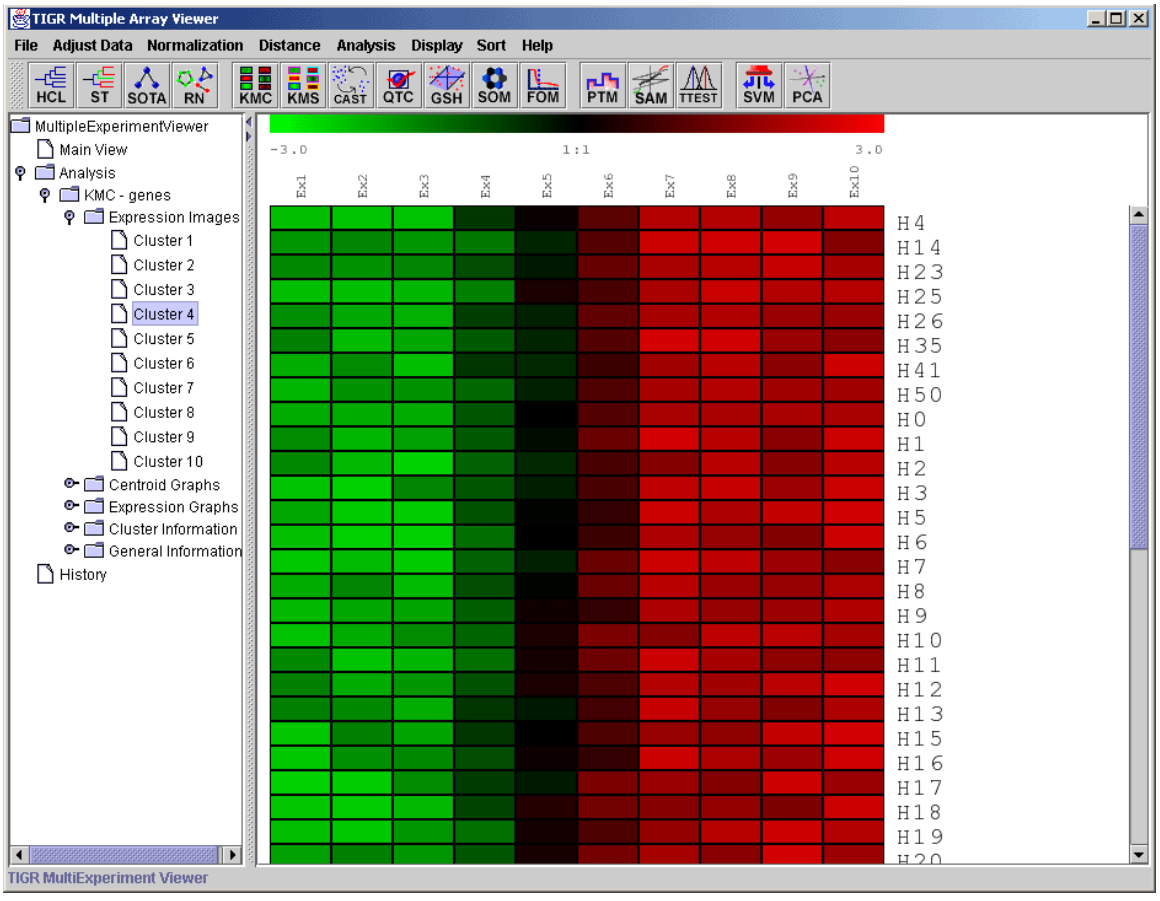
7. Viewer Descriptions

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

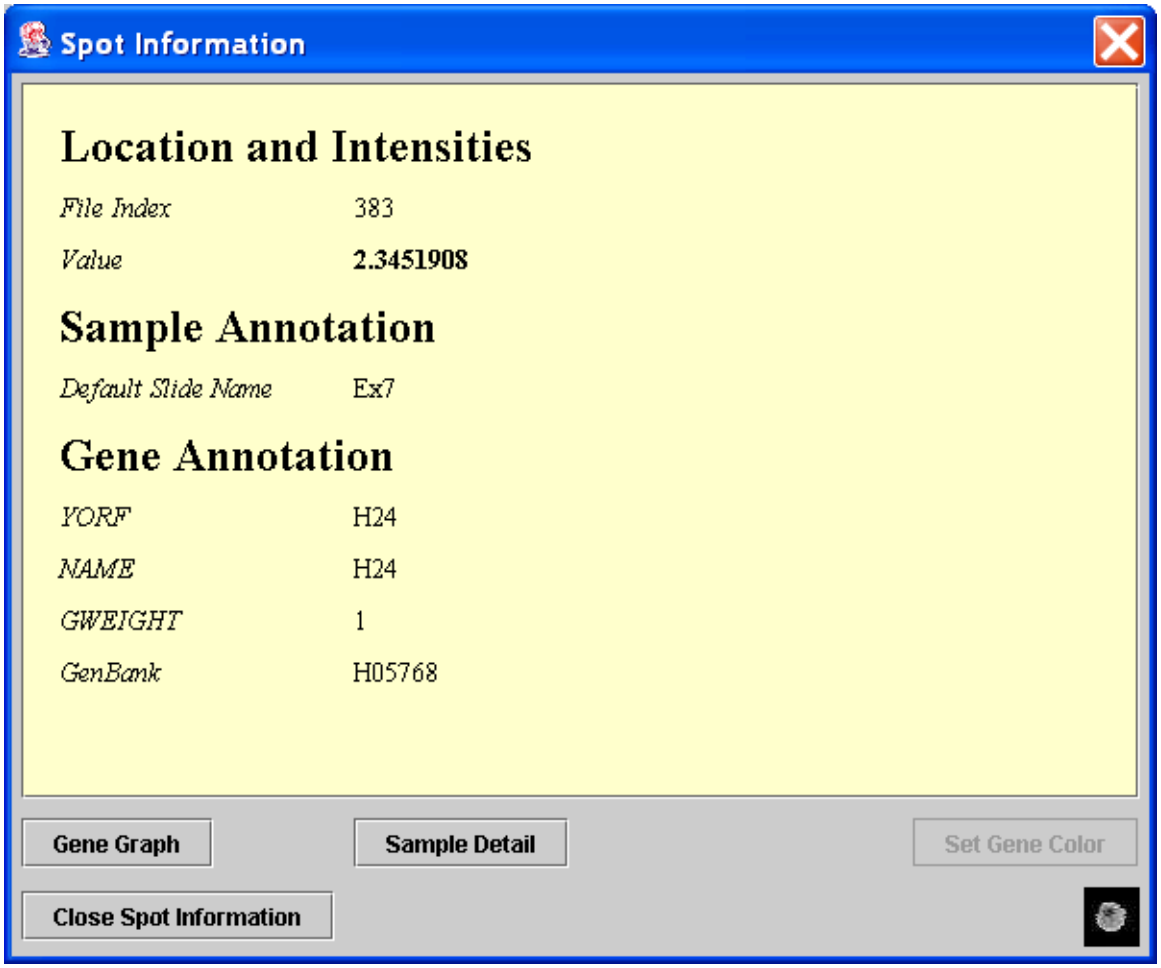
This viewer is used in the main window of the Multiple Experiment Viewer as well as in most of the modules (7.2.1). It consists of colored rectangles, representing genes, in a matrix. Each column represents all the genes from a single experiment, and each row represents the expression of a gene across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



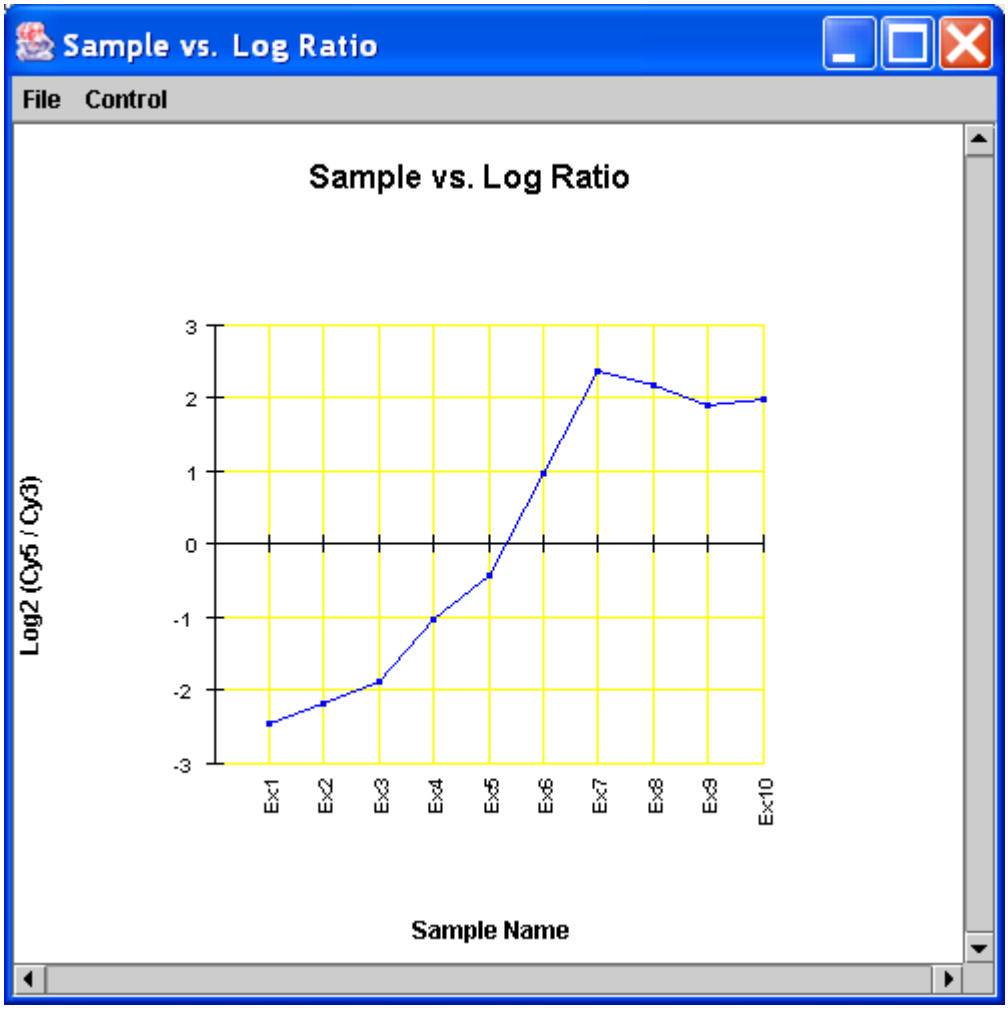
7.2.1. Expression Images Viewer

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

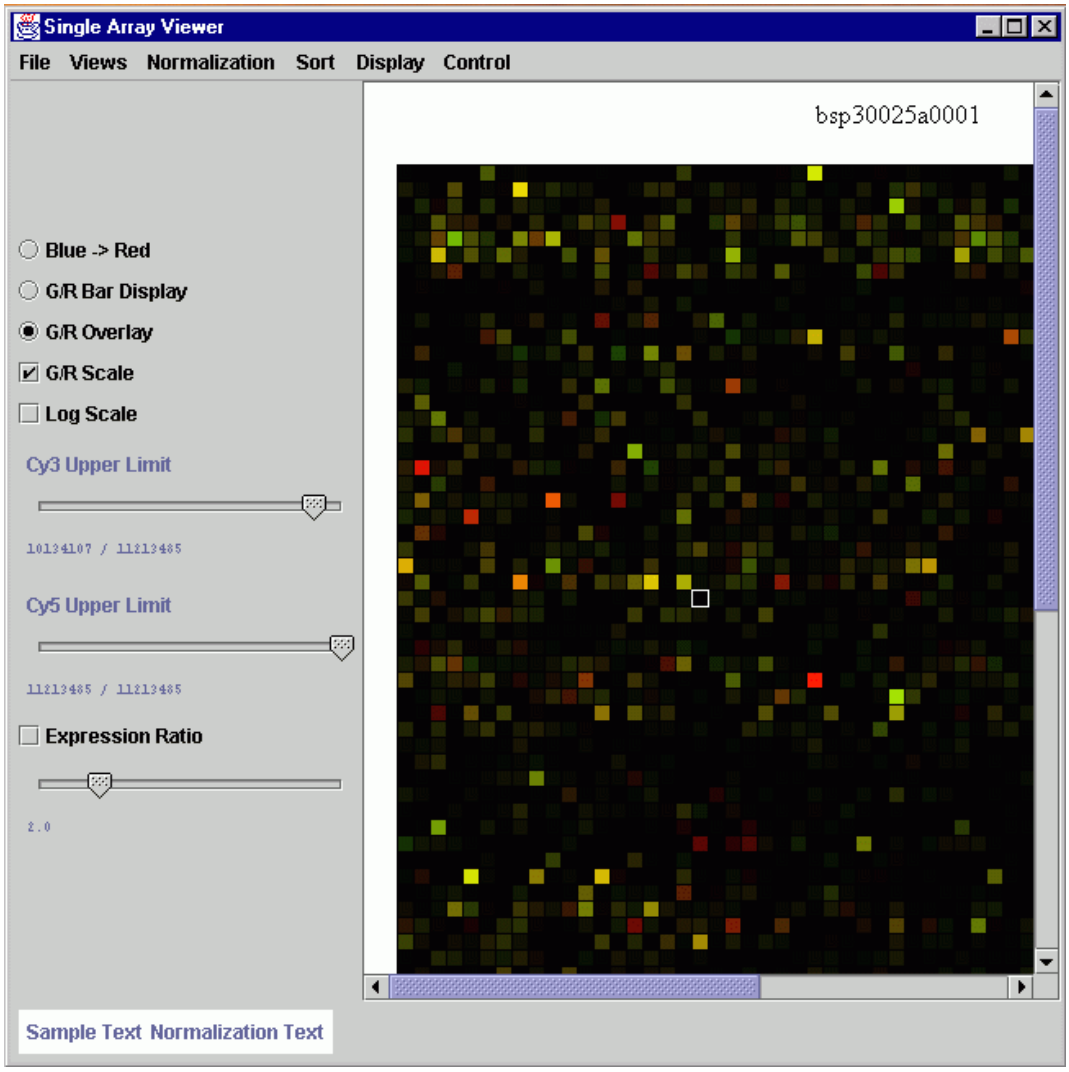
information, including expression values and row/column coordinates. Clicking the *Gene Graph* button will open a window containing a graph of this gene's expression level across all samples (7.2.3). To view a Single Array Viewer displaying the entire experiment of which this spot is a part, click the *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



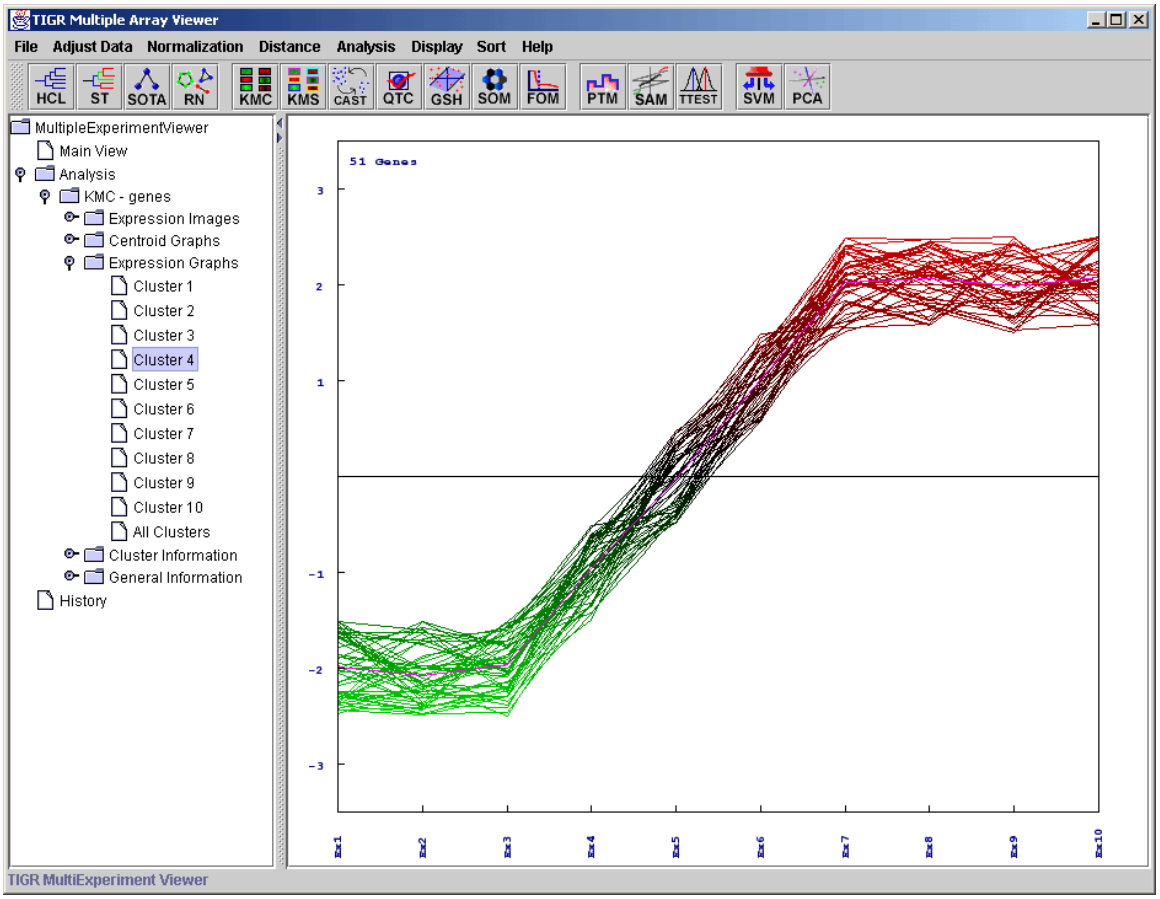
7.2.4. Sample Details displayed in a Single Array Viewer

7.3. Expression Graphs

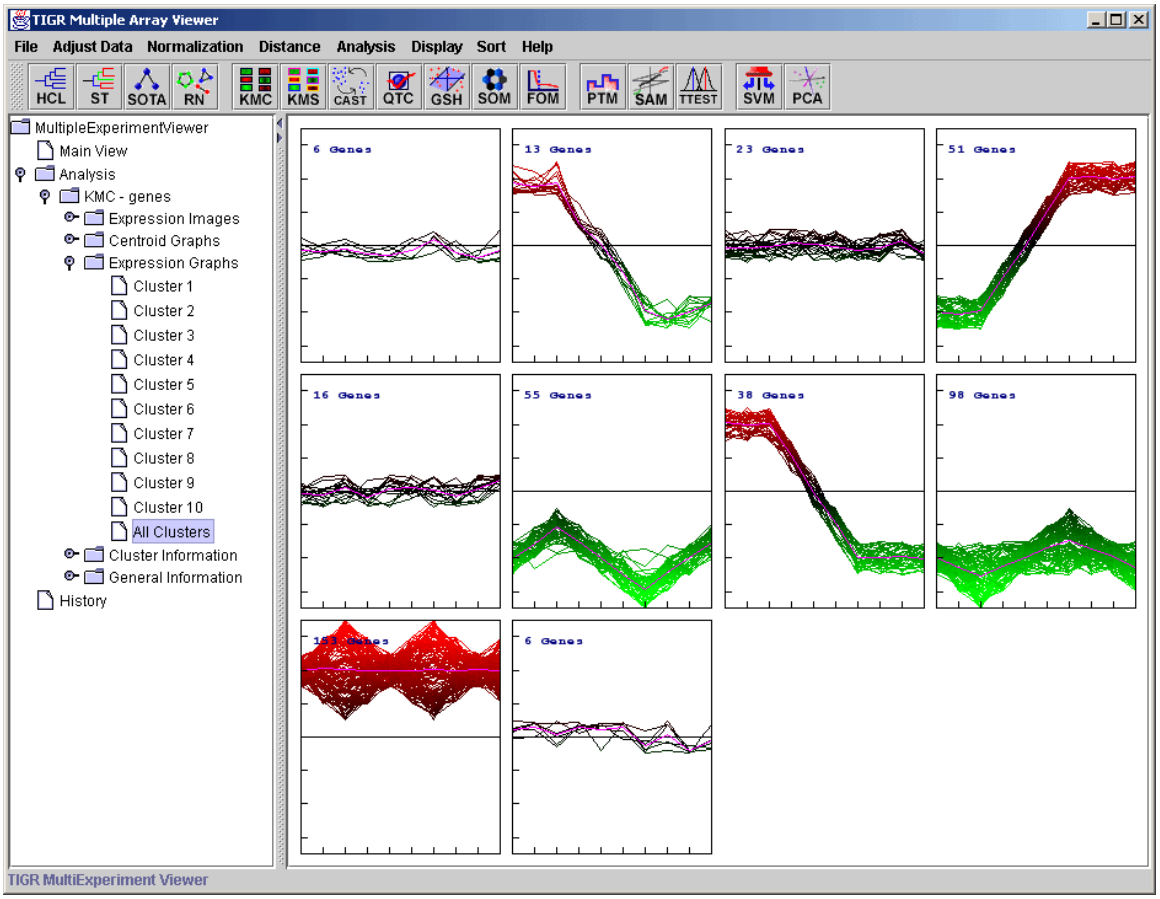
This viewer displays graphs of the expression levels of each gene across the experimental conditions (7.3.1, 7.3.2). By default, the line color is gray or colored according to the cluster membership. The lines can be set to display a gradient coloring option in which the lines are colored according to expression level. This option is found by selecting *Display* → *Color Scheme* → *Use Color Gradient on Graphs.* This will toggle the 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. Each line represents the graph of an individual gene, and combined they created the whole centroid graph. The red line represents the graph of the mean of all the genes across the samples. If you previously saved and color coded your clusters, those colors will appear on the graph if you *haven't* turned on the *Use Color Gradient on Graphs* feature. This allows you to see the patterns, if any, of the behavior of the genes across the samples.

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.



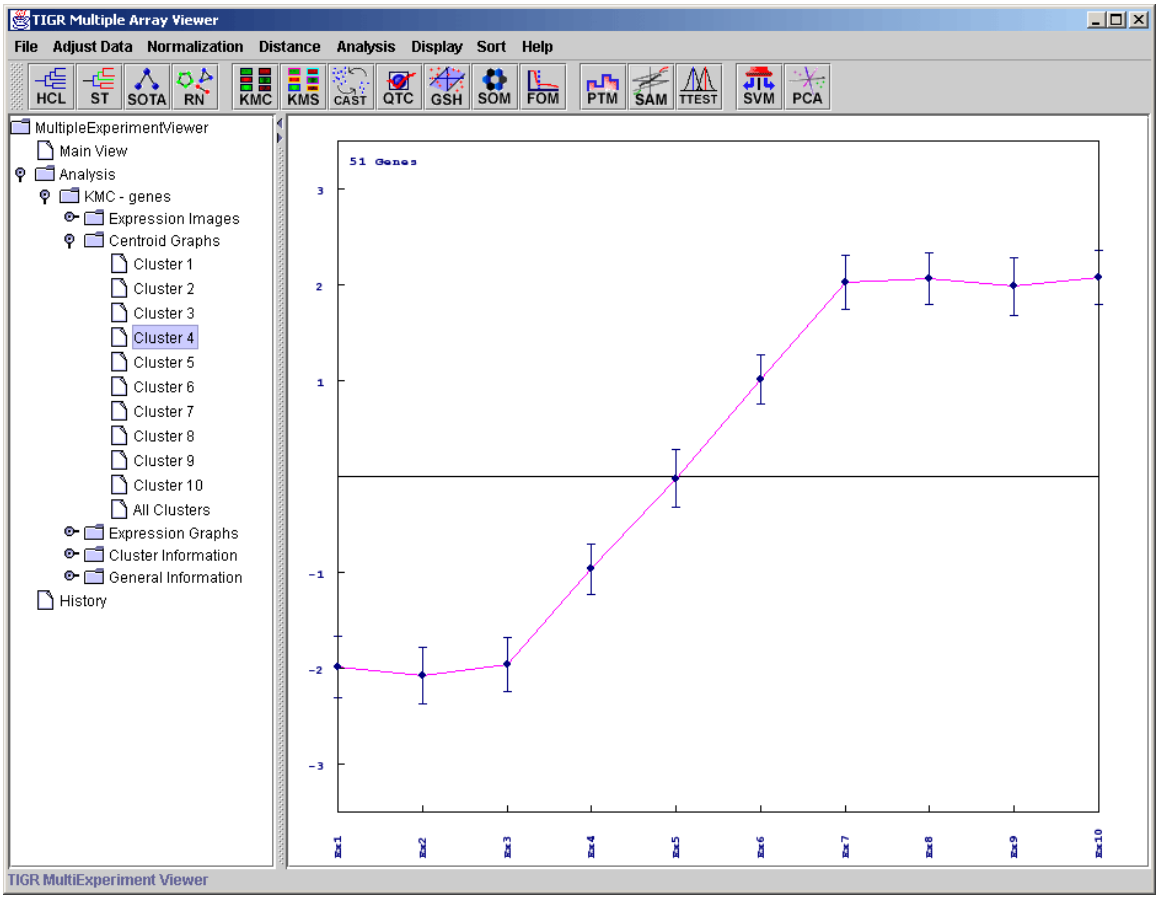
7.3.1. Expression Graph of one cluster (with gradient color selected)



7.3.2. Expression Graphs of all clusters

7.4. Centroid Graphs

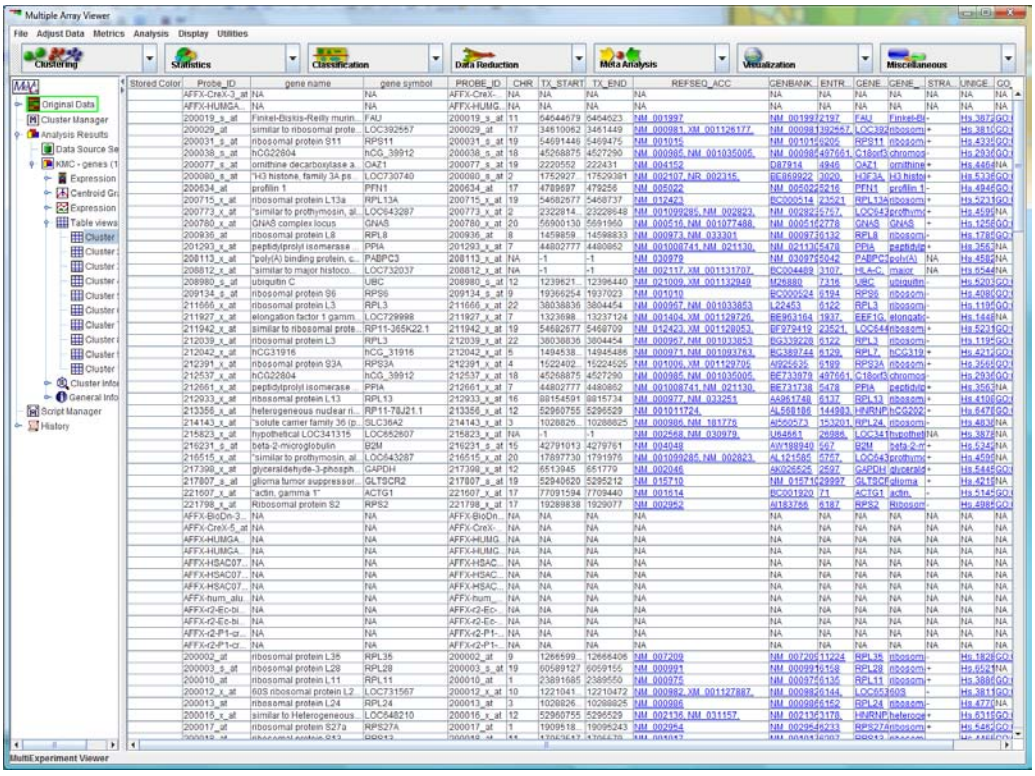
This viewer is very similar to the Expression Graph Viewer, except that line graphs for individual genes are omitted, leaving only the centroid graph. Error bars represent the standard deviation of expression within the cluster (7.4.1). That is, the mean for all the genes of a sample is represented by a dot, and the standard deviation for the genes in that sample are shown above and below it. In fact, the red line connecting the centroids is the same red line as in the expression graphs.



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

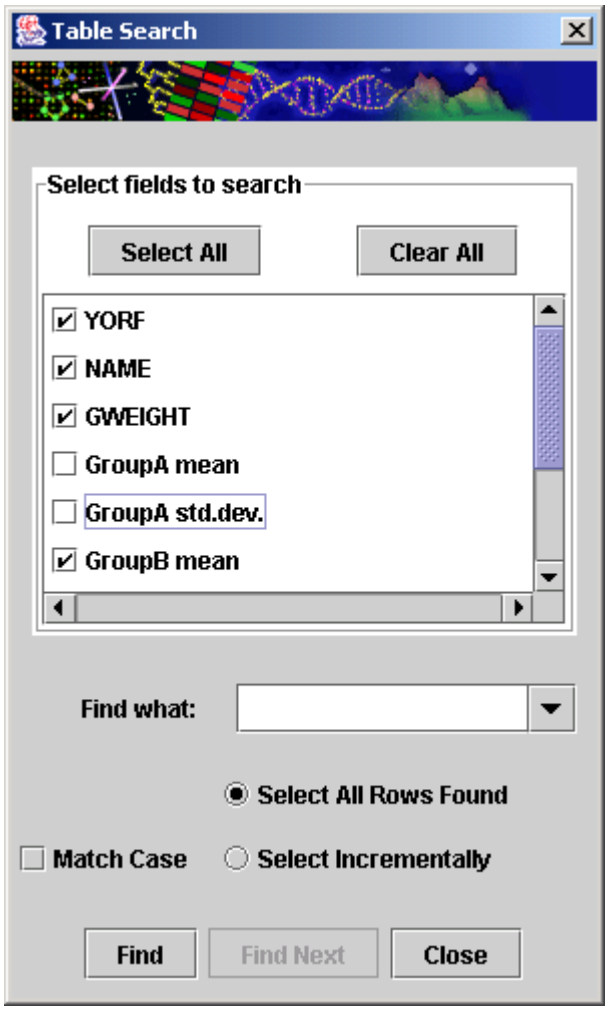
Columns can be dragged horizontally across the table to change their relative ordering. Successive clicks on the header for any column will sort the rows in ascending or descending order of the entries in that column. Sorting of the “Stored Color” column will bring together elements that have been stored 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).

Left-clicking on any of the blue-underlined cells in the table will launch a web browser and display the contents of that cell in a web page appropriate to the type of annotation contained within it. The mappings that determine which web site MeV will attempt to access for each annotation type can be changed my modifying the appropriate configuration file. A template for this file can be found on the TM4.org website, at http://www.tm4.org/mev/annotation_URLs.txt. Simply download the file and modify it to contain annotation name / URL mappings you wish to use. These mappings will **override** the default mappings MeV uses. Load the file using the Preferences -> Select Annotation Linkout File option from the main MeV toolbar. The name and location of this file will be saved in MeV’s user preferences file, and re-loaded in subsequent sessions.

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. 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 an 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 method 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.

Broadcast Matrix to Gaggle

Choose this option to broadcast the entire contents of this viewer to the Gaggle network. To broadcast to a specific goose, choose that goose from the *Utilities*->

Gaggle -> Broadcast Target menu. More details about the Gaggle network are available in the section x of this manual.

Broadcast selected Matrix to Gaggle (some viewers only)
The same as *Broadcast Matrix to Gaggle*, but broadcasts values only for the table rows that are currently highlighted.

Broadcast Namelist to Gaggle
Broadcast the annotation of the genes displayed in the current viewer to the Gaggle network. The annotation values that are currently displayed in the viewer will be sent to the goose selected in the *Broadcast Target* menu in the *Utilities* menu.

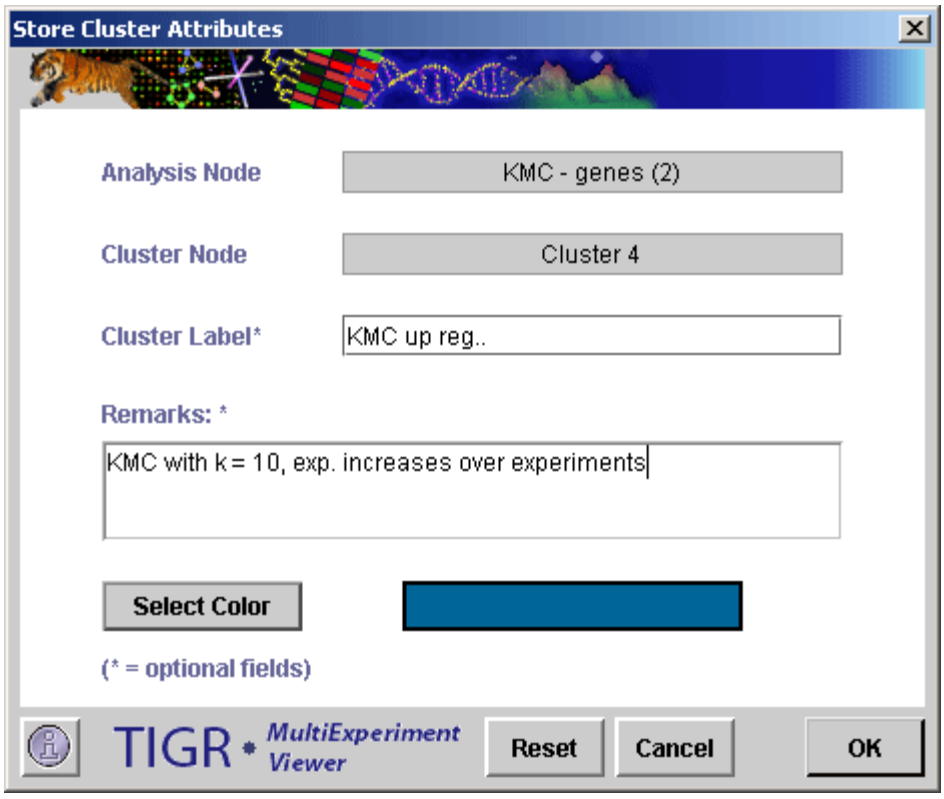
Broadcast selected Namelist to Gaggle (some viewers only)
The same as *Broadcast Namelist to Gaggle*, but broadcasts names only for the table rows that are currently highlighted.

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. Highlight the cluster of interest by clicking it, then in the right click menu select *Store Cluster*. 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. No two clusters may use the same identifying color.



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 or Factor* 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.

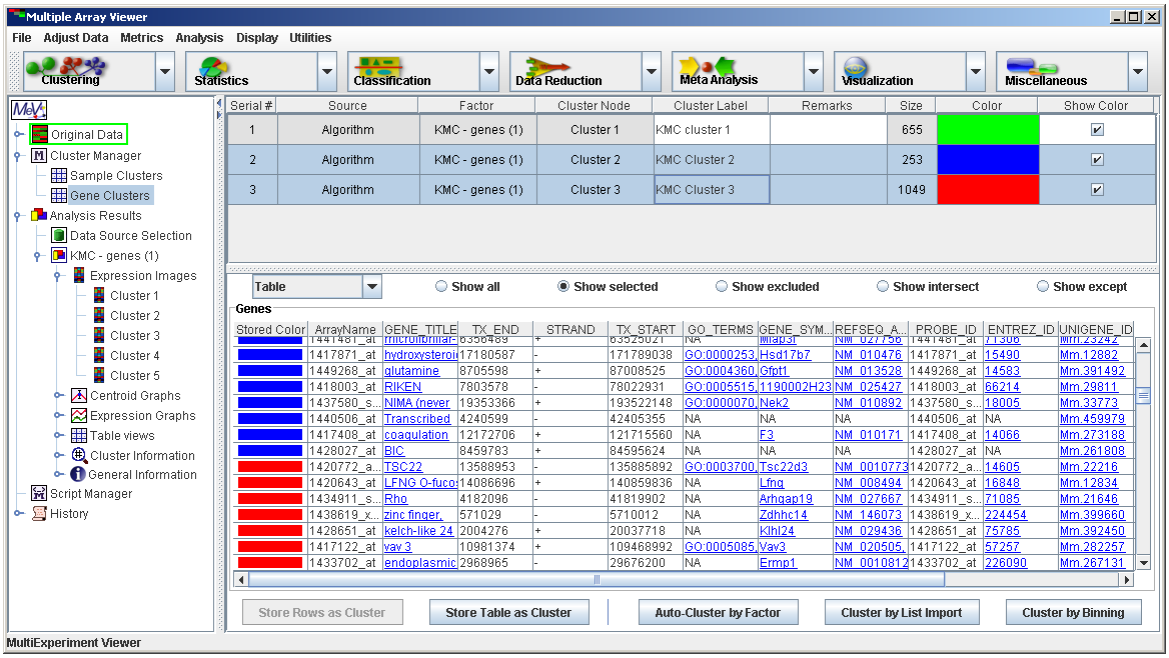
-The *Cluster Label* is an optional user defined name for the cluster.

-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. If you click the color box a screen will show that allows you to change the color if you wish.

-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.



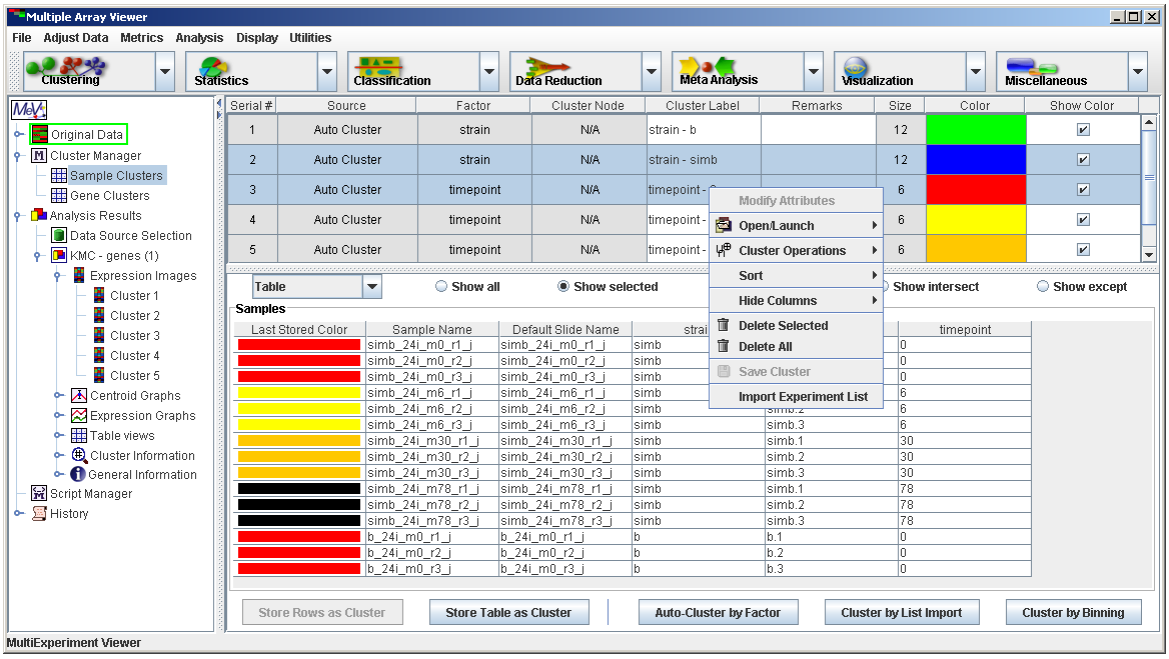
8.1.2. Gene Cluster Manager

Users have 4 options for display types and 5 options for the data to display. Four MeV viewers are accessible from the Cluster Manager-

1.) Table: This option displays a table in the Cluster Manager showing every available annotation for each gene or sample included.

2.) Expression Image: This option displays a heatmap in the Cluster Manager showing expression values for each gene and sample included.
3.) Expression Graph: This option displays a graph in the Cluster Manager showing expression values for each gene and sample included.
4.) Centroid Graph: This option displays a graph similar to the Expression Graph, but individual lines are omitted. Instead, a single mean expression line is displayed and bars representing standard deviation at a particular sample/gene are shown.

Five data options are available based on the clusters selected.
1.) Show All: All elements in the loaded data are shown, regardless of cluster membership.
2.) Show selected: All elements belonging to *any* of the currently selected clusters are displayed.
3.) Show excluded: All elements that belong to *none* of the currently selected clusters are displayed.
4.) Show intersect: All elements belonging to *every* currently selected cluster are displayed.
5.) Show except: All elements belonging to *one and only one* of the currently selected clusters are displayed.



8.1.3. Sample Cluster Manager with 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. Double-clicking on a cluster will open the *Modify Attributes* dialog.

-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 ClusterViewer* 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. *Union* 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.

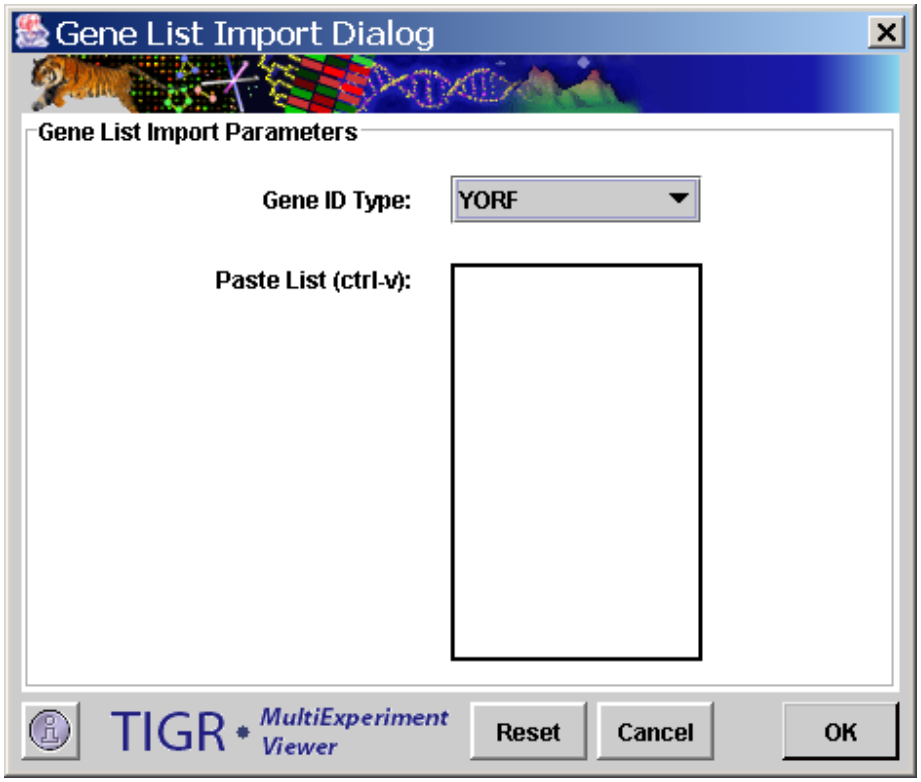
-Options also exist to delete selected clusters or all clusters in the list as well as to save a selected cluster to a specified file.

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.

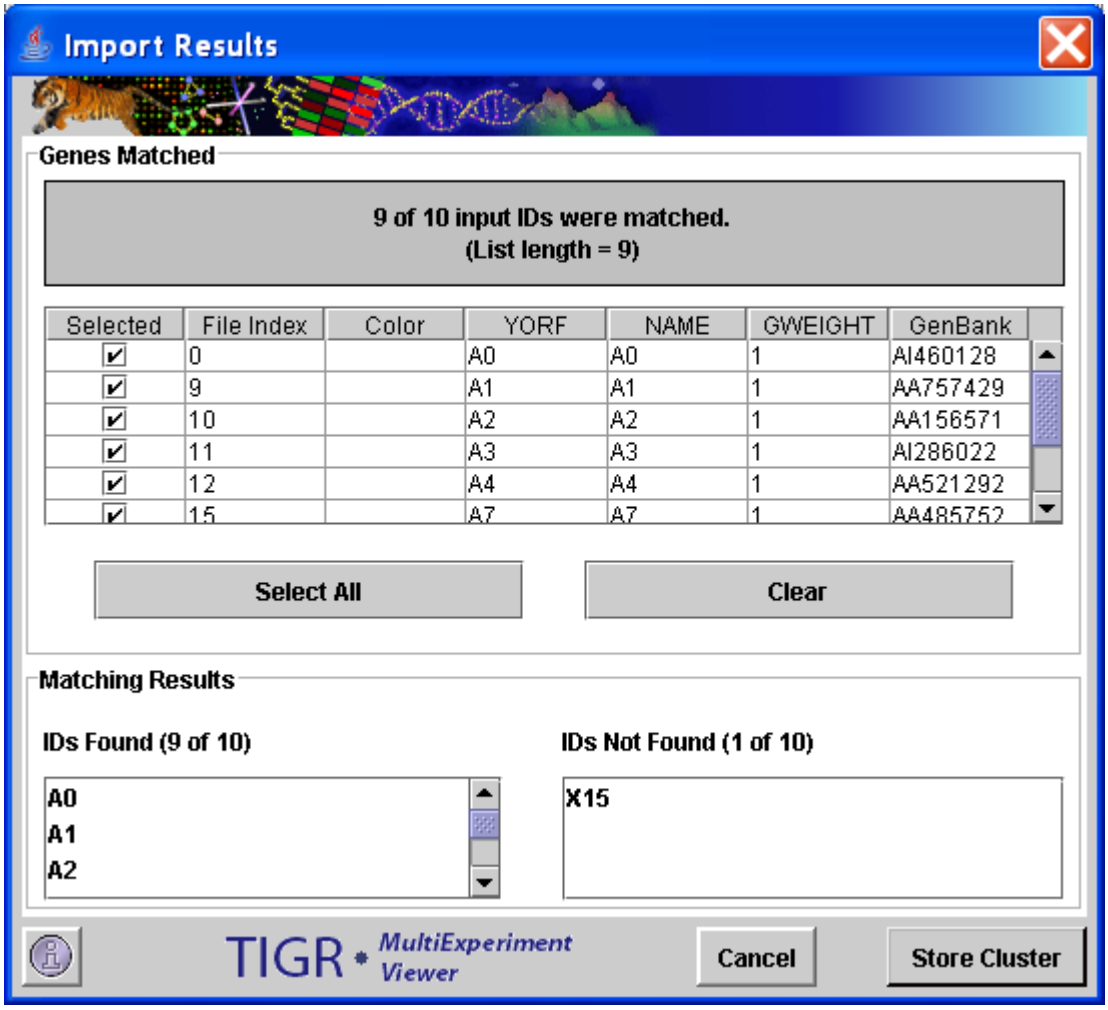
You can also *Save Cluster Data* to a tab-delimited text file. Selecting this option from the right-click menu will cause a file chooser to appear. Select a file name and a place to save row/column data, log ratio expression values, and (optionally) Cy3 and Cy5 values for each gene in the cluster. Selecting *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 to create a cluster based on supplied identifiers. For example, if you wish to make a cluster out of specific genes that you know are important, you can paste those genes into the dialog box and the cluster will be created out of those specific genes you pick. 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. 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. 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.

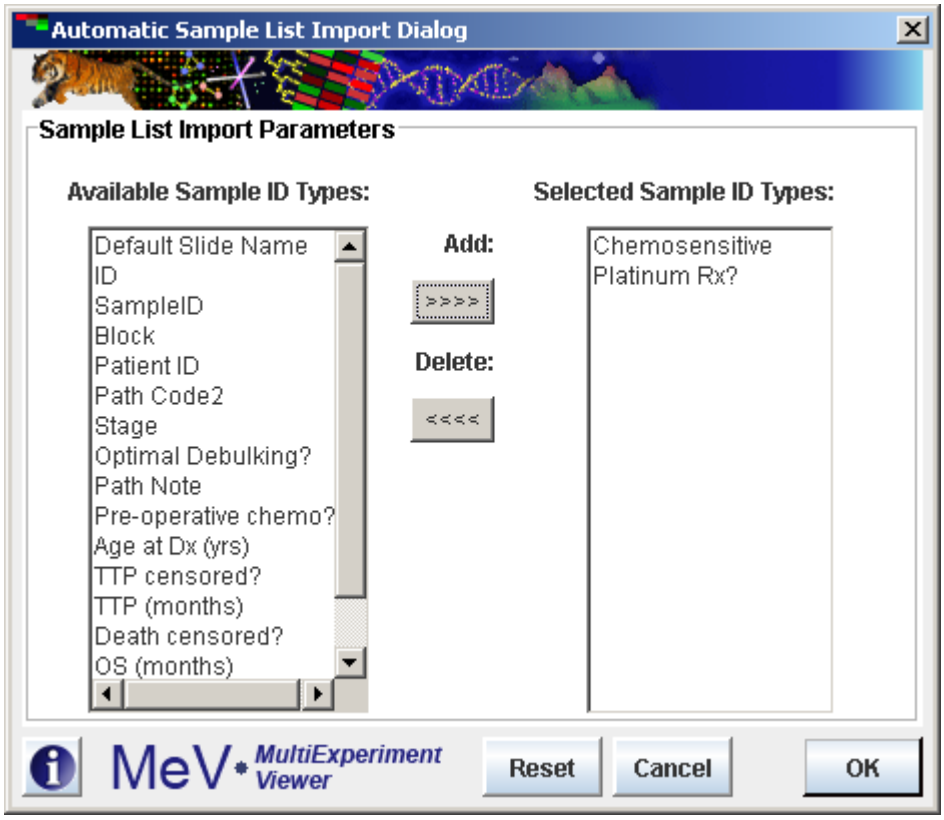


8.1.4. List Import Dialog



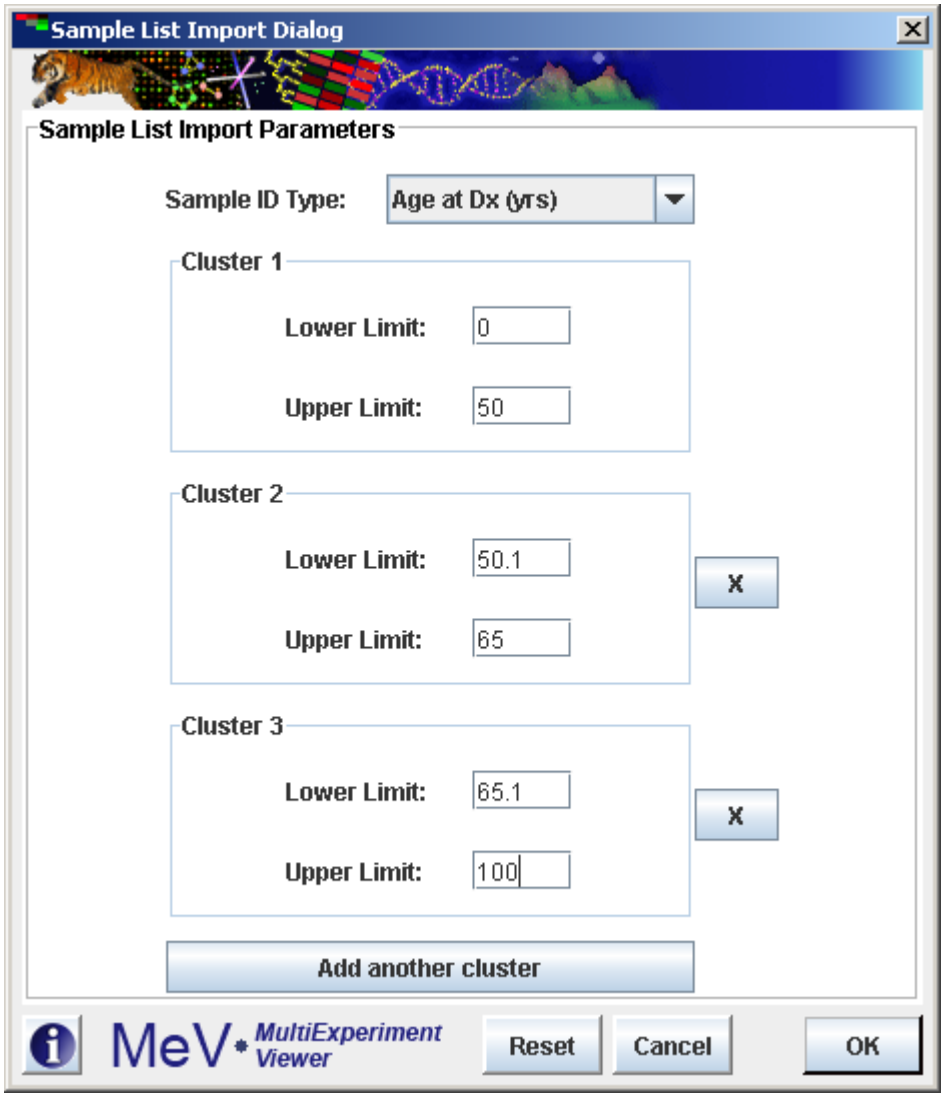
8.1.5. Import Result Dialog

The Automatic Cluster Import Feature allows the user to create clusters by supplied identifiers for a specific annotation type. Select *Utilities* → *Cluster Utilities* → *Automatic Cluster Import* and then choose *by Gene Annotation* or *by Sample Annotation*. A selection dialog box will appear allowing you to add any of your previously loaded annotation types into the "Selected ID Types". Click OK and all samples (or genes) within the selected annotation types will be grouped into unique clusters.



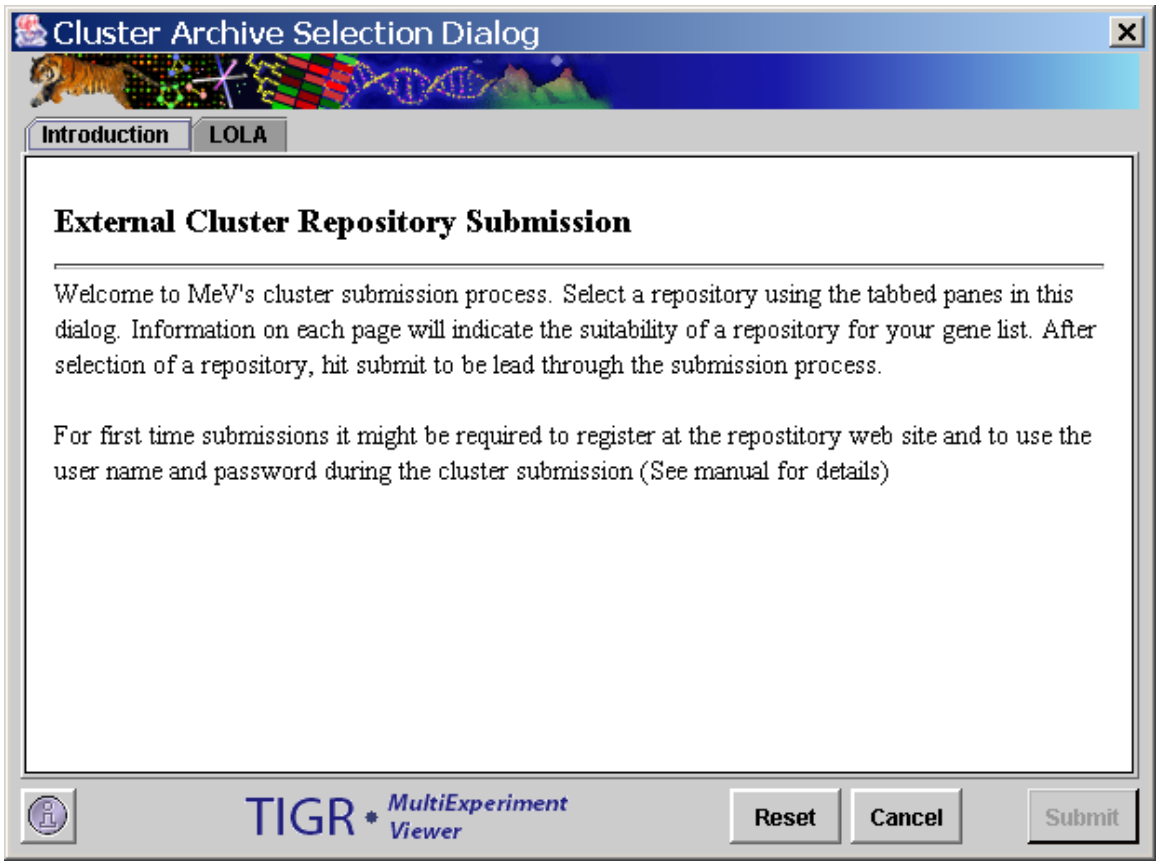
8.1.6. Auto-cluster by annotation dialog

The Binned Clustering Feature allows the user to create clusters using upper and lower limits for annotation type. Select *Utilities* → *Cluster Utilities* → *Binned Cluster Import* and then choose *by Gene Annotation* or *by Sample Annotation*. A selection dialog box will appear allowing you select limits for your cluster based on the annotation type that you've chosen from the drop-down box.
Click OK and all samples (or genes) falling within the given limits will be grouped into clusters.



8.1.7. Binned Cluster 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.



8.1.8. Repository Selection Dialog

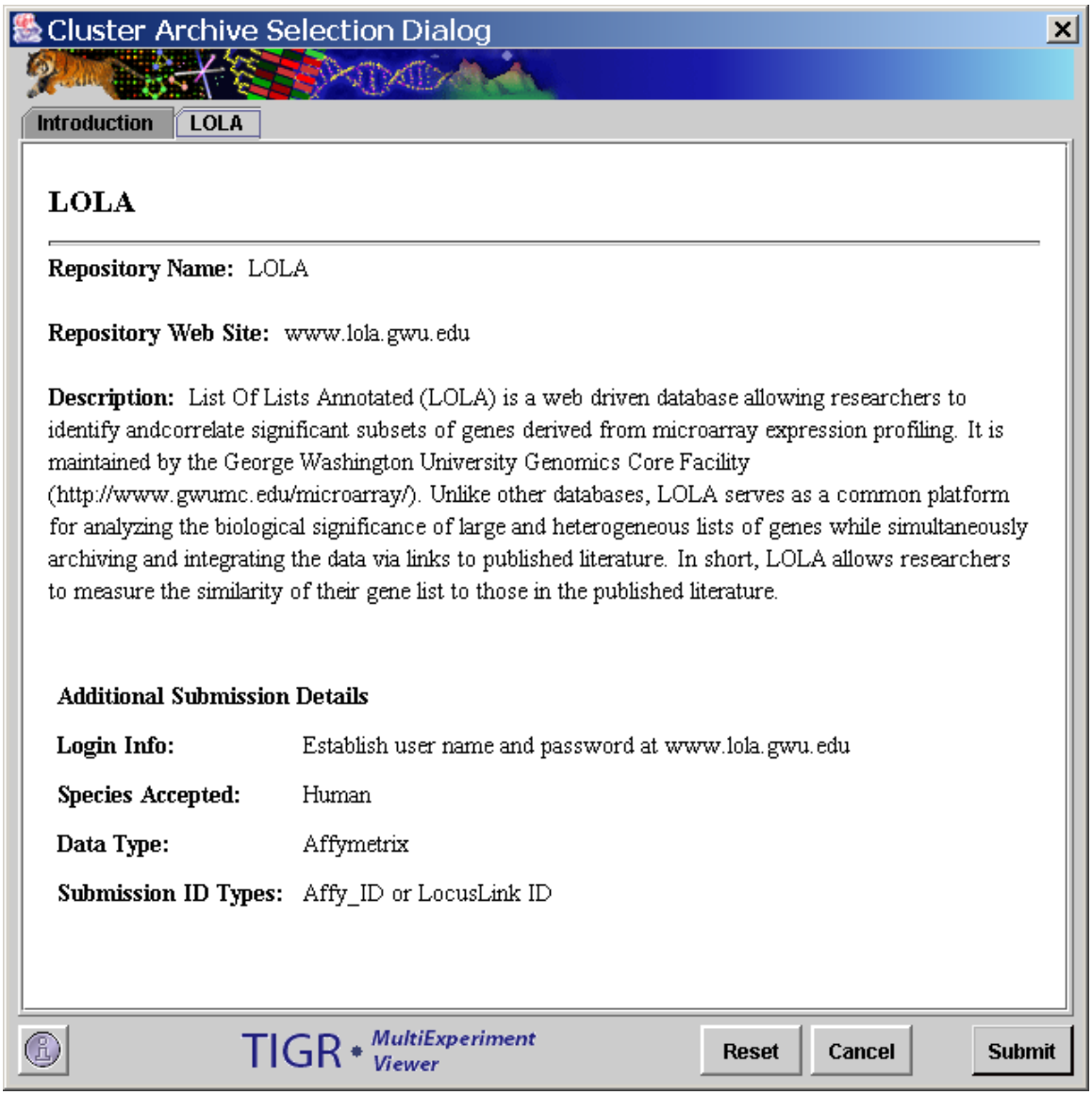
Note that some repositories may require user accounts with passwords to be established before the submission can be enabled. If a password is required, a dialog will be provided to enter a user name and password for the repository. Note that one can enter login information in the repository configuration 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.



8.1.9. Sample Repository Information Page

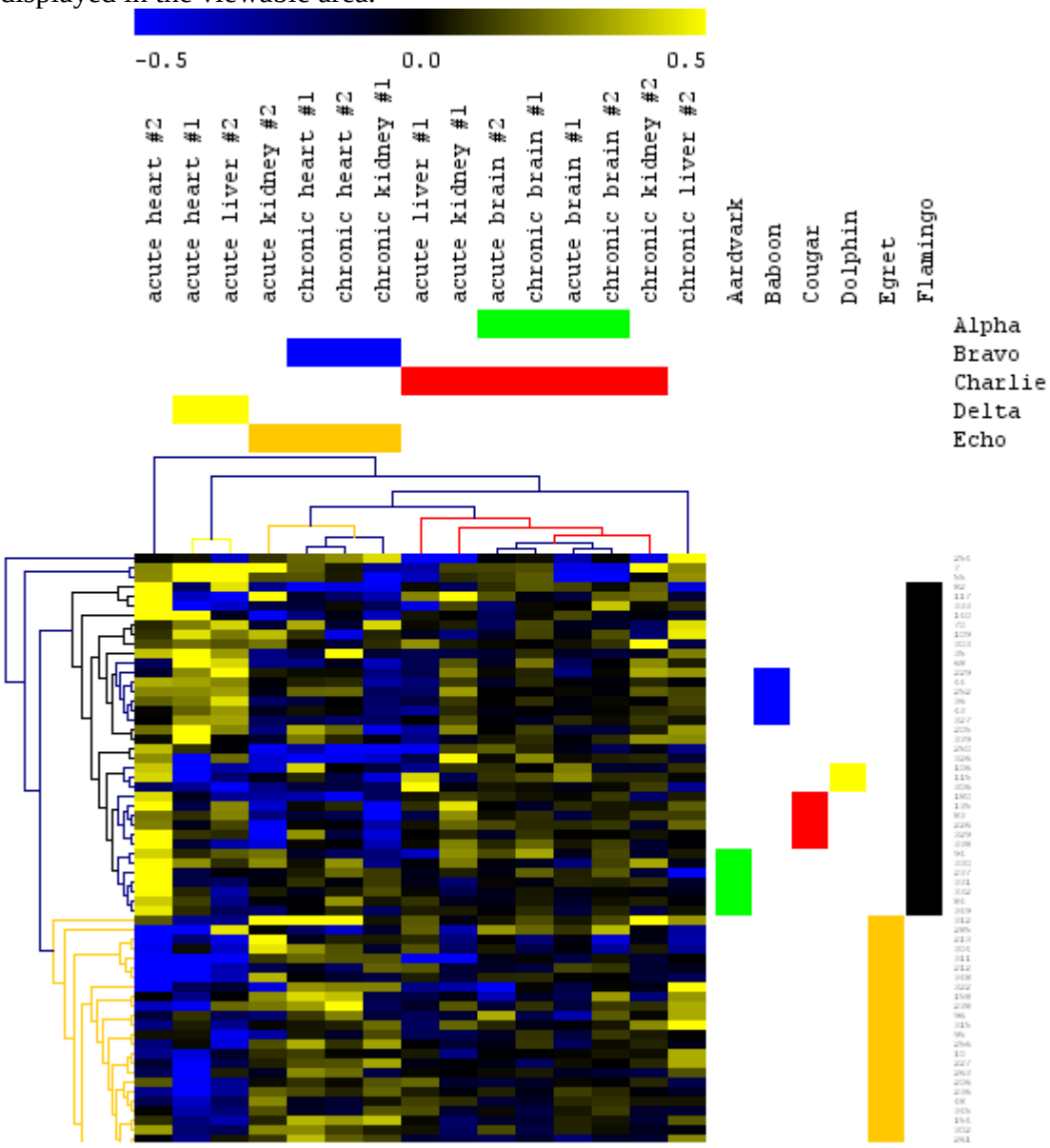
8.2 Cluster Display

A new feature in MeV v4.2 displays cluster assignments in a more informative and useful way on the heat-map. Cluster color-bars are now layered with their cluster labels displayed vertically or horizontally above and to the right of the annotations for genes and samples, respectively.

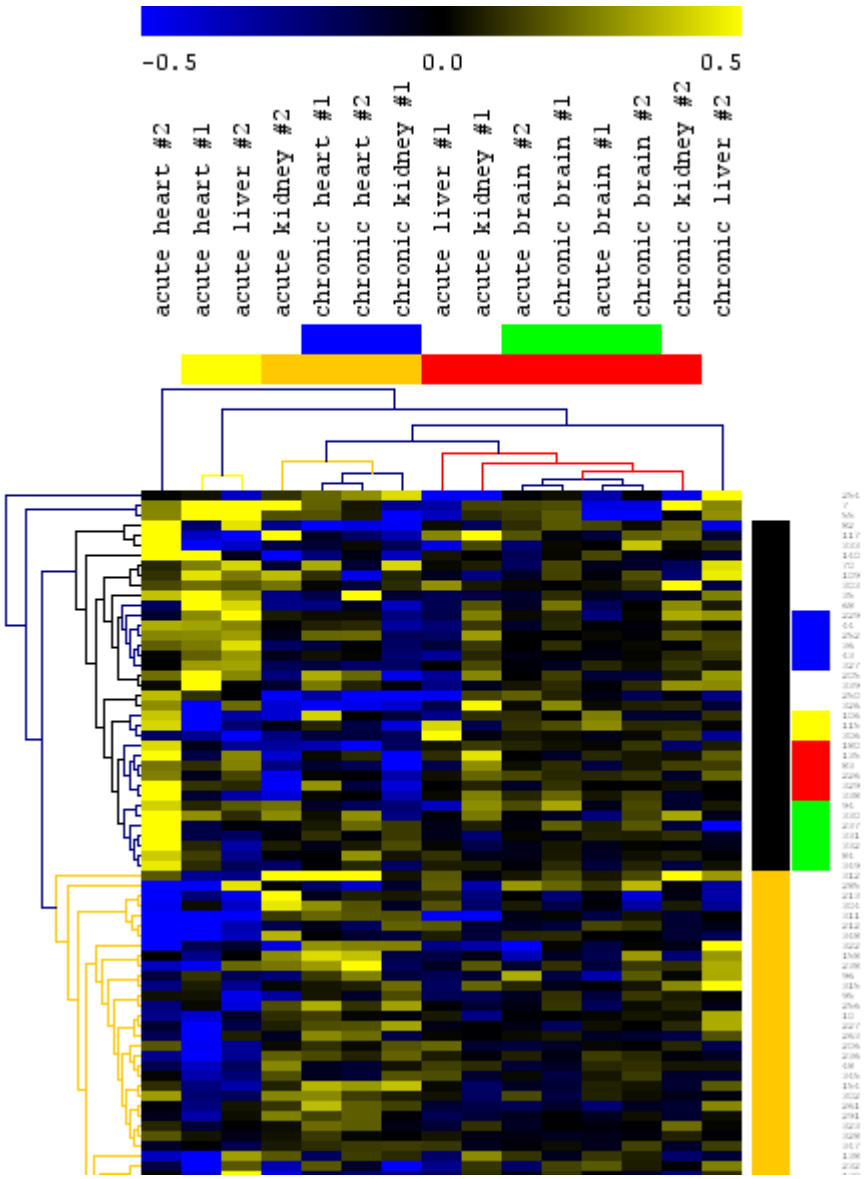
Cluster color-bars are adjustable. Users are able to drag and drop to adjust the order of the display. To maintain an adjusted colorbar layout, uncheck the checkbox in the Display Menu by clicking *Display* → *Cluster Viewing Options* → *Auto-Arrange Cluster Colors*. Auto-arrangement of cluster color-bars is the default setting.

To toggle the grey rectangles that appear as the mouse scrolls over portions of the viewable area, uncheck *Display* → *Cluster Viewing Options* → *Show Mouse Rectangles.* Additionally, users may click on the cluster color-bar area to display non-movable red rectangles as a point of reference. Re-click the intersection of the rectangles to remove them.

To compact the colorbar display into the smallest possible area without producing any overlapping clusters, click the checkbox at *Display → Cluster Viewing Options → Compact Cluster Color Groups*. This option will collapse the color-bar area into a tighter, neatly packed group while maintaining individual color-bar visibility. As a result of the compaction, cluster labels are no longer displayed in the viewable area.



8.1.10. HCL viewer with un-compacted cluster color-bars.



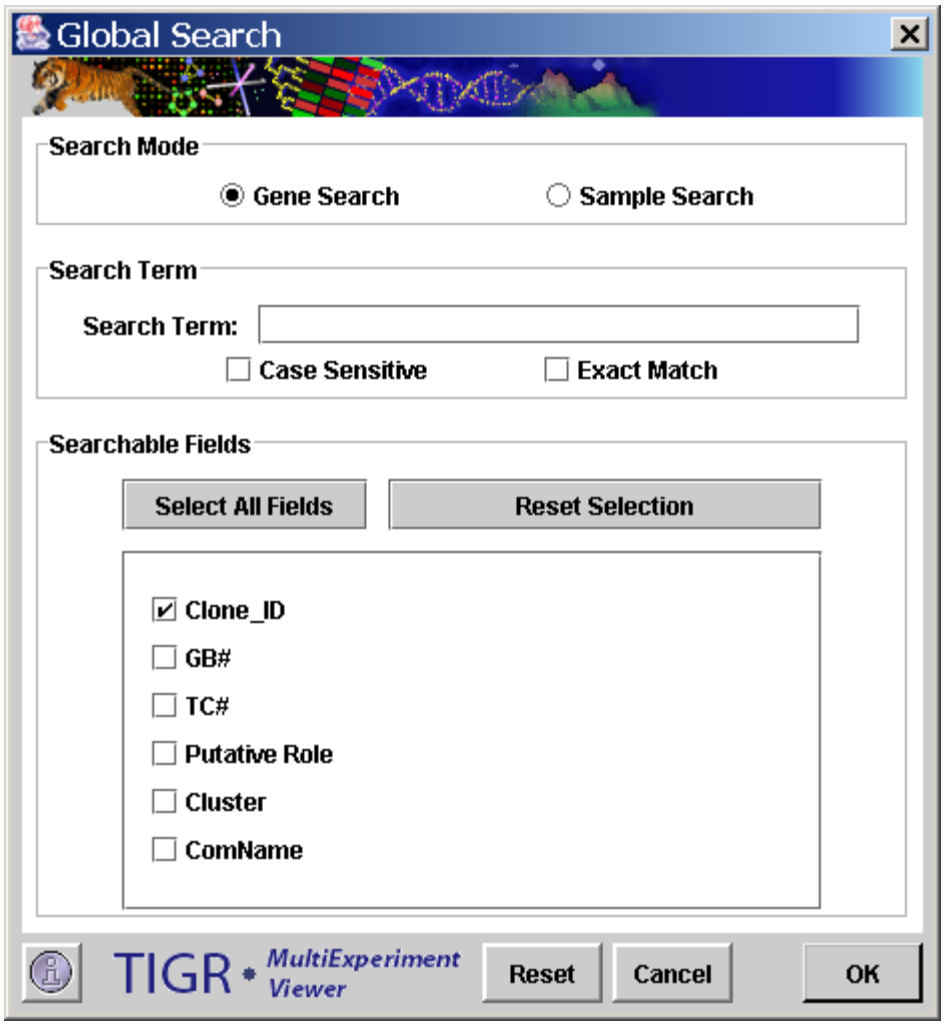
8.1.11. HCL viewer with compacted cluster color-bars.

9. Utilities Menu

9.1 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.

Select *Utility* → *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. This is especially good to use when there is annotation and you want to find specific types of genes.

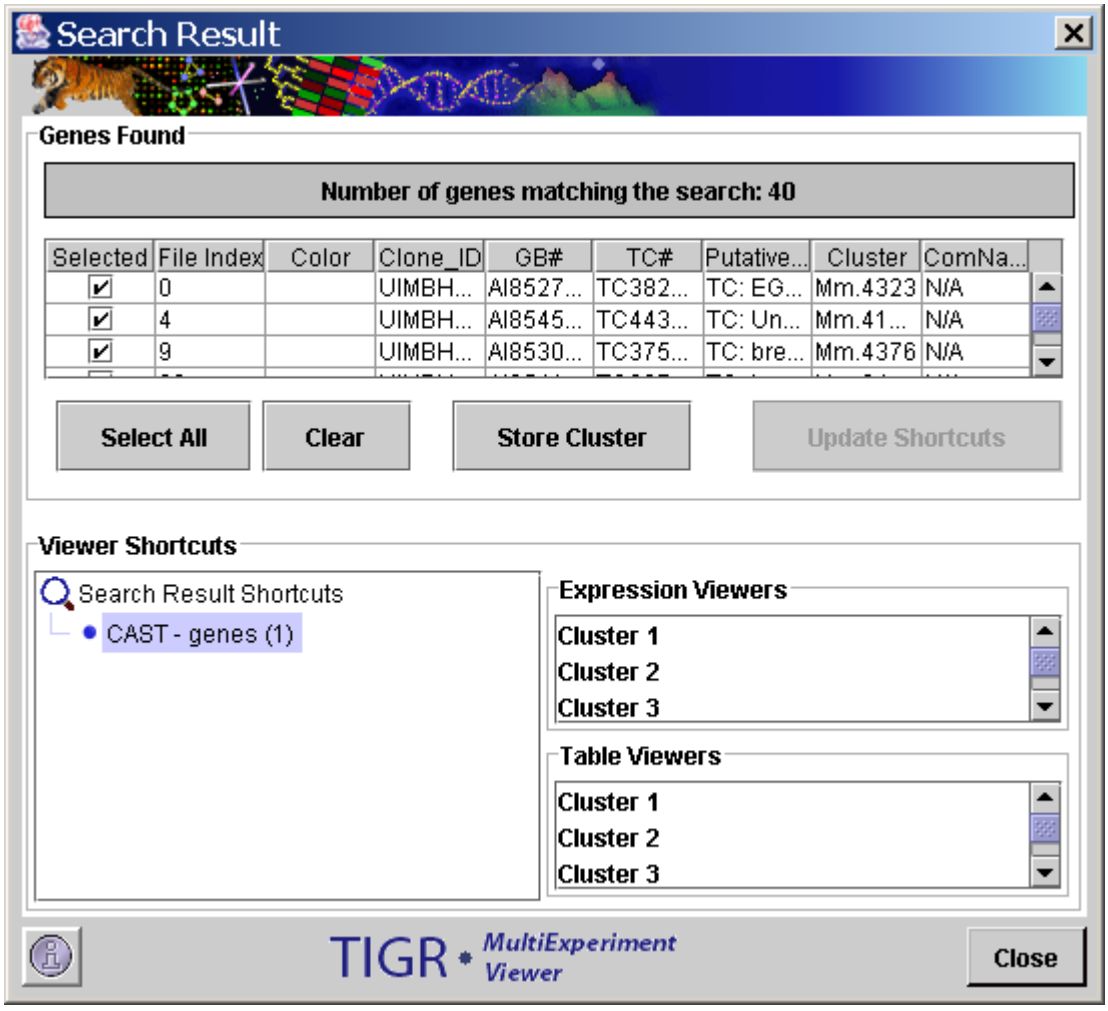


9.1.1. Search Initialization Dialog

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

9.2. Import Gene or Sample List

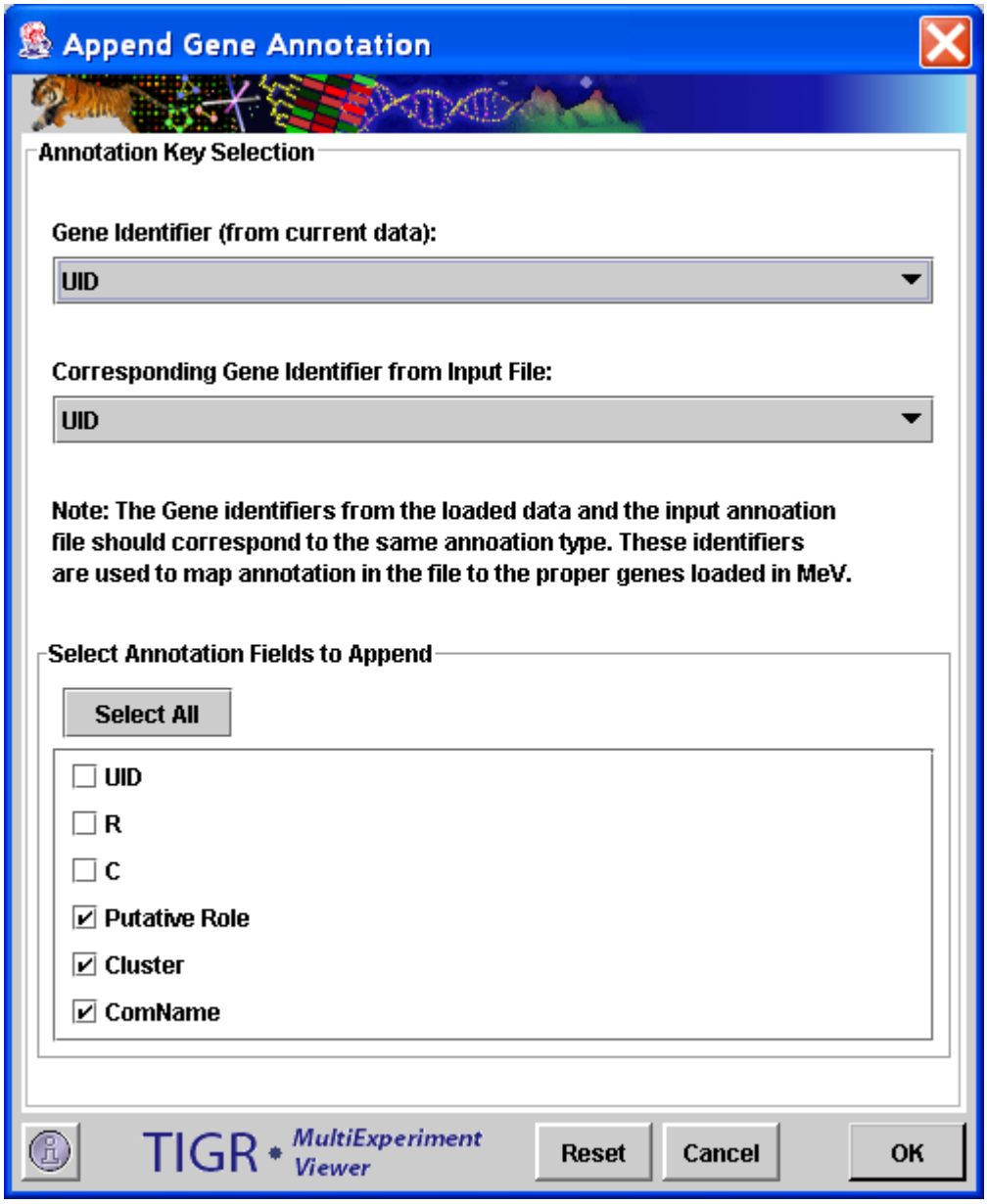
The *Import Gene/Sample List* allows one to create a cluster based on supplied identifiers. To import genes or sample lists, please refer to the explanation in section 8.1.

9.3. 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.

9.4. 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.



9.4.1. Gene Annotation Import Dialog

9.5. The Gaggle Submenu

For a description of the contents of the Gaggle submenu, please see section 11.2.

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 3.10.

10.1. Saving the Analysis

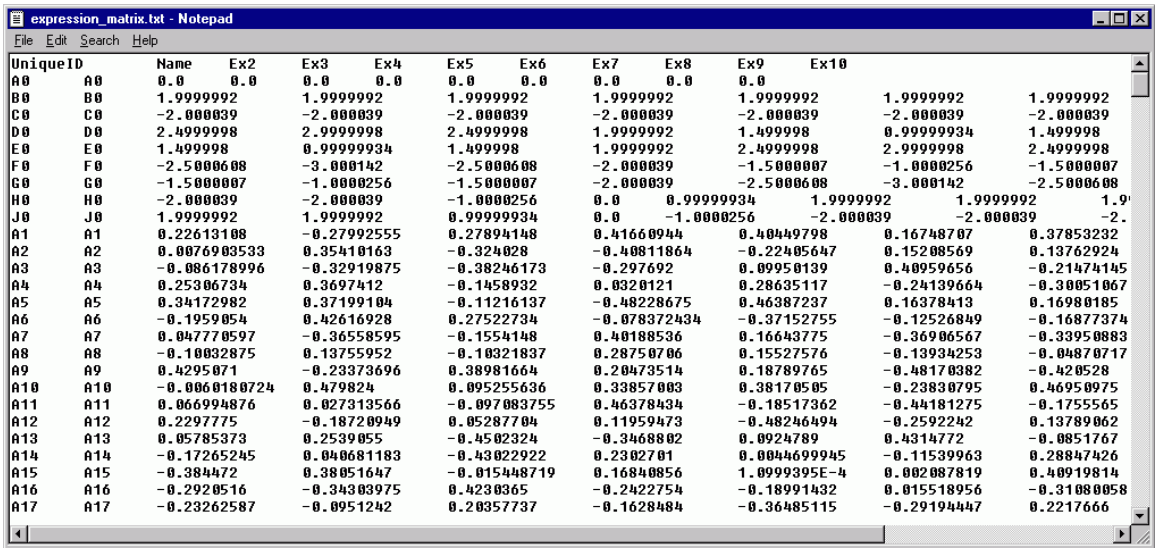
Saving the state of an ongoing analysis to file can be accomplished by selecting *File → Save Analysis As* in the Multiple Array Viewer. The analysis file will contain the loaded data, the current analysis results, and any clusters that are stored in the cluster manager. Once it has been saved, the *Save Analysis* option will be enabled. To reopen an analysis, *select File → Open Analysis*. This will restore the data and all algorithm results as they were last saved.

One word of caution on analysis-saving: this feature was dramatically re-written since MeV v4.0 to improve reliability and portability. Because of these changes, MeV 4.0 and higher cannot open saved analysis files created by MeV 3.1 or earlier, or MeV4.0b. In order to make saving analyses seamless and efficient, we have had to sacrifice backwards-compatibility with previous versions of the analysis saving features of MeV. Versions of MeV higher than v4.0 are able to open analysis (*.anl) files created by MeV v4.0 and later.

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 *File → Save Matrix*. 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			
B0	B0	1.9999992		1.9999992		1.9999992		1.9999992		1.9999992	1.9999992	1.9999992
C0	C0	-2.000039		-2.000039		-2.000039		-2.000039		-2.000039	-2.000039	-2.000039
D0	D0	2.4999998		2.9999998		2.4999998		1.9999992		1.499998	0.99999934	1.499998
E0	E0	1.499998		0.99999934		1.499998		1.9999992		2.4999998	2.9999998	2.4999998
F0	F0	-2.5000608		-3.000142		-2.5000608		-2.000039		-1.5000007	-1.0000256	-1.5000007
G0	G0	-1.5000007		-1.0000256		-1.5000007		-2.000039		-2.5000608	-3.000142	-2.5000608
H0	H0	-2.000039		-2.000039		-1.0000256		0.0	0.99999934	1.9999992	1.9999992	1.9
J0	J0	1.9999992		1.9999992		0.99999934		0.0	-1.0000256	-2.000039	-2.000039	-2.
A1	A1	0.22613108		-0.27992555		0.27894148		0.41660944		0.40449798	0.16748707	0.37853232
A2	A2	0.0076903533		0.35410163		-0.324028		-0.40811864		-0.22405647	0.15208569	0.13762924
A3	A3	-0.086178996		-0.32919875		-0.38246173		-0.297692		0.09950139	0.40959656	-0.21474145
A4	A4	0.25306734		0.3697412		-0.1458932		0.0320121		0.28635117	-0.24139664	-0.30051067
A5	A5	0.34172982		0.37199104		-0.11216137		-0.48228675		0.46387237	0.16378413	0.16980185
A6	A6	-0.1959054		0.42616928		0.27522734		-0.078372434		-0.37152755	-0.12526849	-0.16877374
A7	A7	0.047770597		-0.36558595		-0.1554148		0.40188536		0.16643775	-0.36906567	-0.33950883
A8	A8	-0.10032875		0.13755952		-0.10321837		0.28750706		0.15527576	-0.13934253	-0.04870717
A9	A9	0.4295071		-0.23373696		0.38981664		0.20473514		0.18789765	-0.48170382	-0.420528
A10	A10	-0.0060180724		0.479824		0.095255636		0.33857003		0.38170505	-0.23830795	0.46950975
A11	A11	0.066994876		0.027313566		-0.097083755		0.46378434		-0.18517362	-0.44181275	-0.1755565
A12	A12	0.2297775		-0.18720949		0.05287704		0.11959473		-0.48246494	-0.2592242	0.13789062
A13	A13	0.05785373		0.2539055		-0.4502324		-0.3468802		0.0924789	0.4314772	-0.0851767
A14	A14	-0.17265245		0.040681183		-0.43022922		0.2302701		0.0044699945	-0.11539963	0.28847426
A15	A15	-0.384472		0.38051647		-0.015448719		0.16840856		1.0999395E-4	0.002087819	0.40919814
A16	A16	-0.2920516		-0.34303975		0.4230365		-0.2422754		-0.18991432	0.015518956	-0.31080058
A17	A17	-0.23262587		-0.0951242		0.20357737		-0.1628484		-0.36485115	-0.29194447	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 *File* → *Save Image*. 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. The Gaggle Implementation

11.1. Introduction to the Gaggle

The Gaggle is a framework, developed at the Institute for Systems Biology in Seattle, WA, for exchanging data between software tools. It is specifically designed to handle systems biology data, and allow for the transfer of biological data between independently-developed software tools on a users' desktop computer. The Gaggle defines a set of data types that software tools can "broadcast" to one another, thereby allowing easy communication between applications without the use of intermediate data files.

Applications that implement the Gaggle framework are referred to as *geese*. Each MultipleArrayViewer window can act as a goose. They can broadcast data to each other and to other geese connected to the Gaggle network. They can also receive data from other geese, if they have no data already loaded.

11.2. The Gaggle Menu

The Gaggle submenu is located under the *Utilities* menu in the MultipleArrayViewer toolbar. It contains the functions used to connect, disconnect, and direct traffic to the Gaggle Network.

Connect to Gaggle
Connect the current MultipleArrayViewer to the Gaggle network. This option will only function correctly if the computer is connected to the internet. After the first MultipleArrayViewer has been connected, new MAV windows will automatically connect themselves when they are opened.

Disconnect from Gaggle
Disconnect from the Gaggle network. Choosing this option will only disconnect the currently open MultipleArrayViewer from the Gaggle network. If other MultipleArrayViewer windows are open and connected to Gaggle, they will need to be disconnected separately. MeV will automatically disconnect all MultipleArrayViewers from the Gaggle network when the application is closed.

Broadcast Target
Select the target of Gaggle broadcasts. This submenu will contain the names of all currently-connected geese. If "Boss" is selected, all broadcasts will be sent to the boss and to all other connected members of the Gaggle network. If another goose is selected, the data will be broadcast only to that goose.

Show Goose
Select a currently-connected goose to bring to focus in the screen.

11.3. Using the Gaggle with MeV

MeV is capable of sending and receiving some of these types of data, including expression (matrix) data, lists of genes, and gene networks. Before MeV can send or receive, however, it must be connected to the internet and the Gaggle network. In a Multiple Array Viewer window, select *Utilities->Gaggle->Connect to Gaggle*. The Gaggle Boss program will launch and minimize itself. This program must remain running while you use the Gaggle network.

If MeV is connected to the Gaggle network and has an empty MultipleArrayViewer window open, it will listen for the broadcast of a matrix of data. This data will then be loaded into the MultipleArrayViewer as if it had been loaded via the standard file loaders. If the MultipleArrayViewer window is not empty, incoming matrix broadcasts will be ignored.

To broadcast data from MeV to other members of the Gaggle network, use the right-click menu in the standard module result viewers. From these viewers, expression matrix data and/or gene lists can be broadcast. Please see section 7.6, Common Viewer Activities, for further information.

More details on the workings of MeV with the Gaggle and a tutorial on its use are available on the TM4 website and the Gaggle website.

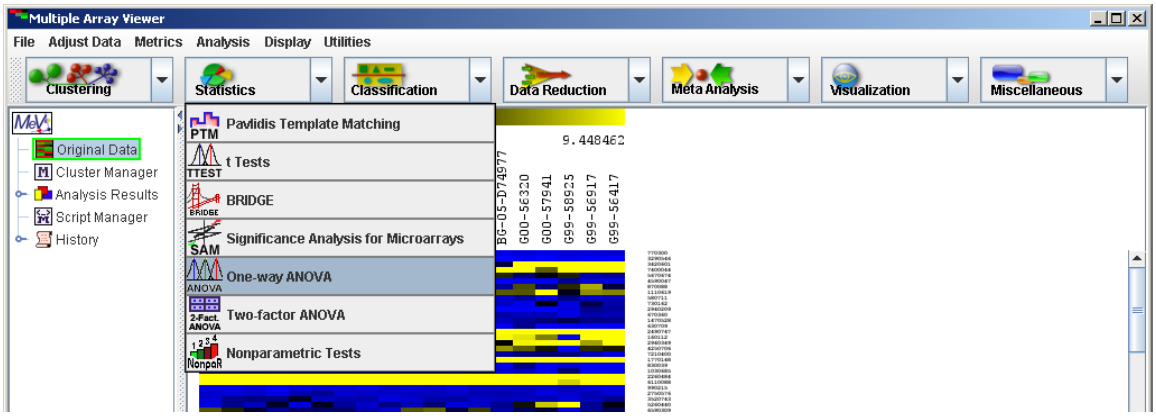
12. Modules

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. New to MeV 4.2, the toolbar groups modules into seven categories: Clustering, Statistics, Classification, Data Reduction, Meta Analysis, Visualization, and Miscellaneous. Clicking on the category icon brings a drop-down list of analyses within that category. 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 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.

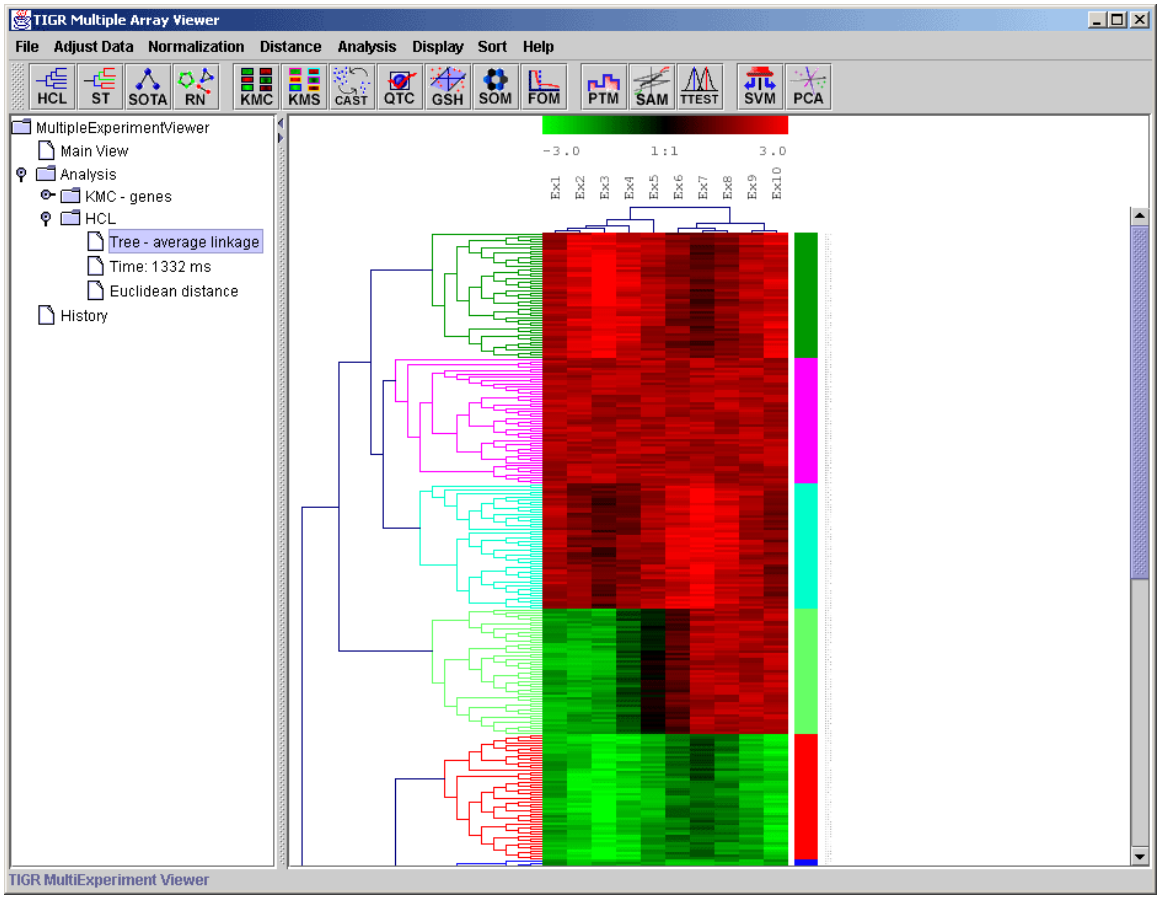


11.1 HCL: Hierarchical clustering

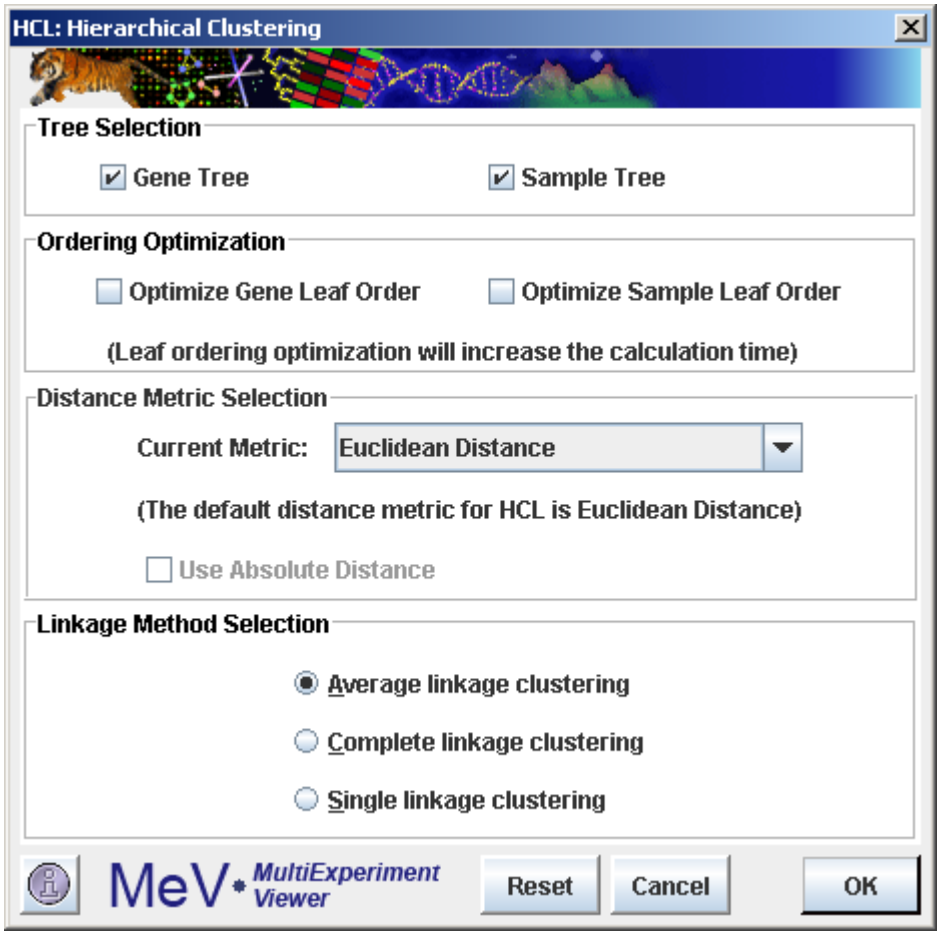
(Eisen *et al*. 1998)

Selecting this analysis will display a dialog that allows different linkages and options to cluster genes, samples or both. Once the computations are complete, select the nodes *Analysis* → *HCL* → *HCL Tree* 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.



12.1.1. Hierarchical tree with clusters selected.



12.1.2. HCL Initialization Dialog

Parameters:

Tree Selection

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

Order Optimization

These checkboxes are used to indicate whether the ordering of the leaves will be optimized for genes, samples or both.

Distance Metric Selection

This menu is used to indicate the distance metric that will be used in calculating the tree. The default distance metric is Euclidean.

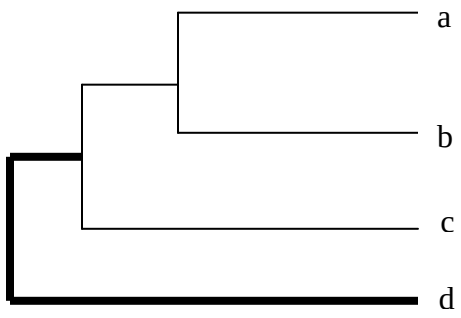
Linkage Method

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

Single Linkage: The distances are measured between each member of one cluster and 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 to measure the 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:

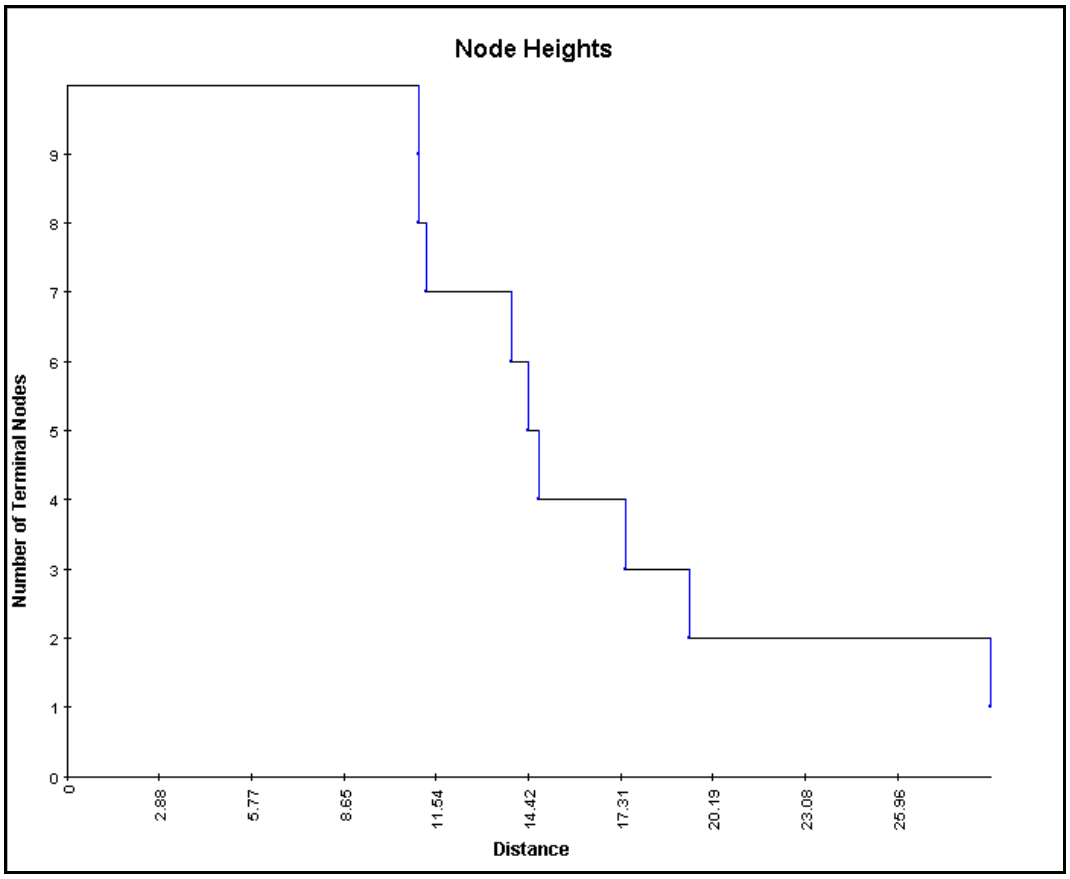
$$d_{d,(a,b,c)} = \frac{\left(d_{d,a} + d_{d,b} + d_{d,c}\right)}{3} = \frac{d_{d,a}}{3} + \frac{d_{d,b}}{3} + \frac{d_{d,c}}{3}$$

Weighted Average Linkage:

$$d_{d,(a,b,c)} = \frac{\left\lfloor (d_{d,a} + d_{d,b})/2 + d_{d,c} \right\rfloor}{2} = \frac{d_{d,a}}{4} + \frac{d_{d,b}}{4} + \frac{d_{d,c}}{2}$$

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

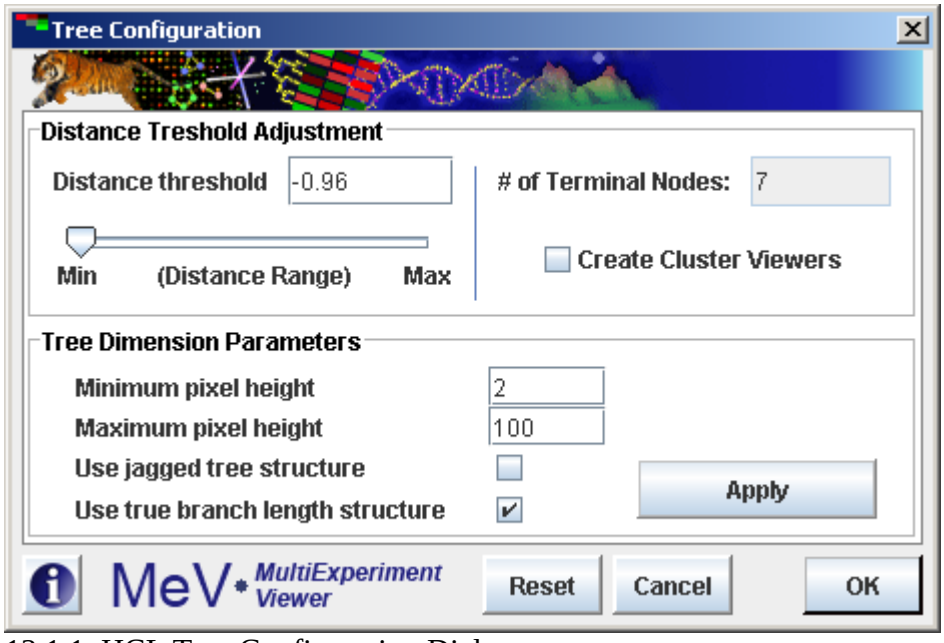
Complete Linkage: The distances are measured between each member of one cluster and each member of the other cluster. The maximum of these distances is considered the cluster-to-cluster distance.



12.1.3. 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, *Gene Tree Properties*. This option allows 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 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.



12.1.1. HCL Tree Configuration Dialog

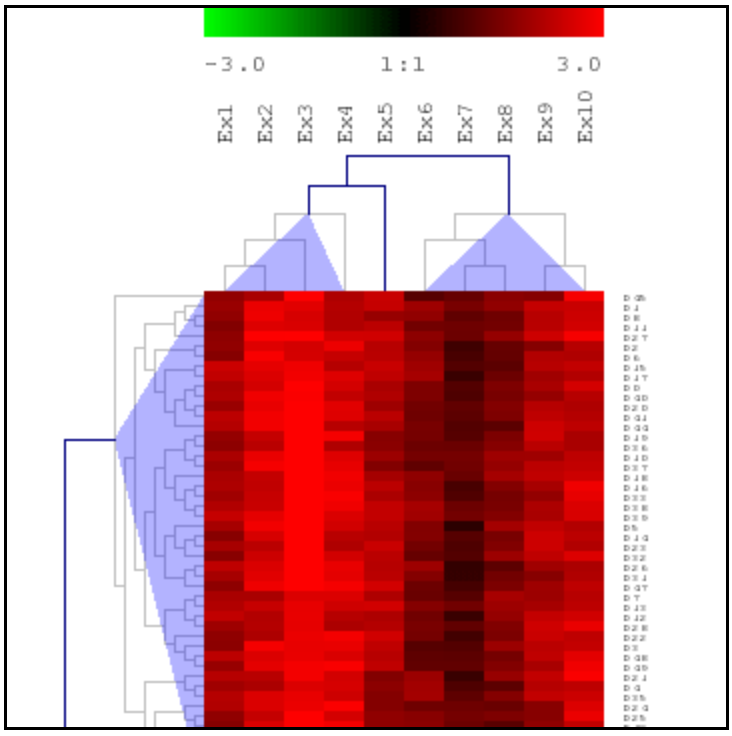
The *Create Cluster Viewers* option allows you to create viewers based on the distance threshold. This option collects groups of elements falling below terminal nodes in the tree using the current distance threshold. The clusters of elements are represented as nodes in the result navigation tree under 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.

By default, tree branches are built from the heatmap up towards the root node. Consequently, a node will have branches of differing heights to reach its two children. To change the tree so that both branches of a node represent the “node height”, check the *Use jagged tree structure* box. Most branches will no longer reach the heatmap from the terminal nodes.

To draw the tree such that the position of every node is exactly the node height, check the Use true branch length structure. Note: For some distance metrics this feature does not display a tree.

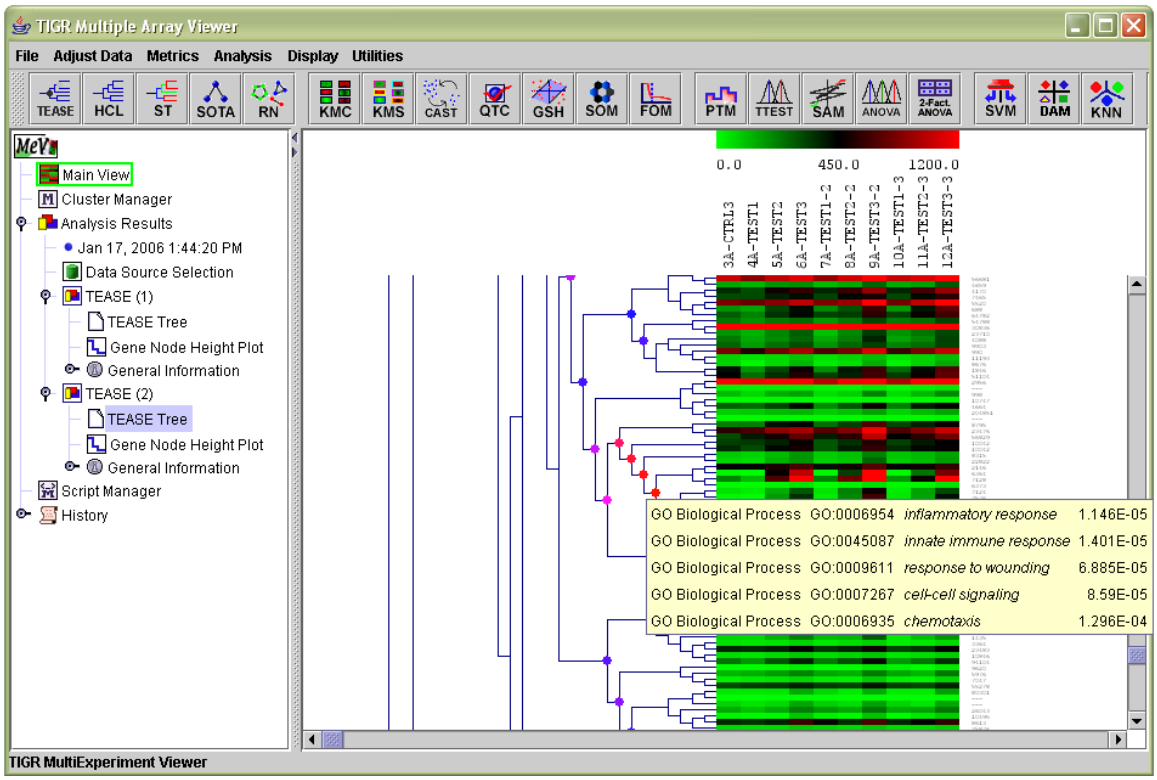
MeV 4.4 features a new option to allow users to rotate nodes. Rotating nodes does not affect the tree structure. Nodes will continue to have the same parents and children, but the two subtrees of the selected node will be displayed in reverse order. To rotate a node, left- click to select the node and right-click to select *Rotate Selected Node*.



12.1.2. HCL Tree with distance thresholds applied.

11.2 TEASE: Tree-EASE

Selecting this analysis will display a configuration window with three main modes and four tabbed panels; each contains essential configurations that are specific for Hierarchical Clustering (HCL) or EASE analysis. Once the analysis is completed, select the tree node under “Analysis Results” to view the hierarchical tree with color dots signified whether the cluster contains over-expressed biological categories (Red: most significant, Blue: least significant). To view EASE analysis information of each cluster, position the mouse over the color dot at the root of the cluster to display a popup window. The first column in the window is the name of annotation file which contains the category. The second to the third to the last column are information about the category that is included in the annotation file. If the default mode “Cluster Analysis” is selected, the last column is the “score” the category receives, that is the probability the certain category is decided over-expressed by chance. If “Annotation Survey” mode is selected, the last two columns would be the hit count and cluster size, respectively.



11.2.1 TEASE hierarchical tree with color coded visualization dots.

Mode selection:

HCL Only

Perform Hierarchical Clustering (HCL). No biological theme exploration.

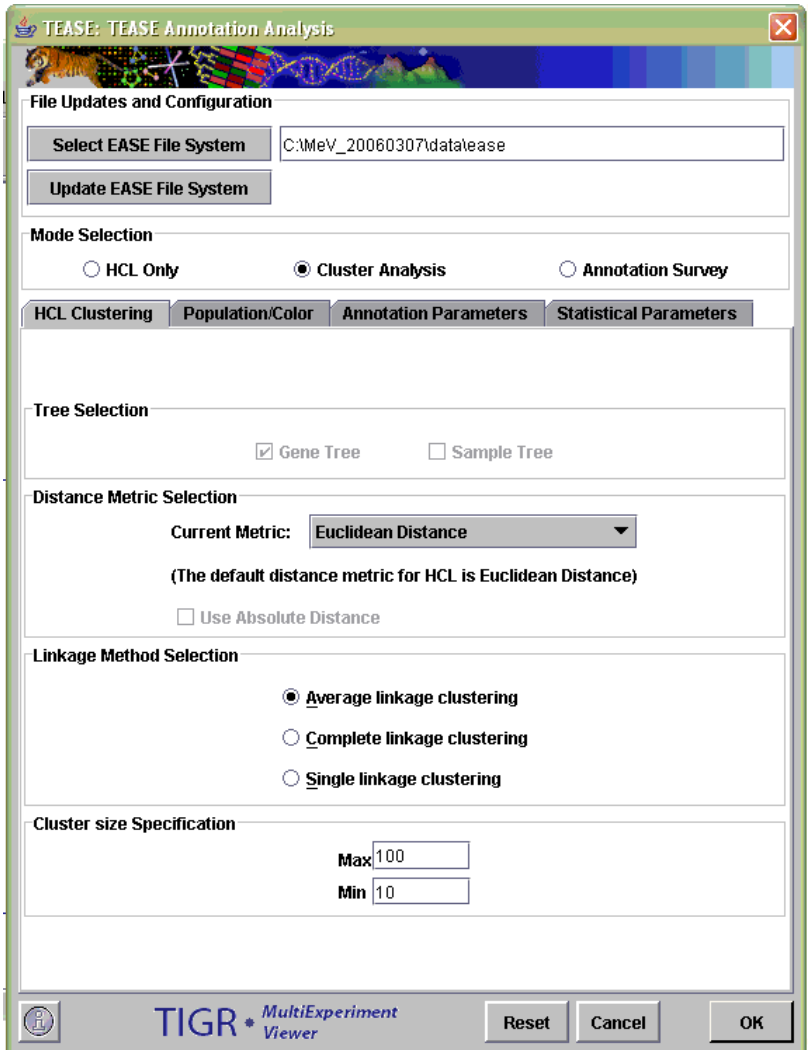
Cluster Analysis

EASE analysis on clusters that fall within the minimum and maximum size of population as specified in the panel “HCL Clustering” – “Cluster size specification”. Calculate the score of each category and rank the categories by score.

Annotation Survey

EASE analysis on clusters that fall within the minimum and maximum size parameters specified by the user. Calculate and rank present categories in each cluster by hit counts. No score is calculated in this mode.

Parameters:



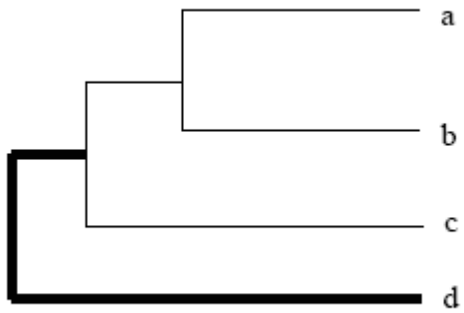
11.2.2 TEASE parameter setting window

<u>HCL Clustering</u>

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:

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

Weighted Average Linkage:

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

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

Cluster Size Specification
State of the size of clusters you want to analyze. You won't get too much information for clusters that are too big or too small. The default size is from 10 to 100.

Population/Color

Population selection
Available only when in "Cluster Analysis" mode. Please refer to EASE documentation for more information.

Assign Color Gradient
Specify the upper and lower score for assigning color gradient. Increasing upper and lower bounds will cause the gradient shifts to red whereas decreasing will

shift it to blue. Gradient can also be adjusted in the graphic view window after analysis is completed and the correct tree code is selected. The default setting for upper bound is 0.1, and 0.00001 for lower bound.

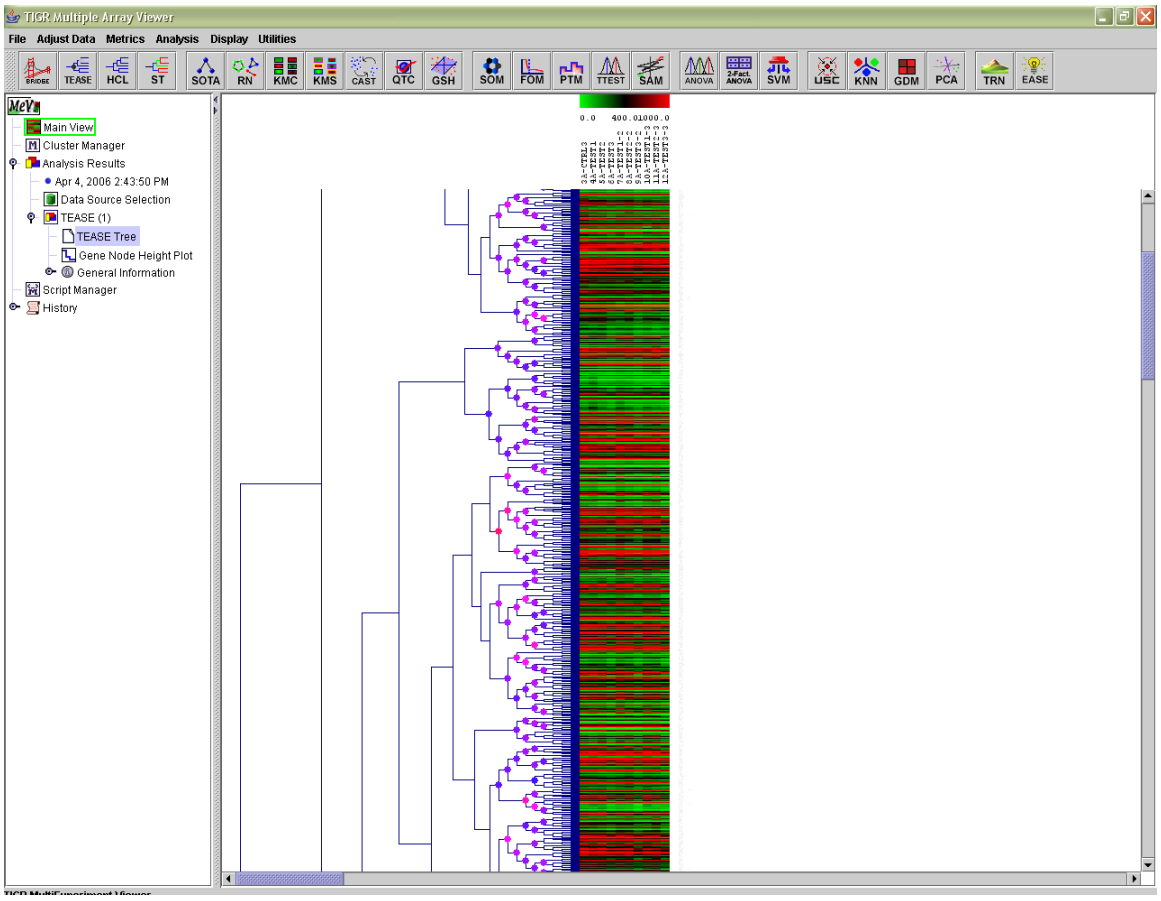
Annotation Parameter

Selections of annotation key, annotation conversion file, and gene ontology/gene annotation/gene linking files. Refer to EASE documentation for more information.

Statistical Parameter

Selections of reported statistics, multiplicity correction and time parameters. Refer to EASE documentation for more information.

Navigating the hierarchical tree



11.2.3 TEASE Tree View

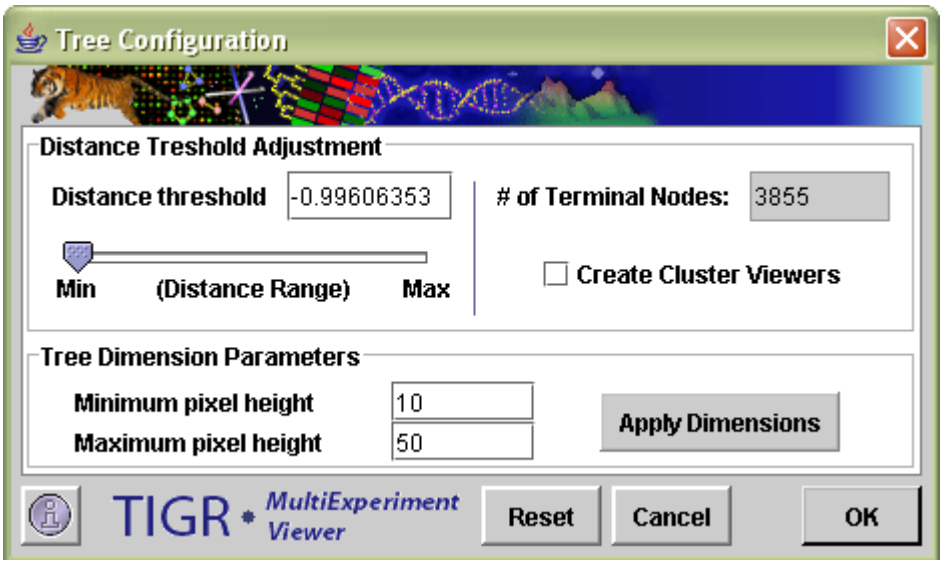
Basic Navigation

A large dataset is likely to have more than a handful clusters that fall within the size range, but only clusters that are “more red” worth attention. It is thus important to assign appropriate color gradient boundaries to save time. Adjusting

color gradient will be in a later section. To view information about each cluster, simple position the cursor over the root of the cluster to reveal a pop-up window. When done, move the cursor away and the window will disappear.

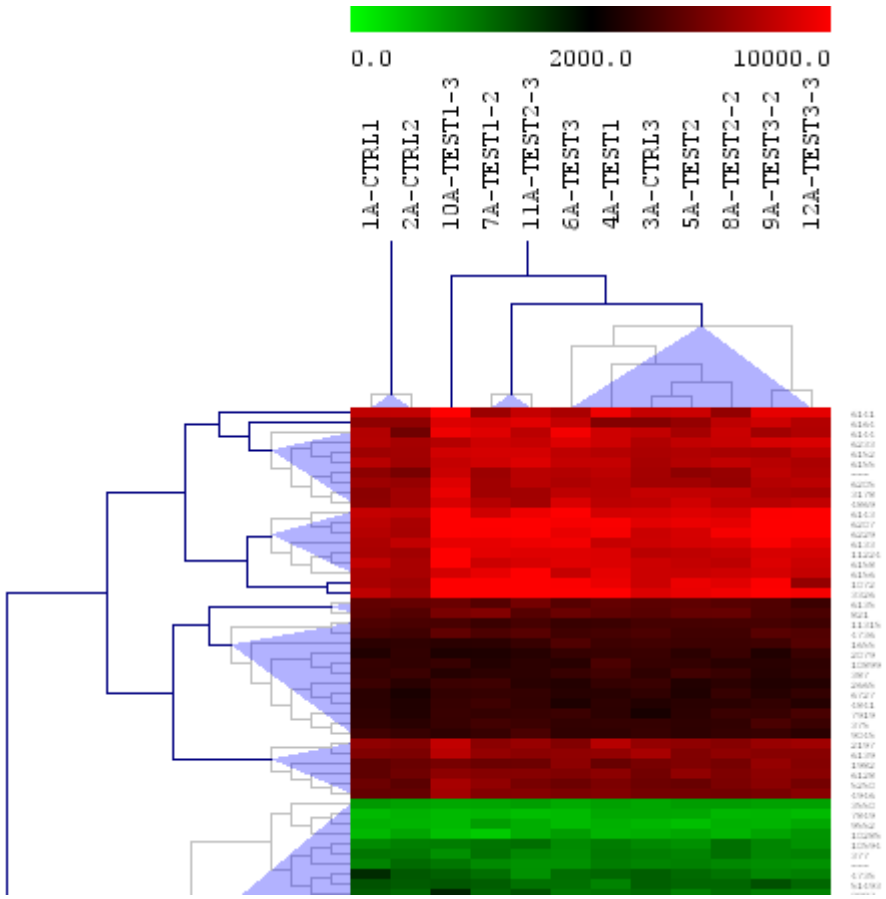
Adjusting Tree Configuration and Viewing Clusters

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 TEASE 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 draw 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.



11.2.4 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 TEASE result node. The results are added once the TEASE 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 TEASE tree. This allows one to fine tune the tree's appearance without dismissing the dialog.



11.2.5 TEASE tree (HCL mode) with distance thresholds applied

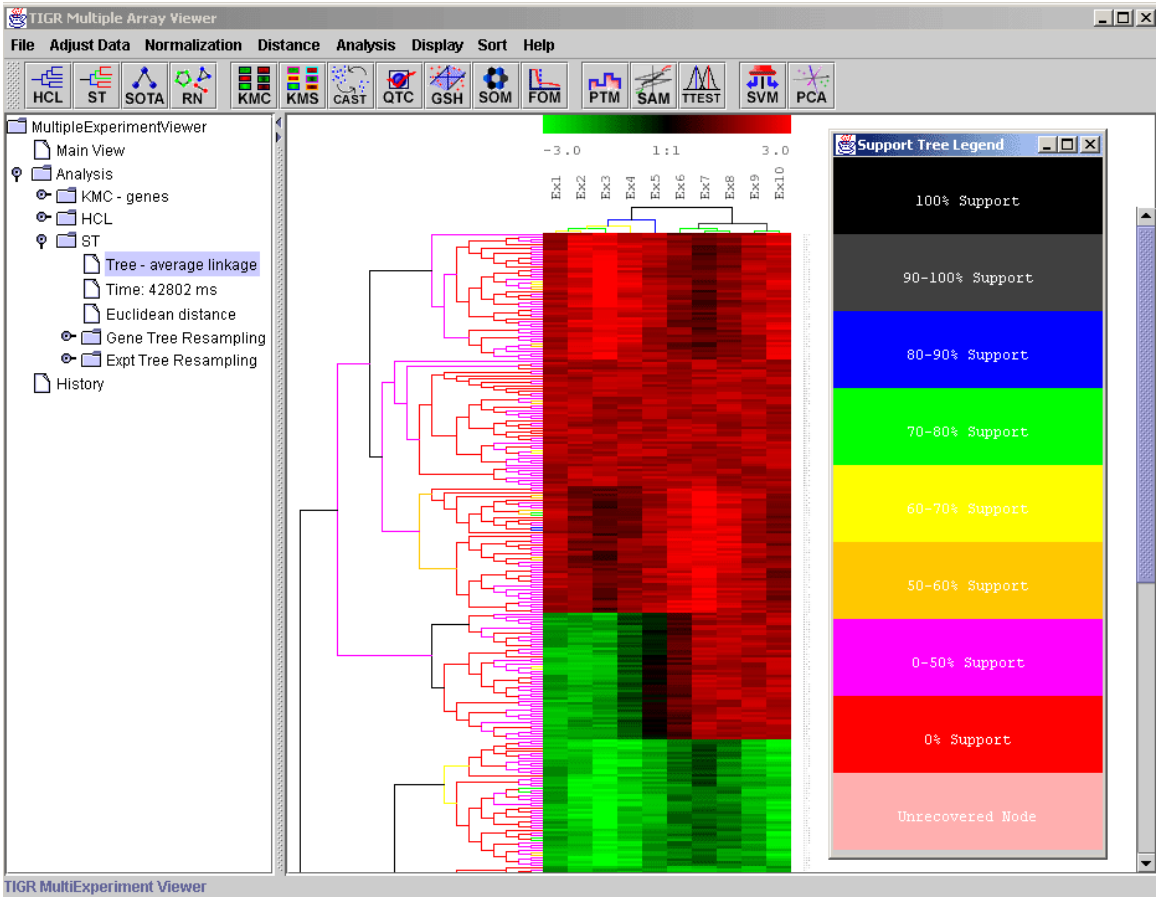
Adjusting Color Gradient

A right click on the TEASE Tree View will reveal a pop-up menu. Near the bottom of the menu is an option that says “Change Score Boundary”. Left click on the option will produce a configuration window. In the window are two editable text fields similar to the “Assign Color Gradient” panel in the initialize window. Enter appropriate number and click on “OK” to apply the changes or “Cancel” to exit the window without any changes. You can view the changes you make in Tree View once you exit the window.

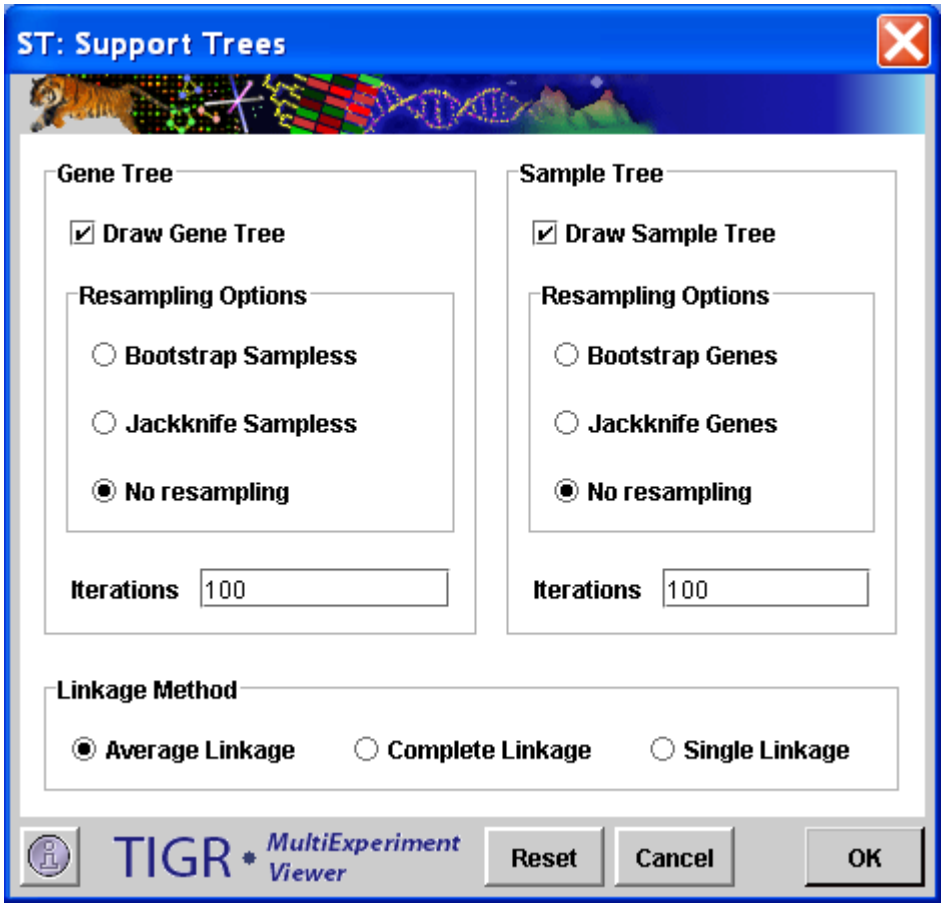
11.3 ST: Support Trees

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.



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



11.3.2 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

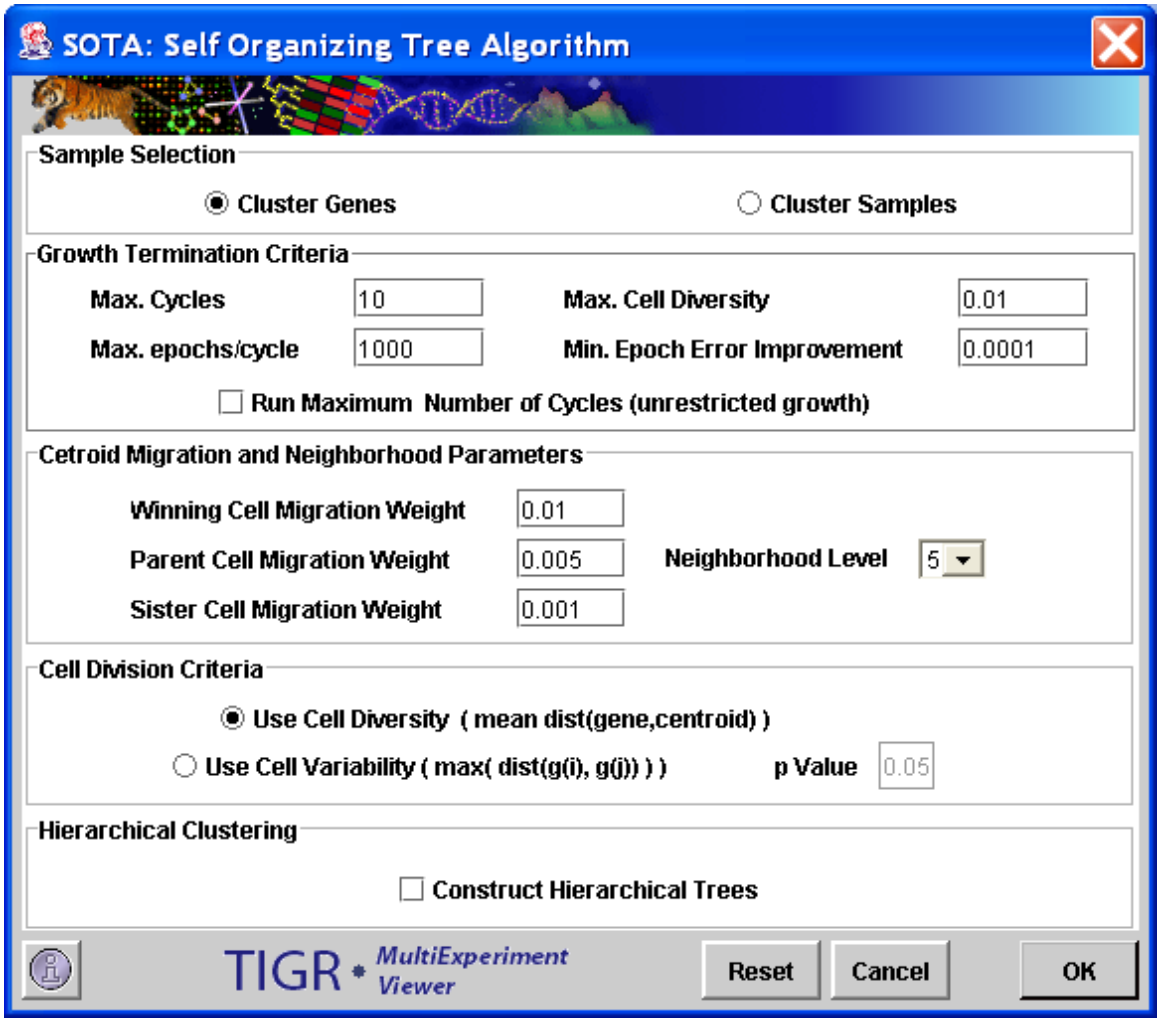
A legend to relate support tree output colors to % support is displayed by selecting the support tree legend menu item in the main help menu.

Default Distance Metric: Euclidean

11.4 SOTA: Self Organizing Tree Algorithm

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

The initialization form shown below (0) is divided into four main areas. The SOTA algorithm constructs a binary tree (dendrogram) in which the terminal nodes are the resulting clusters.



11.4.1 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.

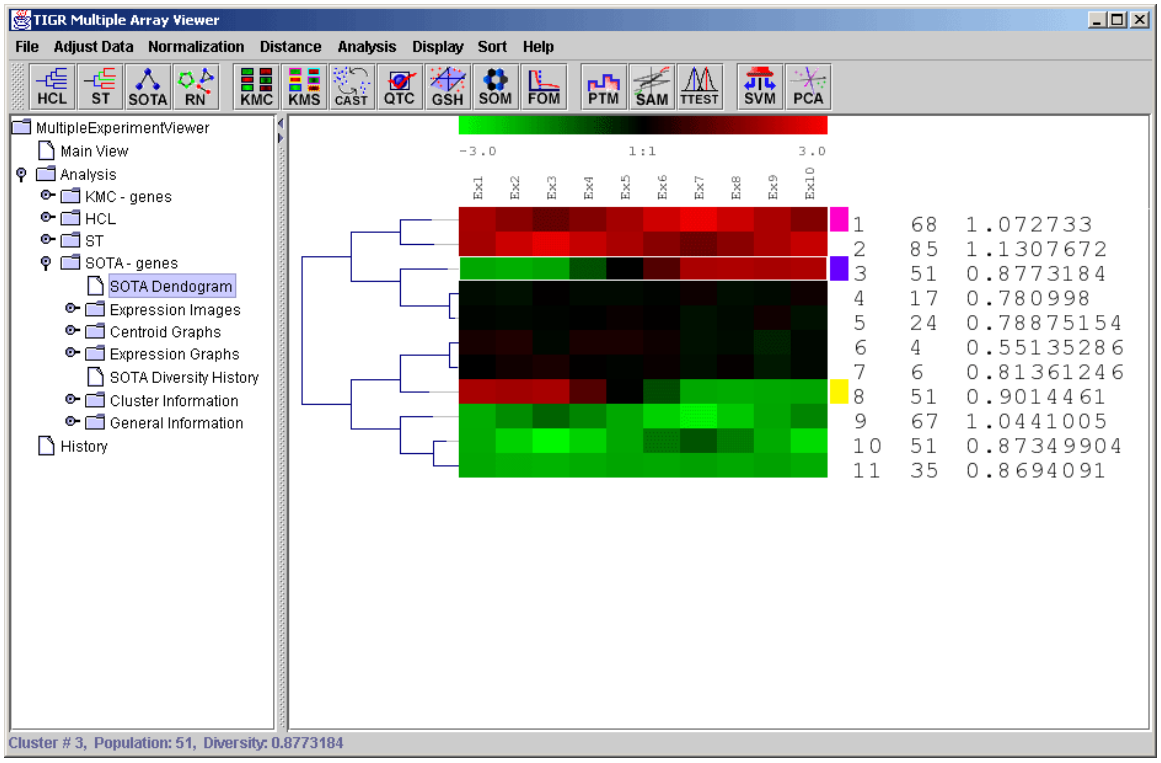
Hierarchical Clustering

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

Default Distance Metric: Euclidean

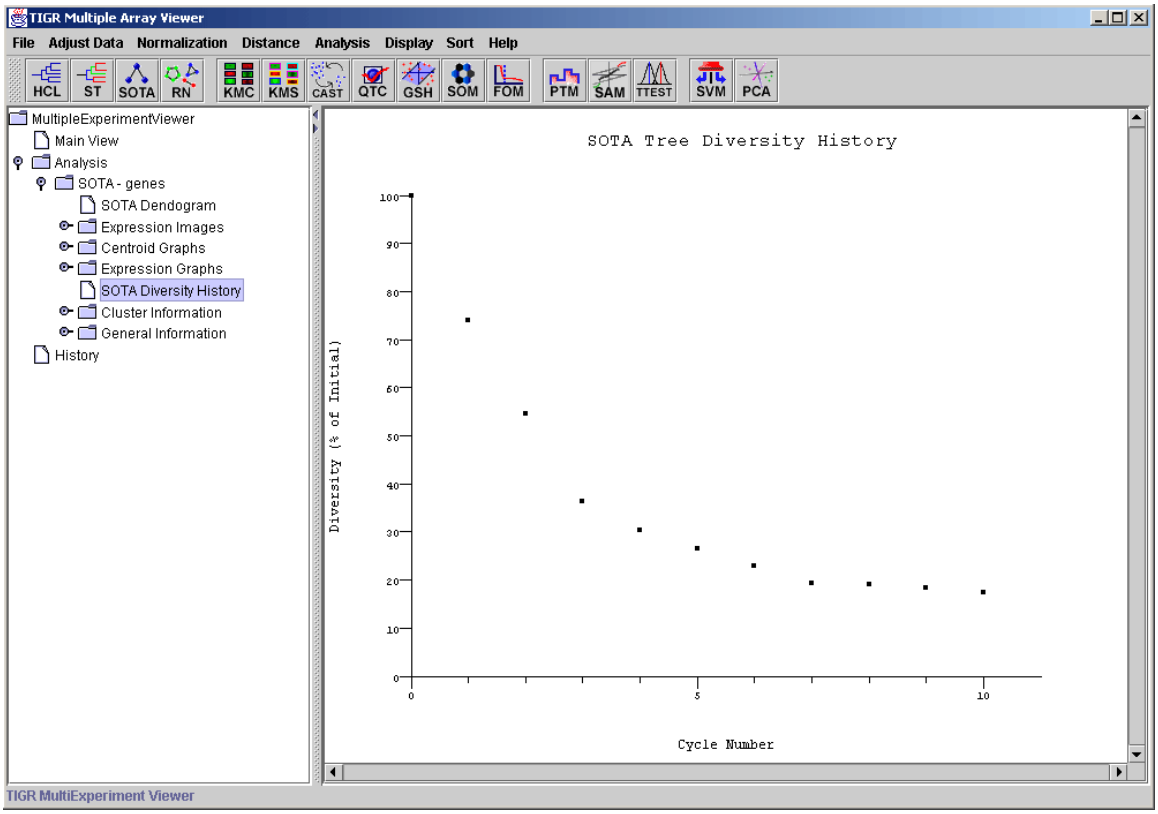
The result views created by the SOTA algorithm include the basic viewers with the addition of two SOTA specific viewers and enhancements to expression image viewers to include more cluster information.

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.



11.4.2 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).

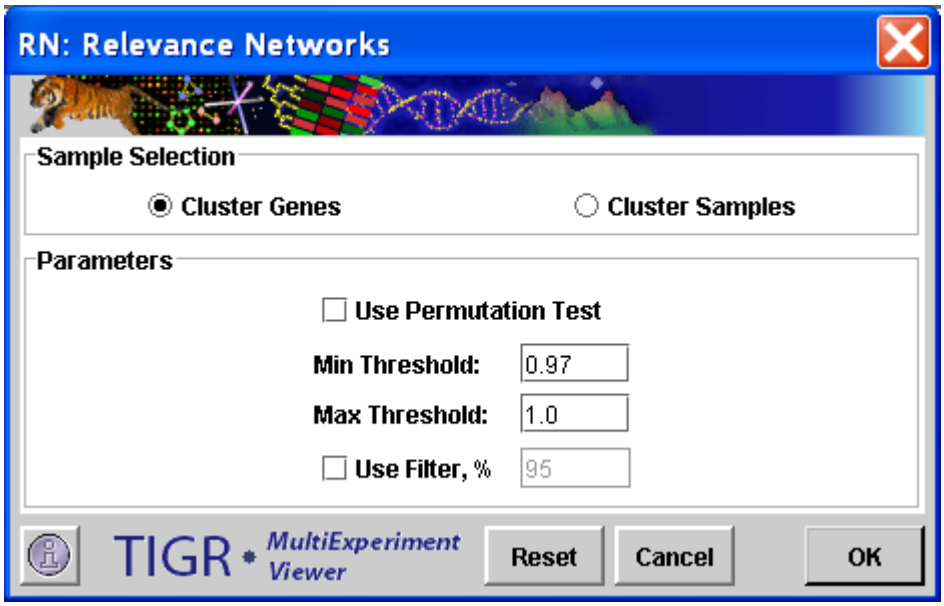


11.4.3 SOTA diversity viewer

11.5 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.



11.5.1 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^2values derived following permutation of the expression vectors.

Min Threshold

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

Use Filter

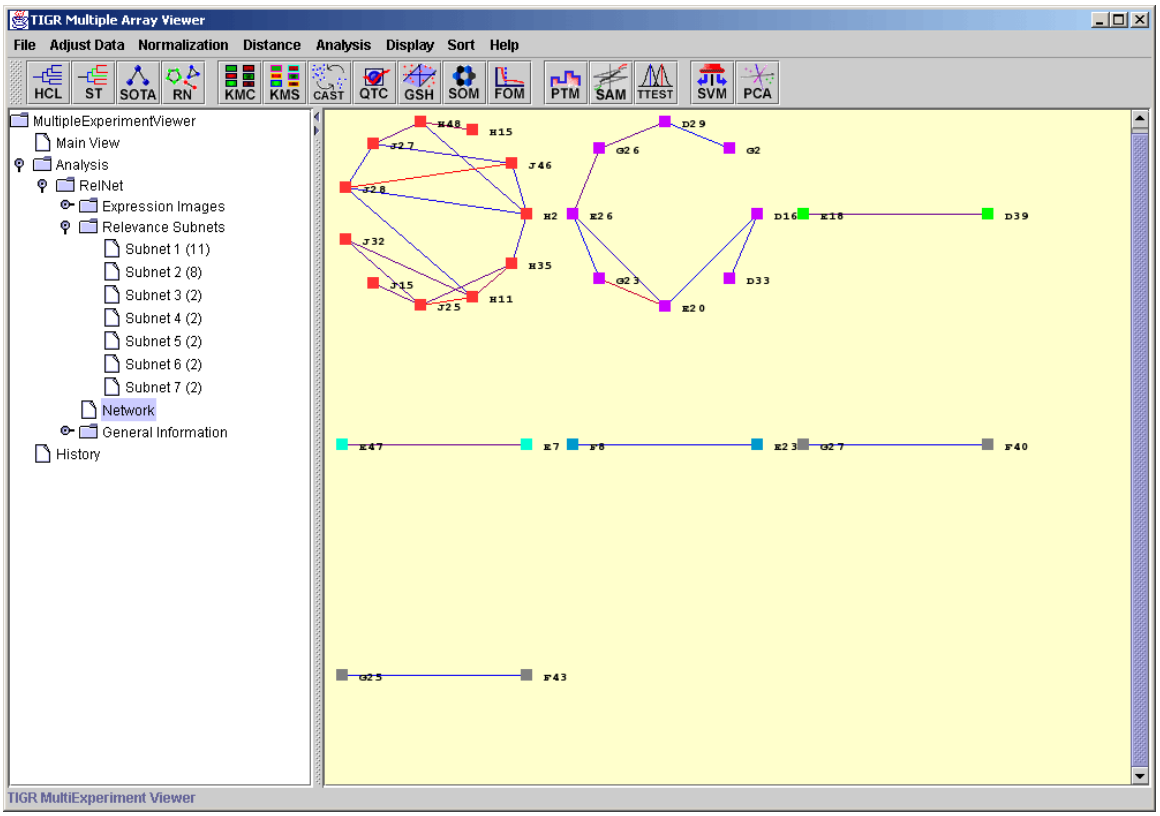
This option allows the user to filter out elements with little dynamic change thus removing flat or uninteresting elements. A measure of entropy is used to rank the elements. The percentage value entered (1 to 100) indicates what percentage of the elements to retain for the construction of the network.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



11.5.2 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.

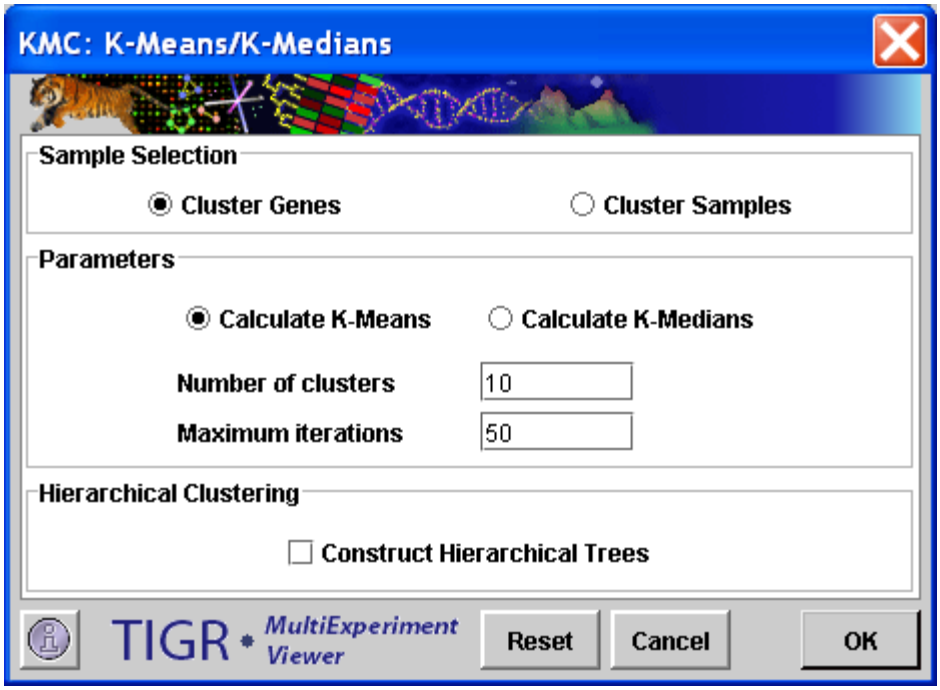
Other options reveal the nature of the subnets. You can select to alter the threshold to make it more stringent. Using this option under the *links-threshold* the viewer is instructed to only show links for elements correlated with R^2 greater that the new threshold. Selection of elements in the network 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).

11.6 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.



11.6.1 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.

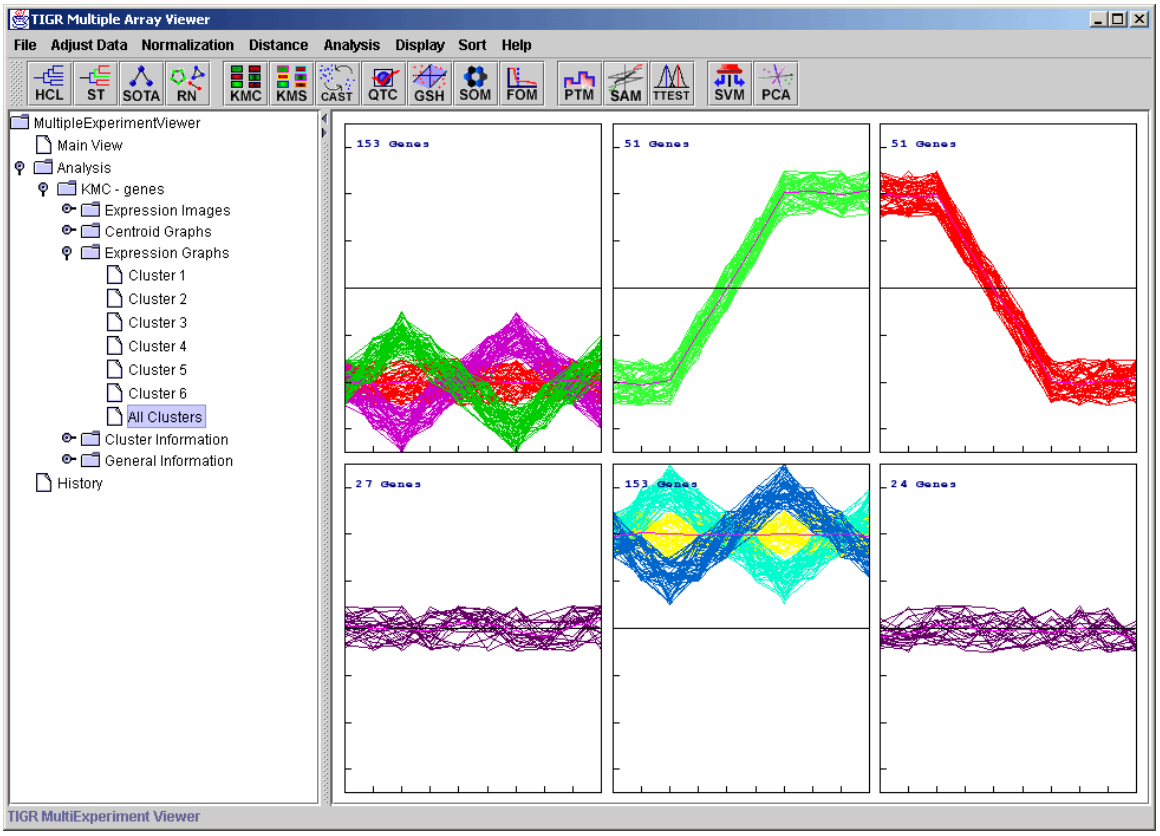
Number of Iterations

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



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

11.7 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.

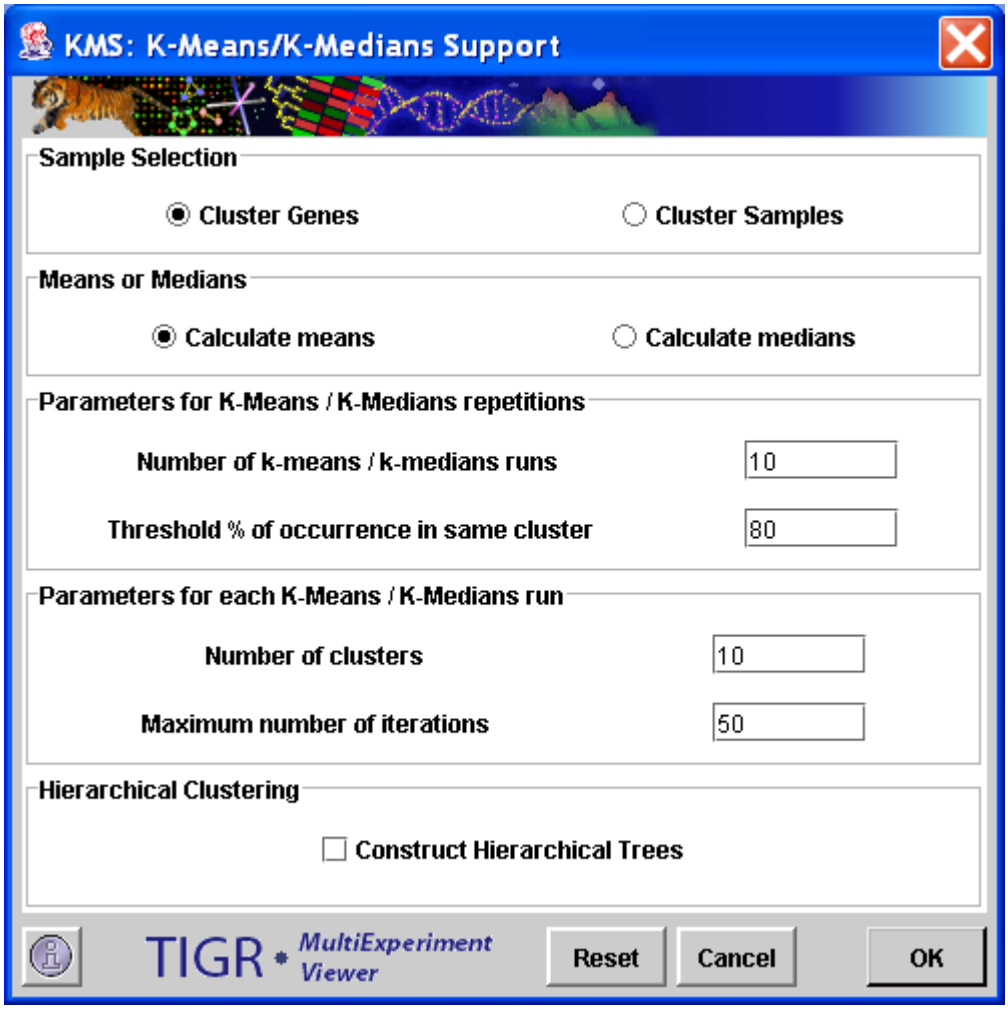
Number of Iterations

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



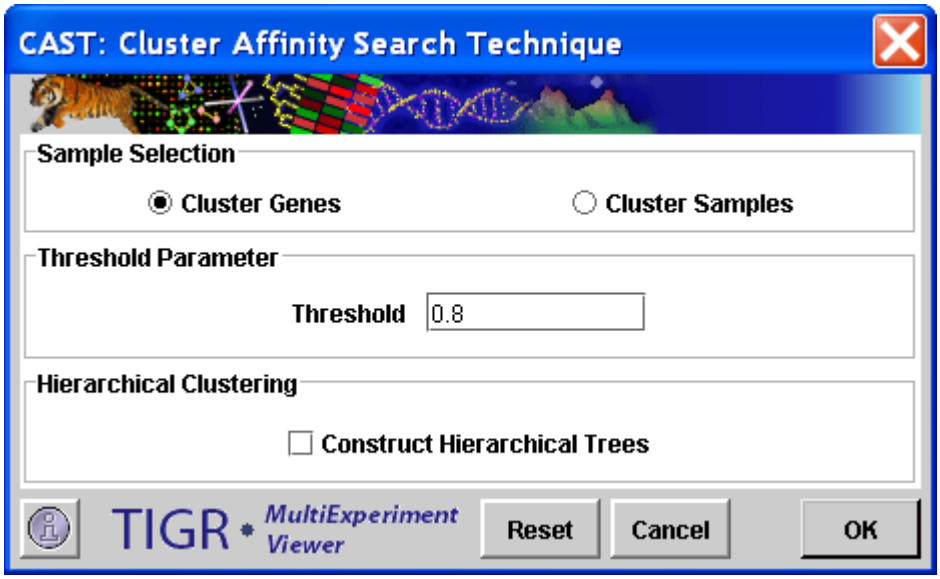
11.7.1 K-Means / K-Medians Support: Initialization Dialog Box.

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.

11.8 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.



11.8.1 CAST Initialization Dialog

Parameters:

Sample Selection

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

Threshold

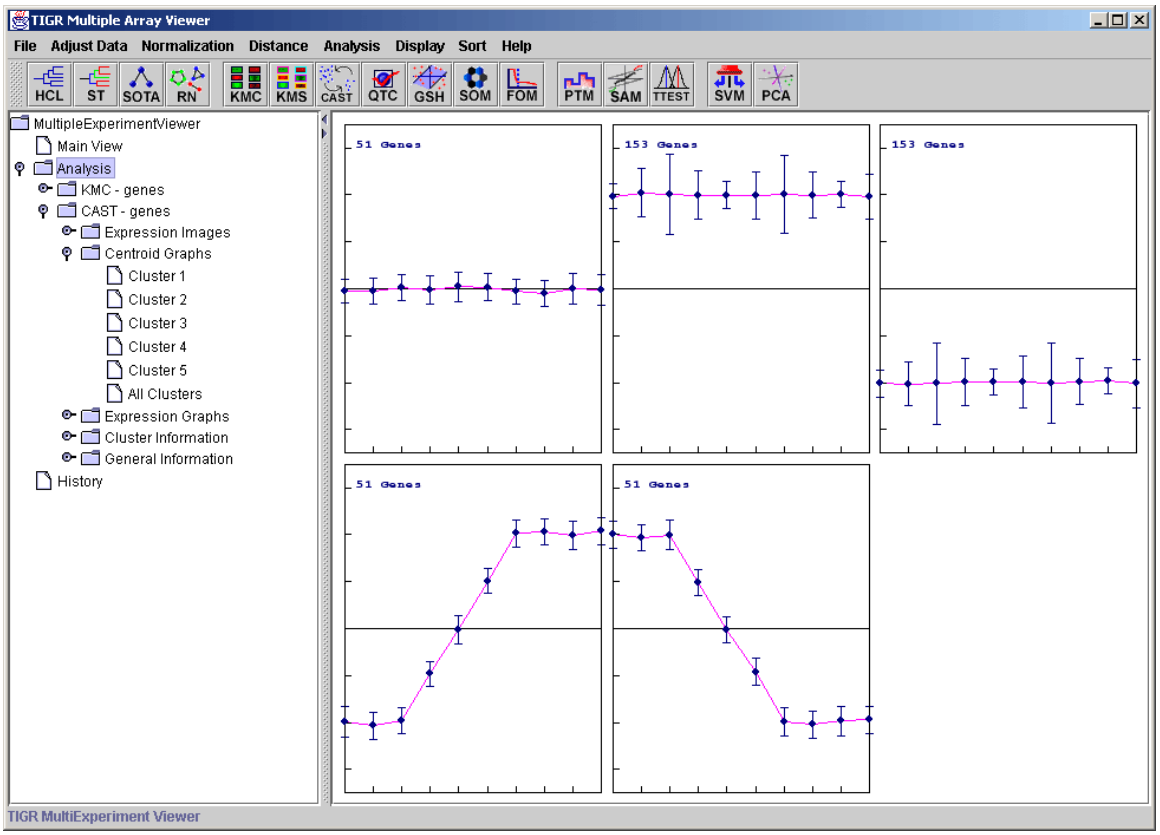
The threshold parameter is a value ranging from 0.0 to 1.0 which is used as a cluster affinity threshold. Each expression element will have an affinity for the current cluster being created based on it's relationship to the elements currently in the cluster. If that affinity is greater than the supplied 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 trail-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

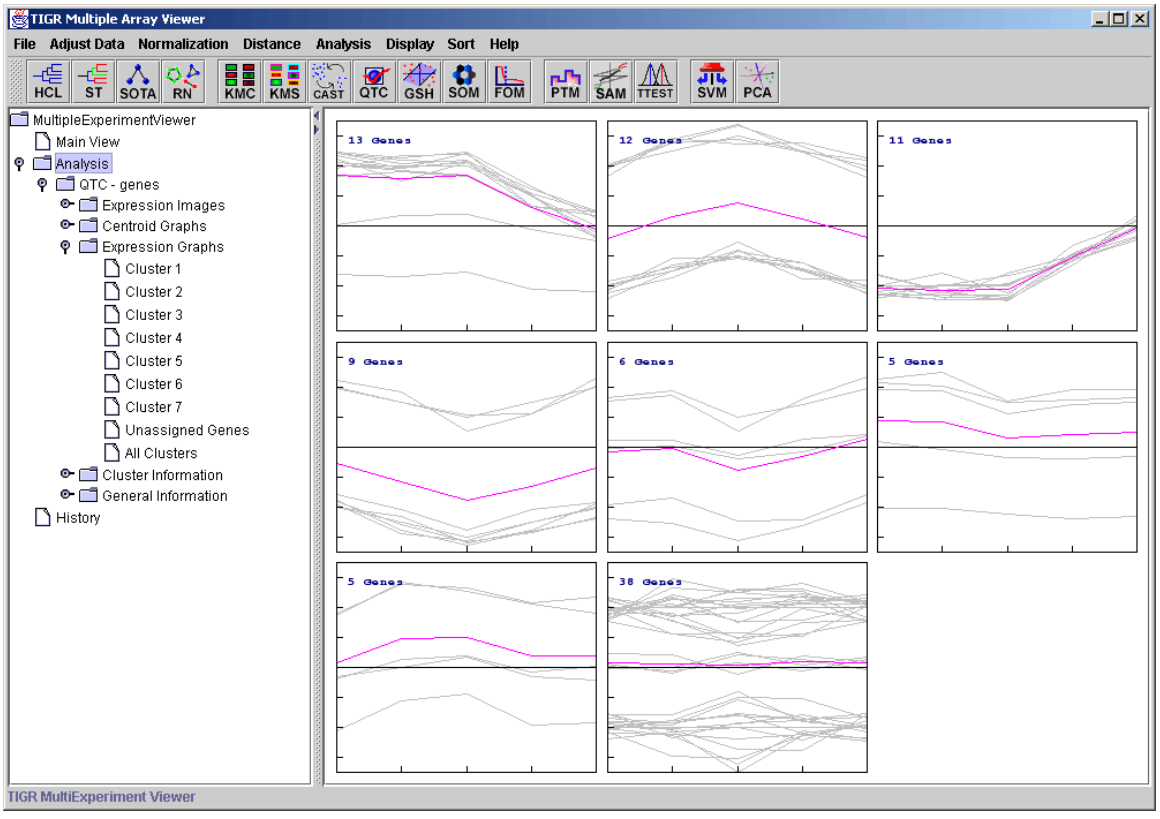


11.8.2 CAST: Centroid View.

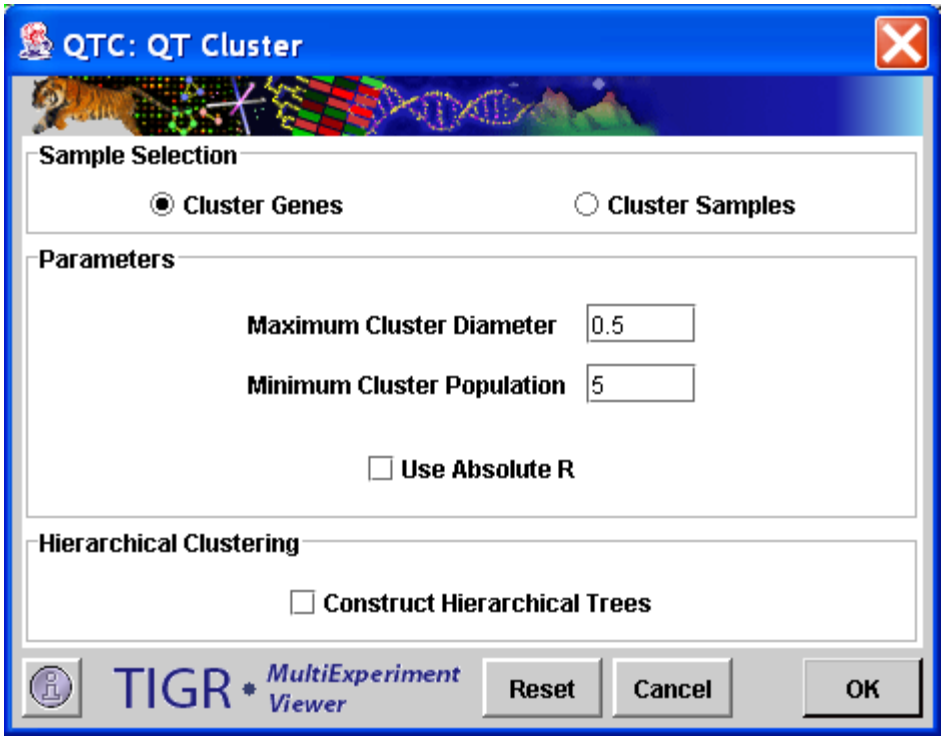
11.9 QTC: QT CLUST

(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.



11.9.1 QT CLUST: Expression Graphs.



11.9.2 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 that 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.

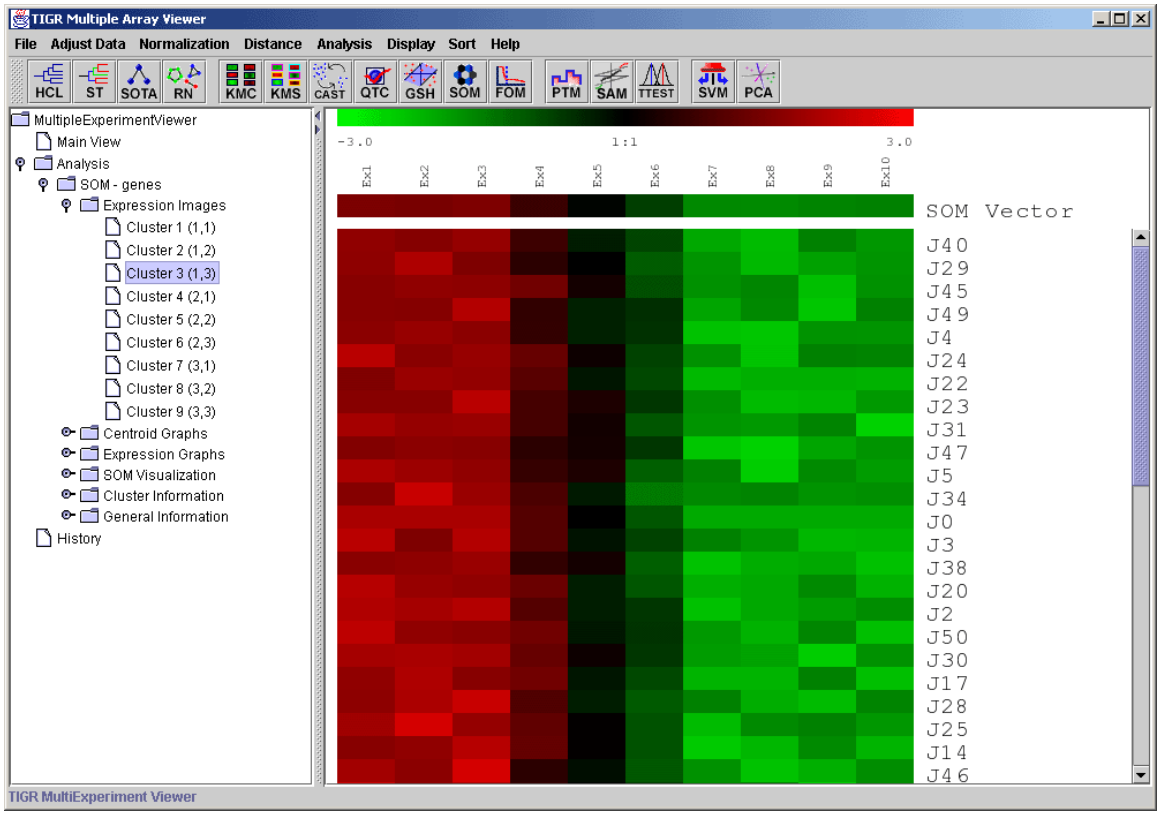
Default Distance Metric: Pearson

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

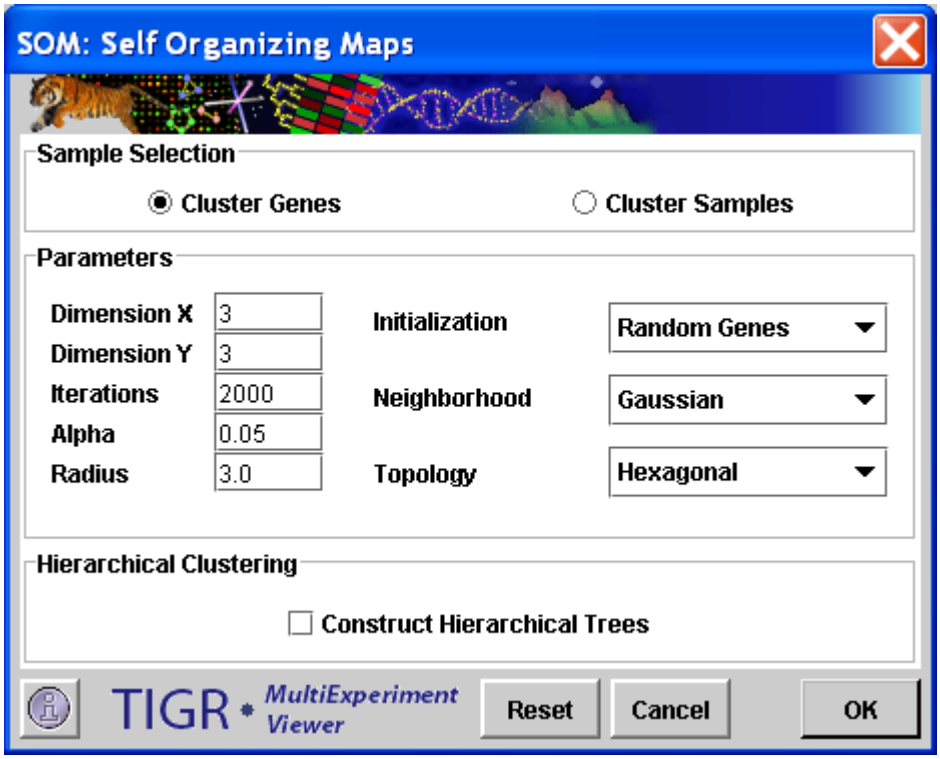
11.10 SOM: Self Organizing Maps

(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.



11.10.1 SOM: Expression Image.



11.10.2 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 it's 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

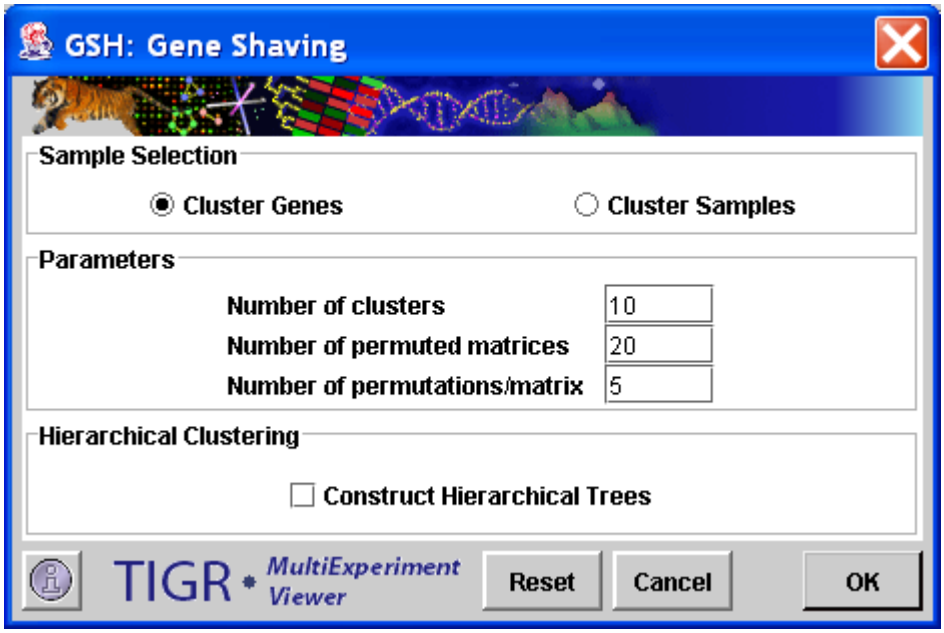
This check box selects whether to perform hierarchical clustering on the elements in each cluster created.

Default Distance Metric: Euclidian

11.11 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.



11.11.1 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.

11.12 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.

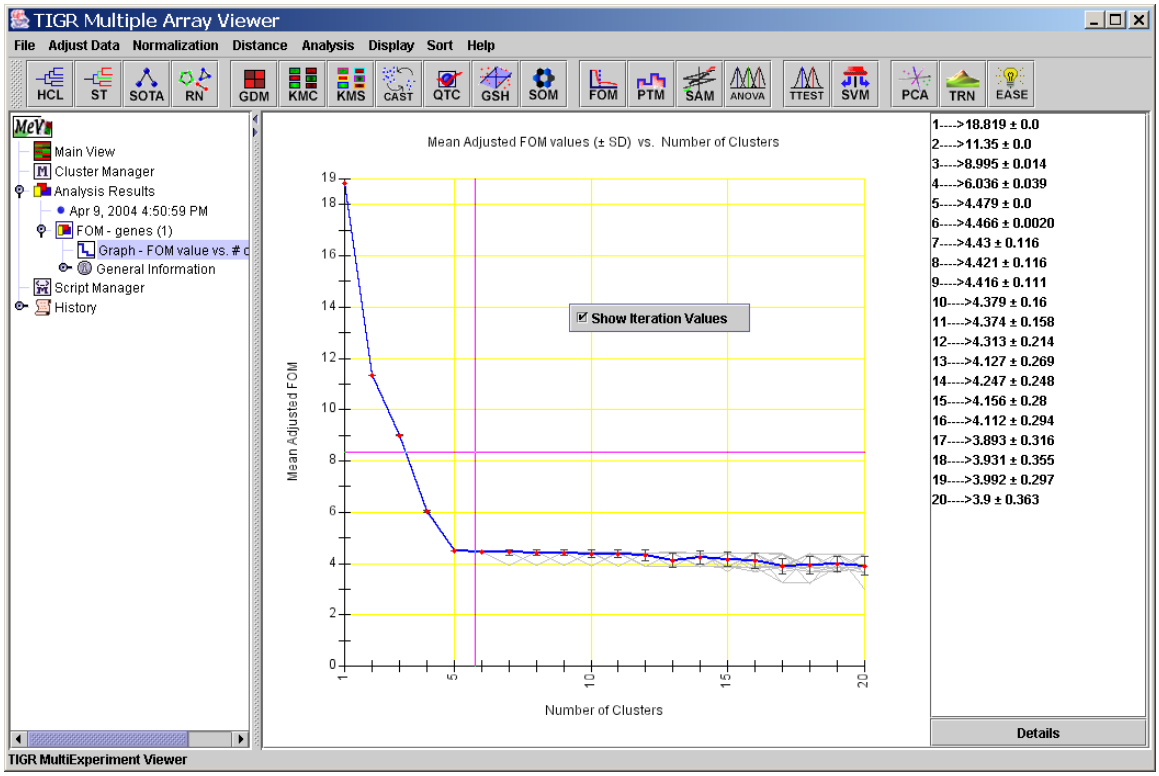
Currently, FOM is available for the CAST, K-means and K-medians algorithms. A figure of merit is an estimate of the predictive power of a clustering algorithm. It is computed by removing each sample in turn from the data set, clustering genes based on the remaining data, and calculating the fit of the 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.

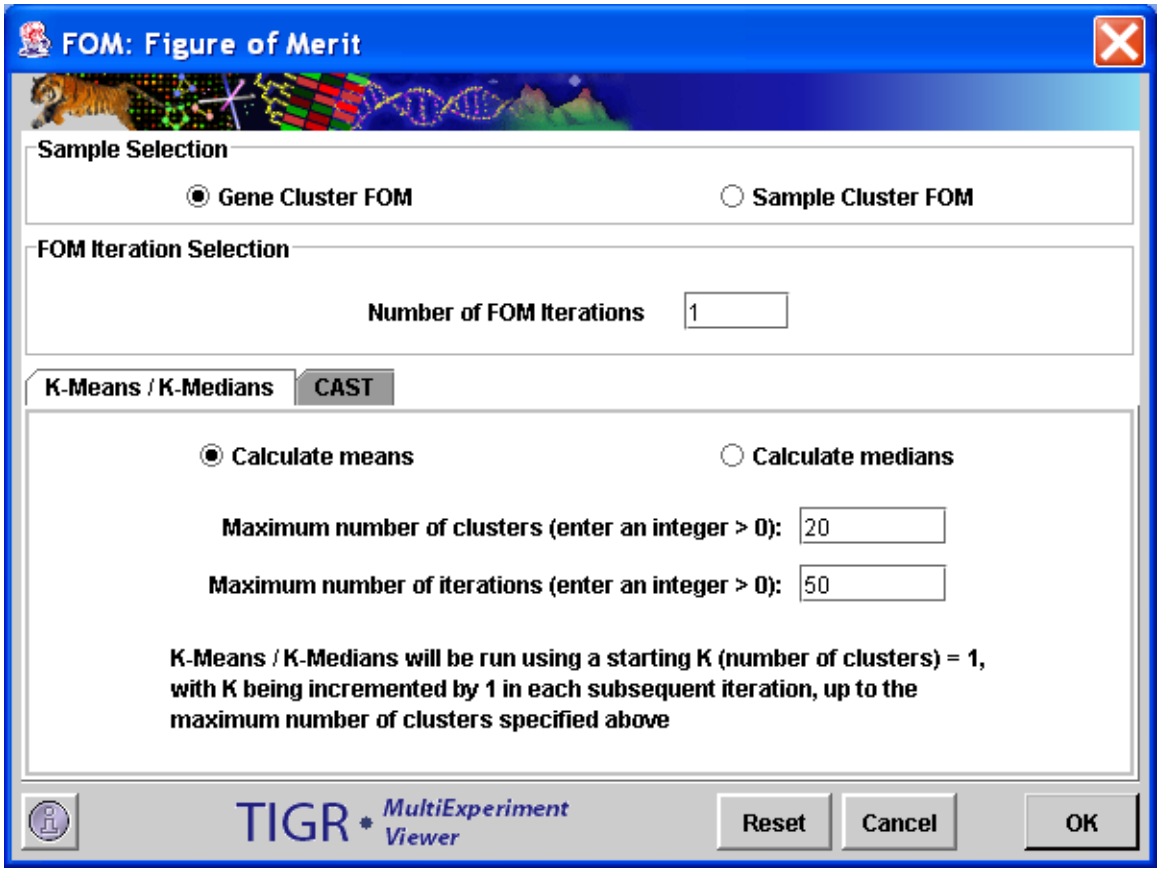
FOM Input Parameters

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.



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



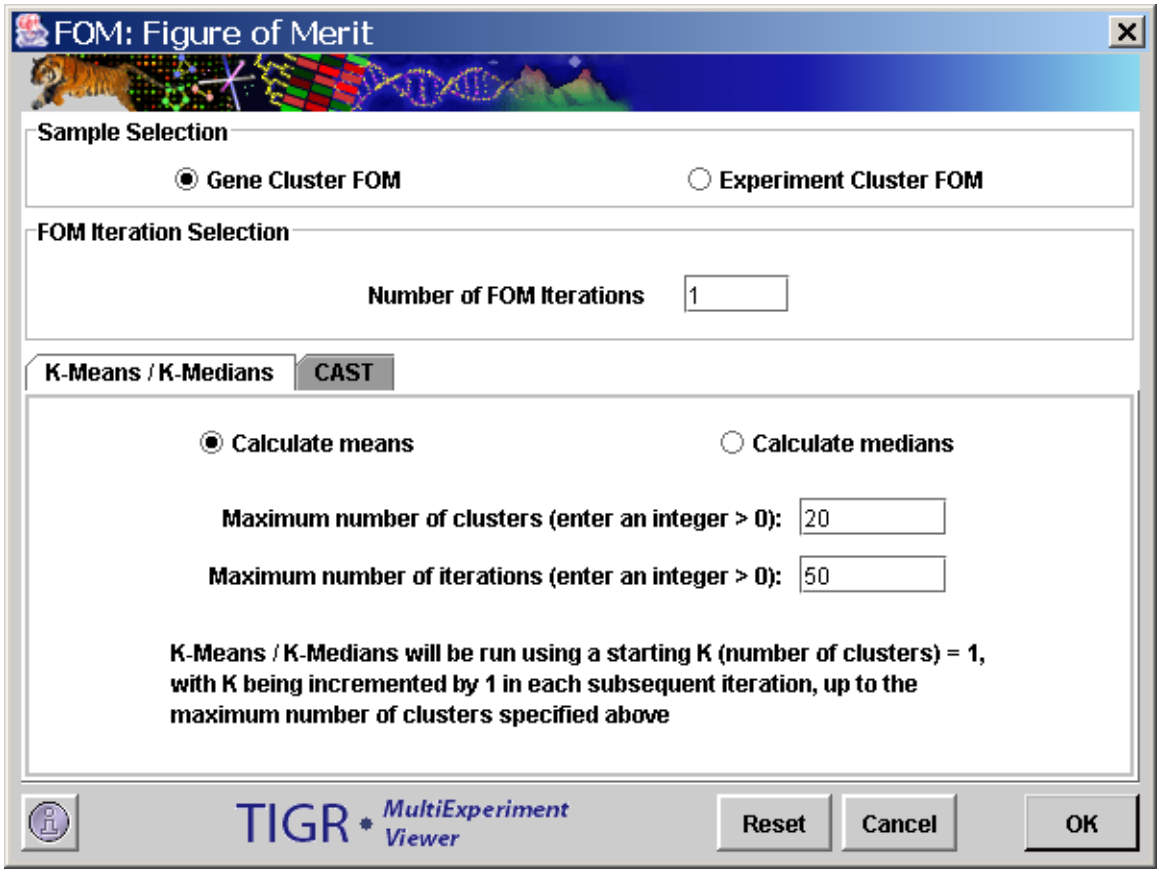
11.12.2 FOM Initialization Dialog (CAST)

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 it's 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.



11.12.3 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.

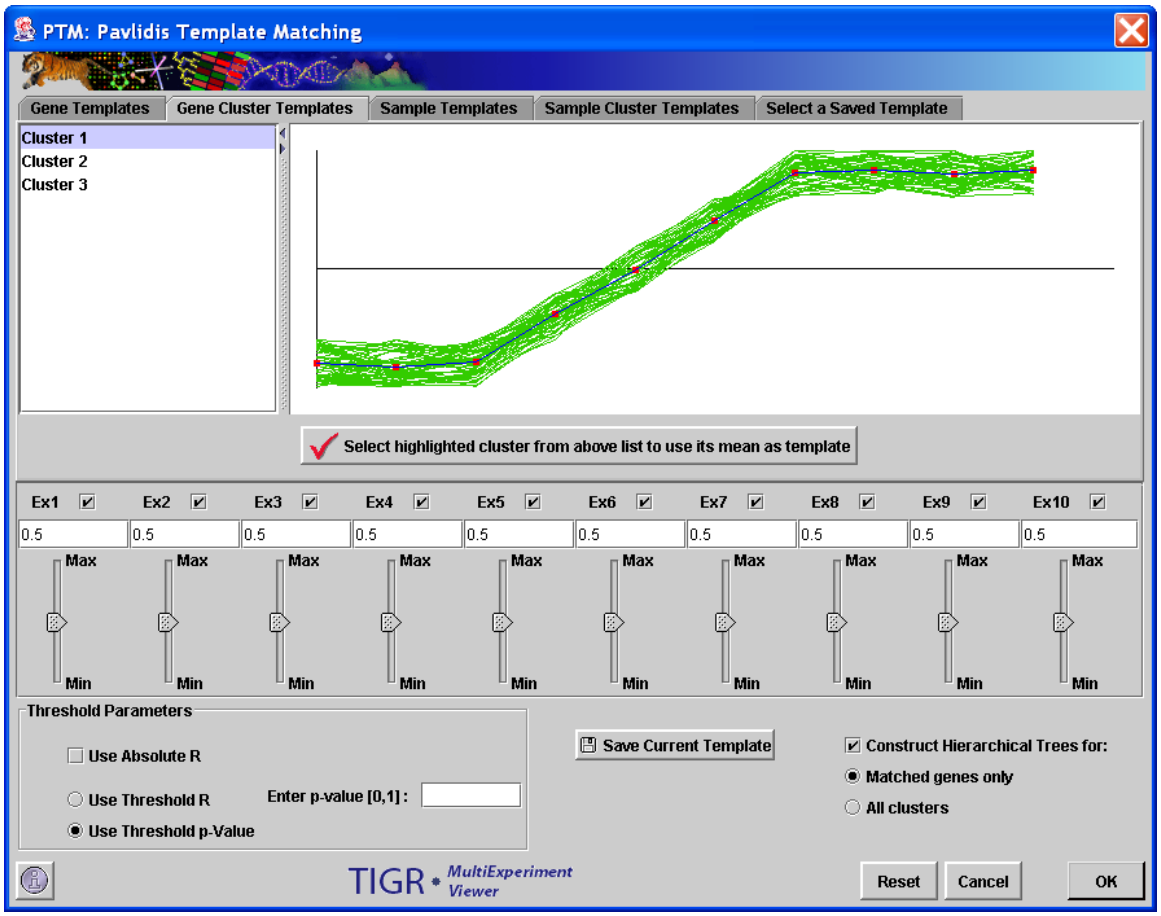
Maximum Number of Iterations

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.

11.13 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.



11.13.1 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.

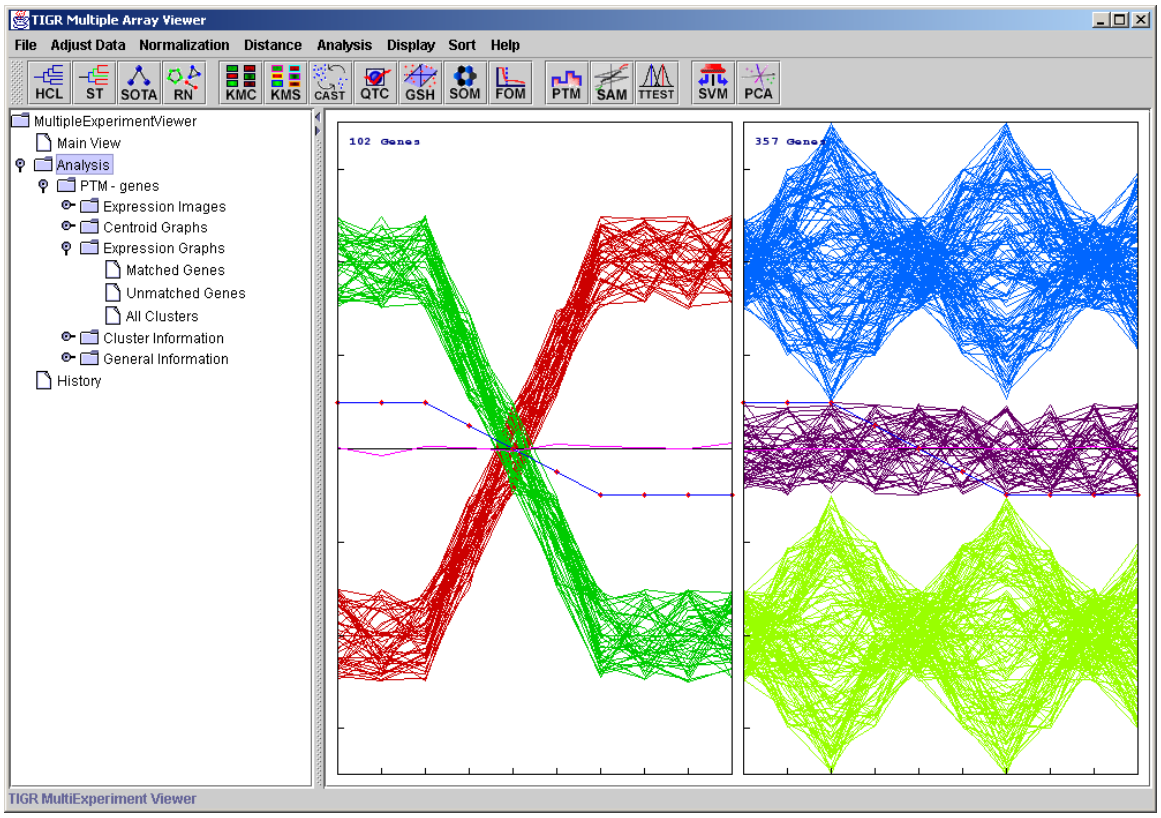
Hierarchical Clustering

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:



11.13.2 PTM results: Expression Graphs.

11.14 TTEST: T-tests

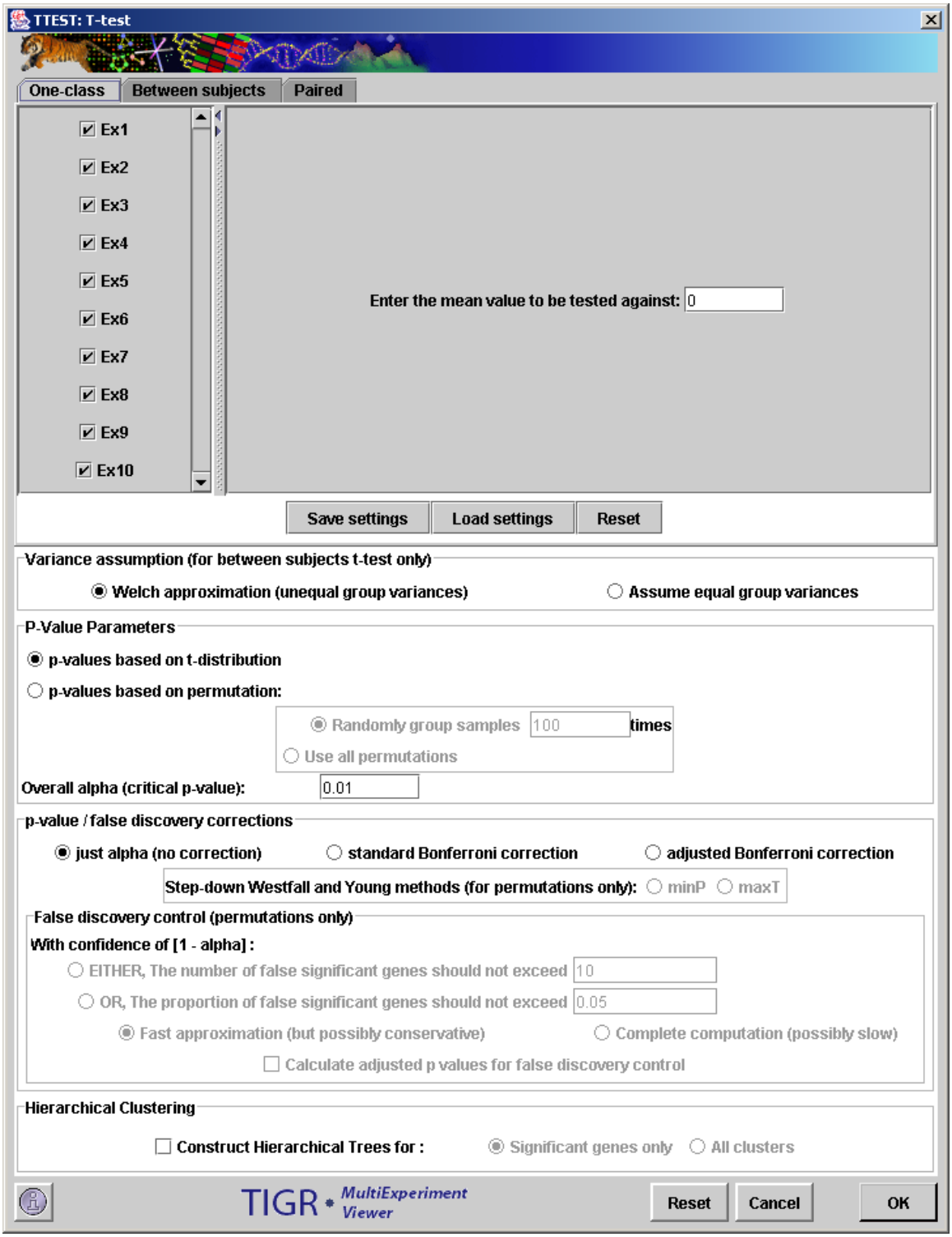
(Dudoit *et al*. 2000, Pan 2002, Welch 1947, Zar 1999, Korn, *et al.* 2001, 2004)

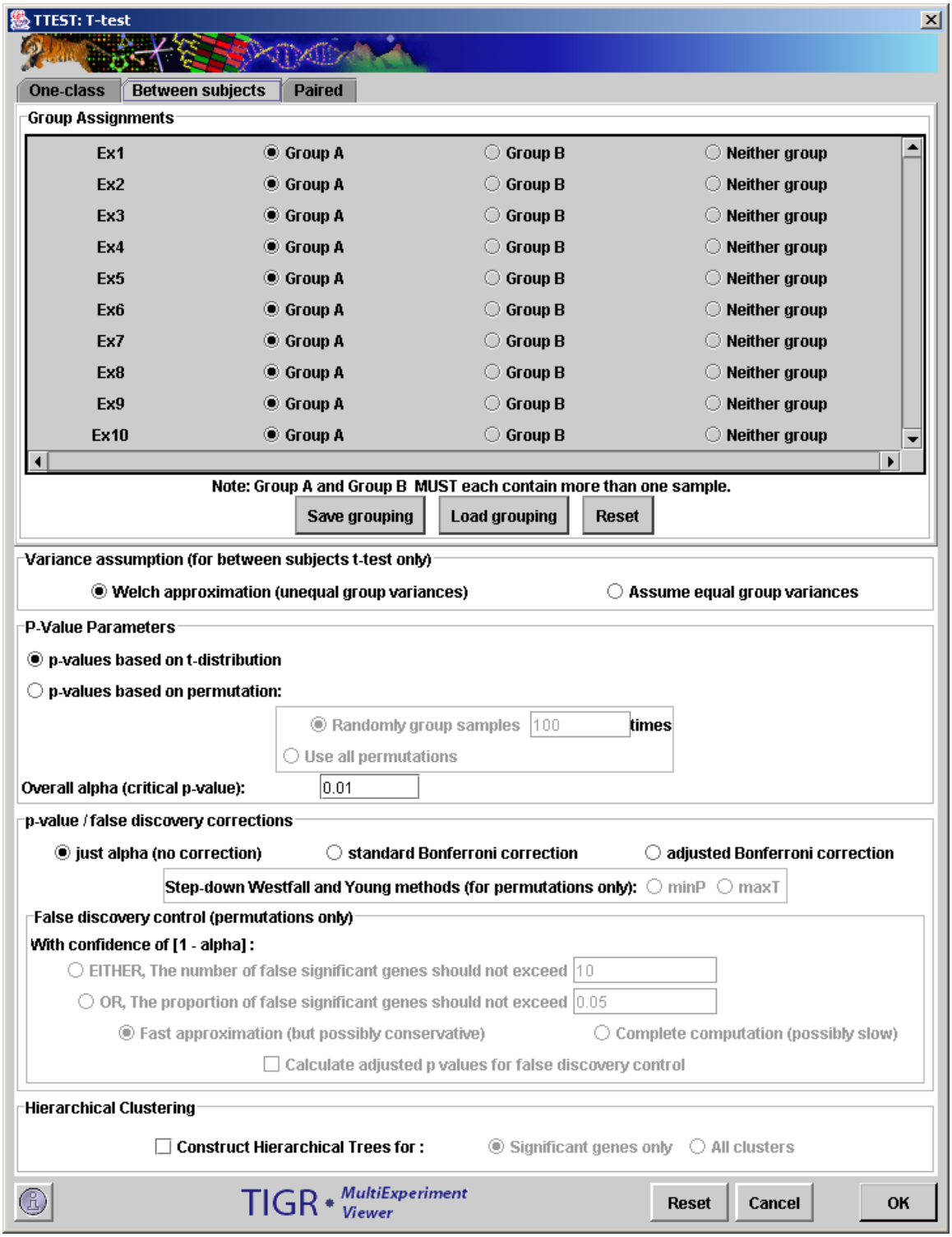
Three t-test designs are implemented: one-sample, paired and between-subjects. In the one-sample design, the user specifies a mean. Each gene whose mean log2 expression ratio over all included samples is significantly different from the user-specified mean is assigned to one cluster, while those genes 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: 0
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 100 times
Use all permutations
Overall alpha (critical p-value): 0.01
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 10
OR, The proportion of false significant genes should not exceed 0.05
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
Reset
Cancel
OK
(a)



(b)

11.14.1 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: (original value) - 2x(original value - 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 (alpha / 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:

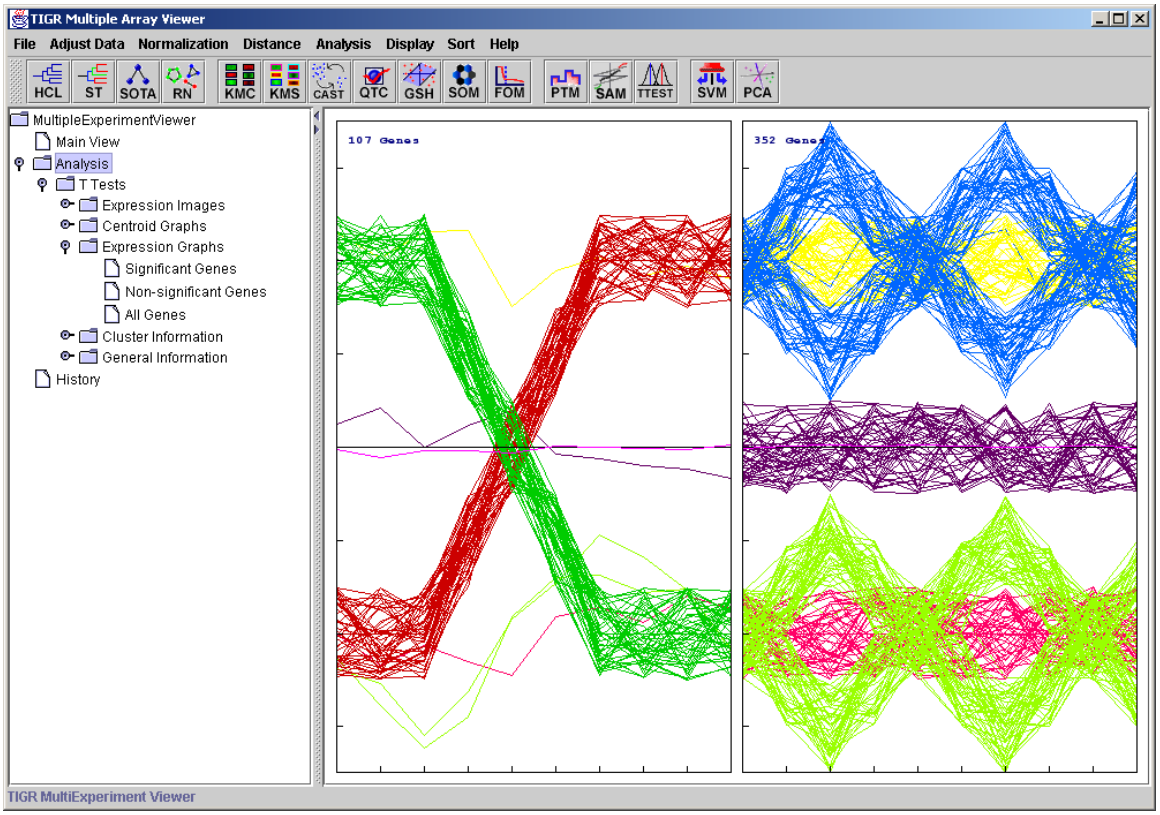
The algorithms are described in Korn *et al.* 2001, 2004.

Hierarchical Clustering

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

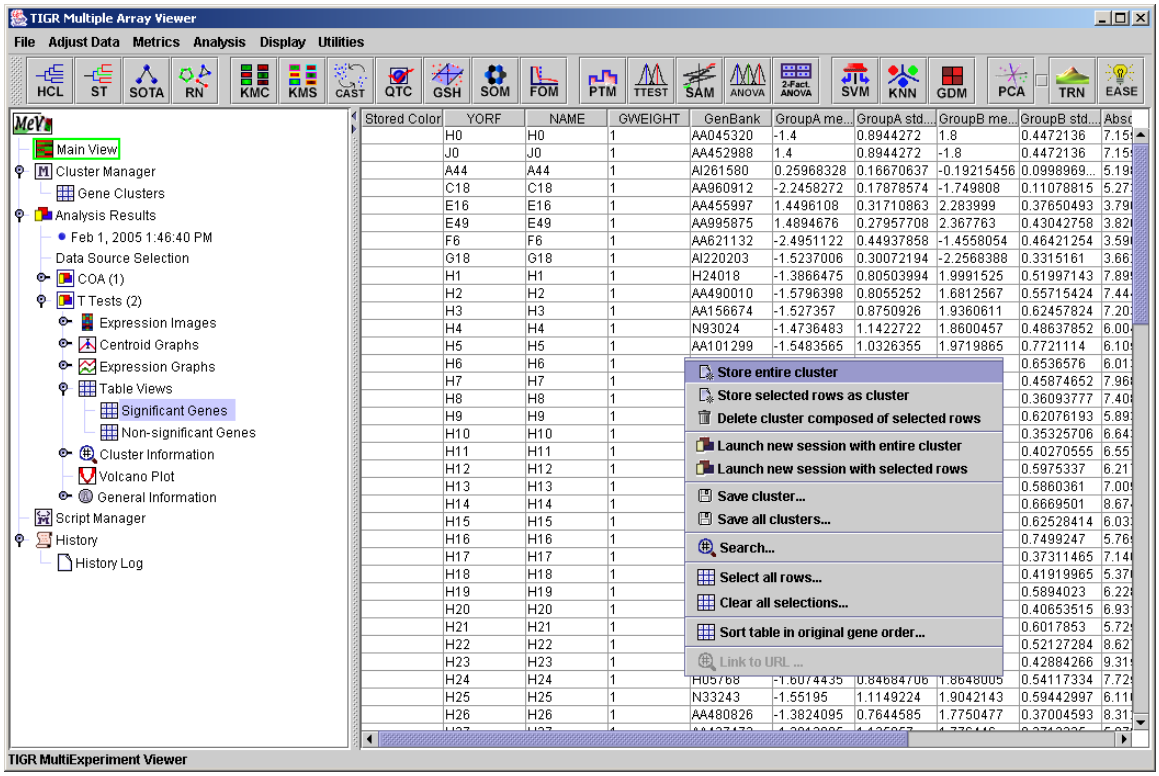
P-value corrections reduce the probability that a non-significant gene will be erroneously picked as significant. This can be a serious issue when many tests are done (which is usually the case in microarray analyses, as there are as many tests as there are genes in the analysis). The standard Bonferroni 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:



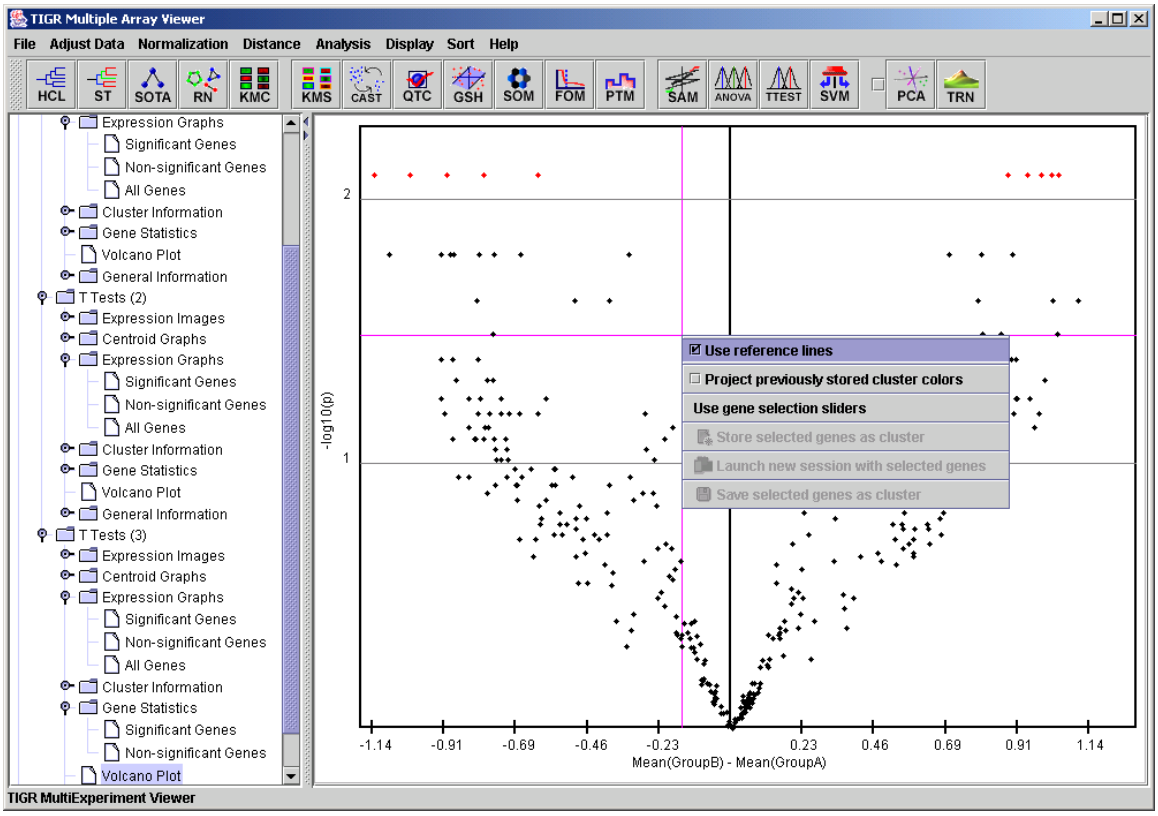
11.14.2 TTEST Results: Expression Graphs (significant genes are in the graph on the left).

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.



11.14.3 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.



11.14.4 TTEST Volcano plot

11.15 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. 0).

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).

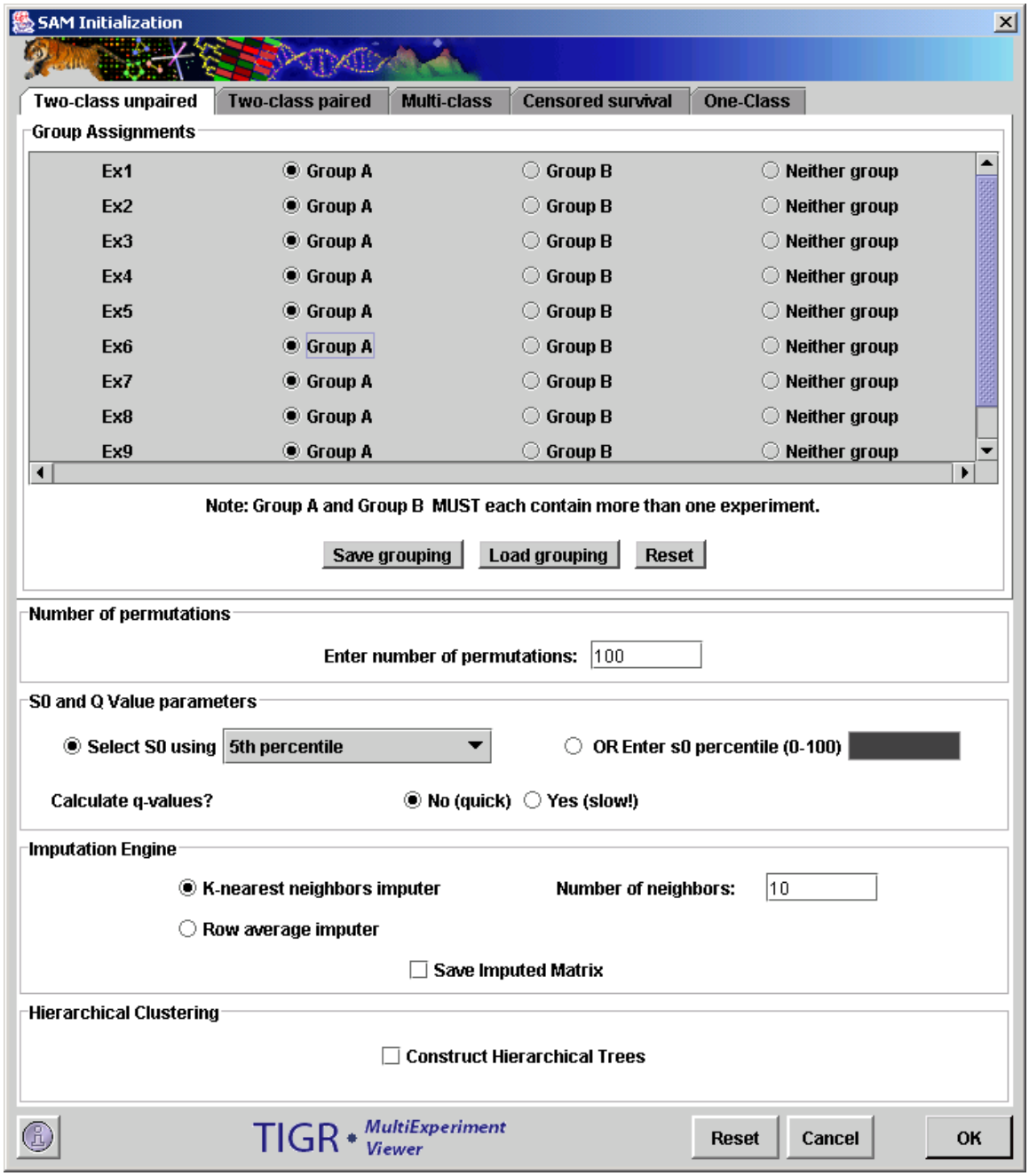
3) **Multi-class,** where the user specifies the number of groups (>2) that samples fall into. Genes will be considered significant if they are significantly different in expression across some combination of the groups.

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.

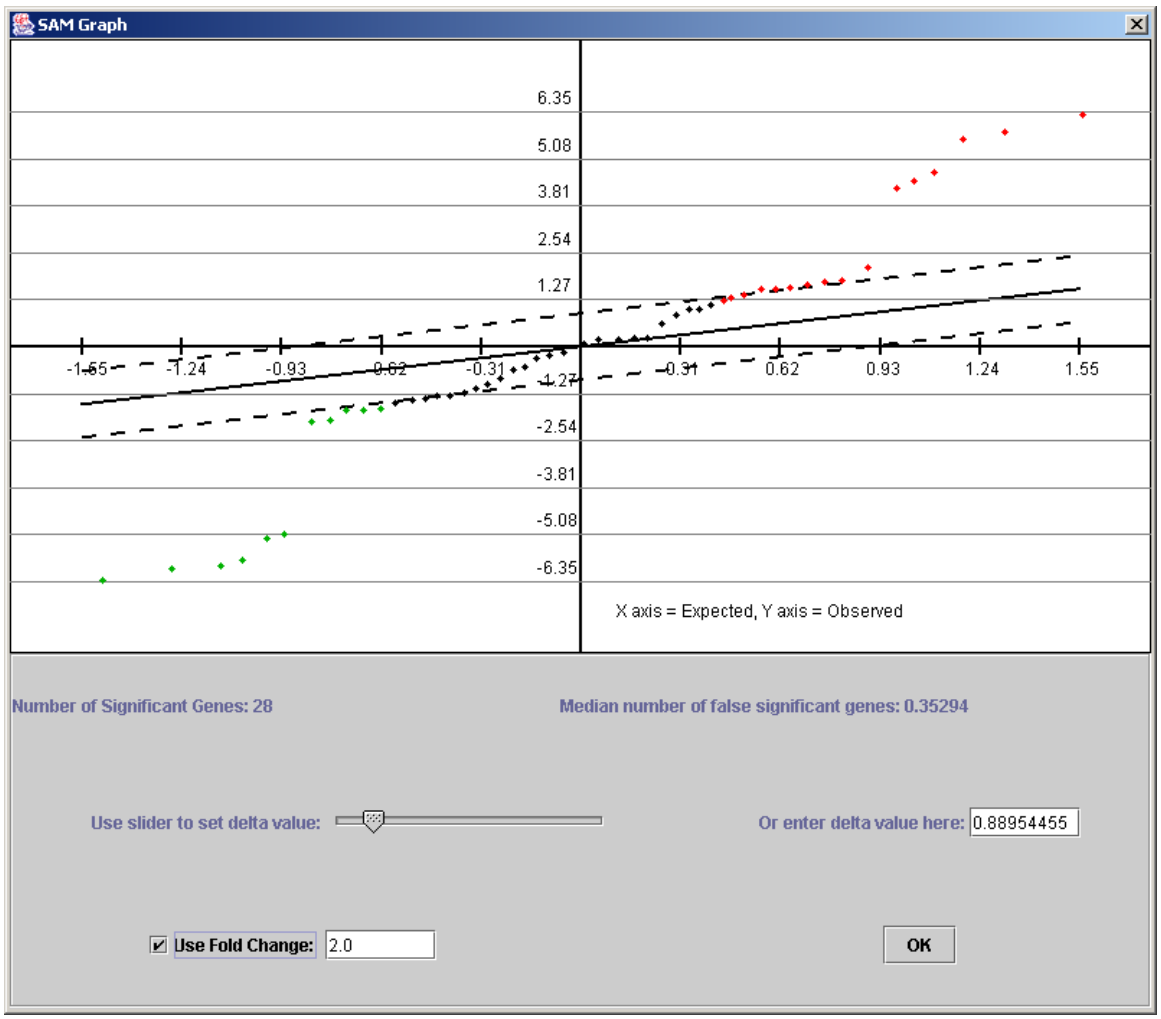


11.15.1 SAM Initialization Dialog

SAM generates an interactive plot (Fig. 0) 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**,

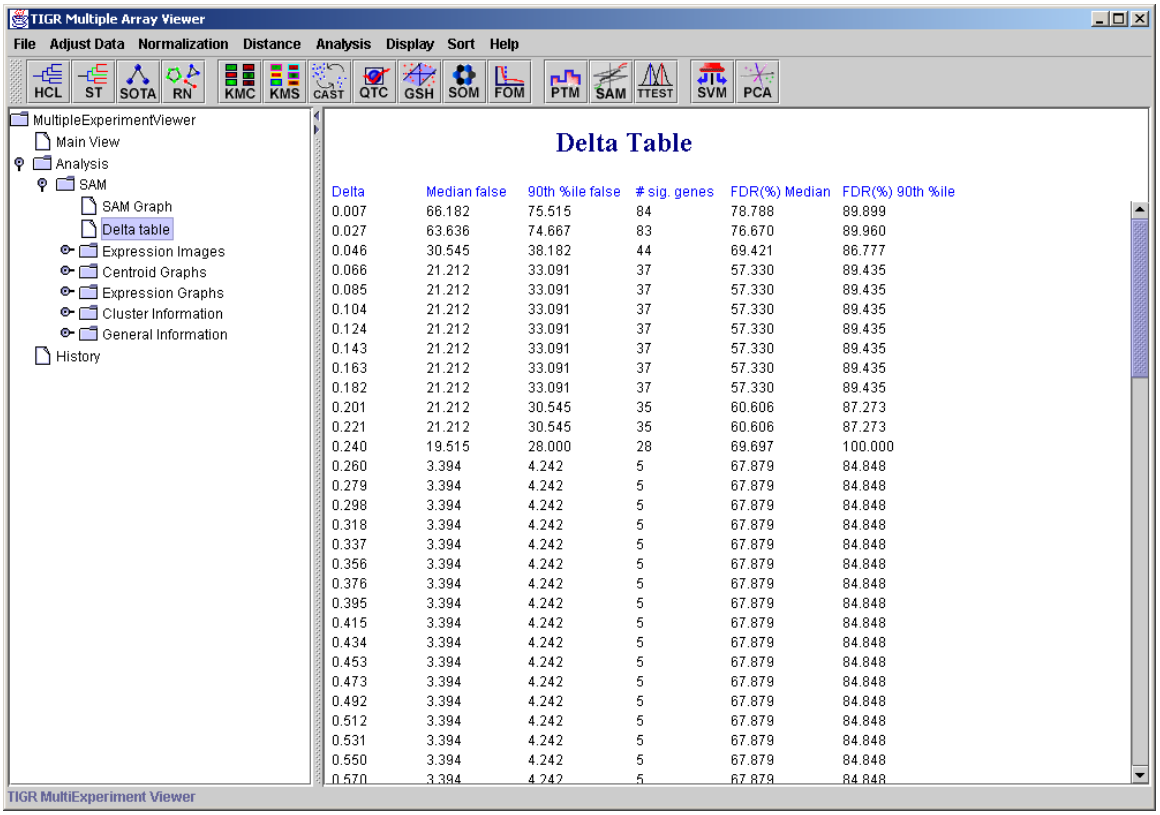
[Mean (unlogged group B) / Mean (unlogged group A values)] ≥ F (for positive significant genes), or ≤ 1/F (for negative significant genes), where F is the fold change.]



11.15.2 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 0), 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.

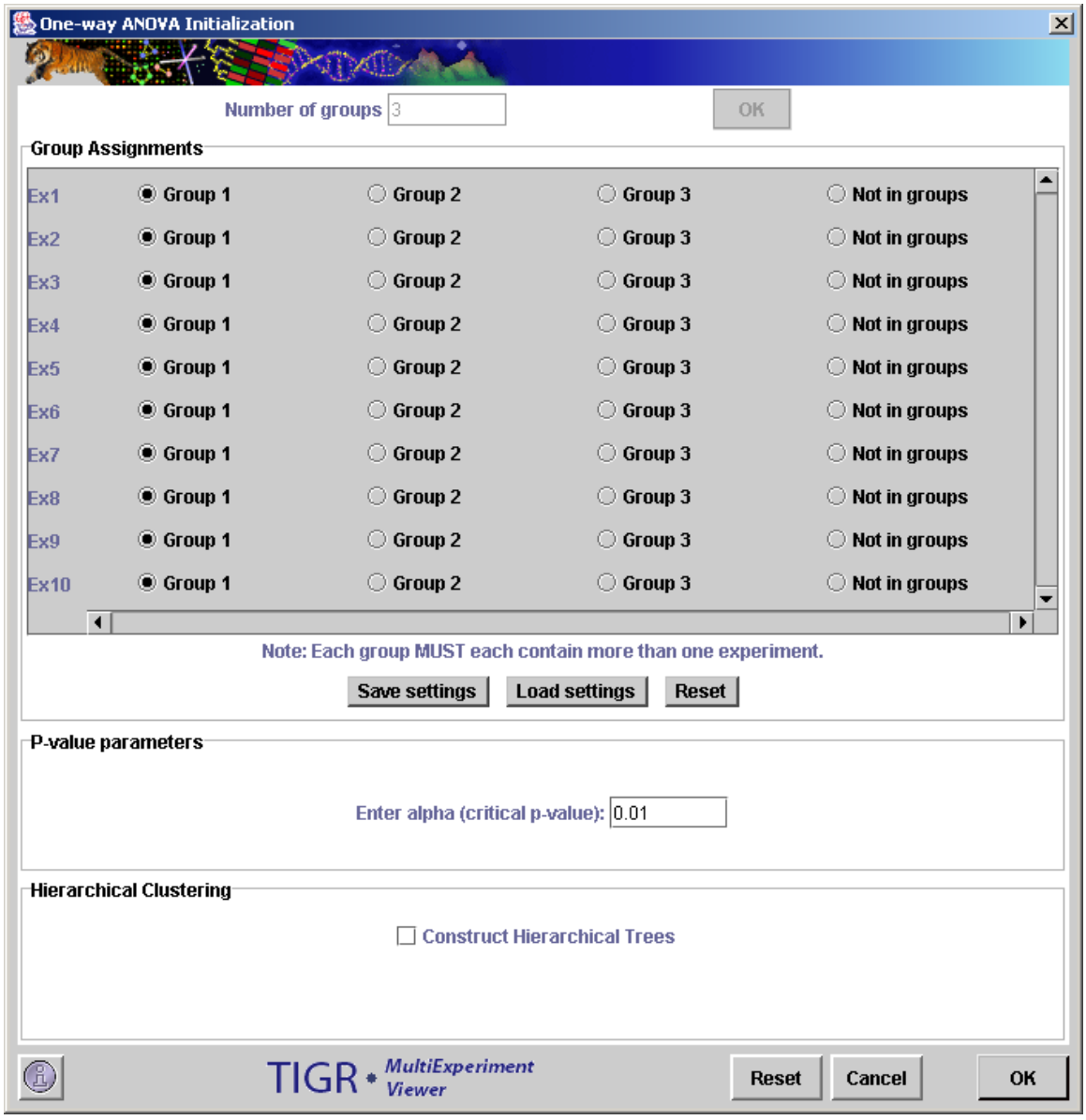


11.15.3 SAM Delta Table Viewer

11.16 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.



11.16.1 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.

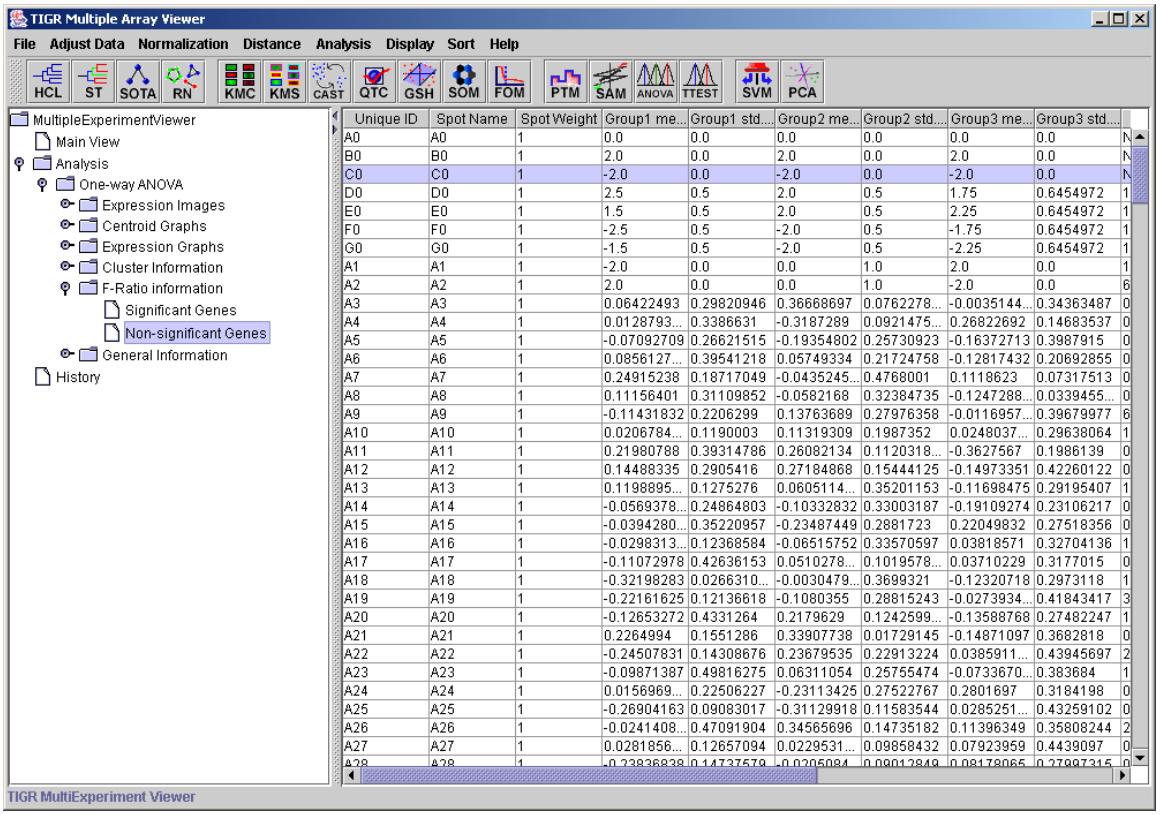
P-Value Parameters

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.



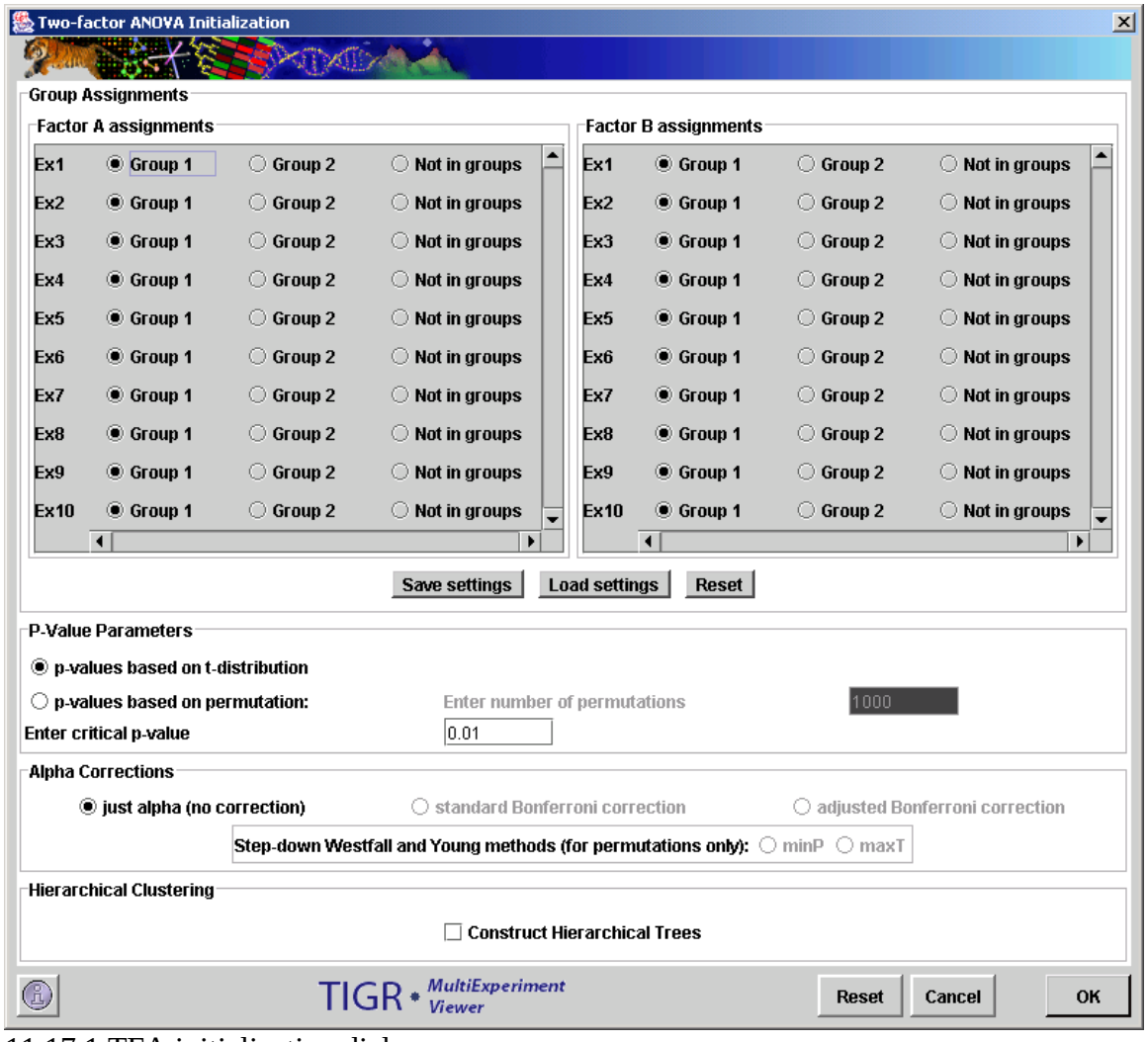
Unique ID	Spot Name	Spot Weight	Group1 me...	Group1 std...	Group2 me...	Group2 std...	Group3 me...	Group3 std...	
A0	A0	1	0.0	0.0	0.0	0.0	0.0	0.0	N
B0	B0	1	2.0	0.0	2.0	0.0	2.0	0.0	N
C0	C0	1	-2.0	0.0	-2.0	0.0	-2.0	0.0	N
D0	D0	1	2.5	0.5	2.0	0.5	1.75	0.6454972	1
E0	E0	1	1.5	0.5	2.0	0.5	2.25	0.6454972	1
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	-0.23836838	0.14737579	-0.0205084...	0.09012849	0.08178965	0.27997315	0

11.16.2 One-way ANOVA: F-Ratio Information viewer

11.17 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:



11.17.1 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 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.

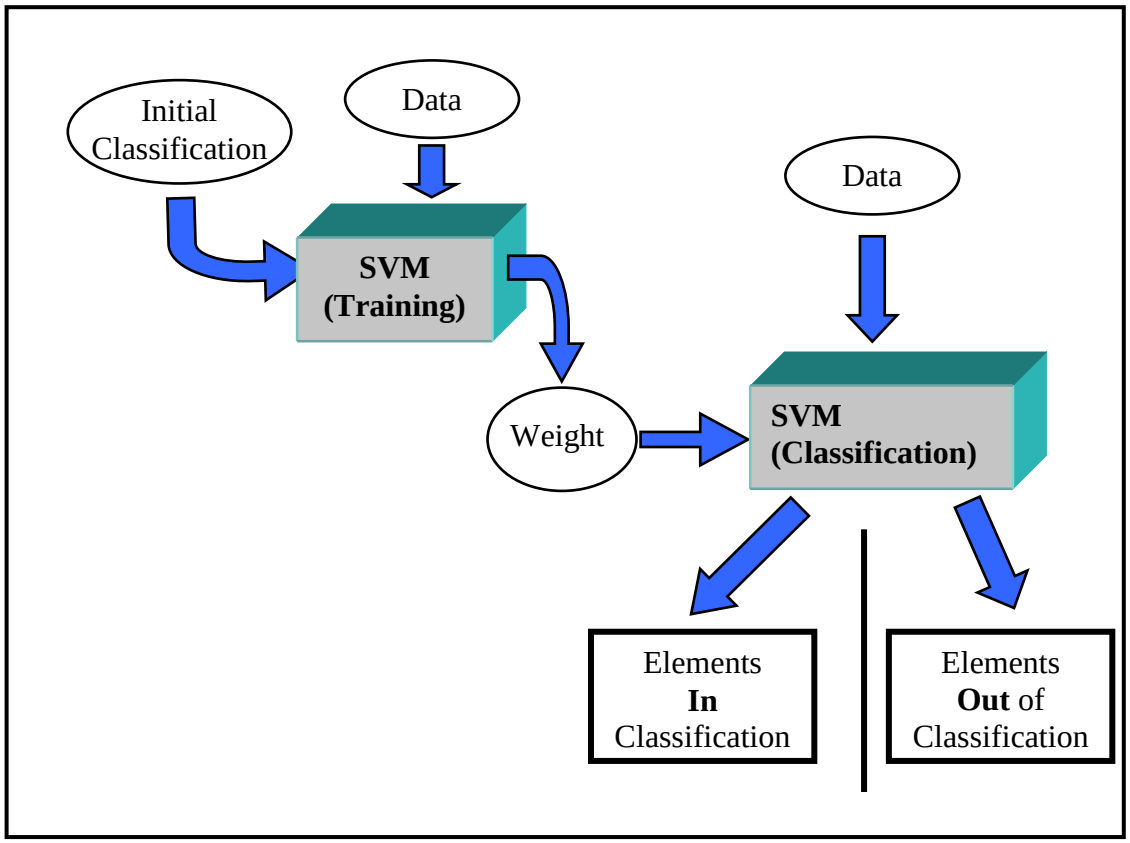
- Unbalanced designs where one or more cells have only one sample, or no samples, are not tested for any effects.

11.18 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. 0)



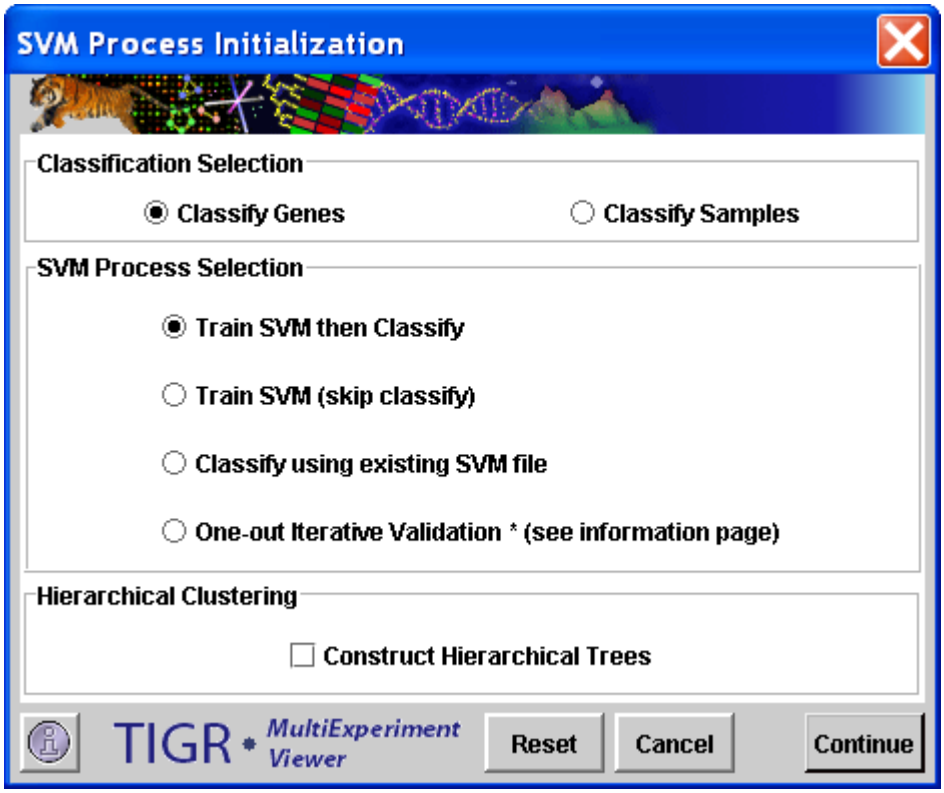
11.18.1 SVM Process Overview

SVM Dialog Overview

The initial dialog (Fig. 0) 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. 0) 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.



11.18.2 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.

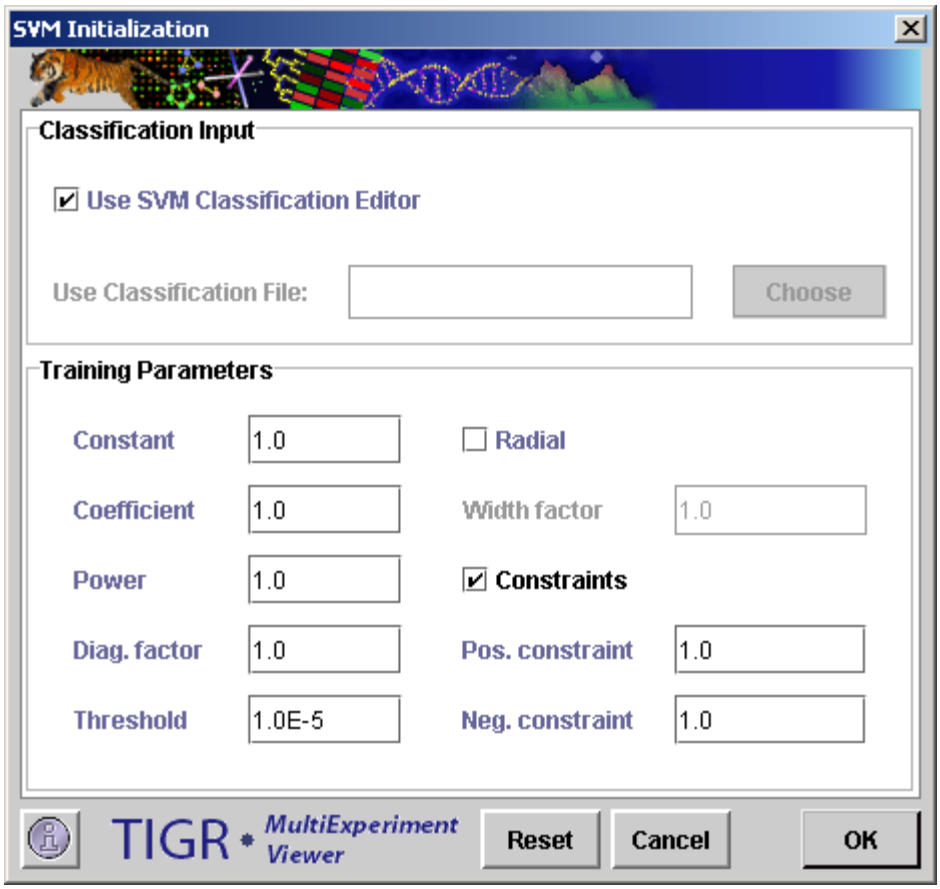
The *Training Only* option results in a set of numerical weights which can stored as an SVM file and used for classification at a later time.

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 notbe 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.



11.18.3 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 kernal 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 yeild 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,

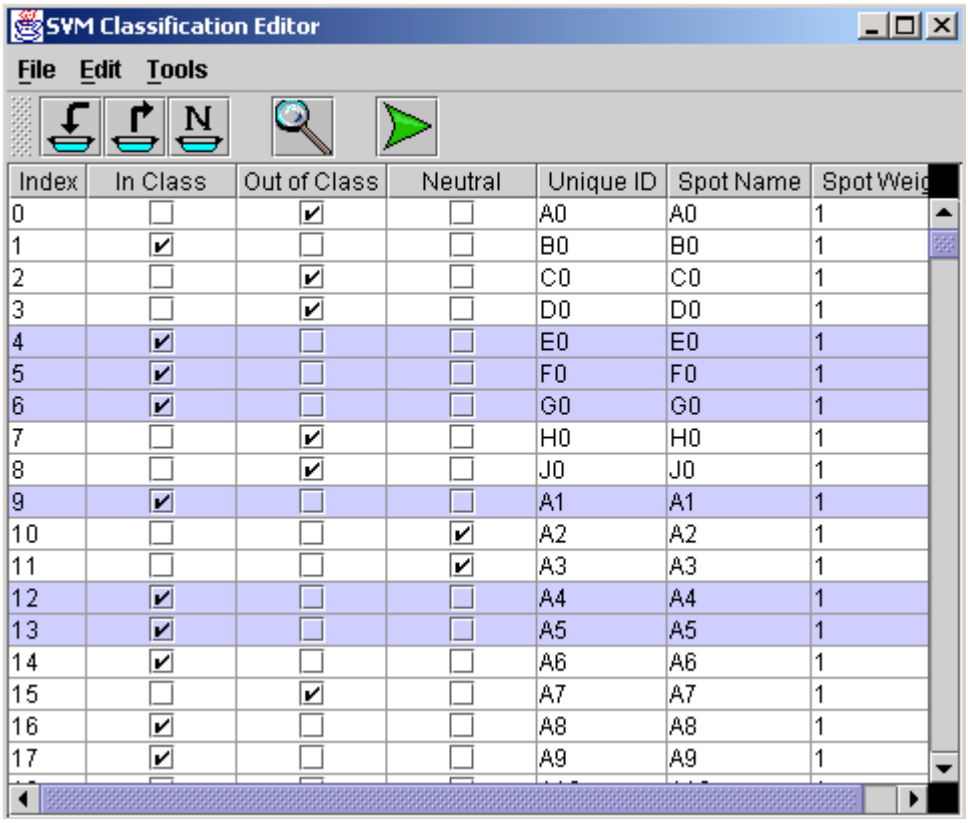
1.) Index - a sequential integer index.

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. 0) 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.



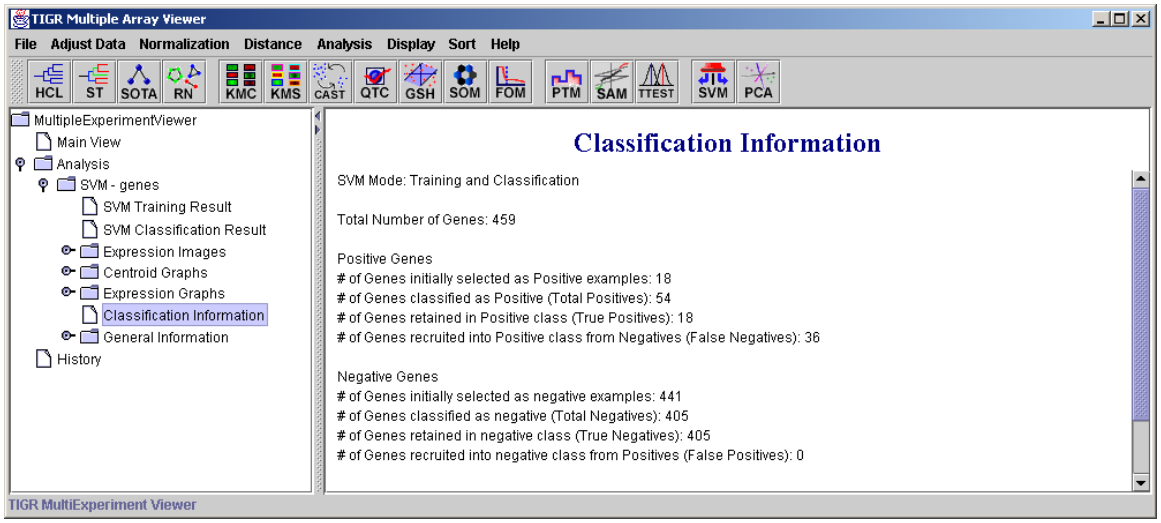
Index	In Class	Out of Class	Neutral	Unique ID	Spot Name	Spot Weig
0	☐	☑	☐	A0	A0	1
1	☑	☐	☐	B0	B0	1
2	☐	☑	☐	C0	C0	1
3	☐	☑	☐	D0	D0	1
4	☑	☐	☐	E0	E0	1
5	☑	☐	☐	F0	F0	1
6	☑	☐	☐	G0	G0	1
7	☐	☑	☐	H0	H0	1
8	☐	☑	☐	J0	J0	1
9	☑	☐	☐	A1	A1	1
10	☐	☐	☑	A2	A2	1
11	☐	☐	☑	A3	A3	1
12	☑	☐	☐	A4	A4	1
13	☑	☐	☐	A5	A5	1
14	☑	☐	☐	A6	A6	1
15	☐	☑	☐	A7	A7	1
16	☑	☐	☐	A8	A8	1
17	☑	☐	☐	A9	A9	1

11.18.4 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.



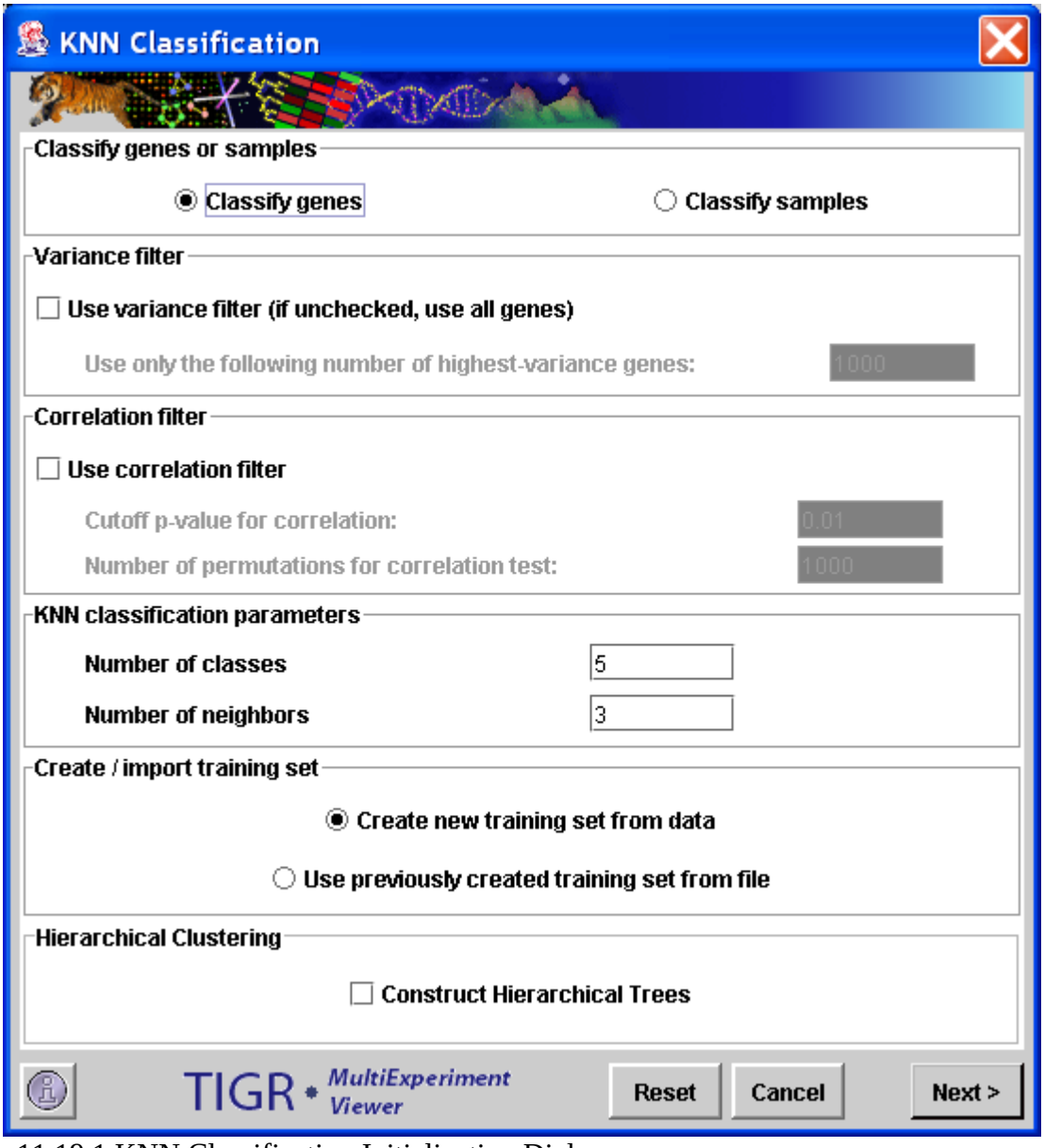
11.18.5 Classification Information Viewer

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.

11.19 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.



11.19.1 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.

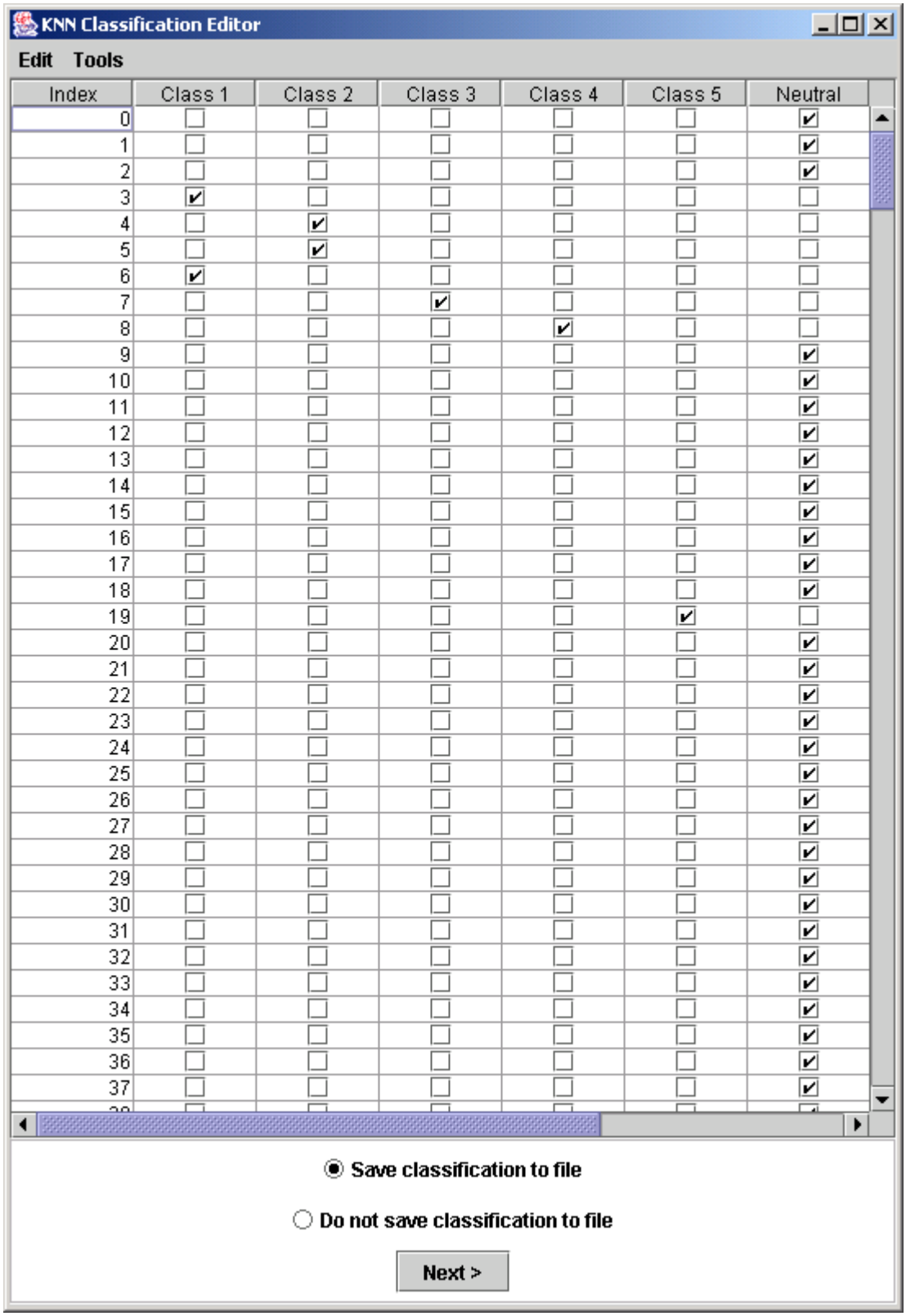
Create / import training set

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.



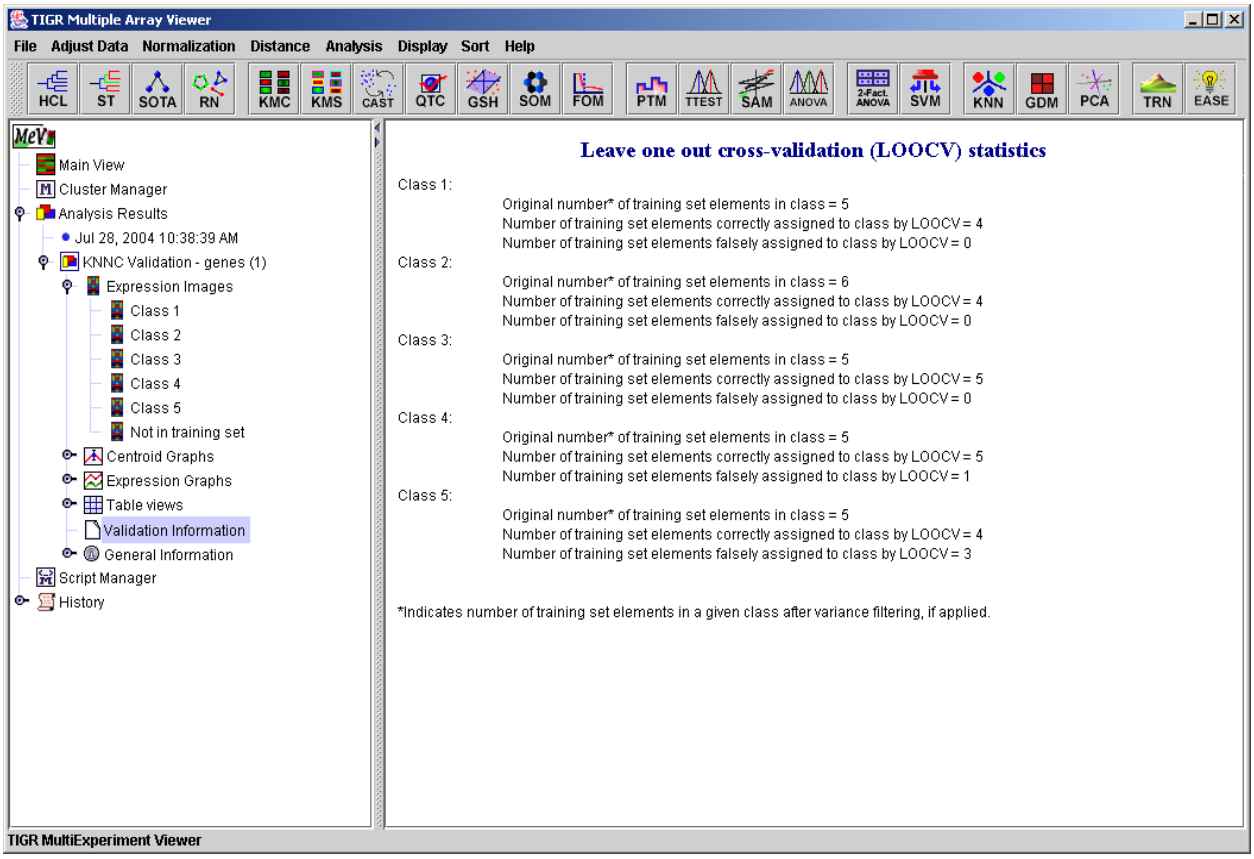
11.19.2 KNN Classification editor

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.

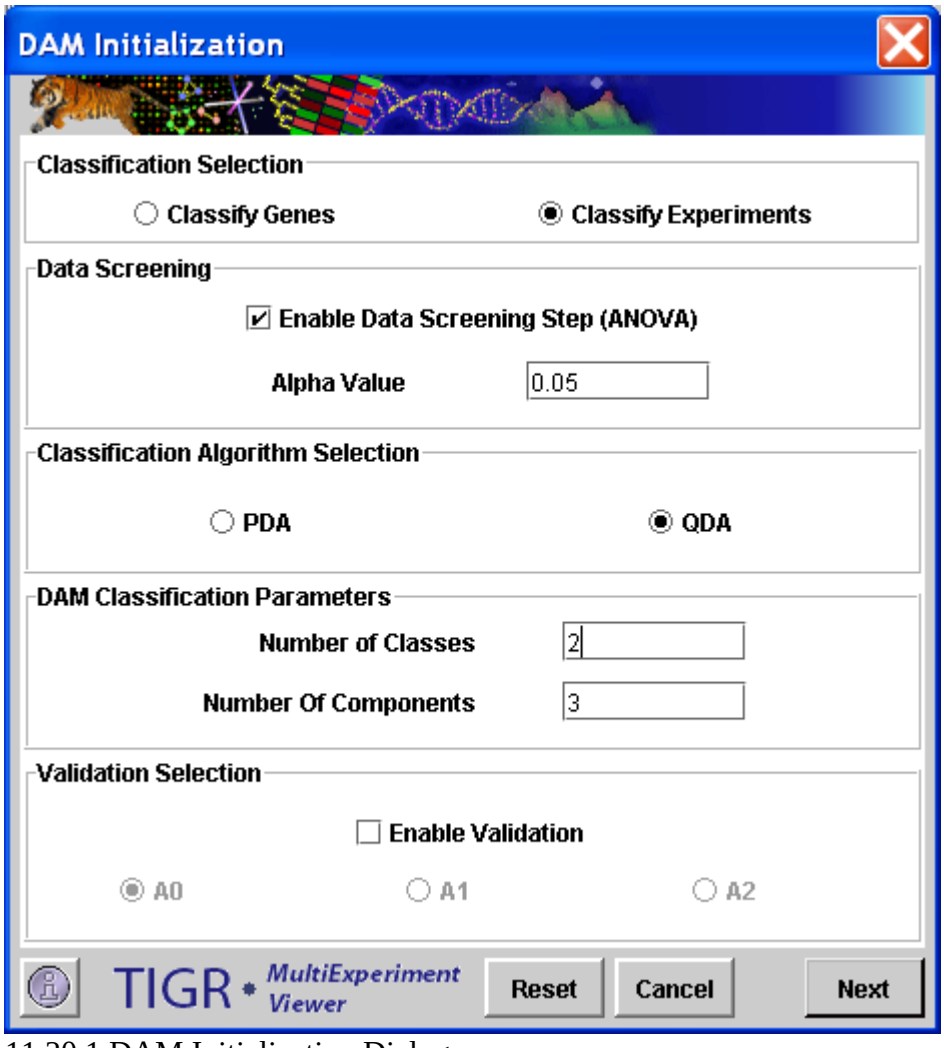


10.18.3 KNN Classification cross-validation statistics viewer

11.20 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.



11.20.1 DAM Initialization Dialog

Parameter Selection

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

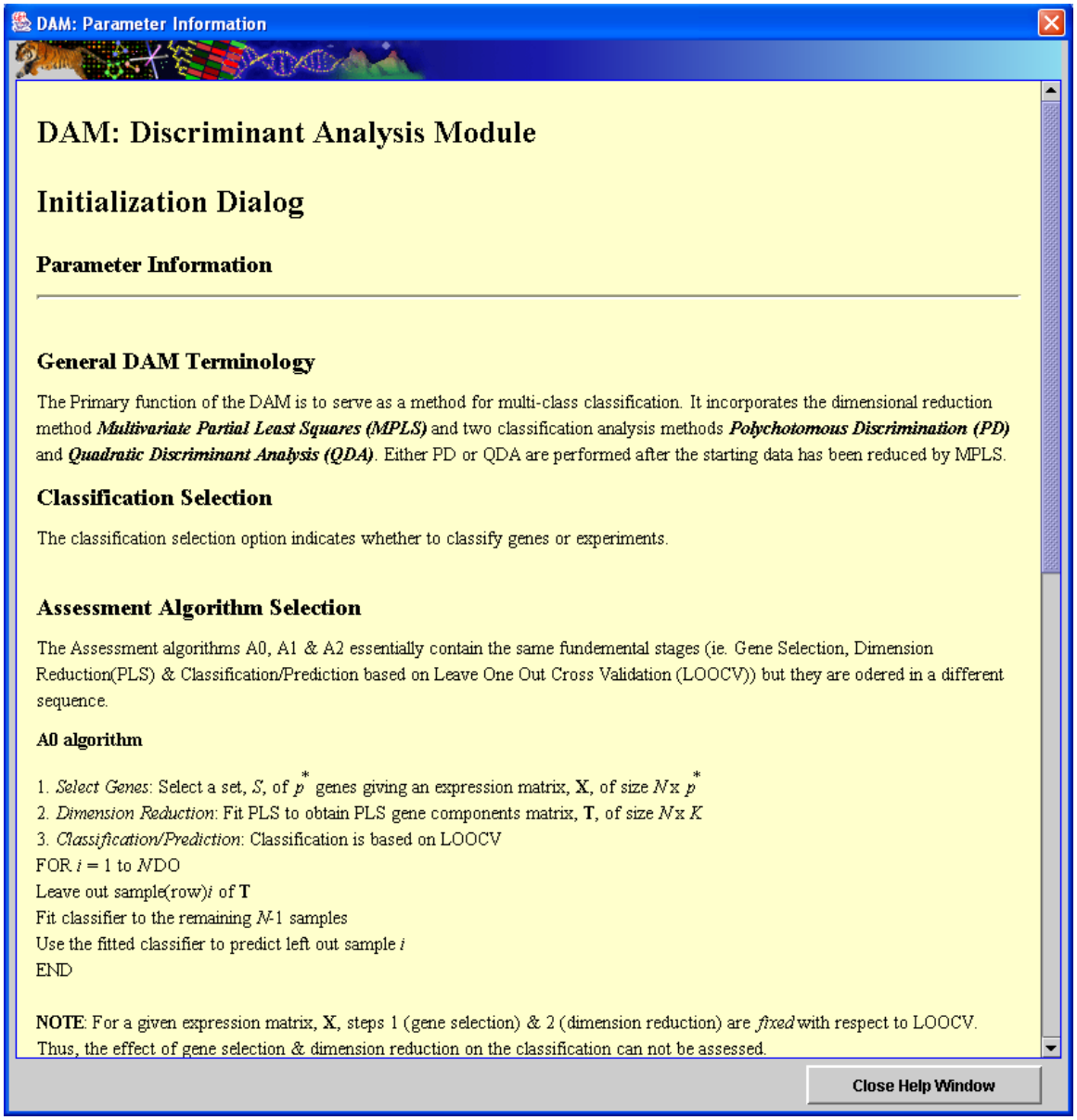
The *Data Screening* panel allows the users to select to apply a filter to select genes (or samples if classifying genes) from the loaded set that should be near optimal for partitioning the elements to classify based on Analysis of Variance (ANOVA) on permutations of the known (training) class members. 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 portioning 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.

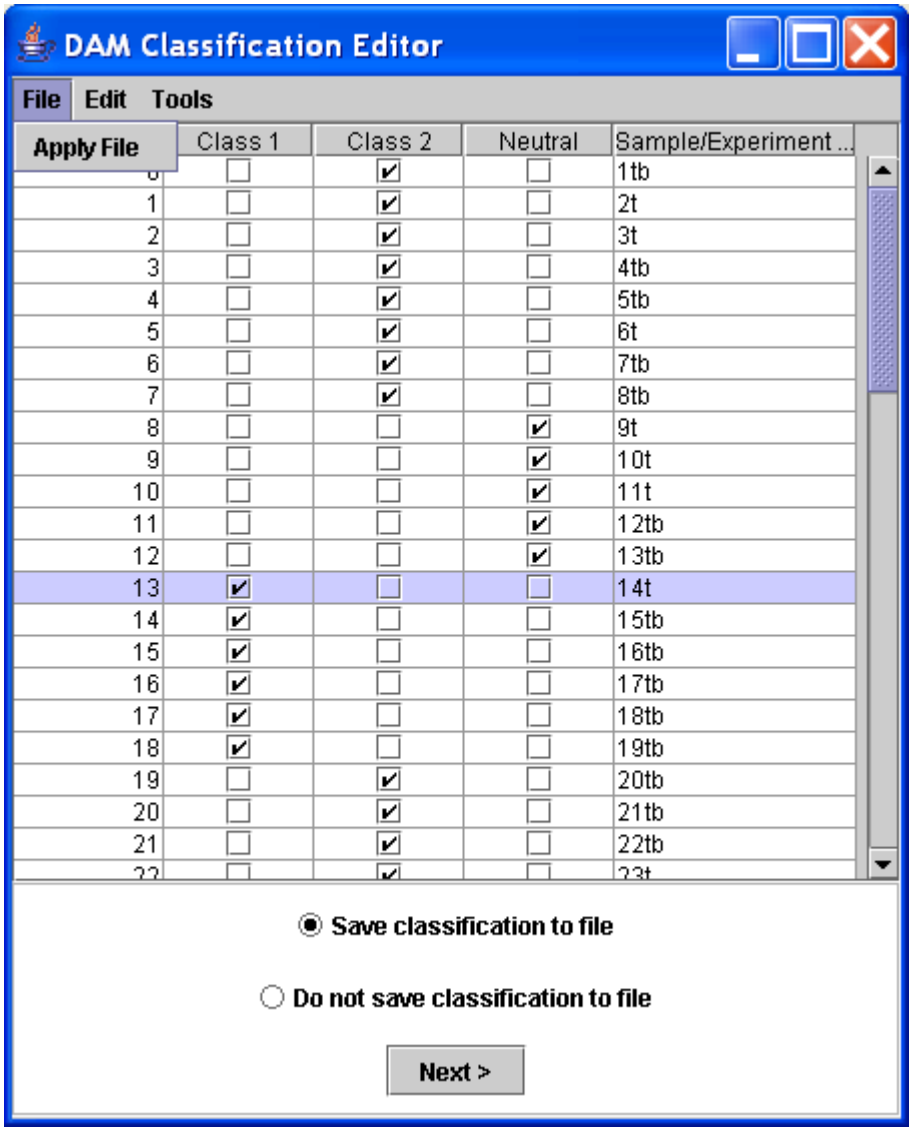
When the Information button at the bottom-left corner of the DAM Initialization Dialog is depressed, an Information Dialog screen is popped up. This Dialog contains brief description of the terminologies in DAM, and references page describing the methods and parameters used in DAM.



11.20.2 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.



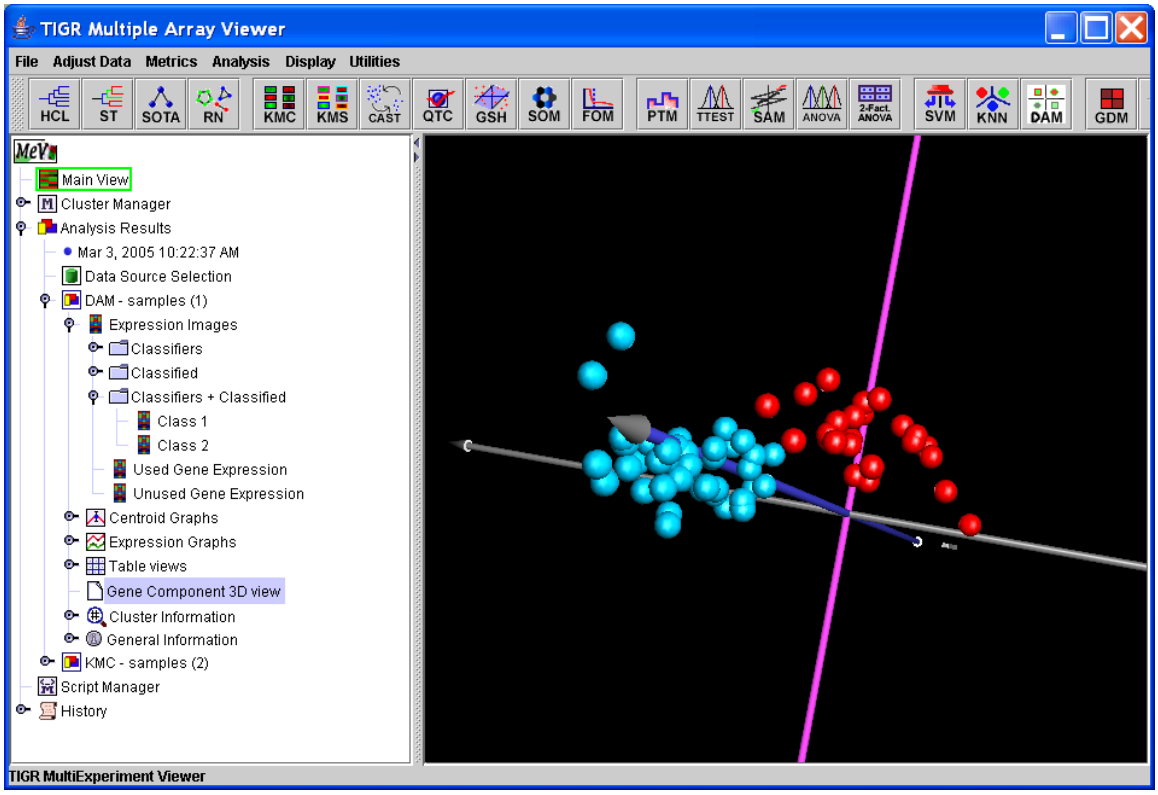
11.20.3 DAM Classification Editor

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.



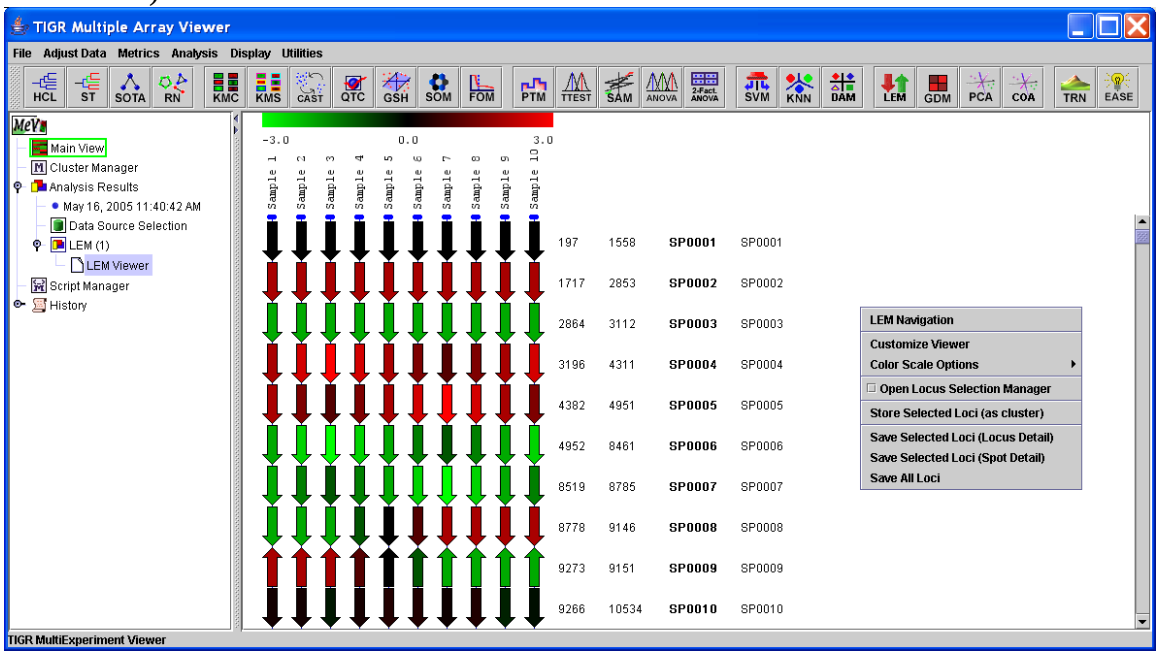
11.20.4 3D Component Projection View

11.21 LEM: Linear Expression Maps

The Linear Expression Map module produces a viewer to display locus level expression information organized by chromosomal location. The LEM aligns the sampled loci from multiple samples in a single viewer so that expression patterns for loci and groups of loci can be visualized. The expression values from all spots on the slide that map to a particular locus are averaged to produce a single value for a locus. The map is highly customizable in appearance and has fine and course navigational controls to assist in stepping through the genome or chromosome being viewed. The LEM features also include options to view locus detail, options to view locus related spot annotation information, and the option to link to web resources for additional information about a selected locus.

General Appearance

The LEM organizes loci by chromosomal location where loci are represented as arrows that indicate the direction of transcription. Initially, the LEM will be configured to have fixed locus arrow lengths and fixed length open areas between sampled loci. The header indicates the sample related to each column of loci on the map. The arrow colors represent the mean expression for the locus in the measured sample. The initial color representation takes the color from a gradient within the displayed range in the header. The annotation for each locus includes the 5' location, the 3' location, and the selected locus identifier. The last column is reserved for a user selected field of annotation that can be selected from the 'Display' menu of MeV (Please see manual section 6.2, Selecting Gene Annotation).



11.21.1 LEM viewer with fixed length locus arrows

Requirements for LEM Construction

- *Locus Identifier* – an annotation key that will map spots to loci.
- *Coordinate Information* – 5' and 3' coordinates that correspond to the set of loci.

- *Chromosome Information* – an annotation key to indicate on which chromosome the loci is located. This information is not needed in the case of organisms that have only one chromosome.

  The *locus identifier* should be an annotation field loaded with the annotation during the expression file load in MeV. If the input data is in mev format then the locus identifier should be a field in the corresponding annotation (MeV .ann format file). For other formats this annotation should be found in the data file where each locus id is on the row corresponding to the associated spot information (annotation and expression data).
  (Please see the manual section 4, Loading Expression Data, for information on file loading for specific file format loading instructions).

  The *Coordinate Information* and *Chromosome Information* can likewise be imported with the annotation during the expression file load or a chromosomal coordinate file can be used to supply the coordinate and chromosomal information. This optional chromosome/coordinate file has a simple tab delimited format described below.

Optional Chromosomal Coordinate File

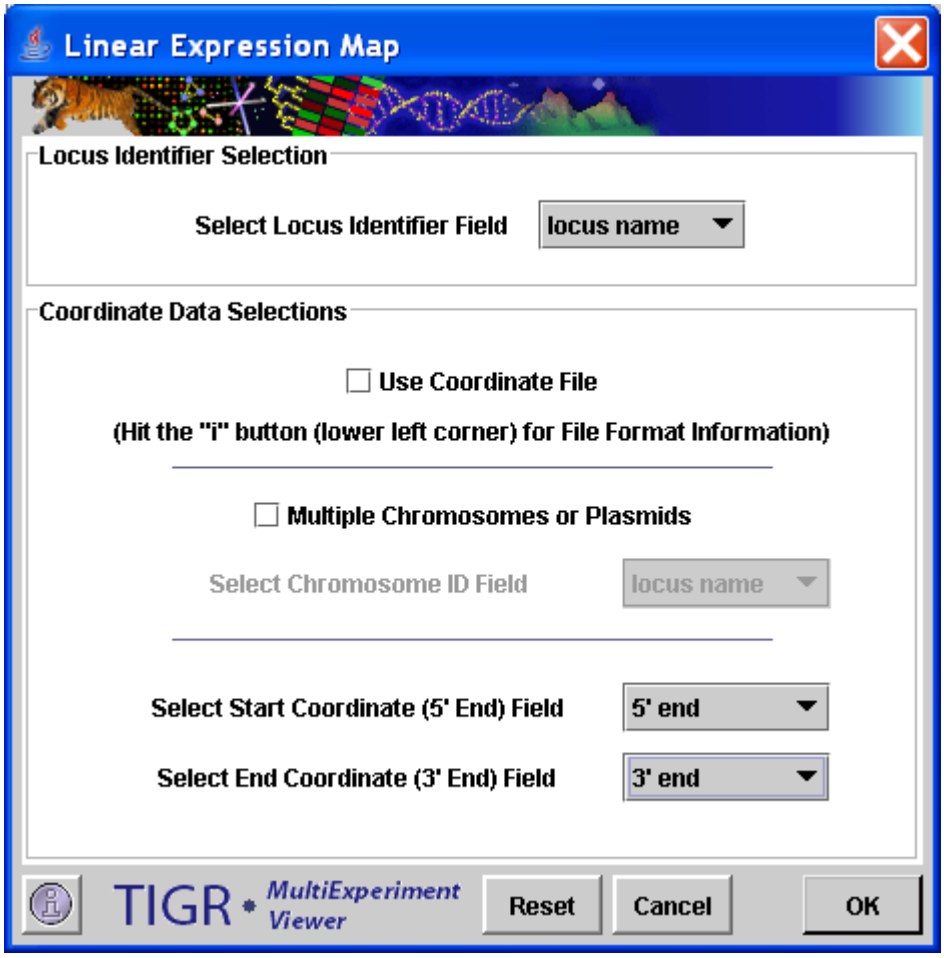
In the case where chromosomal information is not supplied during the initial MeV file load, an auxiliary coordinate file can be supplied for LEM construction. The format of the file is a text tab-delimited format with a row for each locus. The columns have the following order ([] indicates optional column):

<Locus ID> [Chr ID] <5' end> <3' end>

The file should not have a header. Spots in the loaded data set will be ignored if they map to loci for which location information is incomplete or missing.

The LEM Initialization Dialog

The initialization dialog collects information that indicates where the critical information for LEM construction resides and provides information about the nature of the data.



11.21.2 LEM Initialization Dialog

Locus Identifier Field

This option selects the annotation field that maps spots to loci.

Use Coordinate File

This option indicates if coordinate information will be loaded via an auxiliary coordinate file described above.

Multiple Chromosomes Option

This check box indicates if there are multiple chromosomes. If selected, MeV will expect chromosome identification information from the coordinate file or as a loaded annotation field in MeV. If the presence of multiple chromosomes or plasmids are indicated then there will be one map produced for each chromosome or plasmid.

The 5’ and 3’ annotation fields indicate which annotation fields in MeV identify the coordinate information if a coordinate file is not used.

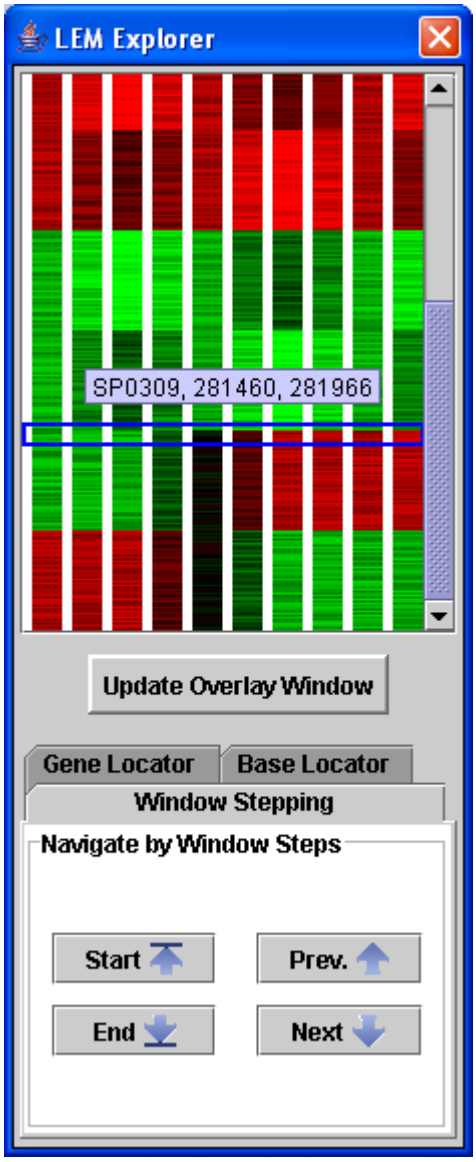
Basic LEM Features and Options

Several features to help navigate over the LEM, to customize the appearance, and to extract information from the LEM are available via a right-click context menu. The following is a list of general features that are described in more detail in later sections. Most of these features correspond to menu options except for Locus Information Panels which are launched with a left mouse click on a locus arrow.

- **LEM Navigation** – provides multiple options to systematically navigate over the map
- **Locus Information Panels** – provides detailed locus expression and annotation information, a link to web resources, and an option to navigate at the locus level
- **Customize Viewer** - sets locus arrow scaling options and viewer layout options
- **Color Scale Options** – sets the mode and constraints for coloring locus arrows to reflect expression
- **Locus Selection Manager** – provides support for the creation and visualization of lists of selected loci, options for list output, and methods for targeted loci selection
- **Store Cluster** – stores spots related to selected loci to MeV's cluster manager/repository
- **Save Selected Loci (Locus Detail)** – saves locus level information to a file
- **Save Selected Loci (Spot Detail)** - saves spot information for selected loci
- **Save All Loci** – saves the all loci expression and annotation information to file.

LEM Navigation

The LEM navigation controller is launched via the right-click menu option titled *LEM Navigation*. The Navigation Control provides several options for moving within the LEM. The upper section of the control is a reduced representation of the LEM where each locus is one pixel high. A blue rectangle indicates the area that is currently visible in MeV's main window.



11.21.3 LEM Navigation Control, with active tool tip (locus id and coordinate info)

Navigation Modes

Click to Location

In this option the main view jumps to a location in the LEM that corresponds to the location of a left mouse click in the Navigation Controller's LEM representation. The blue rectangle outline of the main viewer's range is updated to reflect the current viewable location in the main viewer. Aside from clicking on obvious expression features, the reduced size navigation screen also displays a tool tip to indicate the locus id and location information for the element under the mouse tip. This feature allows one to mouse over the controller in search of a particular locus or base location. The reduced size view can be a quick means to take a survey of the genome or chromosome in the LEM.

Window Stepping

This option allows you to step through the LEM systematically by advancing by one view-screen at a time. During window stepping the blue rectangle will update to indicate the LEM location that is visible in the main viewer. The buttons to control this are located at the bottom of the controller under a tab labeled *Window Stepping*. Shortcuts to the start and end of the LEM are additional options to quickly jump to those end-points.

Gene Locator

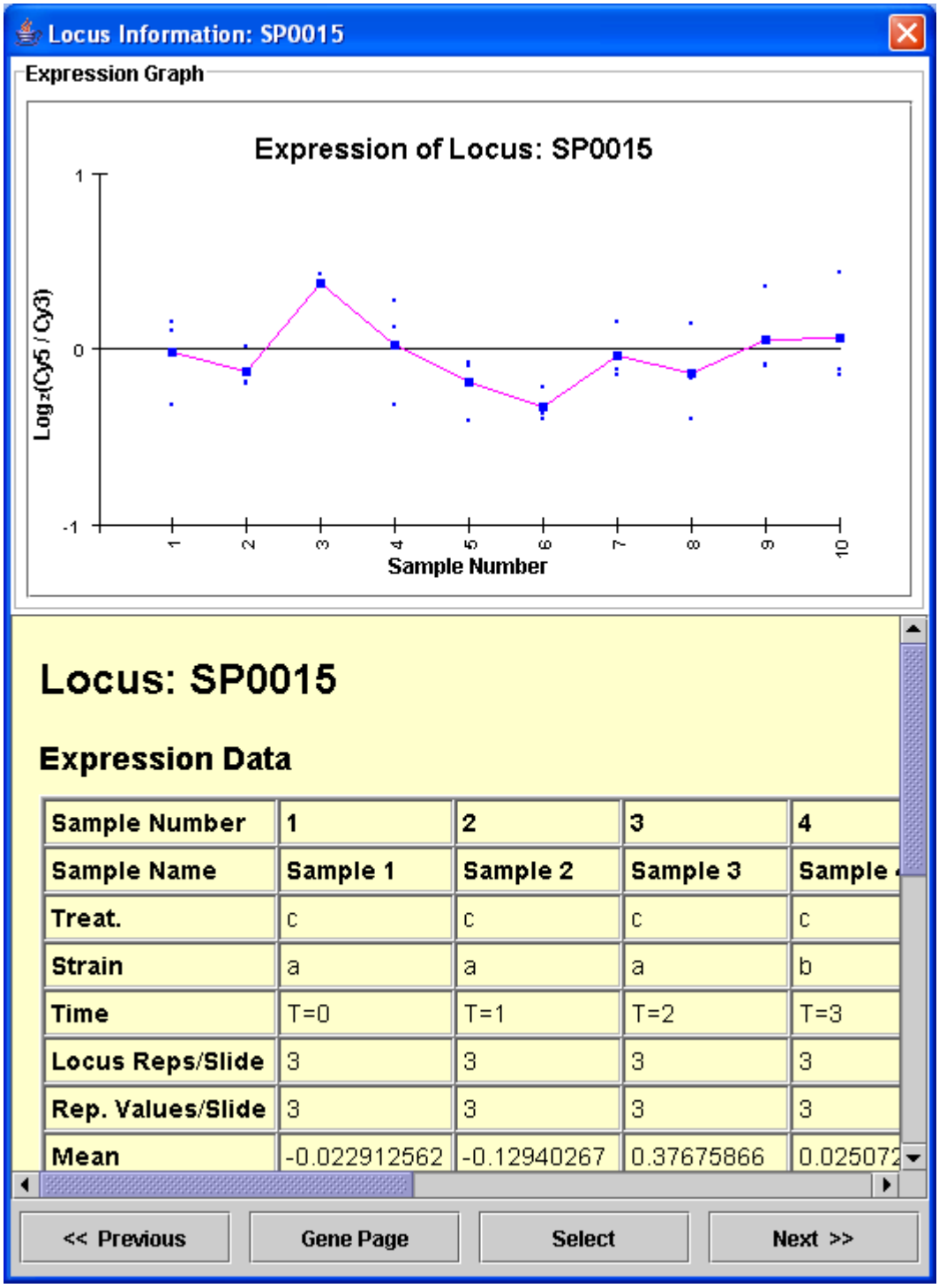
This option moves the LEM to a selected Locus. The controls for this are under a tab in the controller labeled *Gene Locator*. Provide a locus identifier and the LEM will jump to that location with the indicated locus at the top of the viewer. This feature is useful when one has specific loci of interest.

Base Locator

This option moves the LEM to a specified base location. The controls are found under the *Base Locator* tab in the navigation control window. Entering a base location will move the main view to the locus or loci that cover the entered base location.

Locus Information Panel

Locus information panels can be launched by a left mouse click on a locus arrow. The mean expression of a locus for each of the loaded samples is displayed in an expression graph. If there are several spots on the slide that correspond to the locus, each spot will be plotted as a point. The lower half of the panel has a table that contains expression information and another table that contains full annotation information for all spots associated with the locus.



11.21.4 Locus Information Panel

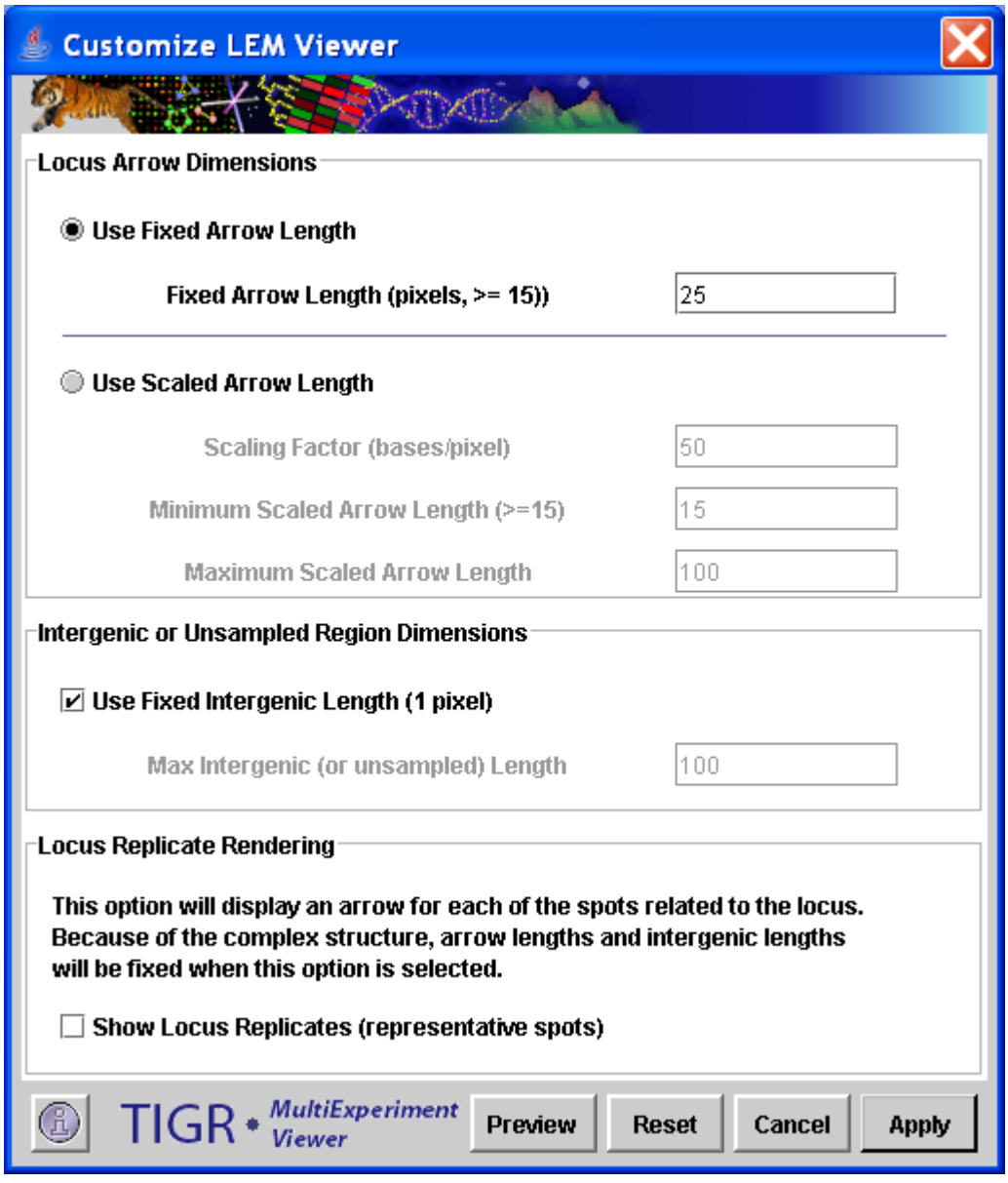
The locus information window contains four buttons at the bottom. The *Previous* and *Next* buttons update the information in the panel to correspond to the next or previous locus relative to the currently displayed locus. This feature allows one to advance one locus at a time in order to compare expression patterns between adjacent loci. This can allow the user to step through sections by viewing loci in order.

The *Select* button is used to push the locus onto the LEM's locus selection list. The locus selection features are detailed in a later section.

The *Gene Page* button launches a web resource relevant to the displayed locus. If the appropriate resource cannot be identified by MeV using the locus ID field name, a list of resources will be presented from which one can identify the proper resource.

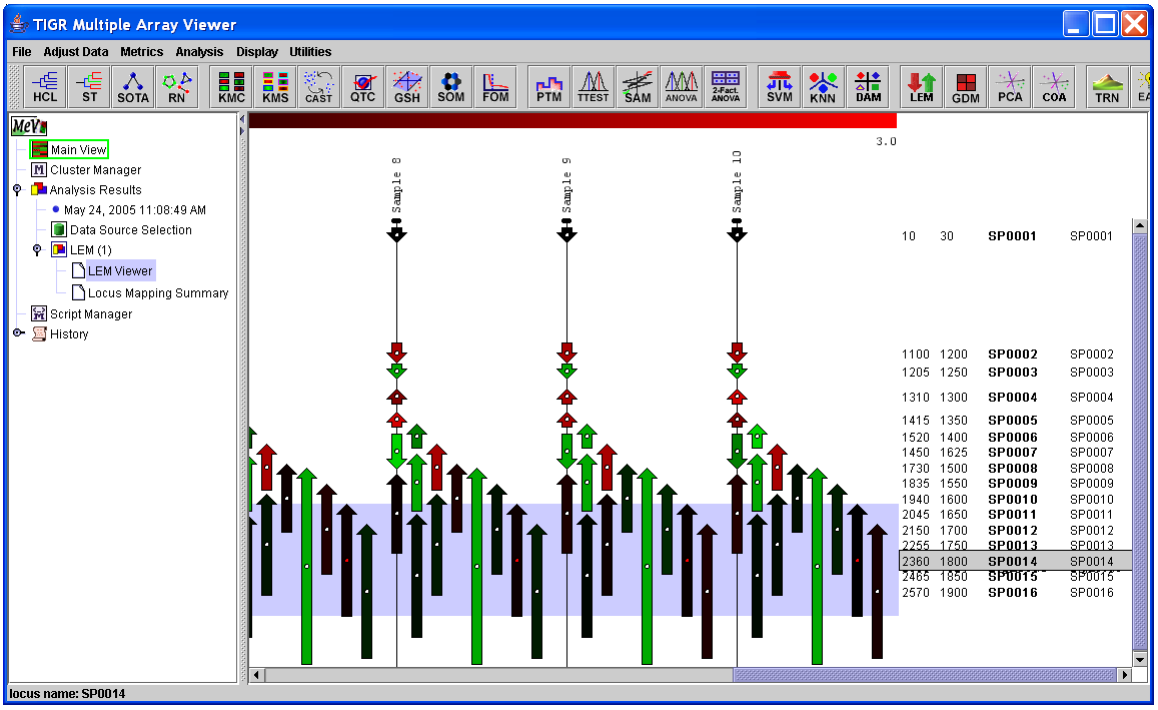
Customizing Viewer Appearance: Scaling the Viewer

The LEM permits one to scale loci based sequence length so that the arrows corresponding to loci reflect the length of the sequence. User defined constraints allow one to limit the length to fall within reasonable bounds. The scaling options are available via the right click context menu by selecting the 'Customize Viewer' menu item. A single dialog box helps to define the view.



11.21.5 LEM Customization Dialog

The main selection to make is whether the loci arrows should be of fixed length, in which case the user selects the fixed length in pixels, or if the arrow length should be scaled to reflect locus sequence length. If the loci arrows are scaled, a scaling factor can be selected to control resolution. Selecting fewer bases per pixel will elongate the viewer and will allow one to distinguish finer differences in length. The minimum arrow length is designed to render very small sequences with at least a small arrow so that associated annotation can be displayed. The maximum scaled arrow length allows one to constrain long sequences to a reasonable length. When using the scaled arrow lengths, loci that have overlapping coordinates are offset to the right so that arrows do not directly overlap.



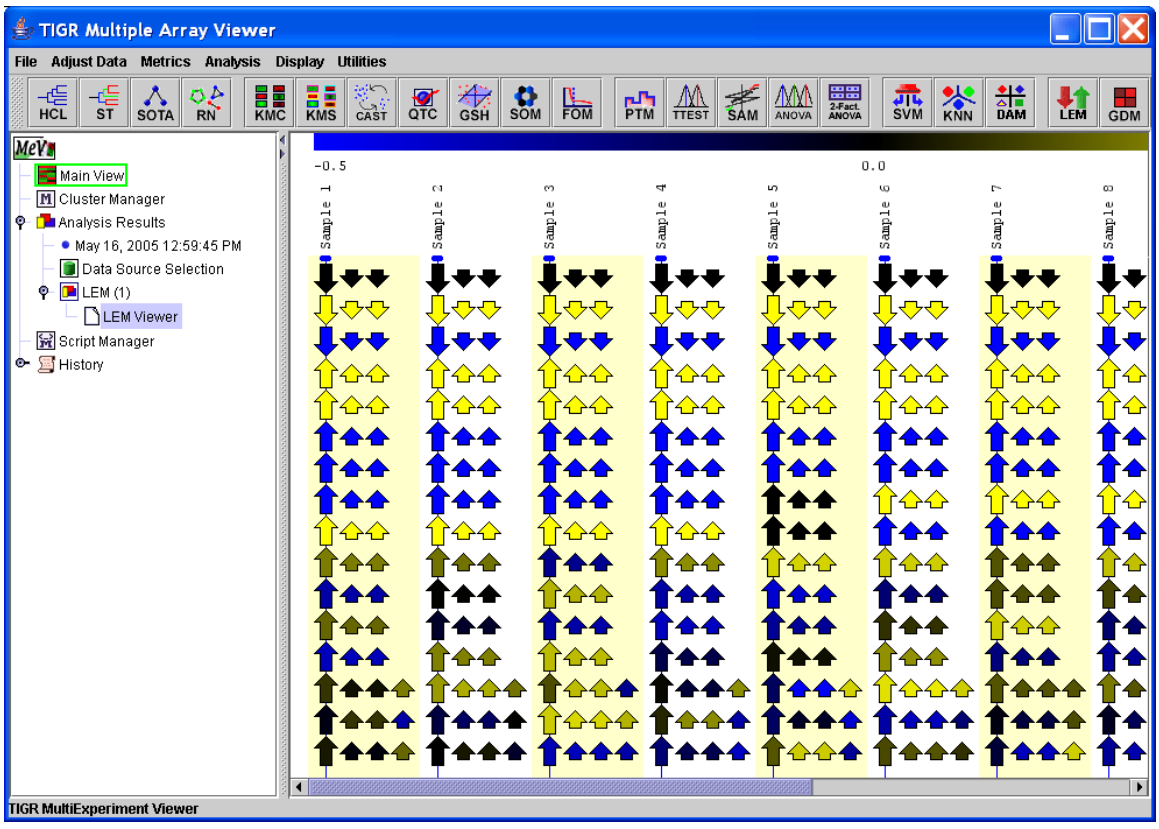
11.21.6 LEM with Offset Overlapping Loci Showing Highlighted Locus and Annotation

When in the scaled locus mode the arrows have an ‘anchor’ point to aid in locus selection. If you mouse over the anchor point, the locus will be highlighted and the anchor point will change from white to red. A shift-left mouse click will select the locus. Another special feature in this mode is that locus annotation that overlaps due to locus overlap will be rendered with the policy of highlighted locus annotation on top.

The un-sampled areas of the chromosome have a separate set of scaling controls. The base to pixel conversion factor is used if these areas are scaled. If these areas are scaled then a maximum length can be selected so that large areas without sampled loci are not extremely long. This prevents the occurrence of long areas without expression values.

Note that during interaction with this dialog one can hit the preview button to view the alterations to the LEM. If cancel is hit, the dialog box will be dismissed and the original settings will be restored to the LEM. The reset button will not dismiss the dialog but will return the values to the values of the LEM at the time the dialog was launched.

The last option on the Customization dialog is the option to expand the viewer to show an arrow for each *spot* related to a loci. If this is selected the arrow that represents the mean expression for a locus is displayed as usual but in addition arrows are displayed that correspond to each spot related to the locus. This displays the expression of the spots that contribute to the locus mean. The additional arrows for locus spots are displayed to the right of the locus mean arrow and are shortened slightly to help differentiate. The background of alternate samples is shaded lightly to help distinguish sample boundaries. When this option is selected, the arrow lengths and open areas are fixed to simplify the view so that spot level arrows are not confused with short loci arrows.



11.21.7 LEM with Locus Spot Replicates Shown (blue/yellow MeV color scheme)

Color Scale Options

The arrows in the LEM are given a color that indicates the level of expression. This color is taken from a scale based on the spot's expression value. Altering the color scale limits can help resolve (or ignore if appropriate) levels of expression

that are closer together. The LEM provides three modes to assign colors to expression values. These color assignment modes can be selected via the right click menu in a submenu labeled *Color Scale Options*.

Gradient Color Mode

The gradient option is commonly used and the value limits can be set using MeV's 'Display' menu. The Display menu also contains options to apply a different gradient color scheme. The Gradient limits and color scheme options are described in other manual sections related to the Display menu. These controls are placed in MeV's menu since they relate to all expression views in MeV. (Please see manual sections 6.4, Setting Color Scale Limits, and 6.3, Color Scheme Selection for information on setting the limits when using the gradient color mode).

Three Bin Mode

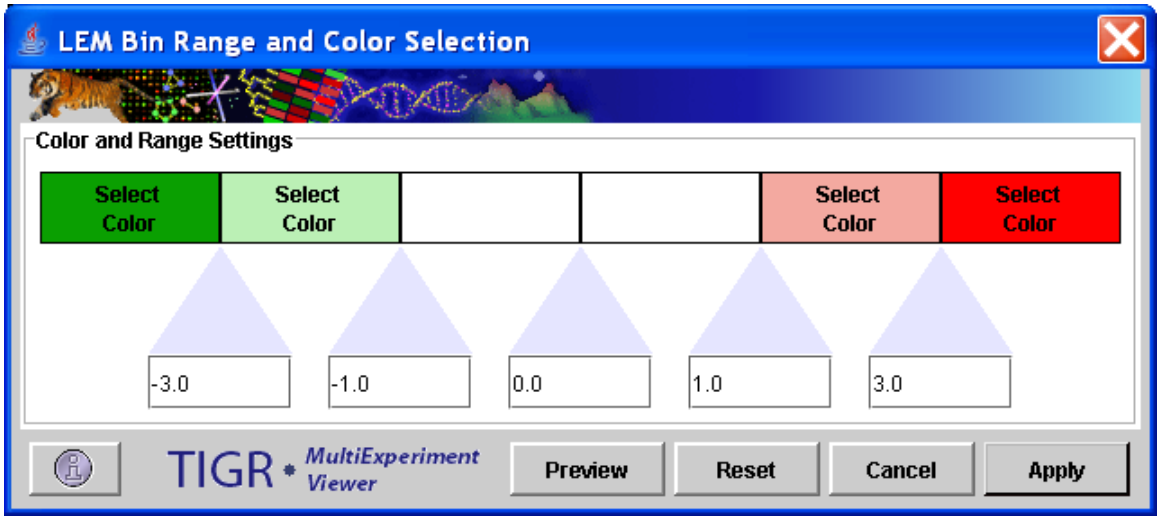
In the three bin color mode, colors are assigned from one of three expression bins. The user sets a high and a low limit on expression and values exceeding these limits are given a color that corresponds to the appropriate bin or left empty if the expression value falls within the set cutoffs.

Five Bin Mode

The five bin mode provides four cutoffs levels and while the LEM still colors the arrows with discrete colors, the two extra bins allow for an intermediate high and low expression bins.

Setting Limits and Colors for the Bin Modes

The discrete color bin modes have cutoff values and color selection options contained in a dialog that can be launched via a right click menu option labeled *Bin Colors and Limits*. The dialog has colored buttons that can be clicked to display a palette for color selection. Five cutoffs for the bins can be altered via text fields. In the case of the three bin mode only the outer most limits and colors are used. The values of the bins should increase from left to right. This convention is enforced to maintain valid limits when switching between 3 bin and 5 bin modes. Note that the preview button will apply the current settings to the LEM. To revert back to the original settings, use *Reset* and then *Preview* or *Apply*.



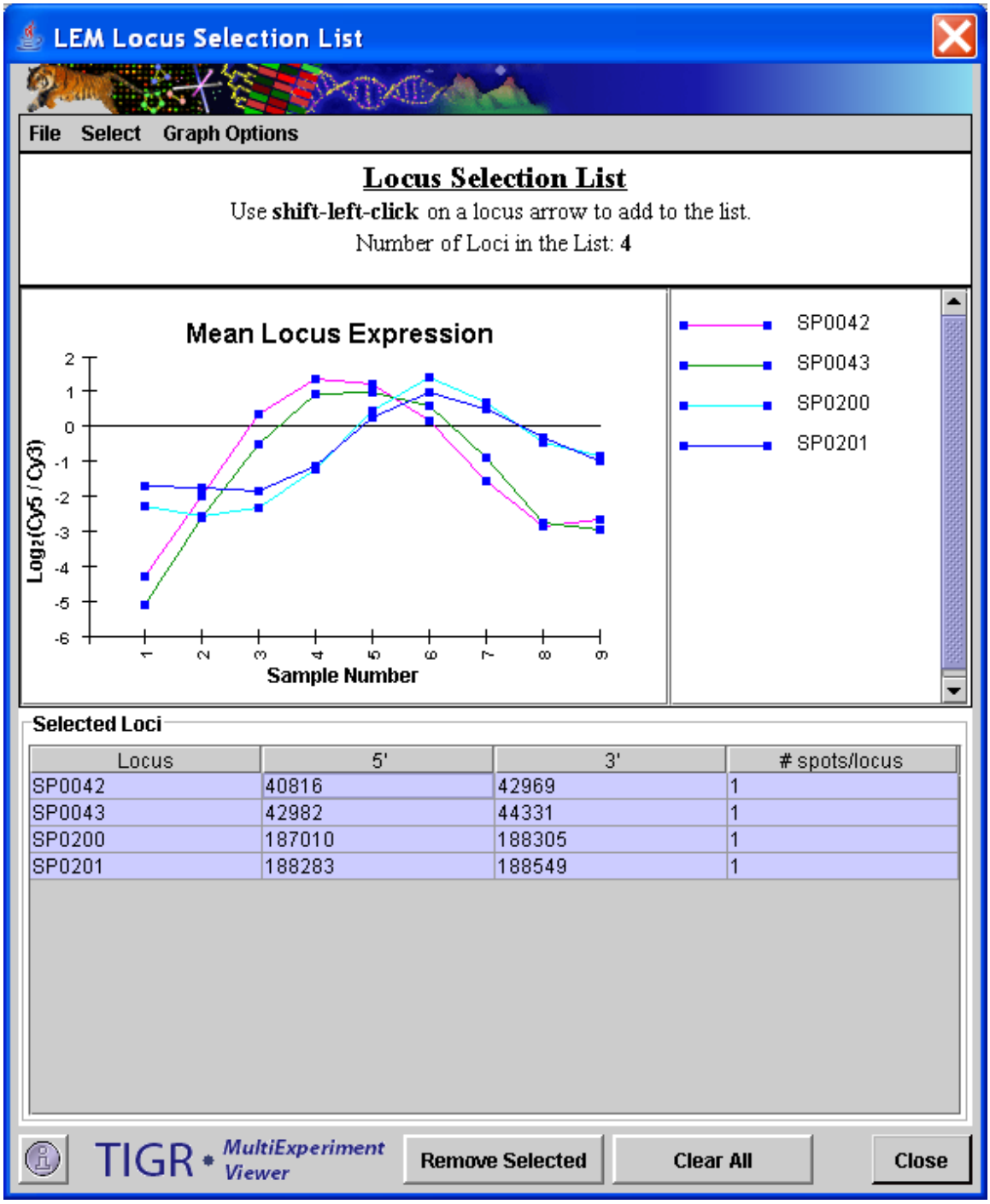
11.21.8 Bin Color and Limit Selection Dialog

Locus Selection List

The LEM viewer allows the user to select loci of interest. These loci can be output to file in various formats and can be stored as an MeV cluster in the cluster manager/repository. A *shift-left-click* on a locus will add the locus to the selection list. one can shift-left-click again on the locus to deselect the locus and remove it from the list. Selected loci will have a red marker to the left of the locus arrow and will have its annotation shaded in red.

The locus selection list can be viewed by selecting the Locus Selection List option from the right click menu in the LEM. The selection list has buttons that allow one to clear selected loci or to clear all loci from the list. The file menu on the selection list provides options to save all loci or selected loci to tab delimited text files that can be loaded into MeV or can be opened in a spreadsheet. These options are similar to options from the main LEM right click menu and will be discussed in detail. The *Select* menu has options to make targeted selections of loci by either selecting a locus or by specification of a base location range. The *Graph Options* menu has options to control locus expression graph rendering. One option is used to hide or show the locus graph while the other option is used to display the graph as either a simple monochrome graph or to render each locus graph as a different color as indicated by the key.

The locus selection graph updates to show the graphs for the loci that are selected in the table section of the window. Using a shift or control click in the table will permit the selection of multiple rows in the table and will overlay the expression graphs of all selected loci.



11.21.9 Locus Selection List (4 loci selected) with Visible Locus Graph and Key

Storing Selected Loci as Clusters

MeV's cluster manager contains many cluster set utilities and cluster operations such as unions and intersections. Loci that are of interest can be stored as a cluster in the manager by selecting the *Store Cluster* option from the LEM right click menu. As usual the user will be prompted to assign an optional label and description to the cluster and will be prompted for a color to associate with the elements of the cluster so they can be tracked during analysis. When storing a set of loci to a cluster, the elements of the cluster are actually the spots on the slide

that map to the set of loci. If multiple spots are associated with a locus in the selection list, each of these related spots are elements in the formed cluster. (Please see section 8 in the manual, Working with Clusters, to learn more about the options within the cluster manager.)

File Output Options

These options output selected loci or all loci to a text tab delimited file. In all cases the files are in the TDMS (tab delimited multiple sample) format and ready for import into MeV as independent data sets. The TDMS format is described in the appendix on file formats.

Save Selected Loci (Locus Detail)

This options saves locus level information to a TDMS file which contains locus identifier, chromosome (if more than one exists), and locus coordinate information. The expression value for each locus is a mean expression value for all spots related to the locus.

Save Selected Loci (Spot Detail)

This option saves spot values for spots related to selected loci. All annotation for the spots is output to the TDMS format file.

Save All Loci

This option provides the output described in the *Save Selected Loci (Locus Detail)* option described above except that it extends to include all identified loci in the viewer.

Additional LEM Output Viewers

The LEM module produces three additional viewers. 1.) Linear Expression Graph Viewers, 2.) Table of Unmapped Spots and, 3.) Locus Mapping Summary.

1.) Linear Expression Graph Viewer (LEG Viewer)

Linear Expression Graphs (LEGs) are also produced during LEM execution. These graphs depict gene expression as a graph where features are segregated by chromosome or plasmid and then ordered based on chromosomal location. The LEG Viewer node is appended to the result tree just below the node for the LEM viewer. A right click will produce

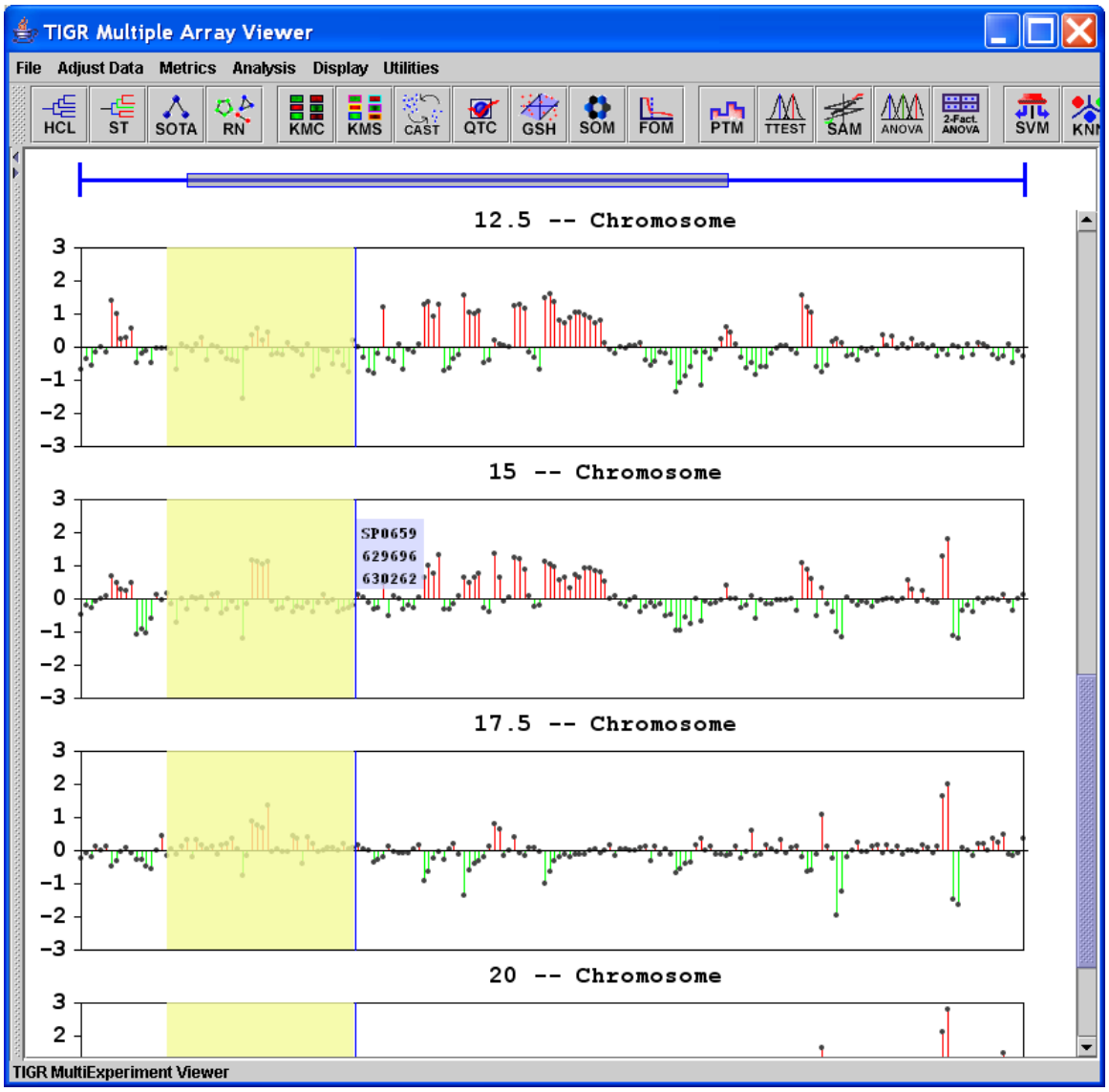
Viewer Options

A right click context menu allows several options for customizing the view. The following sections describe these basic features.

Viewer Modes

Tiled View

The default view for the LEG Viewer, Tiled View, produces a set of graphs, one for each sample that are stacked vertically. A click-drag of the mouse allows you to zoom in on a section of the graph. The header bar will update to indicate zoom level and the current location on the graph. The default color scheme produces 35 unique marker/line color combinations..

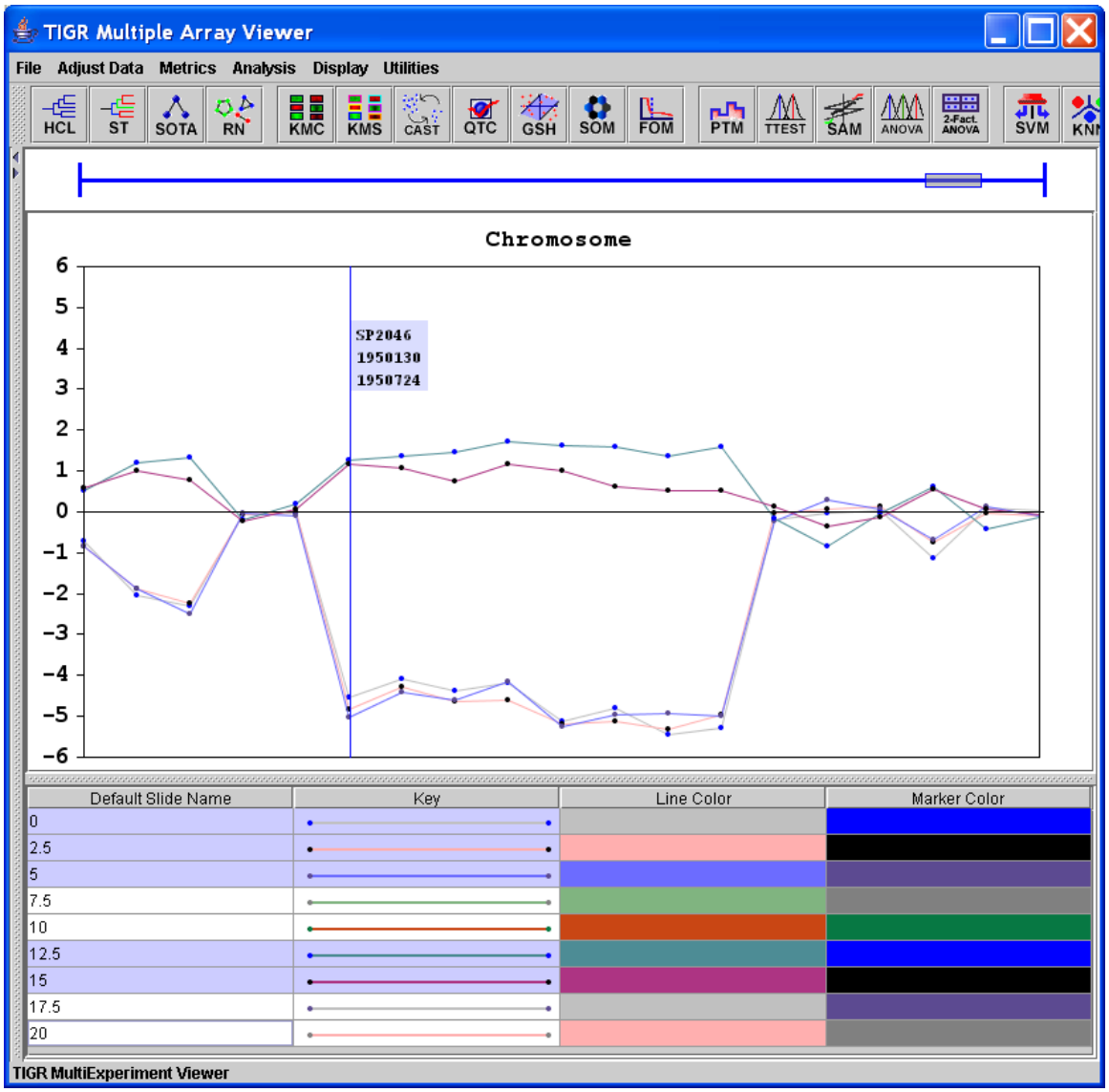


11.21.10Tiled View

Overlay View

The overlay view displays a single graph over which one or more samples can be graphed. The click-drag zoom option is still enabled in this mode. A ctrl-click in the table will permit multiple selections in the table. Clicking on the marker or line colors in the table will permit customization of the color representation for

that sample's graph. In practice, the overlay mode is most practical for viewing a few samples at a time and it is often useful to zoom in on particular regions of interest



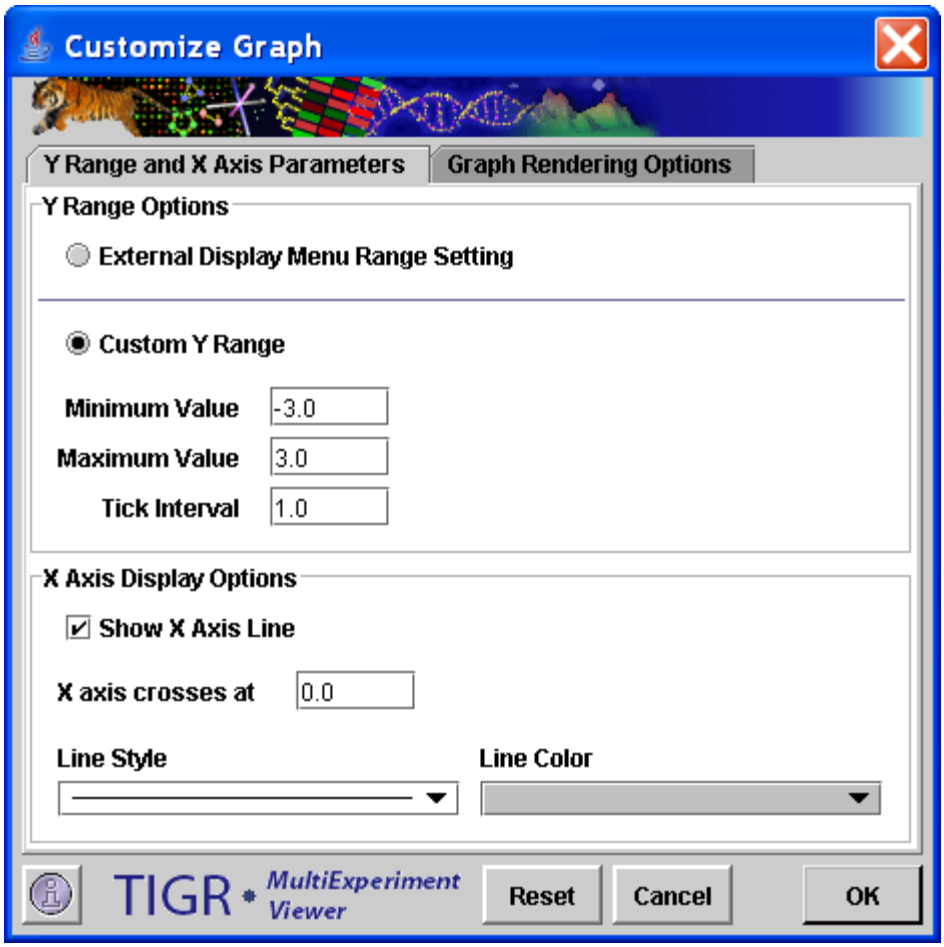
11.21.11 Overlay View

Graph Customization Options

Below viewer mode in the right-click menu there is a customize graph option which allows the y-axis range to be set as well as providing other graph rendering options.

Two options are enabled to permit y scaling. The initial default setting retrieves the range from the color scale range in MeV's *display* menu. A custom range option with tick interval provides a greater level of control.

An optional x axis line can be drawn at a specified y value. This can be a good reference line when dealing with log ratio data. Formatting options include selectable line style and color.

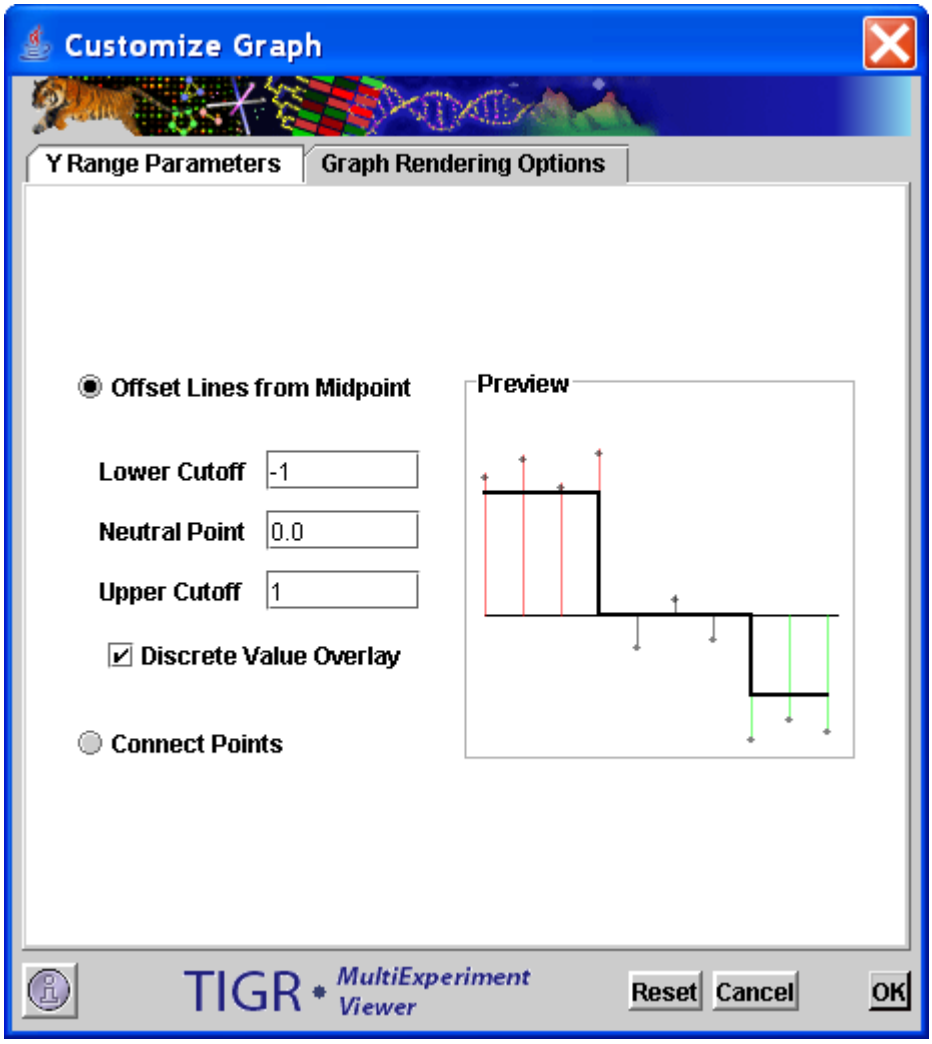


11.21.12 Graph Customization Dialog, Y axis range and X Axis Display options.

The second panel of the Customization Dialog includes two major graph rendering options. The *Offset Lines from Midpoint* mode draws points for expression values and connects those with a neutral line or baseline value. The cutoff values allow one to set limits which affect the color scheme of the offset lines. All points above the upper cutoff have red lines to the baseline, while points below the lower cutoff have green connecting lines. Points that fall within the cutoffs have black connecting lines.

The other option is to render the graph using lines to connect points. When viewing the graph in overlay mode this is the default setting so that the various samples can be easily identified.

The last rendering option is to use a *discrete value overlay*. This option uses the supplied cutoff values and overlays the graph with a square wave pattern where points exceeding the limits force the line through the appropriate limit. This overlay, shown in the preview mode in the dialog figure, can help to visually identify regions of contiguous, at least in ordering, genes that show similar behavior.



11.21.13 Graph Customization Dialog, Rendering Options

The final two menu options in the Graph Viewer include an option to display a locus reference line that provides gene identifier and location information as you mouse over the graph and an option to reset the x range to zoom out to show all data in one window.

2.) Table of Unmapped Spots

In the event that there are control spots on the slide or spots for which the loci and coordinate information is not known, a table is created to collect and identify these spots. This table lists all annotation for unmapped spots and can be reviewed, searched, or saved to confirm that the listed spots are indeed unmapped and cannot contribute to the LEM. The table has the properties and options of MeV's cluster table viewers that are described in section 7.5 of the manual.

3.) LEM Mapping Summary Viewer

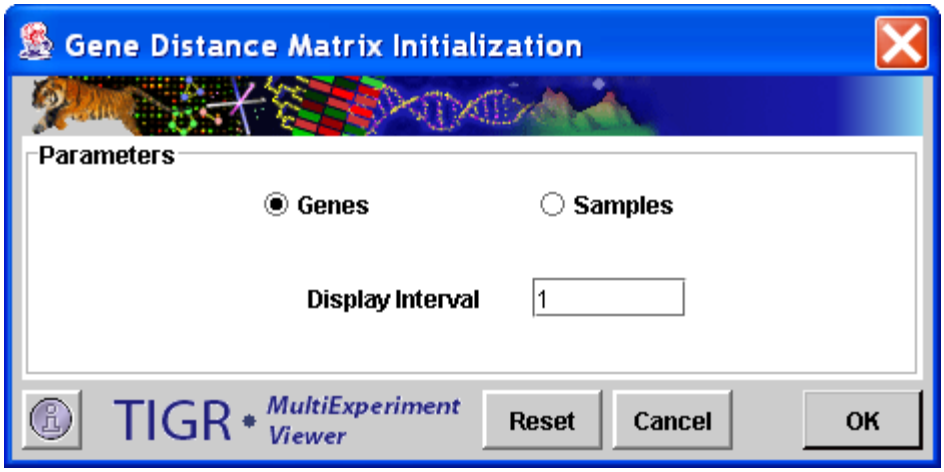
The mapping summary is placed on MeV's result navigation tree below the unmapped spot table, if present, and the LEM viewer(s). This viewer list information about the number of spots that are entering LEM, the number of spots that map to loci, the fraction of spots that are mapped, and lists all of the parameters selected when producing the LEM. If the LEM doesn't appear to render loci appropriately, check the parameter listing to confirm that the proper coordinate information was supplied and other parameters such as the indication of multiple chromosomes or plasmids was indicated correctly. In addition to global mapping data, the number of mapped loci and the number of spots that correspond to each chromosome or plasmid are listed.

11.22 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

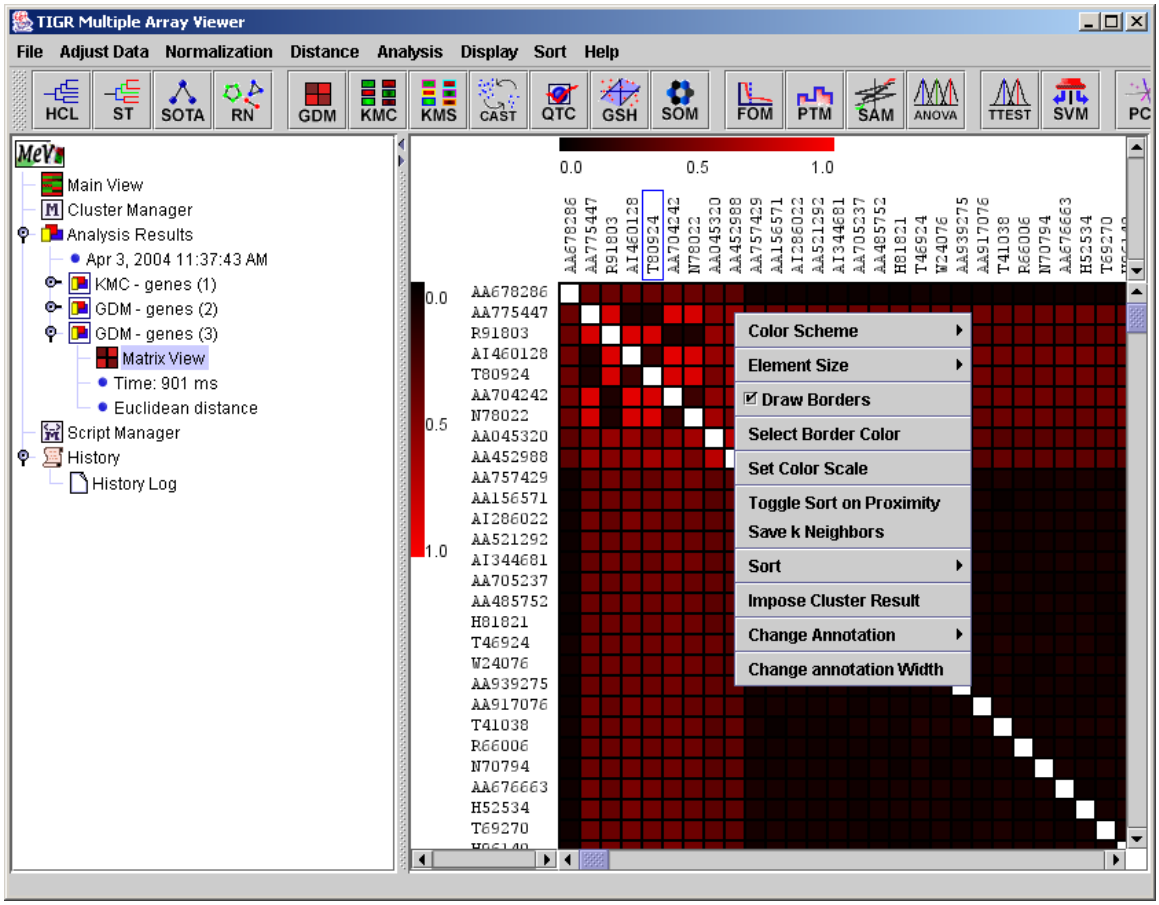
The GDM module can be started by using the GDM button or by selecting the GDM menu item from the analysis menu. When creating a gene matrix it is possible to display a subset of the full data set. The creation of an nxn matrix is expensive from a computer memory standpoint and by using the "Display Interval"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



11.22.1 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.



11.22.2 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

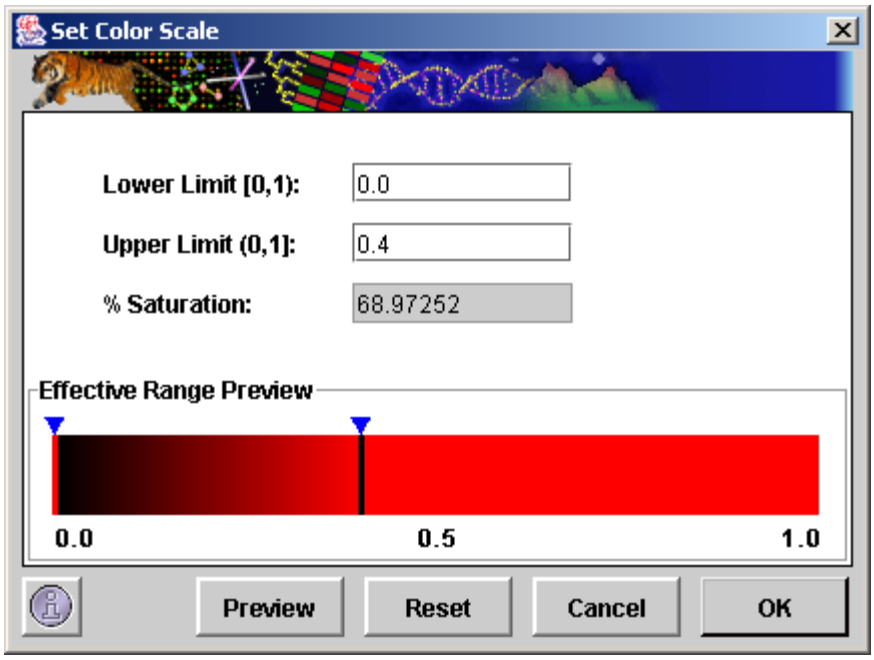
Five preset element sizes are offered as well as the option to select a custom size. Changing the size is a good way to get either a detailed look at specific genes or to take a broad survey of the matrix.

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.



11.22.3 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

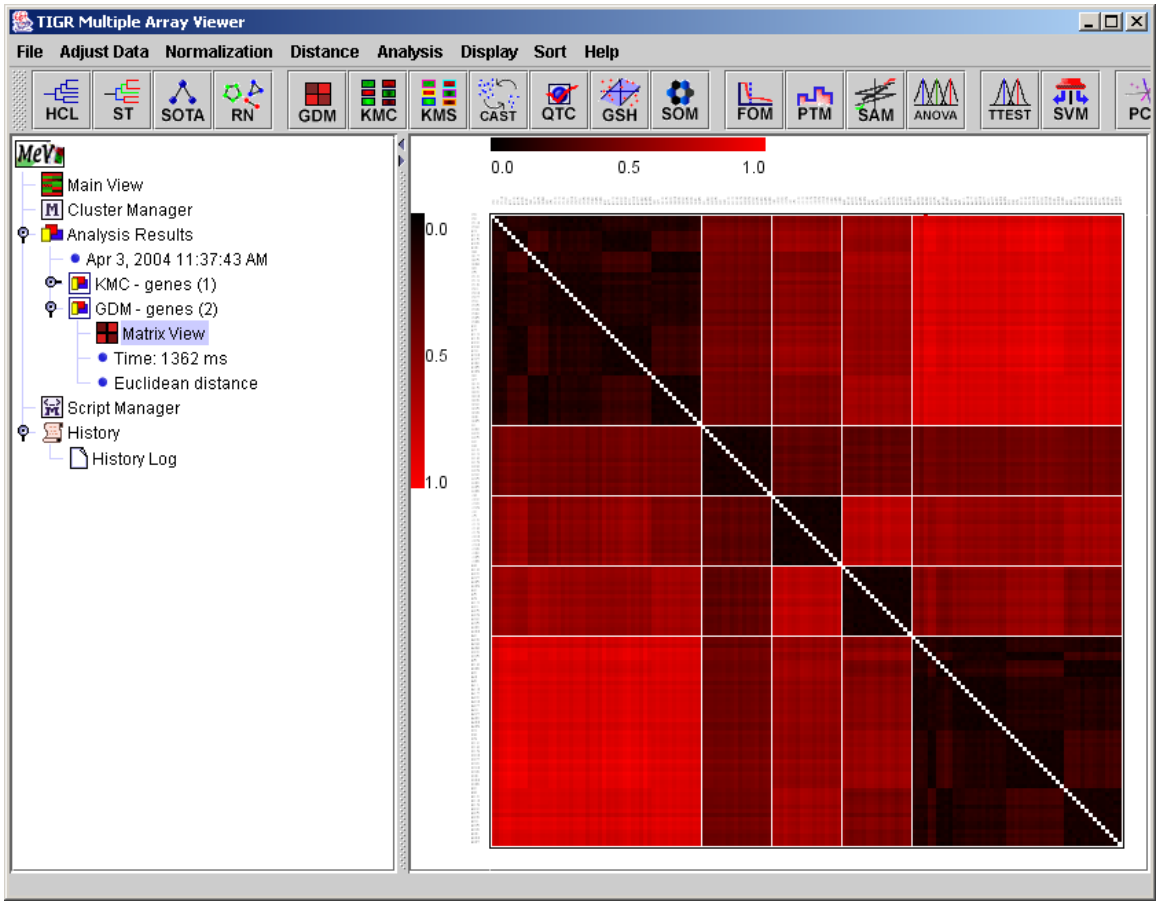
The Impose Cluster Result option is among the most useful features of the GDM. When selected the current MeV session is surveyed for appropriate clustering results to apply to the GDM. Application of the clustering result reorders the genes such that the rows and columns are grouped by cluster membership. 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

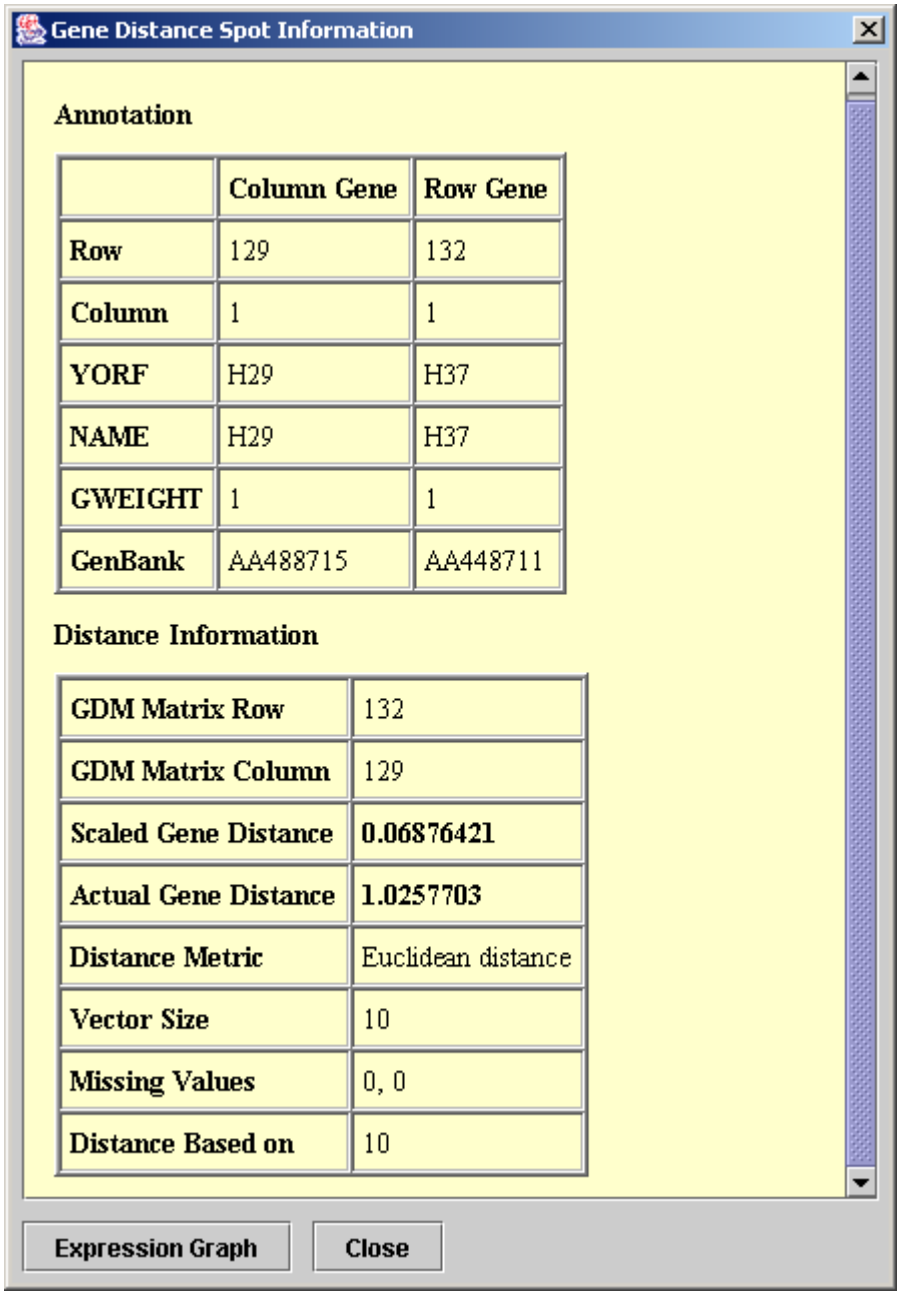
This option allows the expansion or contraction of the header to permit viewing of the header without excessive scrolling or if contracted more of the matrix will be visible.



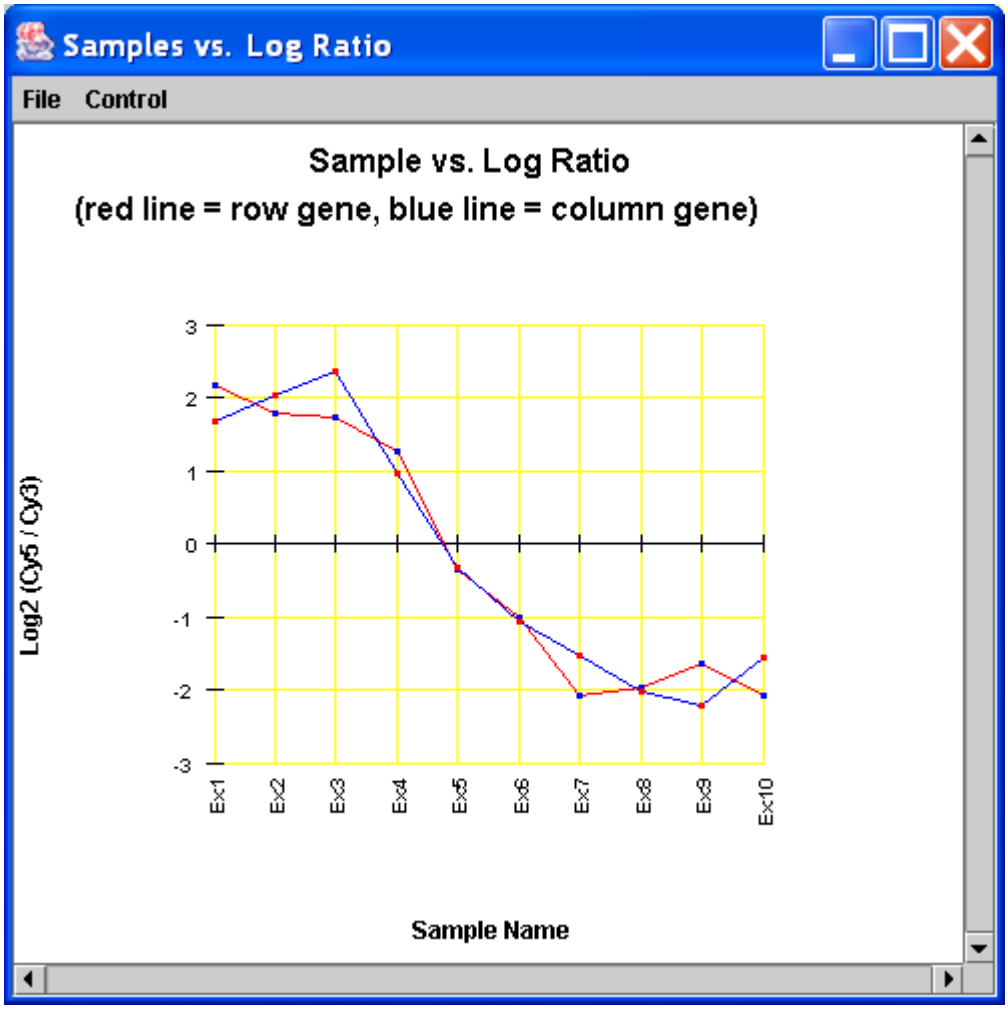
11.22.4 GDM matrix with a KMC clustering result imposed (k = 5)

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.



11.22.5 GDM Element Information Record



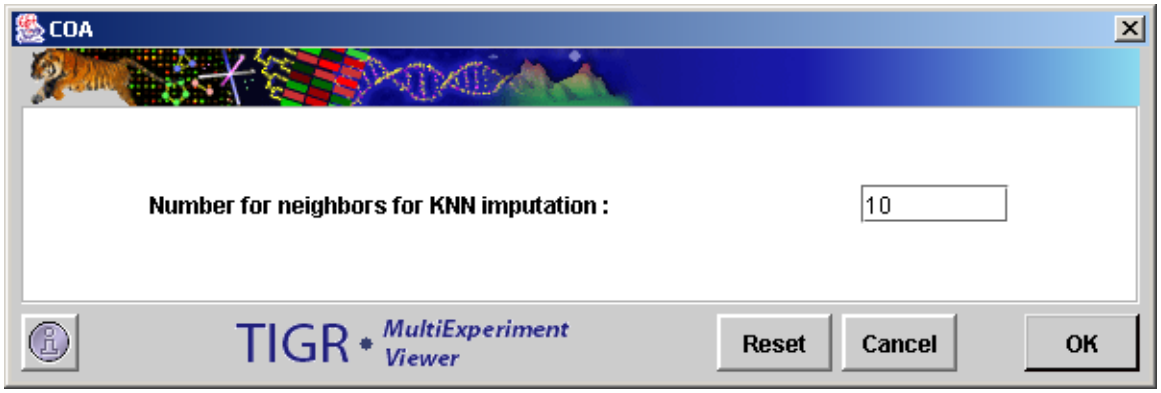
11.22.6 GDM Element Information Expression Graph

11.23 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.

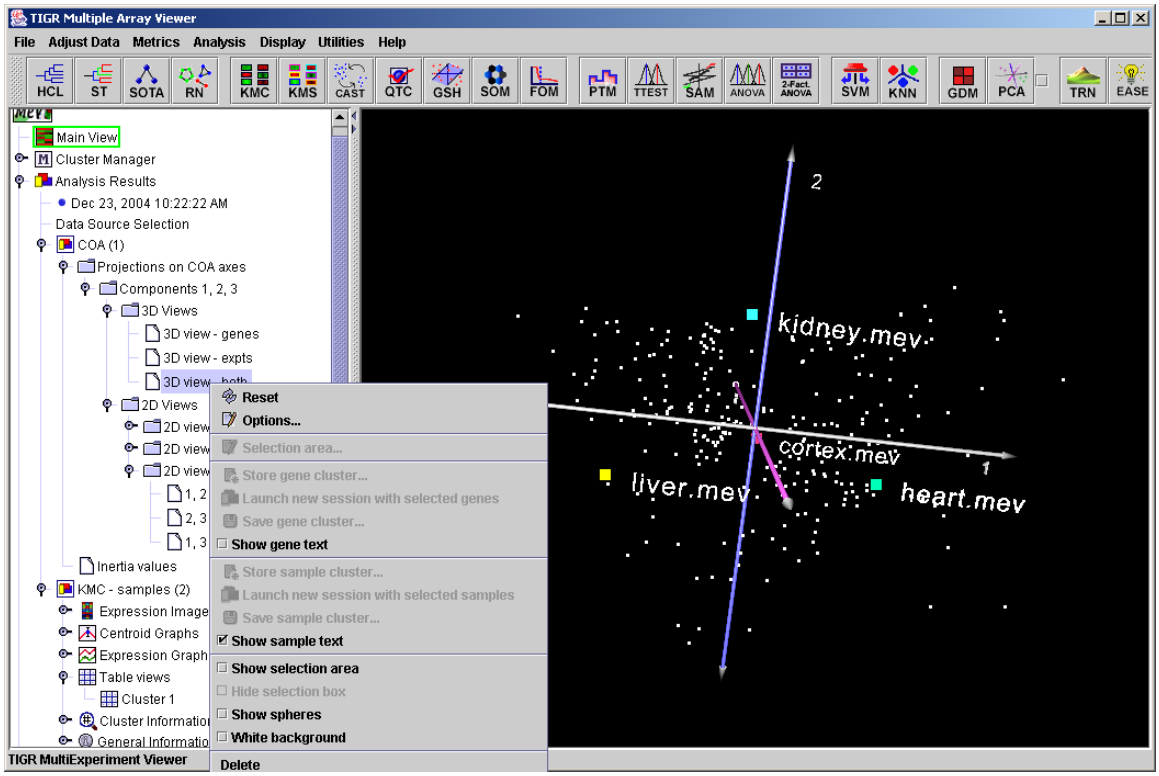


11.23.1 COA initialization Dialog

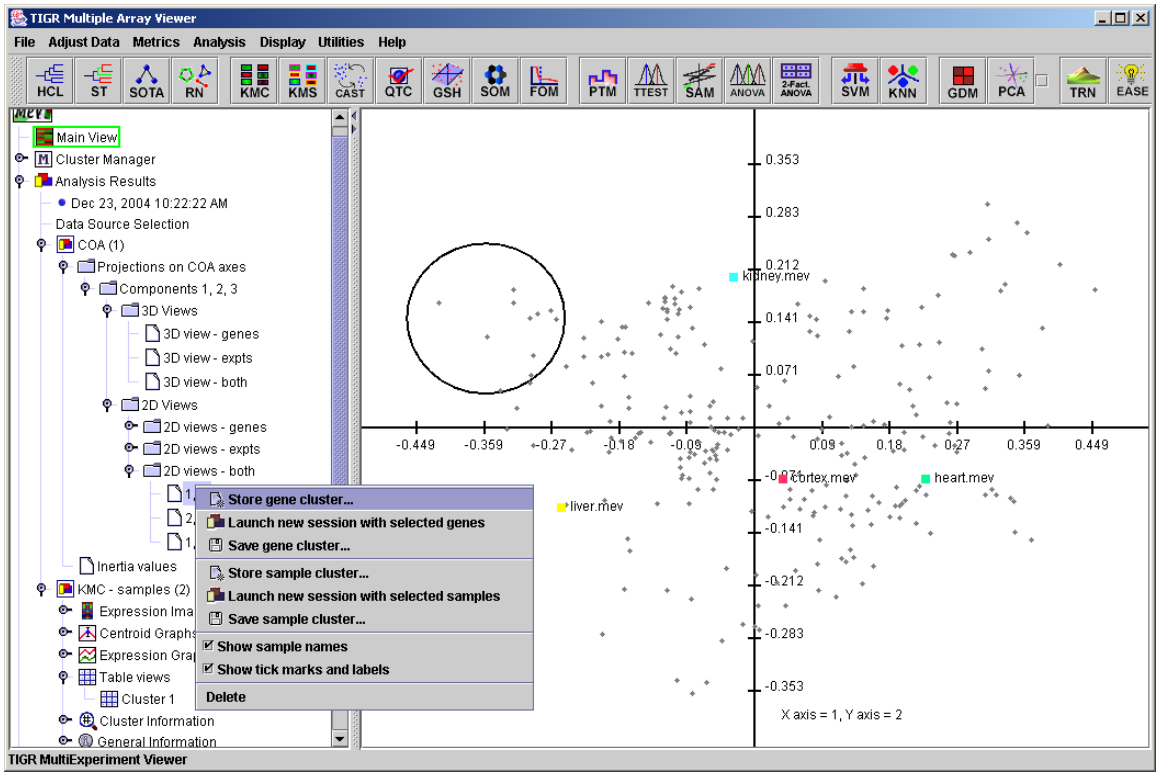
The displays in this module are very similar to the PCA displays, except that 2D and 3D plots are shown for genes and experiments separately as well as together on the same plot. Menus for creating new plots, selecting data points and customizing displays are available from the corresponding nodes on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



11.23.2 COA: 3D display



11.23.3 COA: 2D display

11.24 PCA: Principal Components Analysis

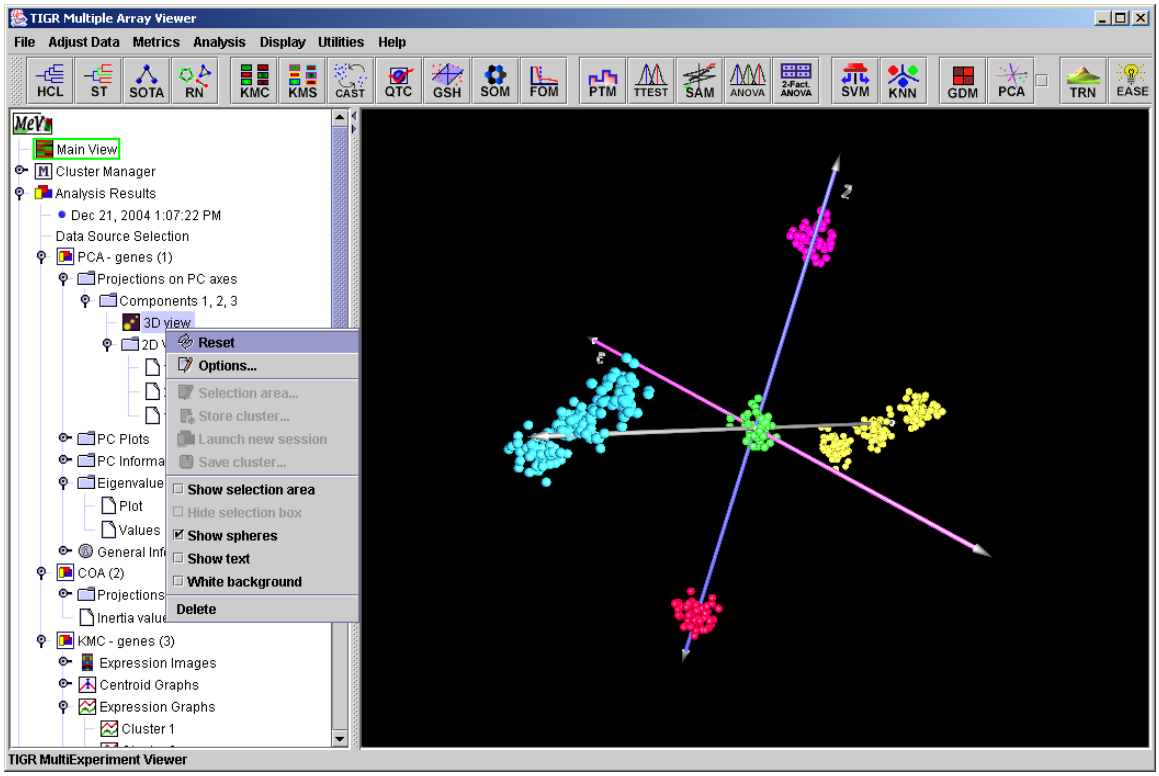
(Raychaudhuri *et al*. 2000)

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.

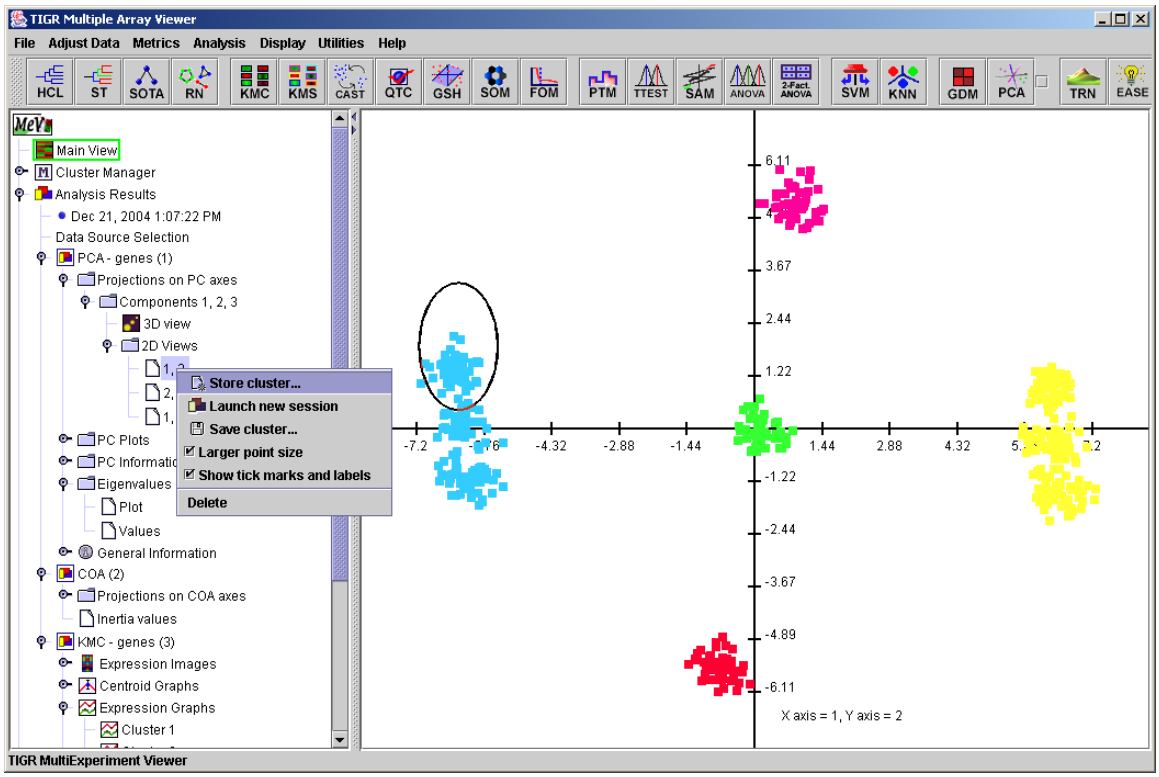
Once the calculations are complete, select the PCA node under Analysis to view the PCA results. Under the node called "Projections on PC Axes" are the default plotting of components 1, 2 and 3. Right-clicking on this node will allow other components to be chosen for plotting. These new plots will show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



11.24.1 PCA: 3D View.



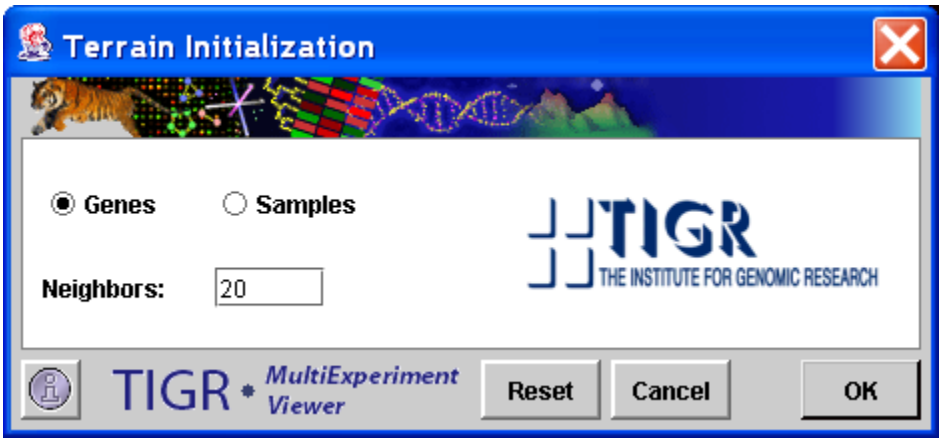
11.24.2 PCA: 2D view

11.25 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.

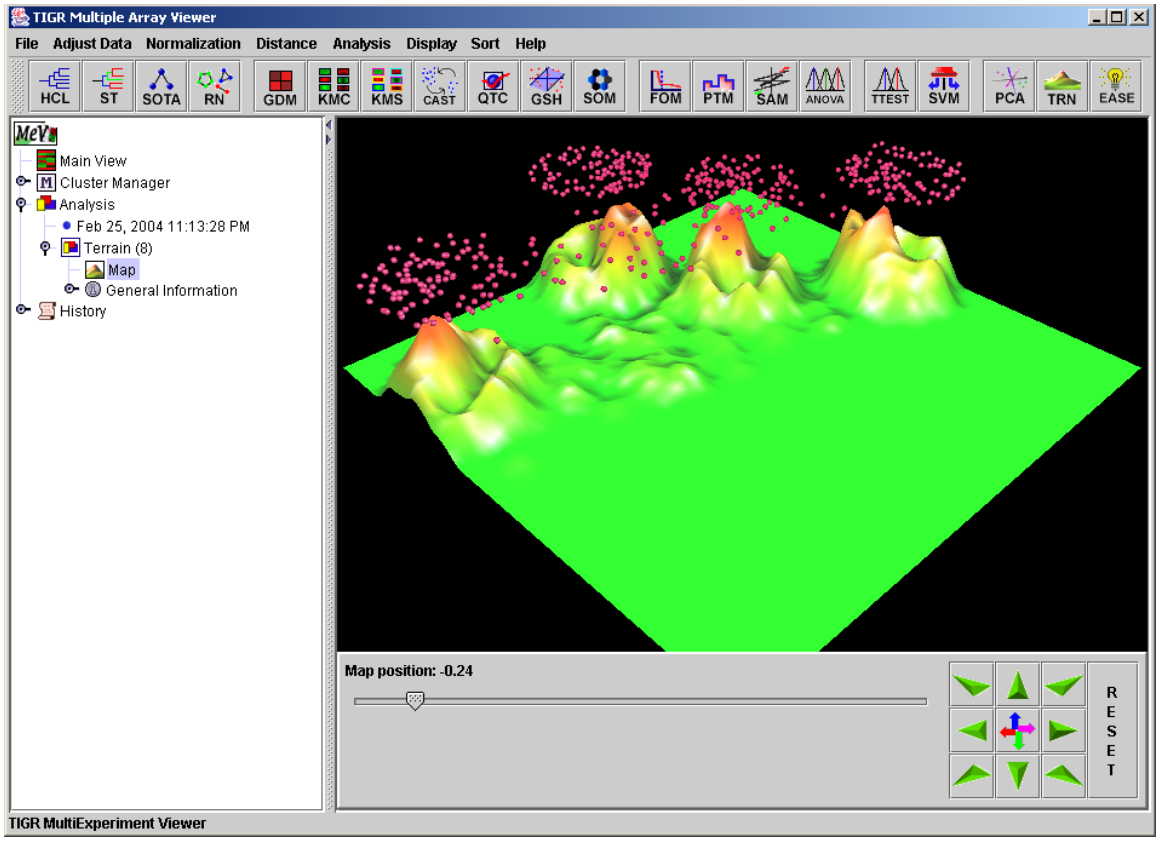
It is very important to note that like many algorithms TRN uses a default distance metric that greatly affects the outcome of the terrain created. The default metric for TRN is Pearson Squared, which groups elements based on correlation and tends to place strongly correlated and strongly anti-correlated 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.



11.25.1 Terrain Map Initialization Dialog.

Data Element Identification

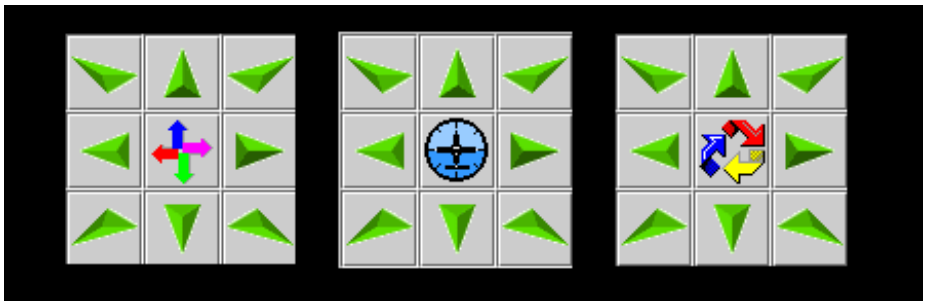
Data element identification can be achieved by four major methods. The right click menu allows one to turn on labels. The utility of the labels depends on how close the elements are to one another. It is also possible to move the mouse cursor over an element and reveal the annotation label currently 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.



11.25.2 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.


11.25.3 Terrain Map Navigational Controls, Various Navigation Modes

Linear Axis Mode

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

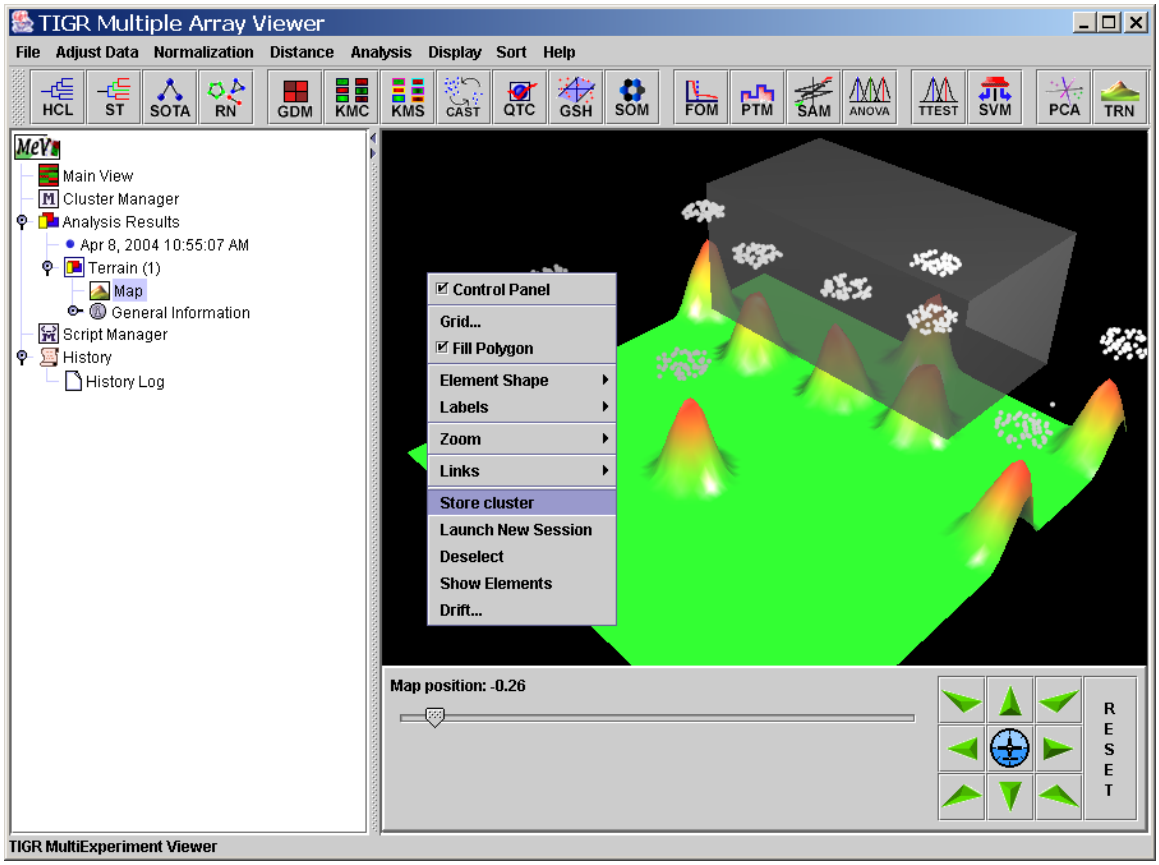
move the pov directly toward the terrain view while the lower right and left arrows move the pov directly away from the terrain. The left and right arrows move the pov left and right relative to the terrain.

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.



11.25.4 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

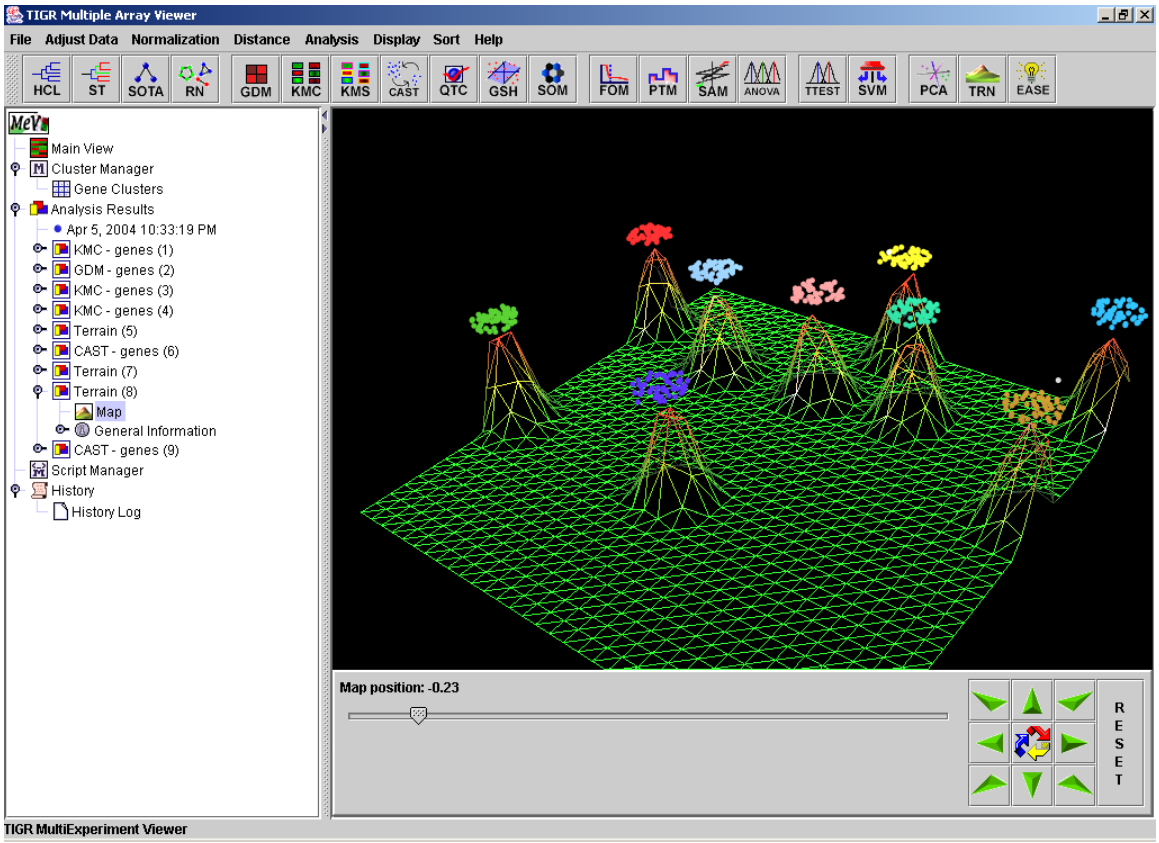
This feature launches a new Multiple Experiment Viewer containing the selected elements. This is perhaps the fastest and most convenient means of viewing the selected elements. An alternative to this is the Show Selection option described below.

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.



11.25.5 Terrain Viewer Without Surface Rendering

11.26 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.

Using EASE

There are two main ways in which users can operate the MeV implementation of EASE annotation analysis on a given gene cluster: a default run and a custom run.

Default Run

This mode is ideal for quick runs. Among the few parameters that need to be specified in the EASE Annotation Analysis dialog (fig 11.26.1) is the data files containing annotation information that match the desired species and array under study, the mode of operation (Cluster or Slide Annotation Analysis) and which subset is to be used from the available gene population. All other data necessary to run EASE such as gene population selection, statistical parameters, and gene ontology files for biological processes, biological roles, cell location and biochemical pathways will be obtained automatically. The following section on "Support Directory / Annotation Parameter Selection" explains in detail the steps to follow.

Custom Run

This mode is designed for tailored runs and can be used by clicking the "Custom" button at the top of the EASE Annotation Analysis dialog (fig 11.26.1). Among the several parameters that can be specified by customizing EASE is the location of the necessary data files containing gene indices for loaded expression data, the selection of the background population against which cluster results will be

compared, annotation keys used to uniquely identify genes, gene ontology files linking indices to biological themes and conversion files in case that the indices used by MeV are different than those in current linking files. Clicking on the "Custom" button on the upper right of the EASE Annotation Analysis window will bring up the EASE Advanced Parameters dialog (fig11.26.2), which will allow the user to adjust default parameters to specific needs. To obtain more instructions refer to the "Advanced Parameters Dialog" section.

As in the case of the default run, if not specified otherwise, advanced statistical parameter for reported statistics, trim parameters and multiplicity corrections remain set to default values and will be used automatically.

EASE Input Parameters

Support Directory/ Annotation Parameter Selection

Selecting an EASE File System

There are two methods of providing EASE with the data files it needs to do its analysis. The first is the simplest, standard method. At the top of the EASE Annotation Analysis dialog (fig 11.26.1) is a panel with drop down menus where you can select the species and array name that corresponds to your loaded data. If you used the automatic annotation loader to get your annotation, these values should already be selected. The button to the right of the species selector will indicate whether MeV already has the data for this array type downloaded. If the button says "Select", then the required data files for the displayed array type are already downloaded to MeV's repository. Click "Select" to select these data files for use. If the button text reads "Download", then MeV has not yet downloaded and stored the required files locally. Click the Download button to begin downloading.

If your array is not listed in the drop-down menus then you will need to manually provide the data files, as described above, and choose the location of those files. To do so, click on the "Custom" button to bring up the EASE Advanced Parameters dialog (fig11.26.2), next click on the "Browse" button in the Directory Selection for Support Files panel. Do not use the Browse button if you have already selected an array with the drop-down menus and the Download/Select button.

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.

Cluster Graph and Cluster Selection

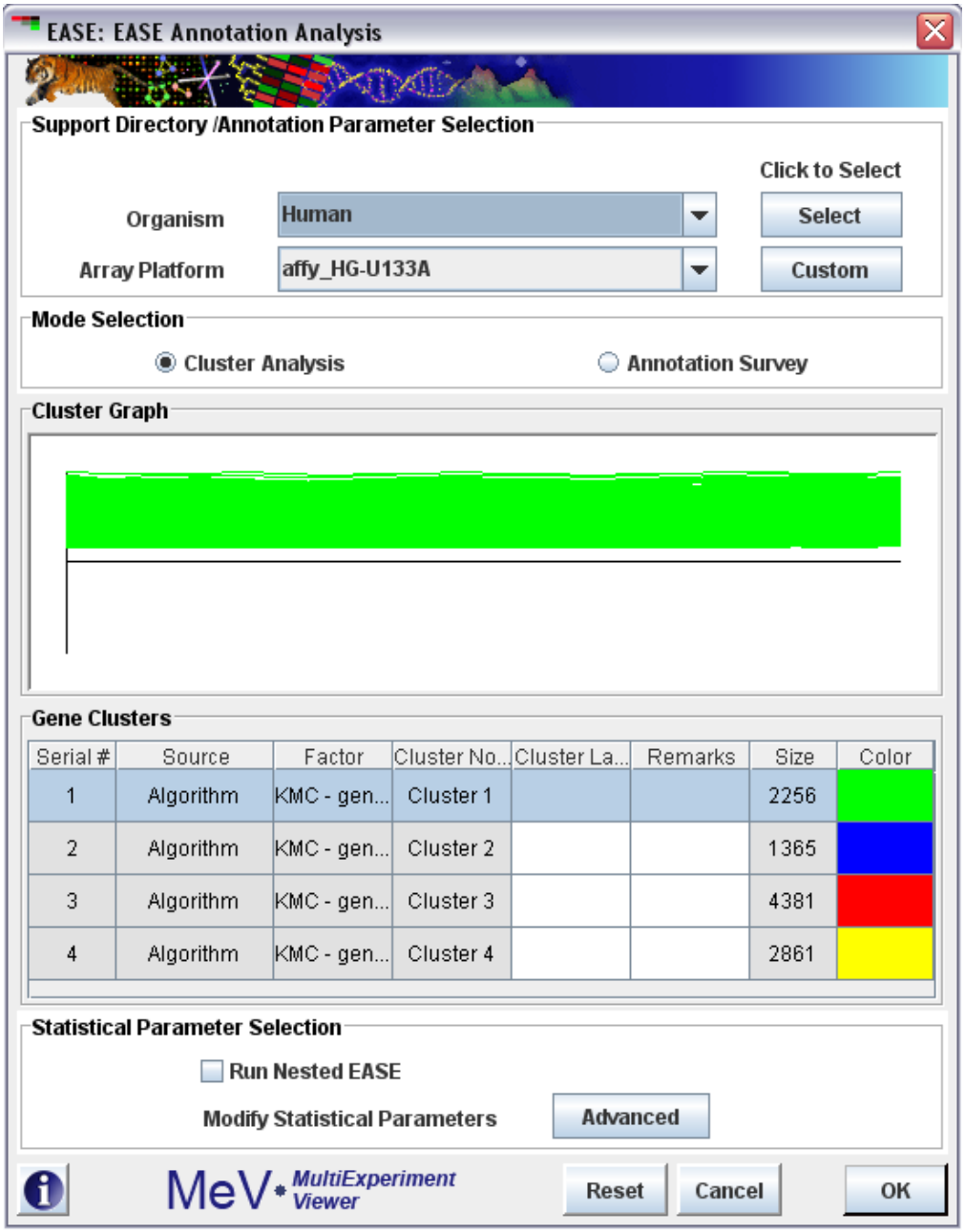
These two panels display gene clusters currently stored in MeV's cluster repository. If no clusters have been saved then a blank browser page will be displayed on this page and the Cluster Analysis mode option will be disabled. Selecting a row in the cluster table will display the cluster in the expression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

Statistical Parameter Selection

Nested EASE

Check *Run Nested EASE* if you would like to run the Nested EASE algorithm after the EASE run is complete. The nEASE algorithm executes a second, sub-level, iterative Fisher's Exact Test on significantly enriched GO terms identified in a first-level EASE analysis. This sub-classification approach provides increased sensitivity for detecting enriched GO terms and thus affords a deeper understanding of possible mechanisms underlying a given condition under study. nEASE was added to MeV as a new feature for version 4.5.

To further specify reported statistical and trimming parameters, click on the "Advanced" button, which will bring up the Statistical Parameters Dialog (fig11.26.3).
.



11.26.1 EASE Annotation Analysis Dialog

Advanced Parameters Dialog

Directory Selection for Support File
This pane allows the user to manually provide an EASE file system from a local directory by clicking on the “Browse” button.

Analysis Population Selection
This panel provides options to specify a gene population list and has two options for list selection.

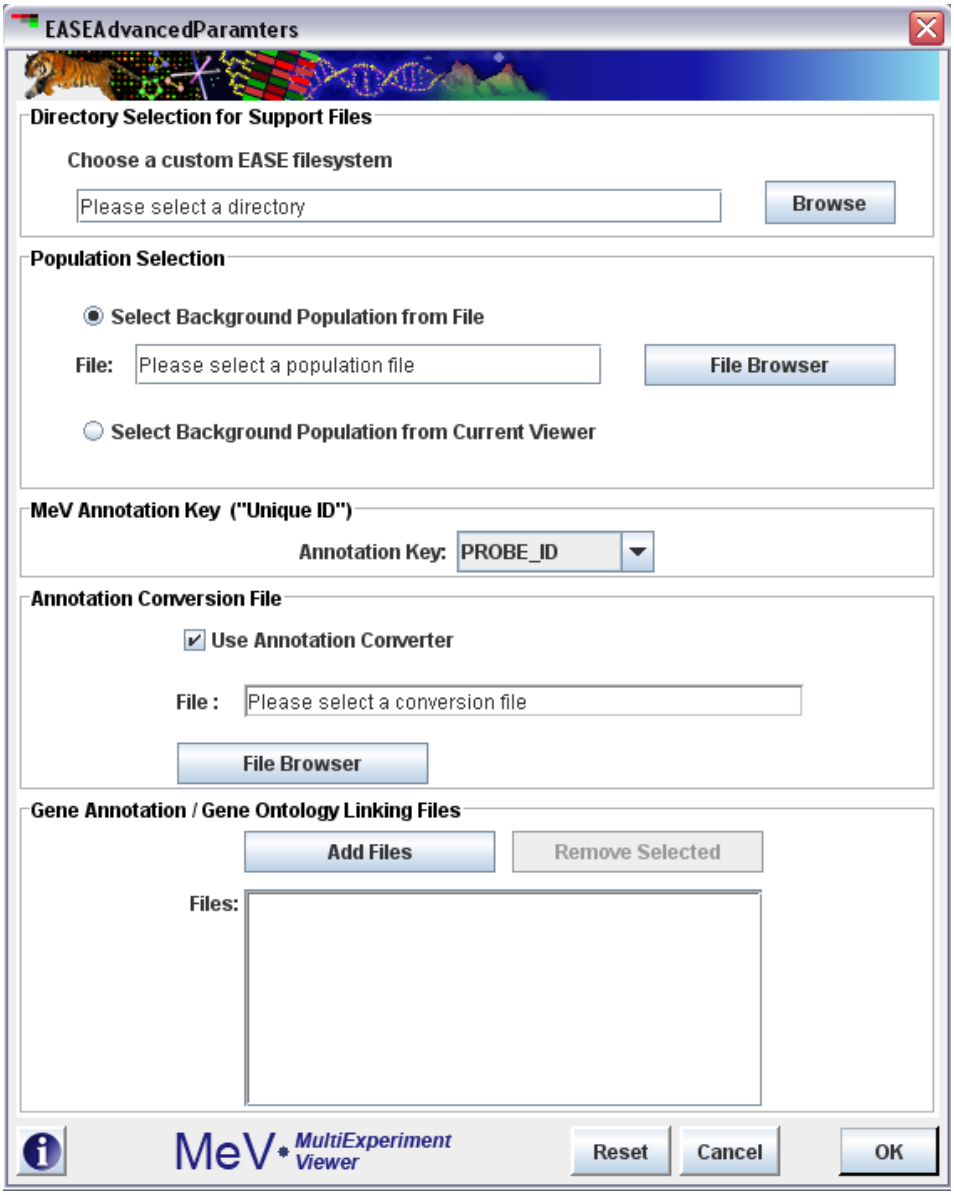
The first option for selecting a background population is to use a population file. This is simply a file containing all gene indices from which the cluster was selected. Often this includes all slide annotation or a large subset of the slide with bad data and control spots removed. The file format is a simple list of indices with one entry per line.

The second option for population definition is to simply use all of the genes currently 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.

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. This file selection tool will be disabled if you have selected "Use loaded array population as background".

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. If the user does not visit the EASE Advanced Parameters dialog when running EASE, all linking files in the current working directory will be loaded.



11.26.2 EASE Advanced Parameters Dialog

Statistical Parameters Dialog

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 sited 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.

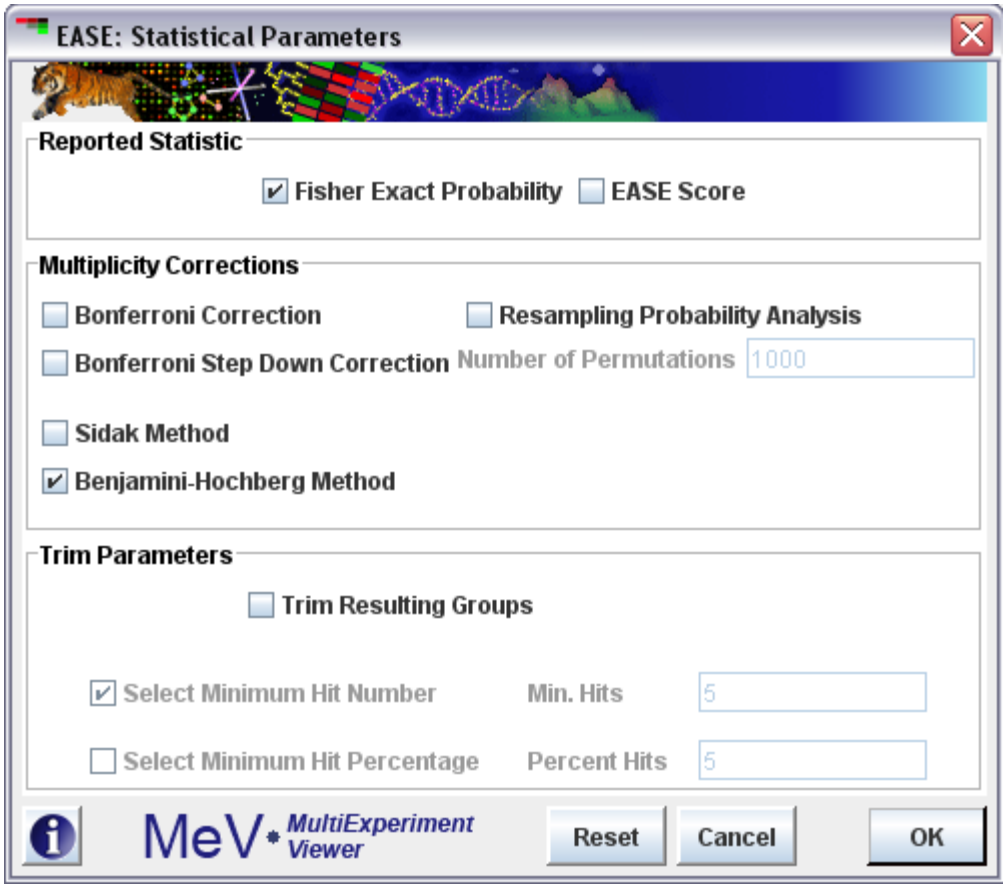
$$v' = 1-(1-v)^k$$ (Sidak method formula)

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.



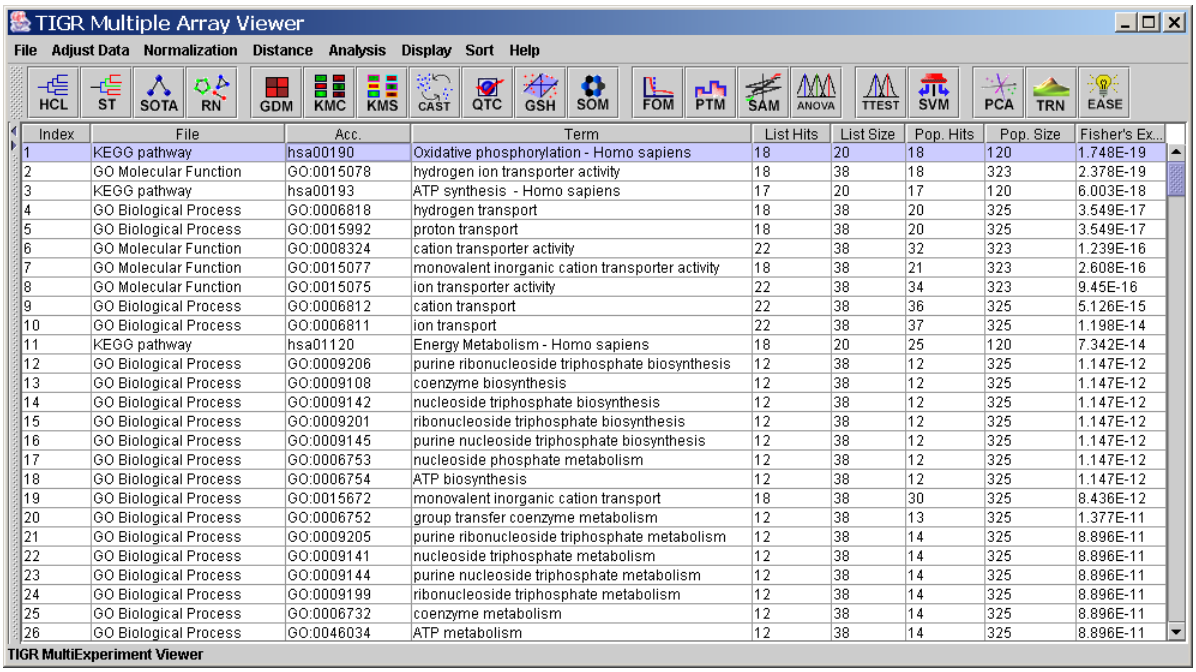
11.26.3 Statistical Parameters Dialog

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.



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

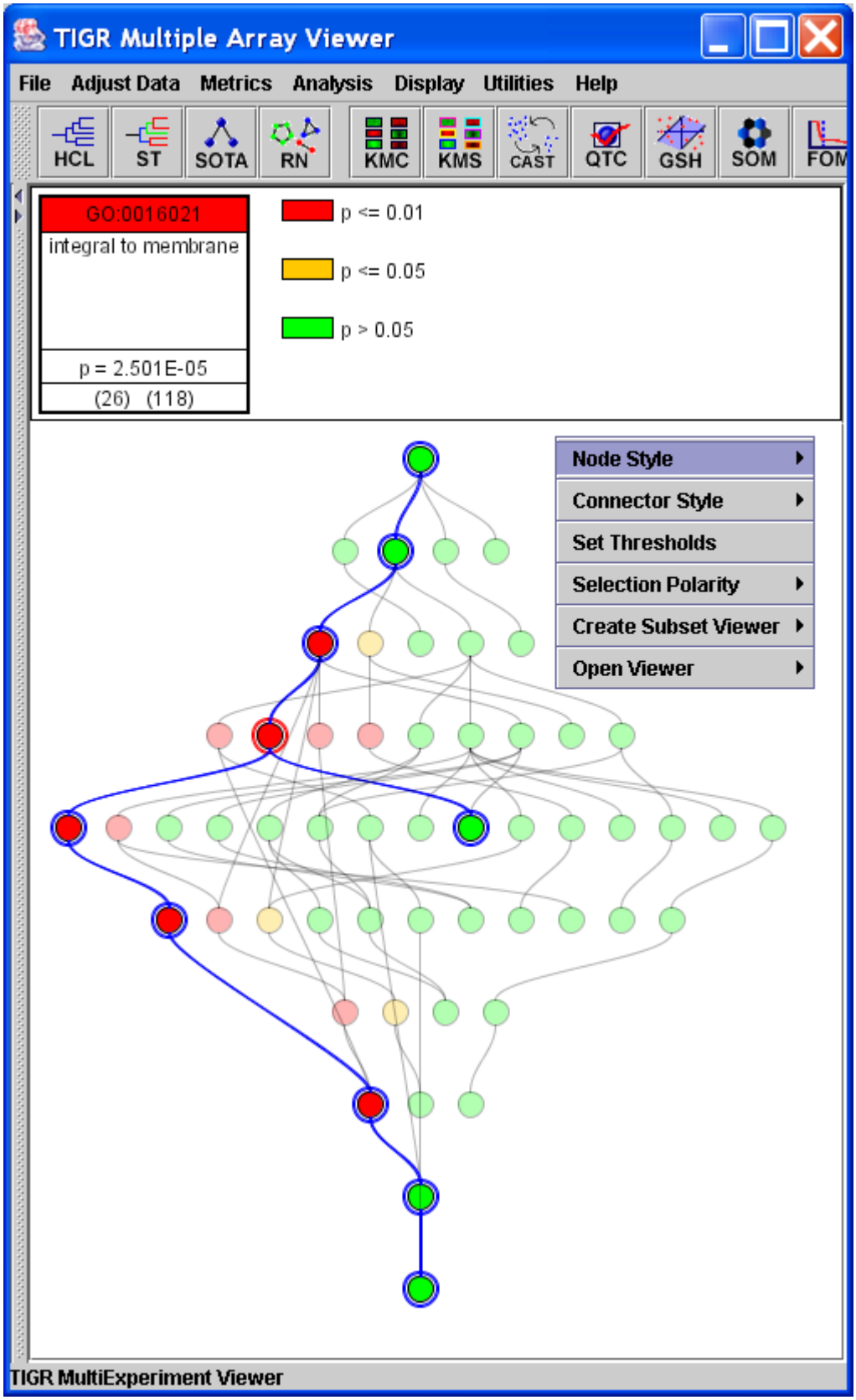
11.26.4 EASE Result Table

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.



11.26.5 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

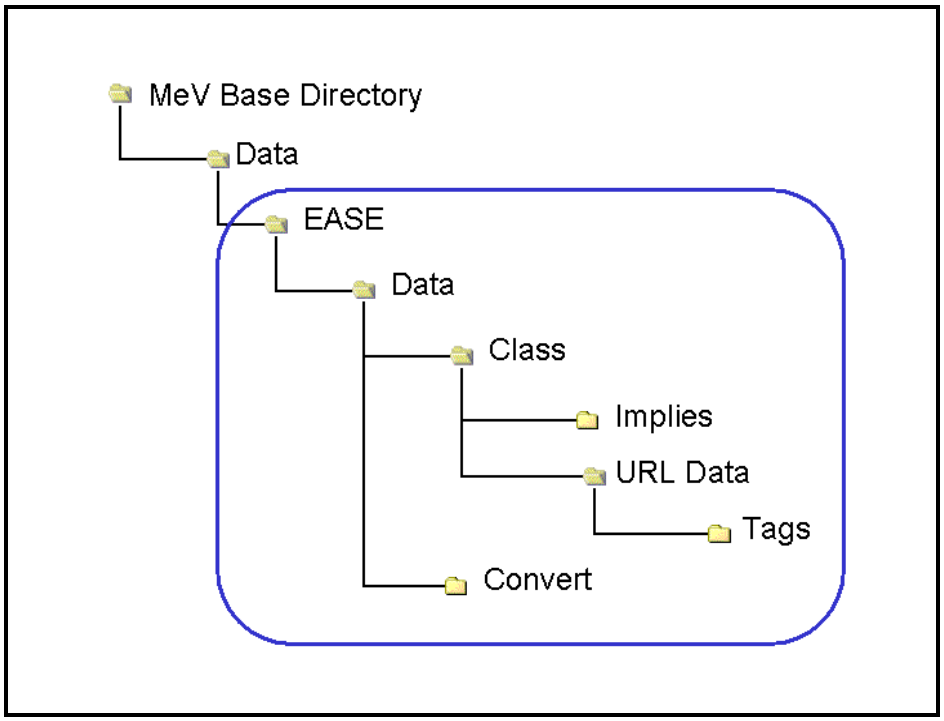
Extracting subsets of nodes from the main hierarchy is perhaps the most important feature of this viewer. After selection of a tree path, it is possible to use this option to build a new GO Hierarchy Viewer from the nodes in the selected path. The extracted trees can be rendered in a new window or docked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

EASE Directory Structure

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. Advanced users interested in creating custom EASE support directories should be especially familiar with the following section.



11.26.6 EASE Directory Structure

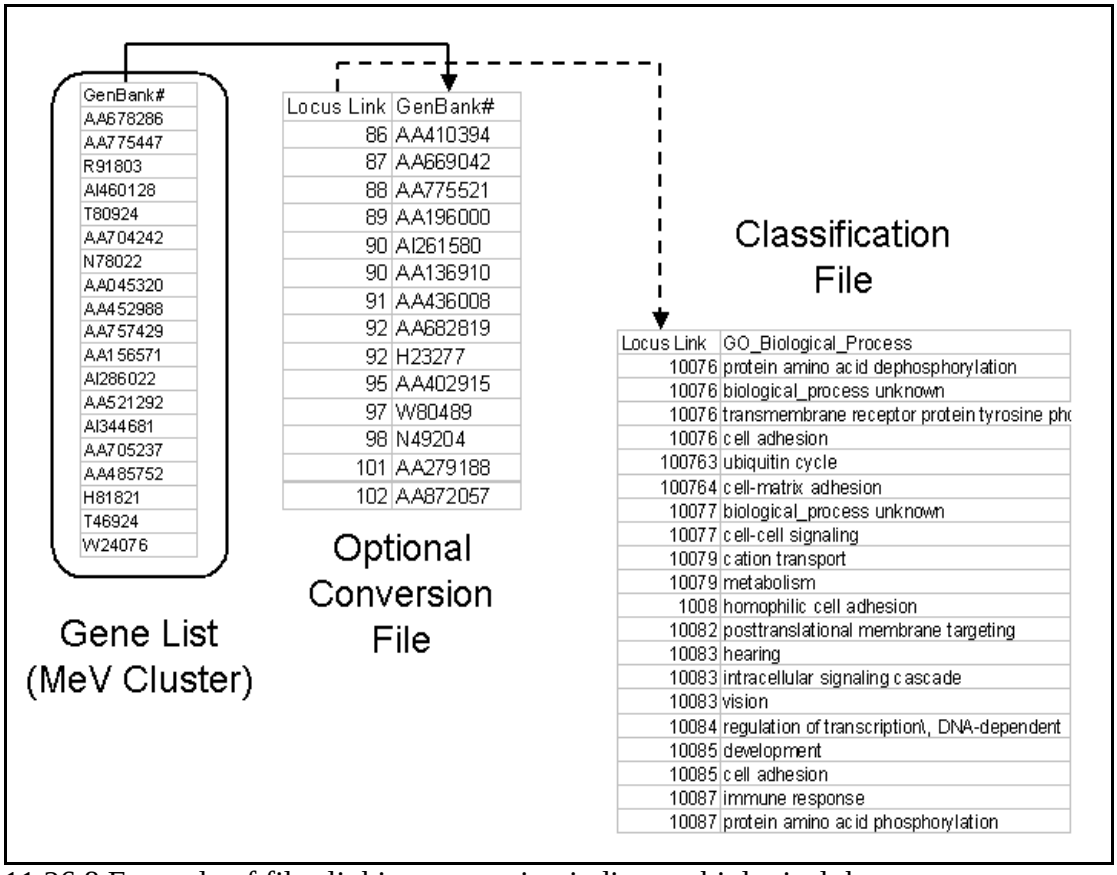
The following table lists the directories used in the EASE implementation within MeV. The primary directories, *Convert*, *Class*, *Implies*, *URL Data*, and *Tags* each contain files used by EASE. The minimal requirement is a file in the Class directory to map gene indices to themes. The optional annotation 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 *optional* 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 *optional* files relating themes to other themes (e.g. hydrogen ion transport "Implies" cation transport)
URL Data	Contains *optional* 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 *optional* files linking biological theme or pathway to accession numbers

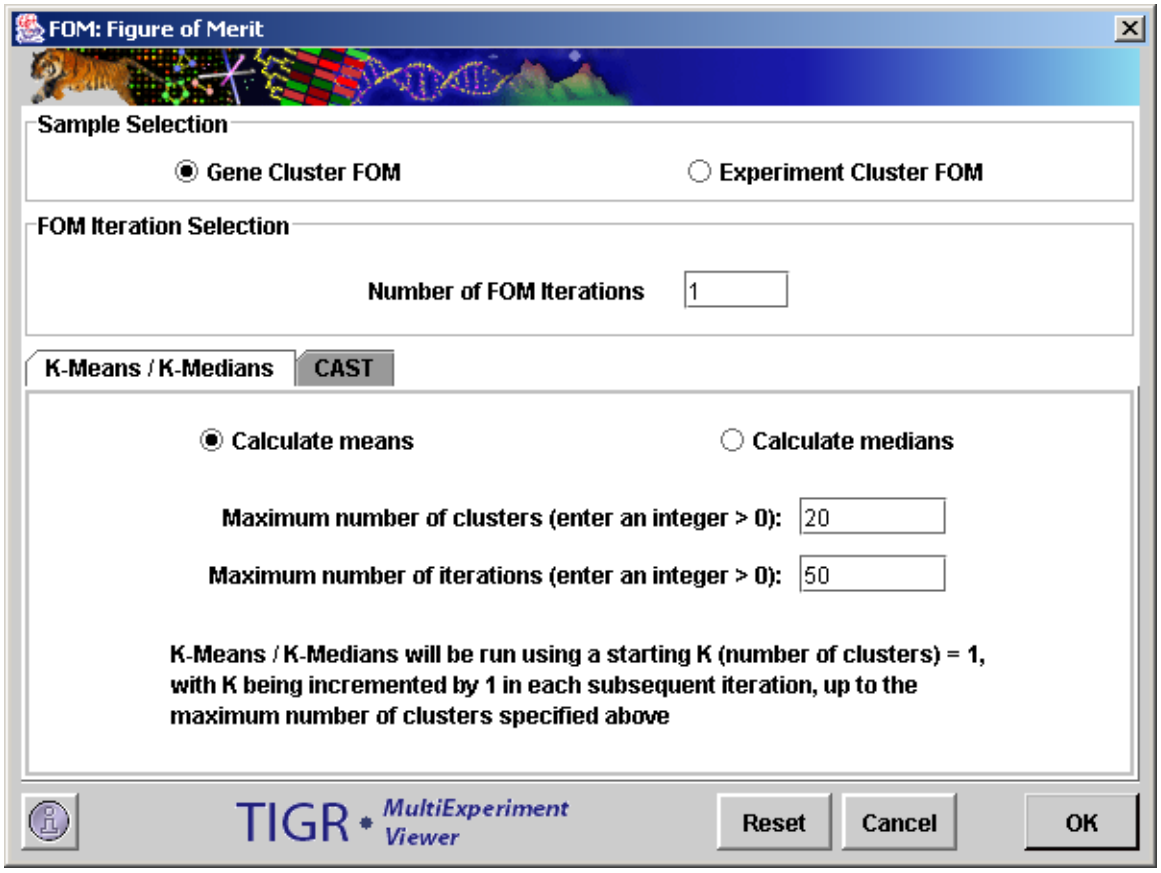
11.26.7 EASE Directory Descriptions

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 (11.26.8) 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.



11.26.8 Example of files linking annotation indices to biological themes

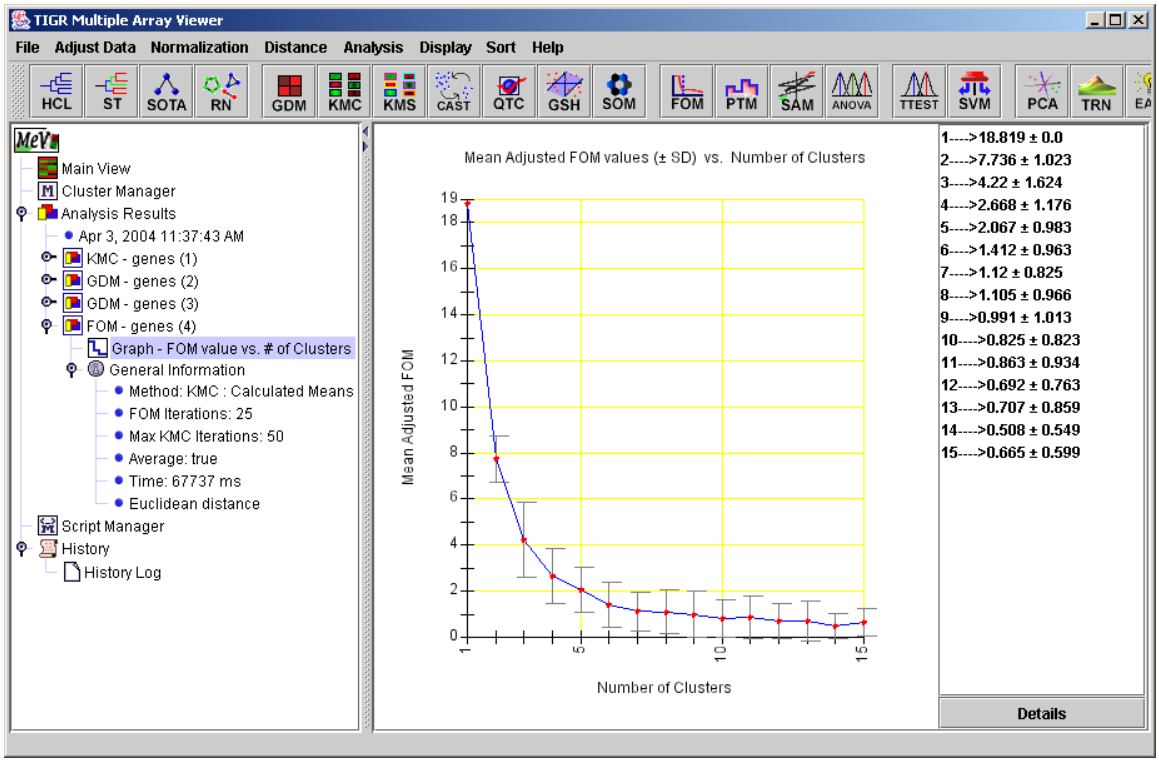
11.27 FOM: Figure of Merit



11.27.1 FOM Initialization Dialog

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.

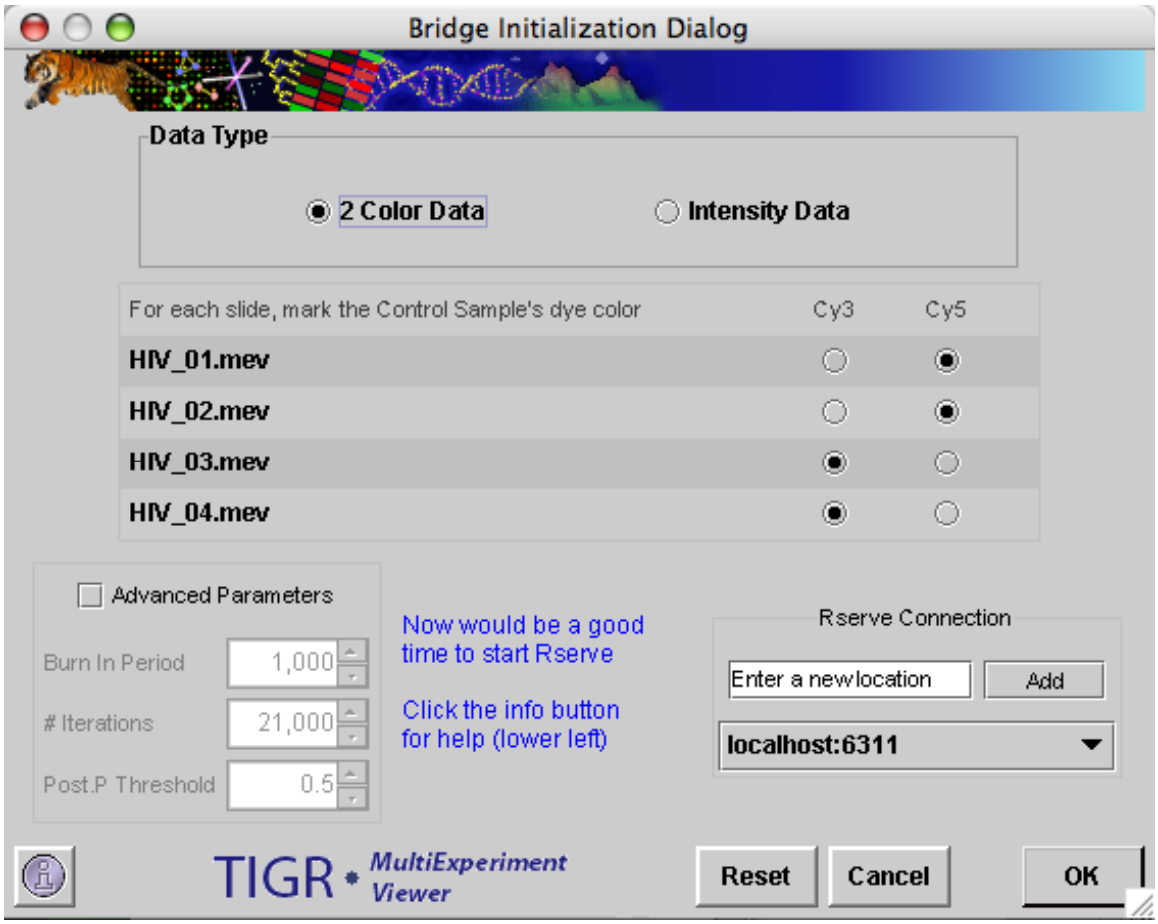


11.27.2 FOM Graph following 25 Iterations of KMC with K = 0 to 15.

11.28 BRIDGE: Bayesian Robust Inference for Differential Gene Expression

(Gottardo *et al*. 2005)

Test for differentially expressed genes with microarray data. This package can be used with both cDNA microarrays or Affymetrix chips. The package fits a robust Bayesian hierarchical model for testing for differential expression. Outliers are modeled explicitly using a t-distribution. The model includes an exchangeable prior for the variances which allow different variances for the genes but still shrink extreme empirical variances. Parameter estimation is carried out using a novel version of Markov Chain Monte Carlo that is appropriate when the model puts mass on subspaces of the full parameter space.



11.28.1 Bridge Initialization Dialog

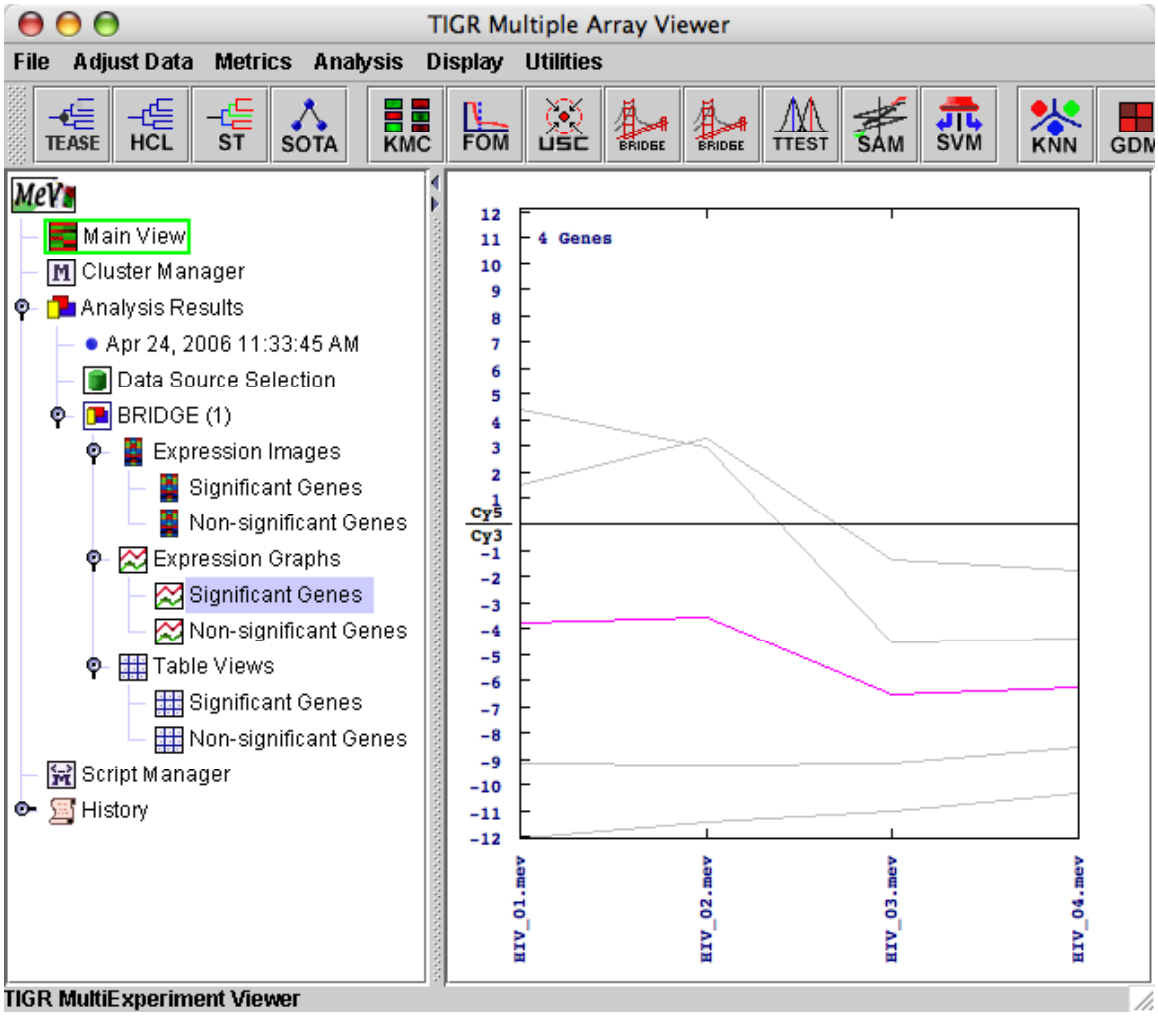
Rserve Connection

BRIDGE is a package written in the R programming language, and requires access to a computer running RServe to function. See Section 8 for details on installing R and RServe.

By default, Bridge will look on the local machine for an Rserve server. However, since Rserve is a TCP/IP server, theoretically it could be running anywhere. The user need only enter an IP address and port number separated by a : in the Text Field “Enter a new location”. By clicking “Add”, the new location will populate

the pull down menu. It will be saved to the user's config file and be available for later use.

Sample output from this module is shown below.



11.28.2 Bridge Results: Expression Graph of Significant Genes.

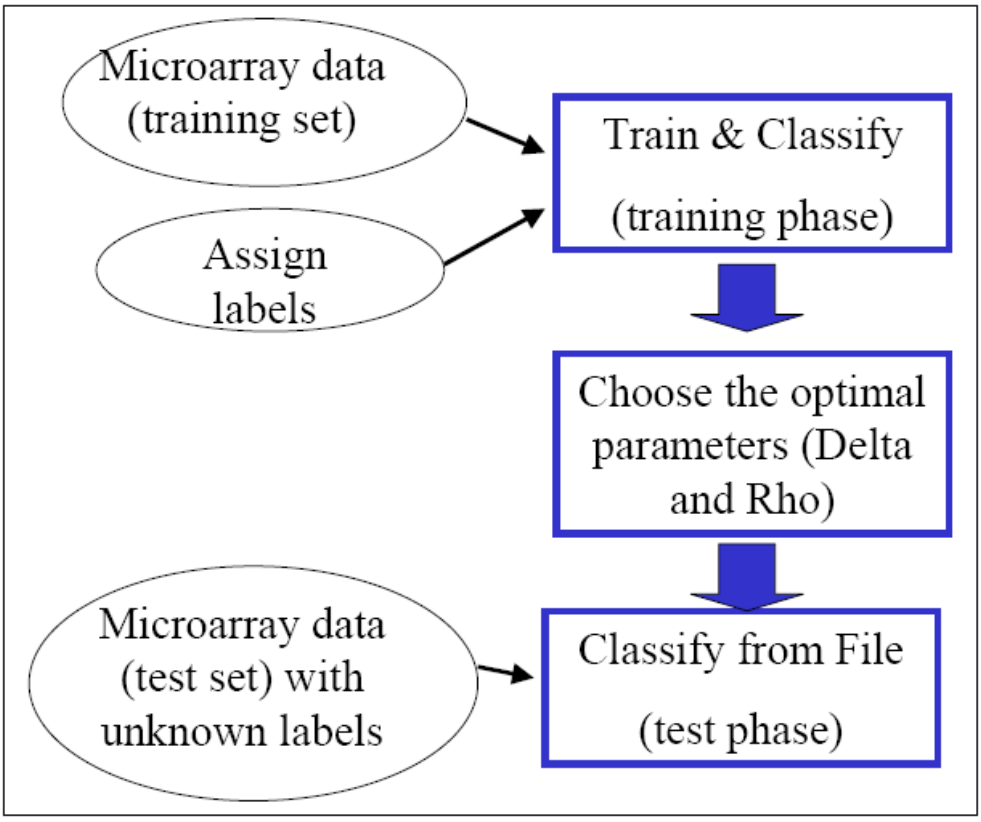
11.29 USC: Uncorrelated Shrunken Centroids

(Yeung et al 2003)

Prediction of the diagnostic category of a tissue sample from its expression profile and selection of relevant genes for class prediction have important applications in cancer research. We developed the *uncorrelated shrunken centroid* (USC) algorithm that is an integrated classification and feature selection algorithms applicable to microarray data with any number of classes. The USC algorithm is motivated by the shrunken centroid (SC) algorithm (Tibshirani et al. 2002) with the following key modification: USC exploits the inter-dependence of genes by removing highly correlated genes. We showed that the removal of highly correlated genes typically improves classification accuracy and results in a smaller set of genes.

As with most classification and feature selection algorithms, the USC algorithm proceeds in two phases: the **training** and the **test** phase. A *training set* is a microarray dataset consisting of samples for which the classes are known. A *test set* is a microarray dataset consisting of samples for which the classes are assumed to be unknown to the algorithm, and the goal is to predict which classes these samples belong to. The first step in classification is to build a "classifier" using the given training set, and the second step is to use the classifier to predict the classes of the test set.

In the training phase, the USC algorithm performs cross validation over a range of parameters (shrinkage threshold Δ and correlation threshold ρ). *Cross validation* is a well-established technique used to optimize the parameters or features chosen in a classifier. In m-fold cross validation, the training set is randomly divided into m disjoint subsets with roughly equal size. Each of these m subsets of experiments is left out in turn for evaluation, and the other (m-1) subsets are used as inputs to the classification algorithm. Since the USC algorithm is essentially run multiple times on different subsets of the training set, the cross validation step in the training phase is quite computationally intensive. The end result of the training phase is a table of the average number of classification errors in cross validation and the average number of genes selected corresponding to parameters Delta Δ and Rho ρ. Depending on the dataset being analyzed, there might be a trade-off between the average number of errors and the number of genes selected. The user will be asked to select one set of parameters (Delta Δ and Rho ρ) to be used in the test phase in which microarray data consisting of experimental samples with unknown classes will be classified.



11.29.1 Overview of USC

Initial Dialog Box

The initial dialog box allows you to choose from 2 modes of operation - '**Train & Classify'** or '**Classify from File'**. The option 'Train & Classify' should be used for the training phase or if both the training and test sets are uploaded as one microarray data. The option 'Classify from File' corresponds to the test phase of the algorithm, and assumes that a classifier has been previously built.

When Training & Classifying, the user is required to enter all the unique class labels of the known (training) experiments. By default, there is space for 2 class labels. If more are needed, use the '# of Classes' spinner. 'Entering Class Labels' is disabled if you are 'Classifying from File'.

You are also allowed at this point to make any adjustments to the default parameters. By default, the parameters are disabled. Clicking on the 'Advanced' checkbox enables adjustment of the parameters.

Advanced Parameters

 # Folds is the number of times to divide the training set in pseudo training and pseudo test sets during a cross validation run. For example: if there are 10 total training experiments to be cross validated and *# Folds* = 5, 2 experiments will be removed as pseudo test experiments during each Cross Validation Fold. After 5 Folds, all 10 experiments will have been used once and only once in the pseudo test set. A higher *# Fol*ds is recommended for smaller class size.

 # CV runs is the number of times to repeat cross validation. Reducing this parameter will reduce computation time in the training phase at the expense of less accurate average number of classification errors and genes selected from the cross validation step.

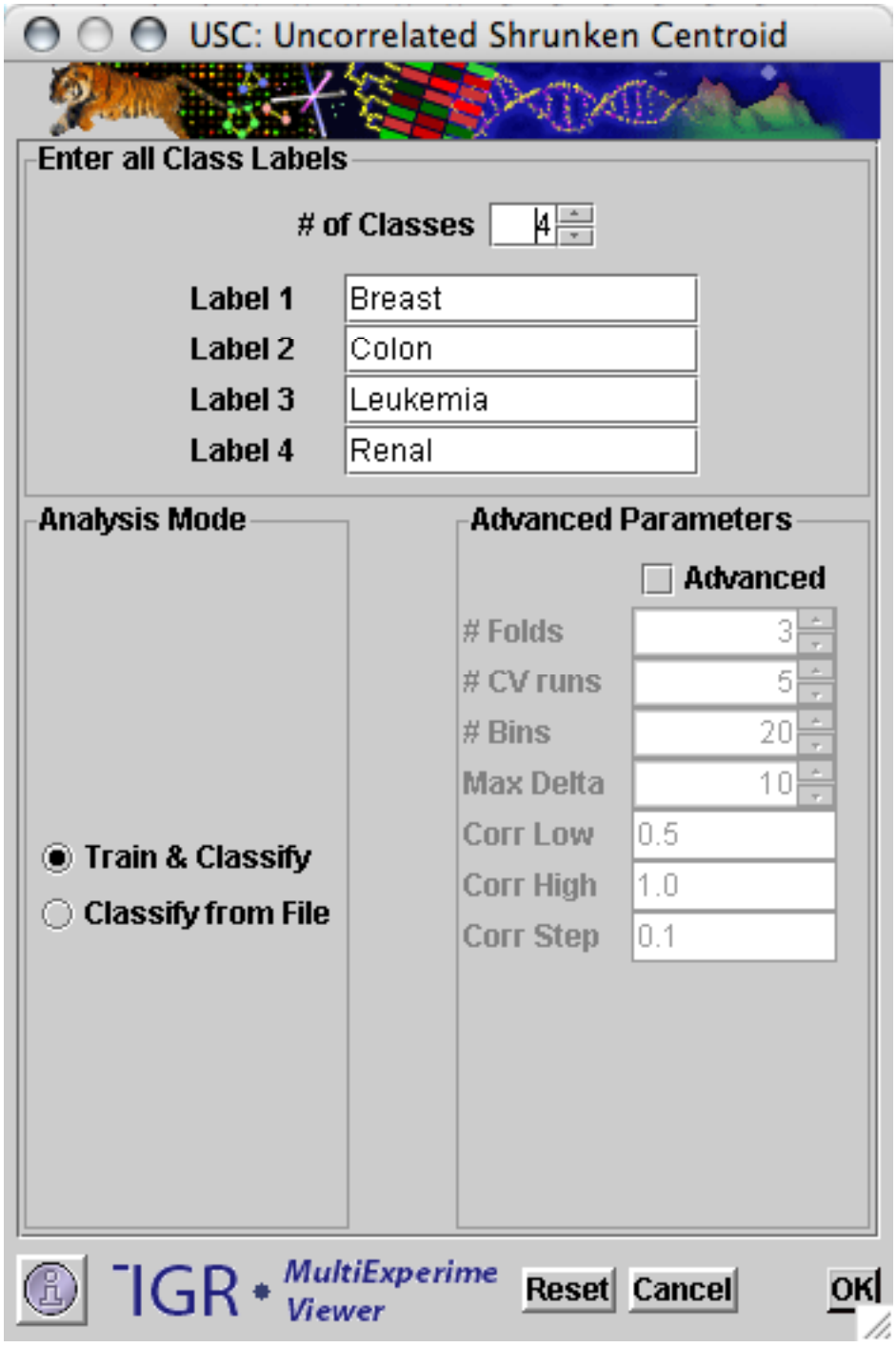
 # Bins is the number of different values to use for Delta.

Max Delta is the maximum Delta value to use. *Deltas* will range from { 0 – *Max Delta* } incrementing by *Max Delta / # Bins*. The user may consider reducing this parameter to get a more precise estimate of the optimal shrinkage threshold Δ if the optimal estimated Δ is significantly smaller than this value. On the other hand, if the number of classification errors from cross validation is unsatisfactory, the user may consider trying a larger *Max Delta*.

Corr Low is the lowest Correlation Coefficient threshold to use. The default is 0.5, which should be sufficient for most cases.

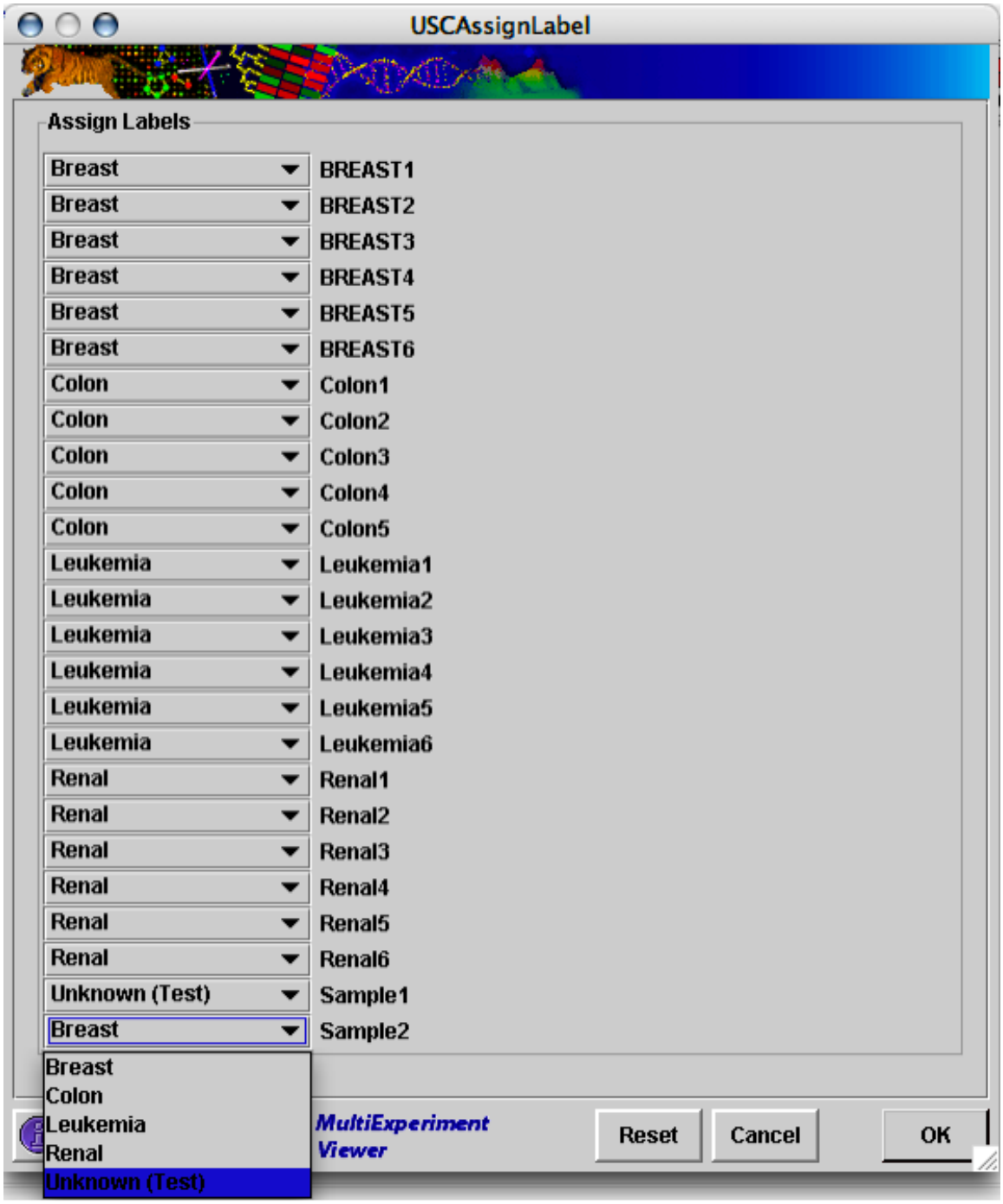
Corr High is the highest Correlation Coefficient threshold to use. The default is 1.0, which is the maximum possible correlation.

Corr Step is the value to increment over going from *Corr High* to *Corr Low*



11.29.2 USC Initialization Dialog

If you are doing *Training & Classifying*, the USC algorithm needs to know the classes of the experiments in the training set. Using the pull down menus, assign labels to each of the experiments that were loaded. Label any test experiments as 'Unknown (Test)'. Keep in mind that you are not required to test any experiments at this point. You may just classify an entire training set, saving the classifier as a file for later use on any test experiments of choice.



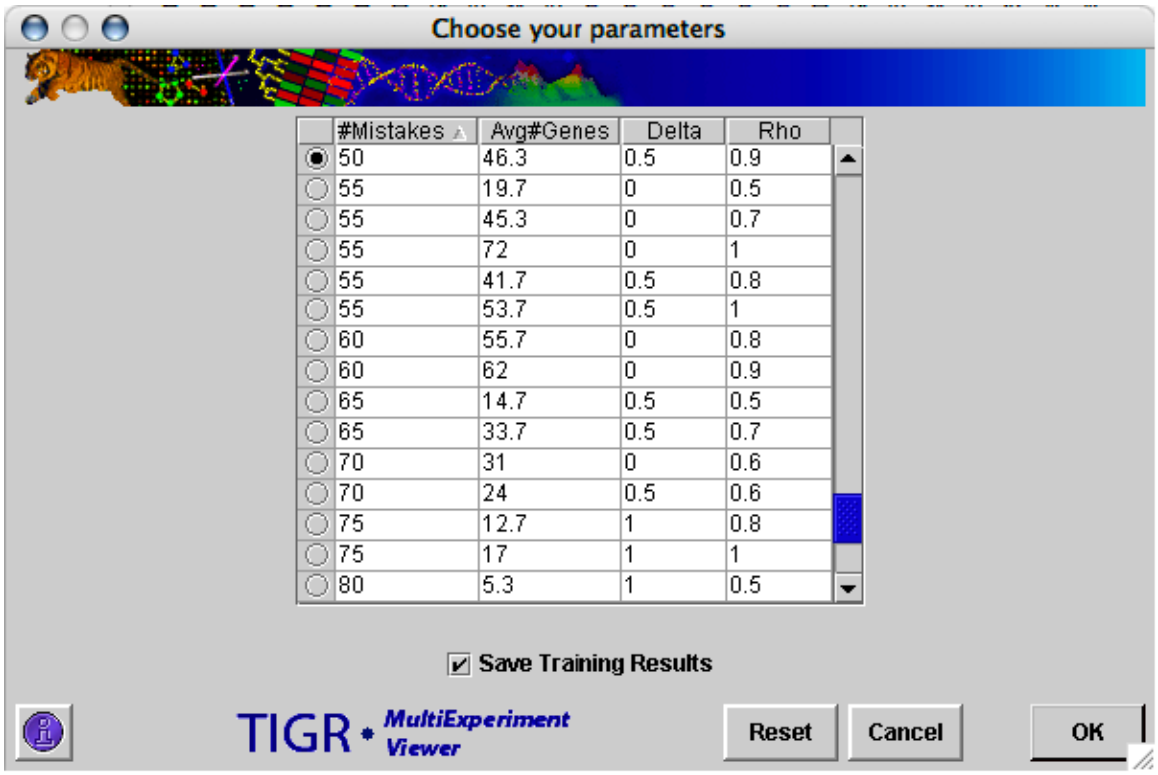
11.29.3 USC Assign Label Dialog

Click 'OK' and wait a short eternity for Cross Validation to run. When Cross Validation is finally finished you'll see the 'Choose Your Parameters' dialog box.

Choose Your Parameters Dialog Box

During Cross Validation, the USC algorithm has compiled a list of results. During each fold of cross validation, each experiment in the pseudo test set has been tested back against the the remaining experiments of the pseudo training set. Here, you'll be asked to choose between accuracy and the # of genes to use during testing. When the 'Save Training Results' checkbox is checked (default), you'll

be prompted to save the training file. If you have any Test experiments, they will be tested now using your chosen Delta Δ and Rho ρ values.



11.29.4 USC Parameters Dialog

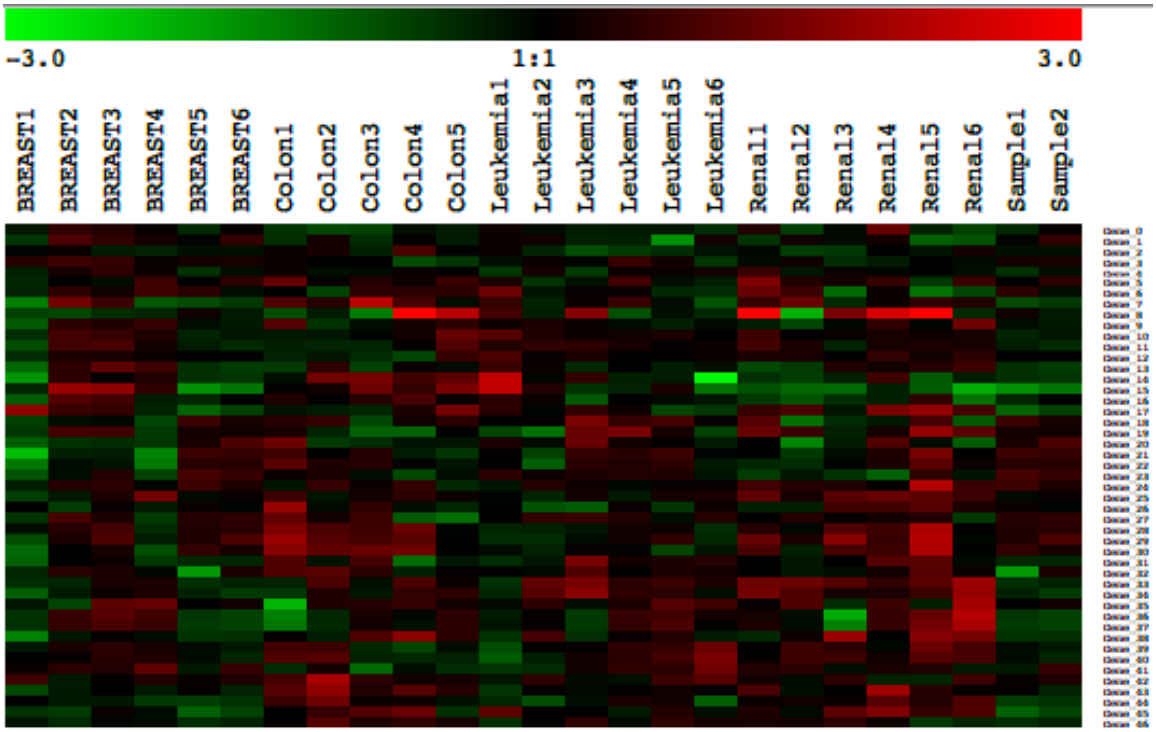
USC Summary Viewer

The results of the USC algorithm are returned to the main Analysis Tree in the left pane of the Multiple Array Viewer window. Clicking on 'Summary' will display the following view. Any test experiments that were loaded are listed along with their class assignment and the Discriminant score of that assignment. Parameters are also displayed as well as the list of the genes that were used for this classification. You can save that gene list if desired.

11.29.5 USC Parameters Dialog

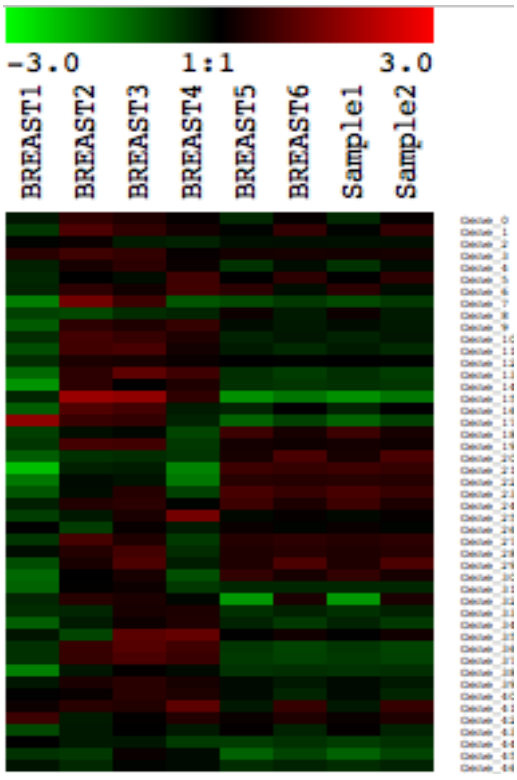
A number of heat map visualizations are also available.

Clicking on ‘All Loaded Experiments – Genes Used’ displays all the experiments that were loaded and the genes that were used during this classification.



11.29.6 USC Heatmap

There is also a heat map visualization for each of the classes in the analysis, again, with the genes that were used during this classification.

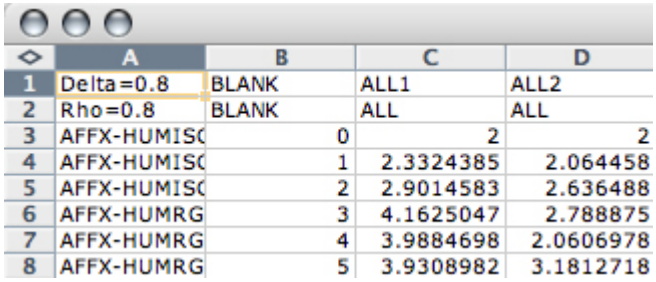


11.29.7 USC Heatmap

Classify From File

Having saved the results of a classification, you may want to test experiments without the time intensive Cross Validation step. It is important that you use different sets of experimental samples in the training and test phases. Keep in mind that you can only test experiments that are of the exact same chip type as the training experiments.

If you would like to experiment with different values for Delta and Rho, you can easily change them in the Training Result File.



	A	B	C	D
1	Delta=0.8	BLANK	ALL1	ALL2
2	Rho=0.8	BLANK	ALL	ALL
3	AFFX-HUMISC	0	2	2
4	AFFX-HUMISC	1	2.3324385	2.064458
5	AFFX-HUMISC	2	2.9014583	2.636488
6	AFFX-HUMRG	3	4.1625047	2.788875
7	AFFX-HUMRG	4	3.9884698	2.0606978
8	AFFX-HUMRG	5	3.9308982	3.1812718

11.29.8 Training Result File

11.30 NonpaR: Nonparametric Statistical Tests

(Hollander and Wolfe, 1999)

The NonpaR module in MeV consists of four nonparametric tests that can be used to analyze several common experimental designs. Three of these tests, Wilcoxon Rank Sum, the Kruskal Test, and the Mack-Skillings Test find genes that are differentially expressed between two or more experimental groups under study. The Fisher Exact Test is applied to data where there are two distinct experimental groups and the data values for each gene are separated into two distinct bins using a numerical cutoff. The test then describes if a data bin (e.g. values above a supplied expression threshold) are over represented in one sample group or the other. Hollander and Wolfe provide a complete description of all of the methods in this module as well as the qualities and the general benefits of nonparametric statistical tests. Please refer to that text for a complete overview of nonparametric tests and the formulae for computing these test statistics.

Nonparametric tests make few assumptions about the underlying distributions of values in the population from which the sample is taken. There is no assumption that the population has a normal distribution of values. Note that there is assumed to be an underlying distribution and that different experimental groups are sampling populations that share the same distribution type. The difference from parametric tests is that the underlying distribution is not necessarily assumed to be normal. In addition to their distribution-free property, nonparametric tests also tend to be less sensitive to outlier measurements than their parametric counterparts since nonparametric tests are based on the ranking of the data.

Brief Test Descriptions

The *Wilcoxon Rank Sum* test handles an experimental design where there are exactly two experimental groups and there are replicate hybridizations for each of these groups. The null hypothesis is that the observations for both samples come from the same probability distribution, that the underlying population means are the same. The test attempts to reject this hypothesis and reports genes that are over or under (two-tailed test) expressed in group 2 relative to group 1.

The *Kruskal Test* is analogous to the parametric one-way ANOVA test. This test handles designs where there are n-experimental groups that vary in some way related to one experimental factor. An example might be three experimental groups representing sets of patients in a clinical trial, ‘control-untreated’, ‘low dose intervention’, and ‘high dose intervention’. The Kruskal tests can be applied to two or more experimental groups. If there are exactly two groups then the Wilcoxon Rank Sum test is typically used.

The *Mack-Skillings Test* is a generalization of the more familiar Friedman test for two factor designs. This test handles design where there are two experimental factors (such as strain and treatment) and there are some number of levels of each factor (e.g. 2 strains and 3 treatment levels, as in a 2 x 3 design). Unlike the Friedman test, the Mack-Skillings test allows for replication and handles balanced

or unbalanced complete designs. This means that each factor combination group should have one or more observation but that factor combination groups (or different cells in the design matrix) can have uneven numbers of replicates. The following table provides an example of an unbalanced complete design having at least one observation per strain/temperature combination but having uneven numbers of replicates across strain/temperature combinations. Two factor design table example: (numbers represent the number of samples in each experimental group)

Strain	Temp = 25C	Temp = 32C	Temp = 37C
Wild Type	4	4	3
Mutant A	3	3	4

The test looks for differences between groups that are related to either of the two factors.

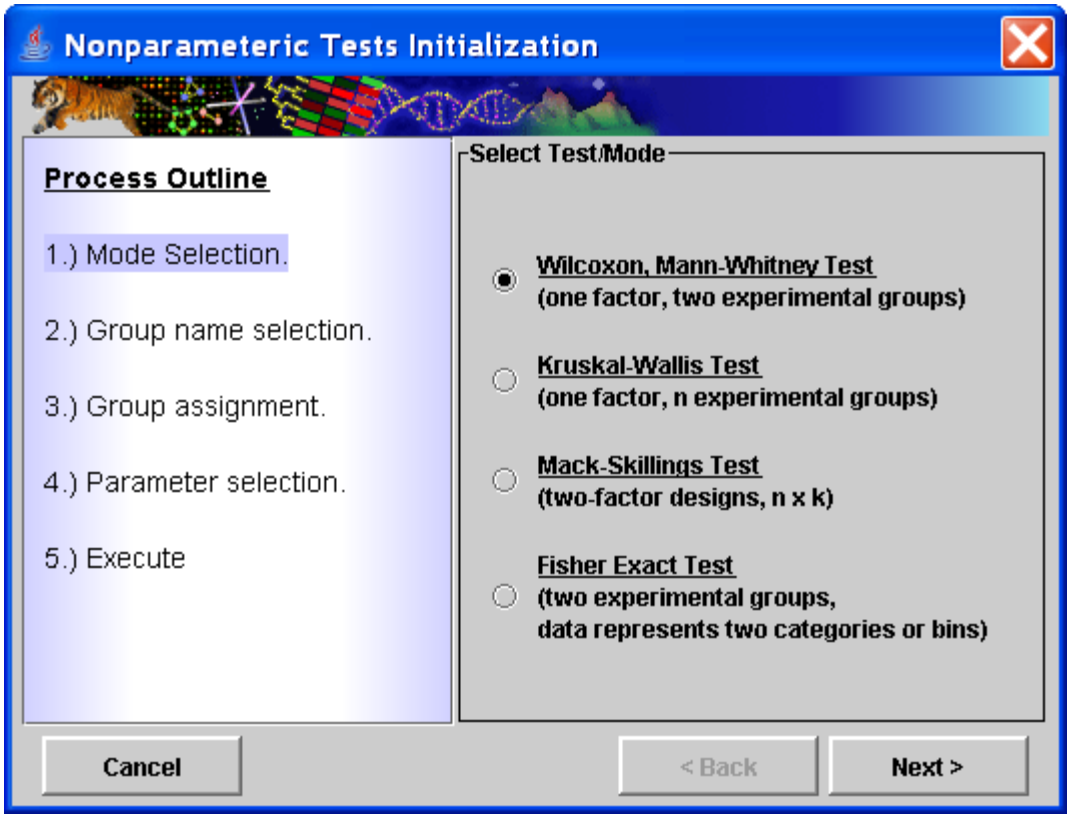
The *Fisher Exact Test* is applied to data where there are two distinct groups of samples and the data can be binned into two groups, called data bins, based on a supplied numerical cutoff. The test then answers the question as to whether there is a non-random association between a sample group and the binned data. Stated in a more another way the question may be asked, are the data values in data bin one associated disproportionately with sample group one or sample group 2? A concrete example of this is where the input data values are from CGH and the data values are numerical representations of absent/present calls (0 or 1) and the samples fall into two distinct groups such as patient set A and patient set B. A contingency matrix classifies each data value for a particular gene according to this sample table:

	Patient Set A	Patient Set B
Present Calls	16	4
Absent Calls	2	19

Patient set A has 18 members, patient set B has 23 members. For this particular gene, there are present calls for 16 of 18 patients in set A while in set B there are only 4 present calls out of 23 patients. This contingency matrix captures the disproportionate number of present calls in patient set A (there are other ways to express the disproportionality). The Fisher Exact Test takes this matrix and reports a p-value describing the probability of this matrix and less likely (more disproportionate situations) occurring by chance. One tailed probabilities are reported for situations where the expected disproportionality is in one direction or the other. Most users will be most interested in the two tailed probability in which reports the probability of having a disproportionality reflected in the matrix in either extreme.

Running NonpaR Tests

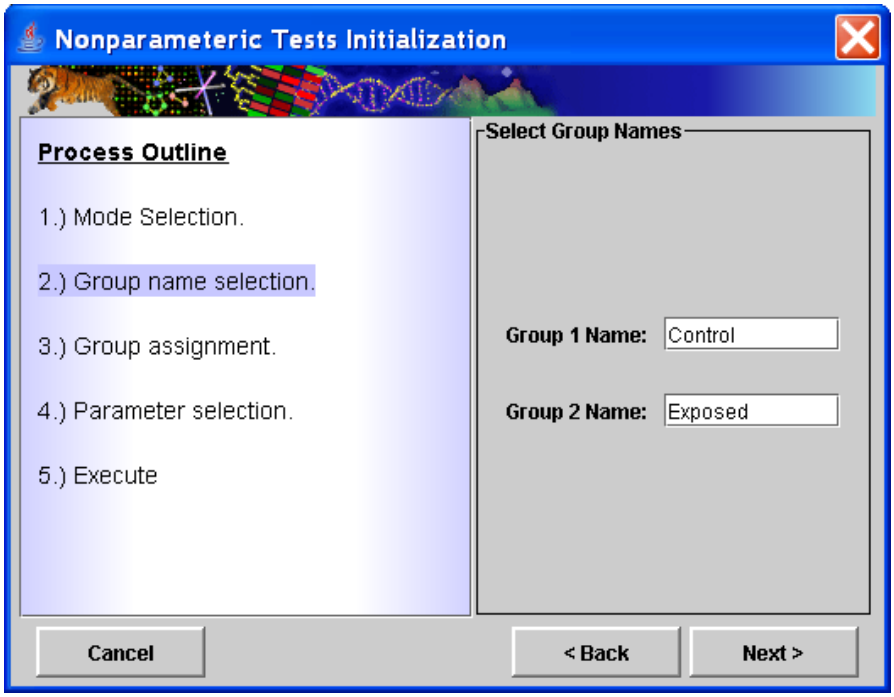
NonpaR uses a set of parameter input dialogs that open sequentially to provide input options that correspond to each step of the process. The first step in the processes is the selection of a test or mode. The test descriptions above will help determine which tests would fit your experimental design and address the specific hypothesis to be tested. The Wilcoxon and Kruskal tests progress using similar input dialogs while the Mack-Skillings test uses parameter panels that support two factor experimental designs. For that reason we first describe the more basic Wilcoxon and Kruskal input process and then turn to the dialogs specific to running the Mack-Skillings test. The Fisher Exact Test will follow.



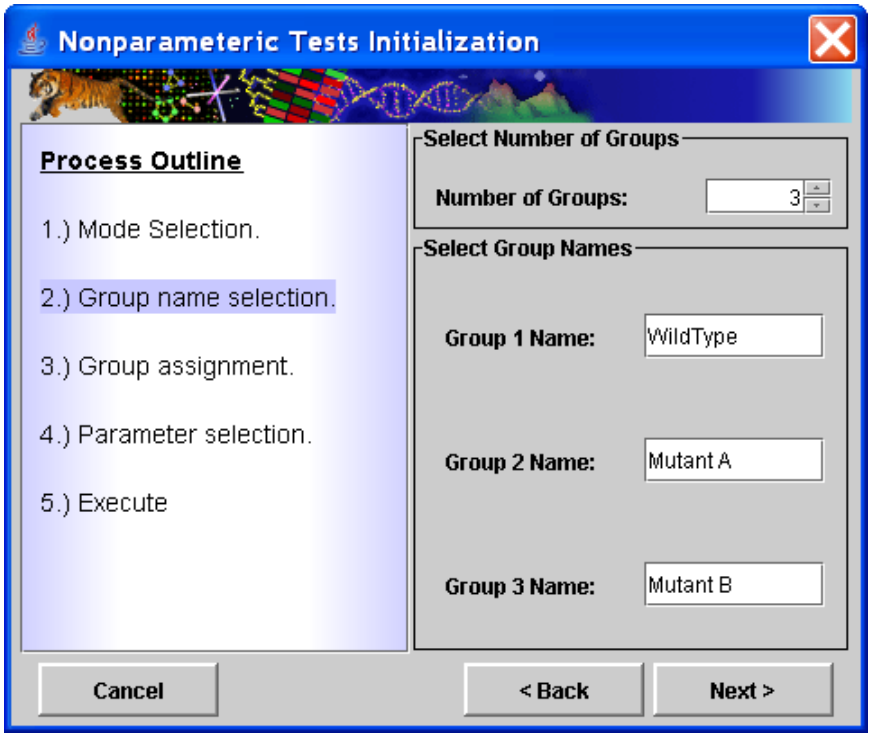
11.30.1 Nonpar Test Selection (Mode Selection) Dialog

Wilcoxon Rank Sum and Kruskal Tests

The second step is to specify group names or labels. By default, all tests start with the names initialized as group1, group2…etc.. These can be modified to reflect the conditions related to each experimental group. In the case of the Kruskal test, the number of groups is selected in addition to the group labels (See figures below).



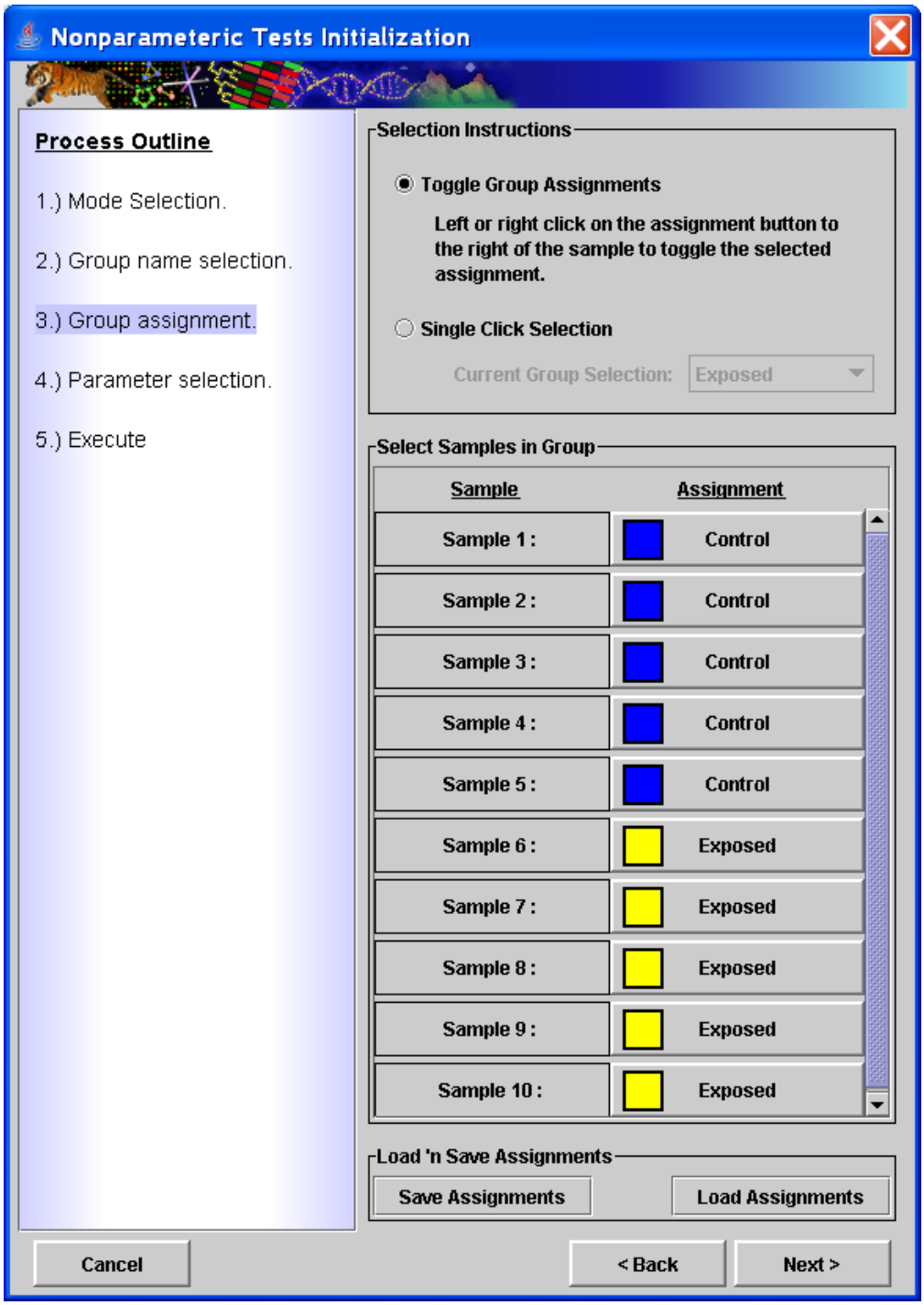
11.30.2 Wilcoxon Group Name Selection



11.30.3 Multi-group Name Selection Dialog (used for Kruskal Test)

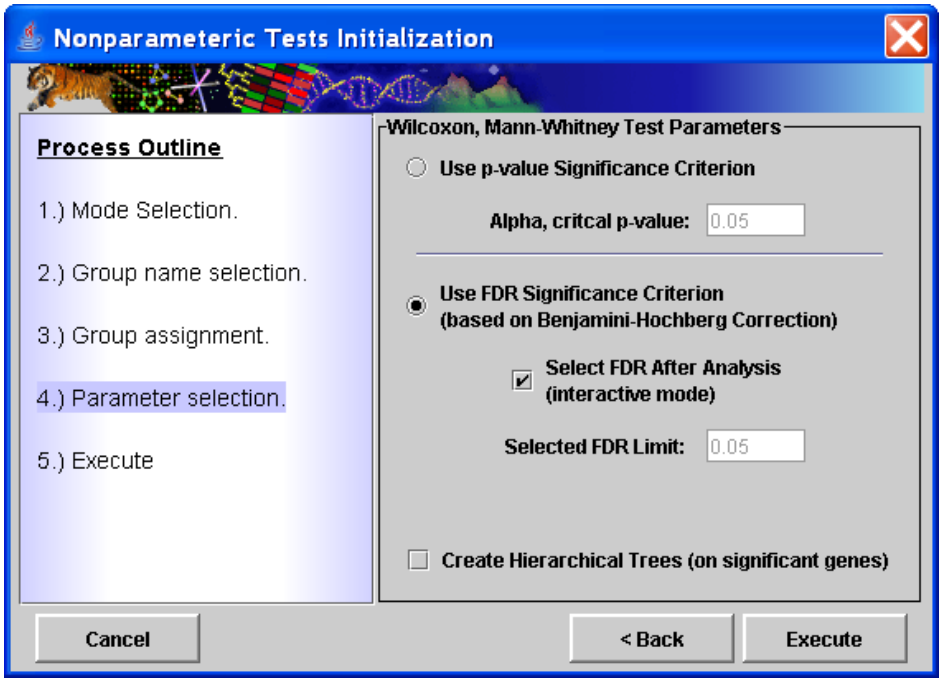
The third step is group assignment where each sample is assigned to a specific experimental group or placed into a group that is excluded from the current run. Left clicking on an assignment button next to a sample will change the group assignment to the next group. Successive left clicks will cycle forward through the possible group assignments. The buttons have been modified so that a right click will cycle through the group options in the reverse order which can be useful in the case where there are multiple groups. Another option is to use the *Single*

Click Selection option. Select a particular group to assign from the dropdown menu and then left click on assignment buttons next to the samples that fall into the selected group. Note that after group assignments are made that a file can be saved that captures the group assignments. This can be loaded on future runs to specify group assignments.



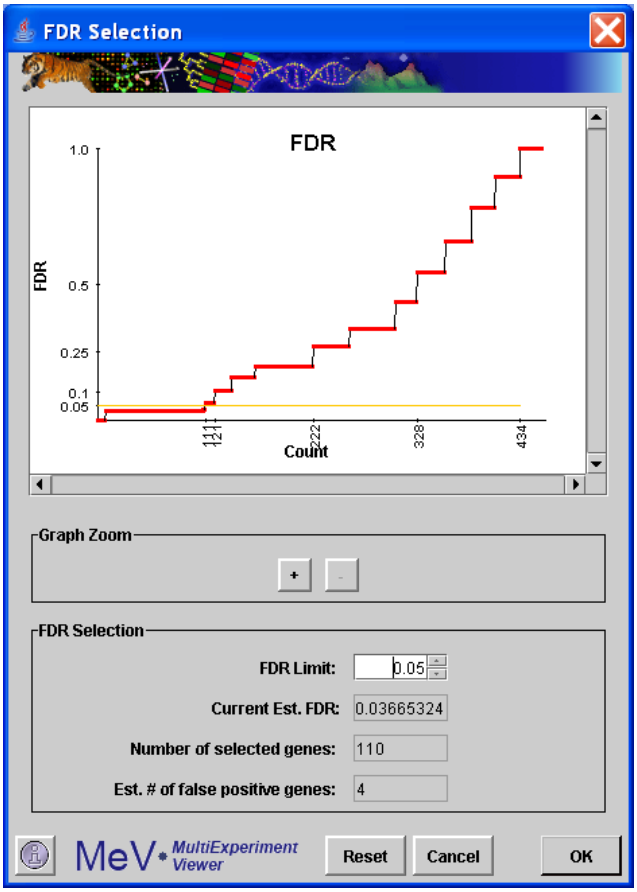
11.30.4 Sample Group Assignment Dialog. Samples are assigned to specified experimental groups.

Step four is parameter selection. For the Wilcoxon and Kruskal tests this dialog present two possible criteria for controlling error rates when collecting 'significant' genes. The first is to supply an alpha value or critical p-value as a cutoff. The other option is to use an estimated FDR or false discovery rate. This attempts to estimate the fraction of false positives among the set of genes selected by the test. Under the FDR option one can either select to enter a specific FDR cutoff or can use an interactive graph to adjust the relationship between number of genes captured and the estimated FDR for that group of genes. The FDR option to control type I errors uses the Benjamini-Hochberg (BH) Correction of p-values and ranks based on these adjusted p-values. In all cases the raw p-values are reported and in the case of FDR, the BH adjusted p-values are also reported.



11.30.5 Test Parameter Dialog. Provides alpha value critical value option or FDR option for setting cutoff for significance.

If the FDR option is selected, an interactive graph is displayed to allow the user to adjust the number of captured genes while observing the FDR. The graph, shown in the figure below allows the user to zoom in to view the more critical behavior below FDR's of 0.1. Each zoom step increases or decreases the displayed *upper* end of the FDR range according to these preset levels, 1.0, 0.5, 0.25, 0.1, 0.05. The lower limit is kept at zero. Adjustments to the FDR cutoff can be made in 0.01 unit increments. The lower portion of the window displays the number of genes captured and the estimated FDR for that set of genes and the number of genes that the FDR represents.

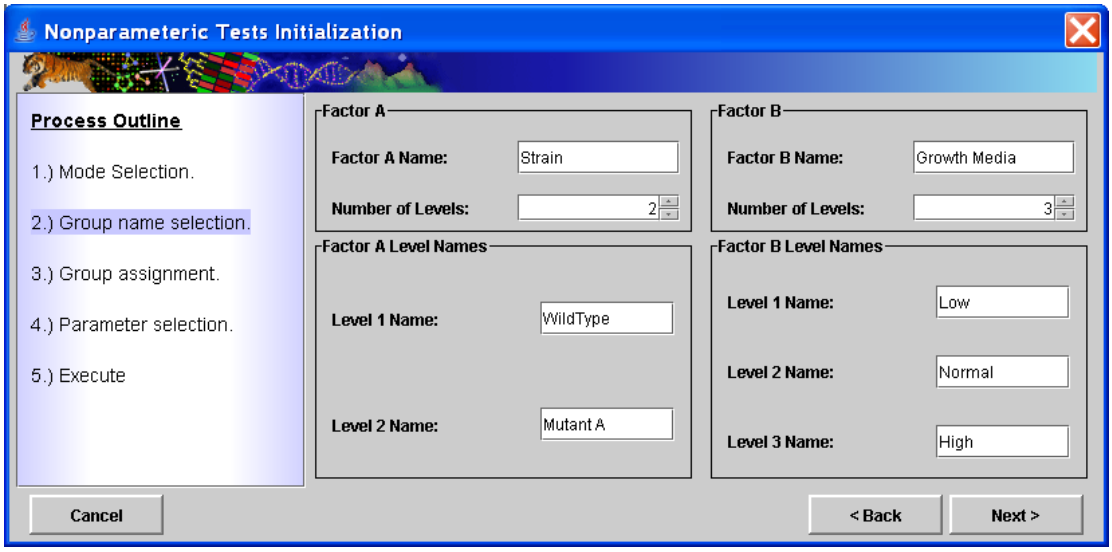


11.30.6 FDR Display and Selection Dialog.

The output from NonpaR consists of the typical cluster viewers, Expression Images, Expression Graphs, Centroid Graphs, and Table Viewers. The Cluster Info viewer will capture the number of significant genes, the test name, and the significance criteria and value. One important note is that clusters saved to file from any of the NonpaR cluster viewers will have the measured statistic and p-values output to the file in addition to the expression and annotation information.

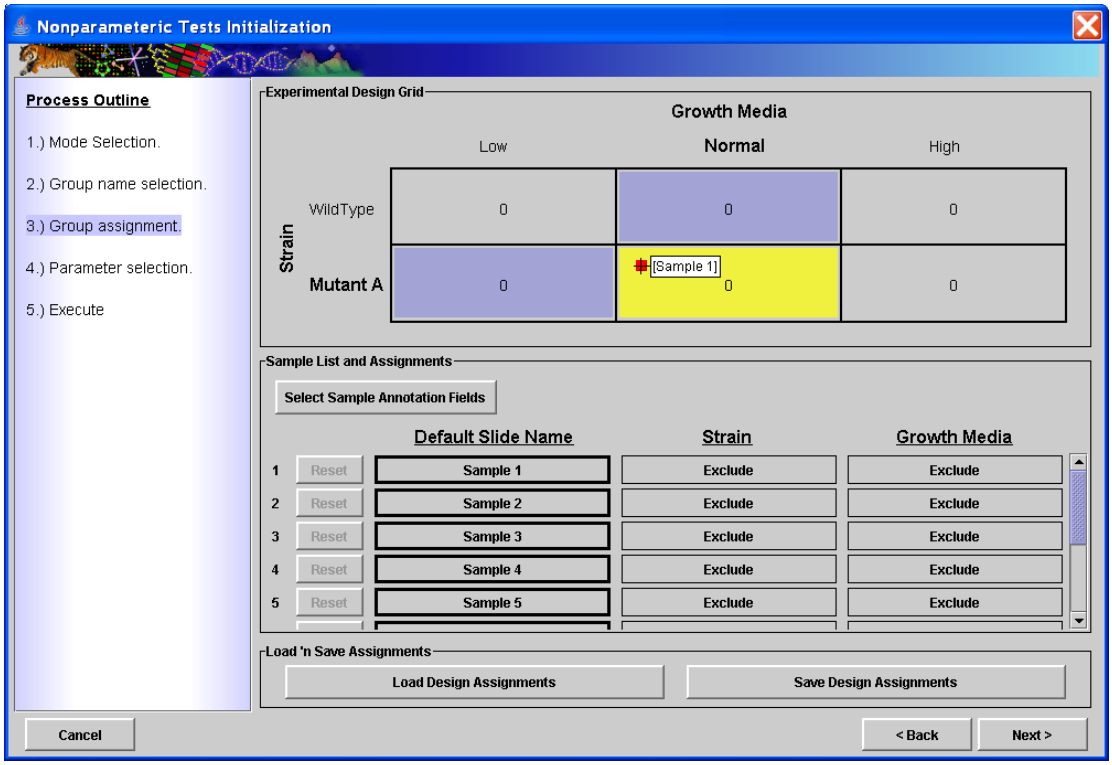
Mack-Skillings (MS) Test Dialogs

After selection to run MS a dialog will be presented that will capture some information regarding the experimental design. This dialog, pictured below, captures labels (names) for the two factors being studied. By default the factor labels are initially labeled as Factor A and Factor B but one can optionally change these to the actual factor names under study. Buttons also allow one to select how many levels there are of each factor. For instance, if the first factor is strain and there are three strains being studied, select three. The same sort of selections are made for the second factor. The figures below show a 2 x 3 design, 2 strains by 3 growth media conditions. This dialog also permits the entry of meaningful names for each factors level. The initial default labels are Level 1, Level 2, etc..



11.30.7 Two factor Name and Level Name selection dialog.

After specifying factor names, number of levels, and level names, select the *Next* button to open a two factor group selection dialog. Note the experimental design grid displayed in the upper portion of the dialog. If this doesn't reflect the experimental design, select the *Back* button to make new selections. This dialog is used to assign samples to experimental groups represented in the design grid. Select the sample name in the table below by pressing the left mouse button while over the bold rectangle. Hold the mouse button down and drag the sample into the grid above to make the assignment. If the assignment is incorrect, select the *Reset* button next to the sample to clear the assignment. The figure below shows 'Sample 1' being dropped into the design grid. The yellow shading in the grid indicates the cell currently under the mouse. Group assignments can be stored to file and reloaded in subsequent runs of MS. A sample group assignment file is shown in the figure below the group assignment figure.



11.30.8 Two Factor Group Assignment Dialog. Samples are assigned to an experimental group in the design grid.

# Two Factor Assignment File				
# User: braisted Save Date: Mon Mar 26 15:11:41 EST 2007				
#				
Factor A Label:	Strain			
Factor A Levels:	WildType	Mutant A		
Factor B Label:	Growth Media			
Factor B Levels:	Low	Normal	High	
#				
SampleIndex	PrimarySam	Strain	Growth Media	
1	Sample 1	WildType	Low	
2	Sample 2	WildType	Low	
3	Sample 3	WildType	Normal	
4	Sample 4	WildType	Normal	
5	Sample 5	WildType	High	
6	Sample 6	Mutant A	Low	
7	Sample 7	Mutant A	Low	
8	Sample 8	Mutant A	Normal	
9	Sample 9	Mutant A	Normal	
10	Sample 10	Mutant A	High	

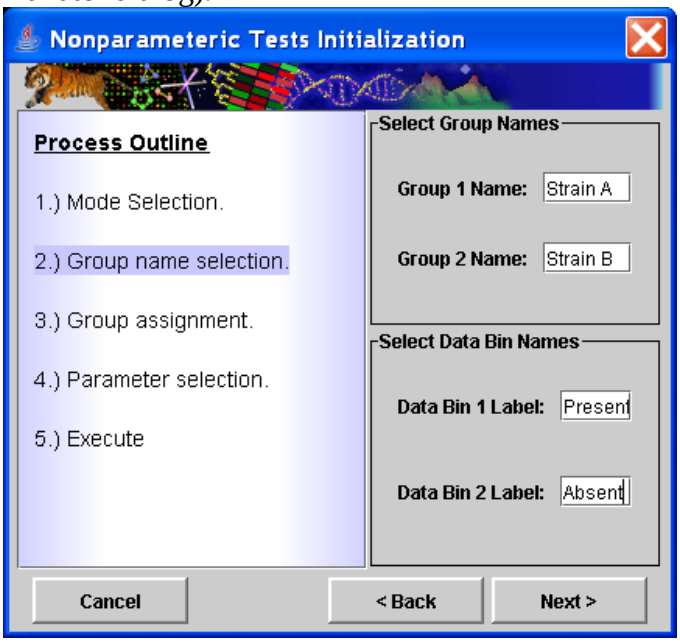
11.30.9 Group assignments can be saved to a file that can be reloaded to supply labels and map the samples to their respective groups.

The parameter selections in the final dialog provide an input field for the alpha value or critical p-value cutoff. The output, as mentioned previously for all NonpaR algorithms are the standard cluster viewers.

Fisher Exact Test

The Fisher Exact Test has some unique dialogs or parameters that must be selected as describe in this section. During execution of the tests, the data is divided first into one of two experimental groups (sets of hybridization results or columns in the expression matrix) and then the data values are partitioned using a numerical cutoff. This partitions the data values for each gene into the contingency matrix described and shown in the overview section on this tests.

The group name section dialog (shown below) includes two extra fields that are used to label the data bins that are partitioned based on the numerical cutoff (specified in a later dialog).



11.30.10 Group name and data bin selection dialog.

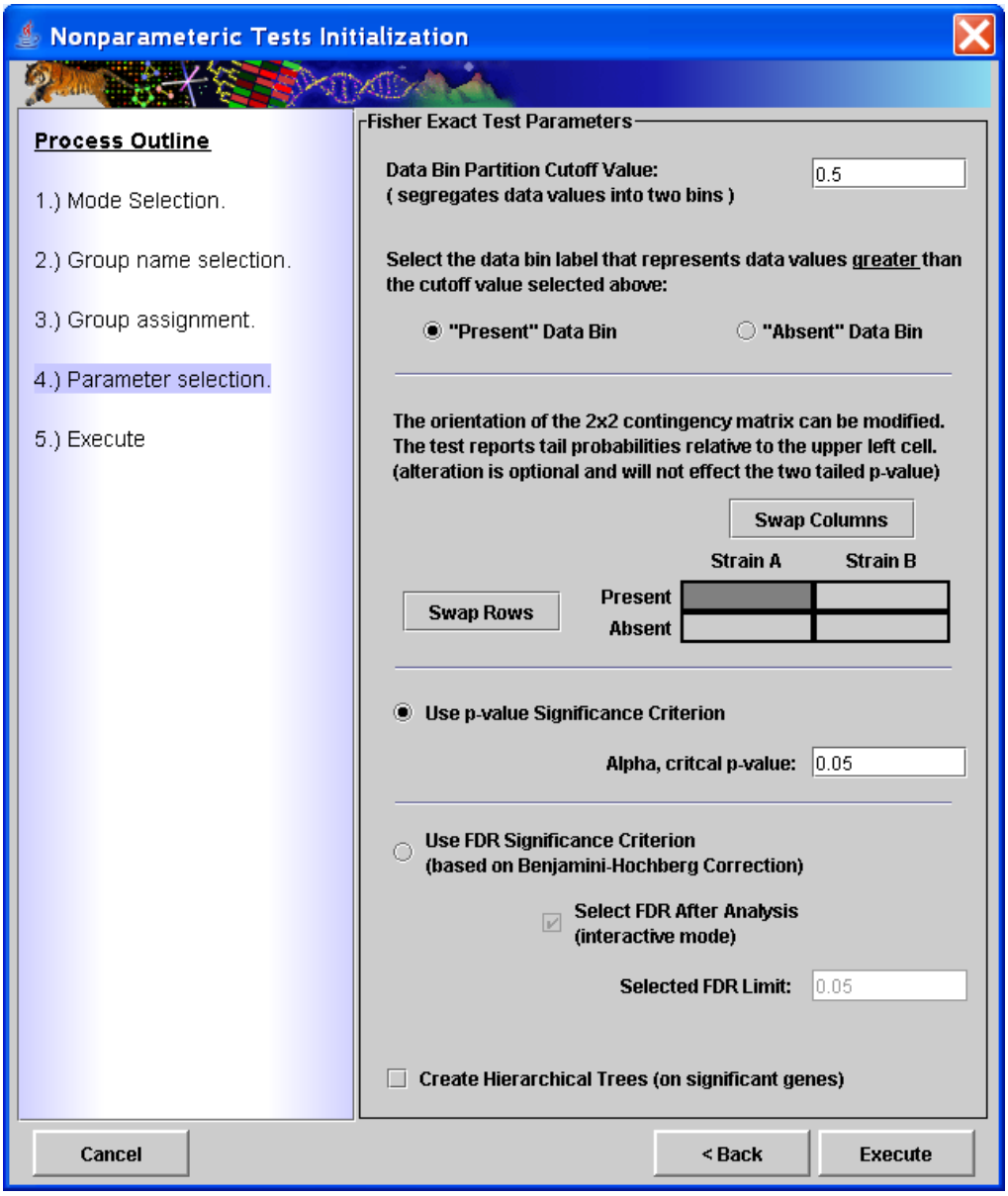
The next dialog to be displayed is used to assign each the loaded samples into one of the two experimental groups by clicking on the assignment buttons to the right of each sample label. This is the same dialog used for sample assignments for the Wilcoxon Rank Sum test shown in an earlier section.

The final dialog captures the data bin cutoff value, the data bin label that should be associated with each data bin, the orientation of the 2x2 contingency matrix, and the various significance criteria options (alpha value or FDR).

The data bin cutoff value is to partition the data based on whether the data falls below or equal or greater than the supplied values. This is most often applied to data were the numerical values are either discrete numerical Present/Absent calls or where it makes sense to partition the data according to a supplied numerical cutoff. A concrete example the application of this test to CGH data, either raw values where a cutoff has been previously determined, or where the data consists of discrete absent/present calls.

The next pair of buttons down on the dialog determine which of the supplied data bin labels should be associated with the data bin representing the values greater than the cutoff value.

The contingency matrix arrangement is displayed and allows the rows and columns to be swapped. Note that for a particular gene, swapping rows and/or columns, simple places the count (number of data values) placed in each matrix cell in a different location in the matrix. The two tailed result, where the test for non-random association between the two factors is the focus, the arrangement of the matrix is not important. The one tailed test will report the disproportionality of the upper left matrix entry. The lower or left tail will indicate when relatively small counts are in the upper left cell while the upper tail or left tail will indicate when relatively large counts are in the upper left cell. The tail probabilities are the same regardless of orientation of the matrix but rather the left and right tail probabilities can be swapped depending on the orientation. Segregation of the significant two tailed results into two groups that are more significant on one tail or the other will illustrate the information gained from the one tailed result and how it relates to the combination of factors falling in the upper right cell.



11.30.11

The output clusters from the Fisher Exact Test are split into significant and non-significant genes lists. The significant gene list also has two additional views, one containing those genes that are significant on the upper tail and another containing genes that are significant toward the lower tail. These tail views show the two directions of disproportionality. Note that one additional table view is created that reports the results of all genes that entered the analysis in one table. This cluster view can be saved to file using the right click 'Save Cluster…' menu option to capture the p-values for all genes being tested.

11.31 BN/LM: Bayesian Networks and Literature Mining

LM: Literature Mining Analysis

(Djebbari and Quackenbush, 2008)

Selection this analysis will display the following window. This window collects all the parameters and inputs required to run the analysis. All the options in this dialog are explained below.

Location of Support File(s): This option allows users to select the location where all support files needed to run BN. A description of files required can be found (see appendix for BN file descriptions).

Network Priors Sources: The checkboxes provide the users to select the source of Bayesian prior probabilities in constructing a seeded network. Currently Literature Mining and KEGG priors are available. The Protein - Protein Interaction as a source of priors is still under development.

As of now, the KEGG support files are automatically downloaded from TM4 website by the application. The user is prompted for Species information if annotation is not available. All other prior sources must be made available.

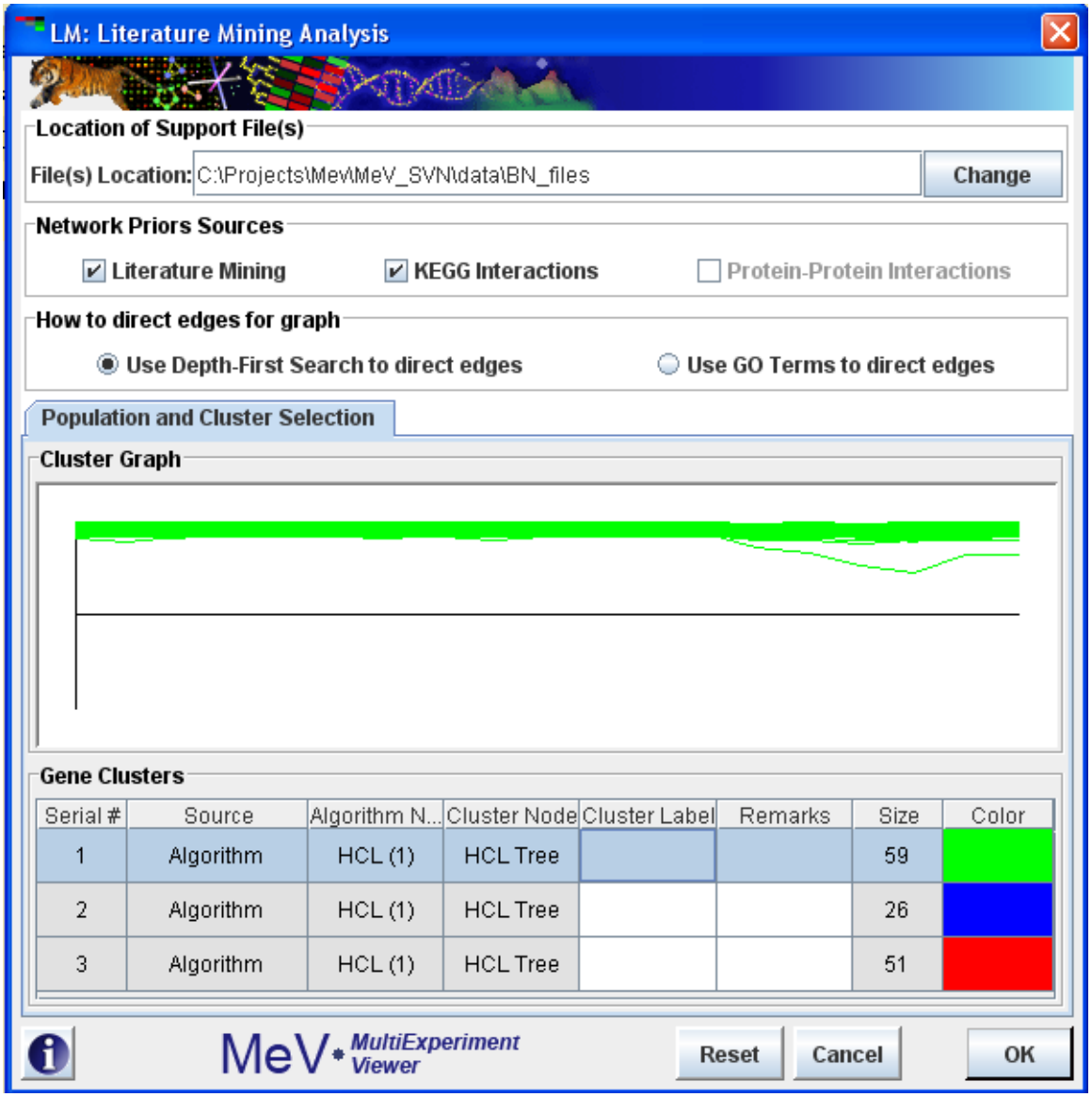
Discretize Expression Values: The data mining algorithm requires that the data be discretized into bins before it can be evaluated for network structure learning. It is strongly recommended that user selects the default value of 3, which means the data can exist in 3 states:

Under expressed
Over expressed
Unchanged

The algorithm functions and reports meaningfully if the 3 state rule is followed.

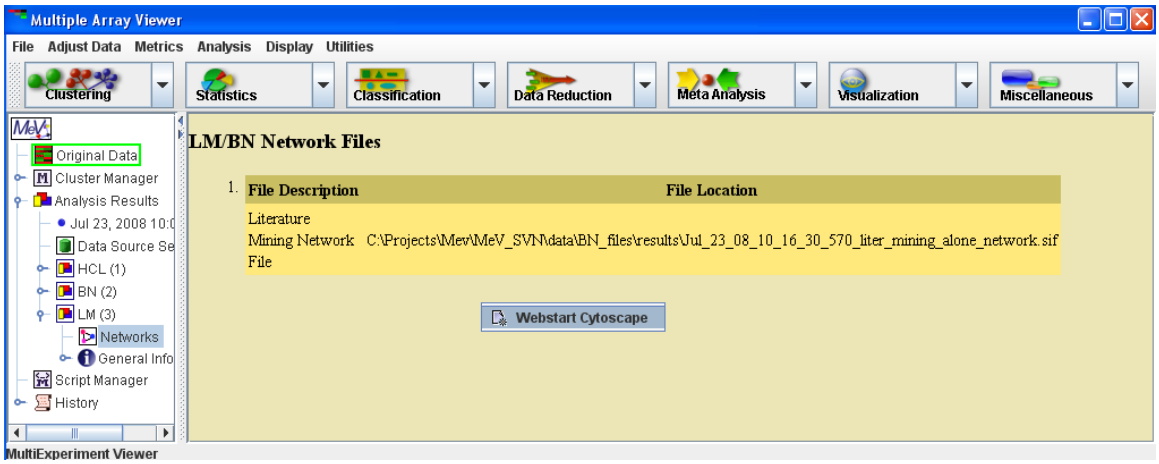
How to direct Edges for graph: The algorithm uses DFS or Depth First Search to connect nodes in the initial seeded network. For large networks with lots of nodes this can take a while to complete. The GO Term option of directing edges is not yet fully developed.

Gene Cluster: The input cluster on which the analysis runs. If the user has saved more than one cluster, one of those clusters must be selected from the clusters panel. By default the first is selected.



11.31.1

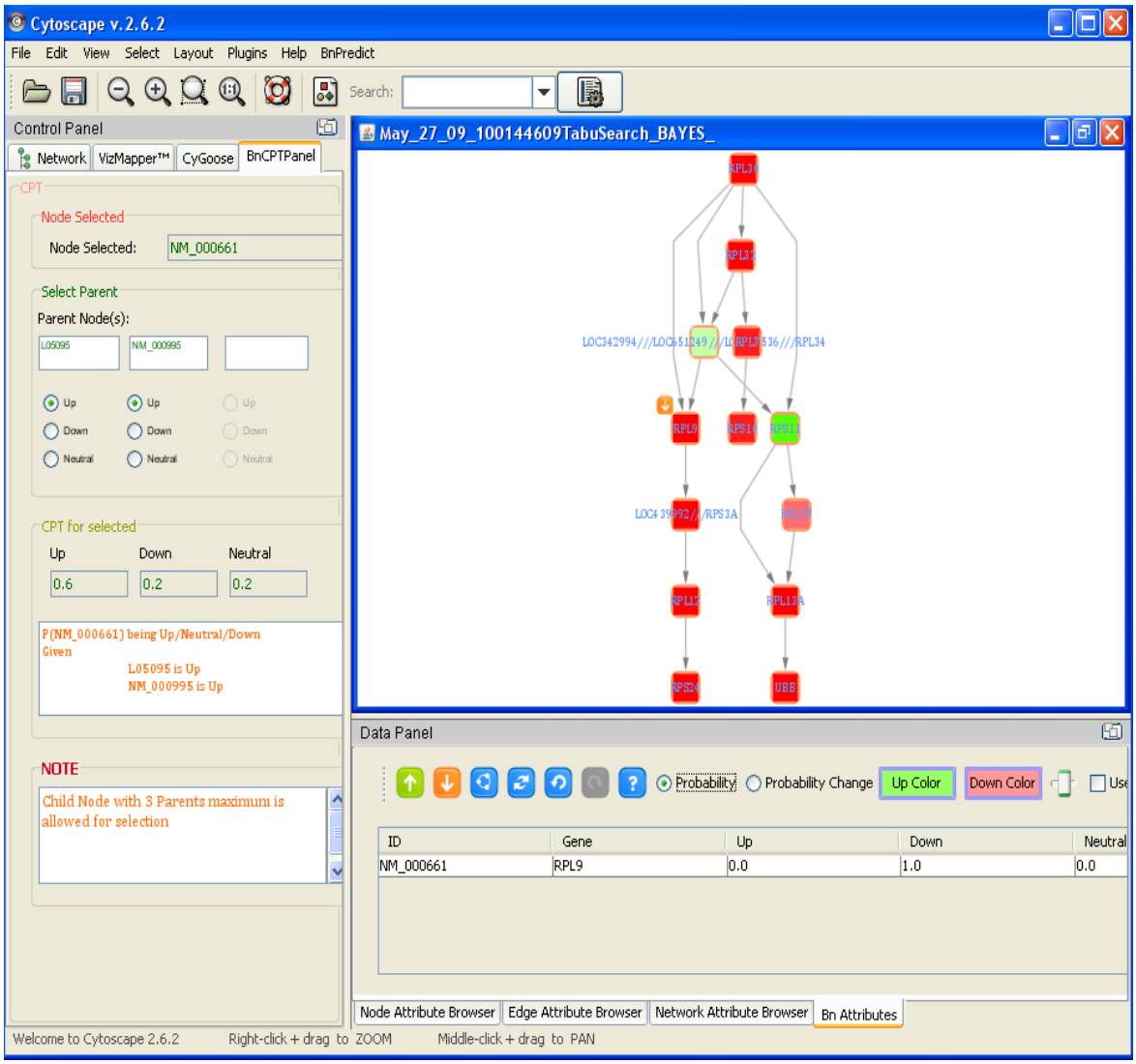
LM Viewer: The LM viewer displays the network file names and locations that were created during the analysis. On Right Click it displays a popup menu to launch Cytoscape via Webstart with all files created.



Results:

All result files are stored in the ***{..}\results folder***. Where {..} represents the directory where the BN/LIT analysis supporting files are located.

Once the analysis completes it automatically launches Cytoscape (http://www.cytoscape.org/) via Webstart with the result networks displayed in it.



11.31.2

<u>BN: Bayesian Network Analysis:</u>

(Djebbari and Quackenbush, 2008)

Selection this analysis will display the following window. This window collects all the parameters and inputs required to run the analysis. All the options in this dialog are explained below.

Location of Support File(s): This option allows users to select the location where all support files needed to run BN. A description of files required can be found (see appendix for BN file descriptions).

Network Seed: If this option is selected the user is expected to provide a file representing network. The file should contain a list of edges, one of each line, and the nodes are separated by a tab. The identifier of the node should be one of the following provided in the drop down identified as "Select Seed UID". Directionality is assumed in each edge specified such that node_A tab node_B is read as node_A to node_B. Cycles are not allowed.

A network seed can also be built from using the "Create Network Seed" button. It allows user to create list of edges selecting nodes from the data directly. This option is limited in feature.

Network seed can be used in one of the three ways:

1. Using the user network seed alone and bypassing literature based network seeding altogether
2. Using the user network seed along with Literature mining seed. In this case conflicts (A to B in user seed vs. B to A in Lit mining) in directions are resolved by giving precedence to used provided network seed.
3. User provided network is used as a complete network and the network structure is not learned, only the Conditional Probability Tables (CPTs) associated with the network is learned for downstream exploration.

Network Priors Sources: The checkboxes provide the users to select the source of Bayesian prior probabilities in constructing a seeded network. Currently Literature Mining and KEGG priors are available. The Protein - Protein Interaction as a source of priors is still under development.

As of now, the KEGG support files are automatically downloaded from TN4 website by the application. The user is prompted for Species information if annotation is not available. All other prior sources must be made available.

Discretize Expression Values: The data mining algorithm requires that the data be discretized into bins before it can be evaluated for network structure learning. It is strongly recommended that user selects the default value of 3, which means the data can exist in 3 states:

Under expressed
Over expressed
Unchanged

The algorithm functions and reports meaningfully if the 3 state rule is followed.

Sample Classification:

The samples or experiments can be classified based on some knowledge that the user might have and the user might want to preserve the classification while learning the network structure. In that case the user might want specify a numerical value > 1 denoting the number of groups the samples belongs to. The default is 1 or all belonging to one class. The same also is true for large sample

size but the user is strongly recommended not to exceed more than 2 or 3 groups. The samples classification can be saved and loaded in to a file. See the label file as described (here).

Note that the user is presented with a Classification Dialog where samples can be assigned to group of users' choice. The sample classification dialog shows up once the user navigates from the main dialog by hitting OK. If number of classes is chosen as one the classification dialog is not shown. A node by the name of "CLASS" shows up in the network which captures the effect of sample groups on the network. Once the network is displayed the "CLASS" node behaves and can be treated as any other node in the network. The CLASS node has no annotation.

How to direct Edges for graph: The algorithm uses DFS or Depth First Search to connect nodes in the initial seeded network. For large networks with lots of nodes this can take a while to complete. The GO Term option of directing edges is not yet fully developed.

Bootstrapping Parameters:
The user has the option of bootstrapping the samples to generate random networks. This feature is optional. This panel allows the user provide the number of time random samples will be generated in the '**Number of Iterations**' box. The '**Confidence Threshold**' box allows defining a confidence level cut-off. The default is 0.7 means the algorithm will select an edge if it appears in 70% of the bootstrap networks.

Note that if bootstrap is chosen, the user is given a chance to play with diffrent cut-offs after the algorithm runs via a small dialog box. It creates new networks for each new threshold that can be viewed in Cytoscape via Gaggle broadcast.

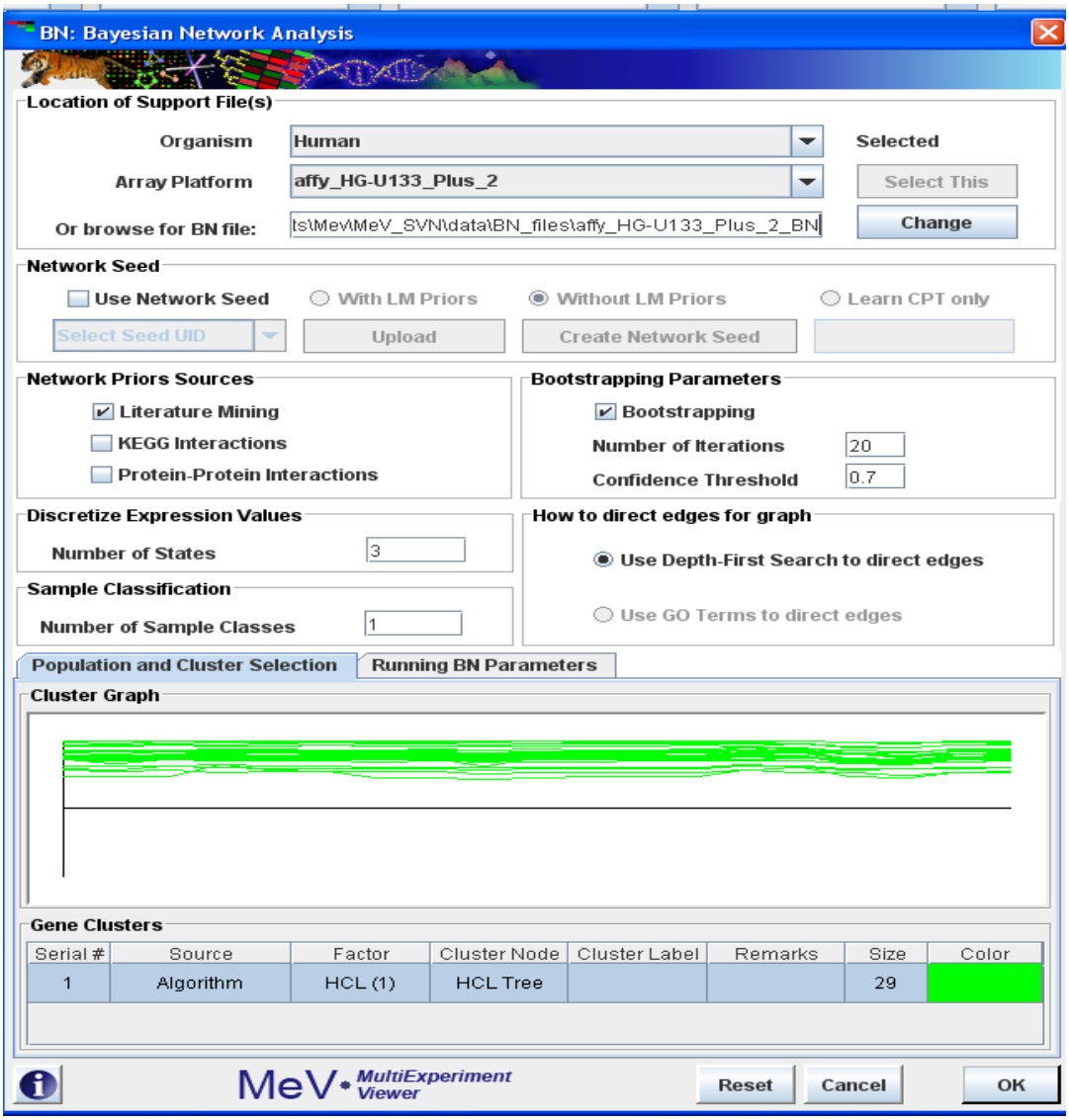
Population and Cluster Selection: The user has to choose a cluster that BN algorithm would use to run the analysis. By default the first cluster is highlighted.

Note, that there is max limit in terms of number of genes that this algorithm can handle. If a cluster is chosen that exceeds the maximum genes limit an error window is displayed showing the maximum allowable number of genes. At this point the user can choose a new cluster, if one is already defined and is below the limit. If a cluster of allowable size is not defined, the user needs to cancel out of BN window, create new cluster(s) and then launch the BN Analysis window again.

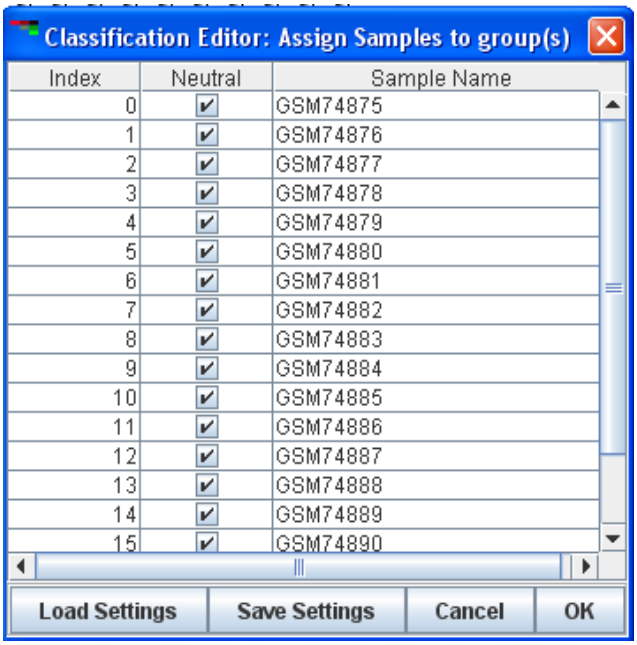
Running BN Parameters: This tab allows the user to customize some advanced options of the algorithm. Most users would be OK to accept the default settings in this panel. Below is a concise description of each available option:

> **Search Algorithm** - The algorithm to search for best network
> **Scoring Scheme** - The scoring mechanism to choose from top networks
> **Max. Number of Parents** - Maximum number of parents each network node may have

Cross Validation Folds (K) - In absence of a training dataset how many cross validation(s) are needed

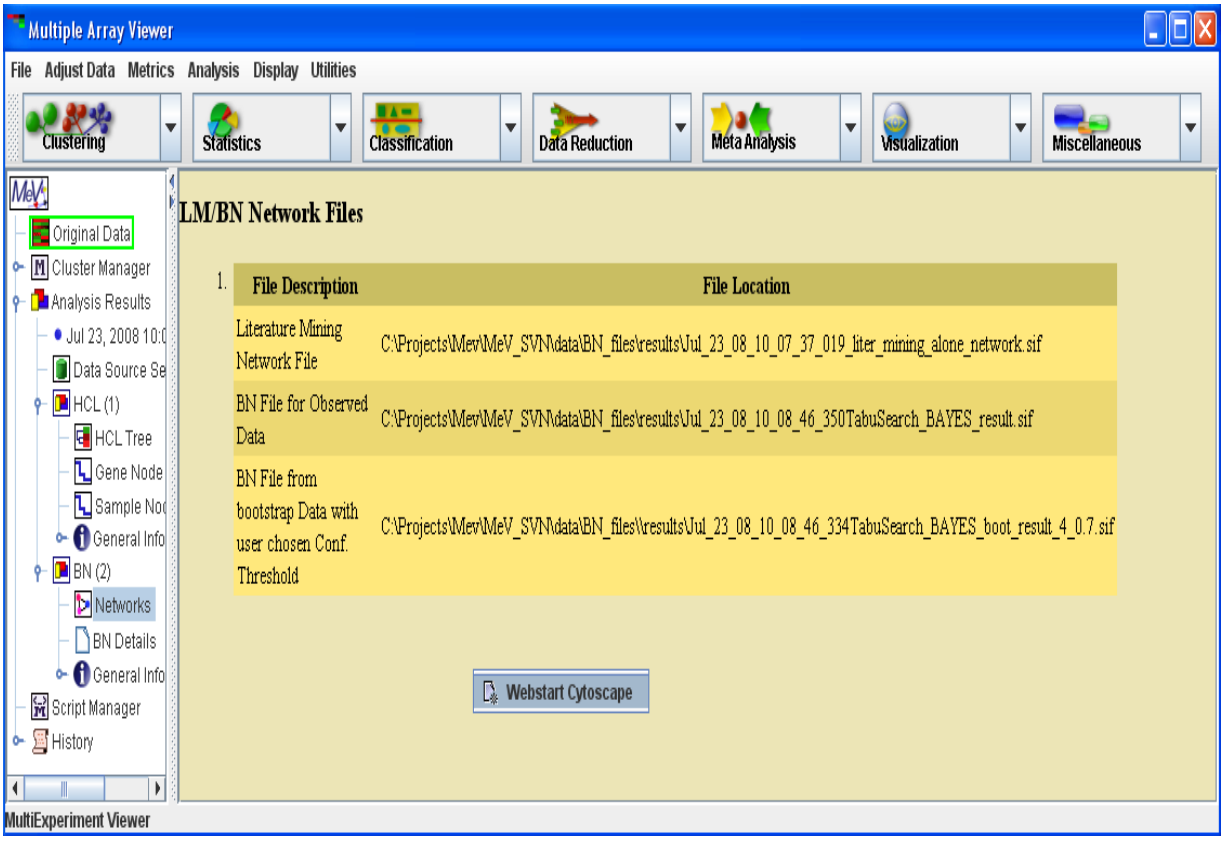


11.31.3



11.31.4

BN Viewer: The BN viewer displays the network file names and locations that were created during the analysis. It also shows the 'final' network that the user selected from the bootstrapped networks. On Right Click it displays a popup menu to launch Cytoscape via Webstart with all files created.



BN Results:

The following result files are stored in the ***{..}\results*** *folder*. Where {..} represents the directory where the BN/LIT analysis supporting files are located.

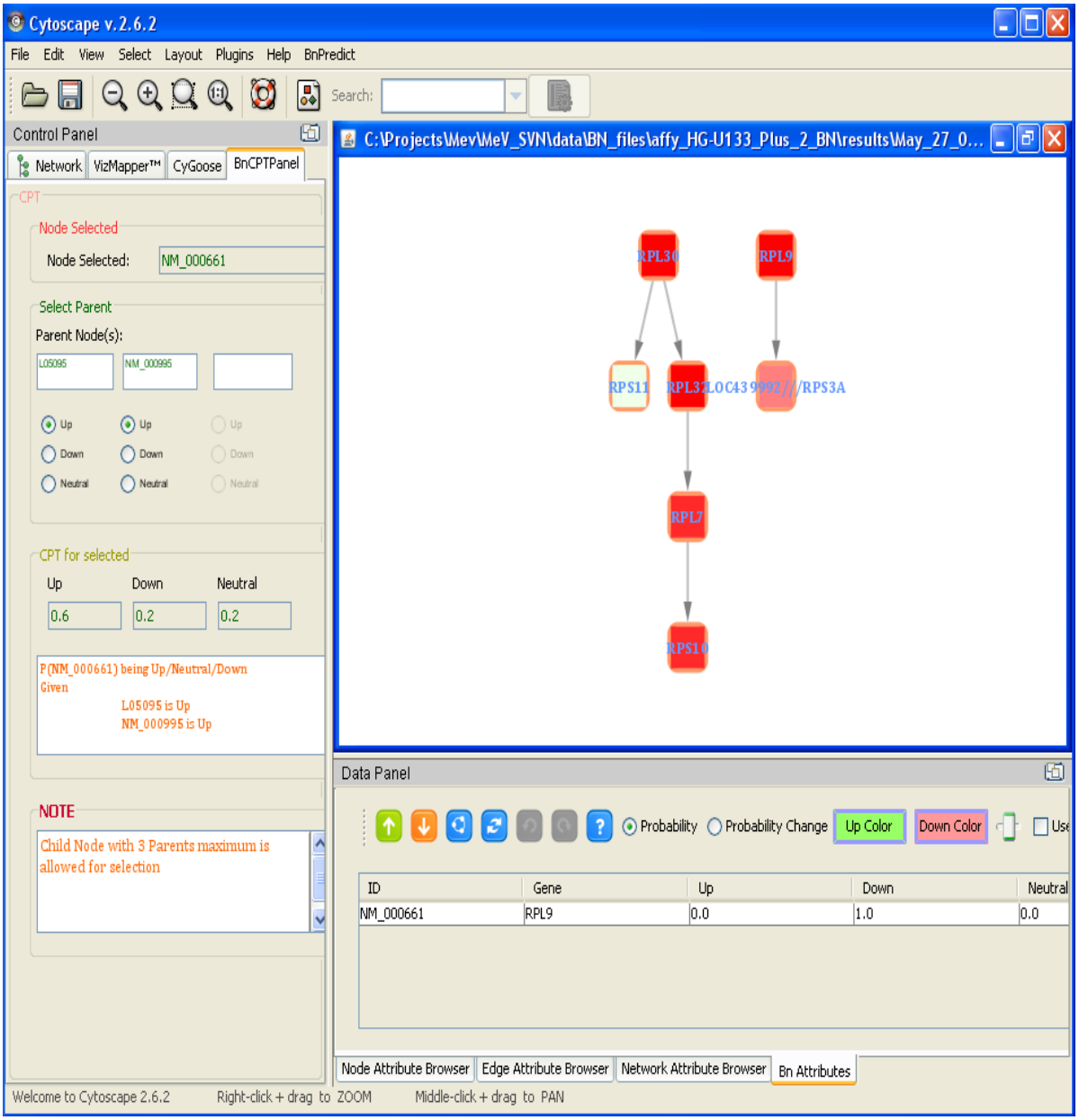
1. Literature Mining Network
2. Observed Network with Priors
3. Network from the bootstrap networks for the default confidence threshold of 0.7 (if bootstrap was chosen)
4. The 'final' network from the bootstrap networks for the confidence threshold of users' choice (if bootstrap was chosen).

Once the analysis completes the following small window is presented. The buttons/options are explained below:

11.31.5

"***Network from Bootstrap***" button: This button shows up only if bootstrapping option was selected as an option in the initial BN parameter window. This shows the bootstrapped resulted network with the confidence threshold as set in the initial BN parameter window. On clicking this button the resulting network is shown in Cytoscape. If Cytoscape is already open it adds another network view or else it launches the application. The result network window will have a title with the following format: ***DateTimeStamp_"SearchAlgorithm"_boot_result_"ScoringScheme"_"#ofBootstraps"_"confidenceThreshold".sif***

"***Update Network"*** button: The bootstrapped network can be viewed for different confidence threshold (70%, 80%, 95% etc) by changing the value in the text box preceding the button. The value should be a float where 0.7 means 70% confidence, 0.95 means 95% confidence. Once the desired threshold is entered, by clicking the "Update Network" button the resulting network can be viewed in Cytoscape. **Note:** It does *not update* previously existing network at a different threshold but adds a *new* network with the specified threshold. The result network window will have a title with the following format: ***DateTimeStamp_"Search Algorithm"_boot_result_"ScoringScheme"_"#ofBootstraps"_"confidenceThreshold".sif***


Cytoscape v.2.6.2
File Edit View Select Layout Plugins Help BnPredict
Search:
Control Panel
Network VizMapper™ CyGoose BnCPTPanel
CPT
Node Selected
Node Selected: NM_000661
Select Parent
Parent Node(s):
L05095
NM_000995
Up Down Neutral
CPT for selected
Up Down Neutral
0.6 0.2 0.2
P(NM_000661) being Up/Neutral/Down
Given
L05095 is Up
NM_000995 is Up
NOTE
Child Node with 3 Parents maximum is allowed for selection
C:\Projects\Mev\MeV_SVN\data\BN_files\affy_HG-U133_Plus_2_BN\results\May_27_0...
RPL30
RPL9
RPS11
RPL3
LOC439992
RPS3A
RPL7
RPS10
Data Panel
Probability
Probability Change
Up Color
Down Color
ID Gene Up Down Neutral
NM_000661 RPL9 0.0 1.0 0.0
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Bn Attributes
Welcome to Cytoscape 2.6.2
Right-click + drag to ZOOM
Middle-click + drag to PAN

11.31.6

11.32 Gene Set Enrichment Analysis

(Zhen Jiang and Robert Gentleman, Bioinformatics, 2007)

What is GSEA?

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).
Ref: *(http://www.broad.mit.edu/gsea/)*

Why GSEA?

Traditional statistics use adjusted P-values with some arbitrary cutoff, treating genes with slightly different P-values as different entities. Also, small differences in mRNA abundance are often not detected, nor are large changes in just a few genes.

GSEA remedies this by using all the genes in your expression data for the analysis. GSEA also compiles per-gene statistics across genes within a gene set, allowing for the detection of small changes in many genes or large changes in few genes.

GSEA algorithm implemented in MeV v4.3 is based on Zhen Jiang and Robert Gentleman's 2007 *Bioinformatics* paper ***(Jiang, Z., Gentleman, R., (2007). Bioinformatics. 2007 Feb 1; 23(3):306-13. Extensions to gene set enrichment analysis)***

Brief description of the algorithm

GSEA algorithm can be roughly divided in to three steps:

1. Calculate the per gene statistic. This is done by fitting a linear model to all the genes, separately and simultaneously.
2. Calculate the gene set statistic.
3. Estimate significance by
    - Permuting factor/phenotype/class labels
    - Compute the per gene statistic for every permutation
    - Compute the gene set statistic for every permutation
    - Calculate and report Unadjusted p-values

How to run GSEA?

GSEA uses a set of parameter input dialogs that open sequentially to provide input options that correspond to each step of the process. The first step in the processes is data selection which lets you assign phenotype/class labels to your samples.

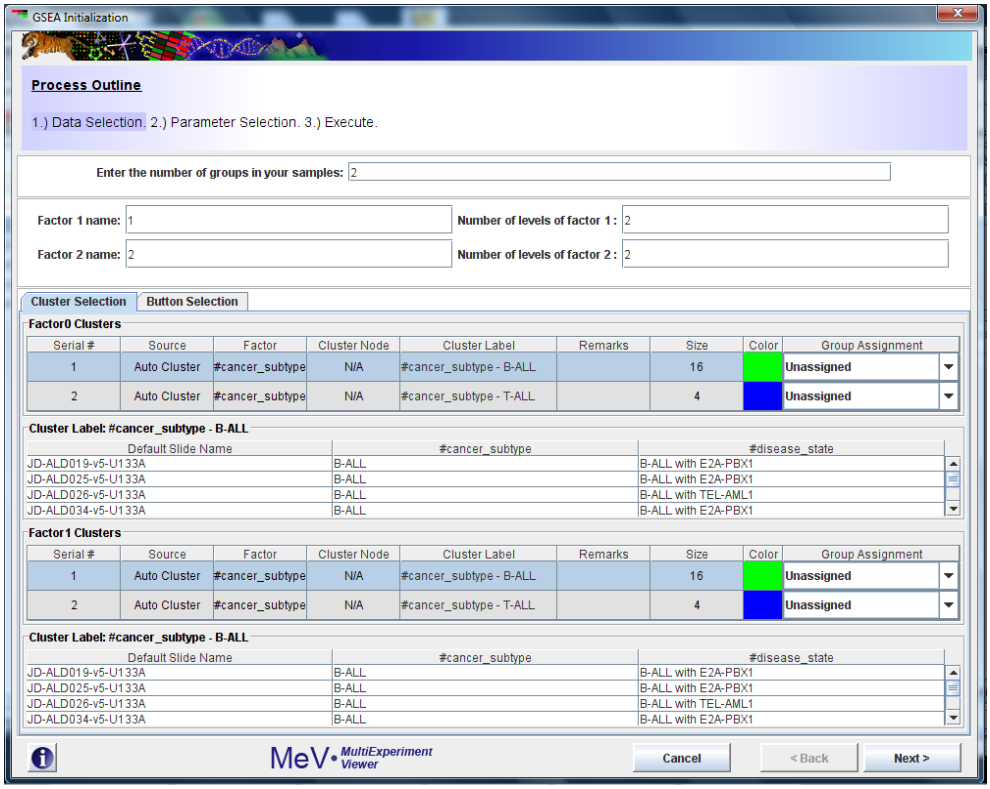


Figure 1

The default assignment (Figure 1)is two groups (factors) with two levels per group. This can be changed to reflect the groups present in your data.

For example, if cancer subtype is the phenotype (factors) that influences your data the most, enter 1 in the "*Enter number of groups in your sample*" textbox.

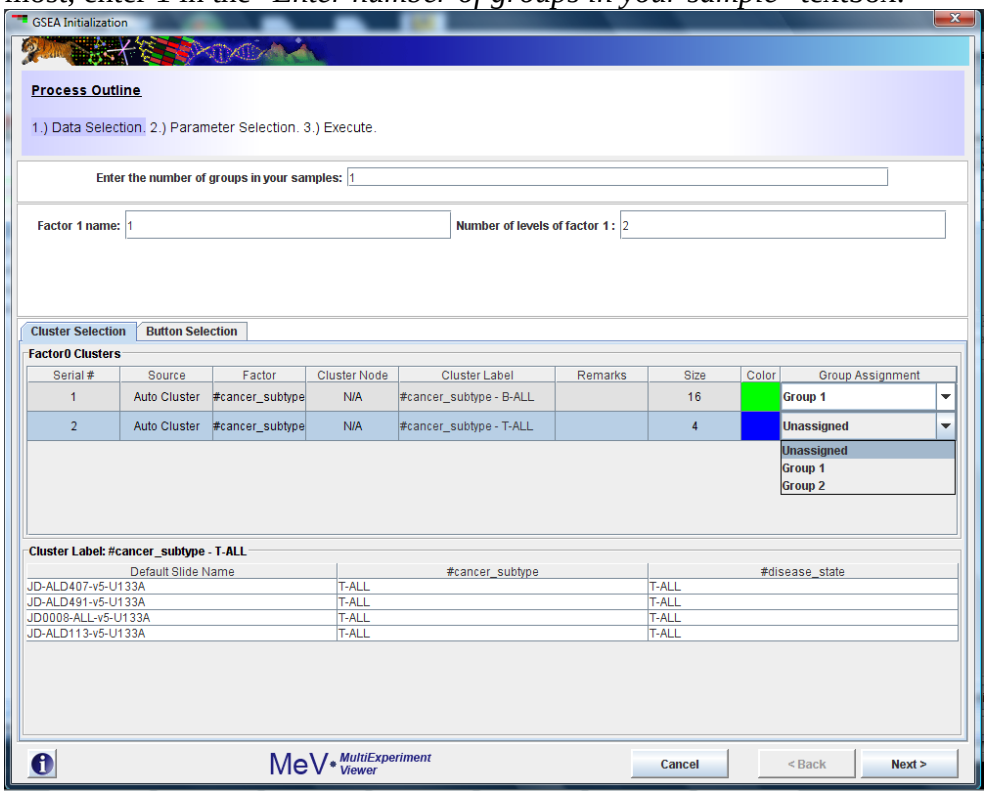


Figure 2

B-ALL and T-ALL are the two levels of this phenotype, so enter 2 in the "*Number of levels of factor*" textbox. If you have pre selected sample clusters and decide to use the "*Cluster Selection*" tab just assign group numbers using the Group Assignment drop down as shown in Figure 2.

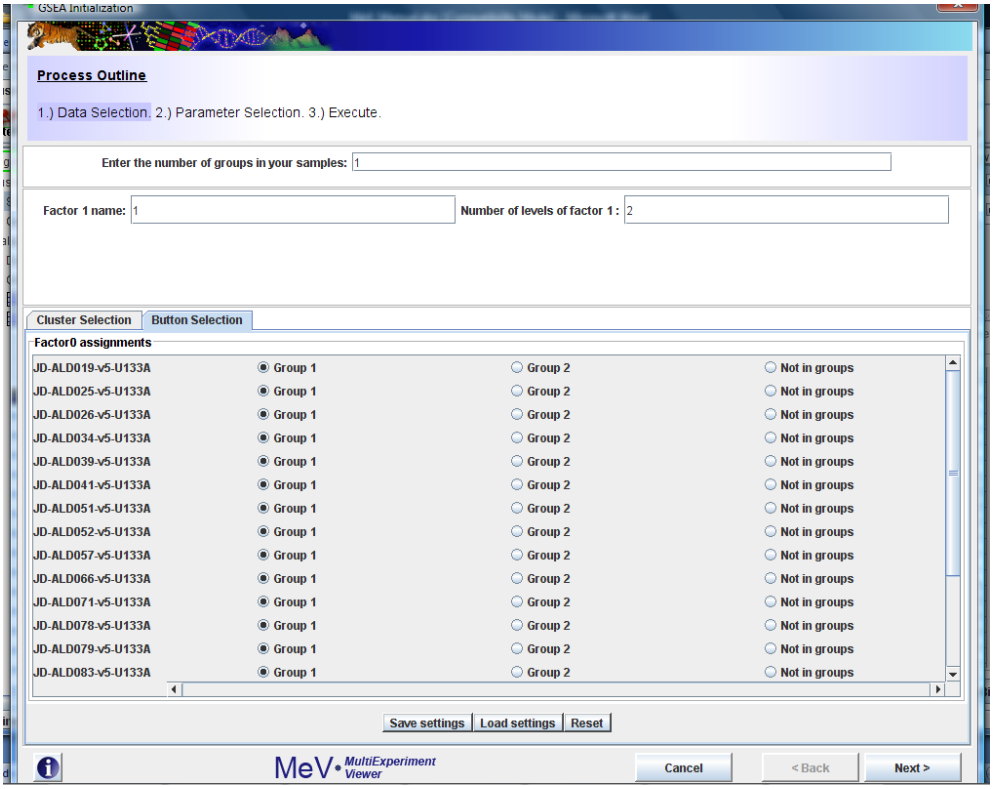


Figure 3

You can also use the "*Button Selection*" tab shown in Figure 3 to achieve the same. Group1 and Group2 symbolize B-ALL and T-ALL. You can save these grouping using the "Save settings" button. To load saved groupings, use the "Load settings" button. Reset button will clear all your choices. Once you are done, click the Next > button.

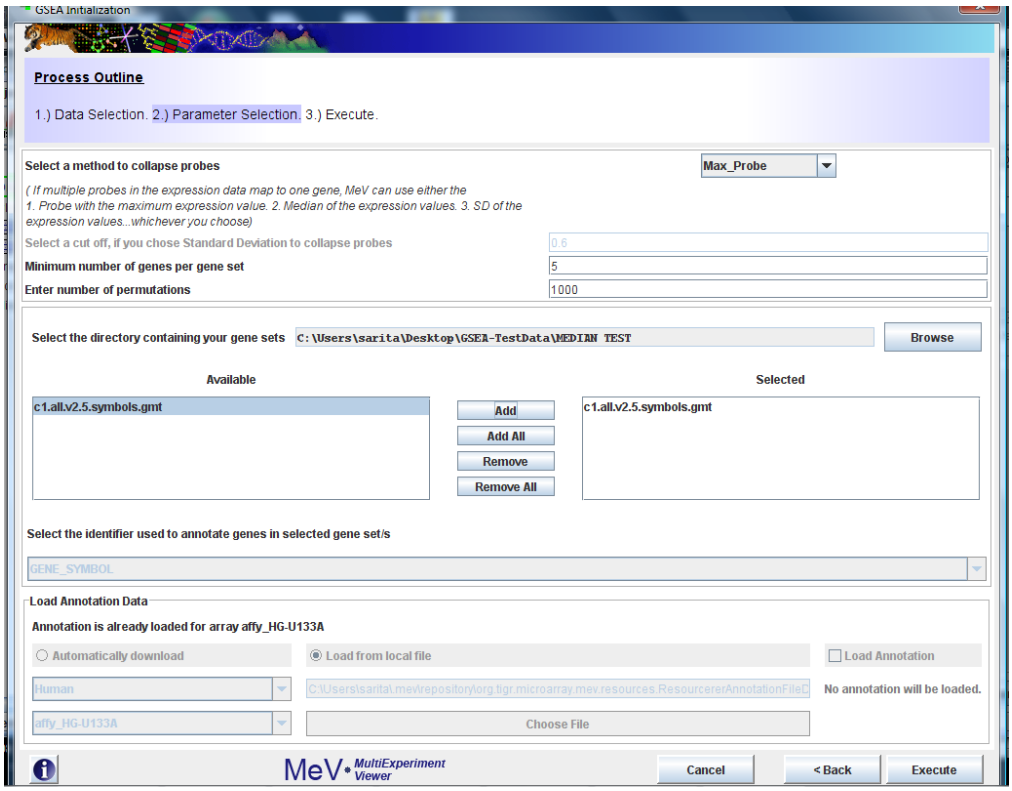


Figure 4

This brings you to the "Parameter Selection" section shown in Figure 4. The GUI is pretty self explanatory, but here is some better clarification about the available methods for collapsing probes to genes, loading gene sets and loading annotations.

```
  MAX_PROBE:
            Sample1           Sample2           Sample3
  Probe_1   10                20                30
  Probe_2   20                5                 10
  Gene_12   20                20                30


  MEDIAN_PROBE:
            Sample1           Sample2           Sample3
  Probe_1   10                20                30
  Probe_2   20                15                10
  Probe_3   20                5                 10
  Gene_123  20                15                10

  STANDARD_DEVIATION:
            Sample1           Sample2           Sample3          SD
  Probe_1   10                20                30                 3
  Probe_2   20                5                 10                 4
  Probe_3   20                15                10                 2
  Gene_123 will be represented by Probe_2, which has the MAX SD value
across samples. So,
  Gene_123   20   5     10
NOTE: SD will be calculated by MeV on the fly. You do not have to do
anything. This is just an example.
```

The '*Browse*' button corresponding to "*Select the directory containing your gene sets*" lets you choose the directory containing gene set files. You can select the files you want to use from the "*Available*" panel. "*Selected"* panel indicates the gene set files that you chose to use for this analysis. In addition to this gene sets can also be downloaded from the MIT/Broad website http://www.broad.mit.edu/gsea/msigdb/downloads.jsp

If your gene set file is *.gmt or *.gmx format, "*Select the identifier used to annotate genes in your selected gene set*" drop down is automatically populated with "GENE_SYMBOL" as shown in figure above. In case of a custom gene set file, you must manually choose the gene identifier from the drop down.

The "*Load Annotation Data*" panel lets you upload annotations. Annotations are a MUST for running GSEA. Details on how to load annotations is described in Using the Annotation Feature

The last step is to hit the Execute button. GSEA outputs besides the standard MeV viewers three new viewers namely "*Test Statistic Graph*", "*Leading Edge Graph Viewer*" and "*Geneset p-value graph"*. "*Significant Gene Sets*" under "*Table Views*" lists gene sets sorted by their Over enriched (upper p values). Upper p values are the probability of seeing a test statistic higher than the observed one. Lower p values are the probability of seeing a test statistic lower than the observed one.

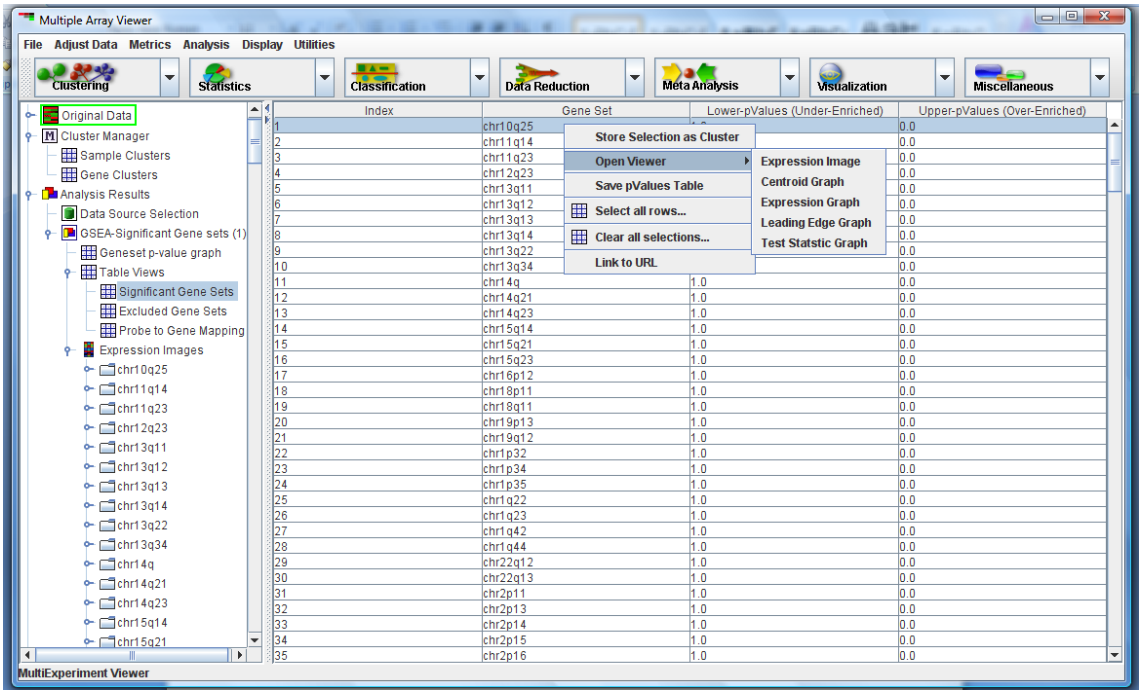


Figure 5

Right clicking on the rows in the table as shown in Figure 5 lets you navigate to different viewers. "*Excluded Gene Sets*" table contains gene sets which do not meet the minimum genes per gene set criteria and hence not included in analysis. "*Probe to Gene Mapping*" table shows all the probes which map to a gene.

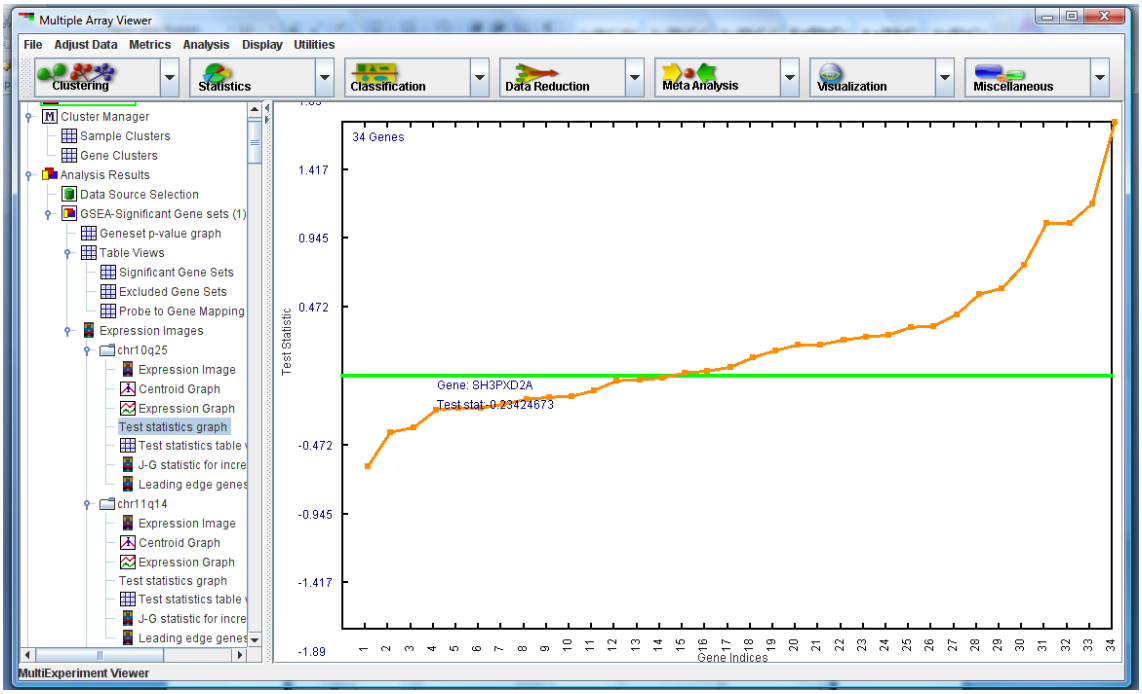


Figure 6

"*Test Statistic Graph*" shown in Figure 6 aims to show how genes within a gene set contribute to the overall gene-set-level metric. This metric is computed by summing the distance from the green line to the orange point and then normalizing this sum by the square root of the number of genes in the gene set.

.

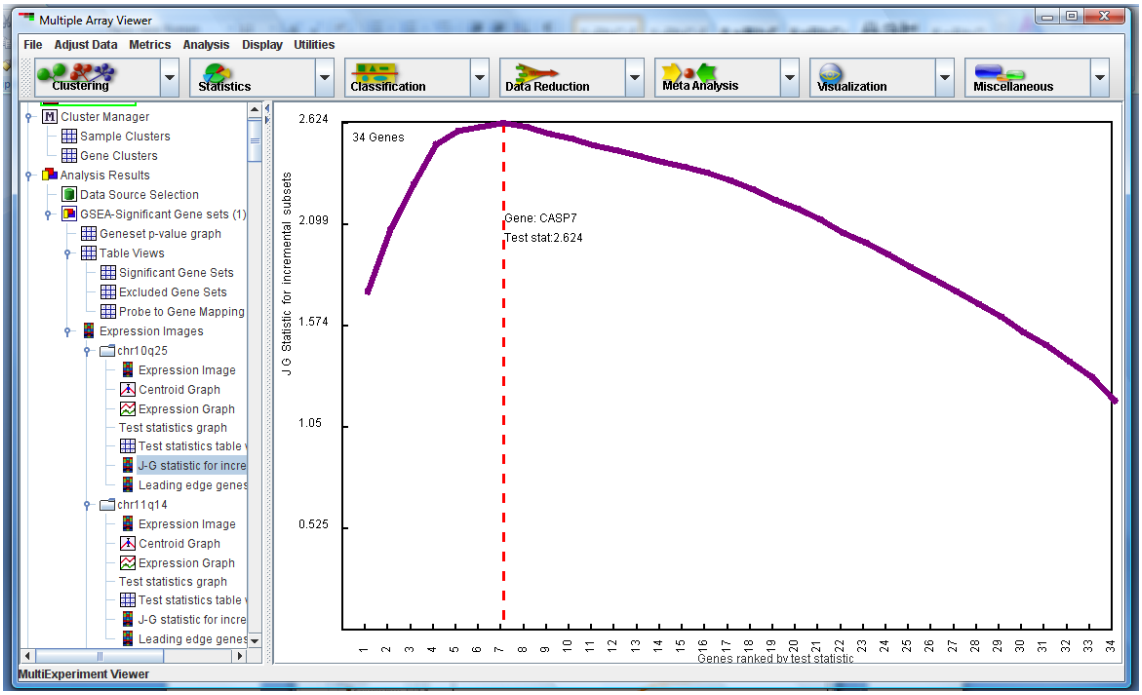


Figure 7

"*Leading Edge Graph***"** in Figure 7 shows which subset of genes within the gene set is contributing to the significance of the gene set level metric. The leading edge subset is calculated by first ranking the genes based on largest to smallest test statistics. We then calculate the Jiang-Gentleman statistic for subsets of the gene set, starting with the first subset containing the gene with the largest t-statistic, and then incrementing the subset to include the next gene with the next largest t-statistic. We iterate through until the final subset contains all the genes in the gene set. The subset which maximizes the Jiang-Gentleman statistic suggests that this group of genes contribute the most to the gene-set level metric

11.33 Bayesian Estimation of Temporal Regulation

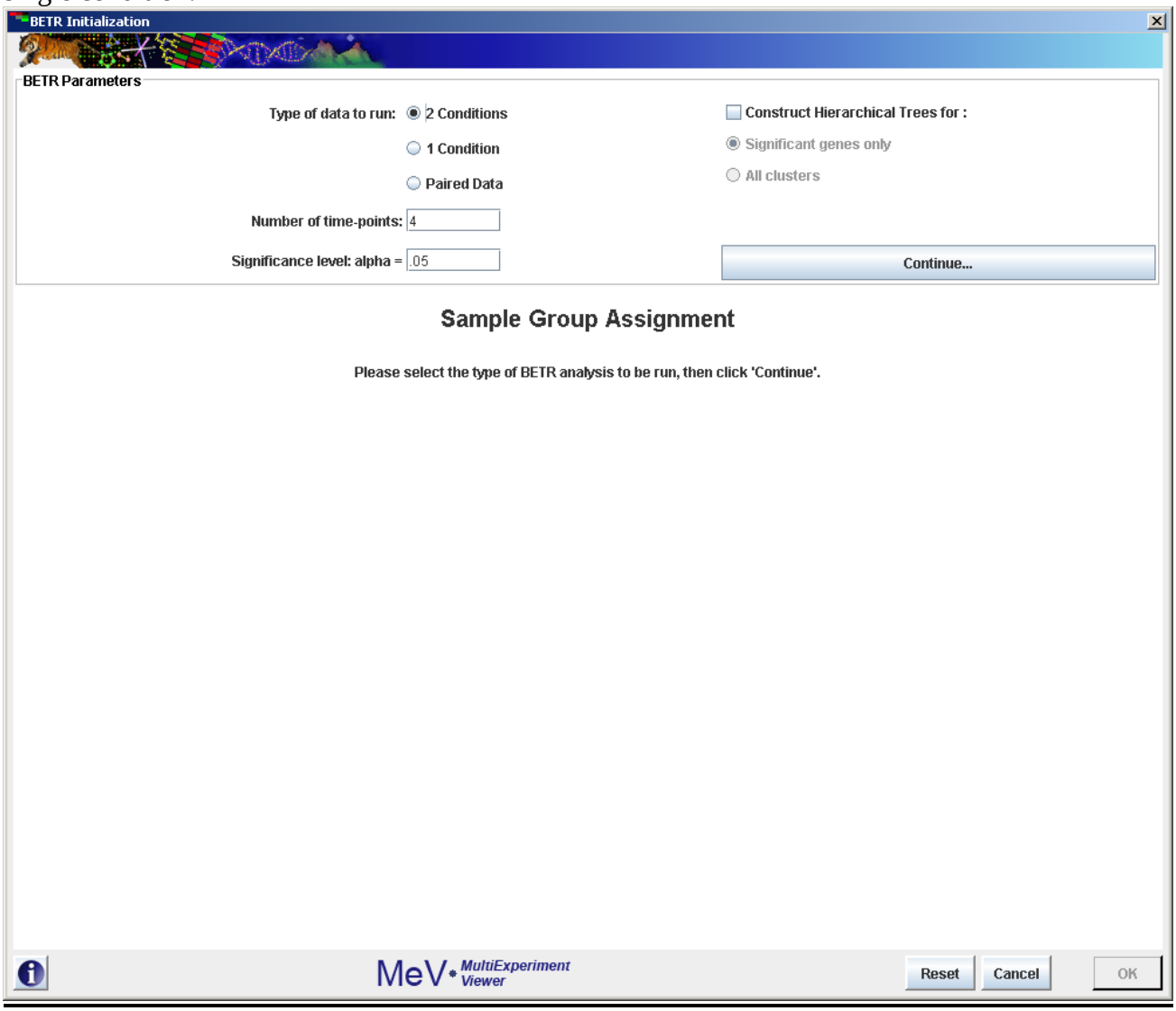
(Aryee, Martin et al , submitted 2008)

Most analyses treat time-points as independent samples and ignore potentially useful information that can be gained if the data contains non-uniform serial correlations. BETR is a flexible linear random-effects modeling framework that takes into account correlations between samples and sampling times. The method makes use of information intrinsic to the data by modeling the joint distribution across all samples.

Each gene is given a probability of differential expression that is derived from an empirical Bayes approach that uses the whole data set to reduce the number of parameters to be estimated.

The BETR method can be applied to one-color or two-color microarray data. The user may also select one-condition which will find differences between a baseline, t = 0, in a

single condition.



Running BETR:
After opening the BETR initialization dialog, select the type of data you intend to run. Your 3 options are:
1.) 2-Condition: Each condition must have multiple replicates for each time-point.
2.) 1-Condition: Only one condition, the initial time-point will be treated as the control.
3.) Paired data: Data exists as a difference between two-conditions.

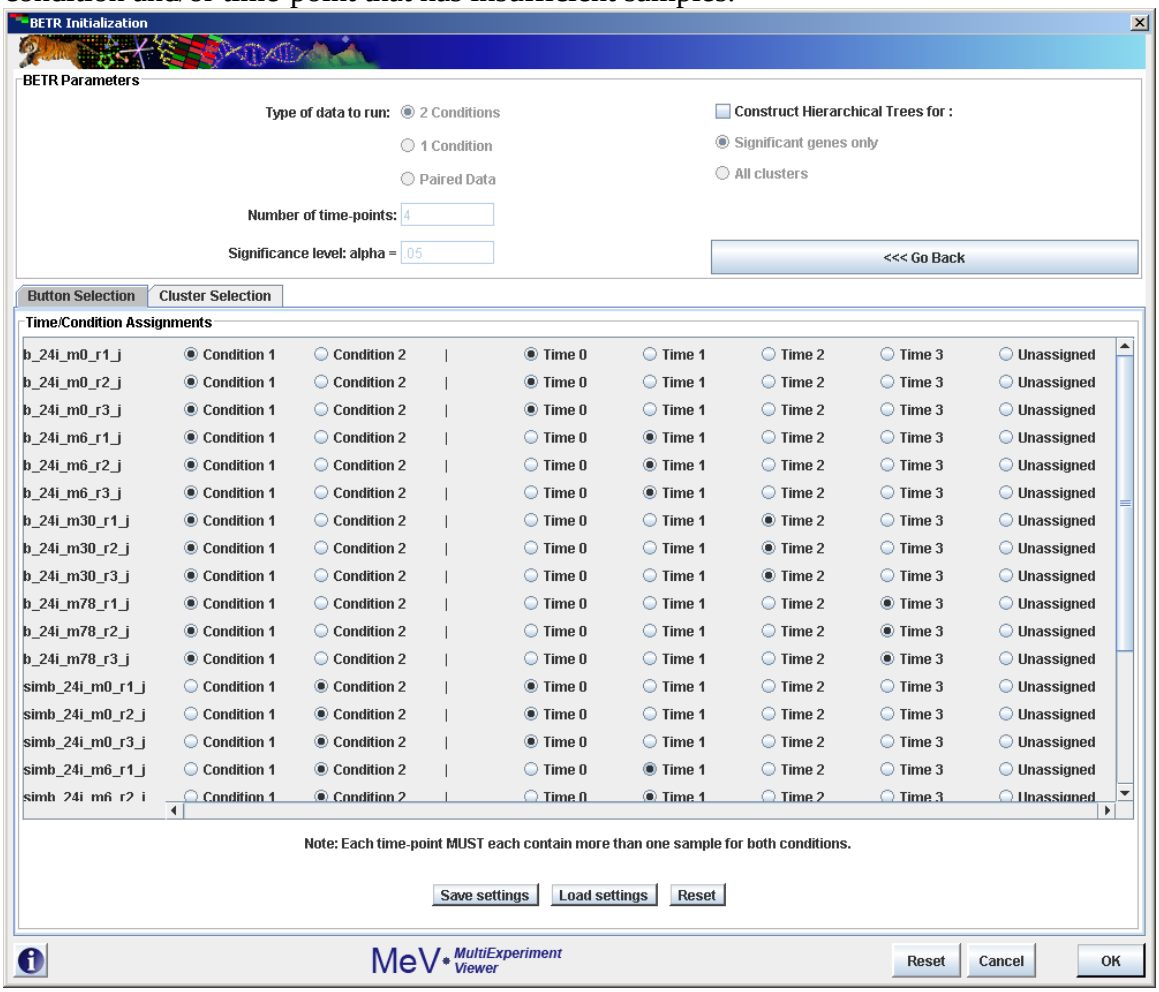
Input the number of time-points that your data contains. If you are running 1-Condition data, the control is treated as the first time-point.

Assign a significance level. Genes assigned a significance greater than 1-alpha will be deemed significant.

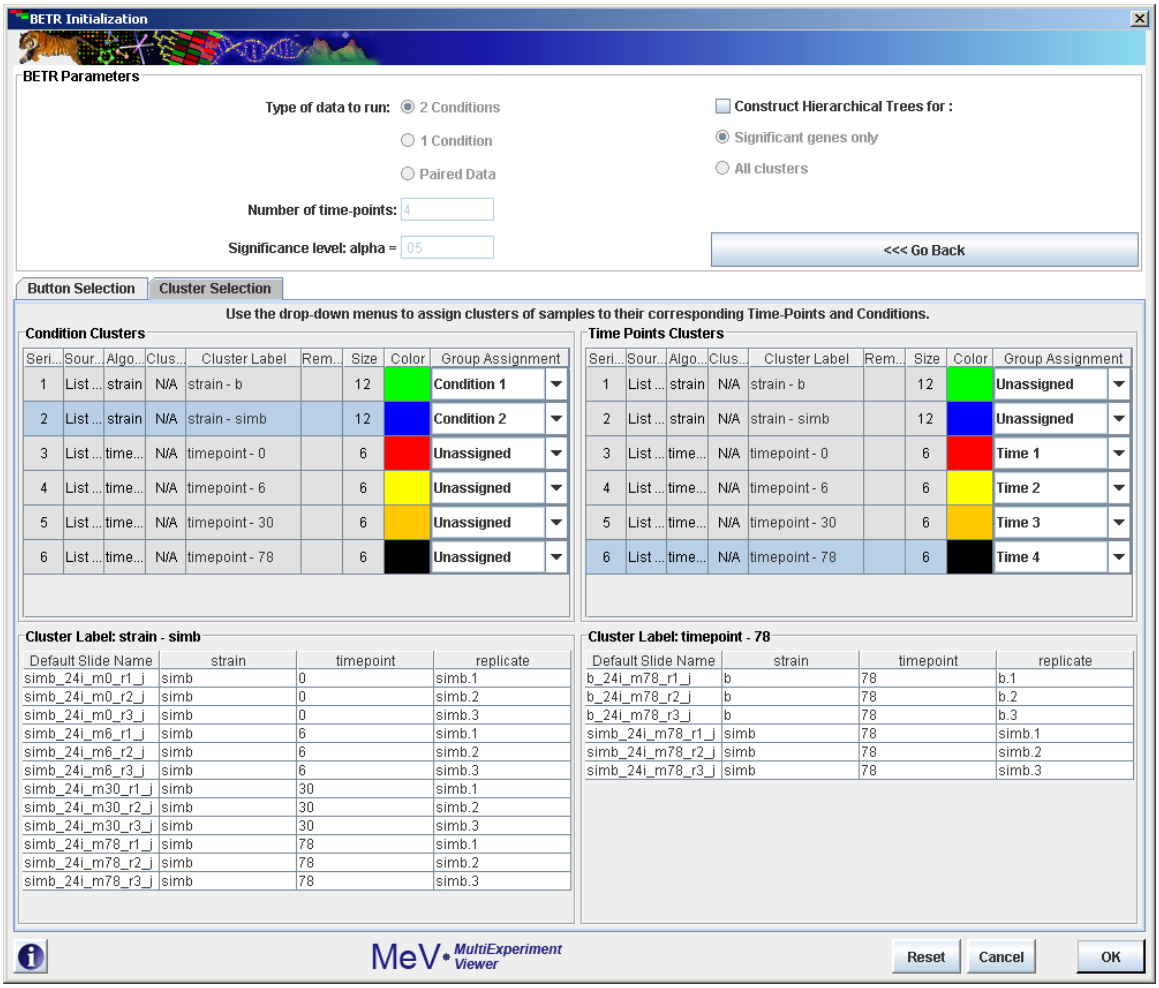
Click “Continue” to move to the next step.

Assign your samples:
Depending on the type of data you have selected, you will need to assign your samples to time-points and conditions, if necessary.

Button Selection:
Use this option if you have not created clusters for your samples and want to assign each sample individually.
Each time-point must contain at least two samples assigned to it.
For 2-Condition data, both conditions must have at least two samples assigned to each of its genes.
If you have failed to adequately assign samples a dialog will pop up helping you find the condition and/or time-point that has insufficient samples.



Cluster Selection:
Use the cluster selection panel if you have previously made clusters and wish to run your analysis based on those clusters. Use the drop-down boxes to choose which clusters' samples are to be assigned to which time-points. For clusters you are not using, leave "unassigned". The same condition and time-point requirements exist for this method, but the cluster selector makes for easier and more organized sample assignment.



Hierarchical Clustering:
To have BETR construct hierarchical trees for your results, check the corresponding check box. Select whether this feature is to be applied to significant genes or significant and non-significant genes. This process may add significantly to the computation time.

The BETR module outputs standard viewers and tables for MeV's statistics analyses.

11.34 Rank Products

(Breitling, Rainer et al, 2004)

Rank Products is a novel test for determining differential expressed genes with multiple replicates. This analysis differs from many other techniques in that it does not apply a sophisticated statistical model, but rather from the calculation of rank products, a faster and simpler method.
Additionally, Rank Products is useful in highly noisy data and can significantly reduce the number of replicate experiments required to obtain reliable results.

Running RP:

MeV's RP currently supports 3 experimental designs.

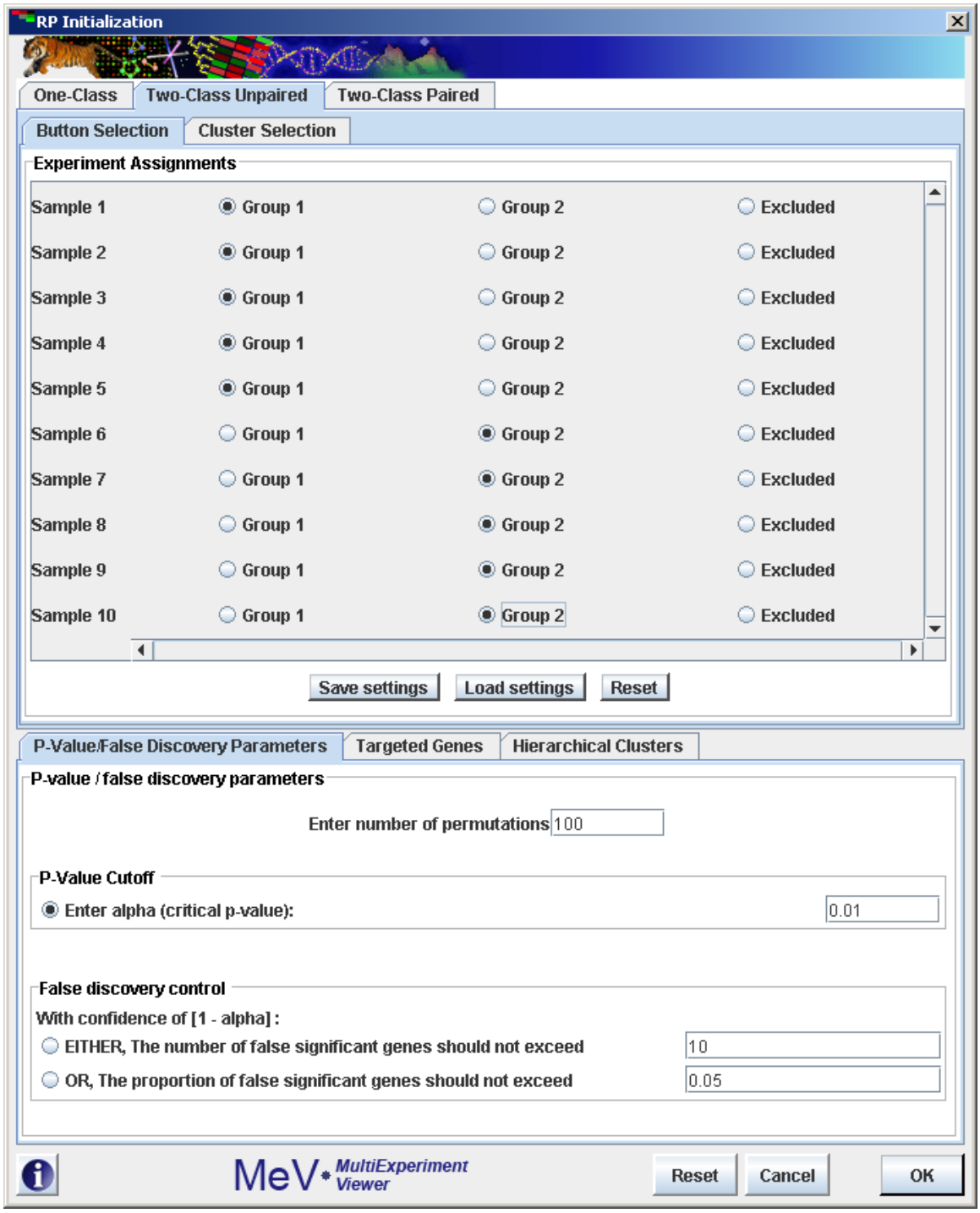
1) **One-class**, typically run on two-color data, this design determines genes that are significantly up or down-regulated within the included group. To exclude a sample from the analysis, uncheck the box next to that sample's name in the left pane of the one-class screen.

2) **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. 0).

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 rank product in group B is significantly higher than in group A. They will be considered "negative significant" if the rank product of group A significantly exceeds that of group B.

3) **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).

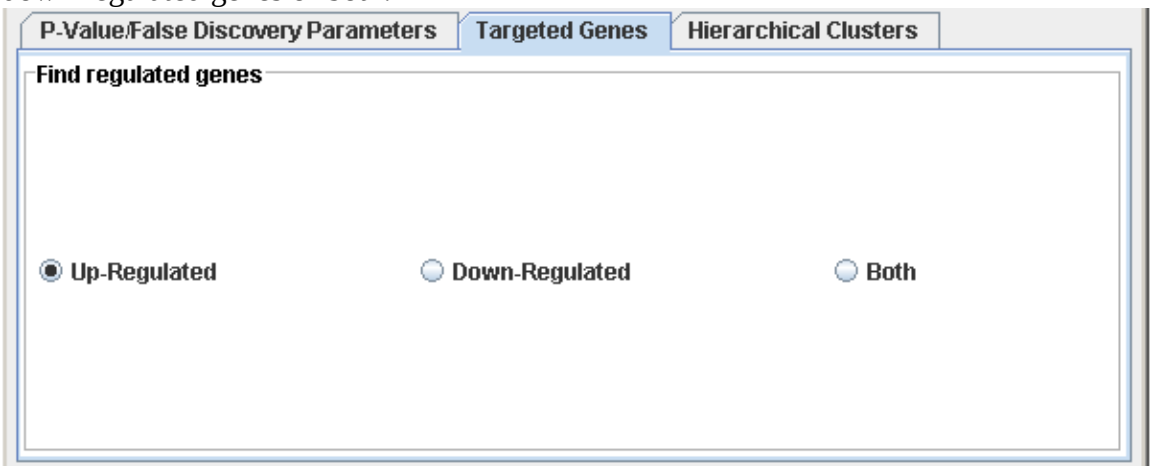
As in most other modules, MeV offers two forms of sample selection: the individual button selectionthe cluster selection tabs to assign your samples to the analysis. Samples left unassigned or unchecked will be ignored in the analysis.



P-Value/ False Discovery Parameters:
For determination of significance levels, specify the number of random permutations you want RP to run.
If setting a significance cut-off using p-values, enter the alpha value you wish to set as the cut-off point.

If determining significance by false discovery rate, check the box next to either the number or percentage of false positives and input the corresponding value.

Targeted Genes:
Check the radio button for determining in the analysis significantly up-regulate genes, down-regulated genes or both.

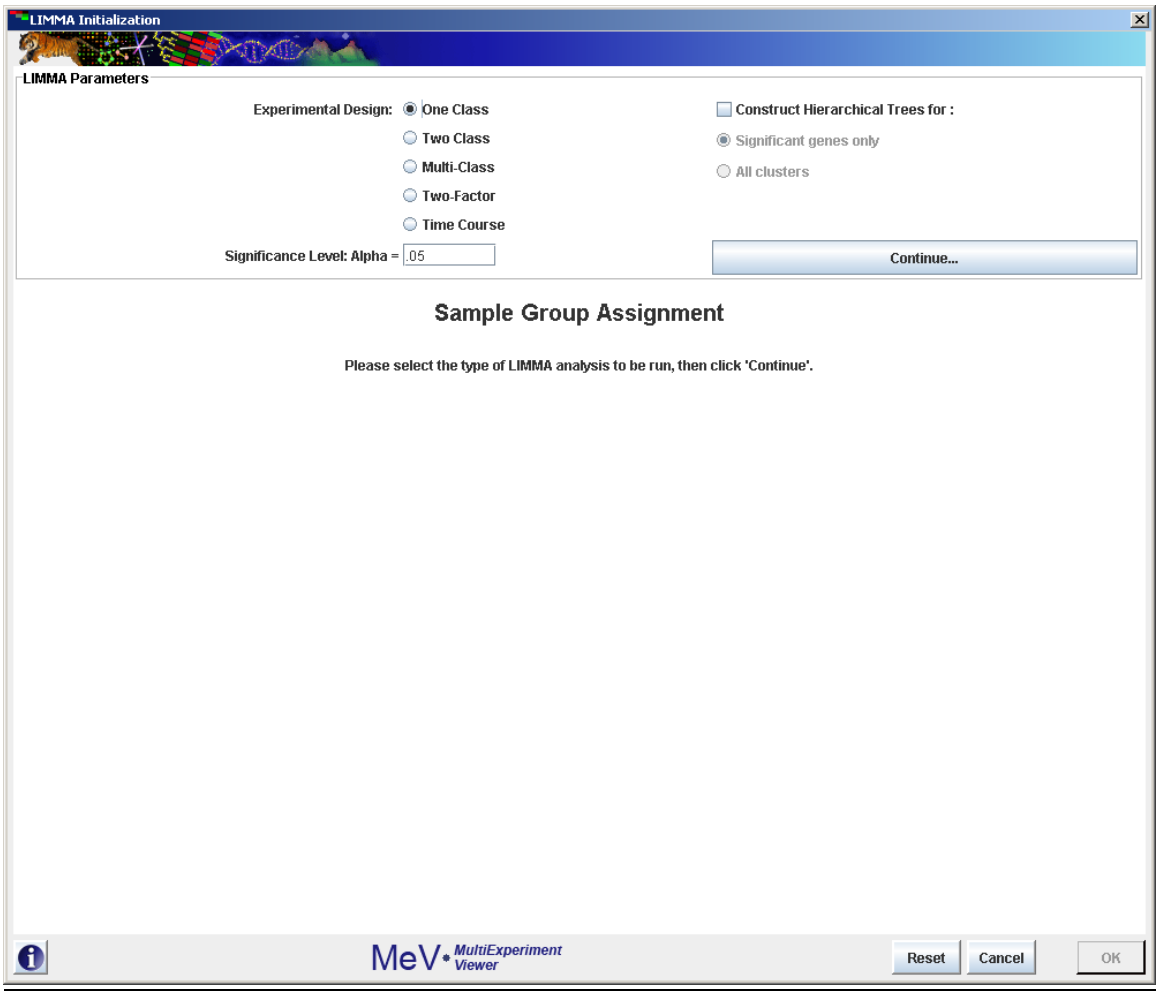


Hierarchical Clustering:
To have RP construct hierarchical trees for your results, check the corresponding check box. Select whether this feature is to be applied to significant genes or significant and non-significant genes. This process may add significantly to the computation time.

The RP module outputs standard viewers and tables for MeV's statistics analyses.

11.35 Linear Models for Microarray Data
(Smyth, G. K. 2004)

LIMMA, a new module as of MeV 4.5, is a technique for identifying differentially expressed genes based off of the fitting of each gene to a linear model. This method can be applied to a number of experimental designs such as single class, two or more groups, as well as factorial and time-course experiments.



Running LIMMA:

After opening the LIMMA initialization dialog, select the type of data you intend to run the analysis on. Your options are:

1.) One Class: Only one group is to be evaluated.

2.) Two Class: The data exists as two groups, and differential expression is to be found between the two groups.

3.) Multi-Class: The data exists as multiple groups, and differential expression is to be found between the groups.

Enter the number of groups in the visible field.

4.) Two-factor: The data will be evaluated to find genes that vary significantly across levels of two independent variables.

Enter factor names, without spaces, and the number of levels for each factor.

5.) Time-course: The data exists as two separate conditions taken at specific intervals of time.

Enter the number of time points in your data.

Assign a significance level. Genes with a p-value less than alpha will be deemed significant.

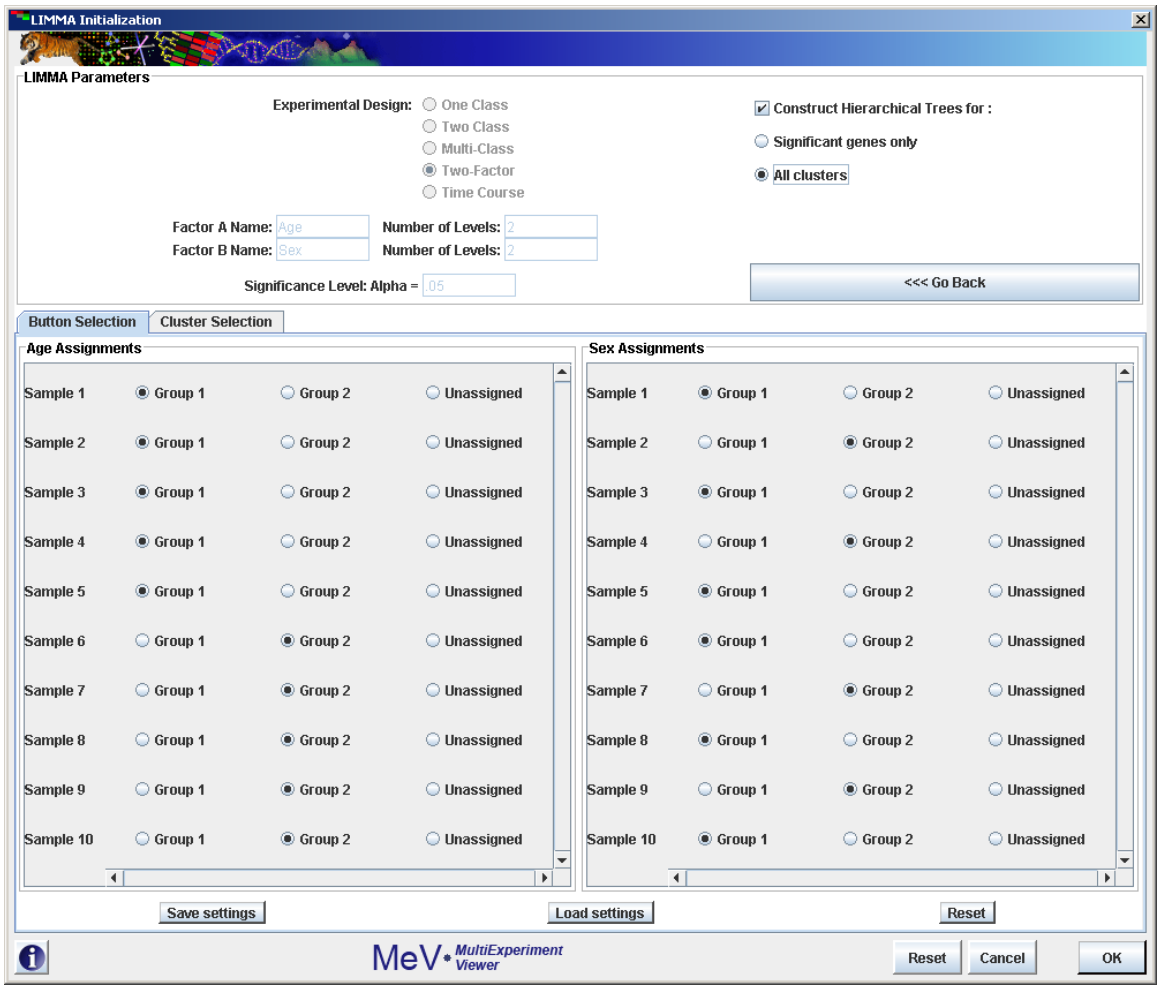
Click “Continue” to move to the next step.

Assign your samples:
Depending on the type of data you have selected, you will need to assign your samples to groups, factors or time-points, if necessary.

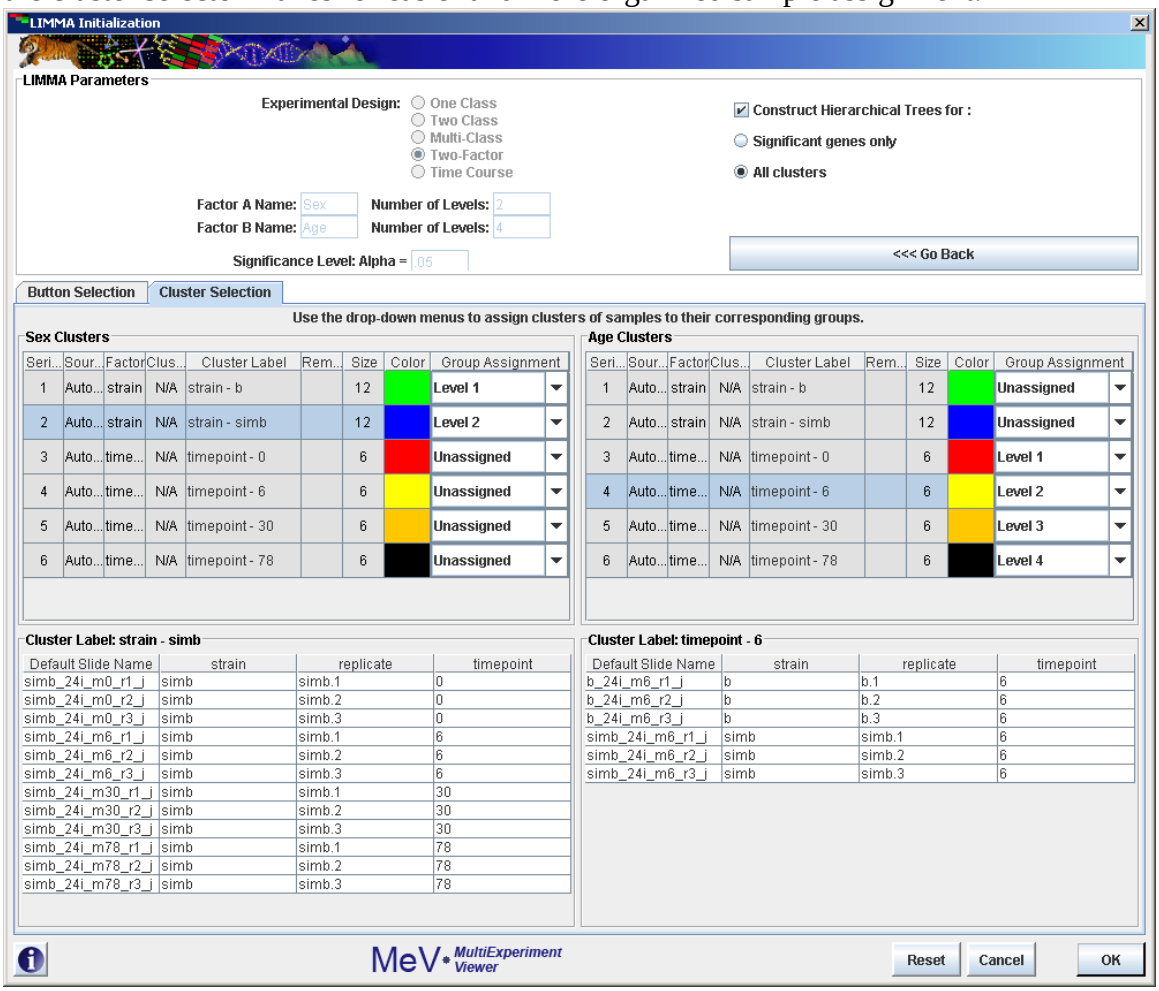
Button Selection:
Use this option if you have not created clusters for your samples and want to assign each sample individually.
Each group must contain at least two samples assigned to it. For factorial and time-course data, each combination of levels in factor A and factor B must contain at least 2 samples.

For example: if an experiment has 3 levels for factor "age" **(<40, 40-55, >55**) and 2 levels for factor "sex" (**male, female**). There must be at least two samples in each of the following:

1.) <40, male
2.) <40, female
3.) 40-55, male
4.) 40-55, female
5.) >55, male
6.) >55, female



Cluster Selection:
Use the cluster selection panel if you have previously made clusters and wish to run your analysis based on those clusters. Use the drop-down boxes to choose which clusters' samples are to be assigned to which time-points. For clusters you are not using, leave "unassigned". The same condition and time-point requirements exist for this method, but the cluster selector makes for easier and more organized sample assignment.



Hierarchical Clustering:
To have LIMMA construct hierarchical trees for your results, check the corresponding check box. Select whether this feature is to be applied to significant genes or significant and non-significant genes. This process may add significantly to the computation time.

The LIMMA module outputs standard viewers and tables for MeV's statistics analyses.

11.36 Non-negative Matrix Factorization

(Brunet et al, 2004)
(Devarajan K, 2008)
(Lee, Seung 2001)

Non-negative Matrix Factorization, a technique which makes use of an algorithm based on decomposition by parts of an extensive data matrix into a small number of relevant metagenes. NMF's ability to identify expression patterns and make class discoveries has been shown to able to have greater robustness over popular clustering techniques such as HCL and SOM.

MeV's NMF uses a multiplicative update algorithm, introduced by Lee and Seung in 2001, to factor a non-negative data matrix into two factor matrices referred to as W and H. Associated with each factorization is a user-specified rank. This represents the columns in W, the rows in H, and the number of clusters to which the samples are to be assigned. Starting with randomly seeded matrices and using an iterative approach with a specified cost measurement we can reach a locally optimal solution for these factor matrices. H and W can then be evaluated as metagenes and metagenes expression patterns, respectively. Using a "winner-take-all" approach, samples can be assigned to clusters based on their highest metagenes expression. Multiple iterations of this process allow us to see the robustness of the cluster memberships. Additionally, running multiple ranks consecutively can allow for the comparison between differing numbers of classes using cophenetic correlation.

NMF is most frequently used to make class discoveries through identification of molecular patterns. The module can also be used to cluster genes, generating metasamples rather than metagenes.

Parameters:

Sample Selection

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

Run multiple ranks

This checkbox determines if NMF will be performed on one rank or compared across multiple ranks. Selecting this option will allow the user to choose a rank range.

Number of runs

Specifies the number of factorization runs to perform on each rank. Typically, this value is between 20 and 100.

Rank Value/ Rank Range

The value or range of ranks for which NMF is performed. This is an integer, or set of integers greater than 1 which will also correspond to the number of clusters.

Maximum iterations

The maximum number of iterations to be completed as W and H approach a local optimization.

Always perform maximum iterations

Specifies whether or not MeV should complete the maximum number of iterations or stop after a certain convergence has been reached.

Cost convergence cutoff

The point at which a run can be halted based on sufficient convergence.

Check Frequency

The frequency, in iterations, of checking for convergence.

Update rules and cost measurement

The algorithmic technique, as described by Lee and Seung, for iteratively updating the factor matrices and the manner in which their cost is measured. The default is “Divergence”.

Data matrix pre-processing

A requirement for running NMF is a data matrix free of negative values. If your data includes negative values, a method must be selected to adjust the data. If “Always adjust data” is checked, the selected operation will be performed, regardless of your data. If unchecked, the selected operation will only be performed if a search through the data reveals negative values. After running your analysis, the preprocessing step (if any) will be recorded in the General Info tab.

Subtract minimum value means that the lowest value in the data, regardless of that value’s negativity, will be subtracted from all values, ensuring a non-negative matrix.

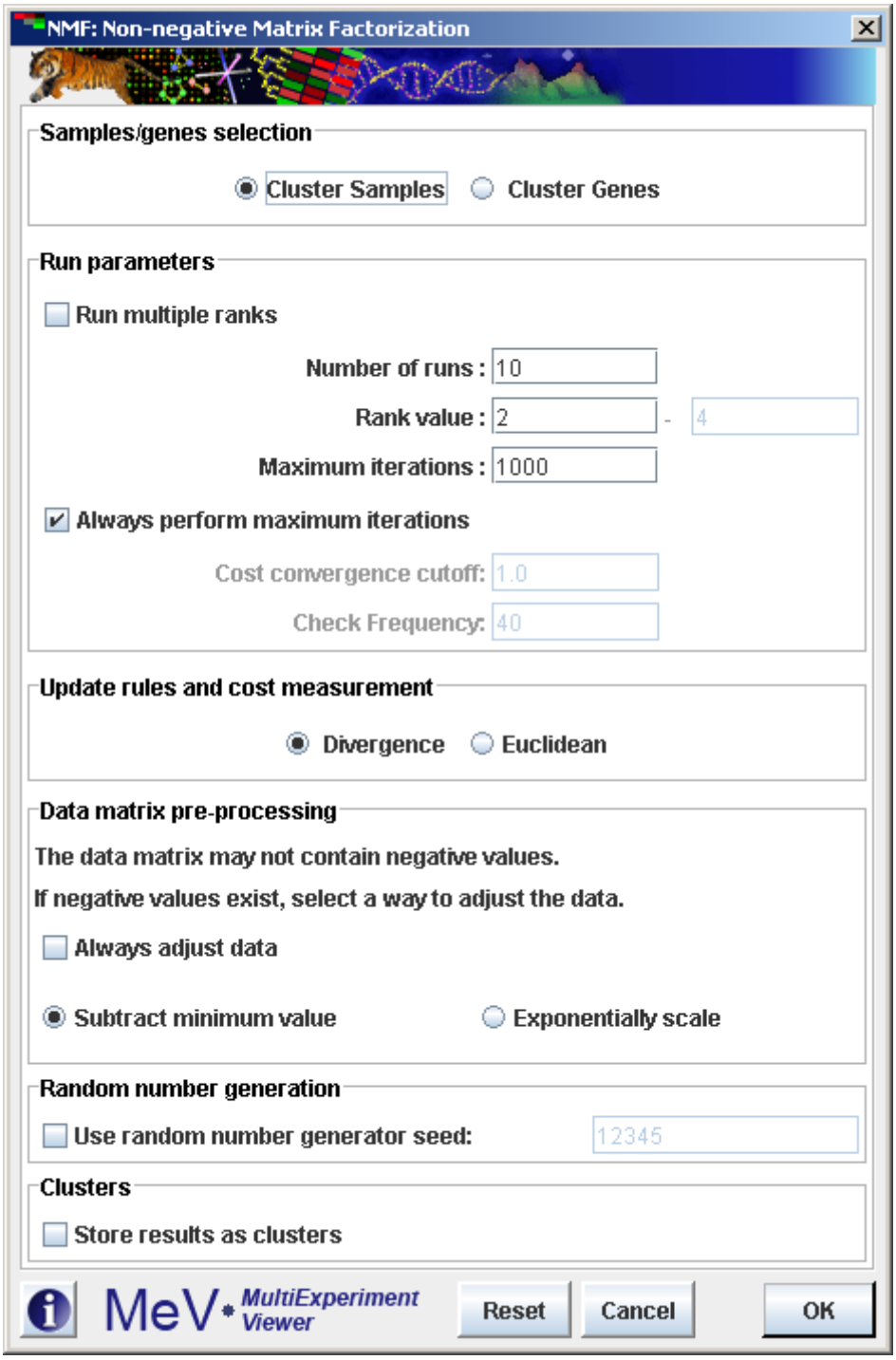
Exponentially scale means that every value will be exponentiated, base 2.

Random Number Generation

For the purpose of reproducibility, NMF allows the addition of a seed value for the creation of initial generation of W and H matrices. Use of the same seed value on runs of NMF with identical parameters will result in identical results.

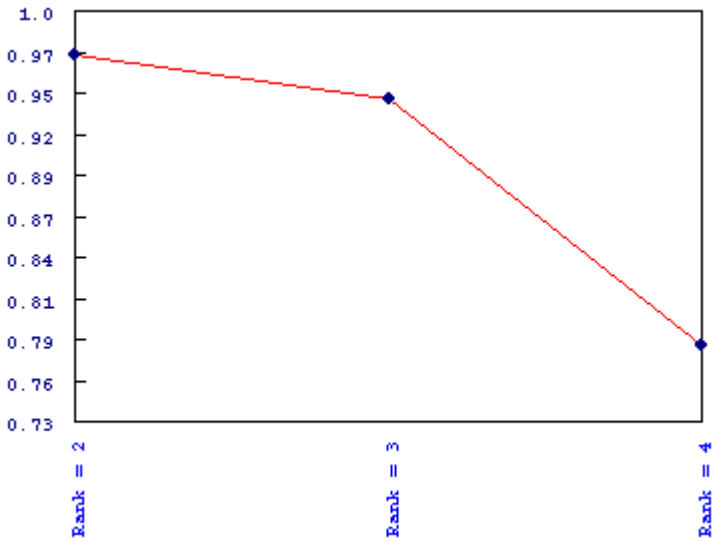
Store results as clusters

Selecting this option will cause the creation and storage of the results as clusters in MeV’s clustering system. These clusters will be visible in the Cluster Manager as well as in MeV’s experiment viewers.



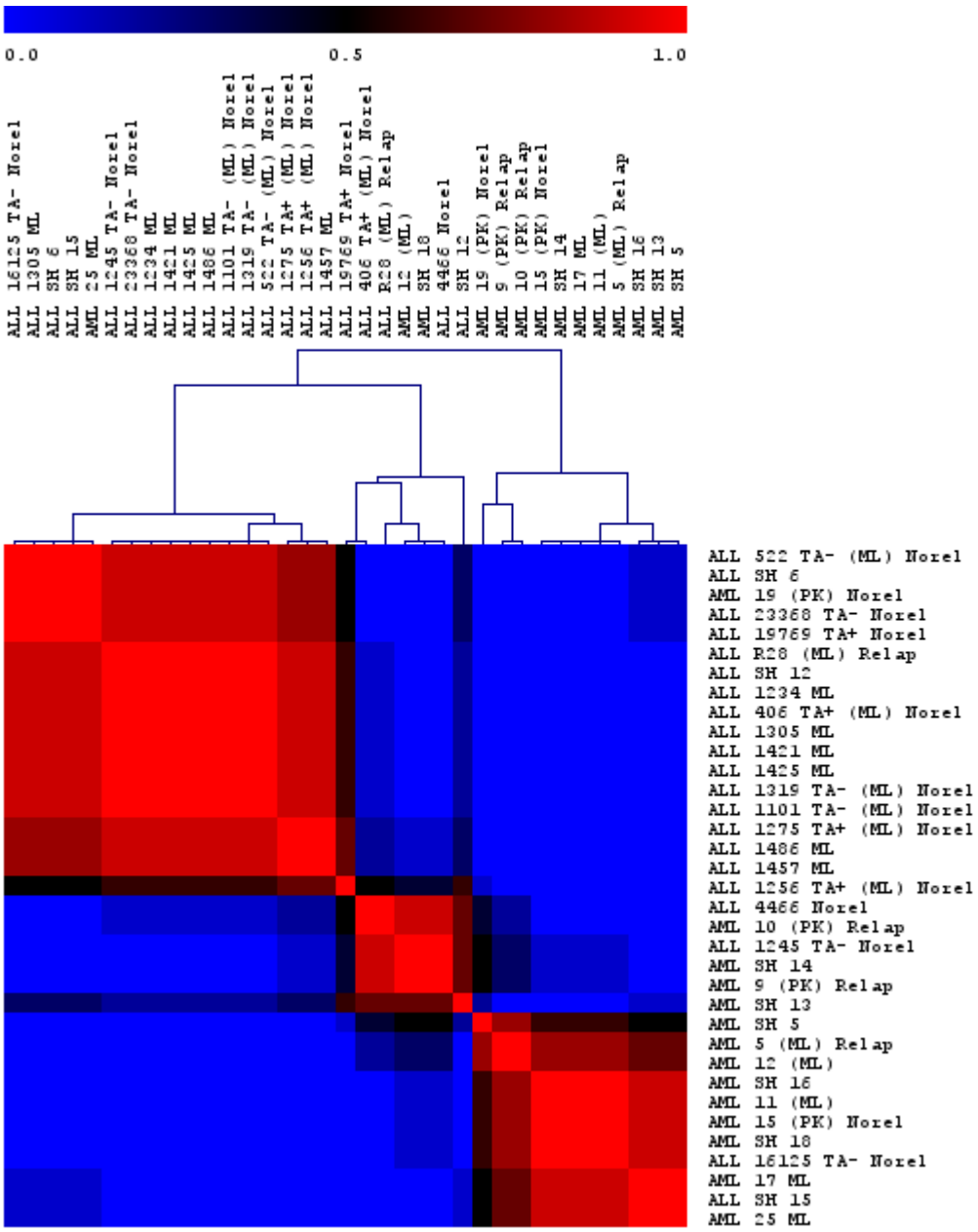
Results:

For each rank analyzed, MeV's NMF calculates a cophenetic correlation that is used to quantify the robustness of the rank's evaluation. This is reported as a node at the top of the result tree for each rank. In the case where multiple ranks are run, a Cophenetic Correlation Graph is created and displayed in the result tree. This gives a visual representation of the relative strength of clustering associated with each rank.



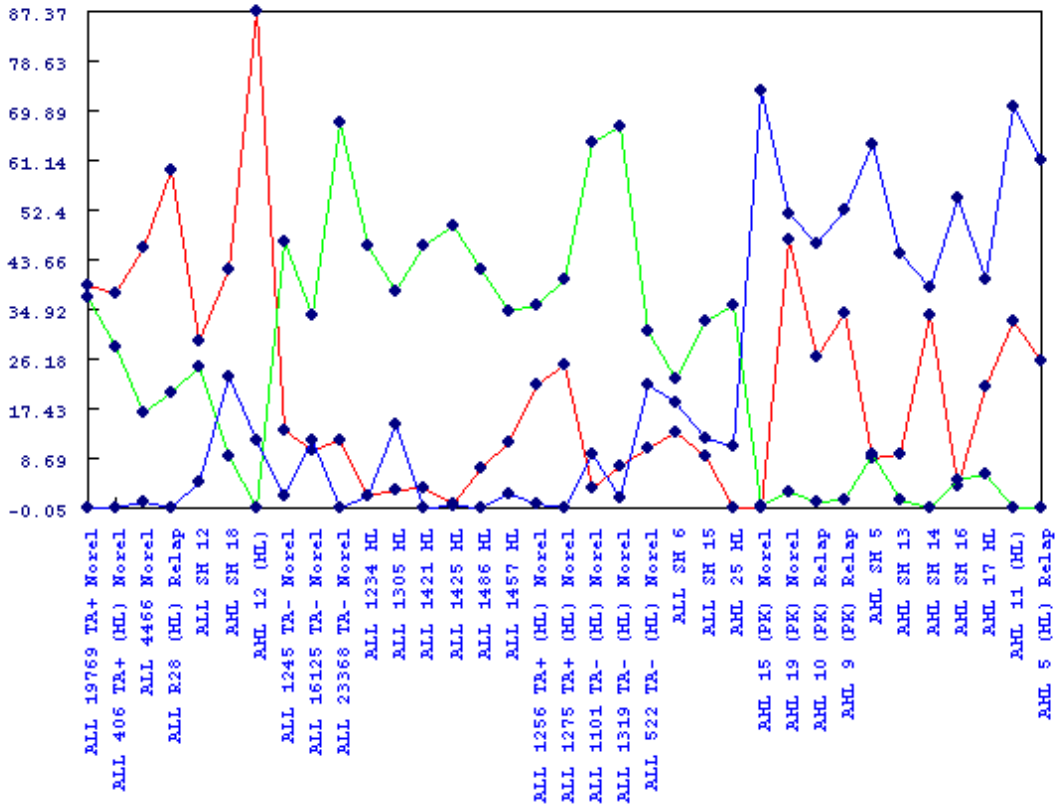
For each rank, MeV displays standard result viewers for each of the clusters, including heatmaps, expression graphs, centroid graphs, tables, and general cluster information.

The Consensus Matrix with Hierarchical Clustering is a visual representation of the clustering and its robustness. The viewer is a square matrix with dimensions equal to the number of samples (or genes, if "Cluster Genes" was selected). In this viewer, each point on the map represents the frequency with which the two samples (or genes) were assigned the same cluster through the runs, with 1 meaning 100% consensus and 0 being 0% consensus. The matrix has been reordered into a hierarchical tree with frequency of co-clustering used as the distance metric. It was then reordered for leaf order optimization.



Metagene and metagene expression patterns are reported in subfolders for each rank. Each rank will have a set of metagenes and metagene expressions corresponding to the factor matrices W and H of each run. Each set are displayed in ranking order of cost, with the most optimal metagenes displayed first. The raw values can be seen and extracted for further analysis in the “Metagenes (W)” and “Metagene Expressions (H)” folders.

Color coded graphs for these values exist in the subfolders beneath the raw values. These offer a visual representation of the strength of each metagene or metagene expression pattern on a particular sample or gene. By right-clicking on the graph, a sorting function can be toggled which, when on, orders the samples (or genes) according to the metagene with the highest expression.


87.37
78.63
69.89
61.14
52.4
43.66
34.92
26.18
17.43
8.69
-0.05
ALL 19769 TA+ Morel
ALL 406 TA+ (HL) Morel
ALL 4466 Morel
ALL R28 (HL) Relap
ALL SH 12
AHL SH 18
AHL 12 (HL)
ALL 1245 TA- Morel
ALL 16125 TA- Morel
ALL 23368 TA- Morel
ALL 1234 HL
ALL 1305 HL
ALL 1421 HL
ALL 1425 HL
ALL 1486 HL
ALL 1457 HL
ALL 1256 TA+ (HL) Morel
ALL 1275 TA+ (HL) Morel
ALL 1101 TA- (HL) Morel
ALL 1319 TA- (HL) Morel
ALL 522 TA- (HL) Morel
ALL SH 6
ALL SH 15
AHL 25 HL
AHL 15 (PK) Morel
AHL 19 (PK) Morel
AHL 10 (PK) Relap
AHL 9 (PK) Relap
AHL SH 5
AHL SH 13
AHL SH 14
AHL SH 16
AHL 17 HL
AHL 11 (HL)
AHL 5 (HL) Relap

11.38 Attract package
(Jessica Mar et al)

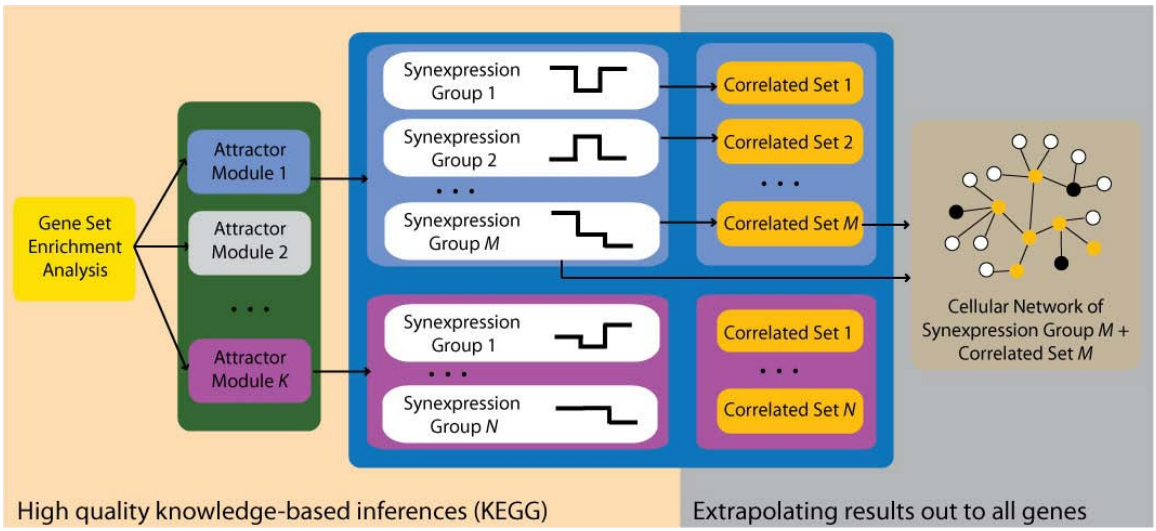
What is Attract?

The Attract algorithm identifies the core gene expression modules that are differentially activated between cell types or different sample groups, and elucidates the set of expression profiles which describe the range of transcriptional behavior within each module.

The algorithm assists in understanding expression changes that are specific to cell types or cell fate transitions. A mammalian organism is made up of over 200 types of specialized cells. Each cell type carries out a specific task integral to maintaining homeostasis of the organism. Cell types can vary by morphology, structure, lifespan, functional ability and much more. Different cell or tissue types acquire their diversity by driving differentially coordinated expression patterns through interacting gene networks. The attractor hypothesis proposed by Stuart Kauffman describes how cell fate transitions between cell types occur through coordinated changes in genome-wide gene expression.

The Attract algorithm starts by using a GSEA step to identify the set of core pathway modules driving the cell types-specific expression changes. For each of these modules, the algorithm then constructs synexpression groups, which consist of genes with correlated expression profiles. The algorithm is also able to extrapolate out to identify genes with similar expression profiles which may not be currently annotated in current pathway databases like KEGG.

Ref: (Mar JC, Wells CA, Quackenbush J. (2010). Identifying Gene Expression Modules that Represent the Drivers of Kauffman's Attractor Landscape)



When should Attract be used?

The Attract algorithm can be used for any expression data set where replicate samples (at least three) have been collected for various cell types or experimental groups such as a panel of drug treatments or a time course experiment. The algorithm is not well-suited for data generated under an experimental design that has only one experimental or covariate group, e.g. samples from 100 identical cell lines with no perturbations.

Brief description of the algorithm

The Attract algorithm can be roughly divided in to four steps:

1. Find core attractor state pathway modules in your data using GSEA.
2. Remove flat or uninformative genes (i.e. genes that do not show expression changes across the different factor levels).
3. Find Synexpression groups. (A synexpression group contains genes that share similar expression profiles across factor levels)
   - Use the informativeness metric to compute clusters (synexpression groups) in each geneset (Mar JC, Wells CA, Quackenbush J. (2010). Defining an informativeness metric for clustering gene expression data. Submitted).
4. For every synexpression group in a geneset, find genes in the expression data which have correlated expression profiles.

How to run Attract?

Attract uses a set of parameter input dialogs that open sequentially to provide input options that correspond to each step of the process. The first step in the process is data selection which lets you assign phenotype/class labels to your samples

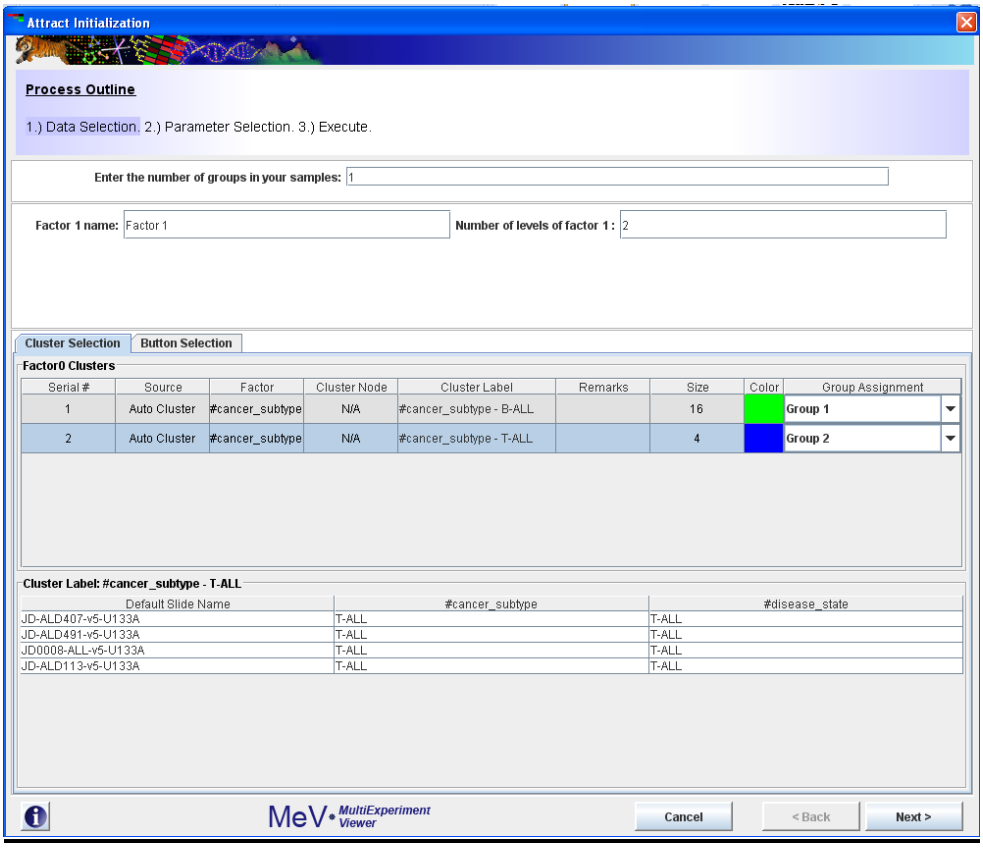


Figure 1

The default assignment (Figure 1) is two groups (factors) with two levels per group. This can be changed to reflect the groups present in your data. For example, if cancer subtype is the phenotype (factors) that influences your data the most, enter 1 in the "Enter number of groups in your sample" textbox.

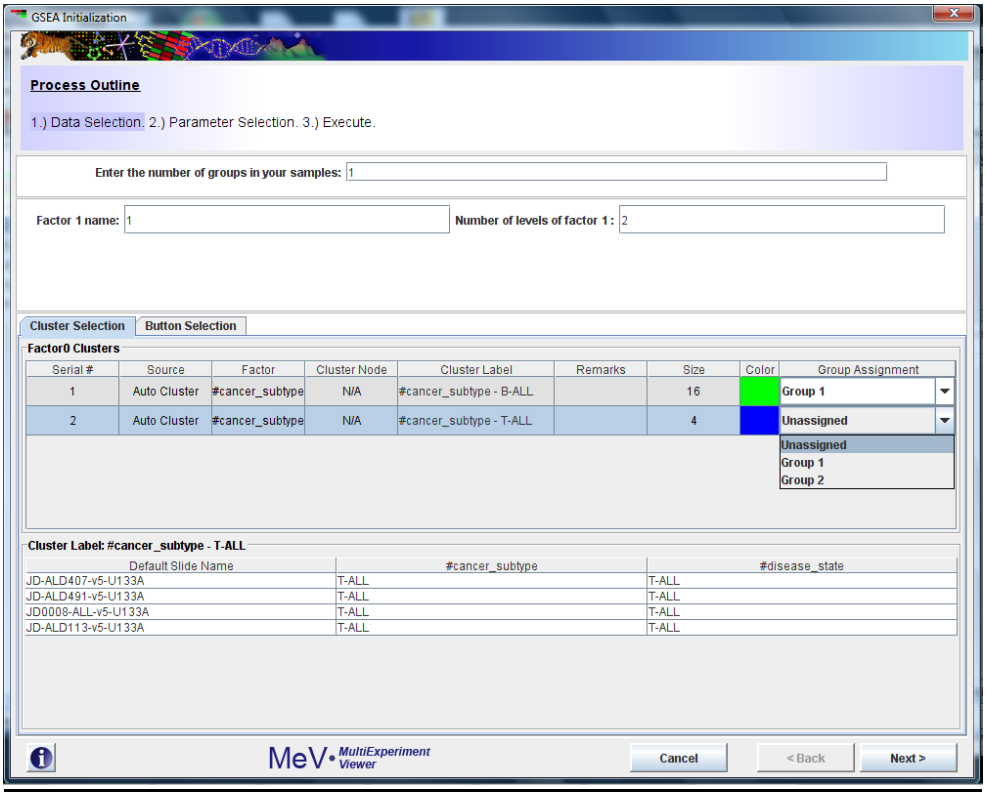


Figure 2

B-ALL and T-ALL are the two levels of this phenotype, so enter 2 in the "*Number of levels of factor*" textbox. If you have pre selected sample clusters and decide to use the "*Cluster Selection*" tab just assign group numbers using the Group Assignment drop down as shown in Figure 2.

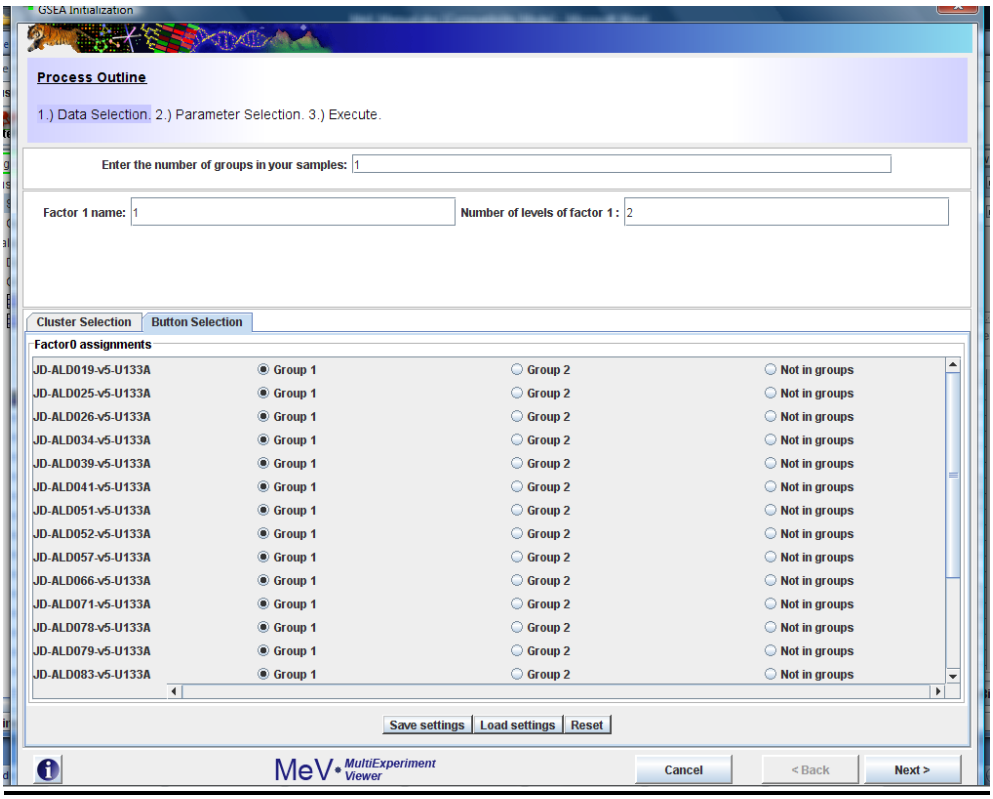


Figure 3

You can also use the "*Button Selection*" tab shown in Figure 3 to achieve the same.
Group1 and Group2 symbolize B-ALL and T-ALL. You can save these grouping using the "Save settings" button. To load saved groupings, use the "Load settings" button. Reset button will clear all your choices. Once you are done, click the Next > button.

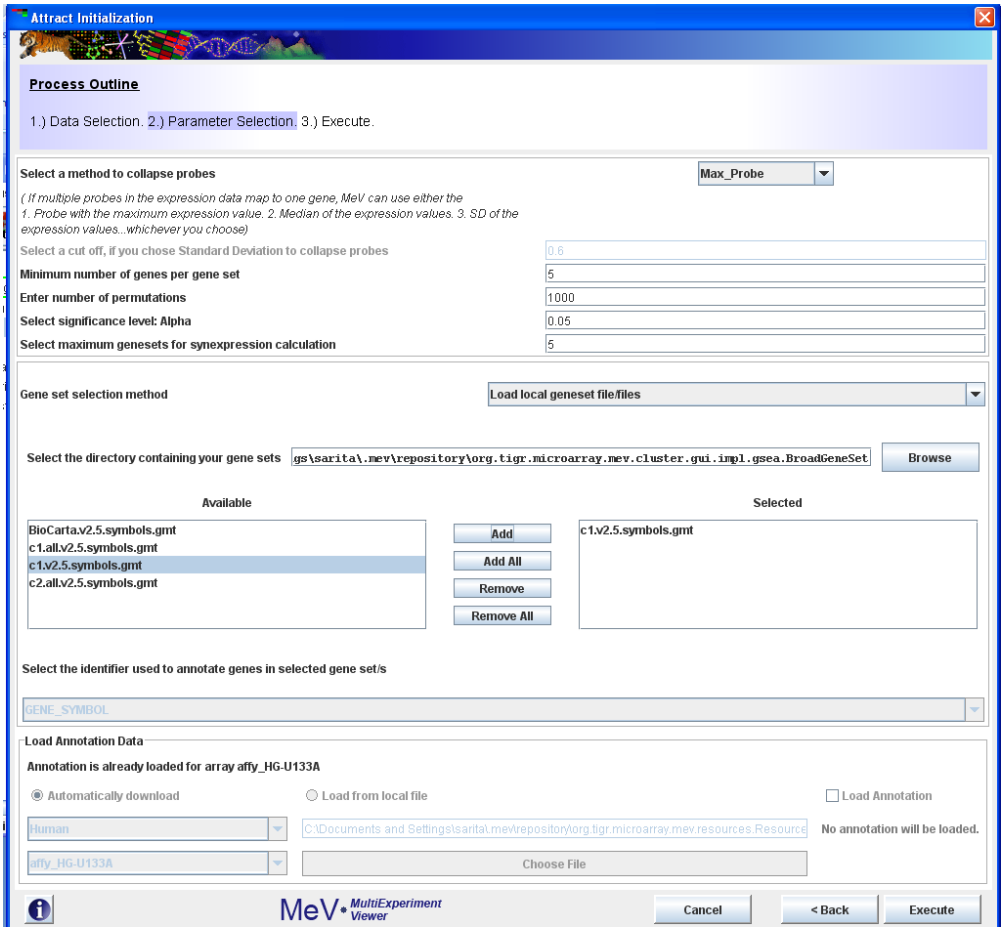


Figure 4

This brings you to the "Parameter Selection" section shown in Figure 4. The GUI is pretty self explanatory, but here is some better clarification about the available methods for collapsing probes to genes, loading gene sets and loading annotations.

```
  MAX_PROBE:
            Sample1          Sample2          Sample3
  Probe_1   10               20               30
  Probe_2   20               5                10
  Gene_12   20               20               30


  MEDIAN_PROBE:
            Sample1          Sample2          Sample3
  Probe_1   10               20               30
  Probe_2   20               15               10
  Probe_3   20               5                10
  Gene_123  20               15               10

  STANDARD_DEVIATION:
            Sample1          Sample2          Sample3          SD
  Probe_1   10               20               30                 3
  Probe_2   20               5                10                 4
  Probe_3   20               15               10                 2
  Gene_123 will be represented by Probe_2, which has the MAX SD value
across samples. So,
  Gene_123   20   5     10
```

NOTE: SD will be calculated by MeV on the fly. You do not have to do anything. This is just an example.

Attract uses LIMMA to remove flat or uninformative genes. "Select *significance level: Alpha text box"* lets you choose a significance level. The default significance level is 0.05. Attract aims to find synexpression groups which contain genes that share similar expression profile across different factor levels. "*Select maximum gene sets for synexpression calculation"* lets you choose the number of gene sets for which you want to calculate synexpression groups. By default we calculate synexpression groups for the top five enriched gene sets.

The '*Browse*' button corresponding to "*Select the directory containing your gene sets*" lets you choose the directory containing gene set files. You can select the files you want to use from the "*Available*" panel. "*Selected"* panel indicates the gene set files that you chose to use for this analysis. In addition to this gene sets can also be downloaded from the MIT/Broad website http://www.broad.mit.edu/gsea/msigdb/downloads.jsp

If your gene set file is *.gmt or *.gmx format, *"Select the identifier used to annotate genes in your selected gene set*" drop down is automatically populated with "GENE_SYMBOL" as shown in figure above. In case of a custom gene set file, you must manually choose the gene identifier from the drop down.

The "*Load Annotation Data*" panel lets you upload annotations. Annotations are a MUST for running Attract. Details on how to load annotations is described in Using the Annotation Feature

The last step is to hit the Execute button. Attract outputs besides the standard MeV viewers a new viewer named "*SynExpression Graph"*. *"Excluded Gene Sets*", "*Excluded Genes*" under "*Table Views*"

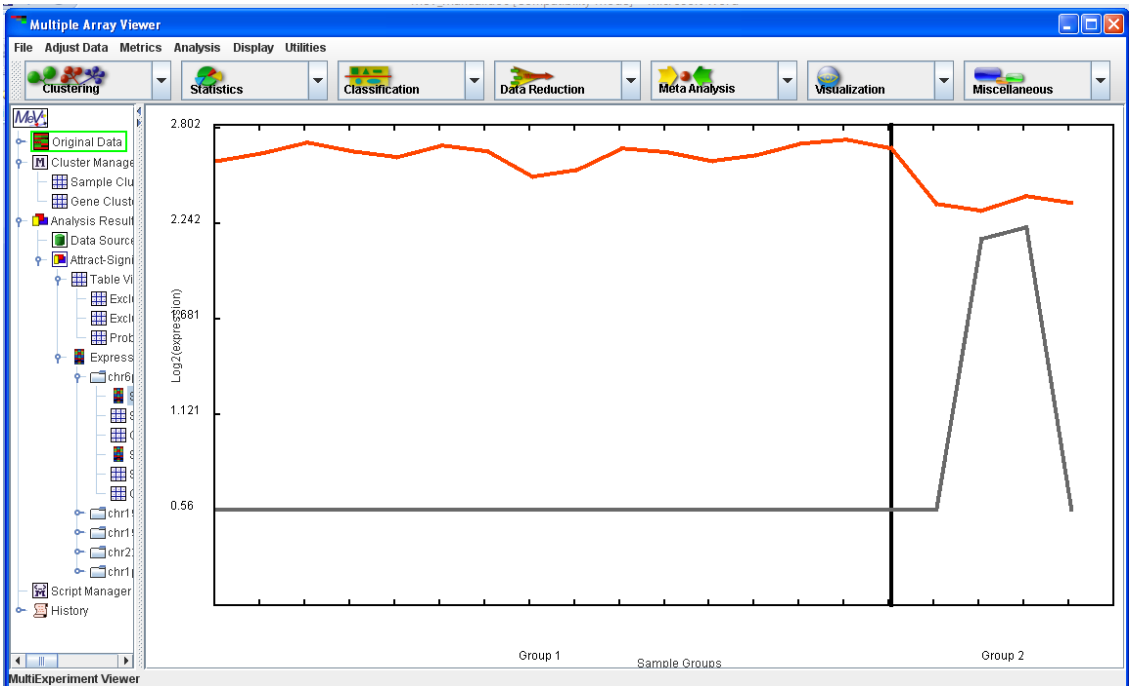


Figure 5

"SynExpression graph" (Figure 5) represents the average expression profile of a geneset within a synexpression group. The orange line represents the average expression profile of genes that belong to the synexpression group. The gray line indicates the average expression profile of genes in the expression data that share a similar expression profile.

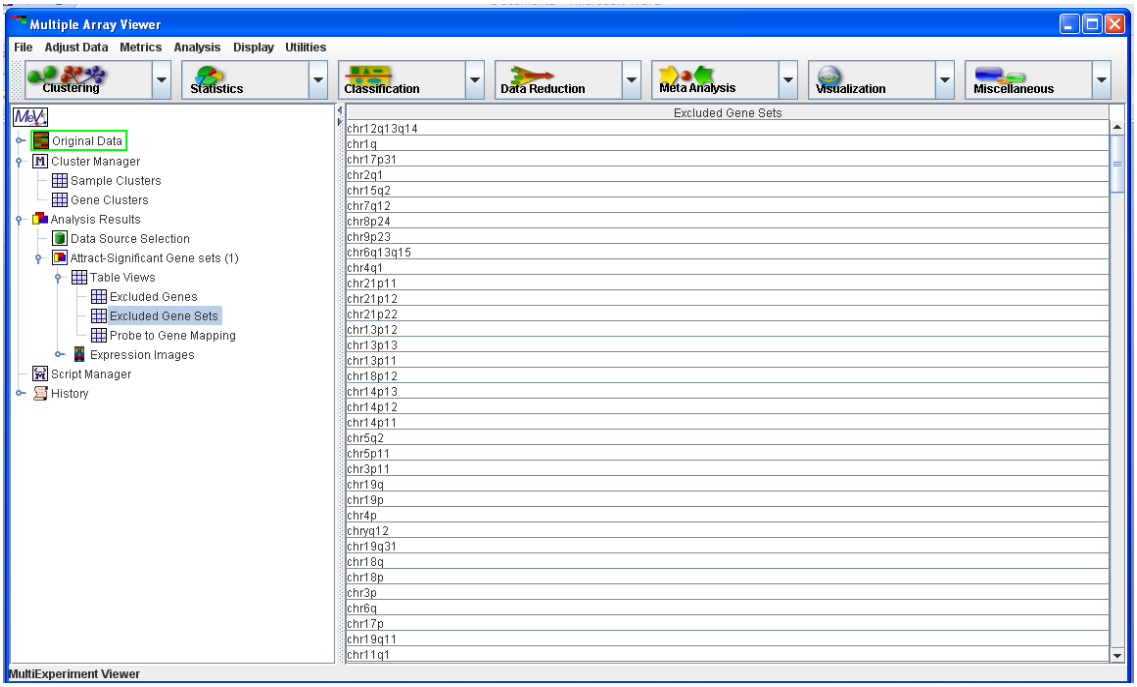


Figure 6

"Excluded gene sets" table (Figure 6) contains gene sets which do not pass the minimum genes criteria and which do not contain at least 5 genes after the algorithm removes flat or uninformative genes.

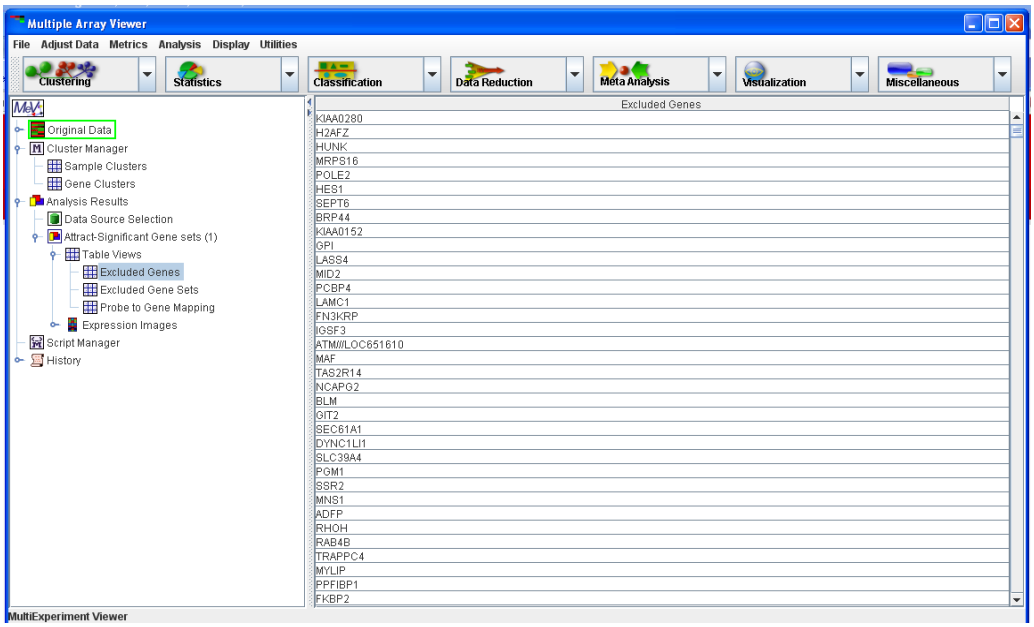


Figure 7

"Excluded Genes" table (*Figure 7*) contains the genes deemed uninformative or flat by LIMMA.

11.38 MINET
(Patrick E. Meyer, Frederic Lafitte, and Gianluca Bontempi. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. BMC Bioinformatics, Vol 9, 2008.)

For a given dataset, minet infers the network in two steps. First, the mutual information between all pairs of variables in dataset is computed according to the estimator argument. Then the algorithm given by method considers the estimated mutual information in order to build the network.

***Parameters*:**
Dataset: A cluster of genes.

Method: The inference algorithm: "clr", "aracne", "mrnet" or "mrnetb"

Estimator: The name of the entropy estimator to be used for mutual information computation: "mi.empirical", "mi.mm", "mi.shrink", "mi.sg", "spearman", "kendall", "pearson".

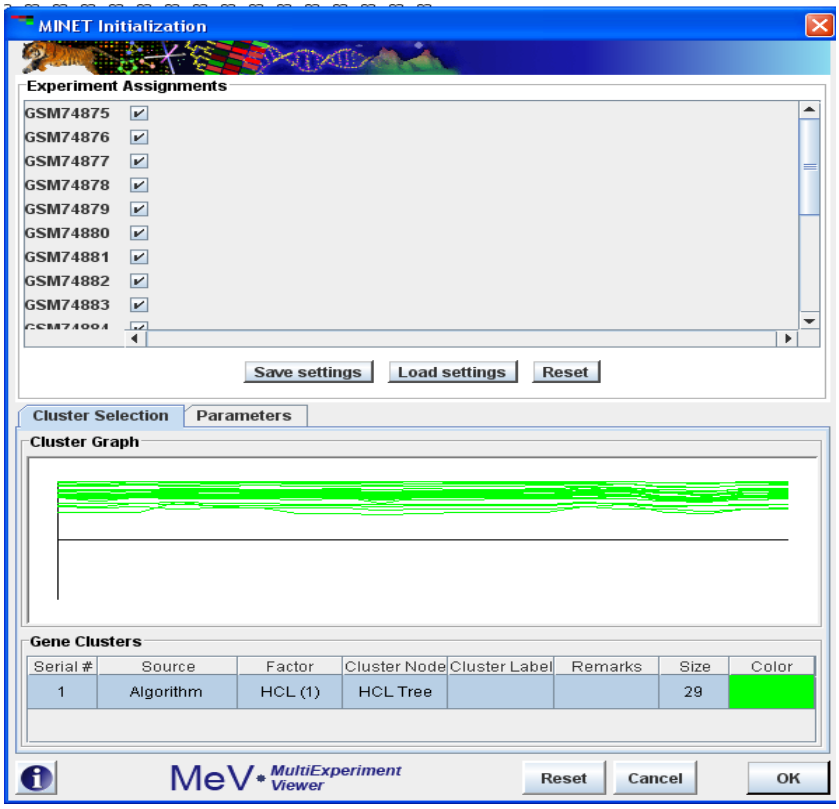
Discretization: The name of the discretization method to be used, if required by the estimator "none", "equalfreq", "equalwidth" or "globalequalwidth".

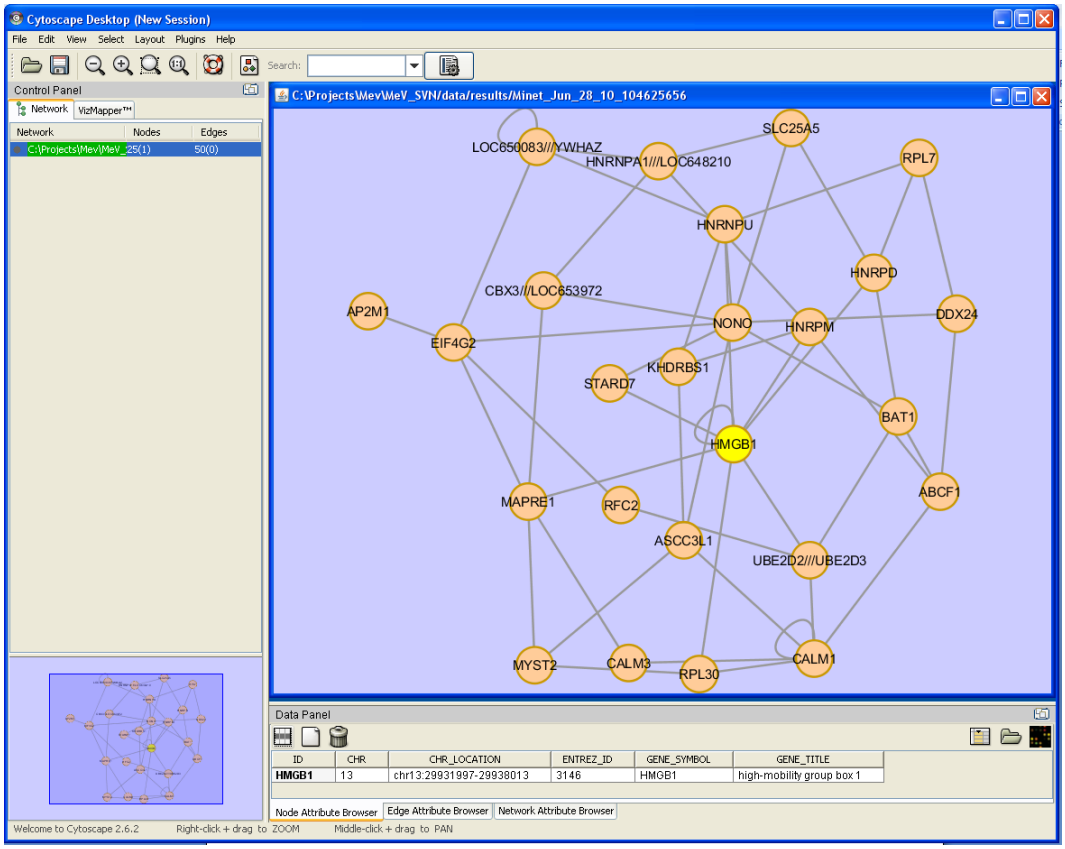
Number of Bins: Integer specifying the number of bins to be used for the discretization if disc is set properly. By default the number of bins is set to $\sqrt{N}$ where N is the number of samples.

Output:
Minet returns a matrix which is the weighted adjacency matrix of the network. The weights range from 0 to 1 and can be seen as a confidence measure on the presence of the arcs. In order to display the network, MeV launches Cytoscape via webstart and displays the network in that application.

How to Run MINET:
1. Create Cluster
2. Launch MINET from ToolBar -> Statistics -> MINET
3. Un-select samples to be excluded from analysis
4. Select one cluster from Cluster selection Panel.
5. Click on Parameters tab to change method, estimator etc or leave them as is to accept default
6. Hit OK

Viewing the network:
The network is displayed in Cytoscape. IN the initial view all nodes are stacked up on top of each other. In Cytoscape follow MenuBar -> Layout -> yFiles –> Organic (or any layout of your choice) to layout the nodes.


MINET Initialization
Experiment Assignments
GSM74875
GSM74876
GSM74877
GSM74878
GSM74879
GSM74880
GSM74881
GSM74882
GSM74883
Save settings
Load settings
Reset
Cluster Selection
Parameters
Cluster Graph
Gene Clusters
Serial #
Source
Factor
Cluster Node
Cluster Label
Remarks
Size
Color
1
Algorithm
HCL (1)
HCL Tree
29
MeV MultiExperiment Viewer
Reset
Cancel
OK


Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
Control Panel
Network
VizMapper™
C:\Projects\Mev\MeV_SVN/data/results/Minet_Jun_28_10_104625656
SLC25A5
LOC650083///YWHAZ
HNRNPA1///LOC648210
RPL7
HNRNPU
HNRPD
CBX3///LOC653972
AP2M1
NONO
HNRPM
DDX24
EIF4G2
KHDRBS1
STARD7
BAT1
HMGB1
ABCF1
MAPRE1
RFC2
ASCC3L1
UBE2D2///UBE2D3
MYST2
CALM3
RPL30
CALM1
Data Panel
ID
CHR
CHR_LOCATION
ENTREZ_ID
GENE_SYMBOL
GENE_TITLE
HMGB1
13
chr13:29931997-29938013
3146
HMGB1
high-mobility group box 1
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.6.2
Right-click + drag to ZOOM
Middle-click + drag to PAN

11.39 Survival

(Goeman, J.J., 2010. L(1) penalized estimation in the cox proportional hazards model. Biometrical Journal. Biometrische Zeitschrift, 52(1), 70-84.; R Development Core Team, 2009. R: A Language and Environment for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org.)

The Survival (SURV) module contains two functions for the analysis of censored survival data. The first is a basic comparison of the survival curves of two groups of samples. Sample data loaded into MeV is compared using the R package survival and the degree to which the curves differ is reported, along with a p-value.

The second feature of the module is the creation of a cox proportional hazards model based on the loaded gene expression data, using survival time as the reporting value. The R packaged penalized is used to build the model and to select the most informative genes. The L1-norm of the expression vector is used as a shrinkage parameter; users can select the lambda value. The module performs cross-validation of the model after it is built and reports a log likelihood as an indicator of model utility.

Running the module

Censored survival data

No matter which of the analysis options you choose, you will need to select the sample annotation types that hold the survival data. There are two fields: survival time should point to a set of sample annotation that contains a floating-point number representing a time to event. The Censored field should point to a sample annotation field containing flags that indicate whether a given datapoint is censored or not. MeV can recognize several censoring flag types, including "Yes" or "No", "Censored/Uncensored", or 1/0. Values of "Yes", "Censored" or "1" indicate that the datapoint should be treated as a censored event, whereas "No", "Uncensored" and "0" are treated as valid, observed events.

After opening the SURV initialization dialog, select the type of analysis you want to run. Your options are:

1.) Kaplan-Meier Plot: Choose this option if you have already selected groups of samples that are of interest to you. This option will compare the survival time of the two groups and calculate a *p*-value indicating whether the difference is significant.

2.) Cox Proportional Hazard Model: Choose this option if you are interested in identifying genes that appear to have predictive value relative to survival time.

After choosing the analysis type, click the Continue button to generate a cluster selection panel. This panel is different depending on which analysis type you have chosen.

For the Kaplan-Meier plot

A sample cluster selection panel will be generated. Select which samples belong in which groups and click ok to compare the survival profiles.

Use the cluster selection panel if you have previously made clusters and wish to run your analysis based on those clusters.

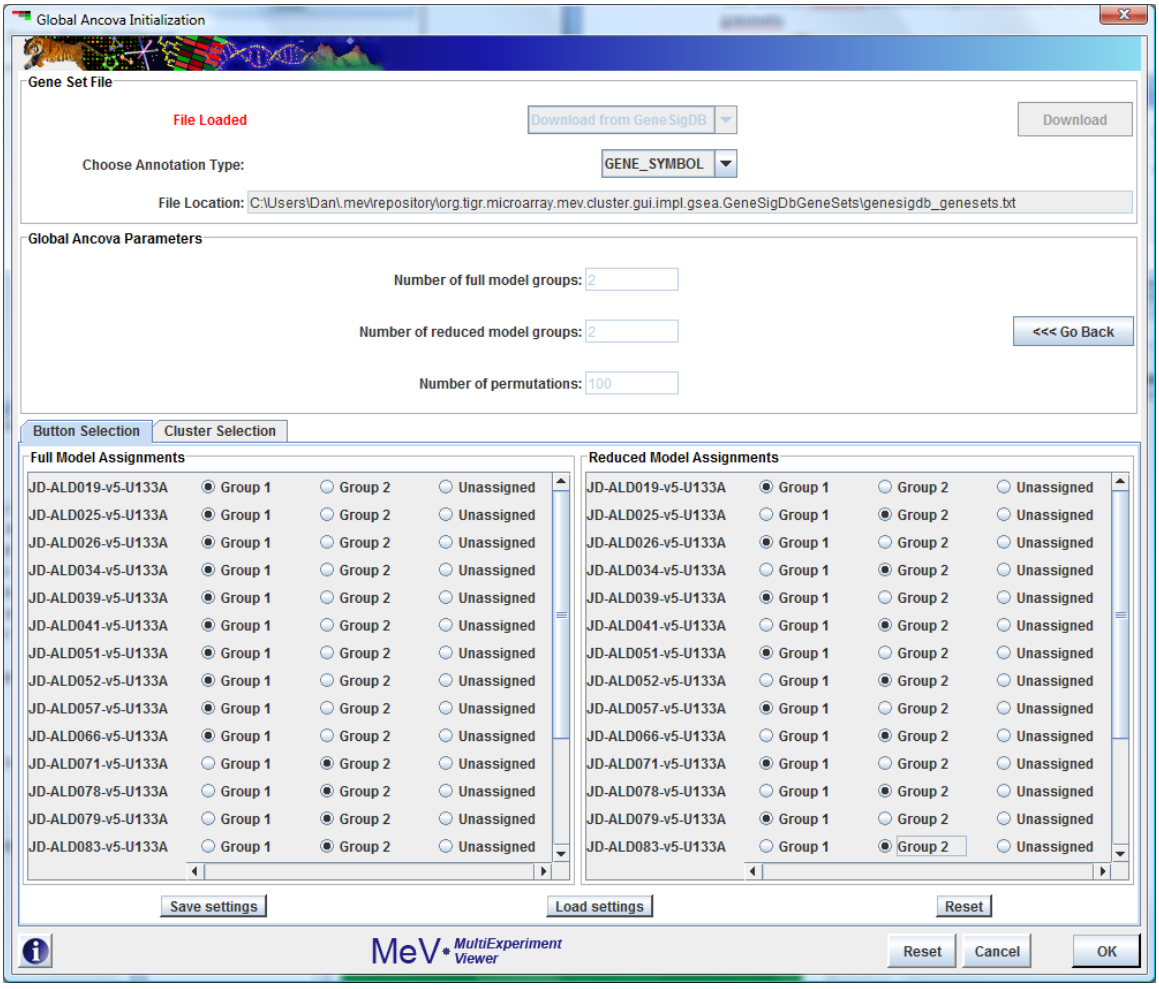
Use the drop-down boxes to choose which clusters' samples are to be assigned to which time-points. For clusters you are not using, leave "unassigned". The same condition and

time-point requirements exist for this method, but the cluster selector makes for easier and more organized sample assignment.

For the Cox Proportional Hazard Model
A gene cluster selection panel with be generated. You can choose to select a group of genes to build the model with (faster) or choose to build the model based on all of the genes in the currently-loaded dataset (slow).
The Cox model will be build using an L1 Norm (lasso) penalty. Select the value of lambda to be used for this penalty. Higher values will result in faster-calculated models with fewer coefficients.
The SURV module outputs standard viewers and tables for MeV's statistics analyses. It also creates survival curves for the selected sample groups.

11.40 Global Analysis of Covariance

Global ANCOVA, a new module to MeV 4.6, is a technique for identifying differentially expressed gene sets based off of the calculation of an F-test between groups of samples. Analyses are typically run in a two-class format but may also be applied to additional groups, time-course or other designs of arbitrary variables. Global Ancova fits linear models to the data and compares them using the extra sum of squares principle. The result table reports p-values, permutation p-values and asymptotic p-values.



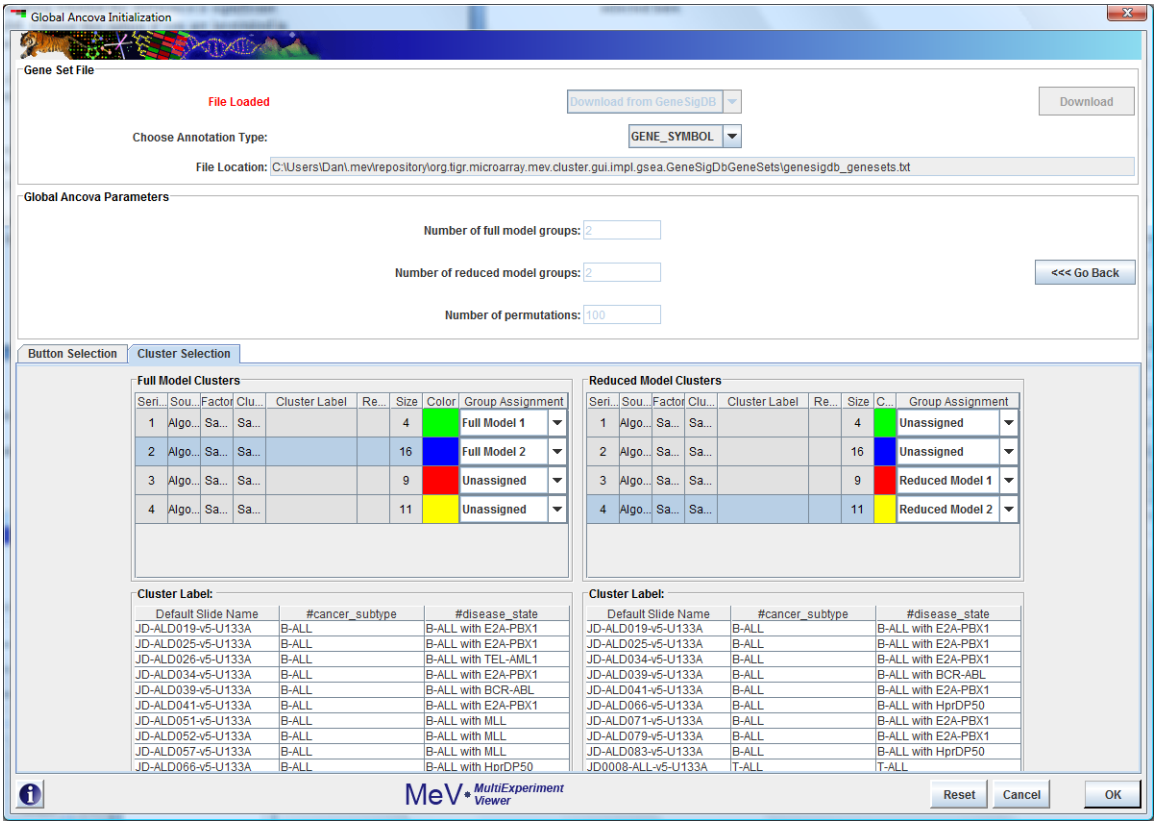
Running Global Ancova:

After opening the Global Ancova initialization dialog, select the file location of your gene set test terms. Your options are:
1.) Local file: A tab-delimited text file located on your machine.
2.) MSigDB: Download gene sets from the MIT/Broad website http://www.broad.mit.edu/gsea/msigdb/downloads.jsp.
3.) GeneSigDB: Download gene sets from GeneSigDB.

Choose an annotation type. The annotation used by your test terms must match those selected here.

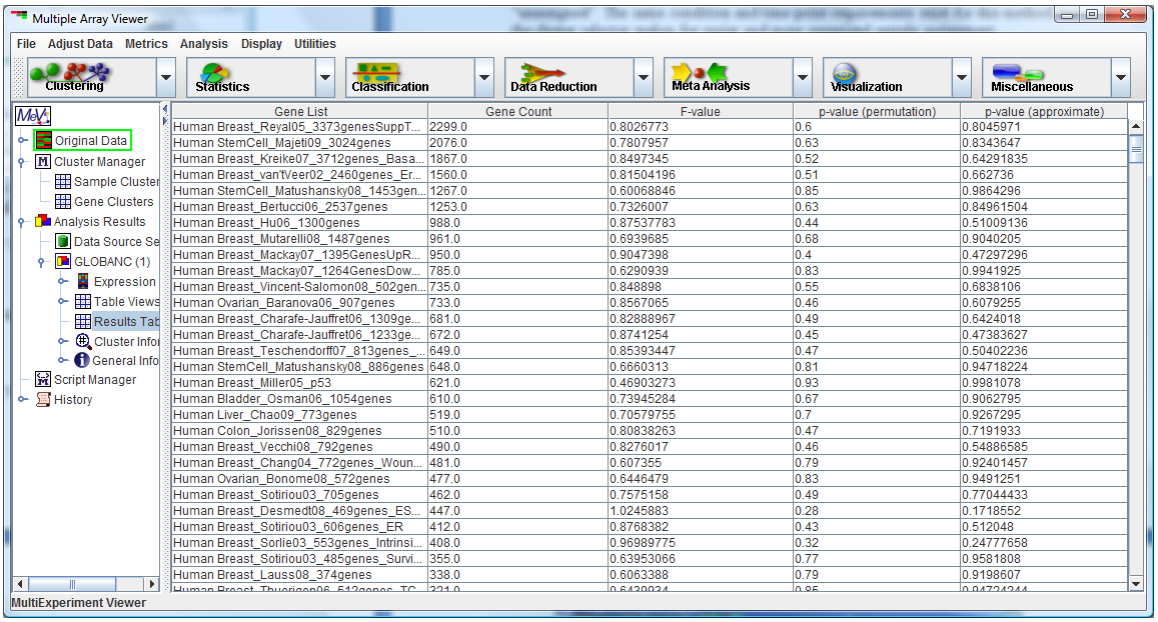
Assign your samples:
Choose the number of groups in your data and use the button or cluster selector to assign the samples to those groups.

Button Selection:
Use this option if you have not created clusters for your samples and want to assign each sample individually.



Cluster Selection:

Use the cluster selection panel if you have previously made clusters and wish to run your analysis based on those clusters. Use the drop-down boxes to choose which clusters' samples are to be assigned to which time-points. For clusters you are not using, leave "unassigned". The same condition and time-point requirements exist for this method, but the cluster selector makes for easier and more organized sample assignment.



The Global Ancova module outputs a table containg F-values, permutation-based p-values and approximation-based p-values.

Additionally, the Global Ancova module outputs standard viewers for each of the supplied genesets.

Scripting

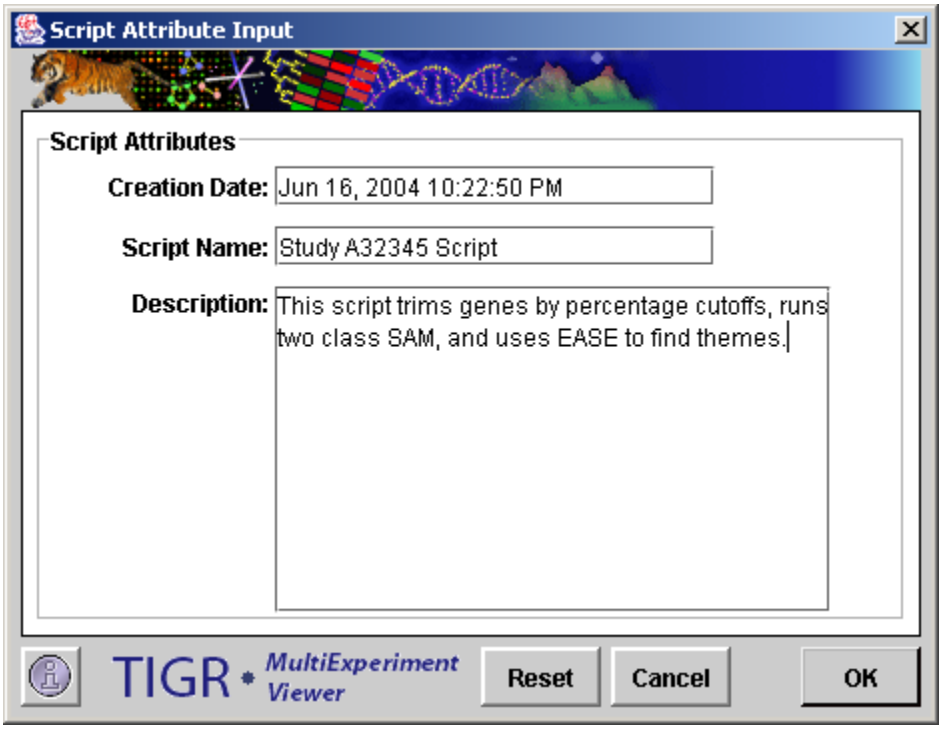
The scripting capabilities within MeV permit the execution of multiple algorithms to be performed without user oversight or intervention once processing begins. The execution steps are dictated by a user-defined script that describes the parameters to use for the selected algorithms. Scripting in MeV 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.

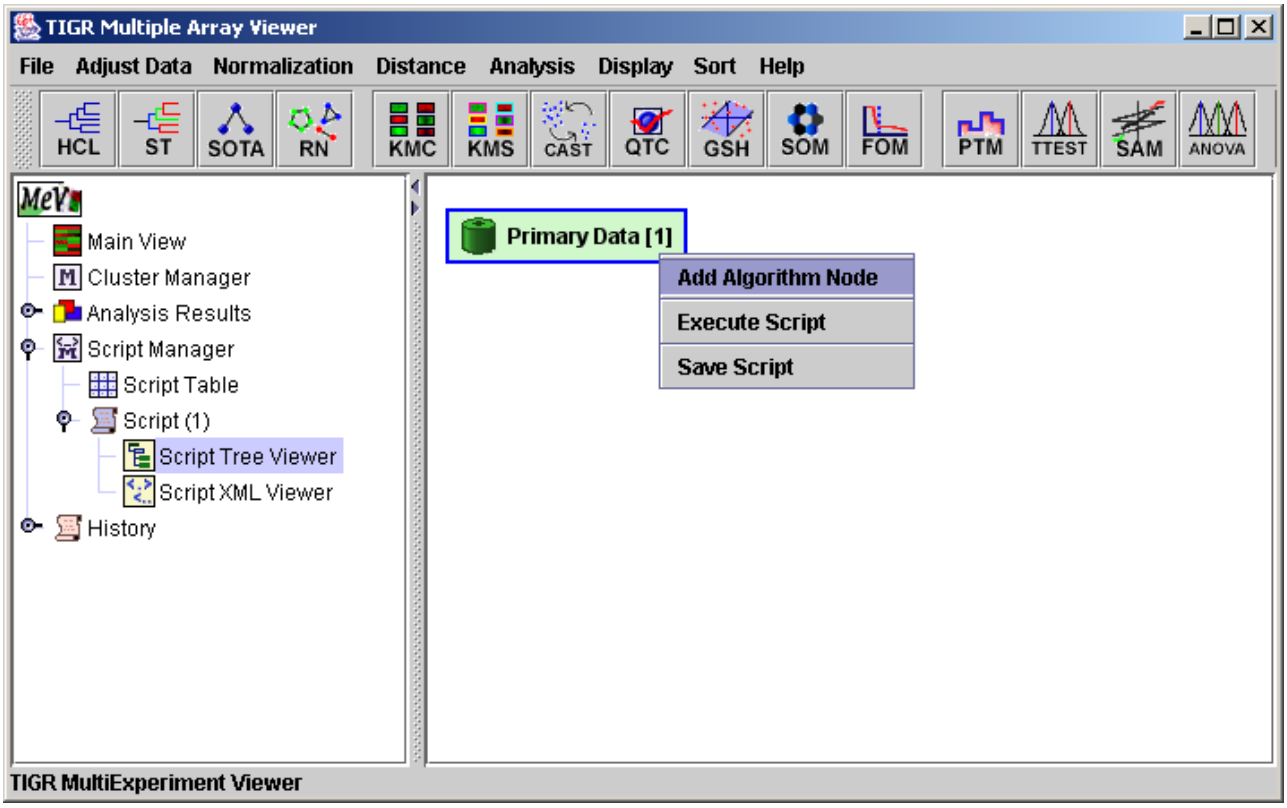


1.1. New Script Attribute Dialog

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

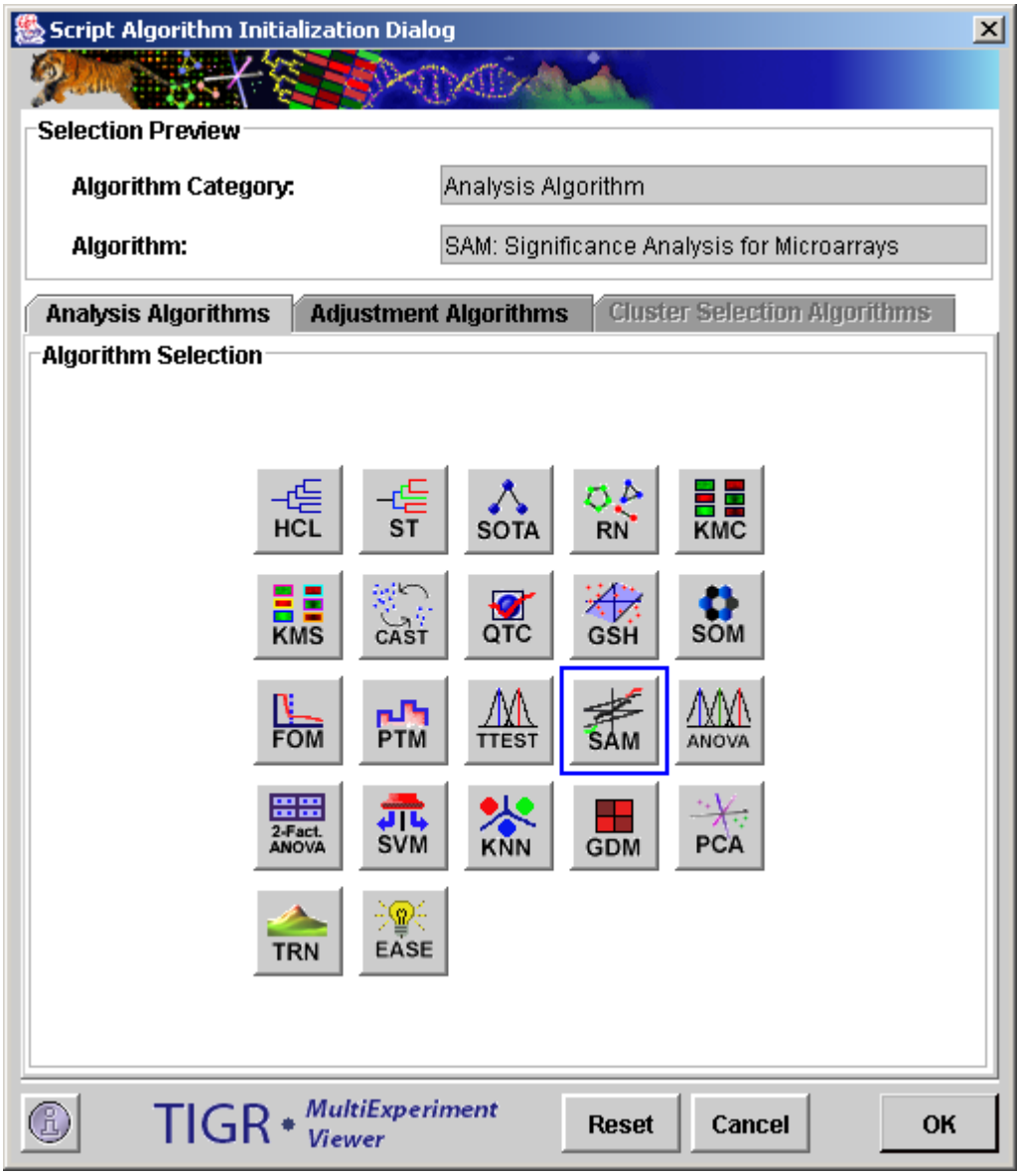
The Script Tree Viewer is the main viewer used to construct the script. The viewer's graphical nature permits the user to focus on script creation without undue consideration of complex script syntax. The Script Tree Viewer represents the script as a set of connected nodes. Each node is either a *data 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.



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

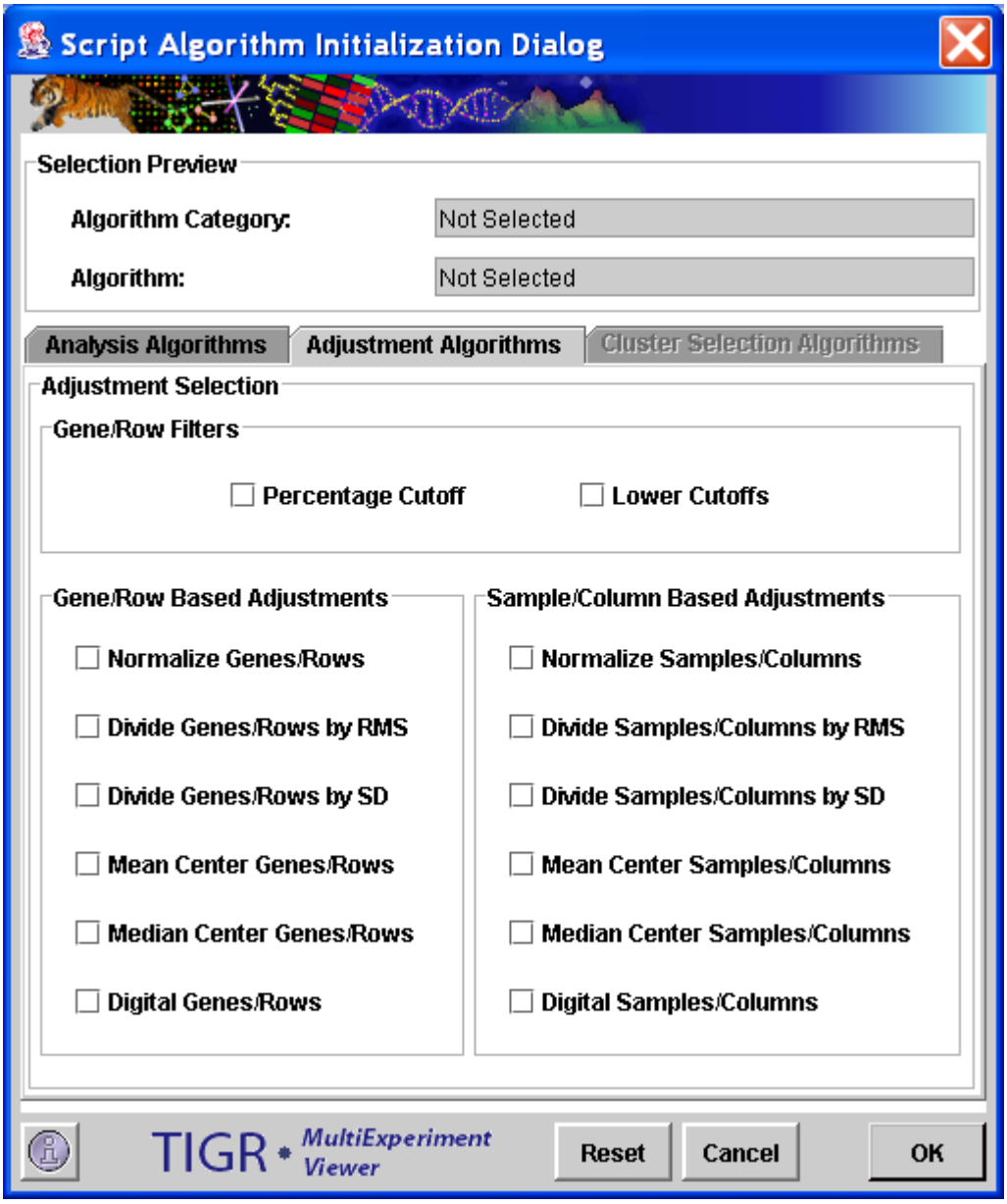
Adding an Algorithm

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.



1.3. 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.



1.4. 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

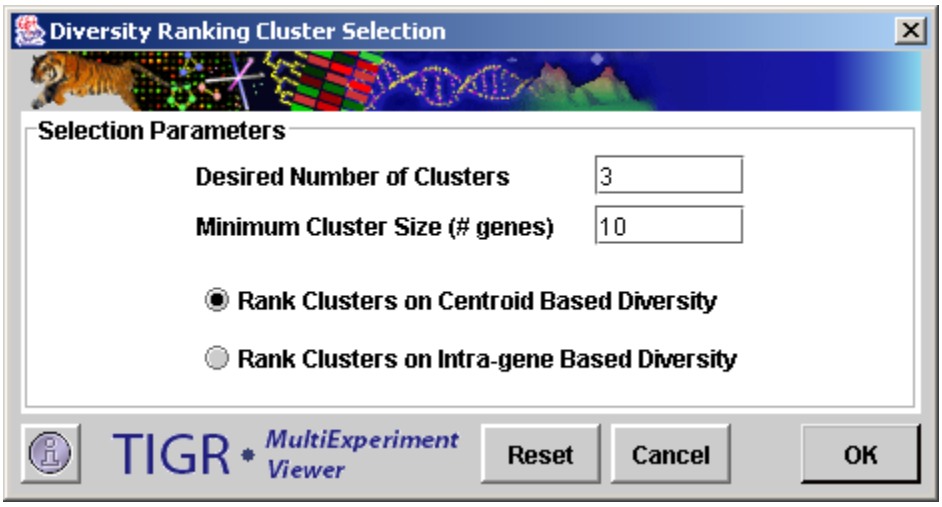
Ranking Cluster Selection two possibilities exist for determination of cluster diversity:

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

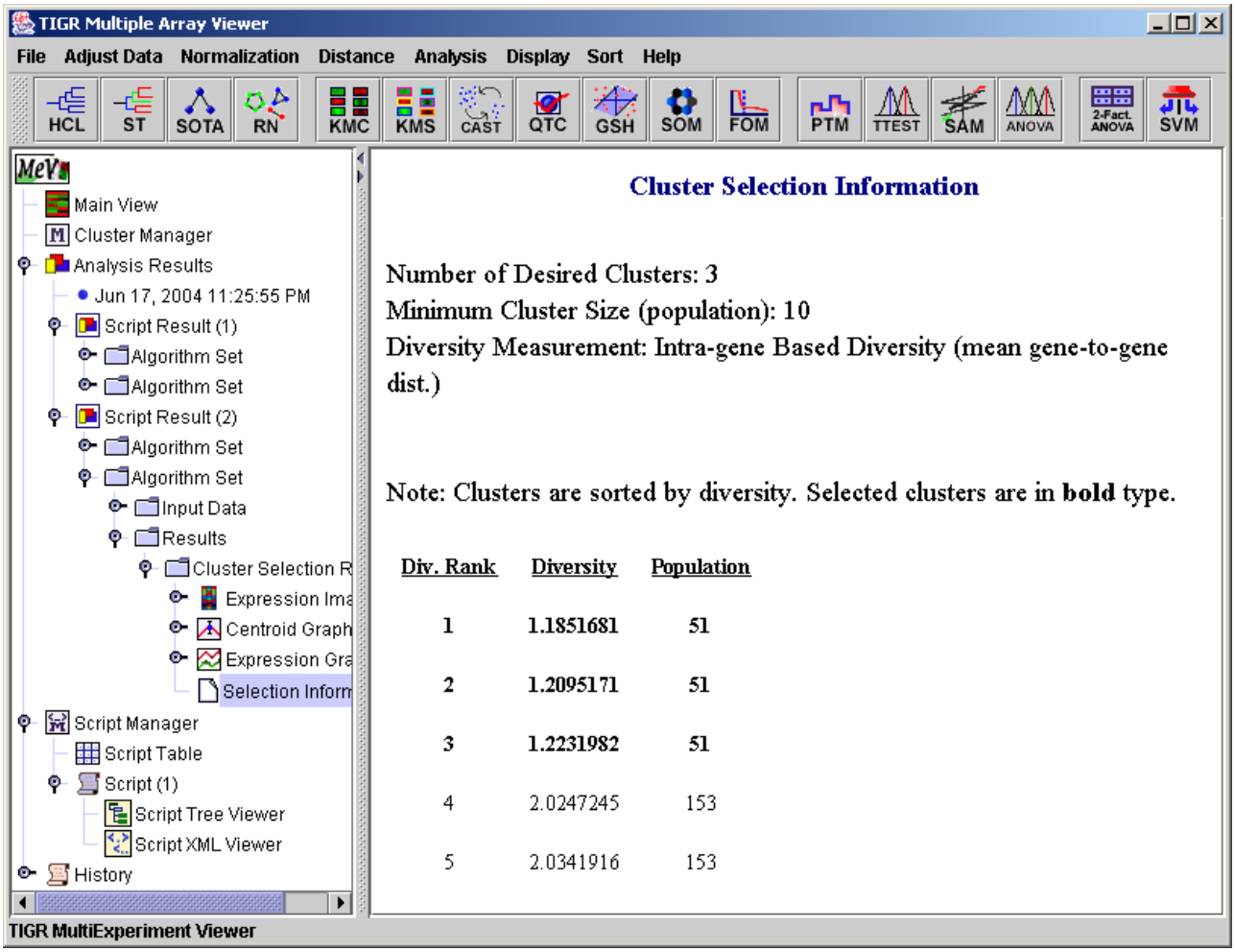
$D = \left[\sum_{i=1}^{n} dist(\vec{g}_i, \vec{c}) \right] / n$; c is the cluster centroid, g_i is the ith expression vector of n vectors.

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

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



1.5. Diversity Ranking Cluster Selection Dialog.



1.6. 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:

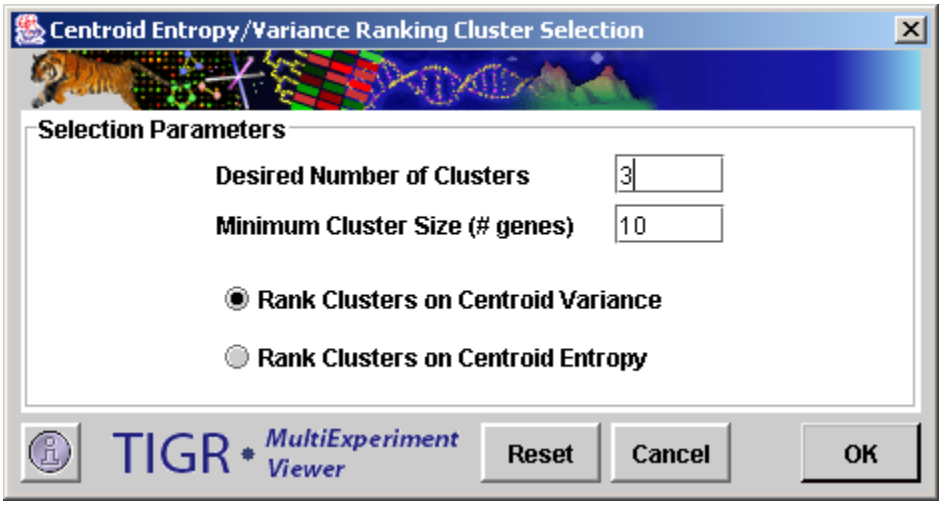
(1) Centroid Variance

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

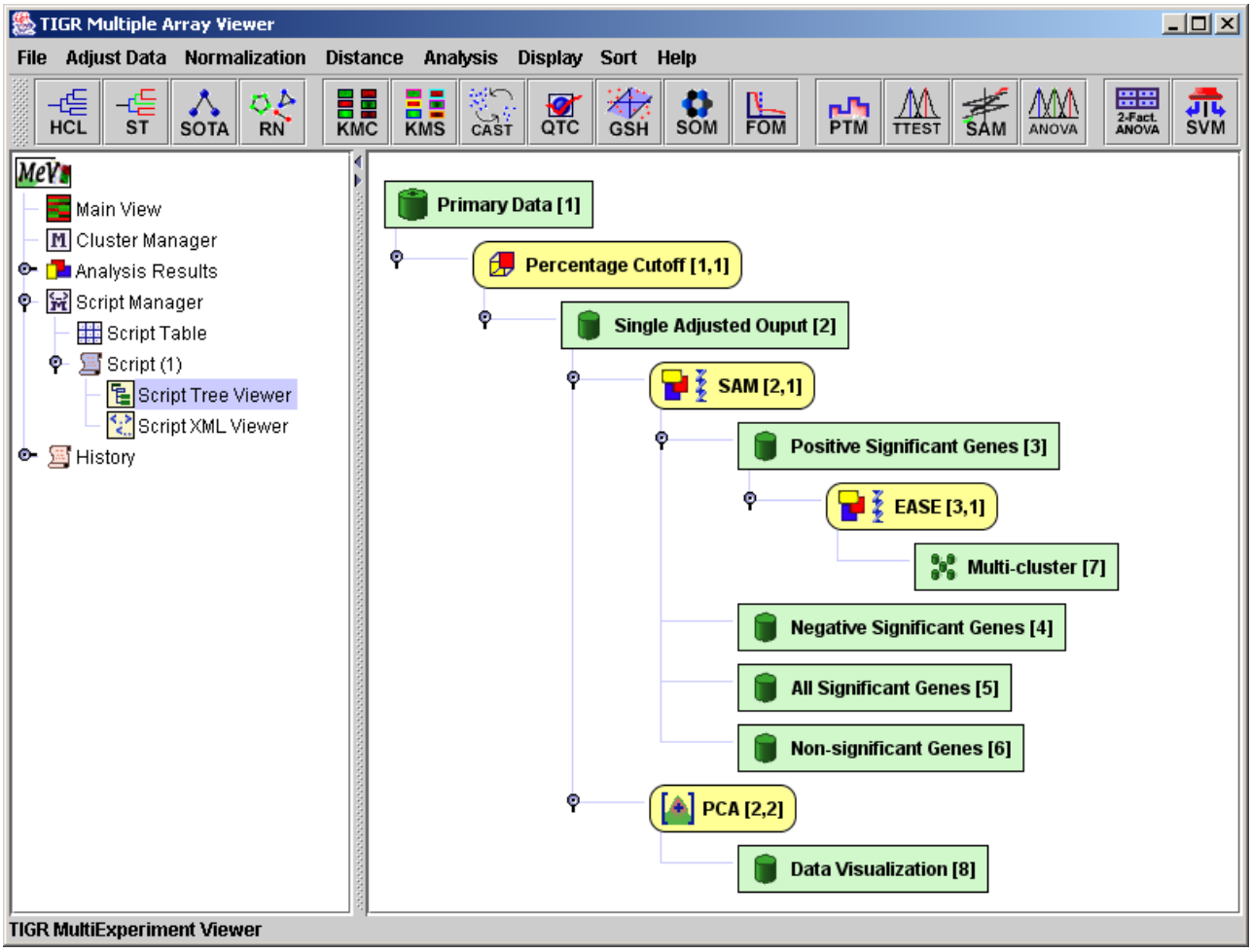
(2) Centroid Entropy

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.

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



1.7. Centroid Entropy/Variance Ranking Cluster Selection Dialog.



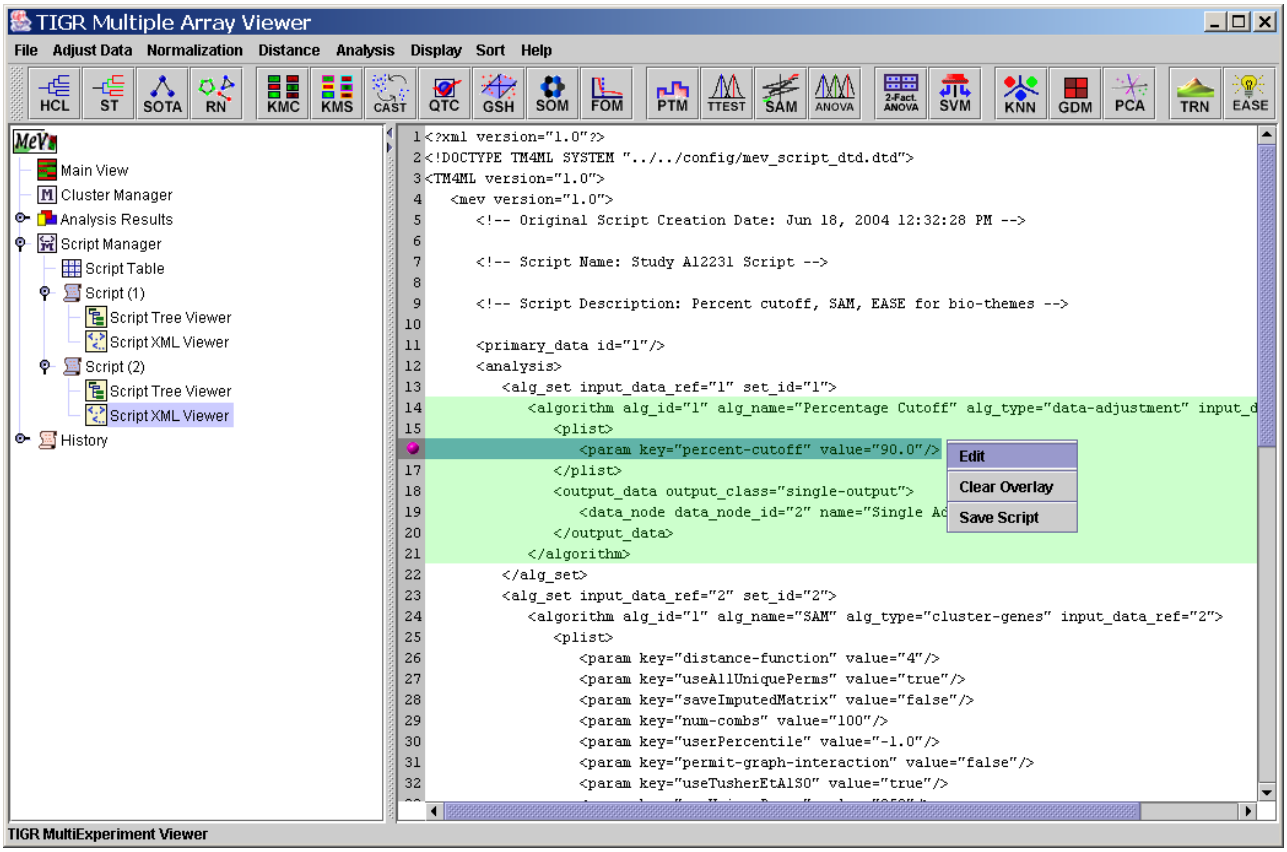
1.8. Script Tree Viewer with Constructed Script

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.

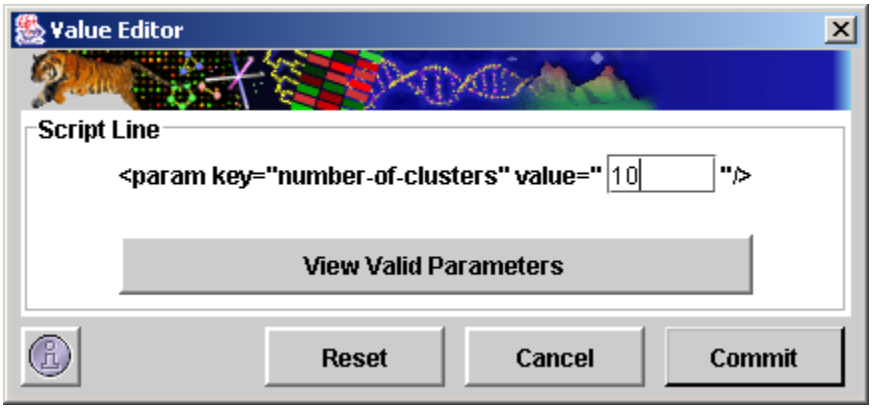


1.9. Script XML Viewer, highlighted algorithm and script line selected.

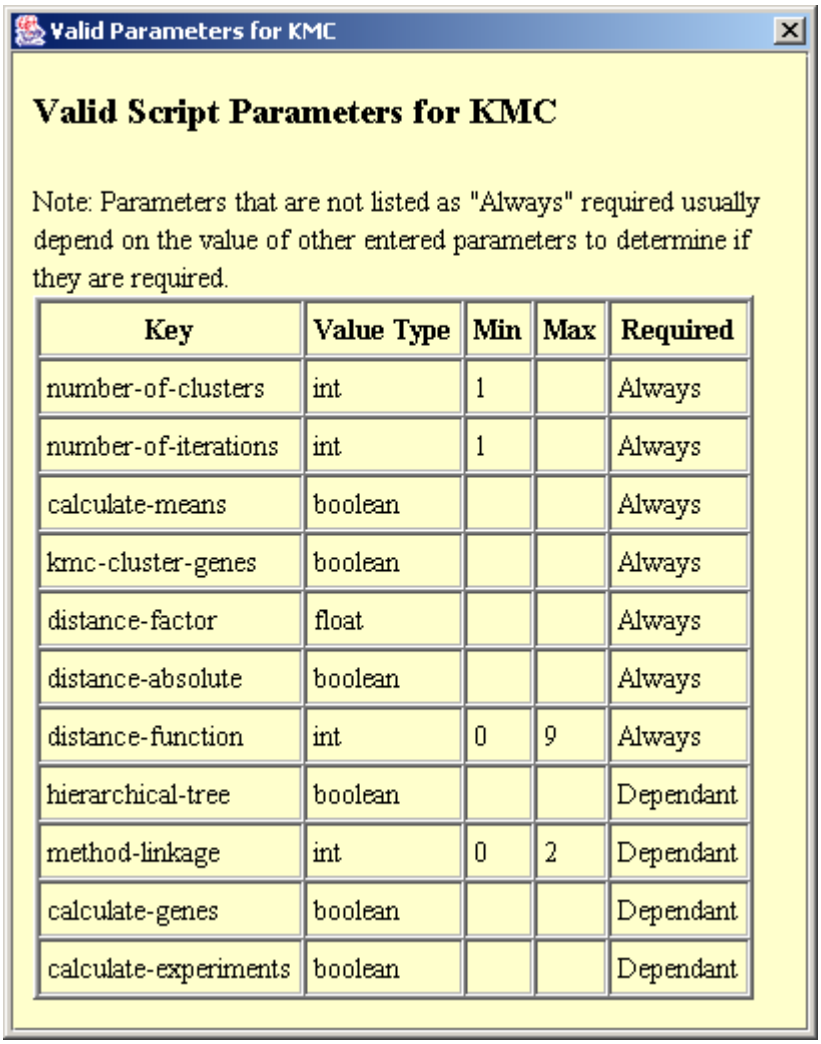
Editing Parameter Values within the XML Viewer

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.



1.10. Script Parameter Value Editor



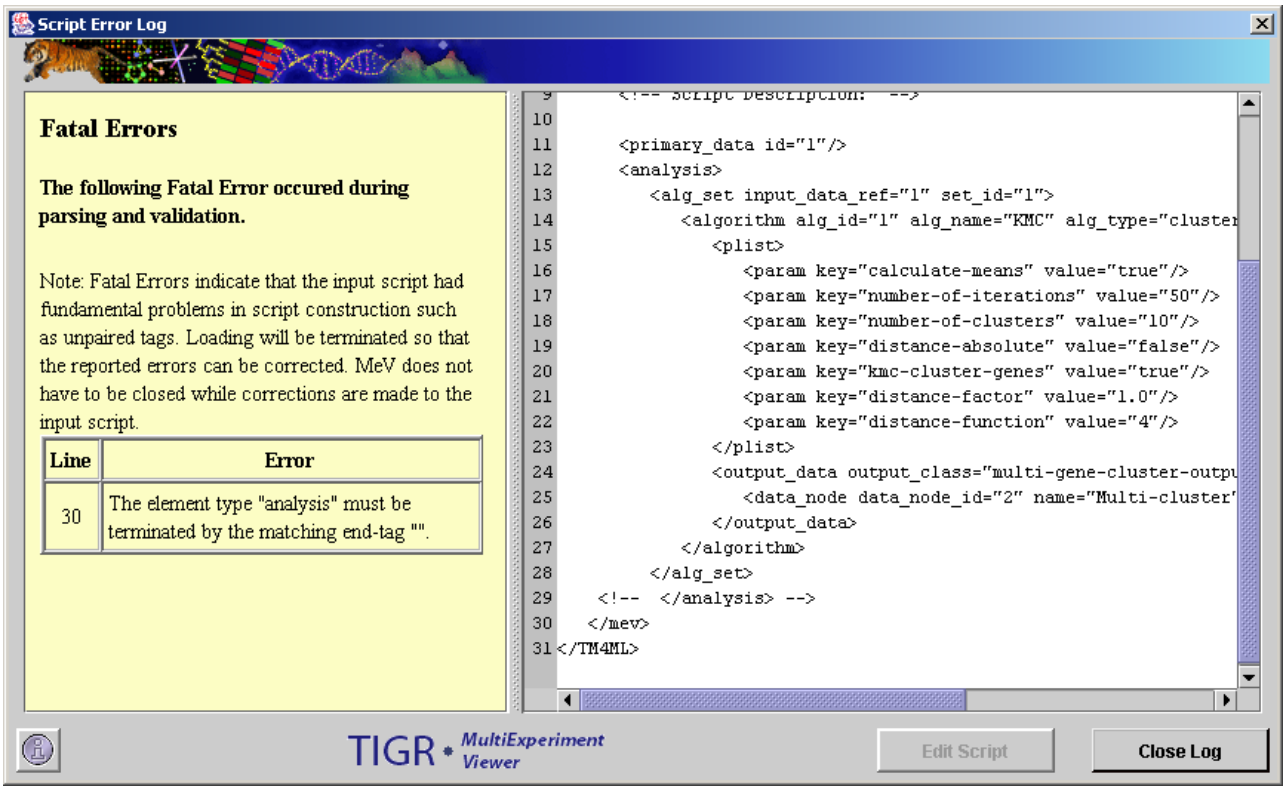
Key	Value Type	Min	Max	Required
number-of-clusters	int	1		Always
number-of-iterations	int	1		Always
calculate-means	boolean			Always
kmc-cluster-genes	boolean			Always
distance-factor	float			Always
distance-absolute	boolean			Always
distance-function	int	0	9	Always
hierarchical-tree	boolean			Dependant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

1.11. 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.

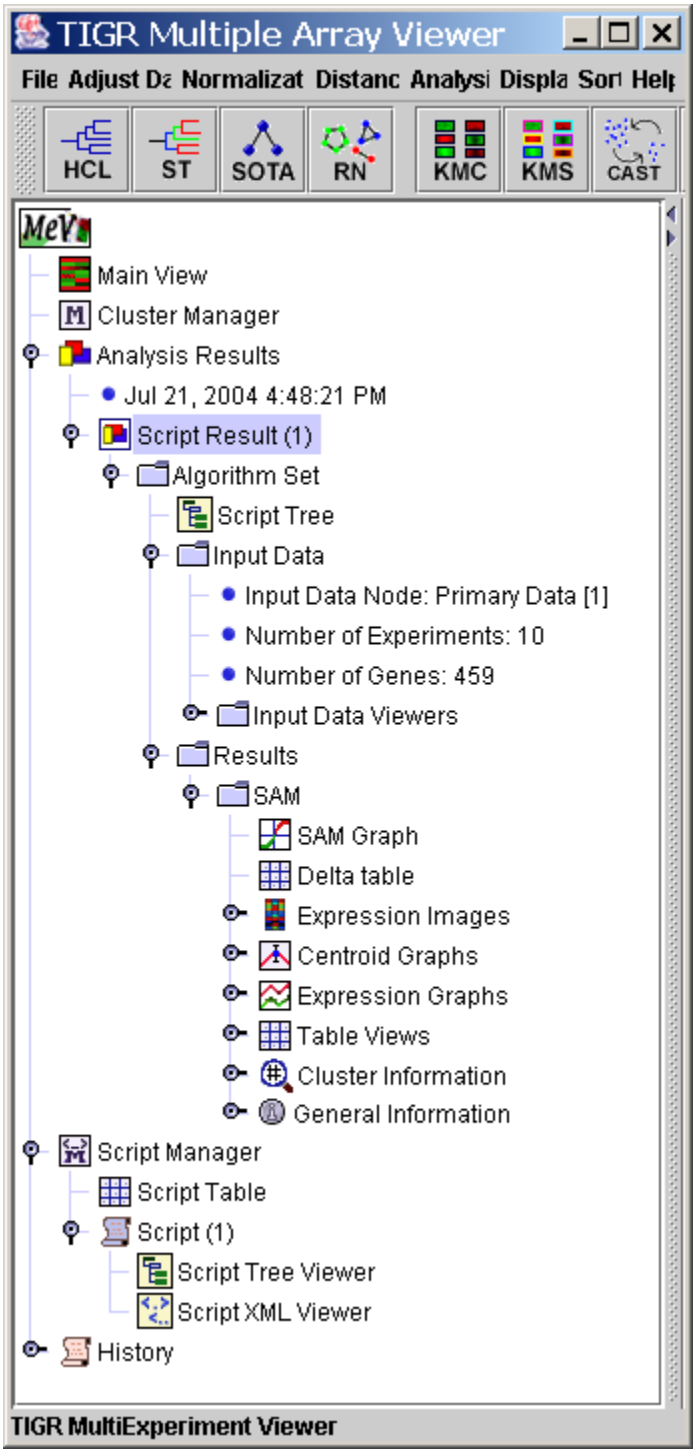


1.12. Script Error Log (with XML editor opened)

Running a Script

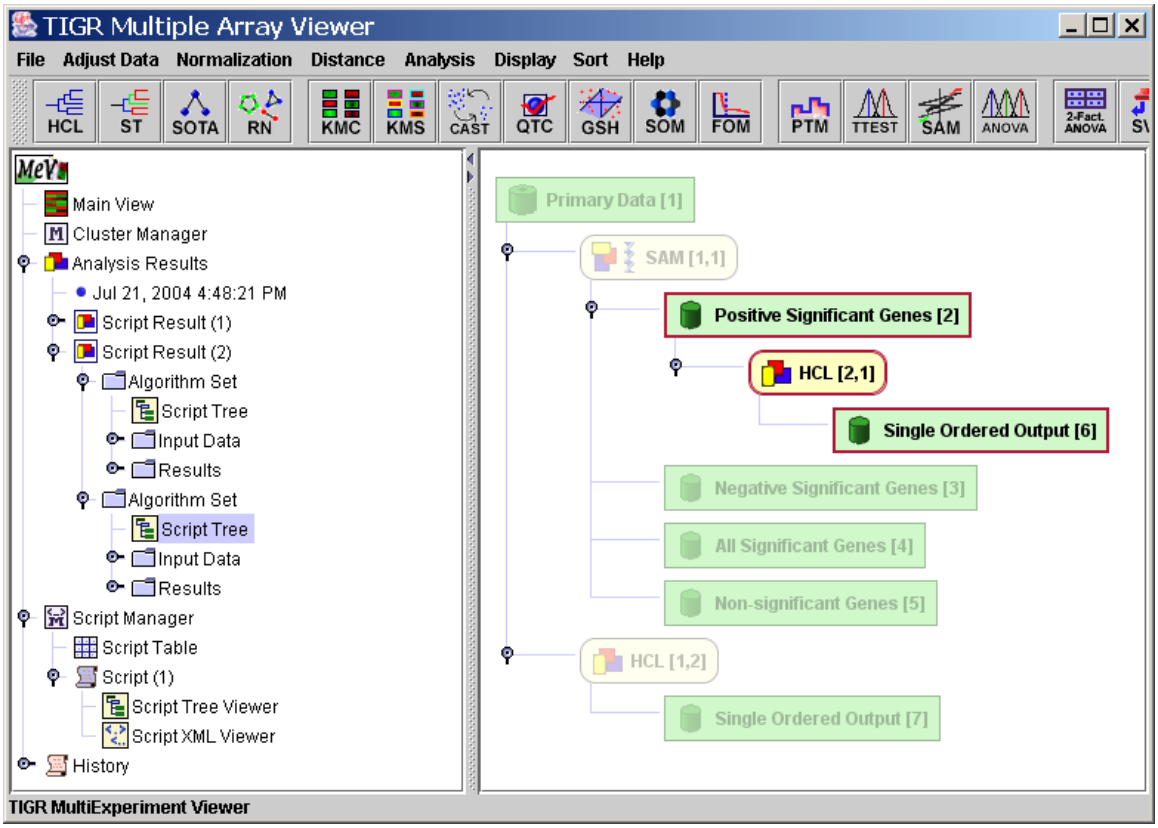
Script execution can be initiated from the Script Table viewer, the Script Tree Viewer, or the Script XML Viewer by using the right click menu option, *Execute Script*. The script is logically split up into units called *Algorithm Sets* that represent one or more algorithms sharing a common input data 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.



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

If an algorithm fails to produces 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.



2. Comparative Genomic Hybridization Viewer

Loading Experiments

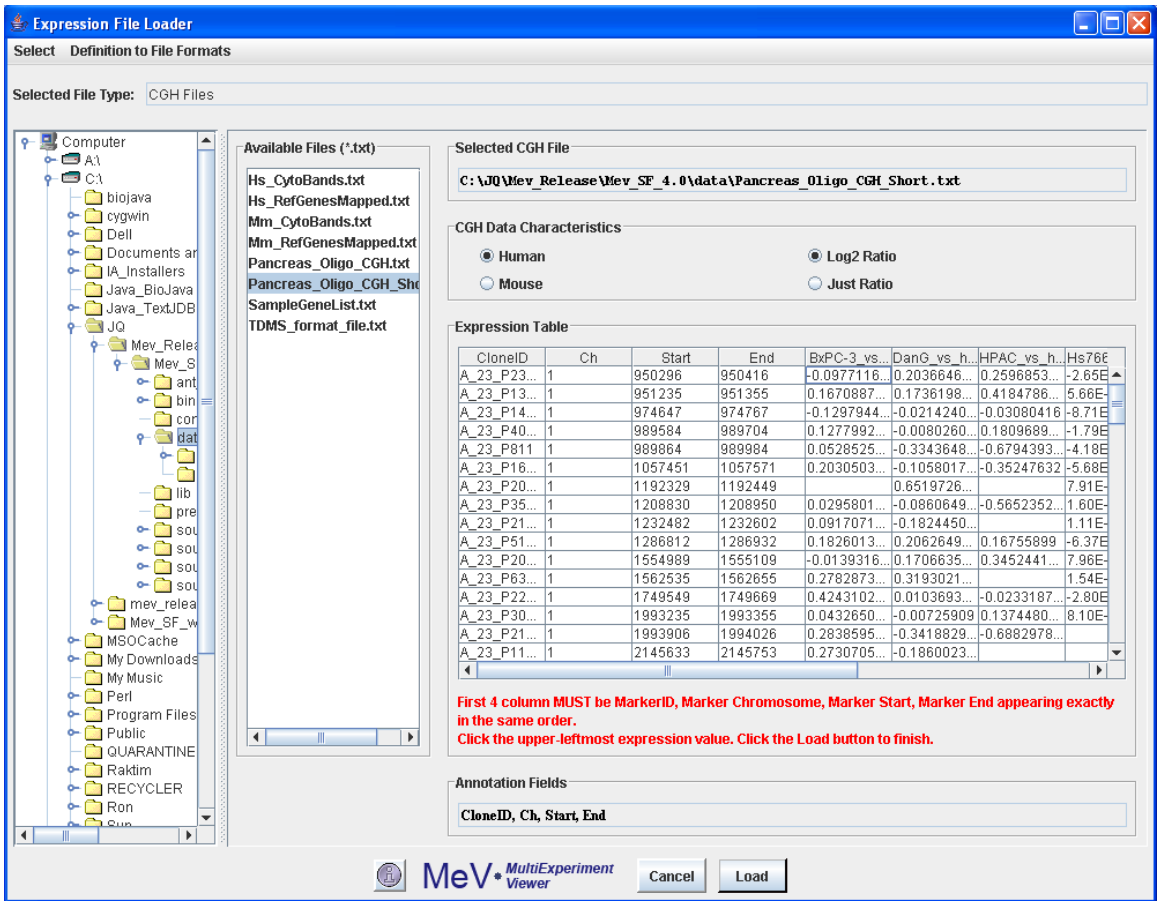
Currently CGH Analyzer is capable of loading experiments only in one generic format. We will provide other loaders to accommodate various other formats in the future. Currently CGH Viewer supports only 2 species data: Human & Mouse.

Loading from Files

CGH Analyzer allows data to be loaded from one format only. The format includes 4 mandatory columns followed by sample columns. The mandatory columns are:

1.) Probe/Marker
2.) Probe Chromosome,
3.) Probe Genomic Start in BP,
4.) Probe Genomic End in BP

The mandatory columns are followed by Sample observations where each observation for each probe is the log2 or simple intensity ratio of Cy3 & Cy5. If the observations are not log2 transformed they are done so by the module.



CGH File Loader

The protocol for loading data from files is as follows:

1.) Launch MeV (see section 1)
2.) Click File->Load Data to launch the file loading dialog.
3.) Clicl Select -> CGH to invoke the CGH loader
4.) Locate directory in which data files are located using the directory tree on left. To load a sample data set included with the MeV distribution, navigate to the installation directory of MeV, expand the Data directory, and select the CGH_sample_data.txt file. Use default settings for Species & Log status.
5.) At the bottom of the window, click "Load"

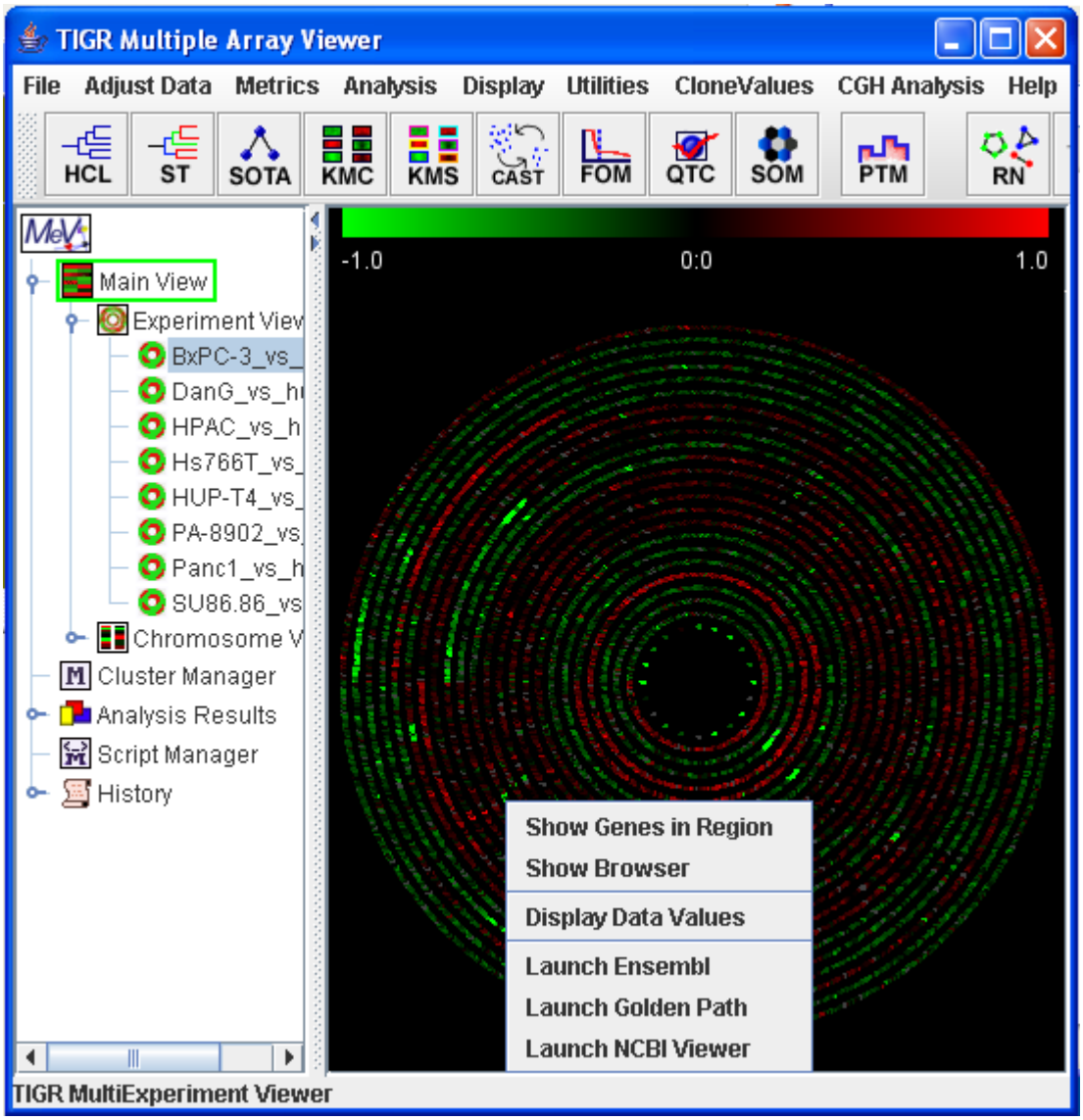
CGH Analyzer Viewers

The CGH Circle Viewer

Once experiments have been loaded, the Main View node of the navigation tree should contain a subtree called Experiment Views. Expand this subtree to display a list of all the samples that have been loaded. Clicking on any of these samples will display the CGH Circle Viewer for that sample (Figure 2-1).

The CGH Circle Viewer is a circular representation of the entire genome of a sample. This view provides easy identification of large scale abnormalities and overall aneuploidy of a sample. The display consists of 24 concentric circles, each representing a chromosome, with chromosome 1 represented by the outermost circle and chromosome Y represented by the innermost circle. Each circle is composed of a series of colored dots, each representing a probe. The probes are arranged by their linear around the genome. The p-arm of each chromosome begins at 180 degrees from the center of the display and subsequent probes are arranged clockwise by their position on the chromosome.

Click on any clone in the circle viewer to display its clone name and chromosome. Right Clicking on any region will display a menu to browse RefSeq genes in the region, Launch CGH browser on a sample, Map out to public domain sites like NCBI, etc.



Circle Viewer view of sample BxPC-2

Clone Values

Log Values

The CGH Analyzer currently allows one method of determining the value for each probe, i.e. the log2 ratio. All displays by default are a red/green ratio gradient color display. Each element is red, green, black or gray. Black elements have a log average inverted ratio of 0, while green elements have a log ratio of less than 0 and red elements have a log ratio greater than 0. The further the ratio from 0, the brighter the element is. Gray elements are missing or were determined as bad by the spot quality filtration criteria and are not used in any analyses.

By default, the lower and upper bounds of this display are -1 and 1, indicating that probes with log ratios less than or equal to -1 are shown with the maximum red intensity, and those greater than or equal to 1 are shown with the maximum green intensity. This scale can be changed, allowing for display of a wider intensity range, by using the Set Ratio Scale item in the Display menu. The colors used

can be changed by selecting Set Color Scheme from the Display menu. Notice how the color bar at the top of the display updates when these values are changed, indicating the current color scheme and ratio scale.

Thresholds

To change to discrete copy number determination based on clone ratio thresholds, select the *Set Threshold* item from the *CloneValues* menu.

Using this determination, each clone is assigned a copy number determination and corresponding color based on the criteria shown in table 2-1.

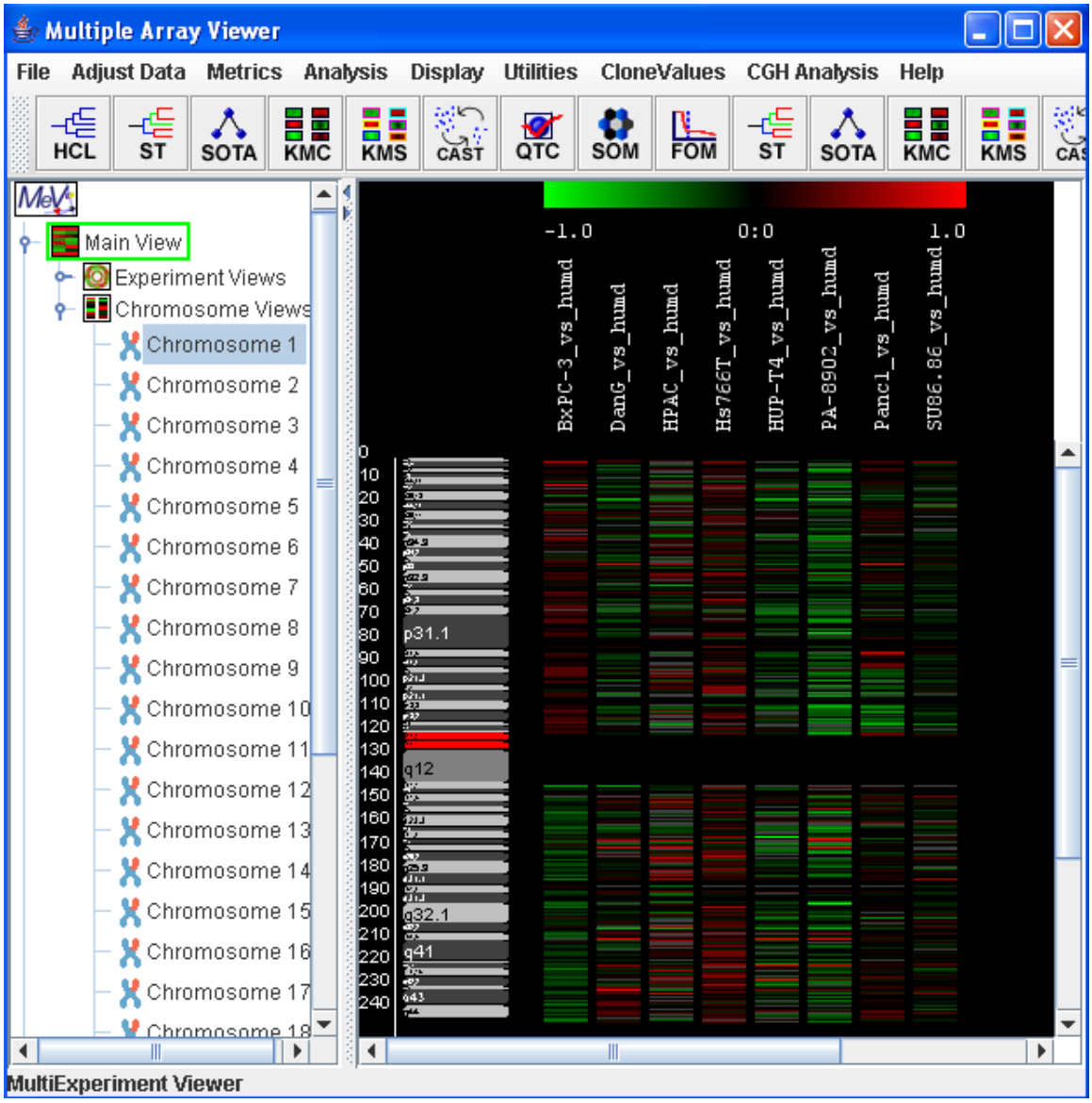
Copy Number	Color	Log2 Ratio	Other
2 Copy Deletion	Pink	< Deletion Threshold 2 Copy	
1 Copy Deletion	Red	< Deletion Threshold	Not 2 Copy Deletion
2 Copy (or greater) Amplification	Yellow	> Amplification Threshold 2 Copy	
1 Copy Amplification	Green	> Amplification Threshold	Not 2 Copy Amplification
No Copy Change	Blue		Not Deleted or Amplified
Bad Clone	Grey		

Discrete copy number determination based on probe log2 ratio thresholds

The CGH Position Graph

The *Main View* node of the navigation tree should contain a subtree called *Chromosome Views* (Figure 2-2). Expand this subtree to display a list of all chromosomes. Clicking on any of these chromosomes will display the CGH Position Graph Viewer for that chromosome.

The CGH Position Graph Viewer is used to display data values for a single chromosome for multiple experiments. The left side of this view displays the cytogenetic bands of the selected chromosome. Positional coordinates, in MB, are annotated to the left of the cytobands. Probes are represented as horizontal bars beginning and ending at positions corresponding to the genomic coordinates of the clone.

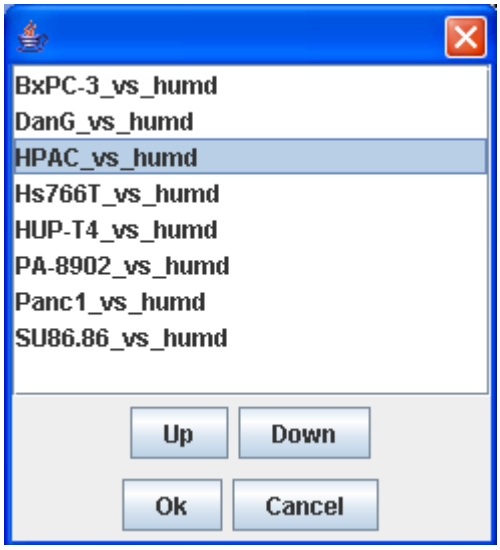


CGH Position Graph view of chromosome 2.

Probes in this display are colored the same way as described for the Circle Viewer. Clone values, color schemes, and ratio scales can be adjusted as described in section **Error! Reference source not found.**.

Changing Experiment Order

The order in which experiment appear in the display can be changed by using the *Display Order* item in the *Display* menu (Figure 2-3). The position of samples can be moved up and down using the buttons on the bottom of this dialog, and selecting *Ok* will cause the experiments to be displayed using the new order.



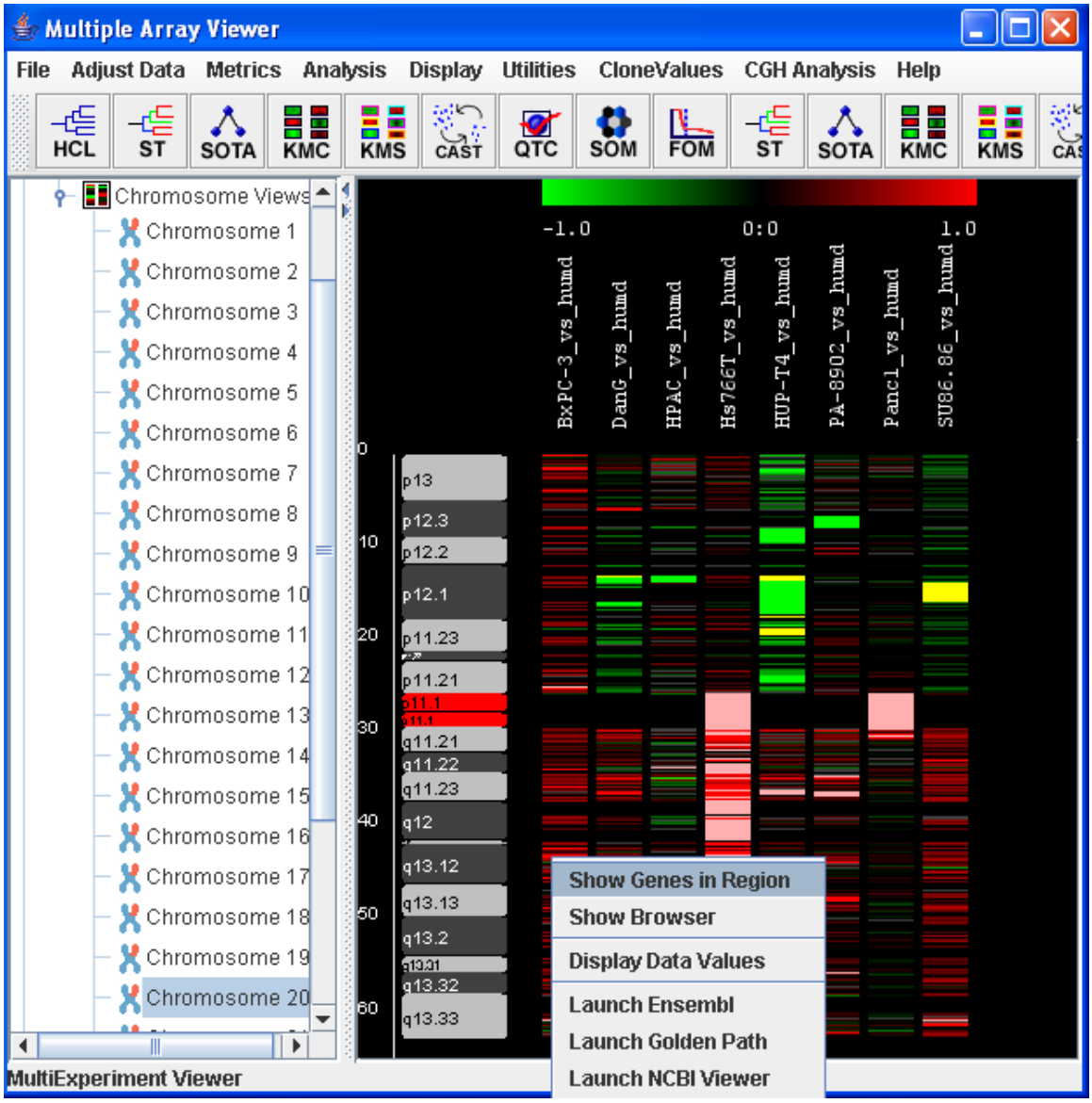
The display order changer

Element Size

The width and length of the probes can be changed through the *Element Length* and *Element Width* items in the *Display* menu. By default the width and length are calculated to fit the entire display on the screen. It is often useful to increase the length to look at a particular region because it is often difficult to distinguish probes that lie close to each other.

Flanking Regions

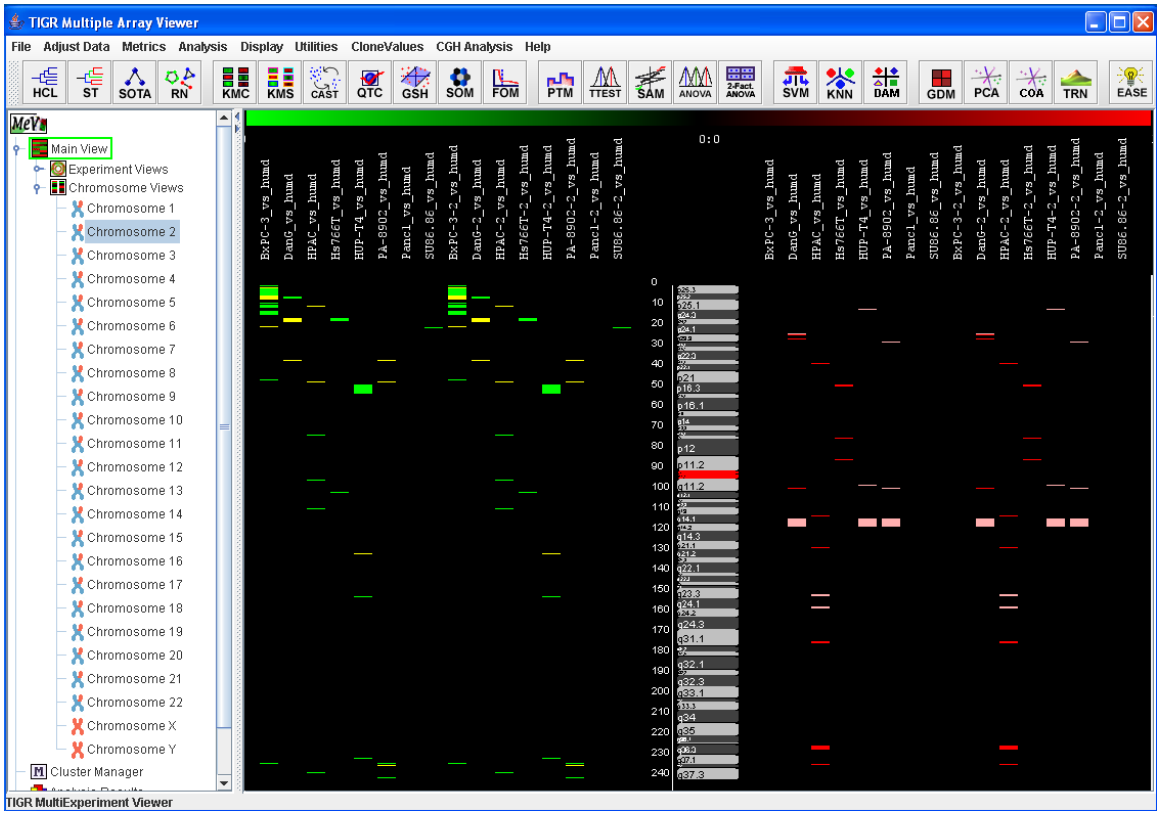
CGH arrays are used to determine a copy number profile throughout the genome. In expression arrays, the values of importance are usually genes that are covered by probes, but in CGH arrays, the regions that lie between probes are often as important as those that are covered. Therefore, unless a CGH array has complete genomic coverage, it is important to interpolate copy number change in the regions not covered by probes. Flanking regions also allow experiments to be analyzed together that were generated using different arrays. Flanking regions are used to approximate a complete genome copy number profile of each sample. Flanking regions rely on assigning a discrete copy determination to each probe. A region between two probes is considered altered if either of the probes that "flank" that region is altered. If a data region is flanked by one or more "deleted" probes, the region is considered deleted, and if it is flanked by one or more "amplified" probes, the region is considered amplified. If a region is flanked by one deleted and one amplified clone, the region is considered as deleted and amplified, allowing for maximum flexibility in algorithms that use flanking regions. Flanking regions can be toggled through the *Flanking Regions* checkbox item in the *Display* menu. Right clicking on any flanking regions (Figure 2-4) will allow for querying of the genes containing in the flanking region, querying the intensity ratios of the probes that make up the region, and to link to the CGH Browser.



CGH Position Graph view of chromosome 20 with flanking regions.

The Separated Viewer

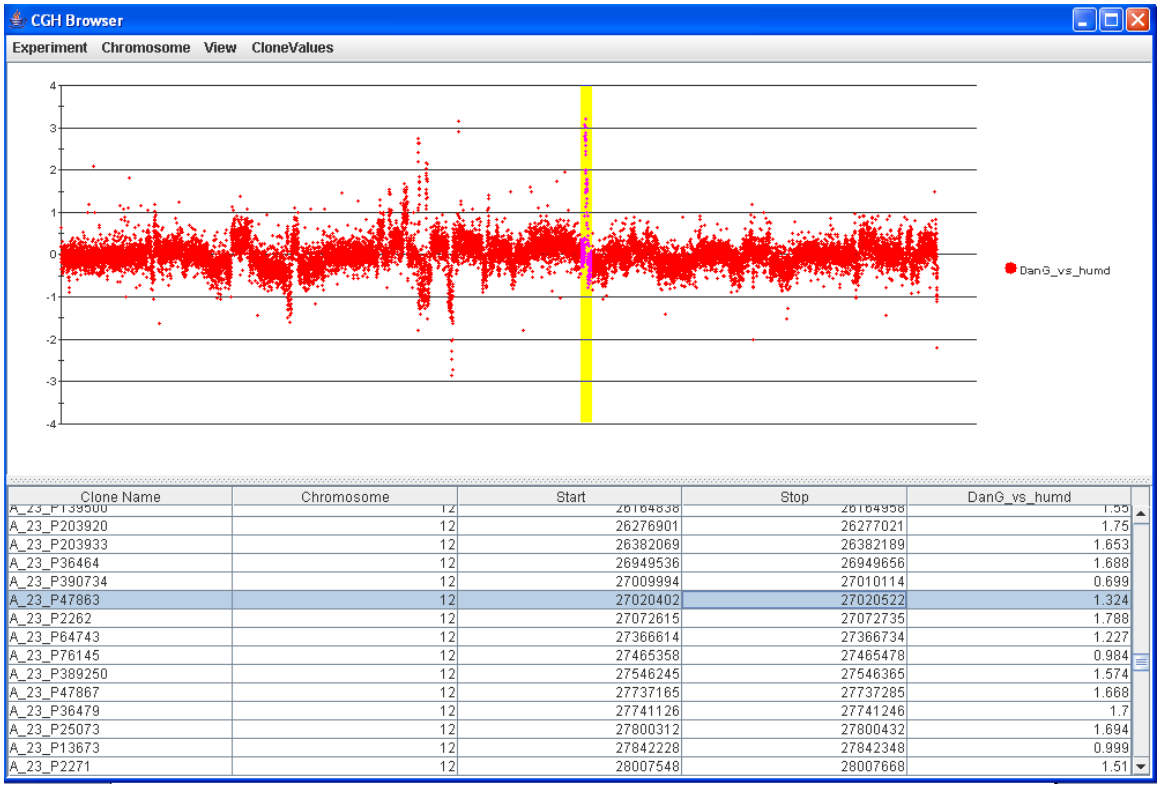
A common way to display CGH data is to draw all deletions on one side of the screen and all amplification on the other side. The separated view of the CGH Postion Graph displays the cytogenetic bands and chromosome positions in the center of the panel, the flanking regions corresponding to deletions on the left of the screen and the flanking regions corresponding to amplifications on the right side of the screen. To display this view, in the *Display* menu, select *Display Type -> Separated* (Figure 2-5). This display often looks better if the element width item is set to a smaller value.



View of chromosome 2.

The CGH Browser

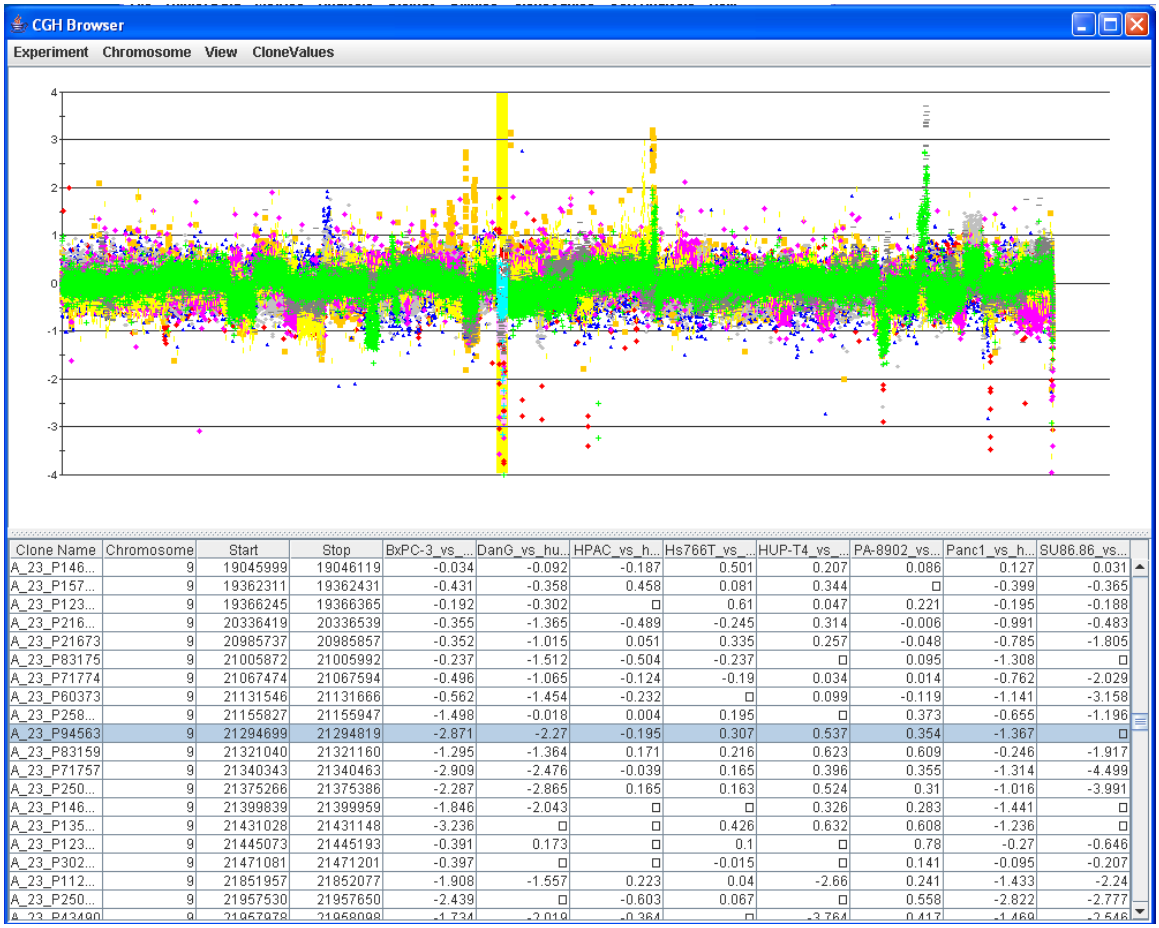
The CGH Browser displays a plot representation of one or more CGH experiments. Right clicking on any flanking region (Figure 2-4) or probe in the CGH Position Graph Viewer, or on any probe in the CGH Circle Viewer and selecting *Show Browser* will launch the CGH Browser with the values corresponding to the selected data region highlighted on both the chart and the table (Figure 2-6).



Clone Name	Chromosome	Start	Stop	DanG_vs_humd
A_23_P139500	12	26164838	26164958	1.55
A_23_P203920	12	26276901	26277021	1.75
A_23_P203933	12	26382069	26382189	1.653
A_23_P36464	12	26949536	26949656	1.688
A_23_P390734	12	27009994	27010114	0.699
A_23_P47863	12	27020402	27020522	1.324
A_23_P2262	12	27072615	27072735	1.788
A_23_P64743	12	27366614	27366734	1.227
A_23_P76145	12	27465358	27465478	0.984
A_23_P389250	12	27546245	27546365	1.574
A_23_P47867	12	27737165	27737285	1.668
A_23_P36479	12	27741126	27741246	1.7
A_23_P25073	12	27800312	27800432	1.694
A_23_P13673	12	27842228	27842348	0.999
A_23_P2271	12	28007548	28007668	1.51

CGH Browser with the data region selected.

The *Experiment* menu of the CGH Browser can be used to toggle the display between each experiment that has been loaded, or all experiments (Figure 2-7). The *Chromosome* menu of the CGH Browser can be used to toggle the display between one chromosome or all chromosomes (Figure 2-6).



Clone Name	Chromosome	Start	Stop	BxPC-3_vs_...	DanG_vs_hu...	HPAC_vs_h...	Hs766T_vs_...	HUP-T4_vs_...	PA-8902_vs...	Panc1_vs_h...	SU86.86_vs...
A_23_P146...	9	19045999	19046119	-0.034	-0.092	-0.187	0.501	0.207	0.086	0.127	0.031
A_23_P157...	9	19362311	19362431	-0.431	-0.358	0.458	0.081	0.344	□	-0.399	-0.365
A_23_P123...	9	19366245	19366365	-0.192	-0.302	□	0.61	0.047	0.221	-0.195	-0.188
A_23_P216...	9	20336419	20336539	-0.355	-1.365	-0.489	-0.245	0.314	-0.006	-0.991	-0.483
A_23_P21673	9	20985737	20985857	-0.352	-1.015	0.051	0.335	0.257	-0.048	-0.785	-1.805
A_23_P83175	9	21005872	21005992	-0.237	-1.512	-0.504	-0.237	□	0.095	-1.308	□
A_23_P71774	9	21067474	21067594	-0.496	-1.065	-0.124	-0.19	0.034	0.014	-0.762	-2.029
A_23_P60373	9	21131546	21131666	-0.562	-1.454	-0.232	□	0.099	-0.119	-1.141	-3.158
A_23_P258...	9	21155827	21155947	-1.498	-0.018	0.004	0.195	□	0.373	-0.655	-1.196
A_23_P94563	9	21294699	21294819	-2.871	-2.27	-0.195	0.307	0.537	0.354	-1.367	□
A_23_P83159	9	21321040	21321160	-1.295	-1.364	0.171	0.216	0.623	0.609	-0.246	-1.917
A_23_P71757	9	21340343	21340463	-2.909	-2.476	-0.039	0.165	0.396	0.355	-1.314	-4.499
A_23_P250...	9	21375266	21375386	-2.287	-2.865	0.165	0.163	0.524	0.31	-1.016	-3.991
A_23_P146...	9	21399839	21399959	-1.846	-2.043	□	□	0.326	0.283	-1.441	□
A_23_P135...	9	21431028	21431148	-3.236	□	□	0.426	0.632	0.608	-1.236	□
A_23_P123...	9	21445073	21445193	-0.391	0.173	□	0.1	□	0.78	-0.27	-0.646
A_23_P302...	9	21471081	21471201	-0.397	□	□	-0.015	□	0.141	-0.095	-0.207
A_23_P112...	9	21851957	21852077	-1.908	-1.557	0.223	0.04	-2.66	0.241	-1.433	-2.24
A_23_P250...	9	21957530	21957650	-2.439	□	-0.603	0.067	□	0.558	-2.822	-2.777
A_23_P43490	9	21957979	21958098	-1.734	-2.019	-0.364	□	-3.764	0.417	-1.469	-2.546

Log Ratios for chromosome 2 of all experiments.

Clicking anywhere on the chart will highlight the data point closest to the selection, as well as the corresponding row in the table. Selecting any number of rows in the table will highlight the corresponding region in the chart. The *View* menu can be used to change annotations and display styles in the browser.

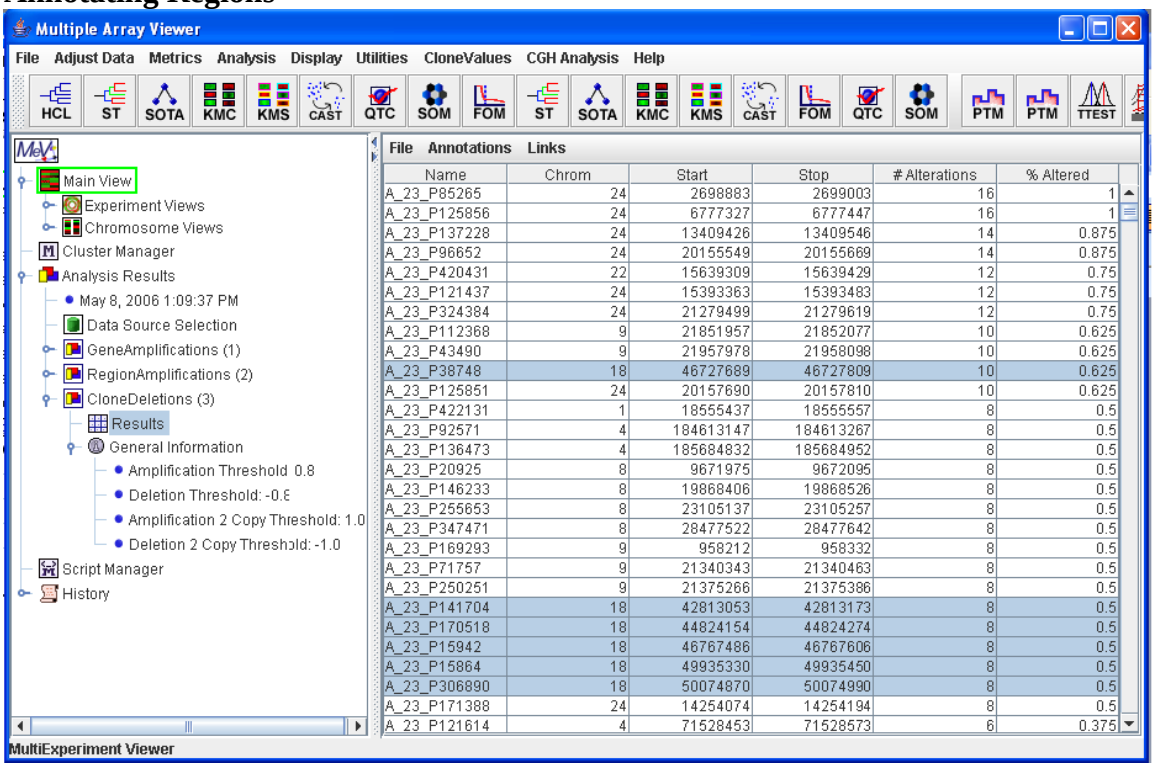
CGH Analysis

The *CGH Analysis* menu contains a number of algorithms for searching for data regions that are consistently altered throughout the experiments. These algorithms can be performed on probes, genes, and data regions (minimal common regions of alteration).

Algorithms on Probes

The items *CloneAmplifications*, *CloneDeletions*, *CloneAmplifications2Copy*, *CloneDeletions2Copy*, are used to search for probes that are commonly altered throughout the experiments. Click on the *CloneDeletions* item. Notice that a subtree has been added to the Analysis node of the navigation tree on the left side of the screen. Expanding this tree and selecting the *Results* node will set the main view to display a table showing the number and percentage of experiments in which each clone is deleted.

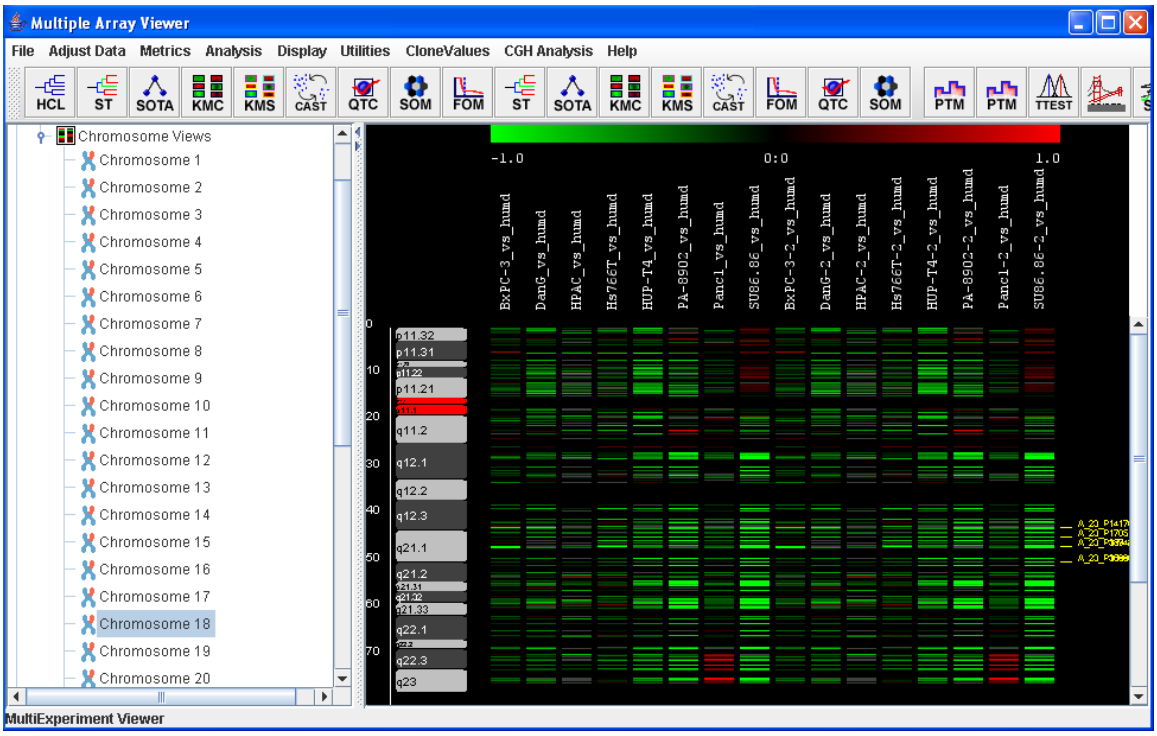
Annotating Regions



Name	Chrom	Start	Stop	# Alterations	% Altered
A_23_P85265	24	2698883	2699003	16	1
A_23_P125856	24	6777327	6777447	16	1
A_23_P137228	24	13409426	13409546	14	0.875
A_23_P96652	24	20155549	20155669	14	0.875
A_23_P420431	22	15639309	15639429	12	0.75
A_23_P121437	24	15393363	15393483	12	0.75
A_23_P324384	24	21279499	21279619	12	0.75
A_23_P112368	9	21851957	21852077	10	0.625
A_23_P43490	9	21957978	21958098	10	0.625
A_23_P38748	18	46727689	46727809	10	0.625
A_23_P125851	24	20157690	20157810	10	0.625
A_23_P422131	1	18555437	18555557	8	0.5
A_23_P92571	4	184613147	184613267	8	0.5
A_23_P136473	4	185684832	185684952	8	0.5
A_23_P20925	8	9671975	9672095	8	0.5
A_23_P146233	8	19868406	19868526	8	0.5
A_23_P255653	8	23105137	23105257	8	0.5
A_23_P347471	8	28477522	28477642	8	0.5
A_23_P169293	9	958212	958332	8	0.5
A_23_P71757	9	21340343	21340463	8	0.5
A_23_P250251	9	21375266	21375386	8	0.5
A_23_P141704	18	42813053	42813173	8	0.5
A_23_P170518	18	44824154	44824274	8	0.5
A_23_P15942	18	46767486	46767606	8	0.5
A_23_P15864	18	49935330	49935450	8	0.5
A_23_P306890	18	50074870	50074990	8	0.5
A_23_P171388	24	14254074	14254194	8	0.5
A_23_P121614	4	71528453	71528573	6	0.375

Clone deletions display with chromosome 4 deletions selected to be annotated.

Highlight all of the probes on chromosome 1 with 4 or more alterations and select the *Annotate Selected* item in the *Annotations* menu (Figure 3-1). This will set the selected data regions to be annotated in the corresponding CGH Position graph (Figure 3-2). Click on the *Chromosome 1* item on the *Chromosome Views* subtree of the navigation tree on the left side of the screen to see the selected probes annotated. The element length may have to be changed to view all annotated probes.



CGH Position graph of chromosome 1 with probes annotated that are deleted in 4 or more samples.

Right clicking on an annotation allows for querying of genes containing in the region, and to link to the CGH Browser with the selected annotation highlighted. If the CGH Browser corresponding to an annotation is displayed, it will display the log average inverted clone values for all experiments for the chromosome corresponding to the annotation.

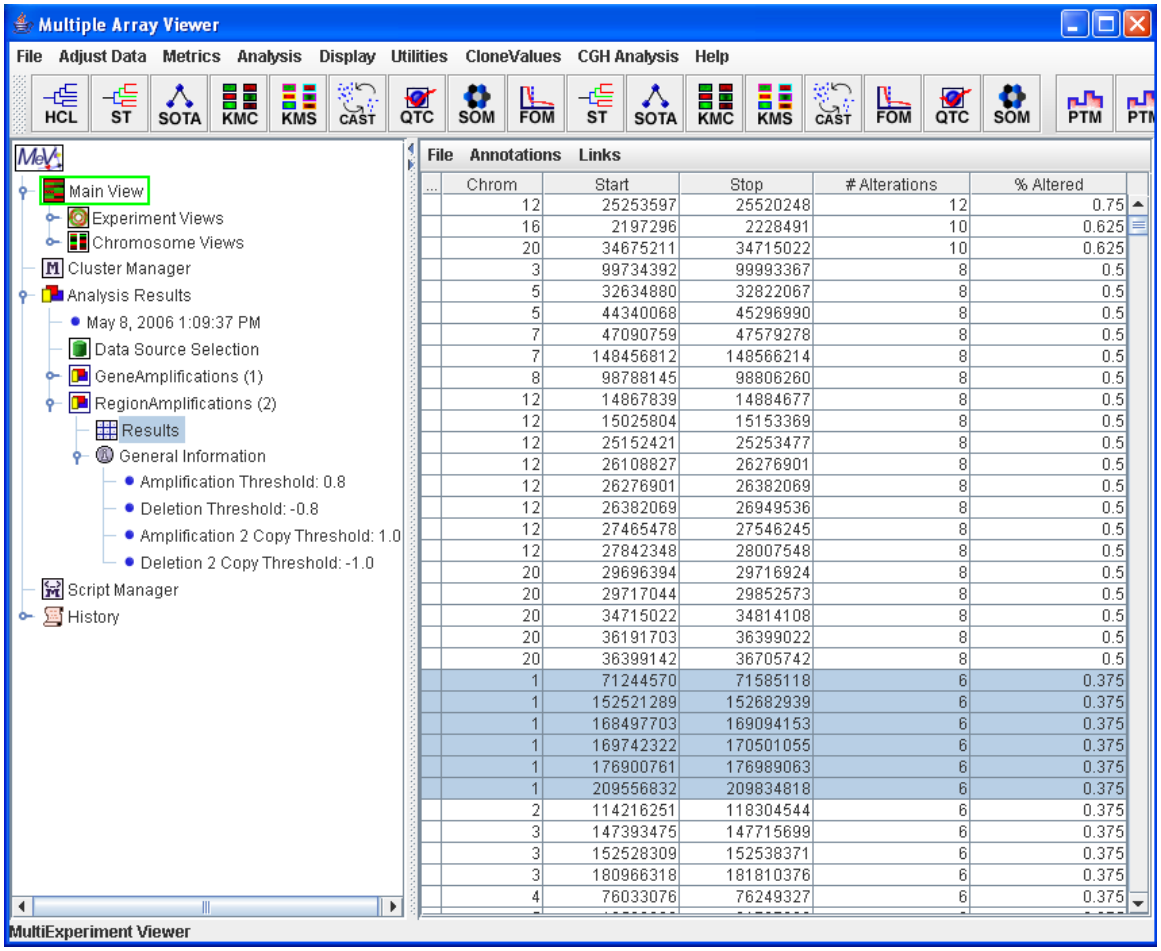
Annotations can be cleared by using the *Clear Annotations* item in the *Display* menu.

Saving Results

The Results of any *CGH Analysis* algorithm can be saved as a tab delimited text file. To do this select the *Save* item from the *File* menu of the algorithm results viewer.

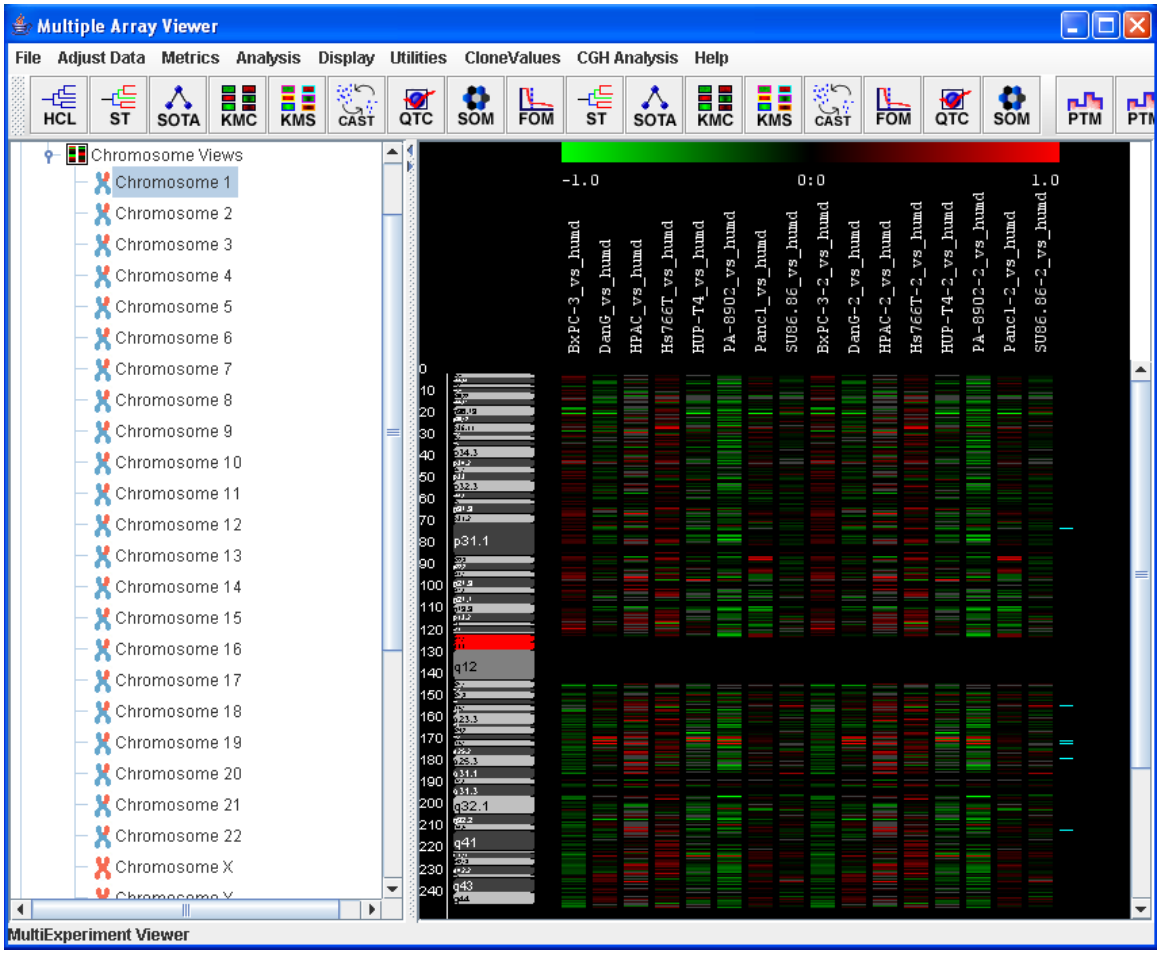
Algorithms on Data Regions

The items *RegionAmplification and RegionDeletions* are used to search for common regions of amplifications and deletions. It is often important to identify minimal regions of alteration that are common between a number of experiments. Select the *RegionDeletions* item. Select the *Results* node in the newly created *Region Deletions* subtree. Notice that there is one region on chromosome 1 that is deleted in all of the samples, and four regions that are deleted in six out of seven samples.



Chromosome 1 deleted regions

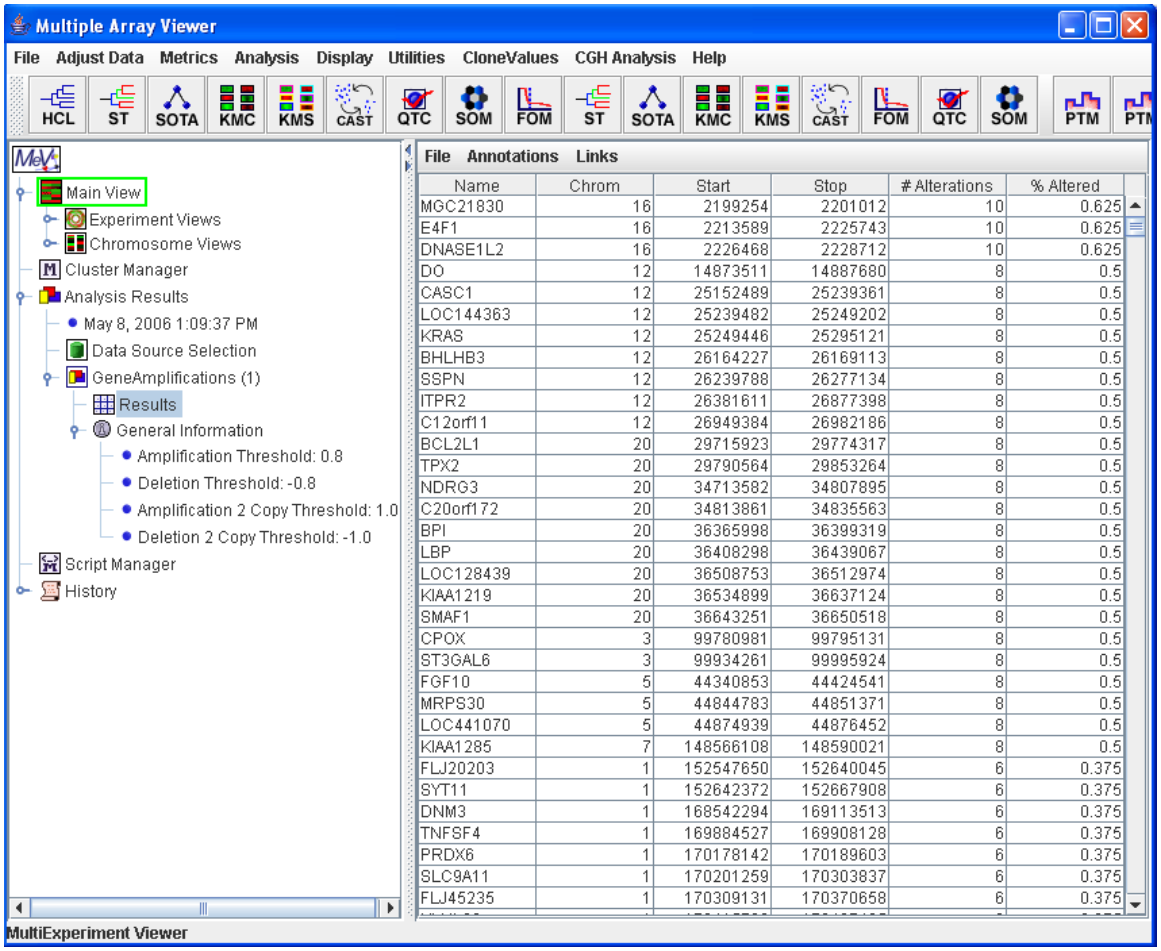
Select and annotate these five regions, and display the CGH Position Graph viewer for chromosome 1 (Figure 3-4). The annotated data regions are represented by light blue rectangles on the right side of the display. This technique can be used to significantly reduce the size of the data regions determined for further investigation. Right-click on any of the blue rectangles and select *Show Genes in Region* to check if there are any consistently deleted genes of interest (Figure 3-5). These are displayed in a tabular format.



Chromosome 1 data regions with six or more amplifications.

Algorithms on Genes

The items *GeneAmplifications and GeneDeletions* are used to search for genes that are commonly altered between experiments. Select the *GeneDeletions* item. Select the *Results* node in the newly created *Gene Deletions* subtree. This view displays the number of deletions for every gene stored in UCSC's Golden Path database.

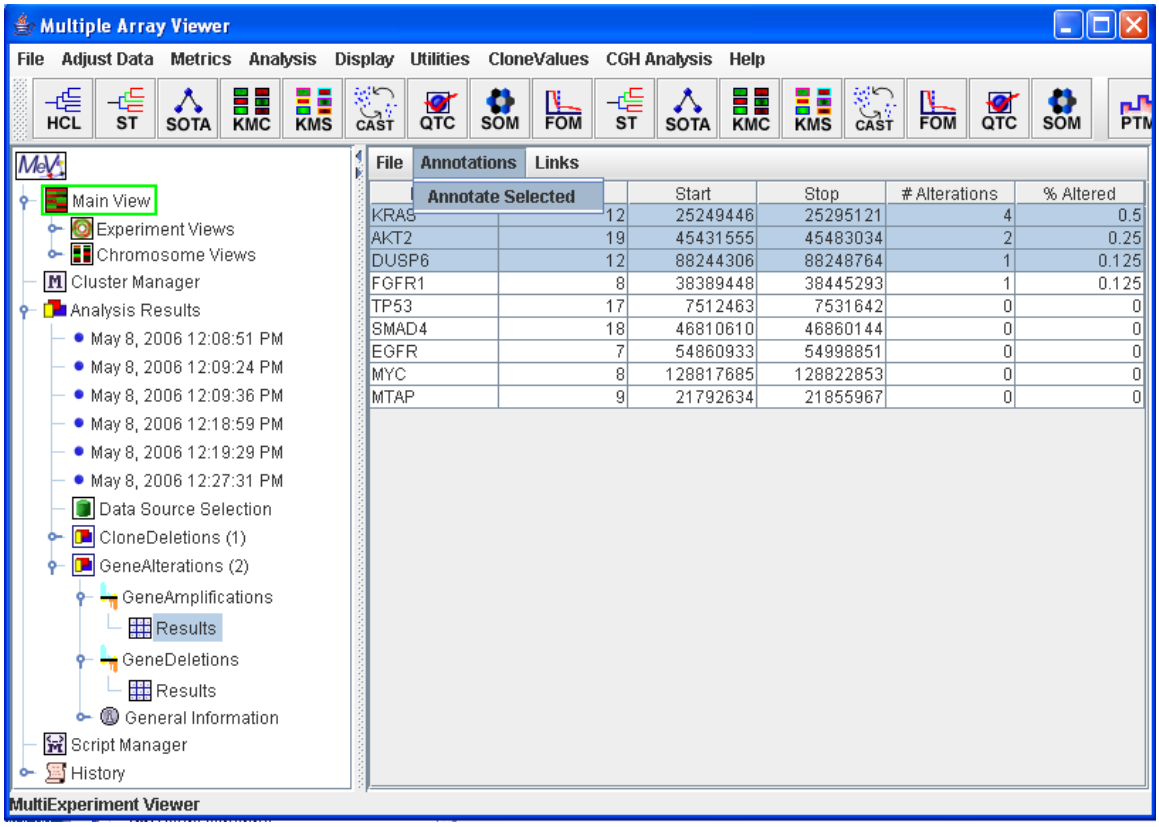


Name	Chrom	Start	Stop	# Alterations	% Altered
MGC21830	16	2199254	2201012	10	0.625
E4F1	16	2213589	2225743	10	0.625
DNASE1L2	16	2226468	2228712	10	0.625
DO	12	14873511	14887680	8	0.5
CASC1	12	25152489	25239361	8	0.5
LOC144363	12	25239482	25249202	8	0.5
KRAS	12	25249446	25295121	8	0.5
BHLHB3	12	26164227	26169113	8	0.5
SSPN	12	26239788	26277134	8	0.5
ITPR2	12	26381611	26877398	8	0.5
C12orf11	12	26949384	26982186	8	0.5
BCL2L1	20	29715923	29774317	8	0.5
TPX2	20	29790564	29853264	8	0.5
NDRG3	20	34713582	34807895	8	0.5
C20orf172	20	34813861	34835563	8	0.5
BPI	20	36365998	36399319	8	0.5
LBP	20	36408298	36439067	8	0.5
LOC128439	20	36508753	36512974	8	0.5
KIAA1219	20	36534899	36637124	8	0.5
SMAF1	20	36643251	36650518	8	0.5
CPOX	3	99780981	99795131	8	0.5
ST3GAL6	3	99934261	99995924	8	0.5
FGF10	5	44340853	44424541	8	0.5
MRPS30	5	44844783	44851371	8	0.5
LOC441070	5	44874939	44876452	8	0.5
KIAA1285	7	148566108	148590021	8	0.5
FLJ20203	1	152547650	152640045	6	0.375
SYT11	1	152642372	152667908	6	0.375
DNM3	1	168542294	169113513	6	0.375
TNFSF4	1	169884527	169908128	6	0.375
PRDX6	1	170178142	170189603	6	0.375
SLC9A11	1	170201259	170303837	6	0.375
FLJ45235	1	170309131	170370658	6	0.375

Gene amplifications on the dataset

Loading a Gene List

Selecting the *LoadGeneList* item in the *CGH Analysis* menu will calculate the number of amplifications and deletions for every gene in a customized gene list. Select this button and load the file named “CGH_sample_genelist.txt”, included with the distribution of the MeV. This list is a text file containing a large number of genes that have been identified as being associated with cancer. Notice that the new Gene Alterations subtree now contains two subtrees, corresponding to the number of times the genes in the list are amplified and deleted.



Gene amplifications on the dataset.

Deleting a Node

Nodes in any tree can be deleted. Right click on the node and select *Delete.*

Searching for a Gene

Selecting the *Find Gene* item from the *CGH Analysis* menu will display a dialog prompting for the name of a gene. Enter the name of a gene of interest and click *Ok*. A dialog will appear showing how many times that gene is deleted and amplified in the dataset. Selecting *Annotate Selected* from the *Annotations* menu of this dialog will display the CGH Position graph corresponding to this gene, with the gene annotated.

Higher Level Analysis

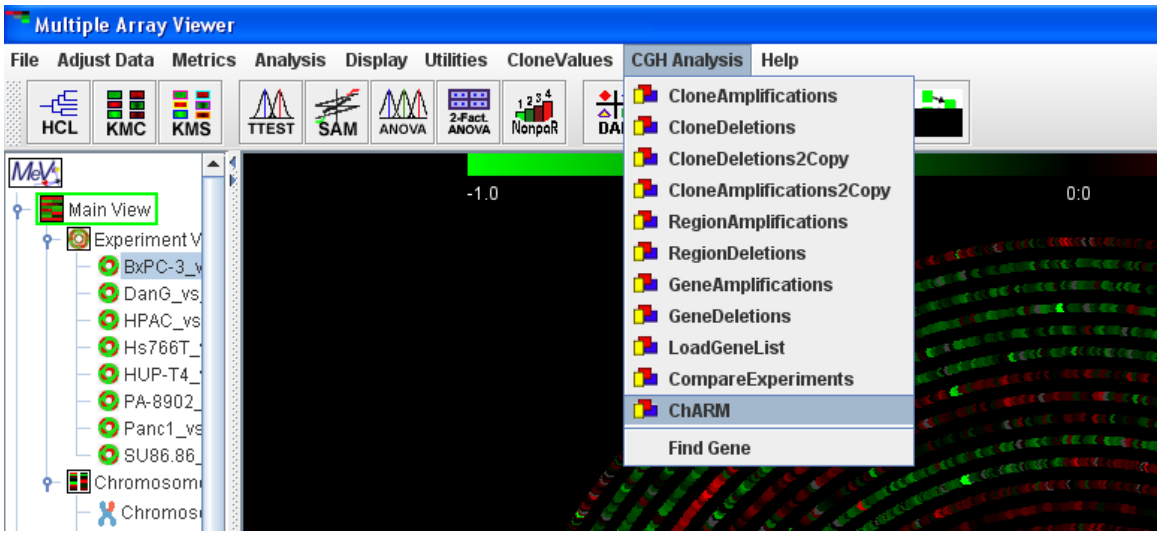
Refer to other sections of the MeV manual for a description of the analysis capabilities of the Multi Experiment Viewer.

ChARM (Chromosome Aberration Region Miner)

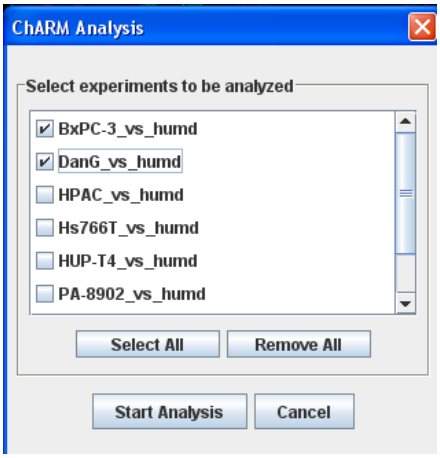
We have integrated a new module called ChARM, a robust expectation maximization based algorithm for identification of segments of chromosomal aberrations from CGH data. (ChARM, Meyers, C.L et al, 2004)

Running ChARM: Please follow the series of screenshots and any instructions associated with it to start the analysis.

1. After loading the CGH data navigate to the main menu "CGH Analysis" option and select "ChARM" from the drop-down.



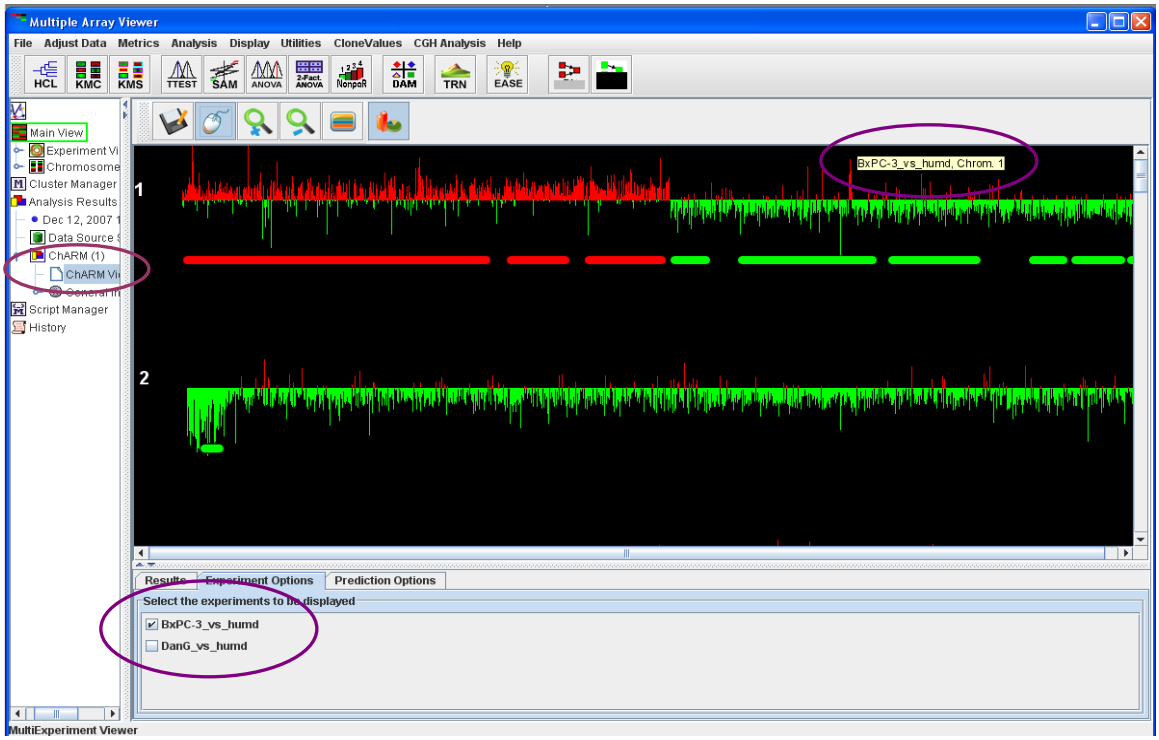
2. A new window opens which displays all CGH experiments that are loaded. Use the check boxes to select the experiments that needs to be analyzed and hit "Start Analysis"



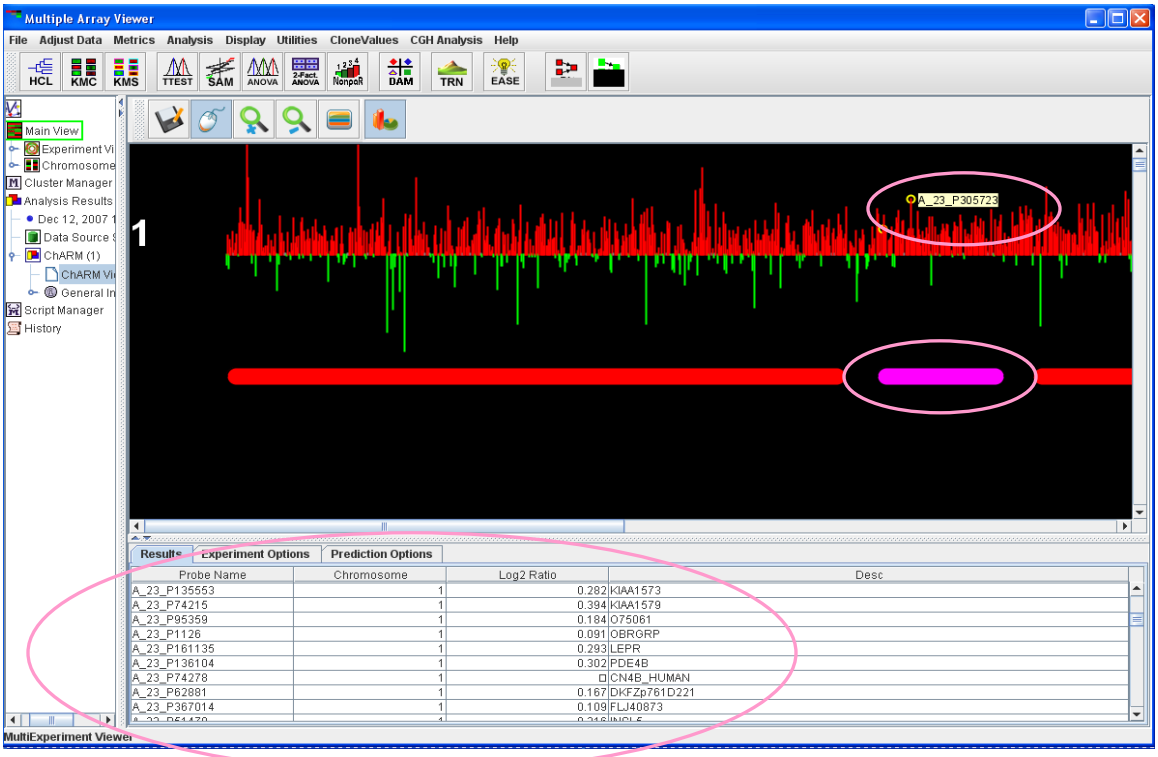
ChARM Results and Result window navigation & options:

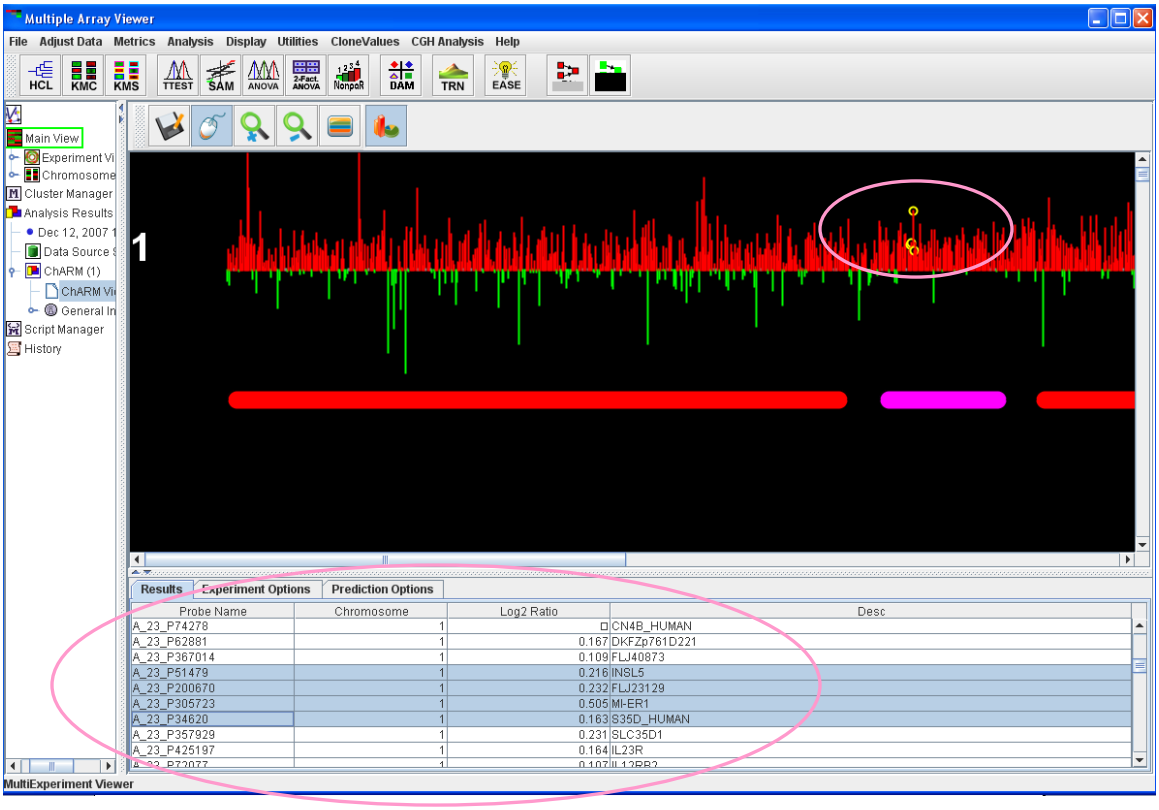
1. Once analysis completes it adds a node called "ChARM" on the left hand panel. Under that "ChARM View" option displays the analysis results as shown below.
2. Holding the mouse cursor on any experiment shows the name of the experiment a shown in the highlighted circle.
3. By default it shows only the 1st experiment. Additional experiments can be loaded by ticking the checkbox against each sample name in the "Experiments Options Panel".

4. The **red** bars below the experiment plots represent amplifications or increase in copy number.
5. The **green** bars below the experiment plots represent deletions or decrease in copy number.

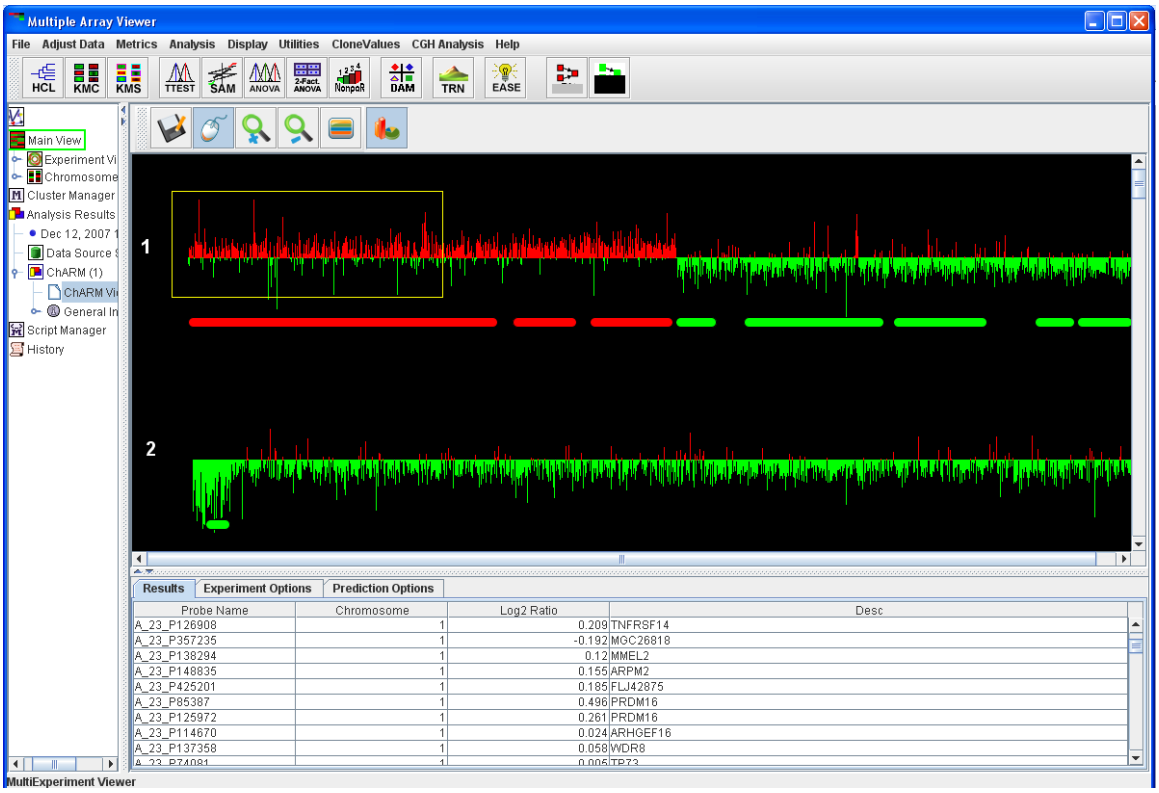


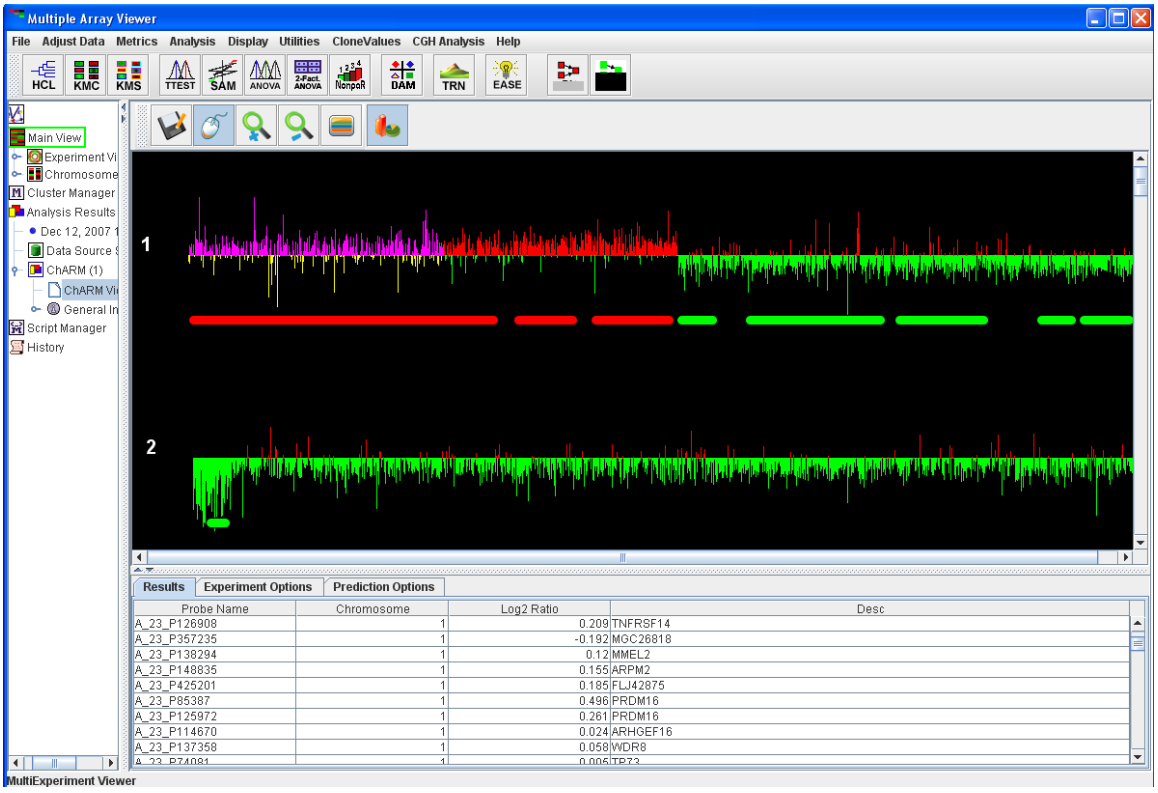
6. By clicking on any of the red/green bars all probes/genes within that segment can be viewed in the "Result" panel below. A segment when selected is displayed in pink.
7. The result panel can be navigated by the up/down arrow key or by the scroll bars.
8. Both single/multiple row selections are possible from the result table. Accordingly the corresponding probe/probes are circled in the plot above. Showed in the following 2 figures.





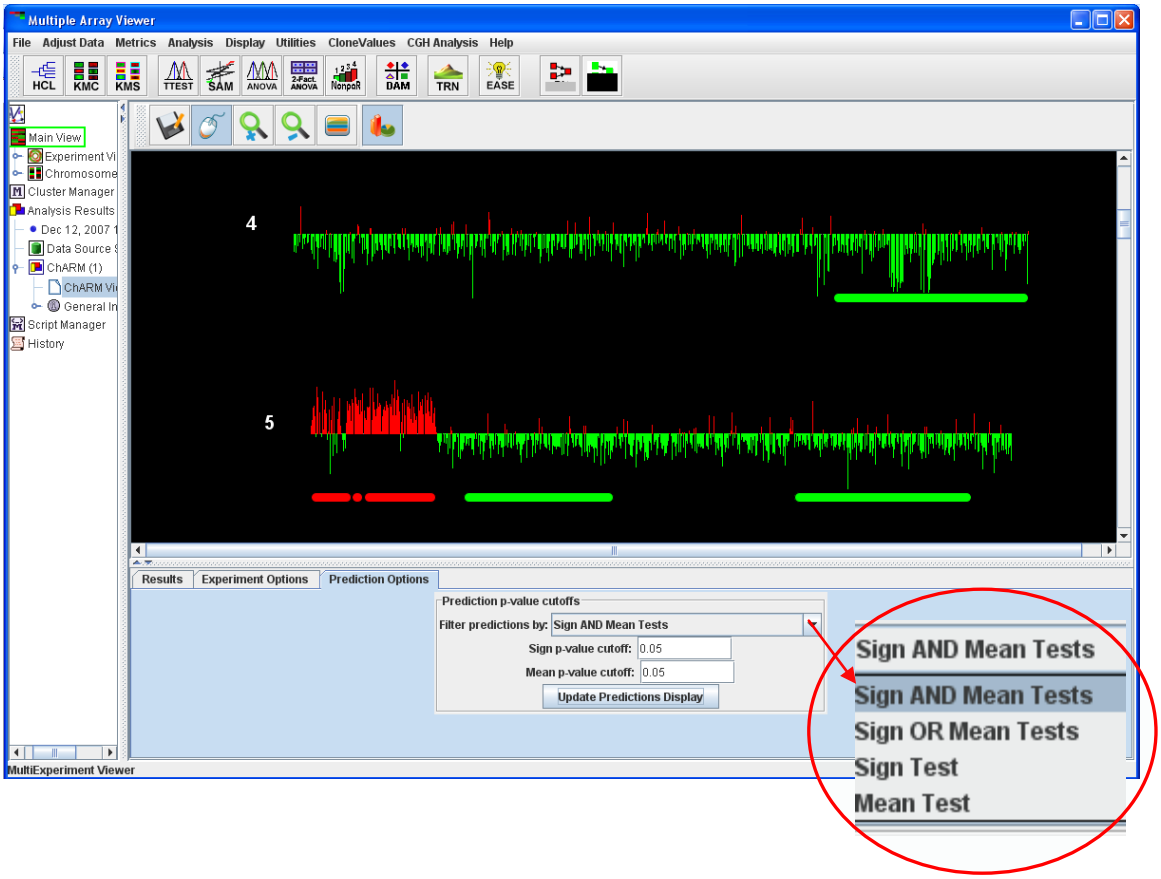
9. Probes or Segments can also be selected by clicking and drawing a box on the canvas around the target. The selections results are then displayed in the Result table as described above.
10. Once selected, the Probes or Segments are displayed in color pink as shown in the 2 figures below.





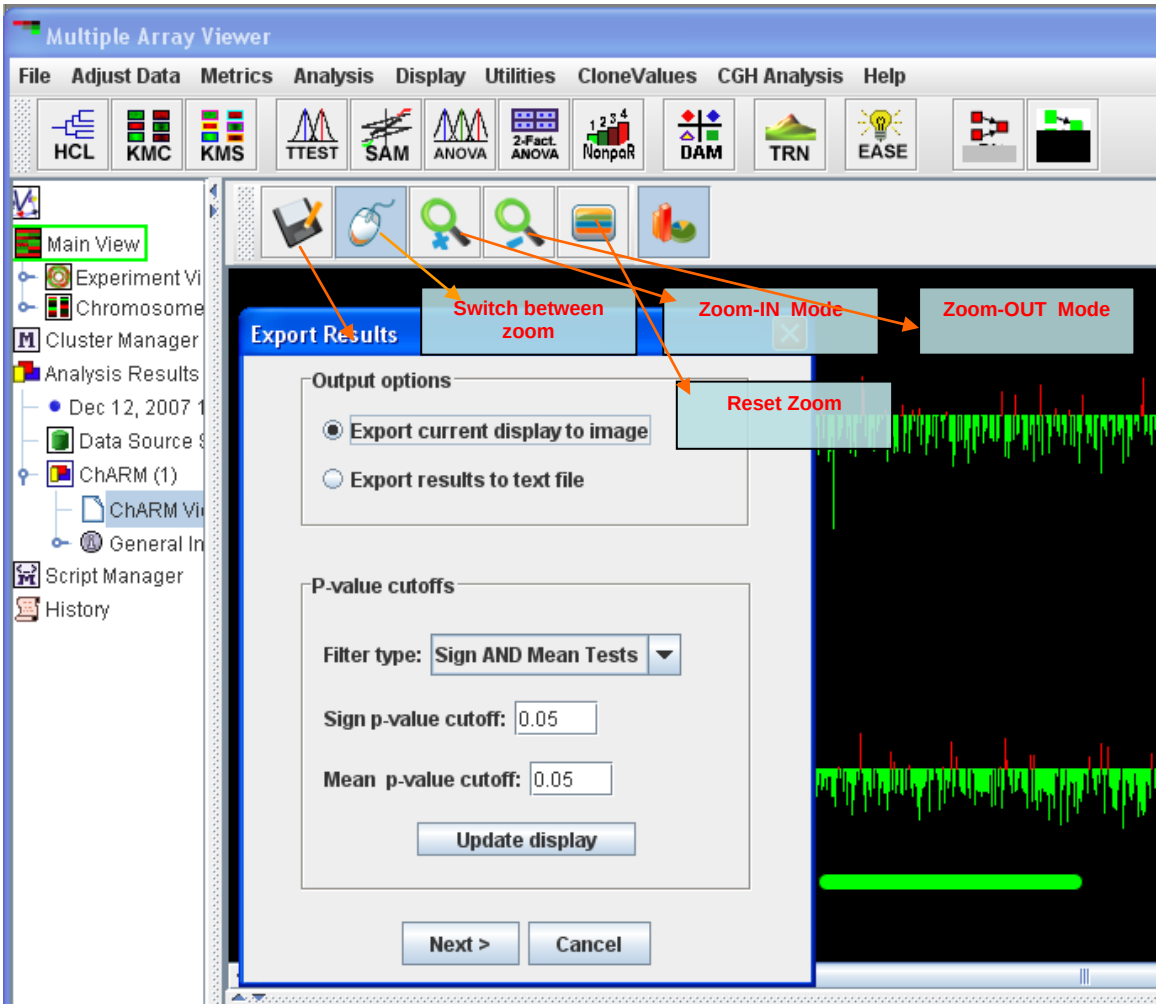
11. Segments are computed based on Sign And Mean Tests for a default p-value. Both, the type of test, their combination and corresponding p-value(s) can be

changed from view for determining significant segments or regions of amplifications & deletions.



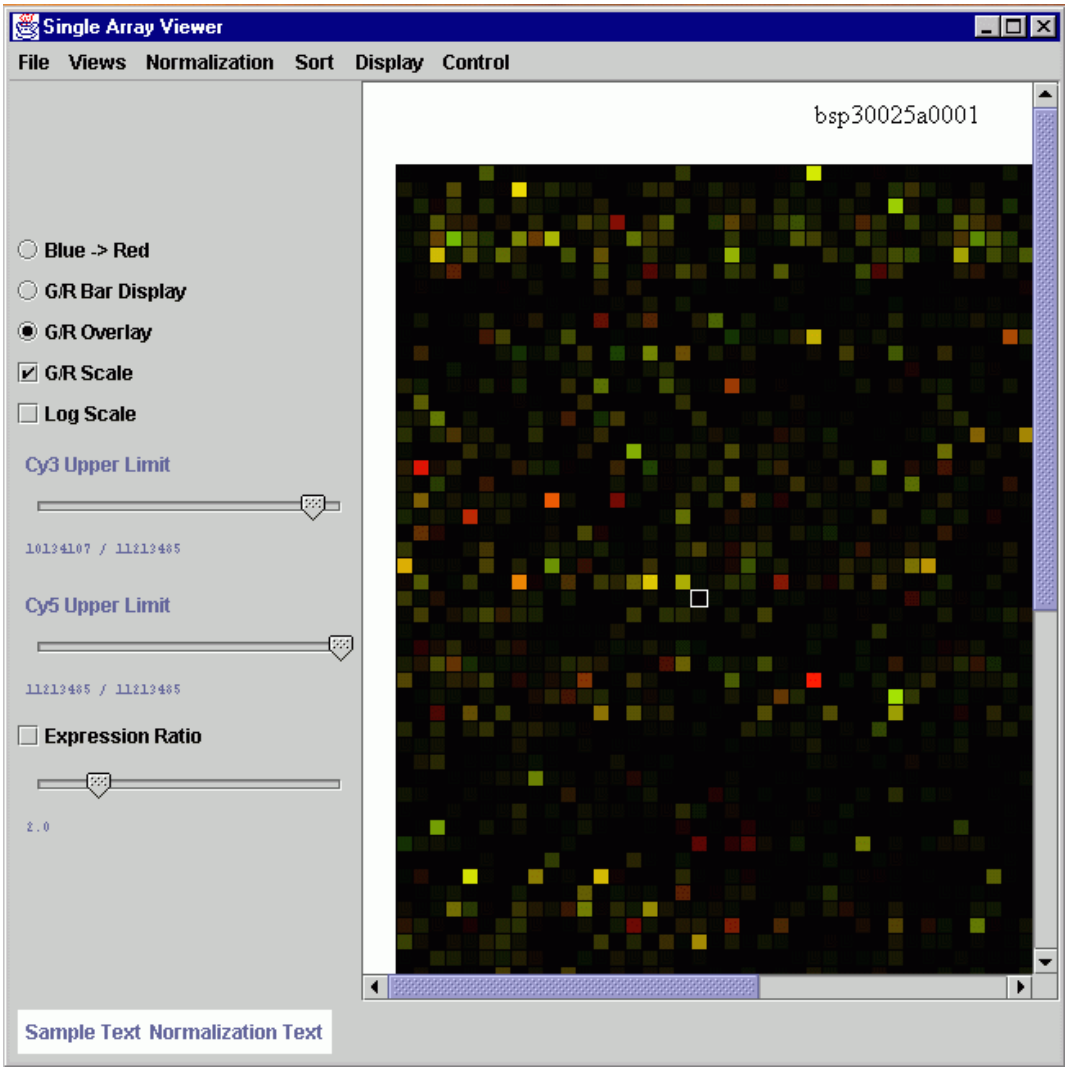
12. The ChARM viewer toolbar buttons can be used to do the following as highlighted in the figure below:

a. Save the results of the analysis
b. Switch between zoom and selection mode.
c. Zoom In/OUT
d. Reset the zoom to fit the window.
e. Toggle view between only the plot and plot with detected segments.


Multiple Array Viewer
File
Adjust Data
Metrics
Analysis
Display
Utilities
CloneValues
CGH Analysis
Help
HCL
KMC
KMS
TTEST
SAM
ANOVA
2-Fact. ANOVA
NonparR
DAM
TRN
EASE
Main View
Experiment Vi
Chromosome
Cluster Manager
Analysis Results
Dec 12, 2007 1
Data Source
ChARM (1)
ChARM Vi
General In
Script Manager
History
Switch between zoom
Zoom-IN Mode
Zoom-OUT Mode
Reset Zoom
Export Results
Output options
Export current display to image
Export results to text file
P-value cutoffs
Filter type:
Sign AND Mean Tests
Sign p-value cutoff:
0.05
Mean p-value cutoff:
0.05
Update display
Next >
Cancel

3. Working with the Single Array Viewer

3.1. 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.

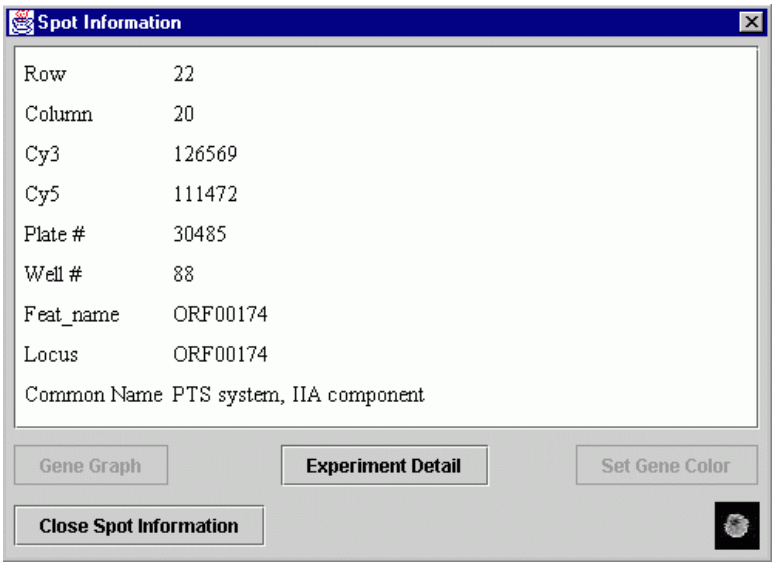


3.1.1. Single Array Viewer

3.2. 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

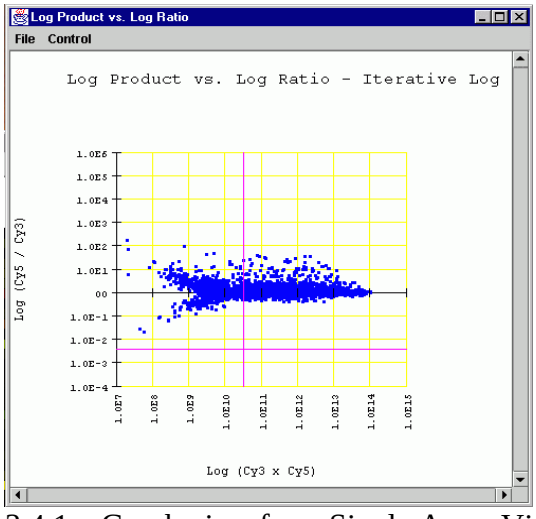
Viewer is the false-color display. It shows the two channels in separate areas, where elements are colored based on a scale where low intensities are dark and blue, while high intensities are bright and red.

3.3. 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.



3.3.1. Spot information

3.4. 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.



3.4.1. Graph view from Single Array Viewer.

3.5. 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.

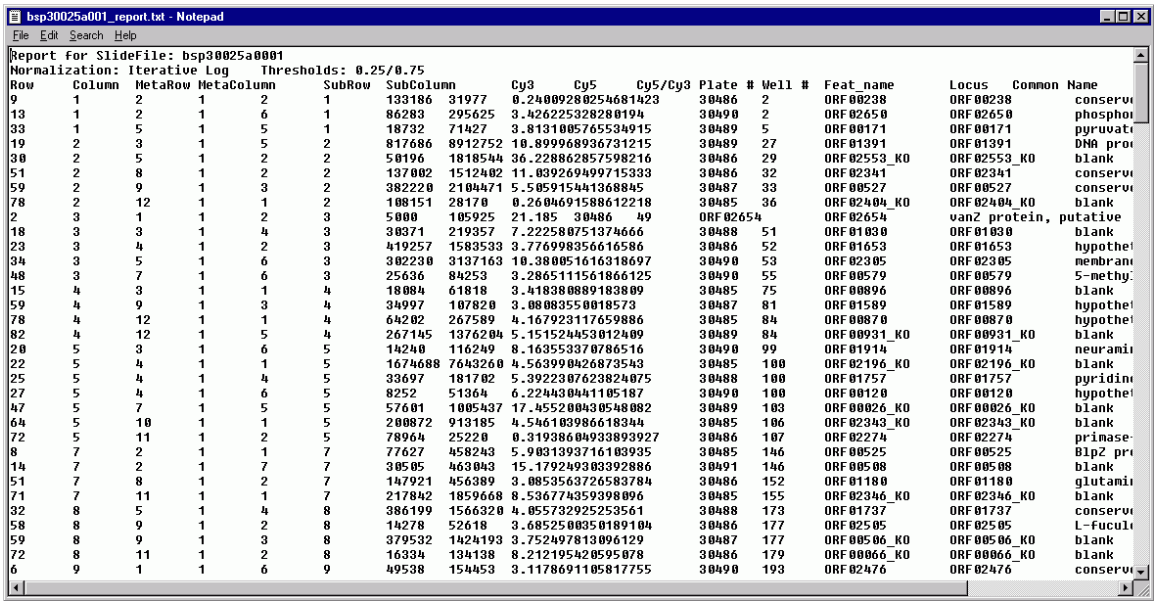
3.6. The View Region item in the *Views* menu displays a dialog for inputting the coordinates of a metablock. A new Single Array Viewer is created, using the targeted metablock as the new dataset. In this way, the user can focus on a particular defined area of the slide.

3.7. 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.

3.8. 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.

3.9. 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.

3.10. 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.



bsp30025a001_report.txt - Notepad

File Edit Search Help

Report for SlideFile: bsp30025a0001
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	0.2400928025468142 3	0486	2	ORF00238	ORF00238	conserv
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	pyruvat
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_KO	ORF02553_KO	blank
51	2	8	1	2	2	137002	1512402	11.039269499715333	30486	32	ORF02341	ORF02341	conserv
59	2	9	1	3	2	382220	2104471	5.505915441368845	30487	33	ORF00527	ORF00527	conserv
78	2	12	1	1	2	108151	28170	0.26046915886122 18	30485	36	ORF02404_KO	ORF02404_KO	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	hypothe
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-methy
15	4	3	1	1	4	18084	61818	3.4183808889183809	30485	75	ORF00896	ORF00896	blank
59	4	9	1	3	4	34997	107820	3.080835500185 73	30487	81	ORF01589	ORF01589	hypothe
78	4	12	1	1	4	64202	267589	4.167923117659886	30485	84	ORF00870	ORF00870	hypothe
82	4	12	1	5	4	267145	1376204	5.151524453012409	30489	84	ORF00931_KO	ORF00931_KO	blank
20	5	3	1	6	5	14240	116249	8.163553370786516	30490	99	ORF01914	ORF01914	neurami
22	5	4	1	1	5	1674688	7643260	4.563990426873543	30485	100	ORF02196_KO	ORF02196_KO	blank
25	5	4	1	4	5	33697	181702	5.392230762382 4075	30488	100	ORF01757	ORF01757	pyridin
27	5	4	1	6	5	8252	51364	6.224430441105187	30490	100	ORF00120	ORF00120	hypothe
47	5	7	1	5	5	57601	1005437	17.455200430548082	30489	103	ORF00026_KO	ORF00026_KO	blank
64	5	10	1	1	5	200872	913185	4.546103986618344	30485	106	ORF02343_KO	ORF02343_KO	blank
72	5	11	1	2	5	78964	25220	0.3193860493389392 7	30486	107	ORF02274	ORF02274	primase
8	7	2	1	1	7	77627	458243	5.903139371610393 5	30485	146	ORF00525	ORF00525	Blp2 pr
14	7	2	1	7	7	30505	463043	15.179249303923886	30491	146	ORF00508	ORF00508	blank
51	7	8	1	2	7	147921	456389	3.085356372658378 4	30486	152	ORF01180	ORF01180	glutami
71	7	11	1	1	7	217842	1859668	8.536774359398096	30485	155	ORF02346_KO	ORF02346_KO	blank
32	8	5	1	4	8	386199	1566320	4.055732925253561	30488	173	ORF01737	ORF01737	conserv
58	8	9	1	2	8	14278	52618	3.685250035018910 4	30486	177	ORF02505	ORF02505	L-fucul
59	8	9	1	3	8	379532	1424193	3.752497813096129	30487	177	ORF00506_KO	ORF00506_KO	blank
72	8	11	1	2	8	16334	134138	8.212195420595078	30486	179	ORF00066_KO	ORF00066_KO	blank
6	9	1	1	6	9	49538	154453	3.1178691105817755	30490	193	ORF02476	ORF02476	conserv

3.10.1. Flat file output from Single Array Viewer

4. Appendix: File Format Descriptions

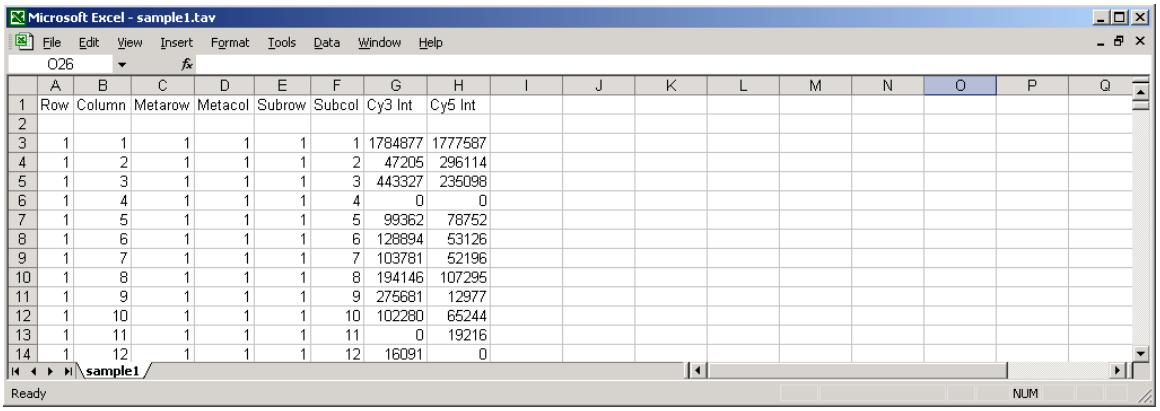
4.1. **TAV Files**

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 4.11). Optional columns can be used to sort the spots in the *Main View*, by choosing the appropriate column from the *Sort* menu.

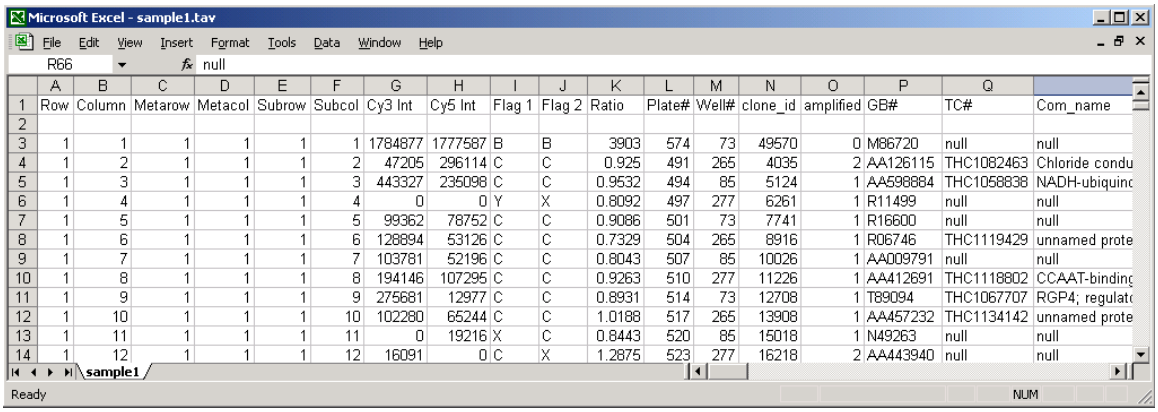
A flag is simply a letter code corresponding to a description of the spot:

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.



	A	B	C	D	E	F	G	H
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

4.1.1. A TAV file containing only the required fields



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
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

4.1.2. A TAV file with several extra fields

4.2. **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.

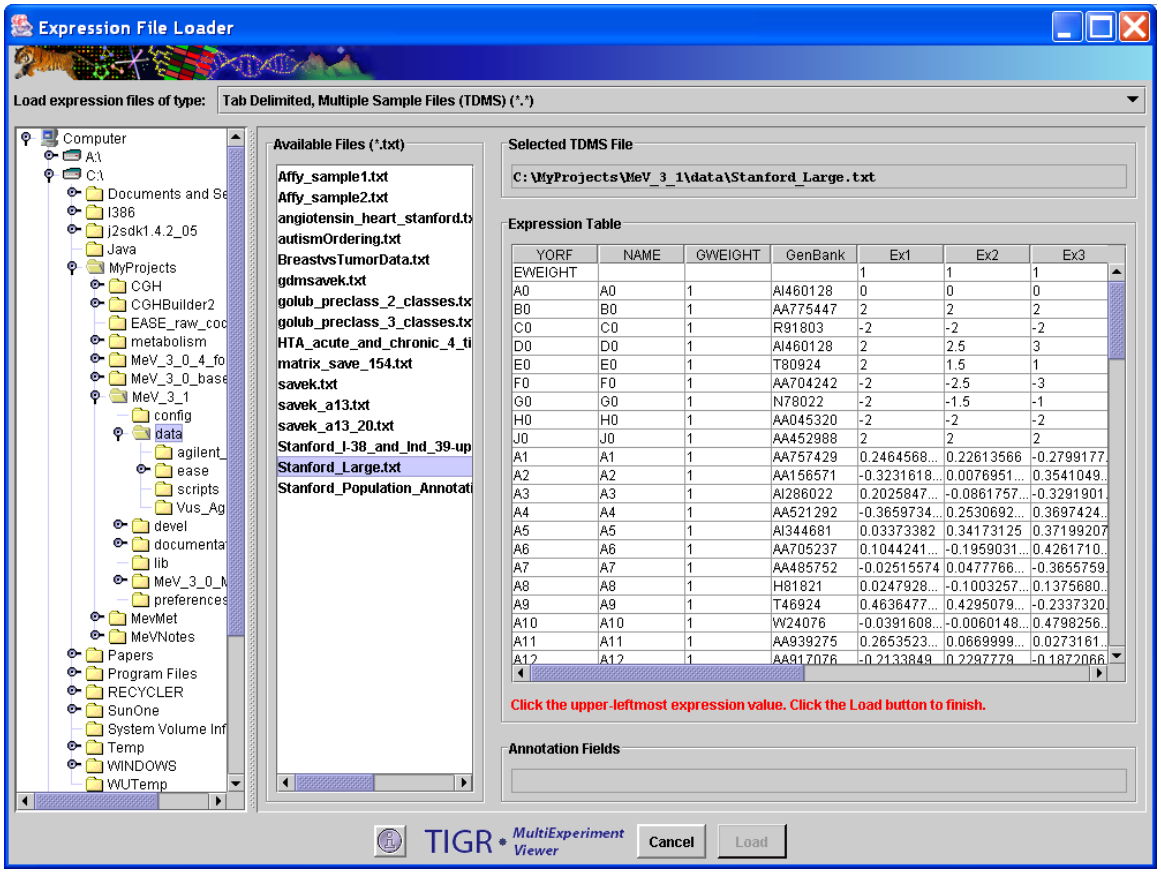
Expression Data (uncolored section with numeric data)

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

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



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

4.3. **GenePix Files**

Microarray data in the Genepix file format can be easily converted to tav 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 tav file.

4.4. **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.

.mev files will typically contain at least one comment at the top of the file with the following information. This information is optional. 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	*id* from the *analysis* table that corresponds to this set of expression values
slide_type	*slide_type* 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

.mev files 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	*TIGR Spotfinder* flag value in channel A
FlagB	*TIGR Spotfinder* 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	*TIGR Spotfinder* flag value in channel A
FlagB	*TIGR Spotfinder* 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.

4.5. **Annotation Files (.ann)**

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	*type* from the *slide_type* 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.

4.6. **Bioconductor (MAS5) Files**

A Bioconductor (MAS5) expression file is a tab-delimited text file that contains Affymatric Gene chip ID and several columns of expression datum. The header line contains all CEL file names you use in the Bioconductor calculation.

```
Sample_A.CEL      Sample_B.CEL      Sample_C.CEL
1053_at     435.013780957768  488.838904739281  435.013780957768
117_at      44.1563783495161  88.9787051028434  44.1563783495161
121_at      1222.87243433892  698.900718145166  1222.87243433892
1255_g_at   58.672587649119   47.0203292843508  58.672587649119
1294_at     336.502704708535  335.169154361150  336.502704708535
1316_at     163.254451578192  92.4927417614044  163.254451578192
1320_at     114.309125038560  105.223093019586  114.309125038560
```

A Bioconductor (MAS5) call file is a tab-delimited text file that contains Affymetrix Gene chip ID and several columns of present/absent detection for corresponding expression file.

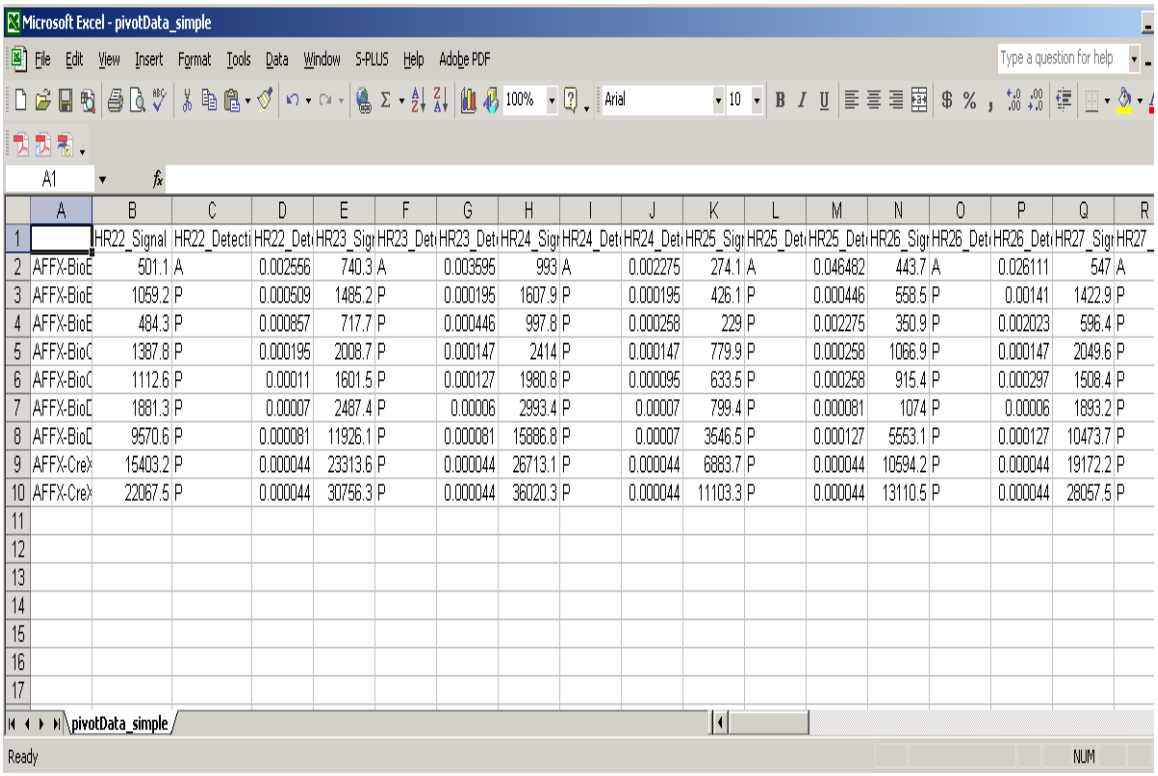
```
Sample_A.CEL      Sample_B.CEL  Sample_C.CEL
1053_at   P       P       P       P
117_at    A       A       A       A
121_at    P       P       P       P
1255_g_at A       A       A       A
1294_at   P       P       P       P
1316_at   P       P       P       P
1320_at   A       A       p       p
```

Users can use following scripts to generate above files by using Bioconductor.

```
library(affy)
data <-ReadAffy()
   mas5data<-mas5(data)
   write.exprs(mas5data,file="affy_mas5.txt")
   mas5call<-mas5calls(data)
   write.exprs(mas5call,file="affy_call.txt")
```

4.7. **Affymetrix GCOS (Pivot Data) File**

An Affymatrix GCOS file is a tab-delimited text file that contains Affymatric Gene chip ID and several experiment datum. In each experiment data it contains one column of intensity, one column of detection call and one column of p-value. The first line is head line.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		HR22_Signal	HR22_Detecti	HR22_Det	HR23_Sig	HR23_Det	HR23_Det	HR24_Sig	HR24_Det	HR24_Det	HR25_Sig	HR25_Det	HR25_Det	HR26_Sig	HR26_Det	HR26_Det	HR27_Sig	HR27_
2	AFFX-BioE	501.1	A	0.002556	740.3	A	0.003595	993	A	0.002275	274.1	A	0.046482	443.7	A	0.026111	547	A
3	AFFX-BioE	1059.2	P	0.000509	1485.2	P	0.000195	1607.9	P	0.000195	426.1	P	0.000446	558.5	P	0.00141	1422.9	P
4	AFFX-BioE	484.3	P	0.000857	717.7	P	0.000446	997.8	P	0.000258	229	P	0.002275	350.9	P	0.002023	596.4	P
5	AFFX-BioC	1387.8	P	0.000195	2008.7	P	0.000147	2414	P	0.000147	779.9	P	0.000258	1066.9	P	0.000147	2049.6	P
6	AFFX-BioC	1112.6	P	0.00011	1601.5	P	0.000127	1980.8	P	0.000095	633.5	P	0.000258	915.4	P	0.000297	1508.4	P
7	AFFX-BioD	1881.3	P	0.00007	2487.4	P	0.00006	2993.4	P	0.00007	799.4	P	0.000081	1074	P	0.00006	1893.2	P
8	AFFX-BioD	9570.6	P	0.000081	11926.1	P	0.000081	15886.8	P	0.00007	3546.5	P	0.000127	5553.1	P	0.000127	10473.7	P
9	AFFX-CreX	15403.2	P	0.000044	23313.6	P	0.000044	26713.1	P	0.000044	6883.7	P	0.000044	10594.2	P	0.000044	19172.2	P
10	AFFX-CreX	22067.5	P	0.000044	30756.3	P	0.000044	36020.3	P	0.000044	11103.3	P	0.000044	13110.5	P	0.000044	28057.5	P
11																		
12																		
13																		
14																		
15																		
16																		
17																		

Pivot Data File

4.8. GEO SOFT Affymetrix File Format

GEO Simple Omnibus Format in Text (SOFT) file format is a kind of flexible tab delimited file format. Users can check the file format in details at http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html#SOFTsubmissionexamples

A template for a single channel Sample file:

```
^SAMPLE=[required]
!Sample_title = [required]
!Sample_source_name = [required]
!Sample_organism = [required]
!Sample_characteristics = [required]
!Sample_biomaterial_provider = [optional]
!Sample_treatment_protocol = [optional]
!Sample_growth_protocol = [optional]
!Sample_molecule = [required][VALUE=total RNA, genomic DNA, polyA
RNA, cytoplasmic RNA, nuclear RNA, protein, other]
!Sample_extract_protocol = [optional]
!Sample_label = [required]
!Sample_label_protocol = [optional]
!Sample_hyb_protocol = [optional]
!Sample_scan_protocol = [optional]
!Sample_description = [required]
!Sample_data_processing = [required]
!Sample_platform_id = [required; provide accession of existing
GEO Platform for array used, e.g. "GPL96"]
#ID_REF = [required; this column should correspond to the ID
column of the reference platform]
#VALUE = [required; typically supplied as normalized signal
intensities]
#HEADER_3 = [optional]
```

```
#HEADER_4 = [optional]
#HEADER_N = [optional; any number of user-defined columns can be
included, and it is recommended that data tables be as
comprehensive as possible excepting annotations that are provided
on the platform entry]
!Sample_table_begin
ID_REF   VALUE   HEADER_3       HEADER_4       HEADER_N
...insert data table here; columns may appear in any order after
the ID_REF column...
!Sample_table_end
```

A template for platform file:

```
^PLATFORM=[required]
!Platform_title = [required]
!Platform_technology = [required][VALUE=spotted DNA/cDNA, in situ
oligonucleotide, spotted oligonucleotide, antibody, tissue, MS,
MPSS]
!Platform_distribution = [required][VALUE=non-commercial,
commercial, custom-commercial]
!Platform_organism = [required]
!Platform_manufacturer = [required]
!Platform_manufacture_protocol = [required]
!Platform_catalog_number = [optional]
!Platform_support = [optional]
!Platform_coating = [optional]
!Platform_description = [optional]
!Platform_web_link = [optional]
!Platform_contributor = [optional; 1 per author; use
'firstname,lastname' or 'firstname,middleinitial,lastname']
!Platform_pubmed_id = [optional]
#ID = [required; a unique id should be provided for each 'spot'
on the array]
#HEADER_2 = [required; all elements of the array should be
identified using one or more columns in addition to the ID
column]
#HEADER_3 = [optional]
#HEADER_N = [optional; provide as many headers as needed to fully
describe the elements of the array]
!Platform_table_begin
ID       HEADER_2       HEADER_3       HEADER_N
...insert data table here; columns may appear in any order after
the ID column...
!Platform_table_end
```

4.9. GEO SOFT two channel file format

GEO Simple Omnibus Format in Text (SOFT) file format is a kind of flexible tab delimited file format for two channel data. Users can check the file format in details at:
http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html#SOFTsubmissionexamples

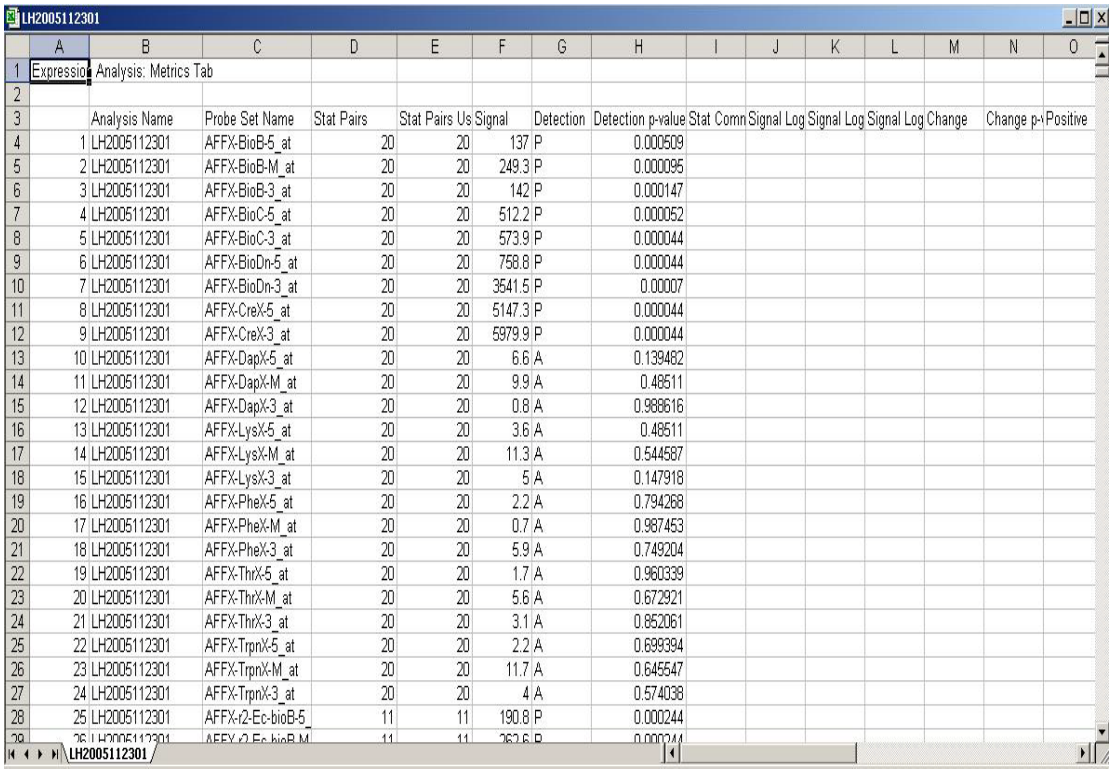
A template for two channel file:

```
^SAMPLE=[required]
!Sample_title = [required]
!Sample_source_name_ch1 = [required]
!Sample_organism_ch1 = [required]
!Sample_characteristics_ch1 = [required]
```

```
!Sample_biomaterial_provider_ch1 = [optional]
!Sample_treatment_protocol_ch1 = [optional]
!Sample_growth_protocol_ch1 = [optional]
!Sample_molecule_ch1 = [required][VALUE=total RNA, polyA RNA,
cytoplasmic RNA, nuclear RNA, genomic DNA, protein, other]
!Sample_extract_protocol_ch1 = [optional]
!Sample_label_ch1 = [required]
!Sample_label_protocol_ch1 = [optional]
!Sample_source_name_ch2 = [required]
!Sample_organism_ch2 = [required]
!Sample_characteristics_ch2 = [required]
!Sample_biomaterial_provider_ch2 = [optional]
!Sample_treatment_protocol_ch2 = [optional]
!Sample_growth_protocol_ch2 = [optional]
!Sample_molecule_ch2 = [required][VALUE=total RNA, polyA RNA,
cytoplasmic RNA, nuclear RNA, genomic DNA, protein, other]
!Sample_extract_protocol_ch2 = [optional]
!Sample_label_ch2 = [required]
!Sample_label_protocol_ch2 = [optional]
!Sample_hyb_protocol = [optional]
!Sample_scan_protocol = [optional]
!Sample_description = [required]
!Sample_data_processing = [required]
!Sample_platform_id = [required; provide accession of existing
GEO Platform for array used, e.g. "GPL1001"]
#ID_REF = [required; this column should correspond to the ID
column of the reference platform]
#VALUE = [required; normalized log2 ratios are typically provided
in this column]
#HEADER_3 = [optional]
#HEADER_4 = [optional]
#HEADER_N = [optional; any number of user-defined columns can be
included, and it is recommended that data tables be as
comprehensive as possible excepting annotations that are provided
on the platform entry]
!Sample_table_begin
ID_REF    VALUE    HEADER_3        HEADER_4        HEADER_N
...insert data table here; columns may appear in any order after
the ID_REF column...
!Sample_table_end
```

4.10. dChip or DFCI core file format

Each dChip data output file contains one experiment data. It’s a tab delimited file which records a lot of information for the experiment. Right now only chip ID, intensity, detection and detection p-value are read into Mev for the analysis.



LH2005112301

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Expressio	Analysis: Metrics Tab													
2															
3		Analysis Name	Probe Set Name	Stat Pairs	Stat Pairs Us	Signal	Detection	Detection p-value	Stat Comn	Signal Log	Signal Log	Signal Log	Change	Change p-\	Positive
4	1	LH2005112301	AFFX-BioB-5_at	20	20	137	P	0.000509							
5	2	LH2005112301	AFFX-BioB-M_at	20	20	249.3	P	0.000095							
6	3	LH2005112301	AFFX-BioB-3_at	20	20	142	P	0.000147							
7	4	LH2005112301	AFFX-BioC-5_at	20	20	512.2	P	0.000052							
8	5	LH2005112301	AFFX-BioC-3_at	20	20	573.9	P	0.000044							
9	6	LH2005112301	AFFX-BioDn-5_at	20	20	758.8	P	0.000044							
10	7	LH2005112301	AFFX-BioDn-3_at	20	20	3541.5	P	0.00007							
11	8	LH2005112301	AFFX-CreX-5_at	20	20	5147.3	P	0.000044							
12	9	LH2005112301	AFFX-CreX-3_at	20	20	5979.9	P	0.000044							
13	10	LH2005112301	AFFX-DapX-5_at	20	20	6.6	A	0.139482							
14	11	LH2005112301	AFFX-DapX-M_at	20	20	9.9	A	0.48511							
15	12	LH2005112301	AFFX-DapX-3_at	20	20	0.8	A	0.988616							
16	13	LH2005112301	AFFX-LysX-5_at	20	20	3.6	A	0.48511							
17	14	LH2005112301	AFFX-LysX-M_at	20	20	11.3	A	0.544587							
18	15	LH2005112301	AFFX-LysX-3_at	20	20	5	A	0.147918							
19	16	LH2005112301	AFFX-PheX-5_at	20	20	2.2	A	0.794268							
20	17	LH2005112301	AFFX-PheX-M_at	20	20	0.7	A	0.987453							
21	18	LH2005112301	AFFX-PheX-3_at	20	20	5.9	A	0.749204							
22	19	LH2005112301	AFFX-ThrX-5_at	20	20	1.7	A	0.960339							
23	20	LH2005112301	AFFX-ThrX-M_at	20	20	5.6	A	0.672921							
24	21	LH2005112301	AFFX-ThrX-3_at	20	20	3.1	A	0.852061							
25	22	LH2005112301	AFFX-TrpnX-5_at	20	20	2.2	A	0.699394							
26	23	LH2005112301	AFFX-TrpnX-M_at	20	20	11.7	A	0.645547							
27	24	LH2005112301	AFFX-TrpnX-3_at	20	20	4	A	0.574038							
28	25	LH2005112301	AFFX-r2-Ec-bioB-5_	11	11	190.8	P	0.000244							
29	26	LH2005112301	AFFX-r2-Ec-bioB-M_	11	11	262.6	P	0.000244							

LH2005112301

dChip File Format

4.11. **Assignment File Saving System**

Initialization dialogs for all modules save their information in a format that is logical, readable by other modules, and readable by humans.

Each module saves the header in the following way, with the abbreviated letters representing the module replacing *Module*.

```
# Assignment File
# User: user Save Date: Mon Jan 01 01:01:01 EST 2009
#
Module:         *ModuleName*
```

MeV then stores a key which is essential for cross-compatibility between modules, as well as preserving label names for those modules that use labels.
Many modules contain a “not included” option for samples. This is not added to the key, and any assignment not associated with the key will be determined to be “not included”. This is important for cross-compatibility.

```
Group 1 Label:      Alpha
Group 2 Label:      Bravo
Group 3 Label:      Charlie
```

The following lines are then written to the file for the purposes of human readability only.

```
#
```

Index Name Group Assignment

Each module will then write the group assignments for each sample, separated by a new line. For each assignment, at least 3 markers will be assigned, separated by tabs.
1.) The sample's index, which will be assigned based on its location within the currently loaded data.
2.) The sample's currently selected annotation, usually the default slide name.
3.) The sample's group assignment that will match a String in the key above or a module defined "excluded" key.

Note: In some cases (BETR, NonPar, GSEA and TFA), additional information is stored for additional factors, conditions, etc. This information is added to the right of each key, and is ignored by other modules.

1	Sample A	Alpha
2	Sample B	Bravo
3	Sample C	Charlie
4	Sample D	Excluded
5	Sample E	Alpha
6	Sample F	Excluded
7	Sample G	Alpha
8	Sample H	Bravo
9	Sample I	Charlie

Special Cases:

BETR:

The BETR module adds an additional parameter, "Conditions \t *int*", which is necessary to tell the BETR reader whether or not to read condition assignments in addition to time assignments.

t-Test:

The reader does distinguish between files saved with "one-class" and "between subjects". Though compatible, the user will be notified that the saved file was not saved for the current method. For "one-class" compatibility, only samples with a key in the first class are added to the "included" checkbox list. All others are excluded. "Paired" does not use this saving system.

Two-class paired:
The loading file format is a tab-delimited file. Samples start from 0. Users define one pair in one line. Following is a sample file. The first sample and second sample are in one pair. The third sample and fourth sample are in one pair and so on.

0	1
2	3
4	5

6 7

SAM:

The reader does distinguish between files saved with “one-class”, “multi-class” and “two-class unpaired”. Though compatible, the user will be notified that the saved file was not saved for the current method. For “one-class” compatibility, only samples with a key in the first class are added to the “included” checkbox list. All others are excluded. “Paired” and “Censored Survival” does not use this saving system.

Two-class paired:
The loading file format is a tab-delimited file. Samples start from 0. Users define one pair in one line. Following is a sample file. The first sample and second sample are in one pair. The third sample and fourth sample are in one pair and so on.

0 1
2 3
4 5
6 7

Censored survival:
The loading file format is a tab-delimited file. Users define one sample in one line. Following is a sample file. The first column is 1(selected) or 0(unselected). The second column is time and third column is 1(censored) or 0(dead).

1 0.0 0
1 0.0 1
1 0.0 1
1 0.0 1

TFA:

Like BETR, TFA assignment files contain an additional column for keys belonging to the second factor. When loading other assignment files that do not contain 2 factors, only the first factor is assigned and the second remains unchanged.

NonPar:

In addition to loading group assignments, the saved group labels replace the currently loaded group labels. If the group label number does not match the current group label number, the loading process is stopped.

GSEA:

Like BETR and TFA, GSEA contains additional columns for additional factors. If the assignment file being loaded does not contain the additional columns

needed to fill the additional factor assignments, GSEA will load only the first factor(s).

SVM:

A "load file" feature has been added to the dialog box and is accessible from the "File" drop-down menu.
All "excluded" samples from outside saved files are added to the "Neutral" group.

USC:

If a file is loaded with a different number of groups, MeV is able (at the user's request) to replace the group label names with the saved names and add or subtract groups to match the saved file's information.

5. 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 4.1), Stanford files (section 4.1.2) 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 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

6. 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.

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

where

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

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.

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

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.

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

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, 3.5, 4.0, 2.0, 3.5) . These ranking vectors are then used to compute the distance via:

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

where

$d = \sum_{i=1}^{m} (x_i - y_i)^2$; 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.

7. Appendix: MeV Script DTD

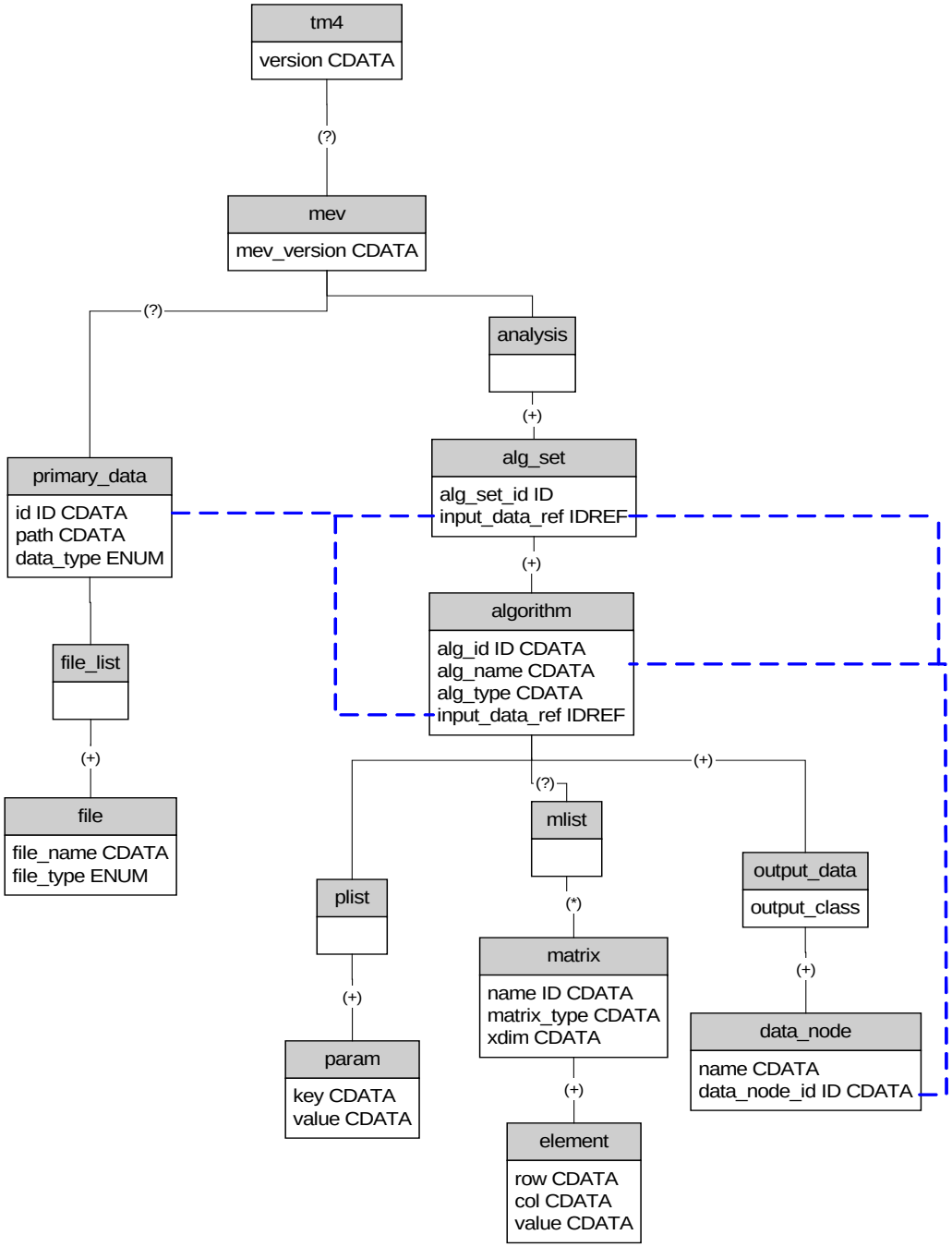
7.1. DTD

```
<!--<?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 paricular 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>
```

7.2. **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)



8. Appendix: MeV R Integration

MeV has integrated a R (CRAN) hook where by R functions and libraries can be called within the Java instance using shared libraries. Mev developed a library around the JRI (rForge) API which uses JNI, to make this happen. The user in *most cases would not have to set up or configure anything to run R dependent modules and it should be completely transparent. Please also note that MeV does not produce a command line interface to run R commands. This integration was done to leverage the R environment where many algorithms are readily available and does not need to be re-implemented in Java. MeV would use R internally as appropriate and the user should not expect to see any change in behavior of MeV.

Notes

1. On Windows and Mac OS X (10.6) where the default JVM is 64 Bit, the user will be thrown a warning to change to 32 bit JVM. This is required because the API is not ready for 64 Bit environment yet and we are working on a solution. However once the default JVM is set to 32 Bit, the R dependent modules would run. To help the users setup the 32 bit JVM we will be providing help and assistance to do the following:
    a. On Windows: To install 32 Bit Java if not already installed and to modify the launch script TMEV.sh to point to it.
    b. On Mac OS X we would provide instructions on how to set up 32bit JVM as default. The default in OS X 10.6 (Snow leopard is 64 bit).
2. On Mac we expect the user to have R 2.9.x universal binary installed in the Application Framework. We do not expect such for Windows and Linux systems.
3. Please use MeV SourceForge page for submitting queries, questions and issues. We have set up a page for this particular JVM issue named R MeV Integration, JVM issues.

9. Appendix: R Serve Package Installation

*****Please note that the Rama/Bridge module is no longer supported by MeV*****

Make sure you are using the latest versions of Rama/Bridge (1.3.0 & 1.3.1 at the time of this writing)

1. Background and Introduction
2. Installing/Updating R
3. Installing Rama/Bridge
4. Updating Rama/Bridge
5. Installing Rserve
6. Running Rserve

1. Background and **Introduction**

The Bioconductor project (www.bioconductor.org) is an open source software project that provides a wide range of statistical tools primarily based on the R programming environment and language (www.r-project.org). Taking advantage of R's powerful statistical and graphical capabilities, developers have created and contributed numerous Bioconductor packages to solve a variety of data analysis needs. However, the use of these packages, requires a basic understanding of the R programming language. Our goal is to provide point-and-click access of these statistically powerful bioconductor packages to the biomedical community through the MeV environment. We have successfully integrated two Bioconductor packages, Rama and Bridge, in the MeV environment. RAMA (Robust Analysis of MicroArrays) [1] uses a Bayesian hierarchical model for the robust estimation of cDNA microarray intensities. BRIDGE (Bayesian Robust Inference for Differential Gene Expression) [2] tests for differentially expressed genes for both one and two-color microarray data. BRIDGE uses a similar Bayesian model as RAMA, but they are two independent bioconductor packages.

In order to make use of the MeV-R integrated environment, every computer that is used to run MeV-R needs the R Environment installed. Furthermore, Rama and Bridge are separate R packages that must be downloaded and installed from Bioconductor. Since MeV is an application written in Java, every computer running MeV-R needs to have the Java Run Time Environment installed. Alone, Java and R are independent environments and mutually ignorant of the other. In order to integrate MeV and R, we use Rserve, written by the University of Augsburg Institute for Mathematics, as the link between Java and R. Rserve is a TCP/IP server. Rserve must also be installed on every computer running MeV-R.

Having installed all the necessary components, the user will be required to start the Rserve server <#runningRserve> each and every time he/she wants to do analysis using MeV-R.

2. Installing/Updating R

Installing under OS X (Build from Source)
If you need the latest version and a precompiled binary hasn't yet been created:

1. Download the latest version (toward bottom of screen - R-2.3.0.tar.gz at time of this writing) from http://cran.r-project.org/src/base/R-2.

2. Be sure that the downloaded file is not in a folder containing a space in the name (like My Downloads) - It will not work. Open a terminal window and cd to the folder containing the file.

3. Unpack it by typing 'tar zxvf R-2.3.0.tar.gz' (or whatever the filename is)

4. CD into the unpacked directory (R-2.3.0 in this case). This directory should contain make and configure files.

5. If you're using OS 10.4 (Tiger), issue the command 'sudo gcc_select 3.3' to force the use of gcc 3.3.

6. Issue the command './configure'

7. Issue the command 'make'

8. Issue the command 'sudo chmod -R g+w /Library/Frameworks/R.framework' to change file permissions.

9. Issue the command 'sudo make install'

Installing under OS X (Precompiled Binary Version)
This is the easiest way, but there is often a lag between when the latest version is available and when a precompiled binary is available.

1. Download R. Get the MacOS X precompiled Binary Distribution from http://cran.fhcrc.org

2. Install R by extracting the downloadeded .dmg file. A new volume will be mounted containing the installer. Run the installer by double clicking on R2.1.1...mpkg. You can place the Application anywhere you like, but the R framework is probably installed into /Libraries/Frameworks/R.framework/versions/x.y.z/resources.

Installing under Windows

1. Download R. Get the Windows precompiled Binary Distribution from http://cran.fhcrc.org

2. Install by double clicking the downloaded installer. You can install anywhere, but remember where.

3. **Installing RAMA/BRIDGE**

Installing under OS X (using the supplied OS X GUI interface)

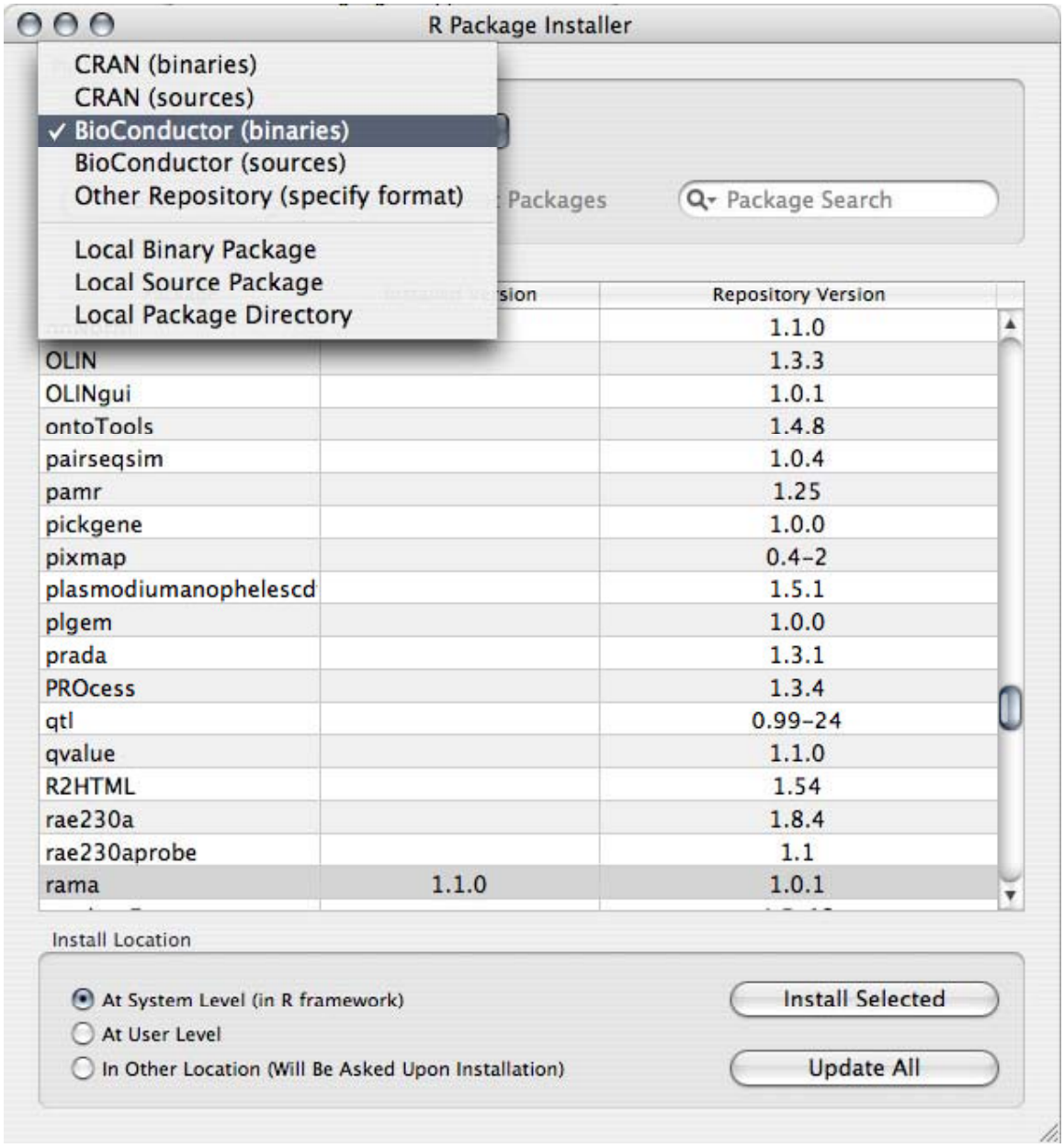
1. Click on Packages & Data in the menu bar. Select Package Installer

2. In the dialog box that appears, choose a Repository. Click on the pull down menu (probably displaying CRAN binaries) and change it to Bioconductor binaries

3. Click Get List

4. Click on rama or bridge to highlight it

5. Click Install Selected



Installing under OS X (using the command line -- RECOMMENDED)

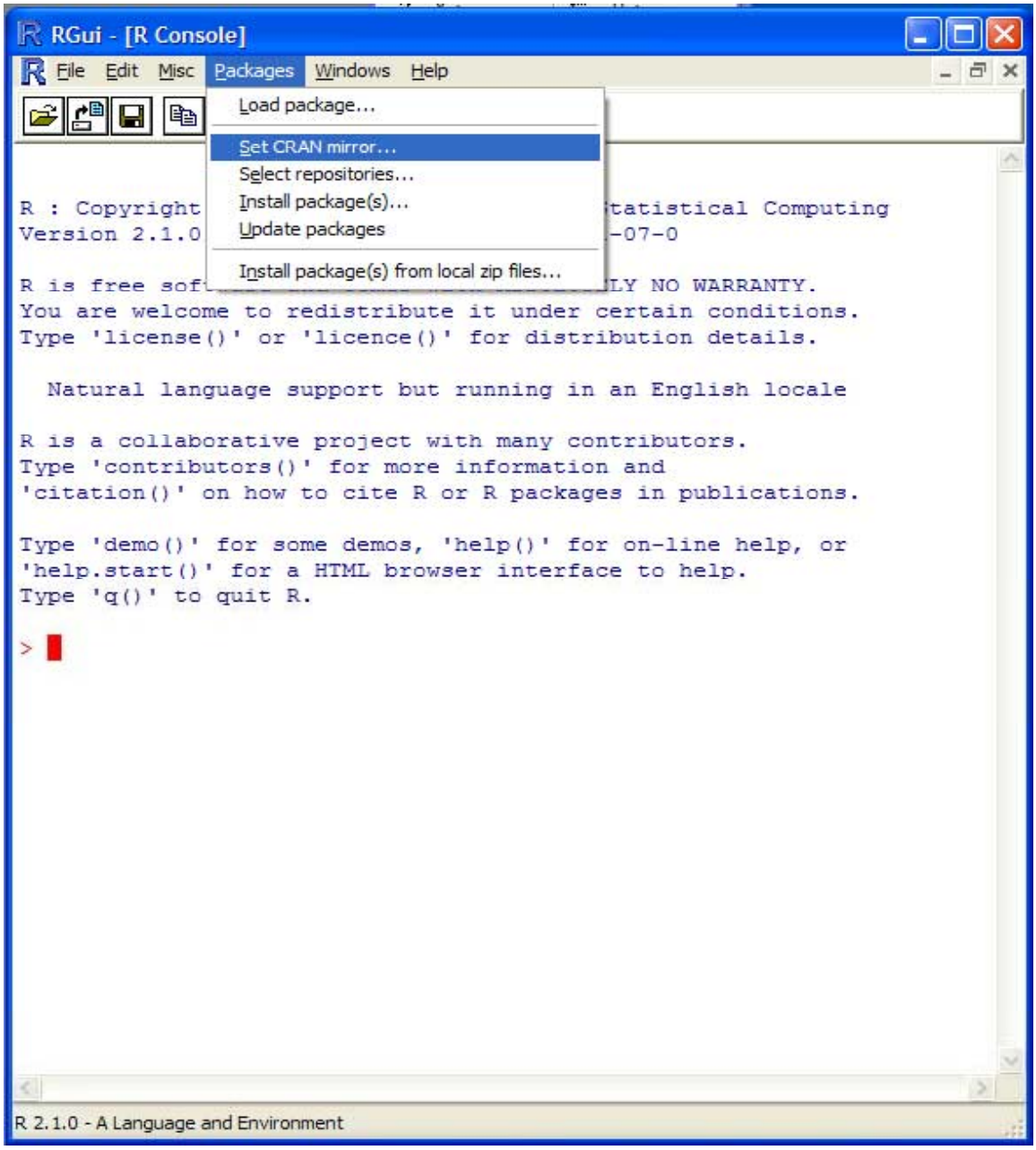
1. Download the Source Package
rama_1.3.0.tar.gz from
http://www.bioconductor.org/packages/bioc/1.8/html/rama.html or
bridge_1.3.1.tar.gz from

http://www.bioconductor.org/packages/bioc/1.8/html/bridge.html

2. Open a Terminal window. This can be found at
/Applications/Utilities/Terminal.
Navigate to the downloaded file. For instance: if you downloaded the
file to the desktop, type
cd /Users/yourusername/Desktop
and hit return. The prompt should now read
MyComputer:~/Desktop username$
To install the package, type
R CMD INSTALL rama_1.3.0.tar.gz (or bridge_1.3.1.tar.gz)

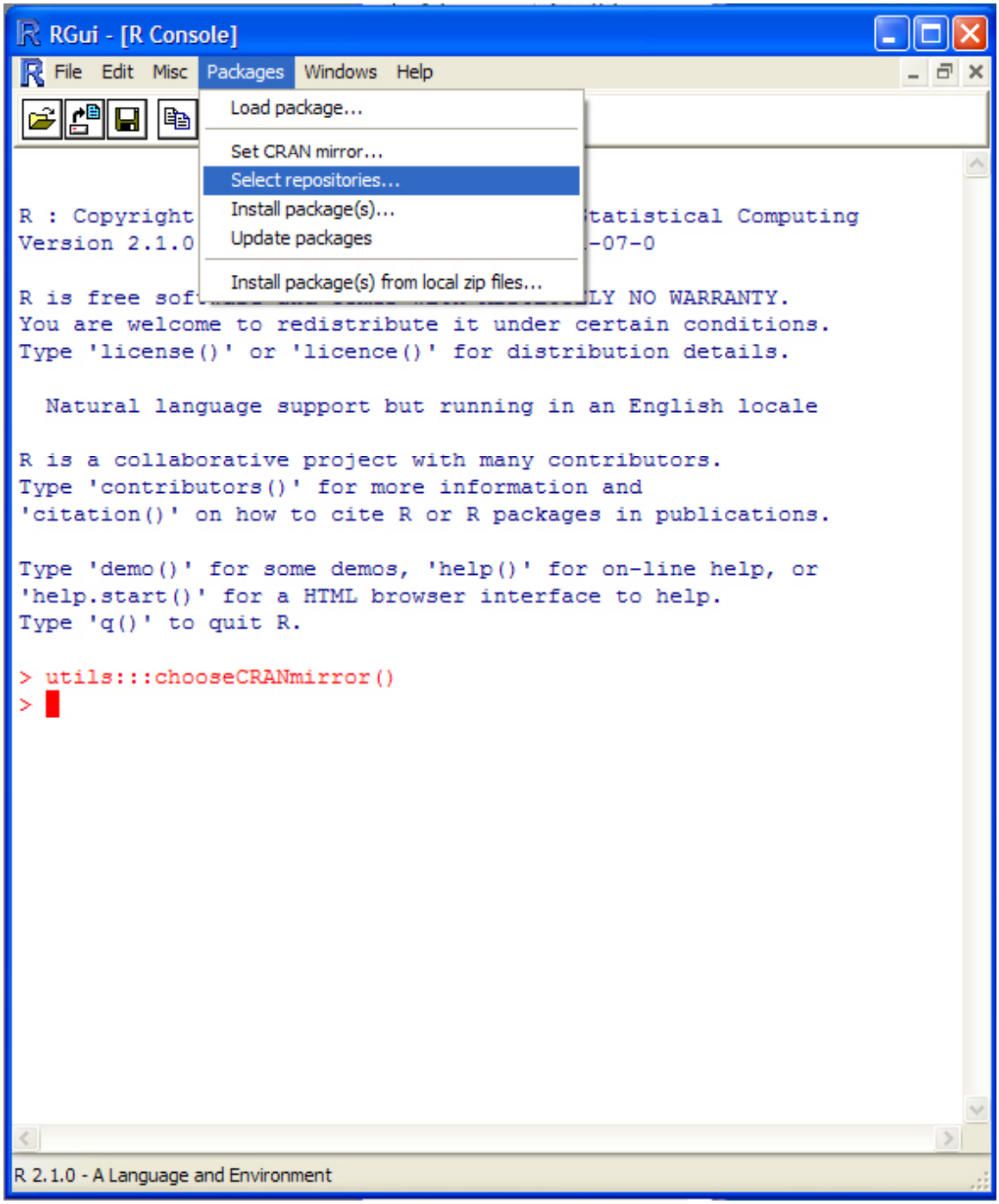
Installing under Windows (using the supplied Windows R GUI interface)

1. (Optional) Choose a Mirror. Click on
Packages > Set CRAN mirror...



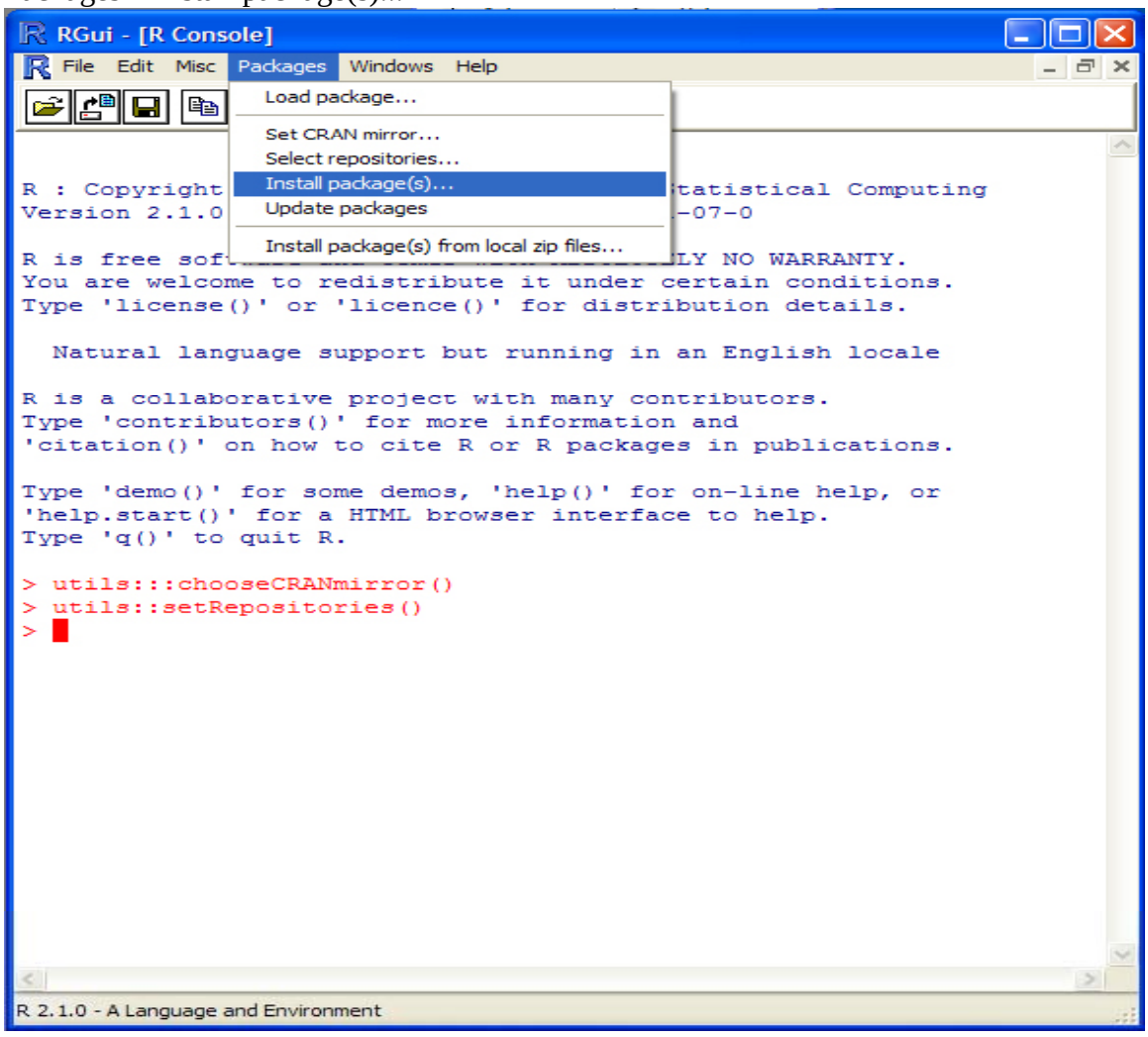
2. In the small dialog that appears, select a location

3. Choose a Repository. Click on Packages > Select Repositories...

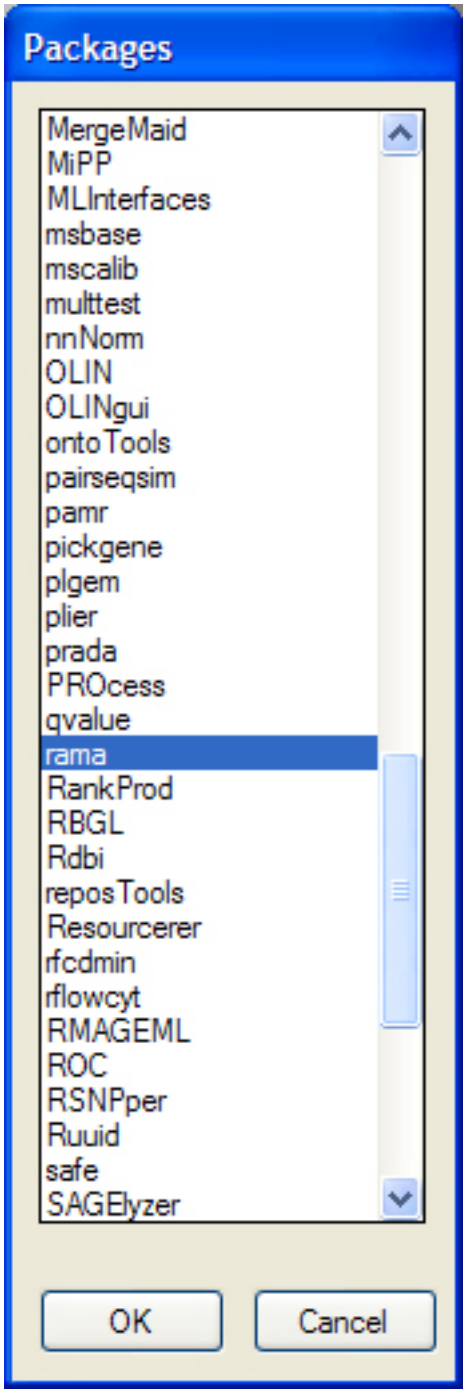


4. In the small dialog that appears, select a repository

5. Install the package. Click on
Packages > Install package(s)...



6. In the small dialog that appears, select *rama* or *bridge*



4. **Updating Rama/Bridge**

Updating under OS X (Command Line -- RECOMMENDED)

1. Download the Source Package
   rama_1.3.0.tar.gz from
http://www.bioconductor.org/packages/bioc/1.8/html/rama.html or
   bridge_1.3.1.tar.gz from

http://www.bioconductor.org/packages/bioc/1.8/html/bridge.html

2. Open a Terminal window. This can be found at
/Applications/Utilities/Terminal.
Navigate to the downloaded file. For instance: if you downloaded the
file to the desktop, type
cd /Users/yourusername/Desktop
and hit return. The prompt should now read
MyComputer:~/Desktop username$
To install the package, type
R CMD INSTALL rama_1.3.0.tar.gz (or bridge_1.3.1.tar.gz)

Updating under OS X (Drag & Drop)

1. Download the Source Package
rama_1.3.0.tar.gz from
http://www.bioconductor.org/packages/bioc/1.8/html/rama.html or
bridge_1.3.1.tar.gz from
http://www.bioconductor.org/packages/bioc/1.8/html/bridge.html

2. Unpack the gzipped file.

3. Replace the old rama directory with your new one in
'/Library/Frameworks/R.framework/Versions/Current/Resources/library'

Note: You may have to chmod your file permissions depending on your
installation.

10. Updating under Windows

1. Download the Source Package
rama_1.3.0.zip from
http://www.bioconductor.org/packages/bioc/1.8/html/rama.html or
bridge_1.3.1.zip from
http://www.bioconductor.org/packages/bioc/1.8/html/bridge.html

2. Unzip the zipped file.

3. Replace the old rama or bridge directory with your new one. The
location depends on where you installed R.
By default it would be 'C:\Program Files\R\R-x.y.z\library'

5. **Installing Rserve**

11. Installing under OS X

1. Download Rserve from
http://stats.math.uni-augsburg.de/Rserve/down.shtml. You will want to
get the Current version (Rserve_0.3-17.tar.gz at time of writing)

2. Open a terminal window and navigate to the directory containing the
downloaded Rserve file. Type R CMD INSTALL Rserve_0.3-16.tar.gz

12. Installing under Windows

1. Download Rserve from
http://stats.math.uni-augsburg.de/Rserve/dist/rserve-win.html.

2. Copy the downloaded file to the same directory where R.dll is located
(by default C:\Program Files\R\rw1080\bin)

Note: The windows version of Rserve is NOT RECOMMENDED. It suffers from
namespace issues namely that parallel connections are not supported.

6. **Running Rserve**

Running under OS X

1. Open a terminal instance. Type R CMD Rserve

Running under Windows

1. Double click Rserve.exe

13. Appendix: Bayesian Network & Literature Mining supporting files description:

Using Support Files created for standard arrays

- **Support file FTP Location:** ***Human, Mouse & Rat***
  - ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/Resourcerer/Human
  - ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/Resourcerer/Mouse
  - ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/Resourcerer/Rat
- **File Naming Conventions:**
  - All BN/LM related files ends with *_BN.zip.
    - E.g.: **affy_HG-U133_Plus_2_BN.zip**
  - All files start with array/chip vendor name, affy for Afymetrix
    - E.g.: **affy_HG-U133_Plus_2_BN.zip**
  - Vendor name is followed by chip/array name
    - E.g.: **affy_HG-U133_Plus_2_BN.zip**
- **Contents of zip files:**
  - All zip files contain 6 files
    1. affyID_accession.txt
    2. res.txt
    3. symArtsGeneDb.txt
    4. symArtsPubmed.txt
    5. all_ppi.txt
    6. gbGO.txt

- **Steps to use the pre-designed support files. Example array chosen for illustration is Affymetrix Human U133 Plus 2. To use any array follow the steps below:**
  1. Download your species & array specific file from the FTP location mentioned above. E.g. **affy_HG-U133_Plus_2_BN.zip**
  2. Extract the contents of the zip under the following MeV directory: ./data/BN_files/
  3. Once extracted, a folder by the array name will be created. In this case if the example file was downloaded, the following location will now exist: ./data/BN_files/affy_HG-U133_Plus_2_BN
  4. Verify all 6 files exist.
  5. From Mev launch LM or BN module.
  6. In the start-up dialogue make sure the 'File(s) Location' box points to the folder where the supporting files are downloaded for the species and array concerned. If the example array was chosen, the text box should point to ./data/BN_files/affy_HG-U133_Plus_2_BN folder.
  7. Now you are ready to start the algorithm.

Rules for creating custom BN/LM support files:

- All files should reside under the same directory. This directory should be chosen as "File(s) Location" during analysis as shown in BN/LM analysis section.
- All file names should be exactly as shown here.
- All column names (headers) should be named exactly as described here.

- Some files may not have column headers/names viz. symArtsGeneDb.txt, symArtsPubmed.txt, all_ppi.txt.
- All columns with or without headers should appear in the exact same order as shown here.

NOTE: Examples of all the files described below are provided at the following location: ./data/BN_files, where '.' Represents the directory where MeV is installed. The supporting files are created based on the dataset provided in the file ./data/BN_files/affy_U133plus2_dataset.txt

Files for BN and LM module:

- **affyID_accession.txt**

	A	B
1	affy_U133Plus2	
2	**Probe ID**	**Genbank Acc**
3	1007_s_at	U48705
4	1053_at	M87338
5	117_at	X51757
6	121_at	X69699
7	1255_g_at	L36861
8	1294_at	L13852
9	1316_at	X55005
10	1320_at	X79510
11	1405_i_at	M21121
12	1431_at	J02843
13	1438_at	X75208
14	1487_at	L38487
15	1494_f_at	M33318

Notes:

1. Both columns are mandatory along with the column names.
2. The entry in the first line is mandatory but its content can be anything that describes the source of the data.
3. "Genbank Acc" contains Genbank Accession numbers associated with each "Probe ID".
4. "Probe ID" here represents Affymetrix probe ids. This ID can be of any other type depending on the source of the data but it must follow these rules:
    - The entry in this column must uniquely identify each row in the loaded dataset.
    - Associated "Genbank Acc" information must be provided.
    - The entry in this column must match the "Probe ID" entry in the **res.txt** file.

- **all_ppi.txt**

	A
1	ACP1 pp SFMBT1
2	ACTB pp ACTB
3	ACTB pp ACTG1
4	ACTB pp CFL1
5	ACTB pp CFL2
6	ACTB pp DSTN
7	ACTG1 pp ACTG1
8	ACTG1 pp CFL1
9	ACTG1 pp CFL2
10	ACTG1 pp DSTN
11	ACTN4 pp PDLIM1
12	ACTN4 pp MYOZ2
13	ACTN4 pp MYOZ1
14	ACTL6A pp EWSR1
15	ACTN1 pp MAGEA11

Notes:

1. This file is only required if protein-protein interaction is selected as the source of priors for network.
2. This file does not have a column header/name
3. This file has just 1 column
4. Each entry represents a protein-protein interaction (pp) where X pp Y means an interaction between gene X & gene Y.
5. Gene X and Y are represented by their official gene symbol.

- **gbGO.txt**

	A	B	C	D	E	F
1	affy_U133Plus2					
2	**Genbank Acc**	**GO**				
3	U48705	GO:0000166(nucleotide binding) GO:0004672(protein kinase activi				
4	M87338	GO:0000084(S phase of mitotic cell cycle) GO:0000790(nuclear ch				
5	X51757	GO:0000074(regulation of progression through cell cycle) GO:000				
6	X69699	GO:0003677(DNA binding) GO:0003700(transcription factor activi				
7	L36861	GO:0005509(calcium ion binding) GO:0007600(sensory perceptior				
8	L13852					
9	X55005	GO:0003677(DNA binding) GO:0003700(transcription factor activi				
10	X79510					
11	M21121					
12	J02843	GO:0004497(monooxygenase activity) GO:0005506(iron ion bindir				
13	X75208	GO:0000166(nucleotide binding) GO:0004672(protein kinase activi				
14	L38487					
15	M33318	GO:0004497(monooxygenase activity) GO:0005506(iron ion bindir				

Notes:

1. This file is only required if GO Terms are selected as an option in the initial BN or LM window for directing network edges (which is not very mature at this stage). ***DFS is strongly suggested as an alternative***.
2. This file has 2 columns which have same name and order as shown in the sample.
3. The "Genbank Acc" column contains genbank accession numbers associated with each probe in the cluster selected for analysis or all the genbank accession numbers associated with all the probes in the entire dataset.
4. The "GO" column contains multiple GO entries separated by whitespace associated with each genbank accession number.
   E.g.: GO:000015(nucleotide binding) GO:0000016(protein kinase)

- **res.txt**

	A	B	C	D	E	F	G	H	I	J	K	L
1	affy_U133Plus2											
2	**Probe ID**	Clone Name	**Genbank Acc**	UniGene ID	**EntrezGene ID**	**Gene Symbol & Name**	**Gene Synonyms**	Human TC	Human GC	**RefSeq Acc**	**TC PubMed Ref**	**GO**
3	1007_s_at		U48705					THC2477693	GC6494			GO:0000
4	1053_at		M87338					THC2468303	GC7718	NM_181471		GO:0000
5	117_at		X51757					THC2462924	GC1853			GO:0000
6	121_at	PROST2013	X69699	Hs.469728	7849	PAX8; paired box gene 8	-	THC2461623	GC2112	NM_003466	NM_013992	GO:0003
7	1255_g_at	TESTI205362	L36861					THC2471218	GC6860	NM_000409		GO:0005
8	1294_at		L13852	Hs.16695	7318	UBE1L; ubiquitin-activating	UBE2 D8 MGC12713	THC2462145	GC3818	NM_003335		
9	1316_at		X55005	Hs.724	7067	THRA; thyroid hormone rec	NR1A1 THRA1 ERBA	THC2480060	GC15863	NM_199334		GO:0003
10	1320_at		X79510	Hs.437040	11099	PTPN21; protein tyrosine pl	PTPD1 PTPRL10	THC2470577	GC13801	NM_007039		
11	1405_i_at		M21121	Hs.514821	6352	CCL5; chemokine (C-C mo	TCP228 SISd MGC1	THC2465035		NM_002985		
12	1431_at		J02843					THC2461325	GC10299	NM_000773		GO:0004
13	1438_at		X75208	Hs.2913	2049	EPHB3; EPH receptor B3	TYRO6 ETK2 HEK2	THC2464788	GC4410	NM_004443		GO:0000
14	1487_at		L38487	Hs.110849	2101	ESRRA; estrogen-related r	NR3B1 ERRalpha E	THC2468691	GC11101	NM_004451		
15	1494_f_at		M33318	Hs.439056	1548	CYP2A6; cytochrome P45	CPA6 P450C2A P4	THC2482745	GC17757	NM_000762		GO:0004

Notes:

1. There are 12 columns
2. All columns should appear in the exact order as shown.
3. All columns colored **red** must have appropriate entries as indicated by the column names.
4. All columns colored **blue** may be empty, but the columns should be present as empty columns.
5. "Probe ID" here represents Affymetrix probe ids. This ID can be of any other type depending on the source of the data but it must follow these rules:
   - The entry in this column must uniquely identify each row in the loaded dataset.
   - Associated "Genbank Acc" information must be provided.
   - The entry in this column must match the "Probe ID" entry in the **affyID_accession.txt** file.
6. "Clone Name" column is mandatory but it can be empty.
7. "Genbank Acc" contains Genbank Accession numbers associated with each "Probe ID".
8. "UniGene ID" column is mandatory but it can be empty.
9. "EntrezGene ID" contains entrez gene Ids (previously locuslink ID) wherever available that best represents the "Probe ID".
10. The Gene Symbol & Name column has Gene Symbol and description. They are separated by the delimiter ";".
11. "Gene Synonyms" are aliases for the official gene symbol separated by whitespace.
12. "Human TC" & "Human GC" columns are mandatory but can be empty.
13. "RefSeq Acc" column contains mRna RefSeq ID wherever available that best represents the "Probe ID".
14. "TC PubMed Ref" column contains?????????
15. "GO" column is formatted in the same way as described in the previous file's description (#3).

- **symArtsGeneDb.txt**

	A	
1	LOC255326	
2	EPN1	10557078,9723620,14702039,12477932,12353027
3	C14orf132	12477932
4	SF3B4P	
5	LOC138864	
6	OR10G1P	8188290,9110172
7	LOC388695	
8	LOC158135	
9	TMSL1	
10	FLJ40365	14702039,12477932
11	IDDM4	8733139,7842018,8072542,8072544
12	LOC389457	
13	TGFBR1	12607775,12145287,14597484,11820800,9661882,12060054,9417915,8242743,14704634,12082094,12446693,1319842,8530052
14	VTI1B	9446565,9636656,12853575,12477932
15	ZNF189	8889548,9653648

Notes:

1. There are 2 columns
2. There are no column headers/names
3. The columns should appear in the exact order as shown.
4. The 1st column contains official Gene Symbol
5. The 2nd column contains PubMed IDs. **The PubMed IDs are obtained by querying the EntrezGene annotation for the given gene.**
6. The 2nd column contains multiple values which are separated by the delimiter ",".

- **symArtsPubmed.txt**

	A	
1	CARD4	[15790594, 15718249, 15771576, 15703626, 15703577, 15839897, 15725060, 15802263, 15845503]
2	CPSF3	[15358515, 15684398, 15550246]
3	TACC2	[15207008, 15304323, 14603251, 14767476, 15675572, 12711550, 12684693, 15226440]
4	RTN3	[15461990, 15765506, 15541434, 15799019, 15858203]
5	CORO1A	[15819430, 15869390, 15823281, 15851765]
6	RBM17	[15526362, 15050977, 14578179]
7	SYCP2	[15730934, 15870106]
8	SULF1	[15080891, 12686563, 14699503, 12368295, 12419802, 15817123, 14973553]
9	PHYH	[12694175, 14672712, 15694837, 12923223, 14713215, 14970690, 14974078, 14713244, 15109262, 12767919, 14713238, 15102880]
10	TOP2B	[15526362, 11528129, 9795238, 7783737, 12197834, 8863738, 15241656, 15543237, 9380682]
11	ATP9A	[9734811, 9838086]
12	NUDT5	[12370179, 10373642, 10722730, 10567213, 12717453]
13	IRF7	[15585412, 15695821, 15743772, 15492278, 15377465, 15842376, 15767254, 15800576, 15851485, 15767370, 15650197, 15664159, 15664995, 15749911]
14	TGFBR1	[15845540, 15860866, 15770714, 15843168, 15797633, 15769863, 15766759, 15850837, 15775557, 15775675, 15833881, 15867212, 15840021]
15	VTI1B	[15797025, 15371541, 15133481, 15640147]

Notes:

1. There are 2 columns
2. There are no column headers/names
3. The columns should appear in the exact order as shown.
4. The 1st column contains official Gene Symbol
5. The 2nd column contains PubMed IDs. **The PubMed IDs are obtained by querying Pubmed abstracts (text query) for the gene symbol, and then collecting all the PubmedIDs of those abstracts that match that gene symbol**
6. The 2nd column contains multiple values which are separated by the delimiter ",".

- **Sample Label file for BN analysis:**

Notes:

1. Once samples are assigned to classes in the classification window, shown (here), the classification can be saved into a text file by clicking the "Save Settings" button.
2. The file names could be anything the user chooses with any extension the user wants. We suggest using .txt as an extension.

3. On subsequent analysis of the same data with same sample classification, the class assignments can be loaded automatically from the saved file by clicking the "Load Settings" button and selecting the previously saved file.

14. Appendix: MeV Dependencies

MeV interacts with several other software packages that enhance its capabilities.

Gaggle Boss

MeV can connect to the Gaggle interaction network through the Gaggle Boss, a webstart application available from the Institute for Systems Biology. MeV requires an internet connection to launch the Boss for the first time, but can cache the application for later use without an internet connection.

Cytoscape

15. License

16. Contributors

J. Craig Venter Institute

Alexander I. Saeed, John Braisted, Wei Liang, Jerry Li, Vasily Sharov, Mathangi Thiagarajan, Chun-Hua Wan

Dana-Farber Cancer Institute / Harvard School of Public Health

Eleanor Howe, Sarita Nair, Daniel Schlauch, Raktim Sinha, John Quackenbush, Weidong Wang, Joseph White

University of Washington

Vu Chu, Annie Liu, Raphael Gottardo, Ka Yee Yeung, Roger Bumgarner

Institute of Biomedical Engineering, Graz University of Technology

Alexander Sturn, Zlatko Trajanoski

DataNaut, Inc.

Mark Snuffin, Aleksey Rezantsev, Dennis Popov, Alex Ryltsov, Edward Kostukovich, Igor Borisovsky

Syntek Systems Corporation, Inc.

Stu Golub, Zaigang Liu, Jane Ruan, Minhas Siddiqui

formerly, The Institute for Genomic Research

Nirmal Bhagabati, Tracey Currier

The George Washington University

Patrick Cahan, Tim McCaffrey

National Center for Genome Resources

Todd Peterson

Fox Chase Cancer Center

Luke Somers

Center for Computational Genomics and Bioinformatics, University of Minnesota

Jim Johnson, Ernest Retzel

National Institute of Allergy and Infectious Disease, NIH, Laboratory of Immunopathogenesis and Bioinformatics.

Glynn Dennis, Douglas Hosack, Richard Lempicki, Wei Gao

Independent Development

Eric Albert

University of Texas, MD Anderson Cancer Center
Sally Gaddis

FDA, National Center for Toxicological Research
Stephen C. Harris

17. References

These references and links to their PubMed records can be found in the main MeV toolbar under *About→Papers/Publication Reference*.

Aryee, M., J. Gutierrez-Pabello, I. Kramnik, T., Maiti, J. Quackenbush 2008. An Improved Empirical Bayes Approach to Estimating Differential Expression in Microarray Time-Course Data: BETR (Bayesian Estimation of Temporal Regulation). *BMC Bioinformatics*. In Press.

Bar-Joseph, Z., D.K. Gifford, T.S. Jaakkola (2001) Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* Vol. 17 no. 90001 2001:S22-S29.

Ben-Dor, A., R. Shamir, and Z. Yakhini 1999. Clustering gene expression patterns. Journal of Computational Biology 6:281-297.

Breitling, R., P. Armengaud, A. Amtmann, P. Herzyk 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. Federation of European Biochemical Societies. 83-92.

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.

Brunet, J-P., Tamayo, P., Golub, T.R., and Mesirov, J.P. 2004. Metagenes and molecular pattern discovery using matrix factorization. */Proc. Natl. Acad. Sci. USA/* 101(12):4164Â–4169.

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.

Chittenden T.W, Howe, E.A., Taylor, J., Mar, J., Aryee, M., Braisted, J., Nair, S.J., Quackenbush, J., Holmes, C. Molecular Subtype-specific Gene Ontology Classification in Human Breast Cancer. In Preparation

Chu VT, Gottardo R, Raftery AE, Bumgarner RE, Yeung KY. MeV+R: using MeV as a graphical user interface for Bioconductor applications in microarray analysis. Genome Biol. 2008;9(7):R118.

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.

Devarajan K, 2008 Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology. PLoS Comput Biol 4(7): e1000029. doi:10.1371/journal.pcbi.1000029

Djebbari, A., Quackenbush, J. In Review. Seeded Bayesian Networks: constructing genetic networks from microarray data.

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.

Goeman, J.J., 2010. L(1) penalized estimation in the cox proportional hazards model. Biometrical Journal. Biometrische Zeitschrift, 52(1), 70-84.

Gottardo R, Raftery A.E, Yeung K.Y, and R.E. Bumgarner, Quality Control and Robust Estimation for cDNA Microarrays with Replicates, accepted for publication in (Journal of the American Statistical Association)

Gottardo R, Raftery AE, Yee Yeung K, Bumgarner RE. 2006. Bayesian robust inference for differential gene expression in microarrays with multiple samples. Biometrics. 62(1):10-8.

Graur D, Li W. Fundamentals of Molecular Evolution. 2nd ed. Sinauer Associates; 2000.

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.

Hollander and Wolfe. 1999. Nonparametric Statistical Methods.Wiley-Interscience. ISBN: 978-0471190455.

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.

Jiang, Z., Gentleman, R., 2007. Extensions to gene set enrichment analysis. Bioinformatics., 23(3):306-13.

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.

Lee DD, Seung SH (2001) Algorithms for nonnegative matrix factorization. Adv Neural Inform Process Syst 13: 556–562.

Manly, B.F.J. 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. 2nd ed. Chapman and Hall / CRC , FL.

Margolin AA, Greshock J, Naylor TL, Mosse Y, Maris JM, Bignell G, Saeed AI, Quackenbush J, Weber BL. 2005 CGHAnalyzer: a stand-alone software package for cancer genome analysis using array-based DNA copy number data. *Bioinformatics.* 21(15):3308-11.

Myers, C. L., Dunham, M. J., Kung, S. Y., Troyanskaya, O. G., 2004. Accurate detection of aneuploidies in array cgh and gene expression microarray data. Bioinformatics 20 (18), 3533-3543.

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

Saeed AI, Sharov V, White J, et al. TM4: a free, open-source system for microarray data management and analysis. BioTechniques. 2003;34(2):374-378.

Saeed AI, Bhagabati NK, Braisted JC, et al. TM4 microarray software suite. Meth. Enzymol. 2006;411:134-193.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, No. 1, Article 3.

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.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**.** Proc. Natl. Acad. Sci. USA 102, 15545-15550.

Tamayo, P., D. Slonim, J. Masirov, Q. Zhu, S. Kitareewan, 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 KY, Bumgarner RE. 2003 Multiclass classification of microarray data with repeated measurements: application to cancer. Genome Biol. 4(12):R83.

Yeung KY, Bumgarner RE. 2005 Correction: Multiclass classification of microarray data with repeated measurements: application to cancer. Genome Biol. 6(13):405.

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.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3, No. 1, Article 3.

Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397-420.